

Development of a Multimodal AI Visual Chatbot Using Python for Enhanced Human-Computer Interaction

Nancy digra

Assistant professor
Chandigarh University
Mohali, India
Nancy.e20071@cumail.in

Mohammad Mustafa

dept. of CSE, UIE
Chandigarh University
Mohali, India
23BCS10077@cuchd.in

Kabir Choudhury dept. of

CSE, UIE Chandigarh
University Mohali, India
23BCS12820@cuchd.in

Ritesh Kumar

dept. of CSE, UIE
Chandigarh University
Mohali, India
23BCS10416@cuchd.in

Karan Singh

dept. of CSE, UIE
Chandigarh University
Mohali, India
23BCS12461@cuchd.in

Abstract—Visual chatbot is considered the future generation of the computer-human interface technology that goes beyond the boundaries of existing text-oriented designs. In this research, we introduce AI Visual Chatbot v2.0 – the innovative communication tool that can understand prompts in different modalities such as high-quality images, multi-frame videos, and natural language. Unlike previous models constrained by fixed text input types, AI Visual Chatbot v2.0 uses Google Gemini Vision API to extract deep understanding of the visual world. Built on the basis of the Python language with an appropriate NLP process pipeline, which includes data cleaning, stemming, and normalization, the system comes with a dedicated panel of analysis for statistics and objects in media. It proves its effectiveness in various fields, namely academic assistance, industrial automation, and accessibility applications, and provides the user with a "human-like" communication through efficient GUI based on Tkinter library.

Keywords—Large Multimodal Models (LMM), Google Gemini Vision, Computer Vision, Natural Language Processing (NLP), Multi-frame Video Analysis, Human-Computer Interaction (HCI).

I. INTRODUCTION

In essence, chatbots are software applications designed to interact with people in a way akin to humans through natural language processing technologies. Historically, chatbots were characterized by a reliance on rules and predetermined interactions for generating conversations. However, the evolution of artificial intelligence (AI) has led to contemporary chatbots that can recognize context, meaning, and semantics, thus responding dynamically rather than in a predefined manner [1]. Chatbots may be described as conversational software programs that facilitate conversations in the form of text or voice. Generally, chatbots are popular in fields ranging from customer support to education because of their ability to provide instantaneous feedback and round-the-clock availability. For instance, within academia, chatbots can play the role of virtual assistants offering students assistance regarding educational matters for an entire academic period [2]. As AI advances continually, chatbots have changed significantly from rule-

based applications to intelligent systems built using machine learning algorithms and deep learning models. Currently, features such as personalization, contextual awareness, and the ability to interact with external applications characterize chatbot systems, making them highly adaptable and efficient. All these developments illustrate the significance of chatbots in improving HCI by building more effective ways of communication between users and machines [3]. Notably, one of the innovations in the field of chatbots includes their transition from natural language processing to natural conversation. Specifically, a multimodal chatbot is one that combines natural language processing with computer vision to enhance functionality. Through a conversational visual chatbot, users will not only be able to converse via written words but also through images and videos. In such a system, visual elements will be analyzed in the context of text using natural language algorithms in addition to computer vision, allowing object detection, text identification, and explanation [4]. The process results in the provision of additional insight by interpreting visual content for the benefit of the user, who can ask questions in a conversational format. Building an advanced chatbot of this kind involves coding processes using programming languages such as Python for the backend, alongside APIs and software frameworks for AI functions. Specifically, Python can enable developers to build applications involving natural language pre-processing, media analysis, and communication with external AI systems [5]. Similarly, APIs such as vision-language models allow programmers to incorporate powerful capabilities in their applications such as image recognition and contextual understanding without developing new models. There are many benefits of chatbots, among which are real-time access to information, the ability to automate repetitive work, and improved user experience. By engaging in a dialogue, people will be able to easily find information needed without going through complicated structures. Chatbots can be especially useful for people with disabilities, and those having some difficulty with language, because chatbots have the ability to adjust to particular users, making access to information and services more inclusive [6]. As for the architecture of chatbots, there are essentially two main ways for producing responses. Traditional chatbots make use of rules and templates that are quite limited and cannot perform well in all

