A Multimodal AI Dialogue System for Unified Document, Visual, and Audio Interaction

Yujun Feng, **Jingyi Huang**, **Yang Zhang**
Department of Computer Science and Software Engineering, Miami University
fengy46@miamioh.edu, huangj84@miamioh.edu, zhang981@miamioh.edu

Abstract

This paper presents a multimodal intelligent dialogue system that seamlessly integrates document analysis, visual media processing, and audio interaction within a unified web interface. The system ensures secure user identity verification through persistent conversational management, leveraging textual document analysis, dynamic context integration, and cross-media interactions via video, image, and real-time speech processing. Our approach introduces three key innovations: (1) context-aware document analysis through text extraction, (2) a multimodal input pipeline supporting images, videos, and audio, and (3) persistent chat history management for maintaining conversational continuity. The system facilitates seamless transitions between audio and text, enabling natural interactions by processing audio input and converting text responses into speech. Additionally, the platform provides an intuitive interface for document uploads, camera capture, and audio recording, while ensuring conversation context is preserved across sessions. This implementation demonstrates the practical integration of multimodal input in an interactive artificial intelligence (AI) system, showcasing its potential for enhanced user engagement and interaction.

1 Introduction

The demand for modern AI systems capable of processing and responding to diverse input modalities within a unified interaction framework is rapidly increasing [Dolgikh, 2024; Zhu *et al.*, 2024]. While numerous dialogue systems exist, most are constrained to single-modal interactions, typically handling either text or images [Ni *et al.*, 2023; Zhai and Wibowo, 2023]. This limitation poses challenges for users who need to reference documents, share visual information, incorporate audio inputs, or seamlessly switch between different interaction modes [Yin *et al.*, 2021]. Recent advancements in multimodal technologies underscore the importance of harmonizing heterogeneous data streams [Deldjoo *et al.*, 2021]. However, practical implementations in conversational systems remain fragmented, limiting their usability in real-world scenarios [Baltrušaitis *et al.*, 2018].

Most contemporary AI dialogue systems operate via web interfaces and offer varying levels of functionality across different input modalities [Lù *et al.*, 2024]. While some systems support document analysis or visual media processing [Hariri, 2023], there is a notable absence of an integrated solution that combines all these capabilities within a single conversational framework. The challenges extend beyond handling diverse input types to ensuring a seamless and context-aware user experience across modalities [Liang *et al.*, 2024].

To address these gaps, this paper presents a multimodal intelligent dialogue system that leverages software frameworks and large language models to process and respond to user queries while maintaining contextual awareness across different input modalities. The system enables users to upload and reference documents, process vision-based inputs, and engage in complex queries that build upon prior conversations. Additionally, it features persistent conversational management, allowing secure storage of user interactions and historical records. This ensures that users can reference past conversations, maintain continuity across sessions, and uphold secure identity verification and data privacy.

2 System Architecture

The system architecture is designed to facilitate seamless multimodal interactions between users and the AI model by implementing a modular framework. Each component processes different aspects of system functionality while ensuring internal integration for a cohesive user experience. The architecture consists of four main components: User Authentication System, Media Processing Pipeline, Dialogue Management System, and Data Persistence Layer, along with a Frontend Interface for user interactions. An overview of the system architecture is shown in Figure 1. A full demonstration of the system can be found at the following link: https://youtu.be/MDSmZq-PwI0.

2.1 Media Processing Pipeline

The Media Processing Pipeline processes multiple input modalities to enhance user interaction. Document processing maintains the contextual relationship between uploaded documents and ongoing conversations, allowing users to reference specific document content in their queries. Visual media

