ChatGPT

Real-Time Multimodal AI for Audio-Visual Understanding

Open-Source Multimodal AI Projects

Meetily – Local AI Meeting Assistant

Meetily (the *meeting-minutes* project by Zackriya Solutions) is a free, self-hosted meeting copilot that transcribes and summarizes meetings **entirely on your device** in real time [1]. It uses OpenAI's Whisper (or alternative local ASR models) to convert speech to text without any cloud services [1], and then applies large language models (e.g. GPT) to generate structured meeting minutes and action items. This local-first design (7.8k stars on GitHub) highlights privacy and offline operation, turning live audio from a webcam or microphone into immediate notes and summaries on Mac/Windows (with Linux support in progress) [2].

Pipecat – Real-Time Voice & Vision AI Framework

Pipecat is an open-source Python framework for building **real-time multimodal conversational agents** that can see and hear. It orchestrates streaming audio, video, and various AI services in a modular pipeline [3]. Developers can plug in speech recognition (ASR), computer vision, and language model components to create voice assistants, meeting companions, or interactive agents that process webcam audio+video in sync [4]. Pipecat emphasizes ultra-low-latency processing and cross-platform support – client SDKs exist for web, mobile, and even embedded devices (ESP32), enabling integration with hardware like smart glasses [5]. This framework powers voice-first interfaces that transcribe speech, analyze visual context, and generate responses or actions in real time.

AI Video Summarizer – WhisperX & GPT Meeting Summaries

The **AI Video Summarizer** is an open-source tool that processes meeting recordings or live streams to produce concise documentation of the content. It transcribes audio using WhisperX (which provides word-level timestamps and speaker info) and then uses GPT-based models to generate summaries tailored to the context – for example, meeting minutes, lecture notes, podcast recaps, or interview highlights [6]. In addition to text summaries, the system can automatically extract **"smart clips"** of key moments, effectively detecting highlights in the video+audio stream [7]. It supports multiple input formats and offers a web GUI for uploading videos, then returns a written synopsis and even short highlight reels of the most important segments.

Meeting Concluder – Automated Summary to Slack

Meeting Concluder is a lightweight project that turns raw meeting audio into a brief conclusion and shares it with your team. It records audio from a meeting (e.g. via a laptop or webcam mic), uses speech-to-text to transcribe the discussion, and then invokes an LLM (OpenAI's GPT) to generate a **concise conclusion and a piece of advice** from the conversation [8]. The summarized conclusion is automatically sent to a Slack

channel via the Slack API, providing teams with immediate meeting outcomes and recommendations. This Python/Go based tool showcases a practical pipeline: real-time transcription, NLP summarization, and trigger of an action (sending a message) – a blueprint for multimodal event understanding that leads to instant communication.

GraphMeeting – Nonlinear Multi-Participant Meeting Map

GraphMeeting is an experimental open-source project reimagining how meetings are recorded and structured. Instead of a traditional single-thread conversation, each participant has their own audio stream ("individual recording booth"), and the system uses voice isolation and real-time transcription to capture everyone's spoken ideas simultaneously [9]. An AI then **automatically organizes the discussion into a dynamic shared mind-map**, rather than a linear transcript. The goal is to visually map out topics, suggestions, consensus points and disagreements in real time, helping teams explore ideas in parallel. Under the hood, GraphMeeting integrates speech-to-text, speaker separation, and an LLM for analyzing relationships – enabling features like consensus detection, controversy highlighting, and action-item generation from the meeting [10]. This multimodal approach (audio + interactive visualization) turns a live meeting into a structured document of ideas and decisions.

PyAnnote-Whisper – Transcription with Speaker Diarization

pyannote-whisper combines OpenAI's Whisper ASR with PyAnnote's speaker diarization to produce rich transcripts of multi-speaker events. Aimed at meeting transcription and analysis, it produces time-stamped text with labels for each speaker, which is crucial for understanding **who said what** in real-world interactions [11]. The pipeline can be run in real time on audio streams (e.g. from a webcam's microphone array), assigning speaker identities to segments on the fly. By tagging segments per speaker, subsequent processing like summarization or action item extraction can be more context-aware. The project even integrates with ChatGPT APIs in examples, showing how you can feed the labeled transcript to an LLM for Q&A or summary generation [11]. This repository is actively maintained (updated in 2025) and serves as a building block for multi-speaker meeting understanding systems.

Voice-Controlled Coding Assistant (Whisper + GPT-4)

A recent demo by developer Zain Ahmad illustrates how real-time audio understanding can drive **program generation**. The project is a voice-controlled Python assistant that listens to natural language commands (captured via microphone), transcribes them with Whisper, and feeds them to GPT-4 to generate and even execute code on the fly [12] [13]. For example, a user can simply say **"Create a Flask app"**, and the system will produce the corresponding Python code and run it, all without any manual coding [12]. This showcases how multimodal AI can turn spoken intent into working software. While the initial implementation runs on a regular PC, the concept could be applied to AR glasses or other hands-free interfaces – a user could describe an application or script in words and see it materialize, bridging voice interaction with code synthesis.

Academic Research and Lab Projects

InteractiveOmni – Open-Source Audio-Visual Dialogue LLM (2025)

InteractiveOmni [14] is a research project from SenseTime Research that released a unified **omni-modal** large language model capable of multi-turn audio-visual interactions. It integrates a vision encoder, audio encoder, core LLM, and speech decoder all in one model, enabling it to both **understand** and **generate** multimodal content [14]. For instance, InteractiveOmni can take video frames plus live audio as input and carry on a conversation about what's happening – listening to spoken language and viewing images/video concurrently. With 4–8 billion parameters, it's relatively lightweight and optimized for real-time use. Experiments show it can handle long conversations with memory and outperform other open models on combined image, audio, and video understanding benchmarks [15]. The project is open-source (with a GitHub repository), providing a foundation model for developers looking to build agents that **see and hear** the world.

