Narrate AI

PROJECT PRESENTATION

Guide

Ms. Amitha Mathew
Asst.Professor
Dept. of CSE

Group 4

Nandhana Suffin (U2103148)
Nikhil Stephen (U2103155)
Niveditha B. (U2103162)
Rachel Jacob (U2103168)

Contents

Problem Definition

The project addresses the challenge of making visual content more accessible to the blind by automating the generation of audio descriptions, using deep learning to produce synchronized, non-overlapping descriptions.

Purpose & Need

To develop a project that automates the generation of Audio Descriptions (AD) for Blind visually impaired (BVI) people, making visual content more accessible.

The **need** arises because manually creating AD is time-consuming, costly, and not widely available.

Project Objective

The objectives of this project are:

- Use **deep learning** to develop a system to automatically generate AD.

- Ensure that the AD is synchronized with the video's **scene change** without overlapping, allowing for smooth integration.

Innovativeness And Novelty

This project introduces an assistive system that automatically generates scene descriptions , enabling visually impaired individuals to experience rich, contextual storytelling.

By integrating deep learning models it ensures real-time, meaningful narration beyond traditional audio descriptions.

Literature Survey

Title	Dataset	Methodology	Result	Advantages	Disadvantages
Machine Generation of Audio Description (2023)	ImageNet (ILSVRC)	Applies machine learning to automate audio descriptions	Generates automated audio descriptions for videos	Automates audio descriptions, making content more accessible	May miss nuances important for full understanding
STAT: Spatial-Temporal Attention Mechanism for Video Captioning (2020)	MSVD, MSR-VTT-10	Enhances video captioning by jointly modelling spatial (object-level) and temporal (frame-level) attention in an encoder-decoder framework	Automatically generates natural language description for video	Reduces errors, Captures fine details	Computationally heavy, Depends on object detection accuracy, Limited gains on MSR-VTT-10K
A Video Captioning Method by Semantic Topic-Guided (2024)	MSRVTT	Uses semantic topic modeling to guide caption generation	Context-based captions enhance user understanding	Provides context-based captions, enhancing comprehension	Requires high-quality input data for effective results

Fine-Grained Image Captioning with Global-Local Discriminative Objective (2020)	MS-COCO	Proposed a global-local discriminative objective with global and local constraints to improve image captioning accuracy and detail.	Outperformed baseline methods significantly, achieving competitive performance on MS-COCO with a notable increase in CIDErscores	Generates more fine-grained and discriminative captions; addresses uneven word distribution issues; enhances the quality of descriptions	Tends to generate captions that may not match ground truth; challenges with adaptive threshold settings for local constraints
TimeChat: A Time-sensitive Multimodal Large Language Model (2024)	TimeIT	Combines multiple modalities and time-sensitive analysis	Improves user experience using diverse data types	Integrates multiple data types for better user experience	Complexity can hinder accessibility for some users

Proposed Method

1.Scene Change and Language Identification.

2.Frame Extraction.

3.Scene description generation.

4. .srt file formation

5. Audio file generation and synchronization.

Proposed Method (Contd.)

1.Scene Change Detection

- The Scene Change Detection module identifies significant visual transitions in the video.
- It determines when a major scene change occurs, ensuring that descriptions are added only when a new scene begins and the corresponding frames extracted.
- This improves contextual accuracy by preventing redundant or unnecessary descriptions.

2.Frame Extraction

- The Frame Extraction module captures the first frame of each detected scene.
- This snapshot represents the visual state of the scene and serves as input for caption generation and object detection.
- By anchoring the description to a single frame, it maintains consistency and focuses on the most relevant visual content.

3.Scene Description generation

- This module generates a detailed textual description of each scene.
- It first analyzes the scene to identify key visual elements and their spatial positions. The initial description is then refined using advanced language processing techniques to ensure clarity and coherence.
- The final output is an informative caption that enhances understanding of the visual content.

