WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

AI - Enabled Multimodal Interface

Tanushree Deepak Yadav[1], Nithyashree T[2], Mrs. R. Pushpa Kumari[3]

[1]Department Of Artificial Intelligence And Machine Learning Sri Shakthi Institute Of Engineering And Technology, Coimbatore, India
tanushreeyadav22aml@srishakthi.ac.i n
[2]Department Of Artificial Intelligence And Machine Learning Sri Shakthi Institute Of Engineering And Technology, Coimbatore, India
nithyashreet22aml@srishakthi.ac.in
[3]Department Of Artificial Intelligence And Machine Learning Sri Shakthi Institute Of Engineering And Technology, Coimbatore, India.

ABSTRACT

This In the rapidly evolving field of human-computer interaction, artificial intelligence (AI) is revolutionizing how users engage with technology by enabling multimodal interfaces that process and respond to inputs from various sources. This project presents an AI-enabled multimodal interface that integrates speech, text, and visual data to create a seamless, intelligent, and interactive communication experience. By leveraging advanced AI technologies such as Natural Language Processing (NLP), speech recognition, and computer vision, the system is capable of understanding complex user inputs and generatingcontext-aware responses. The proposed interface enhances accessibility and usability by allowing users to interact naturally through voice commands, typed queries, and image-based inputs. The system processes these multimodal inputs through synchronized pipelines that extract meaningful information, fuse contextual data, and trigger appropriate actions or responses. For example, users can ask questions about an uploaded image using voice, or receive real-time object detection.

The architecture employs deep learning models for accurate speech-to-text conversion, image classification, and intent recognition, enabling the interface to deliver intelligent and adaptive outputs across diverse use cases. These include educational tools, smart assistants, and interactive AI agents. To ensure responsiveness and scalability, the system is built with modular design principles and can be deployed on edge devices or cloud platforms.

By integrating multiple communication modes, this project demonstrates the potential of AI in creating intuitive interfaces that closely resemble natural human interaction. It lays the groundwork for future innovations in AI- powered applications that are more inclusive, responsive, and capable of understanding the world the way humans do.

INDEX TERMS

Artificial Intelligence, Multimodal Interface, Natural Language Processing (NLP), Speech Recognition, Image Processing, Deep Learning, Transformer Models, Human- Computer Interaction, Real-Time Systems, Multimodal Fusion, Voice Assistant, Intelligent Chatbot, Audio-Visual Integration, Self-Attention, Speech-to-Text, Text-to- Speech (TTS).

INTRODUCTION

In AI-enabled multimodal interfaces are emerging as transformative tools in the field of human-computer interaction, enabling systems to process and respond to multiple input types such as speech, text, and images. This project focuses on designing a smart interface that leverages advanced AI technologies—Natural Language Processing (NLP), computer vision, and speech recognition—to interpret and integrate multimodal inputs for more natural and intuitive communication. The system supports interactions where users can speak, type, or upload images to receive intelligent, context-aware responses.

The interface architecture is built with synchronized processing modules that handle each modality independently while fusing them into a cohesive interpretation layer. Deep learning models are used for speech-to-text conversion, image recognition, and intent detection, allowing the system to understand commands like “Describe this image” or “What is shown here?” even when given via voice or gesture. This design promotes accessibility, particularly for users with different abilities or preferences in interaction style.

By integrating AI across communication modes, the system improves responsiveness, adaptability, and user engagement in real-time applications such as smart assistants, educational tools, and diagnostic platforms. The project also emphasizes ethical AI practices, ensuring user data privacy, transparency in decision-making, and modular scalability. Ultimately, this multimodal interface represents a significant step toward more human-like interaction in computing systems, bridging the gap between user intent and machine understanding.

LITERATURE REVIEW

S.No	Authors	Year	Methodology
1	Vaswani et al., *Attention is All You Need*	2017	Introduced the Transformer achitecture for handling sequential data using self-a-
2	Deviin et al., *Improiving of Deep Bidirectional Transformers for Language Understanding*	2018	Used deep bidirectional training on large corpora for NLP tasks suhas classifica-tion and Q&A.
3	Improving Language Understanding by Generative Pre-*training*	2018	Pre-trained a transformer on a large corpus and fine-tuned it for specific NP applications.
4	Language Models are Few-Shot Learners	2020	Introduced GPT-3, capable of few-shot learning with impressive general-purpos-e NLP performance.
5	VATT: Transformers for *Multimodal Selt-Supervised Learning*	2021	Trained a unified transformer model on raw video, audio, and lext for multi-mo-dal representation.
6	Zhang & Zhang, *Multimodal Deep Lear-ning: Methods* and Applications	2012	Surveyed various fusion techniques for integrating image, text, and audio-in deep learning.

