Project: InsightXR

🚀— Multimodal Video Intelligence System

What It Does

A web platform where users can upload videos (meetings, classrooms, or surveillance footage).

The system:

1. Extracts **audio** → performs **speech recognition (ASR)**
2. Analyzes **visual frames** → detects **objects, people, and actions**
3. Uses **NLP models** to **summarize**, **classify**, and **rank** insights
4. Displays a **dashboard** of insights: keywords, sentiment trends, engagement, and activity highlights.

Updated Tech Stack

Layer	Technology	Role
Frontend	HTML, CSS, JS	User interface for uploading videos, viewing analytics
Backend	Flask	Core API + AI pipeline orchestration
AI Models	Hugging Face Transformers	Vision, NLP, and Audio processing
Audio Processing	Whisper (ASR)	Speech-to-text transcription
Computer Vision	OpenCV + Hugging Face CV models	Object detection and scene description
Database (optional)	SuperBase (SQL)	Store metadata and analytics
Storage	Local or Cloud (AWS S3)	For user video uploads
Visualization (JS)	Chart.js / D3.js	Display data trends interactively

High-Level Architecture

```
[Frontend]
 HTML + CSS + JS
    |
    ▼
[Flask Server]
  ├── /upload (video upload route)
  ├── /process (runs ML pipeline)
  ├── /results (returns processed data)
    |
    ▼
[AI Pipeline]
  ├── Audio → ASR (Whisper)
  ├── Video → Object Detection (HF CV model)
  ├── Text → Summarization + Sentiment (HF NLP model)
    |
    ▼
[Database / Storage]
   Stores extracted text, frames, embeddings, analytics
    |
    ▼
[Frontend Dashboard]
   Visualizes key moments, entities, sentiments, and summary
```

Hugging Face Tasks Used

Domain	Task	Usage
Computer Vision	`Object Detection`, `Image-to-Text`	Identify people, actions, and describe scenes
Audio	`Automatic Speech Recognition (ASR)`	Convert speech to text
NLP	`Summarization`, `Text Classification`, `Sentence Similarity`	Generate meeting summaries, detect tone, group related topics

Example Workflow

1. User uploads a 2-minute meeting video.
2. Flask saves it → extracts frames and audio.
3. Whisper model transcribes audio → text.
4. Vision model detects who’s speaking, gestures, or key visual cues.
5. NLP models summarize the transcript + detect positive/negative tone.
6. Dashboard shows:
    - Key discussion topics
    - Sentiment chart
    - Word cloud of frequently mentioned terms
    - “Action Items” summary