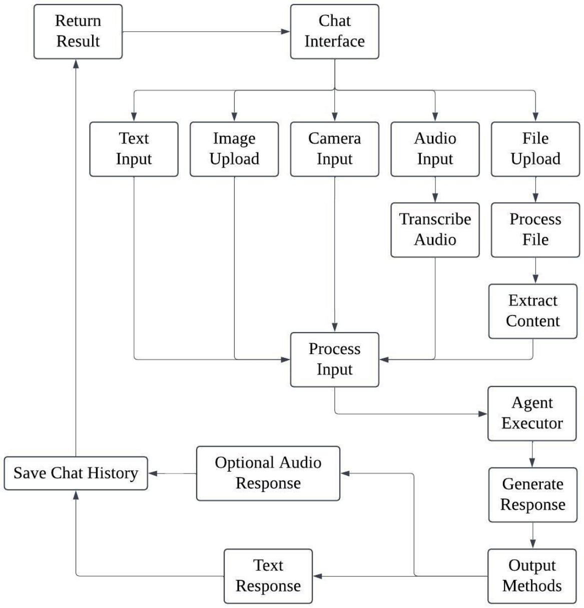


Figure 1: An Overview of System Architecture

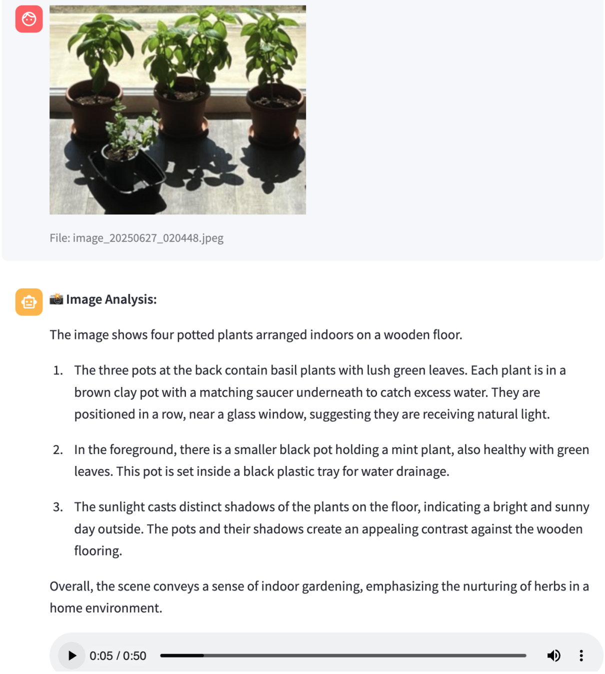
processing supports static images and video content, ensuring consistent data representation and structured storage. Audio processing enables real-time speech recognition and voice synthesis, facilitating seamless audio interactions within the dialogue system. A screenshot of showing the ability of the system to process image, text, and audio is shown in Figure 2. By integrating these modalities, the system ensures that users can interact fluidly using text, images, videos, and speech while preserving contextual continuity.

2.2 Dialogue Management System

The Dialogue Management System orchestrates interactions between user inputs and the AI model. It employs GPT-4o as its core reasoning engine, leveraging structured message formats to maintain contextual awareness across multiple dialogue turns. This system integrates multimodal inputs, ensuring that text, visual, and audio data are processed cohesively within the conversation context. Speech recognition and voice synthesis enable natural spoken conversations through speech-to-text (STT) and text-to-speech (TTS) mechanisms, maintaining coherence across different input modalities. Furthermore, conversation history is dynamically updated, allowing the system to reference previous exchanges and provide contextually relevant responses.

2.3 Data Persistence Layer

The Data Persistence Layer efficiently manages the storage and retrieval of user data, dialogue histories, and media content. Conversation histories are stored in a structured format, allowing for fast retrieval and continuity across multiple sessions. Media file management ensures that uploaded content remains accessible and contextually linked to the conversation. Additionally, robust data structuring and indexing techniques enhance system efficiency while maintaining security and privacy compliance.



File: image_20250627_020448.jpeg

📸 **Image Analysis:**

The image shows four potted plants arranged indoors on a wooden floor.

1. The three pots at the back contain basil plants with lush green leaves. Each plant is in a brown clay pot with a matching saucer underneath to catch excess water. They are positioned in a row, near a glass window, suggesting they are receiving natural light.
2. In the foreground, there is a smaller black pot holding a mint plant, also healthy with green leaves. This pot is set inside a black plastic tray for water drainage.
3. The sunlight casts distinct shadows of the plants on the floor, indicating a bright and sunny day outside. The pots and their shadows create an appealing contrast against the wooden flooring.

Overall, the scene conveys a sense of indoor gardening, emphasizing the nurturing of herbs in a home environment.