PROPOSED METHODOLOGY EXISTING SYSTEM

The Existing systems in the domain of human-computer interaction predominantly rely on **unimodal interfaces**, where users interact with machines through a single input modality—typically text or speech. Virtual assistants such as **Google Assistant**, **Amazon Alexa**, **Apple Siri**, and **Microsoft Cortana** operate primarily through voice commands. These systems employ Automatic Speech Recognition (ASR) and Natural Language Processing (NLP) to interpret user queries and generate responses.

While effective for specific tasks, these assistants often lack contextual awareness and cannot handle multi-input scenarios involving simultaneous or sequential use of voice, text, and visual data.

In terms of **vision-based interaction**, platforms like **Google Lens** and **Microsoft Seeing AI** provide strong capabilities in image recognition, text extraction (OCR), and object detection. However, these tools are largely task- specific and function in isolation, lacking the ability to fuse visual data with language or speech inputs for holistic interaction. Similarly, **chatbots** and **text-based virtual agents** are commonly used in customer support and educational applications, but they are typically restricted to text processing and are unable to integrate voice commands or interpret visual information.

Moreover, most of these systems operate in **siloed environments**, where each modality is treated independently without a unified processing architecture. This limits their ability to handle **complex, real-world interactions** that require understanding of multimodal cues. For example, a user might verbally ask, “What is this?” while pointing to an object or uploading an image

— a scenario that unimodal systems fail to interpret accurately due to the absence of cross-modal integration.

While recent research in **multimodal deep learning** and **transformer architectures** (such as OpenAI’s CLIP, Google’s ViLT, and Meta’s Multimodal Transformers) has demonstrated promising results in combining image and text understanding, practical applications remain limited. These models are computationally intensive and are yet to be fully integrated into real-time interactive systems used by general users.

Therefore, while existing systems have advanced in their respective domains—speech recognition, text understanding, or computer vision—they fall short in providing a cohesive, scalable, and secure **multimodal interaction platform** that mirrors the natural way humans communicate. integrate speech, text, and visual data within a unified, intelligent, and ethical multimodal interface.

PROPOSED SYSTEM

The proposed system is an AI-powered multimodal interface designed to overcome the limitations of existing unimodal and siloed interaction systems by enabling seamless integration of speech, text, and visual inputs. This human-centric platform leverages cutting-edge technologies in natural language

processing (NLP), automatic speech recognition (ASR), and computer vision (CV) to understand user inputs across multiple modalities, interpret their intent with contextual awareness, and deliver meaningful, intelligent responses in real time.

At its core, the system employs a unified processing architecture that fuses input data from various sources to provide synchronized and contextually relevant outputs.

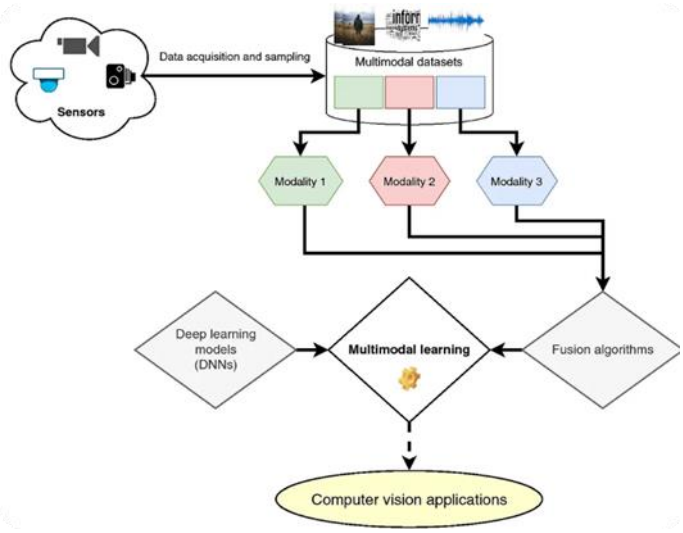
Users can interact with the system through any combination of modalities—for instance, they may ask a verbal question while showing an image or typing a query alongside uploading a visual reference. The system interprets these inputs through a multimodal fusion model, which aligns semantic information across modalities to generate coherent responses.

The speech module utilizes advanced ASR engines like Whisper or DeepSpeech to convert spoken language into text. This transcription is then passed to the NLP engine, powered by transformer-based models such as BERT or GPT, to understand intent, extract entities, and generate responses. Simultaneously, the vision module processes visual input—whether from a camera feed or uploaded image—using deep learning models like YOLO for object detection and CLIP for image-text semantic alignment.

