Multimodal AI Assistant (Vision + Text + Speech)

Project Proposal

1. Project Overview

This project proposes the development of a Multimodal AI Assistant capable of understanding images, processing voice input, generating intelligent responses using Large Language Models, and replying with speech. The system combines Computer Vision, Natural Language Processing, and Speech AI to deliver an interactive assistant experience.

2. Problem Statement

Traditional chatbots rely only on text input and lack multimodal capabilities. Users cannot interact using images or voice, which limits accessibility and real-world applications. This project aims to bridge this gap by enabling natural, multimodal interaction.

3. Objectives

- Develop an AI assistant that understands images using vision models.
- Convert speech to text using Whisper speech recognition.
- Generate intelligent answers using LLMs.
- Convert responses back to speech using Text-to-Speech.
- Deploy the solution using Streamlit for real-time interaction.

4. Tech Stack

Python, PyTorch, Transformers (BLIP, LLMs), Whisper, gTTS, Streamlit, Google Colab, Cloud deployment (AWS/Azure/GCP).

5. Methodology

Step 1: Upload image and audio input. Step 2: Vision model generates image caption. Step 3: Whisper converts speech to text. Step 4: LLM generates response. Step 5: Text-to-Speech converts answer to audio. Step 6: Display results in Streamlit dashboard.

6. Expected Outcomes

• Accurate image understanding and captioning
• Reliable speech recognition
• Intelligent contextual answers
• Real-time web application deployment
• Demonstration of end-to-end AI pipeline skills

7. Use Cases

Customer support automation, accessibility tools for visually impaired users, smart assistants, education tools, and AI-driven help systems.

8. Timeline

Week 1: Data research and model setup
Week 2: Vision + Speech integration
Week 3: LLM integration and testing
Week 4: Streamlit deployment and optimization

9. Conclusion

This project demonstrates practical implementation of modern AI technologies by integrating vision, speech, and language models into a single application. It highlights skills in deep learning, deployment, and full-stack AI engineering, making it highly relevant for current industry demands.