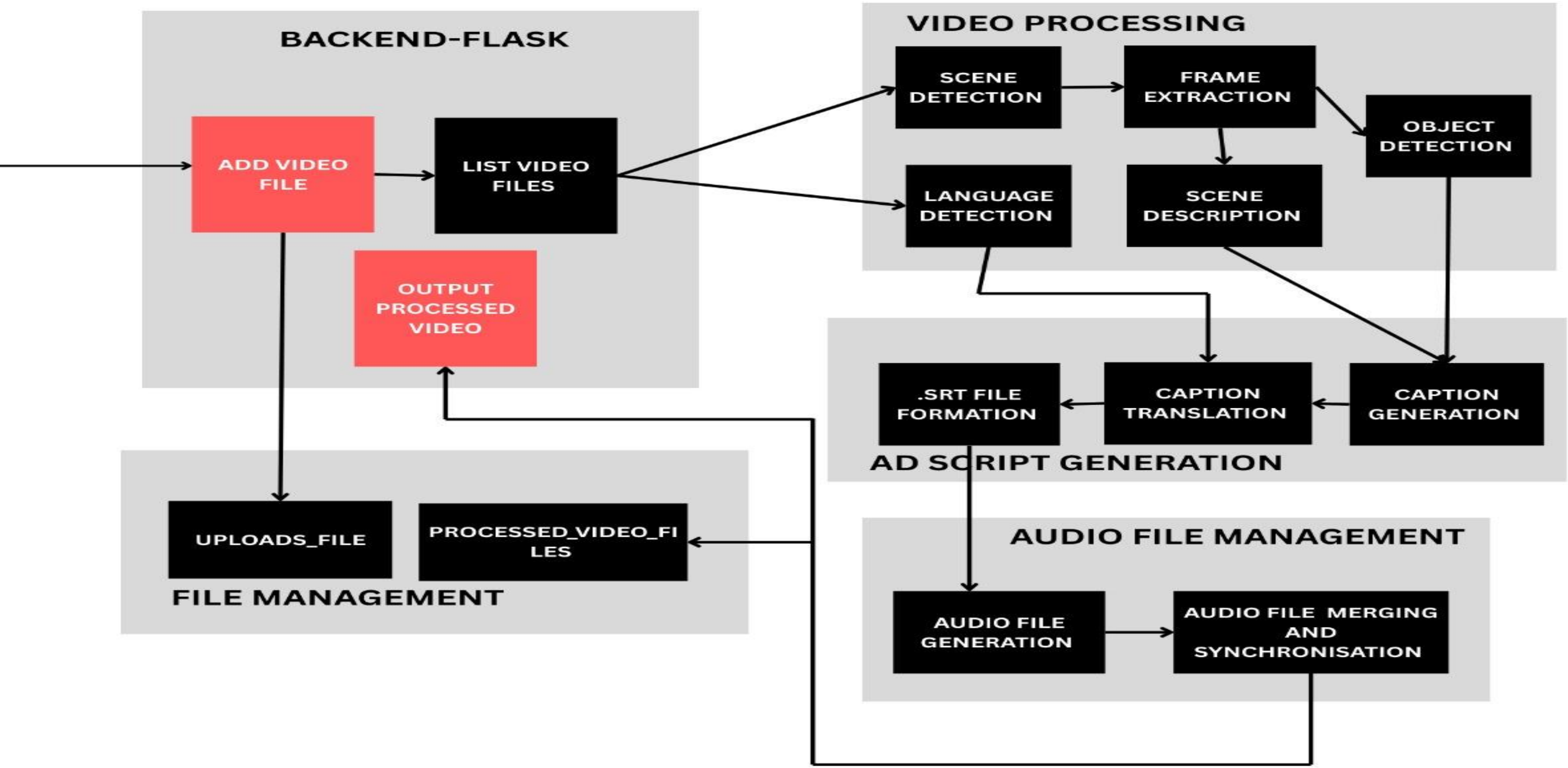
4.srt File Formation

- The Subtitle Generation module timestamps each scene's description and formats it into a standard .srt file.
- Each entry includes a start time, end time, and the corresponding description.
- This allows the captions to be viewed as text alongside the video, improving accessibility and comprehension.

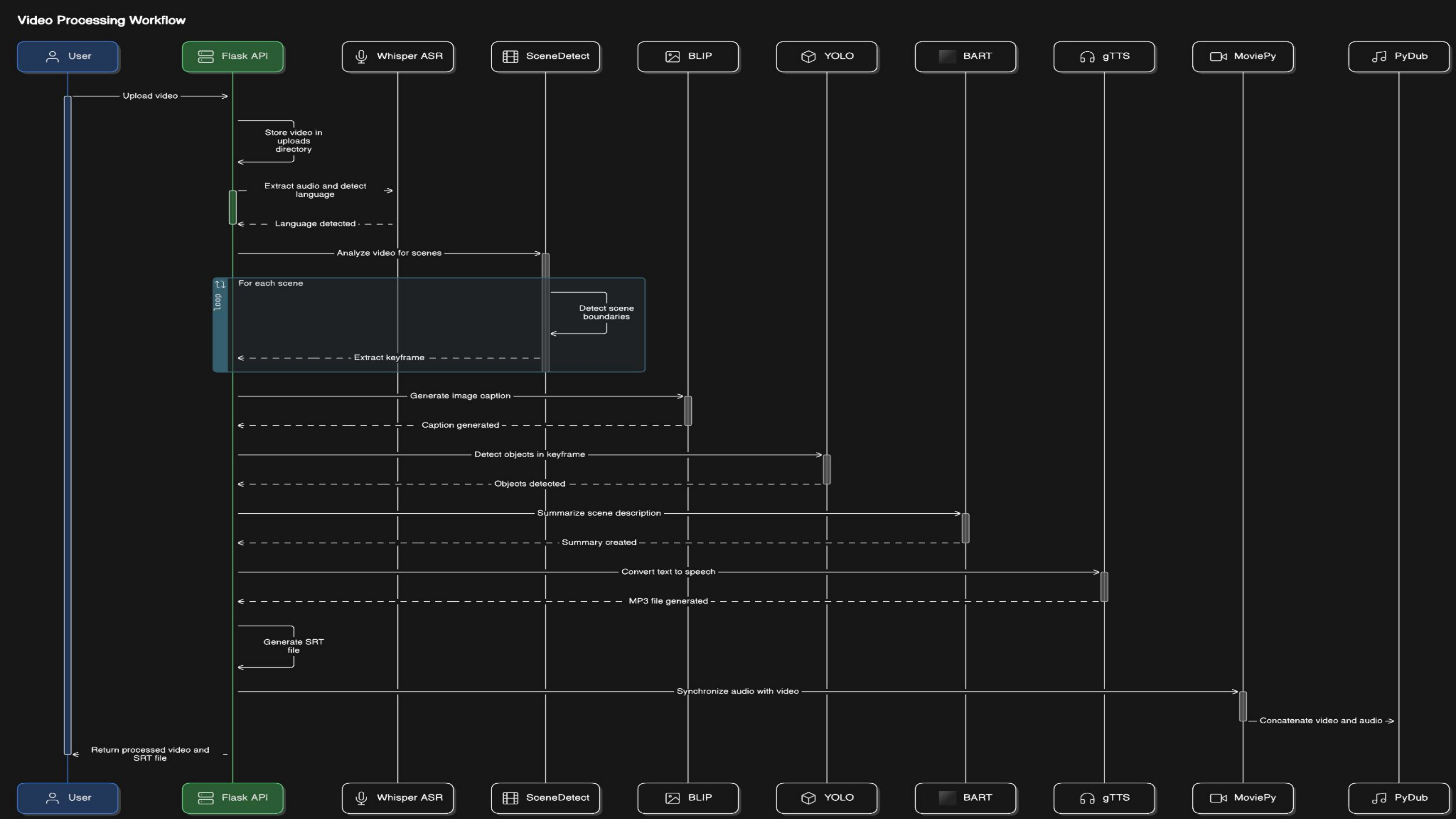
5.Audio file generation and synchronization.

- This module translates the text descriptions into speech using text-to-speech (gTTS) in the appropriate language.
- Each audio clip is synchronized with its corresponding video segment, including both freeze-frames and original scenes.
- This creates a seamless audio description experience, making the video accessible to visually impaired viewers.