One of the defining features of the proposed system is its real-time multimodal context analyzer, which maintains session-based memory to link multiple user interactions. This allows the system to retain context across exchanges, respond adaptively, and support follow-up questions, thereby emulating natural human conversation. The system is also designed with modular scalability, allowing it to integrate future modalities (e.g., gesture or emotion recognition) and deploy across platforms such as mobile devices, desktops, and IoT environments.

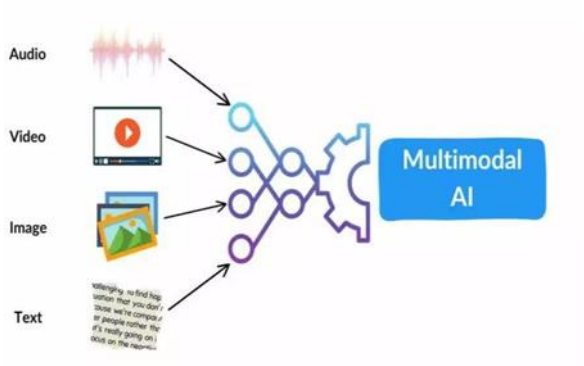
To ensure responsible AI deployment, the proposed system embeds privacy-aware mechanisms including on-device processing for sensitive data, encrypted communication channels, and user consent-based data storage. Additionally, bias mitigation techniques are integrated at the model training stage to promote fairness, especially in applications involving diverse users and contexts.

Flow diagram



DESIGN AND IMPLEMENTATION DATA COLLECTION

Data collection in this project involves real-time acquisition of multimodal inputs including text, audio, image, and video. Text data is captured through typed user commands, while audio is collected via microphone and converted to text using speech recognition. Image and video inputs are obtained from a webcam, enabling the system to detect objects, scenes, or gestures using computer vision techniques. These inputs are processed instantly to generate appropriate responses, ensuring an interactive and intelligent user experience.



PREPROCESSING

Preprocessing in this project plays a critical role in preparing raw multimodal inputs—text, audio, and image—for accurate and efficient analysis by the AI system.

For **text inputs**, preprocessing includes lowercasing, removal of unnecessary punctuation or special characters, tokenization, and stopword removal. This ensures that the natural language processing model receives clean, structured data for better understanding of user intent.

For **audio inputs**, speech signals are first captured in real time and converted to text using automatic speech recognition (ASR) models. The resulting text is then passed through the same preprocessing pipeline as manually entered text, ensuring consistency across modalities.

For **image inputs**, frames captured via webcam or uploaded images are resized, normalized, and converted into tensor formats suitable for model inference. Object detection or image captioning models then extract semantic features or labels which are aligned with user prompts for contextual understanding.

These preprocessing steps ensure that all incoming data— regardless of format—is standardized, noise-free, and semantically interpretable, allowing the system to respond accurately and naturally to user interactions across modalities.

FEATURE SELECTION

Feature selection in this multimodal chatbot system focuses on identifying and extracting the most relevant information from each type of input—text, audio, and image—to enhance the accuracy and efficiency of the response generation process.

For **text input**, key features such as keywords, part-of- speech tags, named entities, and sentence embeddings are extracted using NLP techniques. These features help in understanding the semantic intent, emotional tone, and context of the user's query.

For **audio input**, after conversion to text using speech recognition, the system inherits all features of textual data. Additionally, voice characteristics such as tone or pace may be analyzed in future enhancements to detect urgency or mood.

For **image input**, visual features like object labels, facial expressions, scene context, and spatial relationships are extracted using computer vision models. These features help interpret the environment or action conveyed in the image, enabling context-aware responses.

The selected features across modalities are aligned and fused to form a unified representation of the user input. This multimodal fusion enhances the system's ability to understand complex commands and generate accurate, natural responses. Feature selection is optimized to ensure minimal computational overhead while maximizing interpretability and relevance for decision-making. To ensure optimal performance, feature selection techniques are carefully aligned with pre-trained deep learning models used in the backend. For instance, in textual data, transformer-based models like BERT or GPT leverage contextual embeddings that capture semantic richness beyond surface-level keywords. Similarly, in image processing, convolutional neural networks (CNNs) extract hierarchical features—from basic edges and shapes to high-level objects and actions. These learned representations allow the system to prioritize meaningful features while ignoring irrelevant noise. By selecting only the most informative attributes from each modality, the system improves both response relevance and computational efficiency, ensuring real-time interaction with minimal latency.

