Multi■Modal GenAI Agent – Complete Project Guide

Project Overview

This project is a full■stack Multi■Modal Generative AI Agent that can understand and respond to Text, Image, and Audio inputs. The system is designed for learning purposes and uses free/open■source AI models from Hugging Face, making it suitable for students and beginners.

What This Project Does

- Accepts text prompts and generates intelligent text responses.
- Accepts image files and describes or analyzes the image.
- Accepts audio files and converts speech into text (speech■to■text).
- Provides a modern ChatGPT■style UI for interaction.

Technologies & Tools Used

- Frontend: React.js, Vite, Tailwind CSS
- Backend: FastAPI (Python)
- AI Models: Hugging Face Transformers
- Text Model: Lightweight text■generation model
- Image Model: Image captioning / vision model
- Audio Model: Whisper / Faster■Whisper for speech■to■text
- API Communication: REST APIs (JSON & Multipart FormData)

Project Architecture (Flowchart Explanation)

1. User opens the web application.
2. User selects mode (Text / Image / Audio).
3. Frontend sends request to FastAPI backend.
4. Backend routes request to the appropriate AI model.
5. Model processes input and generates output.
6. Backend sends response back to frontend.
7. Frontend displays the response in chat UI.

Step■by■Step Working

- Step 1: User types text or uploads image/audio.
- Step 2: Frontend sends request to backend API.
- Step 3: Backend identifies input type.
- Step 4: Corresponding AI model processes data.
- Step 5: Output is returned to frontend.

Why This Project is Useful

- Helps beginners understand how GenAI systems work.
- Demonstrates real■world full■stack AI integration.
- Uses free and open■source AI models.

- Easy to extend with databases, authentication, or new models.

How Users Can Practice

Users can clone the GitHub repository, run the backend locally, start the frontend, and experiment with different prompts, images, and audio files to understand multi■modal AI behavior.