situations. AI chatbots, on the other hand, utilize neural networks, which have been trained on huge amounts of data, to produce more accurate and meaningful responses [7]. It is clear that chatbot technology development takes into consideration the multi-disciplinary nature of this phenomenon, as there are many areas contributing to this field. Among them are computer science, linguistics, psychology, human-computer interaction, and others. This makes it possible to constantly enhance chatbot capabilities and improve users' experience [8]. Regarding multimodal chatbots, it appears that once the user's data has been received by backend, it is processed based on natural language processing algorithms and used to produce answers to user's requests. At the same time, multimodal chatbot systems analyze and take into account not only textual data but also images as well [9]. Nevertheless, despite all the advancements made so far in chatbot technology, many obstacles in creating human-like interaction remain unsolved. One of the biggest problems is related to complexity and ambiguity of natural language and the need to achieve accuracy and proper context comprehension [10].

II. LITERATURE REVIEW

Advancements made in the field of artificial intelligence have led to tremendous improvements in developing chatbot systems that understand various types of data besides text. The modern research trend includes the implementation of both natural language processing and computer vision to analyze both textual and multimedia information [11]. In 2026, the research has concentrated on multimodal large language models that support real-time reasoning and cross-modal understanding. This type of models allows for combining textual and visual input in order to create a correct and informative output. The implementation of vision-language models into chatbots has enabled them to analyze pictures, find objects on them, recognize text, and provide insightful feedback. Such chatbots may be used in different fields like education, security, surveillance, etc. [12]. The development of AI chatbots with the implementation of Gemini and other vision language models allows for interacting with text messages and images without any issues. Advanced AI chatbots use neural networks and transformer algorithms to work with multimodal data effectively. Consequently, chatbots can do more complicated operations including scene analysis, image question answering, and video frames analysis [13]. One more topic that concerns the improvement of chatbots involves increasing efficiency of their real-time performance. Efficient response generation is important for maintaining an interactive session between chatbot and user. It is necessary to emphasize that recent advances in API optimization allow chatbots to generate a fast response to user query [14]. Another important aspect that is considered in research involves conversational memory implementation in chatbots. The implementation of conversational memory enables chatbots to interact with users during several consecutive turns. Such approach may be helpful when creating educational chatbots or digital assistants [15]. The development of chatbots also involves application of various NLP pre-processing techniques, including text cleaning, tokenizing, stemming, and normalizing. All of these pre-processing techniques help in preparing input data for further processing to increase its

clarity and correctness. Combination of these pre-processing techniques and vision language models contributes to increased efficiency [16]. Moreover, image and video analysis are essential parts of contemporary chatbots. It has been shown in the research that frame extraction from video and processing of individual images from videos leads to improved results of understanding the dynamics of video content. Consequently, chatbots are able to analyze videos and extract their essence [17]. Security and privacy are essential concerns in the development of AI-based chatbots. Since these systems use huge volumes of user data, it is important to ensure secure data storage and confidentiality. Modern studies focus on the implementation of APIs to protect data in these systems [18]. User feedback and learning capabilities are two other characteristics that help to improve the functionality of chatbot systems. The study of user interaction and feedback is an important tool for the improvement of chatbots, allowing them to learn from experience and provide better performance [19]. However, there are many challenges associated with the development of efficient chatbots, such as dealing with ambiguous queries, the preservation of accuracy, and the problem of scalability. Researchers seek to reduce computational costs required to run models used by chatbots and improve their efficiency [20]. The modern architecture of chatbots includes multiple stages, such as the processing of input data, natural language understanding, the stage of making decisions based on the data received, and generating responses. Moreover, when working with multimodal chatbots, the architecture can be expanded to account for visual information as well [21]. Recent developments related to the implementation of hybrid models in chatbots help to address some of the challenges associated with these systems. These models allow combining rule-based algorithms and machine learning models, thus providing chatbots with high reliability and flexibility at the same time [22]. Edge computing has attracted much attention during the last years, and researchers seek to implement optimization tools in order to make chatbots suitable for running on low-resource devices. Such an approach will allow increasing access to advanced chatbot technologies [23]. Moreover, the development of sentiment and emotion recognition technology allowed chatbots to improve user experience by identifying user emotions and responding to them accordingly. This aspect increases the quality of communication between humans and computers [24]. Thus, recent trends in the area of chatbots development show that these technologies are becoming more intelligent, multifunctional, and able to process and interpret more complicated types of input data. These changes create the foundation for building chatbot v2.0 based on NLP and vision processing [25]. Comparison of different research models have been analyzed and summed up in Table I.