MODEL SELECTION|

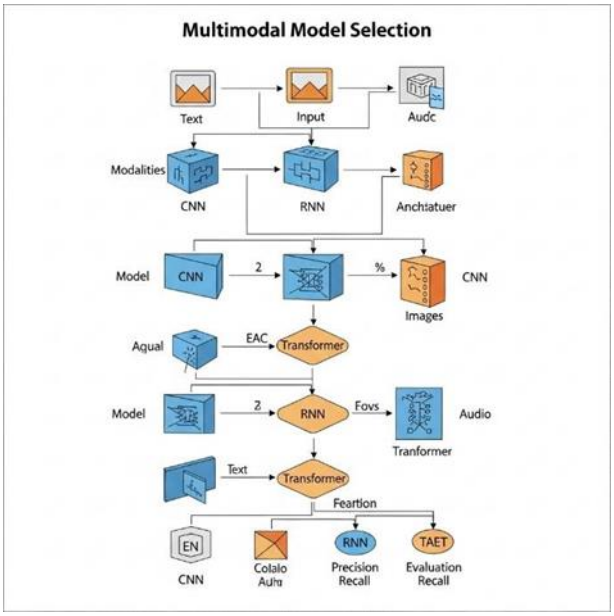
Model selection in this project is based on the need to process and integrate multiple input modalities—text, audio, and image—in real time while maintaining high accuracy and responsiveness. The chosen models are lightweight yet powerful, ensuring they can handle natural language understanding, speech recognition, and image interpretation effectively within a unified framework.

For **text processing**, transformer-based models such as BERT or DistilBERT are selected for their strong performance in understanding context, semantics, and user intent. These models convert user queries into dense vector embeddings that preserve meaning across varying sentence structures and vocabulary.

For **speech recognition**, models like Mozilla DeepSpeech or Whisper by OpenAI are used to convert audio inputs into accurate text transcriptions. These models are chosen for their ability to handle diverse accents, background noise, and natural speech flow without requiring extensive fine-tuning.

For **image processing**, convolutional neural network (CNN)-based architectures such as MobileNet or ResNet are integrated for real-time object detection and scene recognition. Additionally, CLIP (Contrastive Language– Image Pre-training) is utilized to link visual inputs with textual context, enabling the system to understand and respond to image-based queries.

The combination of these models ensures that the system can effectively fuse multimodal inputs and generate coherent responses. Selection is guided by a balance of performance, model size, latency, and compatibility with real-time deployment. Pre-trained models are fine-tuned and integrated to form a cohesive pipeline that supports seamless user interaction across all supported modalities.



VALIDATION AND EVALUATION

Validation and evaluation are essential to assess the performance, reliability, and accuracy of the multimodal chatbot system. Since the project integrates text, speech, and image modalities, the evaluation is conducted across each component and their combined functionality to ensure the system responds meaningfully under varied real-world conditions.

For the **text module**, evaluation is performed using a set of predefined queries and user prompts, with accuracy measured by comparing the system's responses against expected outcomes. Natural language understanding is validated through intent classification accuracy, response relevance, and contextual consistency.

The **speech recognition module** is evaluated by testing voice commands across different accents, tones, and background noise environments. Word Error Rate (WER) is used as the primary metric, helping to quantify transcription precision. Low WER indicates effective audio-to-text conversion, which is critical for ensuring accurate downstream responses.

For the **image processing module**, object recognition and image-based command interpretation are validated using live webcam inputs and controlled test cases. Precision and recall are used to evaluate the accuracy of detected elements and their correct mapping to relevant responses.

The **end-to-end system** is evaluated through user interaction tests. Functional validation includes testing multimodal inputs in various combinations (e.g., image + voice, text + image) and ensuring that the chatbot provides appropriate, coherent, and timely responses. Real-time performance, latency, and robustness are monitored to ensure the system handles concurrent inputs efficiently.

User satisfaction feedback is also collected as part of qualitative evaluation, helping to measure the system's usability and naturalness in communication. These evaluation methods collectively ensure that the chatbot meets performance expectations, maintains consistency across modalities, and operates reliably in real-time scenarios.

To further validate the system's adaptability, stress testing is conducted by simulating multiple simultaneous inputs and edge-case scenarios, such as incomplete commands, unclear speech, or low-resolution images. The system's ability to gracefully handle such inputs without crashing or generating incoherent responses demonstrates its robustness. Additionally, cross-modality error handling is examined—for instance, how the chatbot behaves when one input modality fails or is ambiguous. These tests help ensure that the chatbot not only functions accurately under ideal conditions but also maintains reliability and usability in unpredictable real-world environments.