Multi-RAG – Multimodal Retrieval-Augmented Assistant (2025)

Researchers at Johns Hopkins University and the US Army Research Lab proposed **Multi-RAG**, a system that combines real-time video, audio, and text streams with retrieval-augmented generation techniques [16]. The aim is to create an adaptive assistant that can handle **dynamic, information-rich scenarios** (like a live meeting, a class lecture, or an AR guidance session). Multi-RAG's pipeline uses automatic speech recognition to ingest audio, computer vision to encode video frames, and a retrieval module to pull in relevant external knowledge, all feeding into an LLM that can reason and answer user queries [17]. In a demo scenario, the system observes a scene (through a camera) and listens to a user's question, then retrieves pertinent facts and produces an informed answer grounded in both what it saw and heard [18]. This research highlights how offloading cognitive burden to AI is possible by fusing modalities – for example, a future Mentra Glass app could transcribe a conversation and identify visual cues, then automatically draft a situational report or decision brief for the wearer. Multi-RAG's evaluation on a video understanding benchmark showed it outperforms other vision-language models while using fewer resources [19], indicating promise for efficient multimodal assistants.

Multi-Modal Meeting Summarizer (UIUC/RPI, 2019)

Academic work has also explored using video in addition to audio to improve meeting summarization. Li *et al.* (ACL 2019) developed an **abstractive meeting summarizer** that takes both the transcript and the meeting video as inputs [20]. Their system uses a hierarchical attention model over the transcript (at the segment, utterance, and word levels) and incorporates a novel feature: **visual focus of attention** [20]. Essentially, it tracks when participants are visually focusing on a speaker (using the video feed) under the assumption that utterances receiving more collective attention are more important. By jointly modeling topic segmentation and summarization, and using cues like who is looking at whom, the system produced more focused summaries of multi-person meetings [20]. This research from 2019 demonstrated that adding vision (e.g. recognizing the active speaker or slides on screen) can make automatically generated minutes more on-topic and relevant, compared to using transcript text alone.

Beyond Text: Emerging Multimodal Agents

Multiple labs and companies are now pushing the frontier of **multimodal AI agents** that operate in real-world environments. For example, Meta AI's **SeamlessM4T** and Microsoft's **Phi-4** are new multimodal models that understand both speech and images, aiming for "foundation models" that can drive AR assistants or robotics [21]. Another notable effort is **Ultravox** (by Fixie AI), a multimodal LLM that natively accepts audio waveforms as input and generates responses without a separate ASR stage [22]. By converting speech directly into the token space of a language model, Ultravox can conduct voice-based dialogue faster, preserving nuances like tone or emotion in the audio. These advancements, alongside open datasets and benchmarks for audio-visual tasks, are accelerating research on agents that can observe and act. In practical terms, this means a future system could wear a device like Mentra Glass, *hear* a conversation, *see* the environment, and automatically output useful content – from meeting notes and task lists to generated code or proactive reminders – all in real time. The convergence of Whisper-level transcription, OpenCV/MediaPipe vision (for gesture or scene understanding), and powerful LLMs is making such multimodal "copilots" an active area of both open-source development and academic inquiry.

Sources: The information above is drawn from project documentation and research papers, including GitHub repositories (for open-source code) and academic publications. Key references include the Meetily AI meeting assistant [1] [2], the Pipecat framework [3] [5], the WhisperX summarizer [6], Slack-integrated meeting bot [8], GraphMeeting mind-map tool [9], the pyannote-whisper diarization integrator [11], a voice-to-code assistant demo [12] [13], SenseTime's InteractiveOmni paper [14], the JHU/ARL Multi-RAG study [16], and an ACL 2019 multimodal summary paper [20], among others. These illustrate the state of the art in combining real-time audio and video inputs to generate useful outputs like documents, code, or automated messages.

---

[1] [2] GitHub - Zackriya-Solutions/meeting-minutes: A free and open source, self hosted Ai based live meeting note taker and minutes summary generator that can completely run in your Local device (Mac OS and windows OS Support added. Working on adding linux support soon) https://meetily.zackriya.com/ is meetly ai
https://github.com/Zackriya-Solutions/meeting-minutes

[3] [4] [5] GitHub - pipecat-ai/pipecat: Open Source framework for voice and multimodal conversational AI
https://github.com/pipecat-ai/pipecat

[6] [7] GitHub - sidedwards/ai-video-summarizer: An AI-powered tool for transcribing, summarizing, and creating smart clips from video and audio content.
https://github.com/sidedwards/ai-video-summarizer

[8] [9] [10] [11] meeting-summarization · GitHub Topics · GitHub
https://github.com/topics/meeting-summarization

[12] [13] How I Built a Voice-Controlled Python Assistant That Writes and Executes Code Using Whisper + GPT | by Zain Ahmad | Python in Plain English
https://python.plainenglish.io/how-i-built-a-voice-controlled-python-assistant-that-writes-and-executes-code-using-whisper-gpt-41c016e6e4a7?gi=9ffabe90983b

[14] [15] InteractiveOmni: A Unified Omni-modal Model for Audio-Visual Multi-turn Dialogue
https://arxiv.org/html/2510.13747v1

16 17 18 19 Multi-RAG: A Multimodal Retrieval-Augmented Generation System for Adaptive Video Understanding
https://arxiv.org/html/2505.23990v1

20 Keep Meeting Summaries on Topic: Abstractive Multi-Modal Meeting Summarization - ACL Anthology
https://aclanthology.org/P19-1210/

21 22 GitHub - fixie-ai/ultravox: A fast multimodal LLM for real-time voice
https://github.com/fixie-ai/ultravox