Project Specification

1. Introduction

The modern digital landscape presents an overwhelming volume of information, complicating efficient consumption. This project proposes an AI-driven solution for intelligent text summarization, dynamic content adaptation, and personalized content curation. The system will focus on transforming information from text, audio, and video into concise, coherent summaries.

2. Objective

To develop a multi-modal language model that enhances content consumption by making it faster, more personalized, and engaging. This model will process text, audio, and video inputs to provide summarized outputs, improving users' efficiency in navigating information-dense environments.

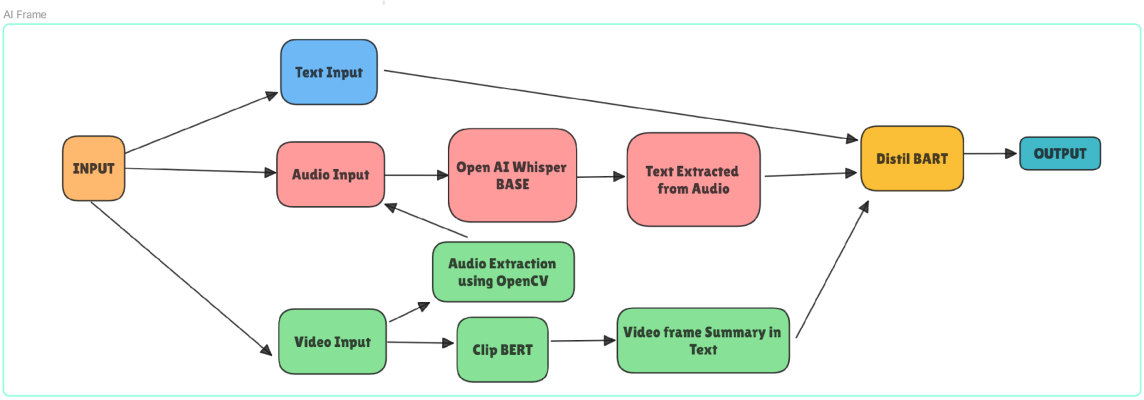
3. Problem Statement

The increasing volume of online content challenges users to stay informed or entertained efficiently. Existing solutions lack the ability to handle multiple input formats or adapt dynamically to user preferences. A system capable of integrating text, audio, and video inputs to produce concise, personalized summaries is required.

4. Proposed Solution

Our proposed solution integrates state-of-the-art language models to handle various input types (text, audio, video) and process them into meaningful, summarized content. The system is designed to be versatile, allowing users to input information in different formats and receive coherent outputs. The flowchart provided illustrates the architecture of our solution, highlighting the key components and their interactions.

Flow Chart



5. Flowchart Explanation:

a. Input Layer

Text Input: Directly accepts text files like .txt ,.pdf,.docx file extensions, which can be processed for summarization.

Audio Input: Accepts audio files. The audio data is first processed using Open AI Whisper BASE, a speech-to-text engine that converts spoken words into text format.

Video Input: Accepts video files. Audio is extracted from the video using OpenCV python library, while video content is analyzed using Clip BERT for generating textual summaries of video frames.

b. Processing Layer

Open AI Whisper BASE: This component takes audio input and transcribes it into text. Whisper BASE is a robust, pre-trained model known for its accuracy in recognizing speech, even in noisy environments. The transcribed text is then fed into the next stage for summarization.

Whisper BASE offers a balanced solution for speech recognition tasks, providing good accuracy and real-time performance while being lightweight and efficient. It is suitable for deployment in environments with limited computational resources and for applications requiring multilingual support. Choosing Whisper BASE over other models is justified when the need for a smaller, faster, and versatile ASR model aligns with the application's requirements. Its ability to handle real-time processing, adaptability through fine-tuning, and strong multilingual capabilities make it a robust choice for a wide range of ASR applications.

Clip BERT: A pre-trained model specifically designed for video understanding. It processes video frames and generates descriptive summaries in text format, which can be further summarized and adapted according to user needs.

Choosing CLIP BERT over other lightweight models depends on the specific requirements of the task. Its ability to handle both image and text data simultaneously makes it a compelling choice for multimodal tasks, offering a balanced combination of efficiency, versatility, and performance. Its design allows for real-time processing, adaptability through fine-tuning, and efficient deployment in resource-constrained environments, making it a versatile tool in the rapidly evolving field of multimodal AI applications

Audio Extraction using OpenCV: For video inputs, OpenCV is utilized to extract the audio track. This extracted audio is then fed into the Whisper BASE model for transcription.

c. Text Integration

Text Extracted from Audio: Once the audio input is transcribed into text by Whisper BASE, it is available as a text entity, similar to direct text input. This step ensures that all content—whether originally in text, audio, or video form—is unified into a single text-based format for consistent processing.

Video Frame Summary in Text: This output from Clip BERT provides a textual representation of the video content, allowing the same summarization process to handle video inputs as well.

d. Summarization Layer

Distil BART: A lightweight version of BART (Bidirectional and Auto-Regressive Transformers), specifically trained for text summarization tasks. Distil BART takes the combined textual data from various input sources and processes it to generate a concise summary.

DistilBART uses a distillation technique, which reduces the number of parameters by approximately 60% compared to the full BART model while retaining around 95% of its language understanding capabilities. This makes it a smaller, faster model that requires less computational resources. This makes it able to handle long text sequences efficiently, making it suitable for summarizing the extracted content.

e. Output Layer

Output: The final summarized content is presented to the user. This output is coherent, condensed, and provides the key information derived from the original input, whether text, audio, or video. The output can be adapted for various applications, such as educational materials, news briefings, entertainment summaries, and accessibility solutions.

Performance Estimates

The proposed model approximately taking the time given in the following table

	CPU	GPU
Text	More than 200 words/ sec	More than 500/sec
Audio	More than 0.8min/sec	More than 1.6min audio processed per sec
Video	More than 9 sec video processed per sec	More than 16 sec video processed per second

Performance is evaluated in a T4 GPU 15 GB VRAM(model only uses 4-5 GB at max out of it) and default Google Colab CPU.

Conclusion

This solution offers a powerful way to consume content from different formats more effectively. By leveraging state-of-the-art technologies like Whisper BASE, OpenCV, Clip BERT, and Distil BART, our system ensures that users receive accurate, personalized, and concise information. This capability enhances the user experience in a digital world overflowing with information, supporting better decision-making, learning, and entertainment.