ARCHITECTURE DIAGRAM

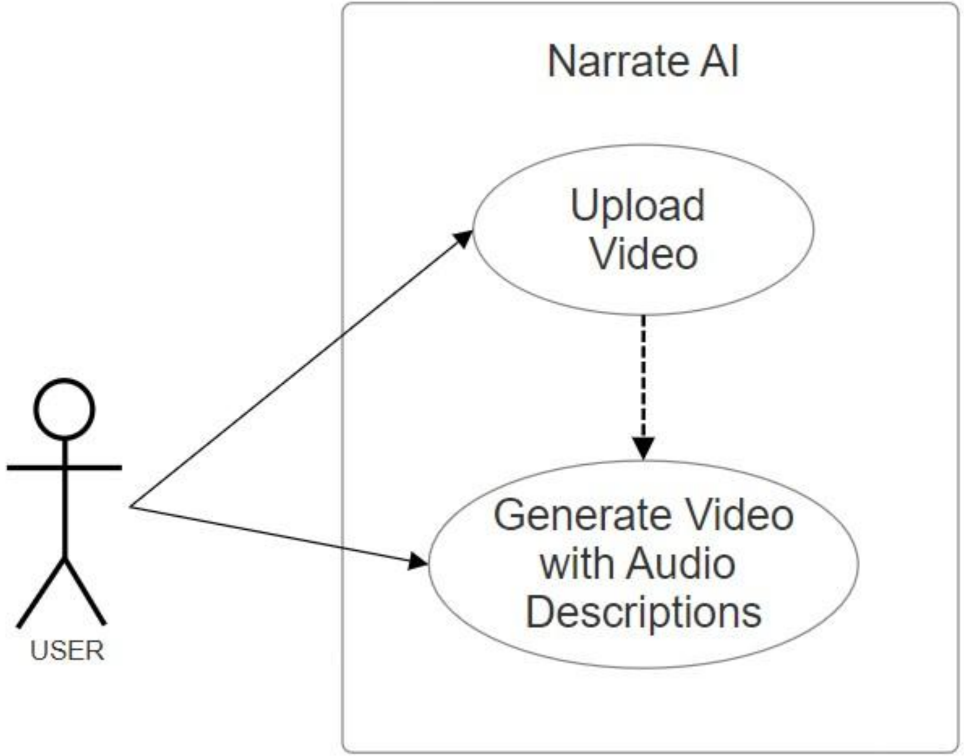

BACKEND-FLASK
ADD VIDEO FILE
LIST VIDEO FILES
OUTPUT PROCESSED VIDEO
VIDEO PROCESSING
SCENE DETECTION
FRAME EXTRACTION
OBJECT DETECTION
LANGUAGE DETECTION
SCENE DESCRIPTION
.SRT FILE FORMATION
CAPTION TRANSLATION
CAPTION GENERATION
AD SCRIPT GENERATION
UPLOADS_FILE
PROCESSED_VIDEO_FI LES
FILE MANAGEMENT
AUDIO FILE MANAGEMENT
AUDIO FILE GENERATION
AUDIO FILE MERGING AND SYNCHRONISATION

SEQUENCE DIAGRAM



Video Processing Workflow
User
Flask API
Whisper ASR
SceneDetect
BLIP
YOLO
BART
gTTS
MoviePy
PyDub
Upload video
Store video in uploads directory
Extract audio and detect language
Language detected
Analyze video for scenes
Loop
For each scene
Detect scene boundaries
Extract keyframe
Generate image caption
Caption generated
Detect objects in keyframe
Objects detected
Summarize scene description
Summary created
Convert text to speech
MP3 file generated
Generate SRT file
Synchronize audio with video
Concatenate video and audio
Return processed video and SRT file

Use Case Diagram



Modules

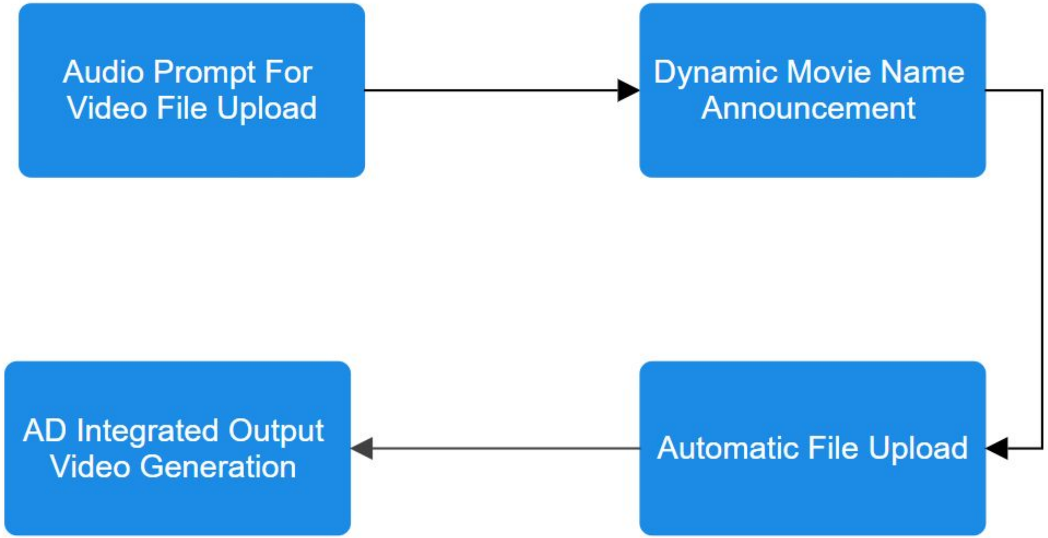
1. Web Interface
2. Scene Change and Frame Extraction
3. Object Detection and Image Captioning
4. Scene-Level Caption Generation
5. SRT File Update
6. Audio Description Generation and Enhancement
7. Appending Audio to Video and Output Generation

Web Interface

- **Frontend**:
  - **HTML/CSS** for layout and styling.
  - **JavaScript** for interactivity and audio feedback using the **Web Speech API**.
- **Backend**:
  - **Flask (Python)** for handling file uploads and processing.

Web Interface(Contd.)