Figure 2: A Screenshot of Image Description and Audio Output

2.4 Frontend Interface

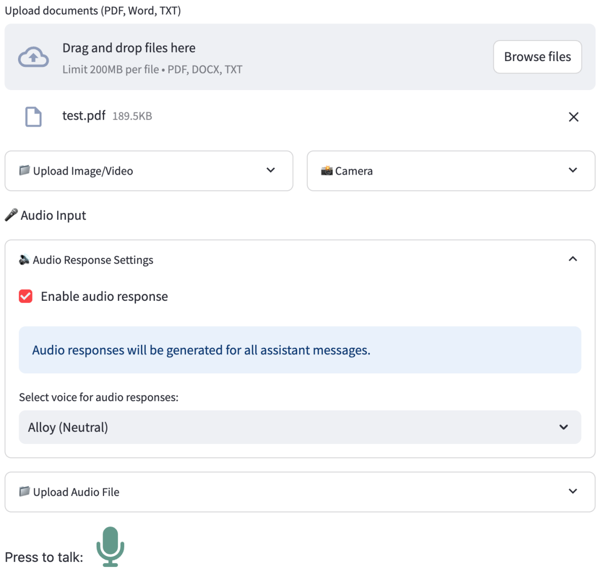
The Frontend Interface serves as the primary user access point, providing an intuitive and responsive platform for multimodal interactions. Users can upload documents, capture images and videos, and record audio within a single interface. A screenshot of showing the multimodal file handling interface of the system is shown in Figure 3. Dialogue history is displayed in a clear, chronological format, updating dynamically as new inputs are processed. The interface also incorporates adaptive UI components that adjust to different input types, ensuring a consistent and seamless user experience across various interaction modalities.

3 Technical Implementation

The technical implementation of the multimodal dialogue system was integrated with multiple types of technologies focusing to forge a cohesive interaction platform. This system combines document processing, visual media processing, audio processing, and AI-driven dialogue management through a modular implementation approach.

3.1 Document and Media Processing

The document processing achieves a flexible approach handling multiple file formats. For PDF documents, the system takes advantage of PyMuPDF to extract text content, and it maintains structural information. The system supports multiple document formats including PDF, Word (DOCX), and plain text files. This implementation supports single and multiple uploading documents to maintain dynamic integration to



test.pdf 189.5KB

☑ Enable audio response

Audio responses will be generated for all assistant messages.

Select voice for audio responses: Alloy (Neutral)

Figure 3: A Screenshot of Multimodal File Uploading

single flow of content with conversation context. Extracted content which is pre-processed and maintained in conversation sessions is being allowed to be accessed continuously during the whole dialogue process.

Visual media processing implementation processes both static and dynamic visual content with enhanced session persistence. Image processing takes advantage of PIL to do the basic image processing which transfers image to base64 format for high-efficient storage and retrieval. The system includes automatic image analysis functionality that generates detailed descriptions when images are uploaded, utilizing OpenAI's GPT-4o vision capabilities. This system supports uploading images and camera capturing through Streamlit's native components. The video processing includes format verification and mechanisms of temporary storage. which enables users to reference visual content during the whole dialogue process. All visual media is processed synchronously with conversation flow and maintaining temporal alignment with textual and audio interactions.

3.2 Interface Integration

The speech interface implementation combines real-time audio recording with asynchronous processing and advanced text-to-speech capabilities. By using Streamlit's recording component for audio input, this component generates binary audio data. The system integrates seamless audio capture and supports uploaded audio files in multiple formats (MP3, WAV, M4A, OGG). This data transfers audio to text using OpenAI's Whisper model meanwhile maintaining high accuracy under different accent and speaking styles. Additionally, the system features configurable voice synthesis with six different voice options using OpenAI's TTS-1 model for generating audio responses. The audio processing system maintains synchronization with conversation based on text to make sure that audio interaction maintains coherence of context.

User interface implementation takes advantage of Streamlit's component system to establish an intuitive interaction environment. The interface will dynamically update to reflect the current conversation state to provide user actions immediate feedback. This implementation includes specialized components of document uploading, image/video uploading, camera capturing, and audio recording, and it contains clear visual indicators which are used to process status and system responses. The interface maintains responsive actions on different sizes of screen and input methods to ensure all types of equipment and interaction modalities maintain consistent usability.