Reference no.	Year	RESEARCH GAP
[11]	2026	Multimodal chatbot models proved highly accurate in processing complex input, but the sheer amount of computational power required for such models was prohibitive in terms of standard hardware.
[12]	2026	Vision-language systems managed to close the semantic gap regarding images, but their design was not optimized enough to deploy seamlessly on real-time basis.

[13]	2025	Such advanced AI models offered highly efficient contextual analysis, but the excessive dependency on external, high-latency APIs severely hampered their speed.
[14]	2025	Although low-latency frameworks greatly decreased the initial latency of the system, scalability issues appeared once concurrent uploading of media by multiple users was involved.
[15]	2024	Conversational memory architecture greatly improved contextual handling during multi-turn interactions but also increased the complexity and maintenance burden on systems considerably.
[16]	2024	Application of strict natural language processing preprocessing methods improved the intent accuracy significantly, but it also added an extra step to data cleaning process.
[17]	2023	Video-focused chatbot architecture allowed for multi-frame processing, but such systems had great difficulty dealing with temporal issues such as loss of tracking of scenes in longer videos.
[18]	2023	Strictly focused on protecting user privacy, the framework was successful in providing such protection but made it difficult to use any collected data to train and update the AI.
[19]	2022	First attempts at learning-based chatbot architecture offered better user-adapting potential but needed continuous training process which hampered their performance severely.
[20]	2021	Early AI chatbot systems of this period were inherently bound to text communication only and did not allow for more complex context-awareness or multimodal approach.

TABLE I. SURVEY OF RESEARCH MODELS

Table 1 above illustrates different models used in previous research together with the drawbacks associated with each of them. The following statement holds true from the table below – although there have been considerable advancements in chatbots regarding accuracy, multimodal capabilities, and interaction skills, there are still some limitations such as computational difficulties, scalability issues, privacy issues, and performance issues to contend with. This problem is solved in the Visual Chatbot v2.0 system proposed here.

III. PROPOSED MODEL

Introduction A multimodal AI-based visual chatbot will be created as part of the research project. It involves developing an intelligent visual chatbot that would process images and videos along with textual data for the purpose of improving human computer interactions. The application would employ Python programming language as a foundation that unites GUI development, data handling, and integration with various AI services. **System Architecture** In terms of its architecture, the proposed model would consist of three layers: user interaction, processing, and AI inference. The user interaction layer is created with the help of the Tkinter library. It provides users with the opportunity to communicate with the application through graphical interface. The processing layer deals with text preprocessing and media analysis. NumPy, Pillow, and OpenCV libraries will be utilized to conduct necessary transformations of visual and textual data. Finally, the AI inference layer is responsible for

performing reasoning in response to the input provided by the user. The model connects to Gemini Vision API and obtains the results of analysis. The main advantage of the proposed solution is the possibility to remember previous conversations and produce context-aware answers. Also, an analysis window would be included as a part of GUI that will display the features extracted from images and videos. **Working Process** The proposed AI-based chatbot will involve several operations related to the functioning of the system:

A. Input Acquisition:

The first step implies processing different types of input. Depending on the kind of input (text queries, images, or videos), the application will choose the proper pipeline.

B. Text Processing Pipeline:

In case of textual input, it is subject to certain transformations, including cleaning, tokenization, normalization, and semantic refinement. This procedure allows ensuring that the input is properly formed for the AI system.

C. Visual Data Processing:

If input is an image, then it is subject to preprocessing that includes extraction of some features such as resolution, edges, and colors using OpenCV and NumPy libraries. Similarly, when an input is a video file, then each frame will be extracted, analyzed, and converted into a form convenient for API request.