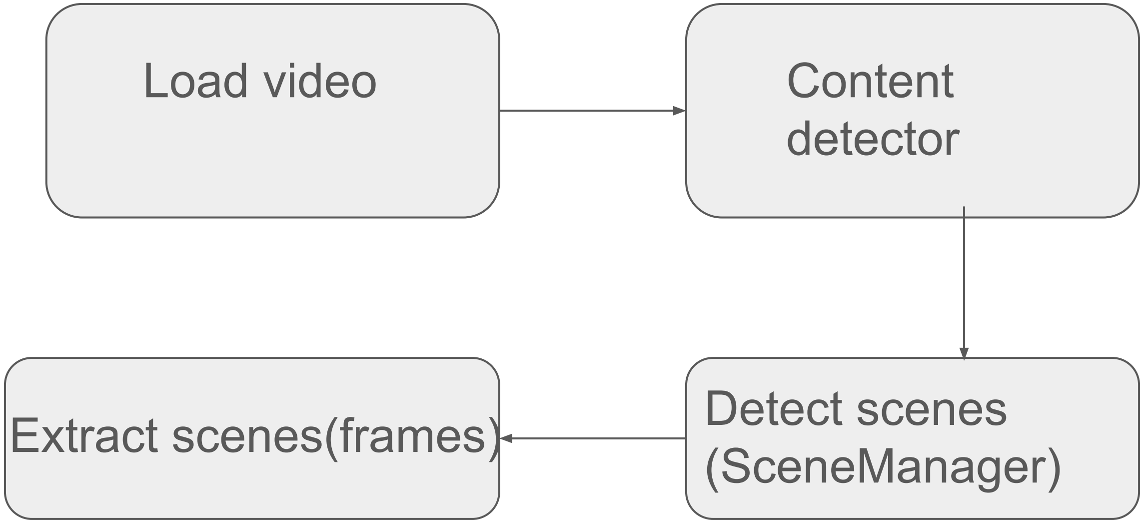
A web application that provides audio feedback for file uploads and movie selections, enhancing accessibility for visually impaired users.



Scene Change Detection and Frame Extraction

- It determines when a major scene change occurs, ensuring that descriptions are added only when a new scene begins.
- When a scene change is detected, the Frame Extraction module captures the first frame of the new scene.
- These frames serve as input for the object detection and caption generation models.
- A temporary storage system ensures efficient handling of extracted frames.

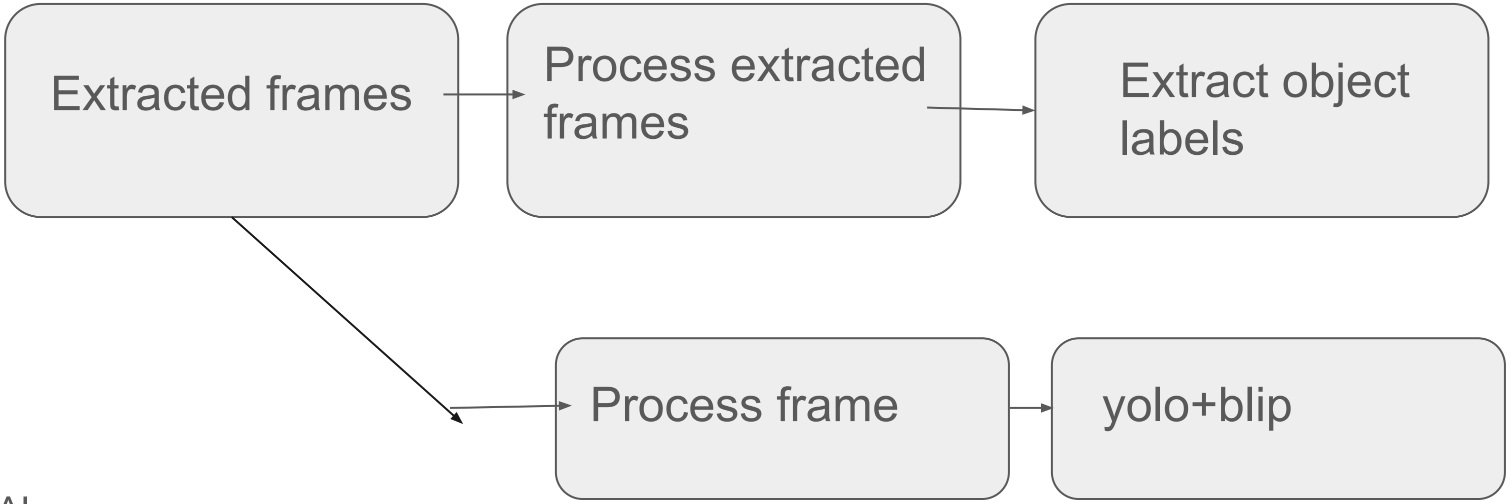
Scene Change Detection and Frame Extraction (Contd.)



Object Detection and Image Captioning

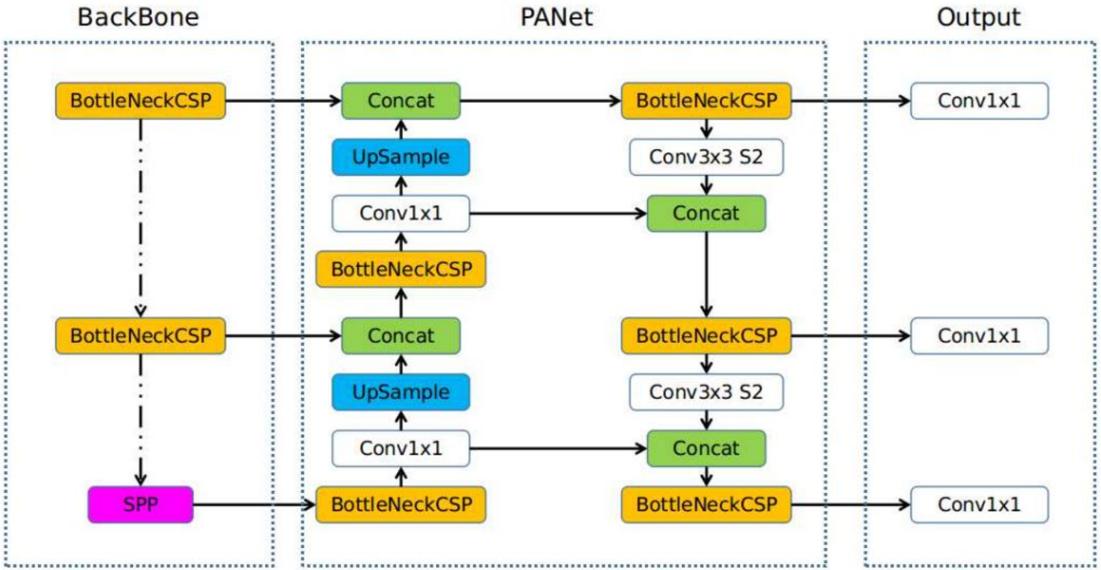
- Extracted frames are processed using a YOLOv5 (You Only Look Once) model to detect objects and visual elements.
- The detected objects are then fed into a BLIP (Bootstrapped Language Image Pretraining) model to generate preliminary image captions.
- By combining YOLOv5 and BLIP outputs, a detailed scene description is created.

