Chatbot for Youtube video

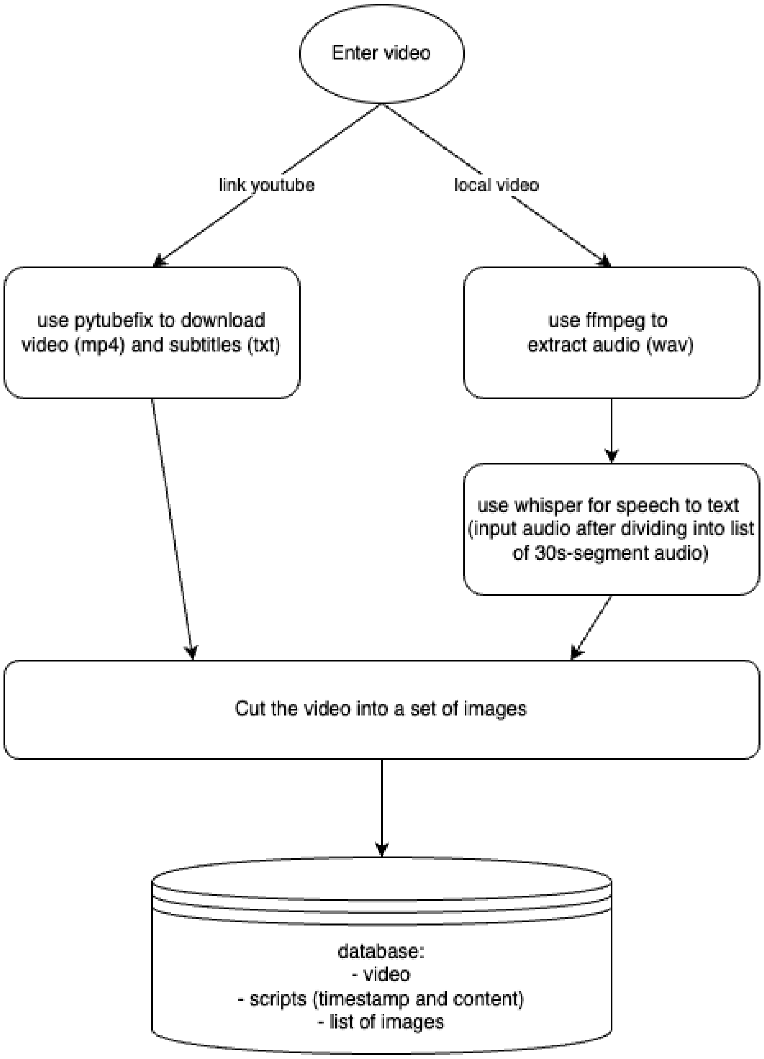
My pipeline consists two phase:

- Init stage: User enters link of Youtube video or path to the video file in user's computer.
- Question-answering stage: User enters query and system outputs answer

You can see file **pipeline.py** to understand more about system.
Source code: https://github.com/PhongNTDo/Video_chatbot/tree/main

Init phase:

My system has two different processing streams depending on whether the user enters a Youtube link or a local path to video.


Enter video
link youtube
local video
use pytubefix to download video (mp4) and subtitles (txt)
use ffmpeg to extract audio (wav)
use whisper for speech to text (input audio after dividing into list of 30s-segment audio)
Cut the video into a set of images
database:
- video
- scripts (timestamp and content)
- list of images

At the end of initiative stage: the system has a database includes: video, list of image (extracted from video), script of video (timestamp: text)

Questions answering phase:

1. The system can answer user's text questions. The system receives the question and prompts the LLM model. The prompt includes the video script information and the user's question. Then, the system receives the answer from the LLM and responds to the user.
   Note: Due to time constraints, I thought this could be improved by adding some of the following processing:
   - The component rejects answers, questions, and normalizes answers to be more to the point and concise for the user's question.
   - An LLM fine-tuned with instructed data better fits the format of the scenario.
   - A RAG to retrieve the correct script containing the answer content, avoiding the case where the LLM has to read a script that is too long, leading to hallucination.
2. User can enter image, then the system will find similar frames that appear in the video and its time. (but i cannot complete it on time)
   Note: Due to time constraints and lack of in-depth research on this problem, I have not been able to fully complete the system. Some essential features:
   - User enters an image and can ask along with a question. At this point can use some model like GPT 4o to handle multimodal tasks.
   - OCR task to get text information in video that is not mentioned in script.