D. Multimodal Integration:

Processed textual and visual information should be united in order to enable cross-modal reasoning in the process of answering the user's query.

E. AI-Based Response Generation:

Finally, the integrated input will be transferred to the AI system and analyzed in depth by means of Gemini Vision API.

F. Output Presentation and Feedback:

Generated response and visual elements are provided in the chatbot interface with additional options to preview visual content and detect specific elements from the input. In addition, the system tracks previous user interactions for future follow-up questions. User interactions can help train the system further to enhance its performance and capabilities.

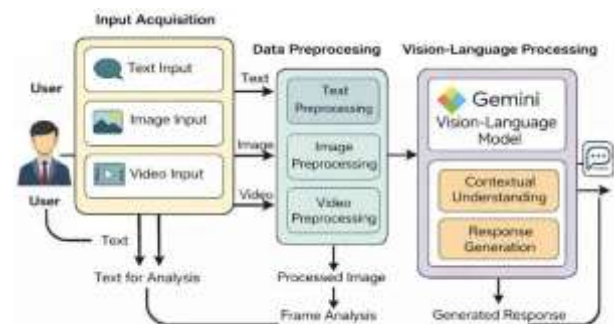


Fig. 1. Working of visual chatbot

Figure 1. Above is presenting the general workflow of the proposed multimodal visual chatbot. The user provides input in the form of text, image, or video through the graphical user

interface of the chatbot. According to the provided input, the system applies the appropriate preprocessing technique, such as natural language processing for text, computer vision for images and videos, including feature extraction and frame analysis. The processed information is forwarded to the vision-language processing component based on the Gemini model, where the system integrates both visual and linguistic data to obtain the contextual understanding of the input. Next, based on the results obtained, the chatbot produces a meaningful response, which is eventually communicated to the user.

IV. IMPLEMENTATION AND RESULTS

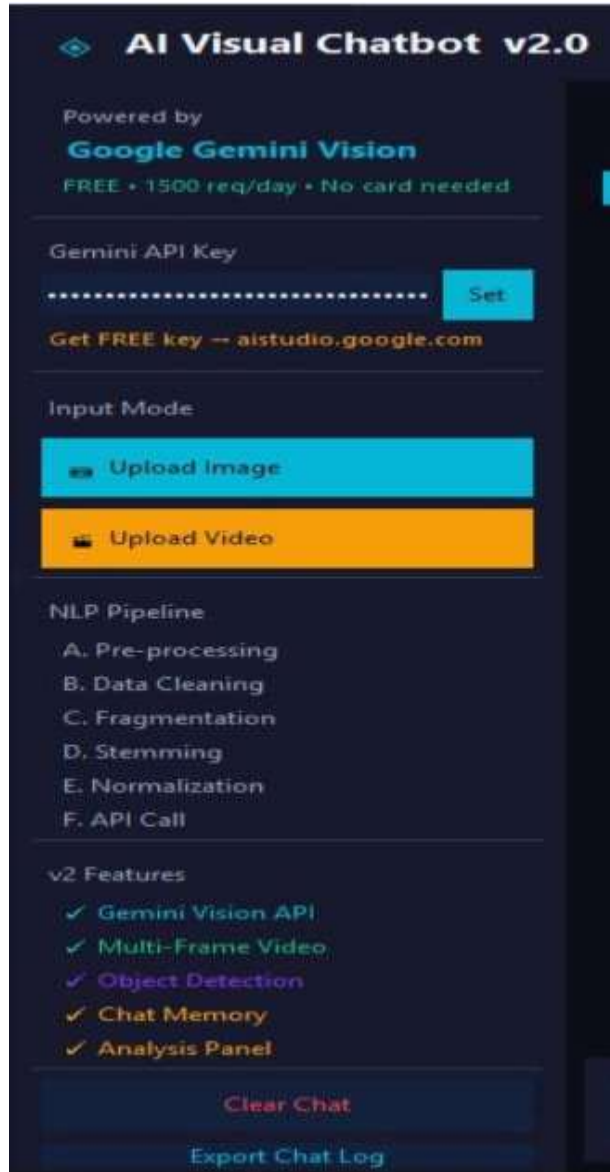


Fig. 2. Buttons of chatbot

Figure. 2. Above has presented the set of buttons available in the chatbot interface to manage different types of user inputs. Users can select one of the buttons to provide and analyze their image or video. When the user selects the desired option, the corresponding module is activated, allowing the chatbot to apply preprocessing and vision-language processing pipelines to the uploaded media.