Object Detection and Image Captioning



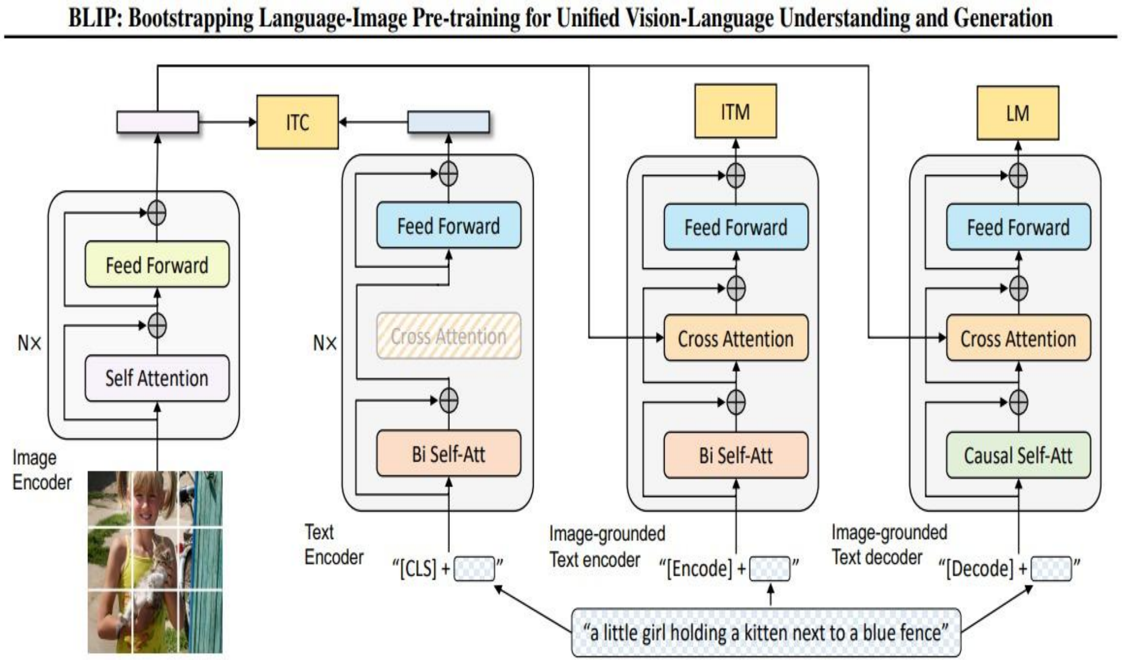
Object Detection(YOLOv5s)

YOLOv5s is a fast and lightweight deep learning model for real-time object detection. It detects and labels multiple objects in images or videos using a single forward pass. It's built with PyTorch and widely used in applications like surveillance and robotics.



BLIP

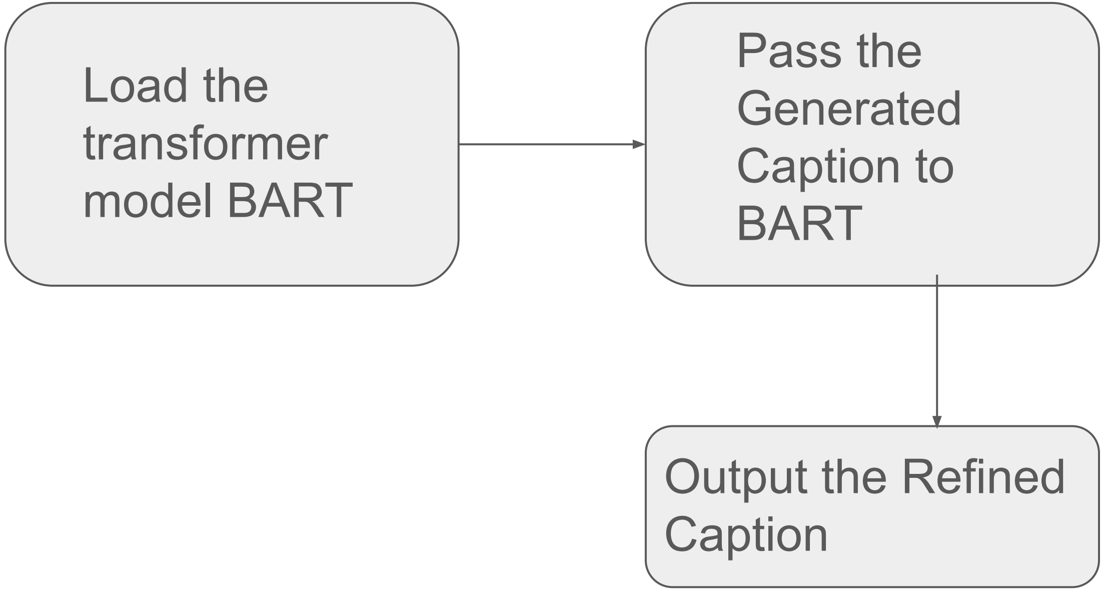
BLIP (Bootstrapping Language-Image Pre-training) is a model for image captioning and vision-language tasks. It uses a Vision Transformer and a language decoder to generate natural, context-aware captions from images.