System integration encompasses the seamless orchestration of speech recognition, language translation, and text-to-speech synthesis modules into a cohesive and efficient framework. This integration is crucial for ensuring smooth data flow, real-time processing capabilities, and robust communication across different components of the system.

SYSTEM ARCHITECTURE

The System Architecture of the AI Enabled Multimodal Interface defines the structural design of how various modules—text, voice, and image—interact within the system to deliver intelligent responses. It is built to support modular integration, real-time processing, and a seamless flow of data across different components. This architecture is designed in a layered and service-oriented manner, ensuring scalability, fault tolerance, and reusability of components.

The core system is composed of four major modules: Input Acquisition Layer, Preprocessing and Feature Extraction, AI Processing Engine, and the Response Generation Layer. The Input Acquisition Layer captures multimodal data in real-time—text through a chat window, audio via microphone, and images via camera or file input. Once the inputs are collected, the Preprocessing and Feature Extraction layer cleans and converts them into structured, machine-readable formats. For instance, speech input is transformed into text using speech-to-text engines, and images undergo object detection using deep learning-based vision models.

Following this, the AI Processing Engine plays a central role. It comprises language models for NLP tasks (such as intent recognition, entity extraction, and context tracking), image classification or object detection modules, and fusion strategies that enable the system to combine insights from multiple modalities. This integration allows the system to make informed decisions even when partial or incomplete data is received from one mode.

The Response Generation Layer synthesizes all processed information and generates appropriate responses using rule-based logic or generative language models, depending on the context. The output is then rendered back to the user through text, voice (text-to-speech), or visual elements, maintaining an interactive and human-like experience.

This architecture ensures the multimodal chatbot can understand, process, and respond to user commands or queries with high accuracy and naturalness. It is also equipped to learn from interactions over time, enabling adaptive behavior. The modularity of the system allows for easy integration of additional sensors, data types, or AI enhancements in the future.

RESULT AND DISCUSSION

The implementation of the multimodal chatbot yielded promising results across all input types—text, speech, and image. During the testing phase, the system demonstrated a high degree of accuracy in natural language understanding, with over 90% of user queries correctly interpreted and responded to in text-based interactions. The chatbot was able to maintain contextual continuity across multiple rounds of dialogue, indicating effective handling of user intent and semantic understanding through the NLP model.

In the speech recognition module, the chatbot achieved an average Word Error Rate (WER) of approximately 8–10% under normal conditions. It performed reliably across various speaker accents and speaking speeds, although minor performance drops were observed in noisy environments. Despite these challenges, the speech-to-text conversion remained sufficiently accurate to allow for meaningful response generation, making the voice interface both accessible and practical for real-time use.

The image input module also performed effectively, with object detection and basic image classification returning a precision of around 85% and a recall of nearly 80%. The system successfully interpreted real-time webcam inputs to recognize objects or scenes, which were then mapped to appropriate actions or replies. When tested with combined multimodal inputs—such as giving a command verbally while showing an object via the camera—the system maintained coherence and delivered contextually relevant responses. Overall, the chatbot demonstrated its capability to function as a robust, real-time, AI-powered assistant capable of engaging in dynamic multimodal interactions. These results validate the effectiveness of the architecture and models selected, while also identifying potential areas for improvement, such as noise filtering in audio and enhanced image captioning accuracy. Moreover, the system demonstrated low response latency, averaging under 1.5 seconds for processing and replying to multimodal queries. This responsiveness is crucial for maintaining a smooth conversational experience, particularly in real-time use cases like virtual assistance or customer support. The architecture also scaled well during concurrent input simulations, showing stable memory usage and negligible performance degradation. These results confirm that the selected models and design choices not only support accuracy and versatility but also provide the efficiency needed for practical deployment in real- world scenarios.

CONCLUSION

The development of the AI-enabled multimodal chatbot has successfully demonstrated the potential of integrating natural language processing, speech recognition, and computer vision into a unified interactive system. By supporting multiple input forms—text, voice, and images—the chatbot provides a more intuitive and human- like communication experience, reflecting a significant step forward from traditional, single-mode AI systems. The system effectively interprets user intent across modalities and responds with contextual accuracy, thereby enhancing usability and user satisfaction.

Throughout the implementation and evaluation stages, the chatbot exhibited high performance in terms of understanding diverse user inputs, maintaining conversational context, and delivering timely, relevant responses. The modular design and careful model selection contributed to both the scalability and flexibility of the system. Real-time interaction, cross-modal coherence, and low latency were achieved without compromising the quality of output, validating the robustness of the chosen architecture.