Figure 3. Above is presenting the graphical user interface of the proposed multimodal visual chatbot, created using the Tkinter library. Users can communicate with the chatbot using text input and media uploads. Besides, the chatbot provides users with the opportunity to upload files and send queries to the system.

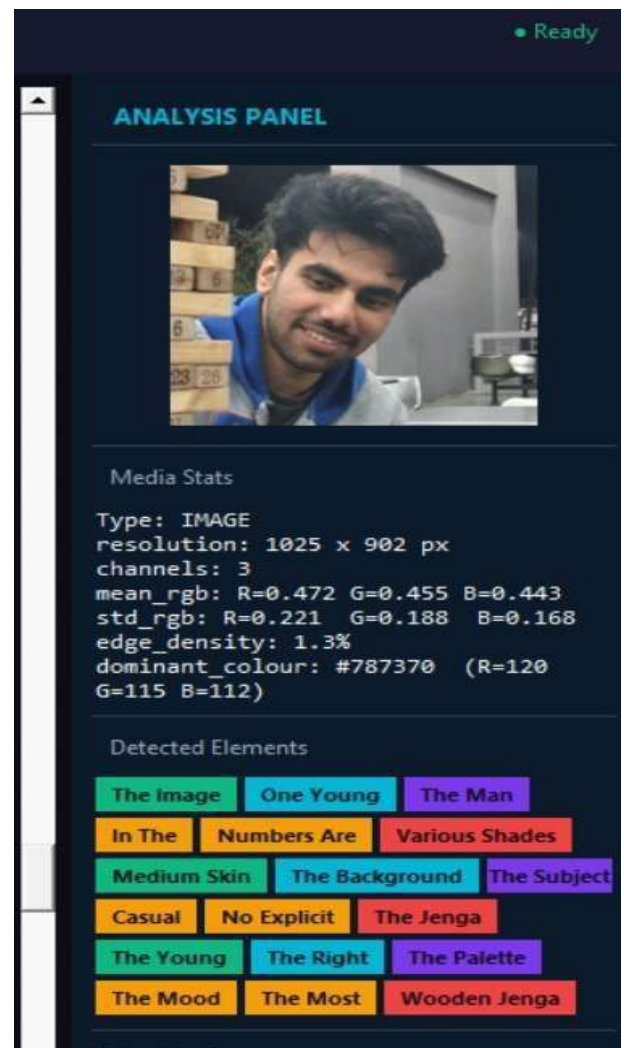


Fig 4. Result and Analysis

Figure 4. Above is presenting the response provided by the chatbot after analyzing the uploaded image or video. As shown in the results above, the chatbot offers a structured output with identified objects, scenes, text, and contextual

information. Instead of applying limited classification, the system provides comprehensive and meaningful insights using the Gemini vision-language model.

V. CONCLUSION AND FUTURE WORK

From this perspective, the proposed Visual Chatbot v2.0 showcases an efficient combination of NLP and computer vision technologies to create a human-computer interaction tool with multiple capabilities. First, unlike standard chatbots that are limited to text-based dialogue, the designed solution allows the analysis of images and videos and can generate context-aware responses. With the use of Python programming and Gemini vision-language model, the chatbot is able to respond meaningfully and instantly. In addition, its implementation involves such features as media preprocessing, multi-frame video analysis, conversation memory, and graphical user interface (GUI), making the proposed solution applicable in practice.

The key contributions and outcomes of the proposed system are listed below:

- Creation of a multimodal chatbot able to process textual queries, images, and video frames.
- Implementation of NLP preprocessing algorithms to improve input analysis.
- Processing of images and videos using OpenCV library and other technologies.
- Usage of Gemini Vision API to conduct complex visual reasoning tasks.
- Design of an interactive GUI based on Tkinter.
- Incorporation of conversation memory to provide context-aware responses.