Scene-Level Caption Generation

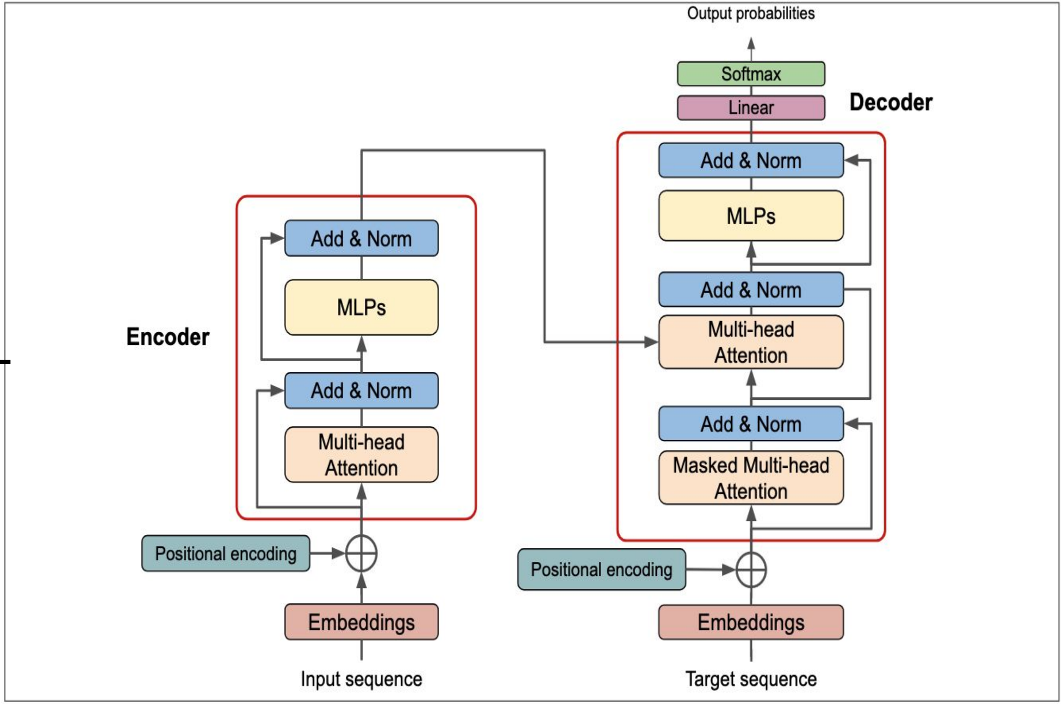
- To improve temporal coherence, a BART (Bidirectional and Auto-Regressive Trans-
- former) model refines the captions by incorporating linguistic structure and scene context.
- This ensures that captions are not only accurate but also readable and natural.

Scene-Level Caption Generation



BART

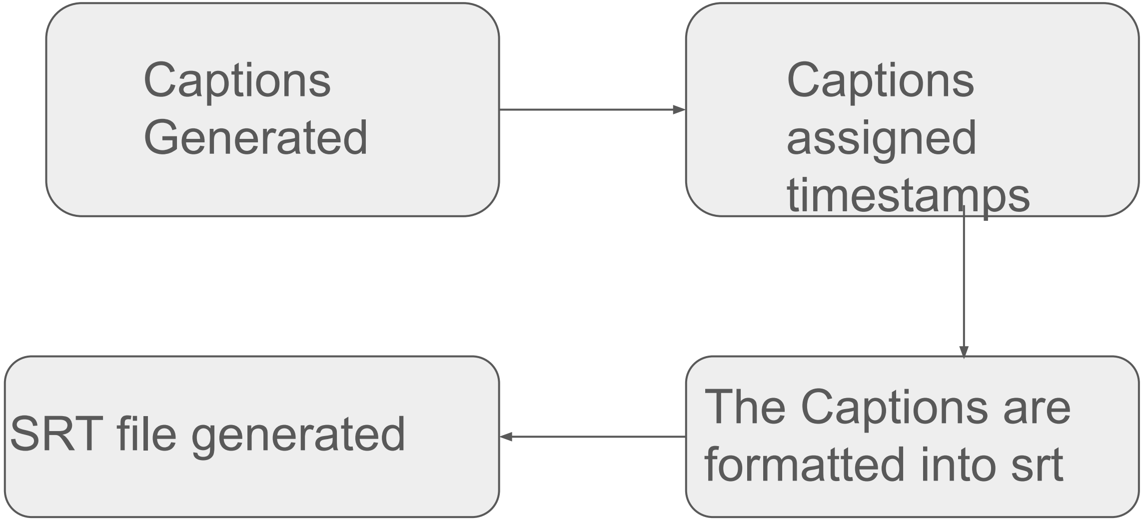
BART (Bidirectional and Auto-Regressive Transformers) is a sequence-to-sequence language model. It combines the strengths of BERT and GPT . BART is commonly used for tasks like text summarization, translation, and caption generation.



SRT File Update

- The generated captions are converted into timestamped subtitles and integrated into an SRT (SubRip Subtitle) file
- This ensures that descriptions align properly with scene changes in the video.