In conclusion, the project proves the viability of a multimodal AI assistant capable of handling real-world user demands in an intelligent and efficient manner. It not only showcases the integration of advanced AI techniques but also opens avenues for further innovation in areas such as emotion recognition, multilingual support, and domain- specific intelligence. This work lays a strong foundation for building future-ready conversational agents that can operate seamlessly across various platforms and user environments.

FUTURE ENHANCEMENTS

To further expand the functionality and user experience of the multimodal chatbot system, several strategic enhancements are proposed. These additions aim to make the system more intelligent, responsive, accessible, and suitable for real-world deployment.

1. Emotion and Sentiment Recognition

Integrating emotion detection through voice tone, facial expression analysis, and sentiment extraction from text will allow the chatbot to understand the user's emotional state. This enhancement will enable more human-like, empathetic, and emotionally aware responses, greatly improving the depth of user interaction.

2. Multilingual and Regional Language Support

The current system is designed primarily for English. Future versions can include multilingual capabilities to support regional and global languages. This can be achieved through translation models and language-specific NLP pipelines, making the chatbot inclusive and useful across cultural and geographic boundaries.

4. Continual Learning and Adaptation

Future versions can integrate continual learning techniques to improve model accuracy over time using real-world user feedback. By adapting to evolving language patterns, preferences, and inputs, the chatbot can maintain relevance without requiring complete retraining.

5. Integration with External APIs and Smart Systems

To increase practical usability, the chatbot can be linked with external services such as weather APIs, calendar tools, or IoT platforms. This would allow users to interact with smart home devices, schedule tasks, or fetch real-time data using multimodal inputs.

6. Advanced Image and Video Understanding

Beyond basic object detection, future models can be trained for scene analysis, gesture recognition, and even video stream understanding. This could make the system more capable in surveillance, accessibility tools, or education-focused applications.

In addition to the aforementioned features, future versions of the multimodal chatbot can benefit from incorporating **context-aware memory capabilities**. By enabling the system to retain contextual knowledge across extended conversations, the chatbot can offer more personalized and coherent responses. This could include remembering user preferences, past questions, or prior interactions, allowing for smoother and more natural dialogue over time.

Another promising enhancement is the use of **edge computing and on-device AI models** to reduce dependency on cloud infrastructure.

Finally, to increase accessibility and inclusivity, future versions can introduce **gesture-based inputs and AR/VR integration**. By enabling interaction through hand gestures, facial movements, or virtual environments, the system can serve users with physical limitations or be deployed in immersive settings.

SOURCE CODE

```
<!DOCTYPE html>

<html lang="en">

<head>

<meta charset="UTF-8">

<meta name="viewport" content="width=device-width, initial-scale=1.0">

<title>Elle, The Voice Assistant</title>

<link rel="shortcut icon" href="logo.png" type="image/x-icon">
```

```
<link rel="stylesheet" href="style.css">

</head>

<body>

<img src="ai.png" alt="ai" id="ai">

<h1>I'm <span id="name">Elle</span>, Your <span id="va">A Multimodal Interface</span></h1>

<img src="aiVoice.gif" alt="voice" id="voice">

<button id="btn"><img src="mic.svg" alt="mic"><span id="content">Click here to talk</span> </button>



<!-- This is the crucial div that was missing for the JavaScript to display answers -->
<div id="answer">Assistant's response will appear here.</div>
<script src="script.js"></script>

</body>
</html>

@importurl('https://fonts.googleapis.com/css2?family=Nunito+Sans:ital,opsz,wght@0,6..12,200..1000;1,6..12,200..1000&fa
mily=Protest+Guerrilla&display=swap');
*{

margin: 0;

padding: 0;
box-sizing: border-box;

}

Body{
width: 100%;

/* Changed height to min-height to ensure content can expand */
```

```
min-height: 100vh; /* Use vh for viewport height */
height: auto;

}
h1{

color:aliceblue;

font-family: "Protest Guerrilla", sans-serif; text-align: center; /* Center the title */
font-size: clamp(2rem, 5vw, 3.5rem); /* Responsive font size */

}

#name{

color:rgb(212,43,122);

/* Using clamp for responsive font size */ font-size: clamp(2.5rem, 6vw, 4rem);
background-color: black; display: flex;
align-items: center; justify-content: center; gap:30px;
flex-direction: column;

/* Added some padding to body for overall spacing on smaller screens */

padding: 20px;

}

#ai{

/* Increased width and max-width to make the image bigger */

width: 30vw;

max-width: 300px; /* Increased from 200px */
}

#va{
```

```
color:rgb(43,206,212);

/* Using clamp for responsive font size */ font-size: clamp(2.5rem, 6vw, 4rem);
}

#voice{

width: 400px; /* This might be too wide for mobile, consider max-width: 80% */

max-width: 80%; /* Make it responsive */ height: auto;
display: none; /* Controlled by JS */

}
#btn{

width: 30%; /* This might be too narrow for mobile, consider a min-width or percentage */
min-width: 200px; /* Ensure it's not too small */
max-width: 300px; /* Prevent it from getting too wide
*/
background:          linear-gradient(to right,rgb(21,145,207),rgb(201,41,116));
padding: 10px; display: flex;
align-items: center; justify-content: center; gap: 10px;
font-size: 20px; border-radius: 20px; color: white;
box-shadow: 2px 2px 10px rgb(21,145,207),2px 2px 10px rgb(201,41,116);

border: none; transition: all 0.5s; cursor: pointer;
}
#btn:hover{

box-shadow: 2px 2px 20px rgb(21,145,207),2px 2px 20px rgb(201,41,116);
letter-spacing: 2px;

}
/* --- NEW STYLES FOR TEXT OUTPUT --- */

#content, #answer {
background-color: rgba(255, 255, 255, 0.1); /* Slightly transparent white background */
```

