ChatVid AI

Ever wish your video could answer back? Now it can.

A multimodal video-analysis system that lets users chat with YouTube videos

Headstarter SWE Residency Sprint 2 | May 2025

by Misbah Ahmed Nauman

Problem & Motivation

Why We All Struggle With Videos

Ever watched a 10-minute lecture and wasted 3 minutes scrubbing around just to find that one key concept?

Current Frustrations

• Endless Scrubbing: You drag the bar back and forth, hoping to land on the right moment.

• Disconnected Tools: You might open a transcript-only site –but then you can't click a timestamp to watch.

• Missing Visual Search: "I know I saw a red car at 2:15... but where?"

Transcript

0:05
0:12
0:19
0:26

There has to be a better way.

Solution Overview
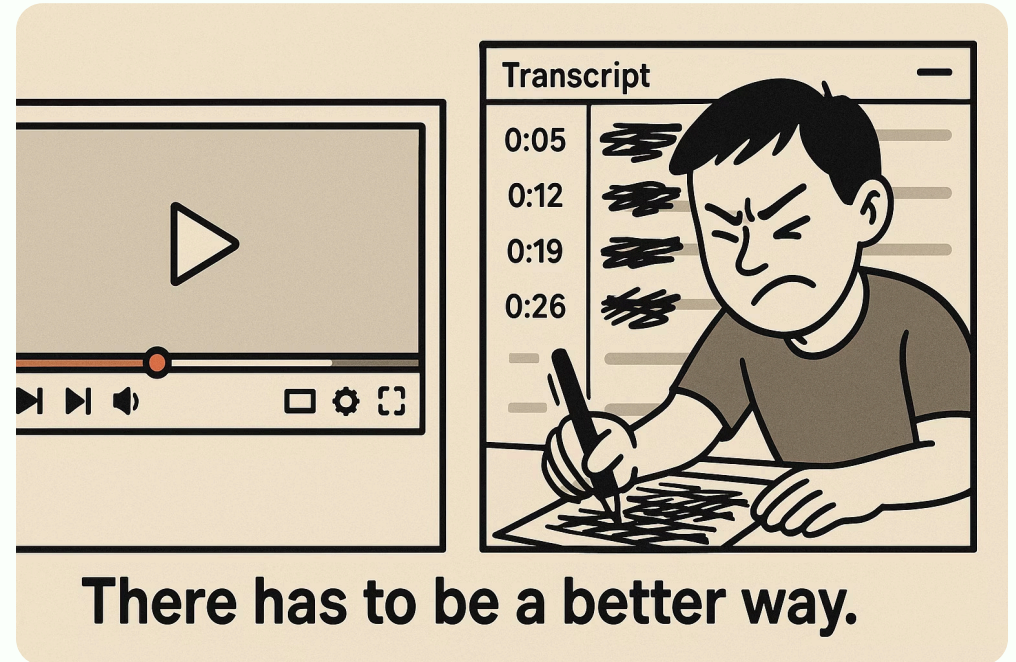
Sections

0:00 – Intro to shade balls and questions

1:07 – Driving a boat through shade balls

3:01 – Bromate, the reason for shade balls

5:31 – Original purpose and material

7:49 – Driving, algae, and evaporation

10:05 – Cost, other shapes, and boba

Visual search: e.g. "red car" — Search

Ask a question about the video… — Send

Fetch YouTube Transcripts
Using youtube-transcript-api to extract text content

AI-Powered Summaries & Q&A
Generate timestamped summaries and handle questions

Extract Visual Data
Use yt-dlp + OpenCV to capture frames every 5 seconds

Present in React Frontend
Next.js + TypeScript + Tailwind for seamless experience

System Architecture

Frontend (Next.js)

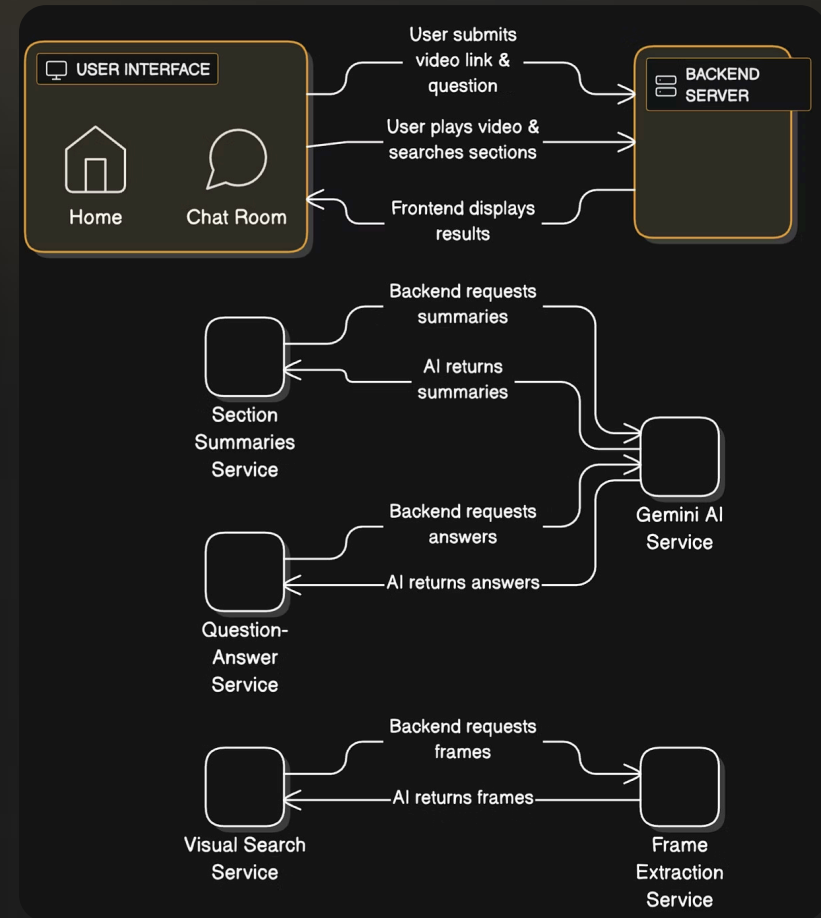Two main routes: Single-page app with two views

- **Home**: paste your YouTube link + AI key
- **Chat Room**: video player, searchable sections, chat box, and visual search.