3.3 Conversation Management

The conversation management system implements structural methods to maintain dialogue context. User interactions are stored using a comprehensive message format that captures content type, media references, and temporal information for each exchange. The system utilizes LangChain's agent framework with OpenAI's GPT-4o as the core inference engine, supplemented by GPT-4o-audio-preview for enhanced audio interactions. Conversation context is maintained through a sliding window approach that considers contextual relevance, while the system supports multimodal context awareness by detecting image-related queries and utilizing vision capabilities when appropriate. This system maintains conversation history through conversation status management and persistent storage to enable users to reference past interactions while maintaining context from different input modalities.

3.4 Data Persistence Strategy

The Data Persistence Strategy implements a comprehensive file-based storage system with hierarchical organization. User authentication data was stored based on CSV format, and basic secure implementation was adopted, meanwhile conversation history records were stored based on JSON format to achieve easy accessing and higher flexibility. This implementation includes establishing and managing mechanisms of user-specific directories, conversation history recording and temporary media files.

4 Conclusion

This paper presents a multimodal intelligent dialogue system that integrates document analysis, visual media processing, and audio interaction within a unified web interface. Developed using the Streamlit framework and leveraging GPT-4o through LangChain's agent framework, the system showcases practical innovations in multimodal interaction. By efficiently combining diverse technologies, it provides a seamless and context-aware interaction experience applicable to both educational and professional settings. Its modular and persistent architecture enables secure dialogue management and conversation history tracking while offering an intuitive interface for document uploads, media capture, and audio interactions. This work underscores the importance of multimodal AI systems in enhancing real-world applications, paving the way for more advanced and context-aware interaction frameworks.

References

[Baltrušaitis *et al.*, 2018] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[Deldjoo *et al.*, 2021] Yashar Deldjoo, Johanne R Trippas, and Hamed Zamani. Towards multi-modal conversational information seeking. In *Proceedings of the 44th International ACM SIGIR conference on research and development in Information Retrieval*, pages 1577–1587, 2021.

[Dolgikh, 2024] Serge Dolgikh. Self-awareness in natural and artificial intelligent systems: a unified information-based approach. *Evolutionary Intelligence*, 17(5):4095–4114, 2024.

[Hariri, 2023] Walid Hariri. Unlocking the potential of chatgpt: A comprehensive exploration of its applications, advantages, limitations, and future directions in natural language processing. *arXiv preprint arXiv:2304.02017*, 2023.

[Liang *et al.*, 2024] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Computing Surveys*, 56(10):1–42, 2024.

[Lù *et al.*, 2024] Xing Han Lù, Zdeněk Kasner, and Siva Reddy. Weblinx: Real-world website navigation with multi-turn dialogue. *arXiv preprint arXiv:2402.05930*, 2024.

[Ni *et al.*, 2023] Jinjie Ni, Tom Young, Vlad Pandelea, Fuzhao Xue, and Erik Cambria. Recent advances in deep learning based dialogue systems: A systematic survey. *Artificial intelligence review*, 56(4):3055–3155, 2023.

[Yin *et al.*, 2021] Yifang Yin, Harsh Shrivastava, Ying Zhang, Zhenguang Liu, Rajiv Ratn Shah, and Roger Zimmermann. Enhanced audio tagging via multi-to single-modal teacher-student mutual learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 10709–10717, 2021.

[Zhai and Wibowo, 2023] Chunpeng Zhai and Santoso Wibowo. A systematic review on artificial intelligence dialogue systems for enhancing english as foreign language students' interactional competence in the university. *Computers and Education: Artificial Intelligence*, 4:100134, 2023.

[Zhu *et al.*, 2024] Bin Zhu, Munan Ning, Peng Jin, Bin Lin, Jinfa Huang, Qi Song, Junwu Zhang, Zhenyu Tang, Mingjun Pan, Xing Zhou, et al. Llmbind: A unified modality-task integration framework. *arXiv preprint arXiv:2402.14891*, 2024.