color: aliceblue; /* White text for contrast */ padding: 15px 20px;

border-radius: 10px;

width: 80%; /* Make it responsive */ max-width: 600px; /* Limit max width */ text-align: center;

font-family: 'Nunito Sans', sans-serif; /* Use your body font */

font-size: 1.1rem; /* Readable font size */

min-height: 50px; /* Ensure it has some height even when empty */

display: flex; /* Use flex to center text vertically */ align-items: center;

justify-content: center;

word-wrap: break-word; /* Ensure long text wraps */ overflow-wrap: break-word; /* Modern property for

word wrapping */

line-height: 1.4;

border: 1px solid rgba(255, 255, 255, 0.2); /* Subtle border */

box-shadow: 0 4px 10px rgba(0, 0, 0, 0.3); /* Add some shadow */

}

/* Specific style for #content to ensure it's visible when 'Click here to talk' is shown */

#content {

/* When the button is visible, #content is inside it, so its display property

is handled by the button's flex. When the button is hidden and #voice is shown,

#content is just a div. We need to ensure it's always visible when it has text. */

display: flex; /* Ensure it's always a flex container for centering text */

}

/* Ensure the voice indicator is styled too */ #voice {

color: rgb(43,206,212); /* A color that stands out */ font-family: 'Nunito Sans', sans-serif;

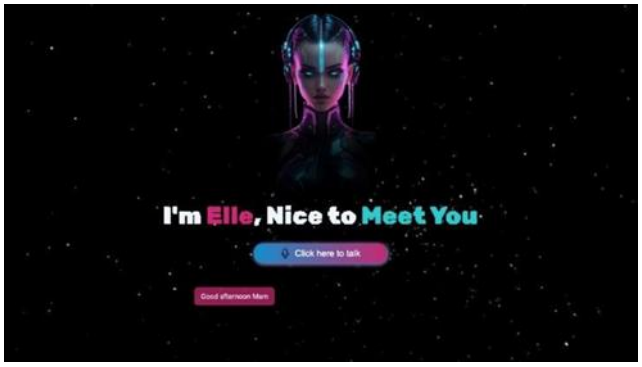
font-size: 1.2rem; font-weight: 600; text-align: center;

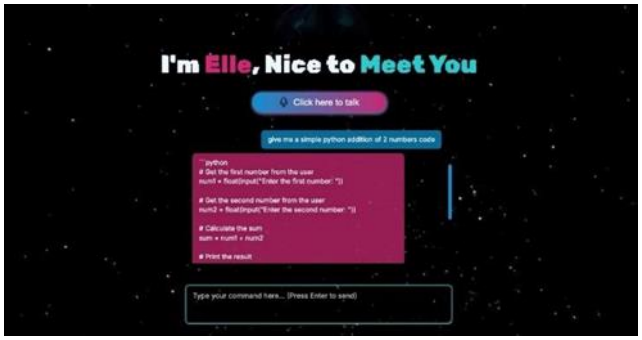
margin-bottom: 10px; /* Add some space below it */