Backend (FastAPI)

Routes: Transcript Service, Section Summaries Service, Question-Answer Service, Visual Search Service

Third-party Services

YouTube API, Gemini API, yt-dlp + OpenCV for frame extraction

Key Features

Timestamped Sections
Automatically divides videos into navigable sections with summaries

Chat with Video
Ask questions about content with timestamped hyperlink answers

JSON Structure
Clean data format for sections, timestamps, and search results

Visual Frame Search
Find specific visual elements across the entire video
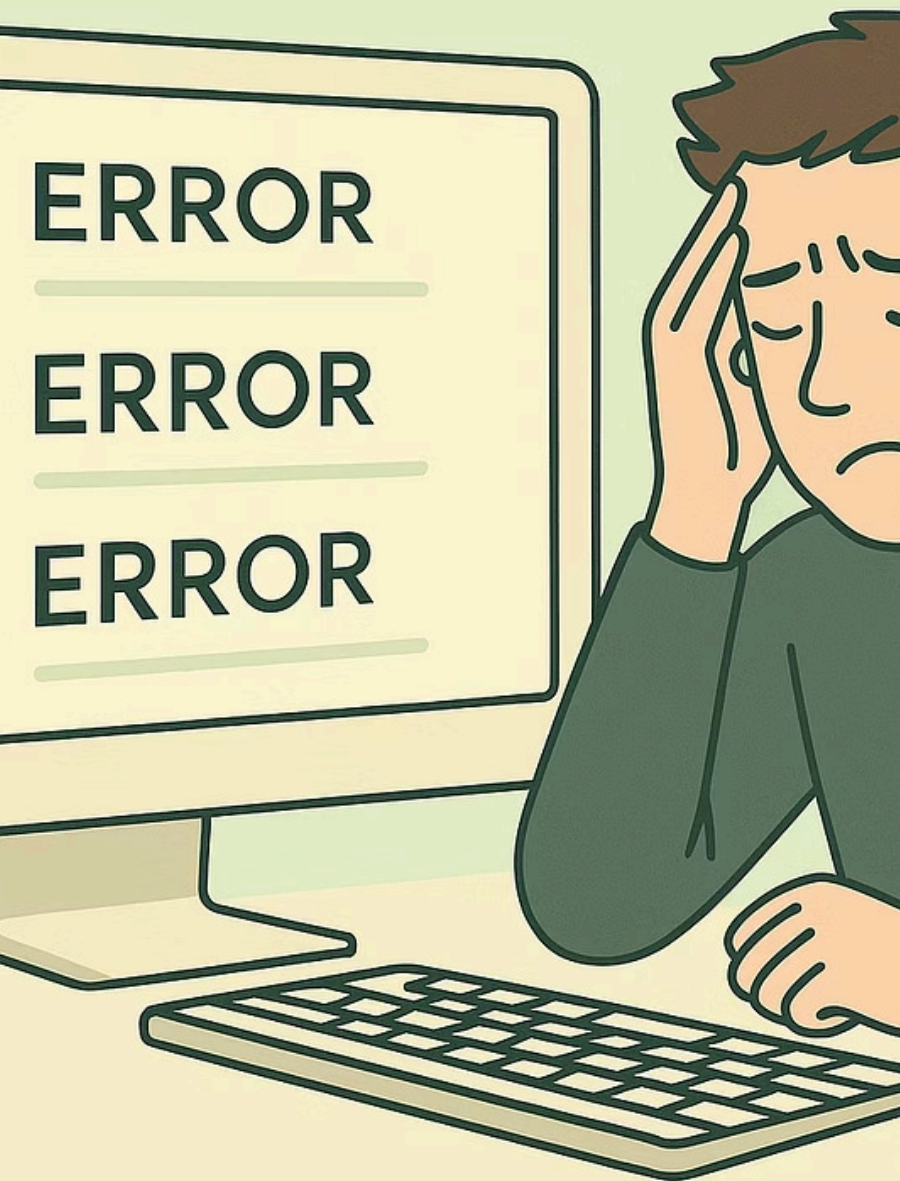
Challenges & Solutions

Async Frame Embeddings

Switched from blocking CPU to asyncio.gather() with httpx.AsyncClient

Timestamp Seeking

Replaced unreliable ReactPlayer with native iframe

Gemini Integration

Leveraged multimodal capabilities instead of juggling separate models

Prompt Engineering

Crafted precise prompts for structured JSON responses

What's Next?

Can we tell you who's speaking?

Yes–speaker labels are coming soon

Want to compare two videos?

Future updates will let you query multiple URLs at once

Ready to save your chat history?

Next version will let you review past Q&A sessions anytime

Contact: misbahahmed2005@gmail.com |
MisbahAN.com