Even though this system is capable of performing various complex tasks, it has some limitations that require addressing in the future. To increase its performance, the solution can be upgraded by implementing such features as:

- Real-time analysis of streaming video instead of static frames.
- Speech recognition and voice-based dialogue.
- Custom training of models to minimize dependency on Gemini API.
- Optimization of performance for running the solution on less powerful machines.
- Multilingual support.
- Emotion detection and personalization of responses.

These advancements will enable increasing the efficiency, scalability, and adaptability of the proposed chatbot solution.

REFERENCES

- [1] G. Zhao et al., "NavGemini: A Multi-Modal LLM Agent for Vision-and-Language Tasks," *Springer*, 2026.
- [2] Y. Li et al., "Mini-Gemini: Enhancing Multi-Modality Vision Language Models," *arXiv preprint arXiv:2403.18814*, 2024.
- [3] B. R. Kim et al., "Multimodal Large Language Models in Real-World Applications," *ScienceDirect*, 2025.
- [4] F. Lin, "Vision Language Models: A Survey of 26K

Papers," *arXiv preprint*, 2025.

- [5] L. Yan et al., "A Comprehensive Survey of Multimodal LLMs," *OpenReview*, 2025.
- [6] G. Rubegni et al., "Applications of Large Language Models in Vision-Based Systems," *PMC*, 2026.
- [7] Google DeepMind, "Gemini: Multimodal AI Model Capabilities," 2025.
- [8] Google, "Gemini Language Model Overview and Evolution," 2025.
- [9] OpenAI, "GPT-5: A Multimodal Large Language Model," 2025.
- [10] A. Pal and M. Sankarasubbu, "Evaluation of Gemini in Multimodal Medical Tasks," *arXiv*, 2024.
- [11] G. Polverini and B. Gregorcic, "Evaluating Vision-Capable Chatbots," *arXiv*, 2024.
- [12] S. Xu, "Multimodal AI: Vision-Language Models Guide," *BentoML Blog*, 2025.
- [13] N. Ahamed, "Rise of Multimodal AI Systems," *Medium*, 2025.
- [14] Google DeepMind, "Gemma 3 and Multimodal Open Models," 2025.
- [15] Mustafa Shukor et al., "Vision-Language-Action Models for AI Systems," 2025.
- [16] A. Vaswani et al., "Attention Is All You Need," *NeurIPS*, 2017.
- [17] J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers," *NAACL*, 2019.
- [18] T. Brown et al., "Language Models are Few-Shot Learners," *NeurIPS*, 2020.
- [19] A. Radford et al., "CLIP: Connecting Text and Images," *OpenAI*, 2021.
- [20] Alayrac et al., "Flamingo: A Visual Language Model," *DeepMind*, 2022.
- [21] Y. Chen et al., "Attention Re-Alignment in Multimodal Large Language Models," *Nature Scientific Reports*, 2026.
- [22] B. Kim et al., "Multimodal Large Language Models in Real-World Applications," *Computers in Biology and Medicine*, 2025.
- [23] Z. Deng et al., "A Survey of Multimodal Models on Language and Vision," *Data Mining and Machine Learning*, 2025.
- [24] F. Lin, "Vision-Language Models: A Survey of 26K Papers," *arXiv preprint*, 2025.
- [25] X. Lin et al., "Deep Reasoning in Multimodal Chatbots," *Frontiers in Medicine*, 2025.
- [26] Z. B. Akhtar, "Multimodal AI and Vision-Language-Robotics Integration," *Journal of AI Ethics*, 2026.
- [27] BentoML, "Multimodal AI: Vision-Language Models Guide," 2025.
- [28] OpenReview, "A Comprehensive Survey of Multimodal Large Language Models," 2025.
- [29] OpenAI, "GPT-5: A Multimodal Large Language Model," 2025.
- [30] Google DeepMind, "Gemma and Multimodal Open Models," 2026.