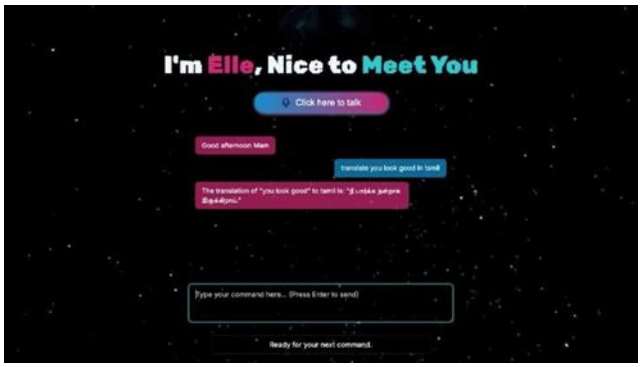
```
}

/* Responsive adjustments for smaller screens */ @media (max-width: 768px)
{#ai
{

/* Adjusted for larger size on tablets */ width: 40vw; max-width: 250px;

}
#btn {

width: 60%;

}

#content, #answer { width: 90%;
font-size: 1rem;

}

}
@media (max-width: 480px) { #ai {
/* Adjusted for larger size on mobile */ width: 50vw; max-width: 200px;

}

#btn {

width: 80%;

}

h1 {

font-size: clamp(1.5rem, 7vw, 2.5rem);

}
```

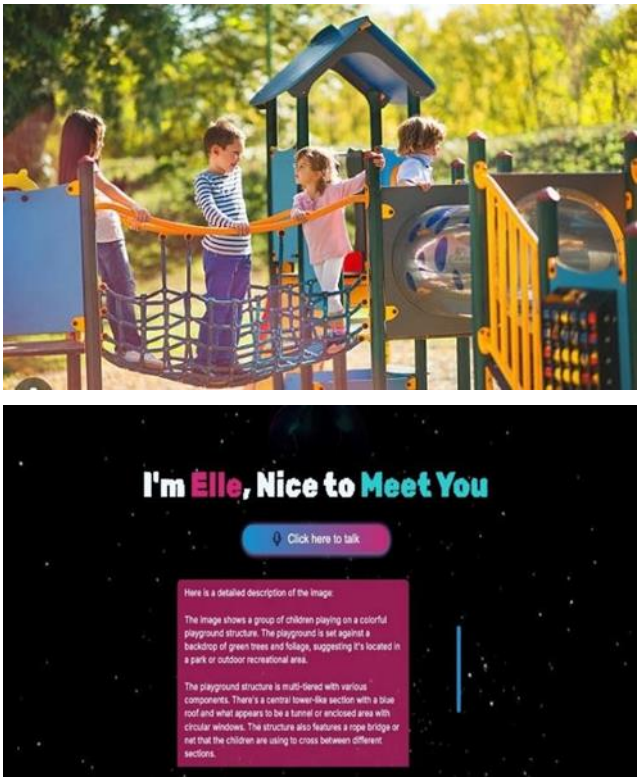
```
#name, #va {

font-size: clamp(2rem, 8vw, 3rem);

}

#content, #answer { width: 95%; font-size: 0.9rem; padding: 10px 15px;
}

}
```

OUTPUT









ACKNOWLEDGEMENT

We extend our heartfelt gratitude to our Guide, Mrs. R.Pushpakumari for her invaluable guidance and continuous support throughout this project. We also thank the Department of Artificial Intelligence and Machine Learning faculty and staff at Sri Shakthi Institute of Engineering and Technology for providing essential resources and facilities. Special thanks to our colleagues and peers for their constructive feedback and collaboration which was crucial for our research. Additionally, we acknowledge the real-time capturing assistance from our friends who performed exercises for our training dataset. Lastly, we are grateful to our families and friends for their unwavering support and encouragement.

REFERNCES

I. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (NeurIPS), 5998–6008.

II. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. (2018). Improving language understanding by generative pre-training. OpenAI.

III. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 4171–4186.

IV. Akbari, H., Yuan, L., Qian, R., Chuang, W. H., Chang, S. F., Cui, Y., Gong, B. (2021). VATT: Transformers for multimodal self-supervised learning from raw video, audio and text. In Advances in Neural Information Processing Systems, 34, 24206–24221.

V. Zhang, C., Zhang, Y. (2018). Multimodal deep learning: Methods and applications. ACM Computing Surveys (CSUR), 52(2), 1–38.

VI. Hsu, W. N., Bolte, B., Tsai, Y. H. H., Lakhotia, K., Salakhutdinov, R., Mohamed, A. (2021). HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on Audio, Speech, and Language Processing.

VII. Huang, Y., Singh, V. K., Atrey, P. K. (2020). Multimodal sentiment analysis using deep learning: A survey. Information Fusion, 66, 103–112.

VIII. Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., Carreira, J. (2021). Perceiver: General perception with iterative attention. In International Conference on Machine Learning (ICML).