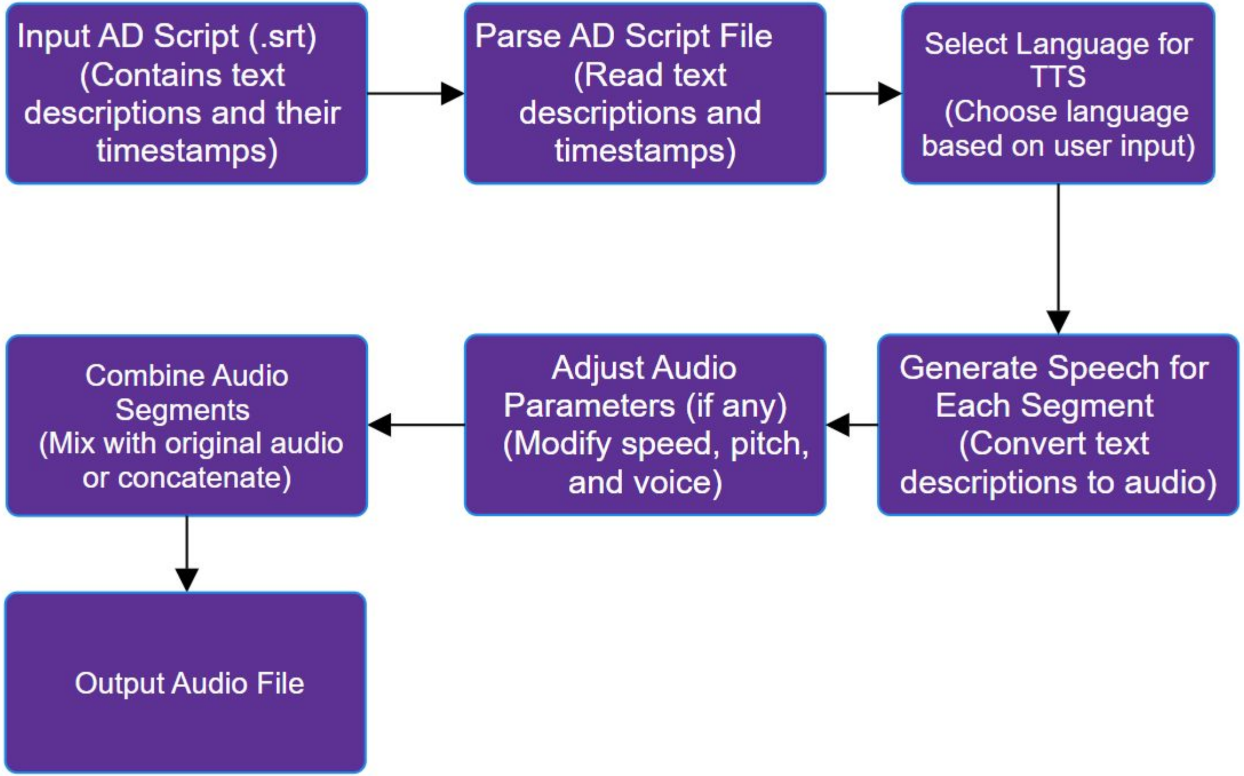
SRT File Update



34

Audio Description Generation and Enhancement

- The finalized text descriptions are passed to a Text-to-Speech (TTS) engine, which synthesizes audio.
- The system takes into consideration the size ofthe description and speed to provide clear and engaging audio descriptions.


Input AD Script (.srt) (Contains text descriptions and their timestamps)
Parse AD Script File (Read text descriptions and timestamps)
Select Language for TTS (Choose language based on user input)
Generate Speech for Each Segment (Convert text descriptions to audio)
Adjust Audio Parameters (if any) (Modify speed, pitch, and voice)
Combine Audio Segments (Mix with original audio or concatenate)
Output Audio File

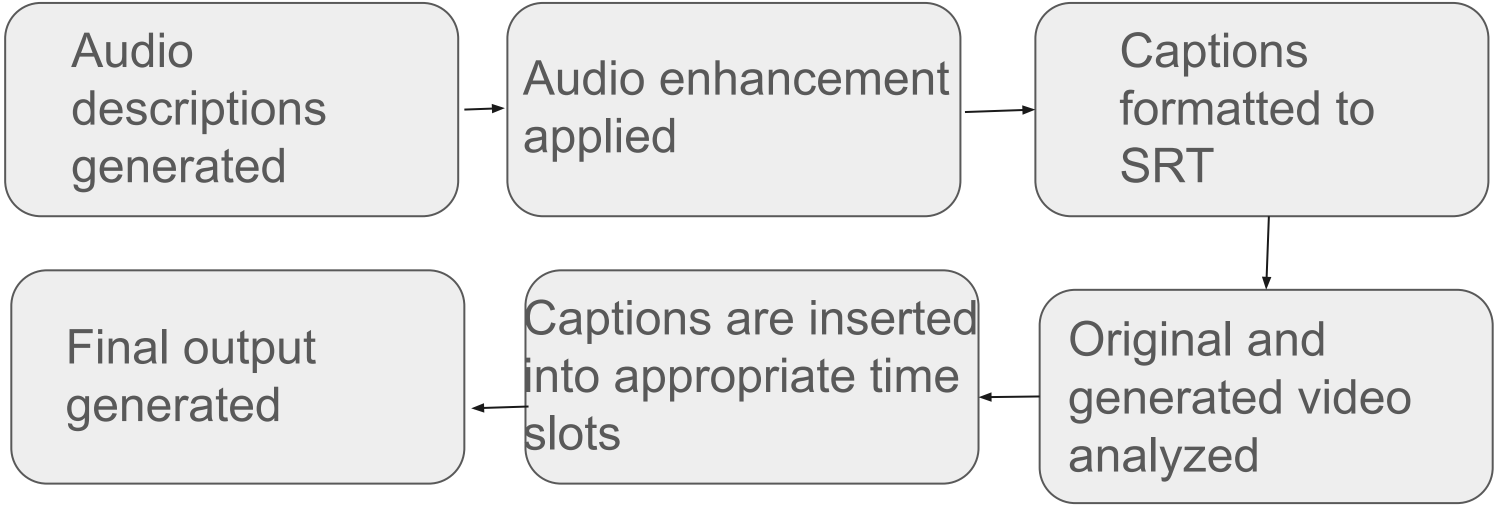
Appending Audio to Video and Output Generation

- Audio descriptions are added without disrupting the original sound.

- The system syncs them with scene transitions for smooth playback.

- The final video is optimized for accessibility and clarity.

Appending Audio to Video and Output Generation



Assumptions

- The input videos are assumed to be of **sufficient resolution** to allow accurate object and scene recognition.
- It is assumed that the project operates within **legal boundaries**, and appropriate permissions for using video content for generating audio descriptions are in place.
- The videos provided are **suitable** for generating audio descriptions.

Work breakdown and responsibilities

1 Nandhana Suffin Video Processing and Object Detection	**2 Nikhil Stephen** Audio Description Script Generation
3 Niveditha B Speech Synthesis and Audio Integration	**4 Rachel Jacob** Web Application and User Interface

Hardware & Software requirements

Hardware:

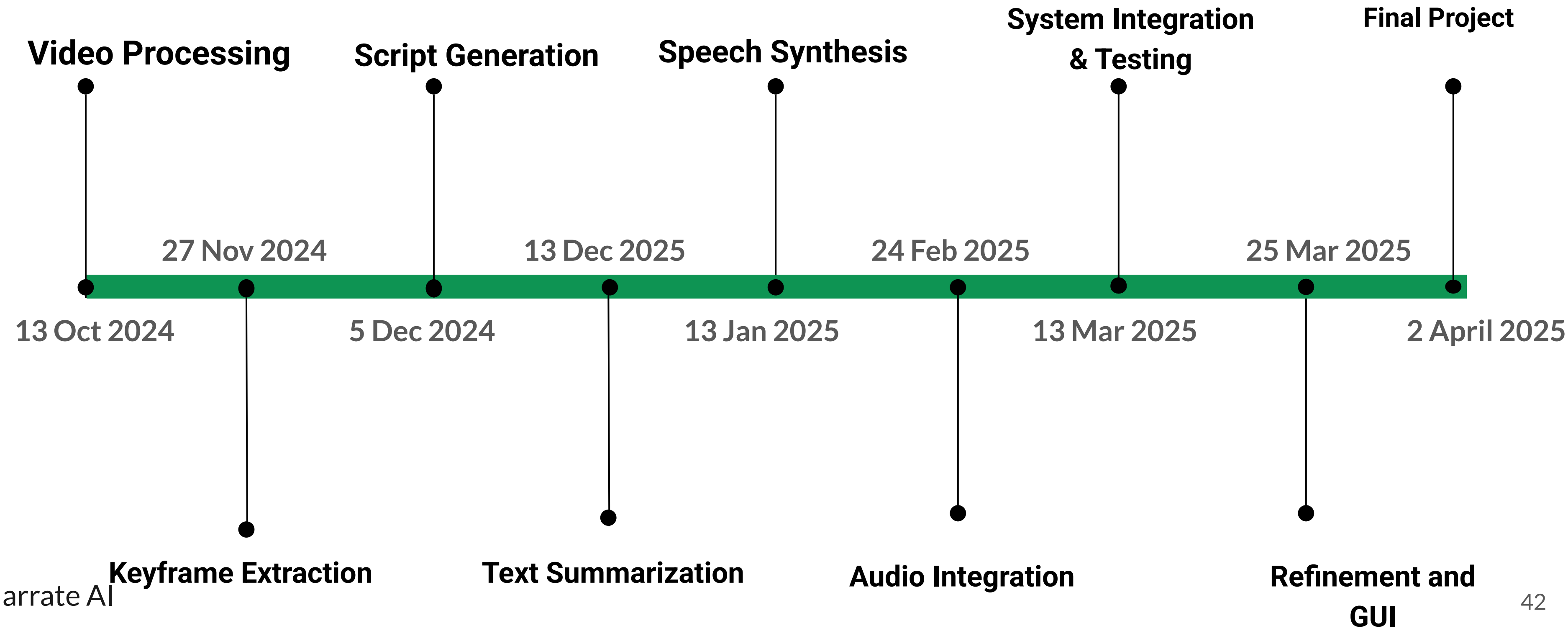
Minimum Specification:

- i5 or Ryzen 5 processor
- 16 GB RAM
- 512 SSD
- OS: Windows 11 64-bit

Software:

- Development environment (Visual Studio Code)
- Framework : Flask,OpenCV,TensorFlow/ Pytorch ,YOLOv5
- Audio Processing: gTTS,PyDub,Speech Recognition

GANTT CHART



Video Processing
Script Generation
Speech Synthesis
System Integration & Testing
Final Project
27 Nov 2024
13 Dec 2025
24 Feb 2025
25 Mar 2025
13 Oct 2024
5 Dec 2024
13 Jan 2025
13 Mar 2025
2 April 2025
Keyframe Extraction
Text Summarization
Audio Integration
Refinement and GUI

Risks & Challenges

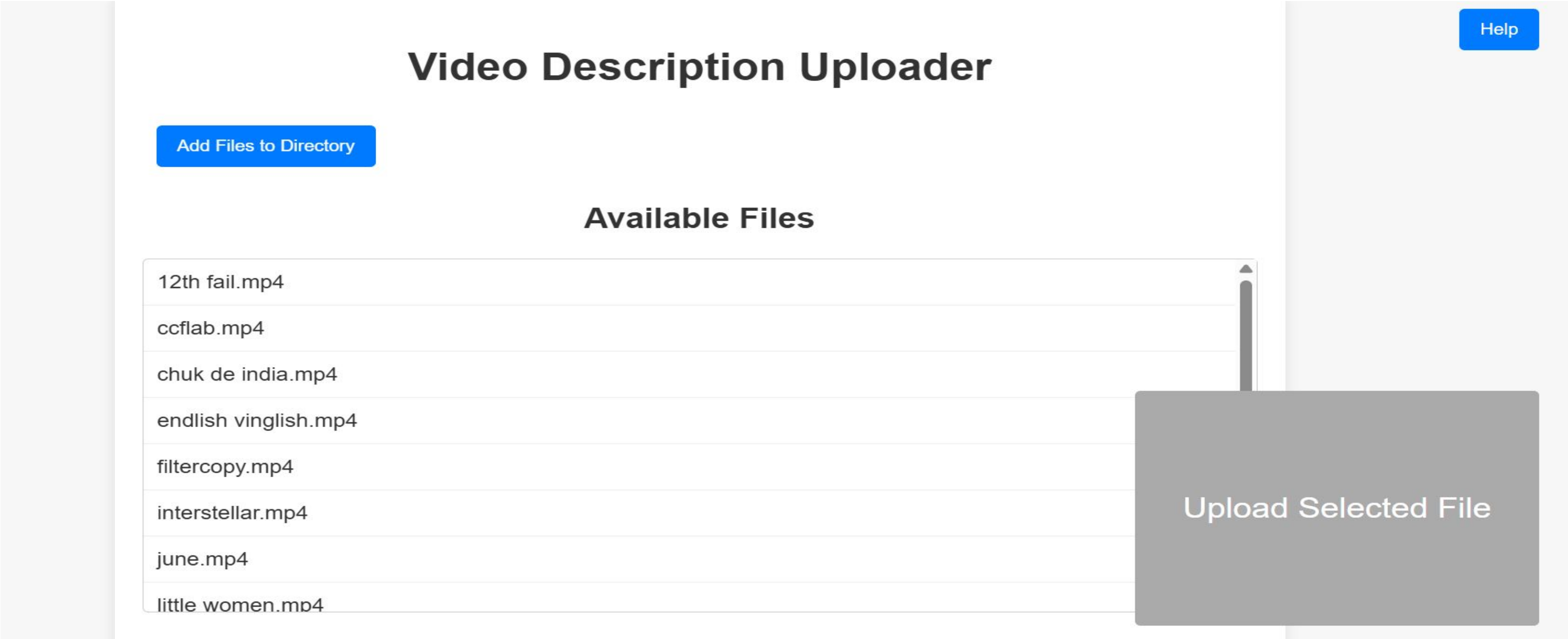
Fitting Descriptions :ADs must be inserted during the scene Change of a video to avoid overlapping with dialogues or sound effects.

Performance and Scalability: Processing large amounts of video data efficiently and quickly to generate ADs could pose performance challenges, especially when dealing with diverse video types and lengths

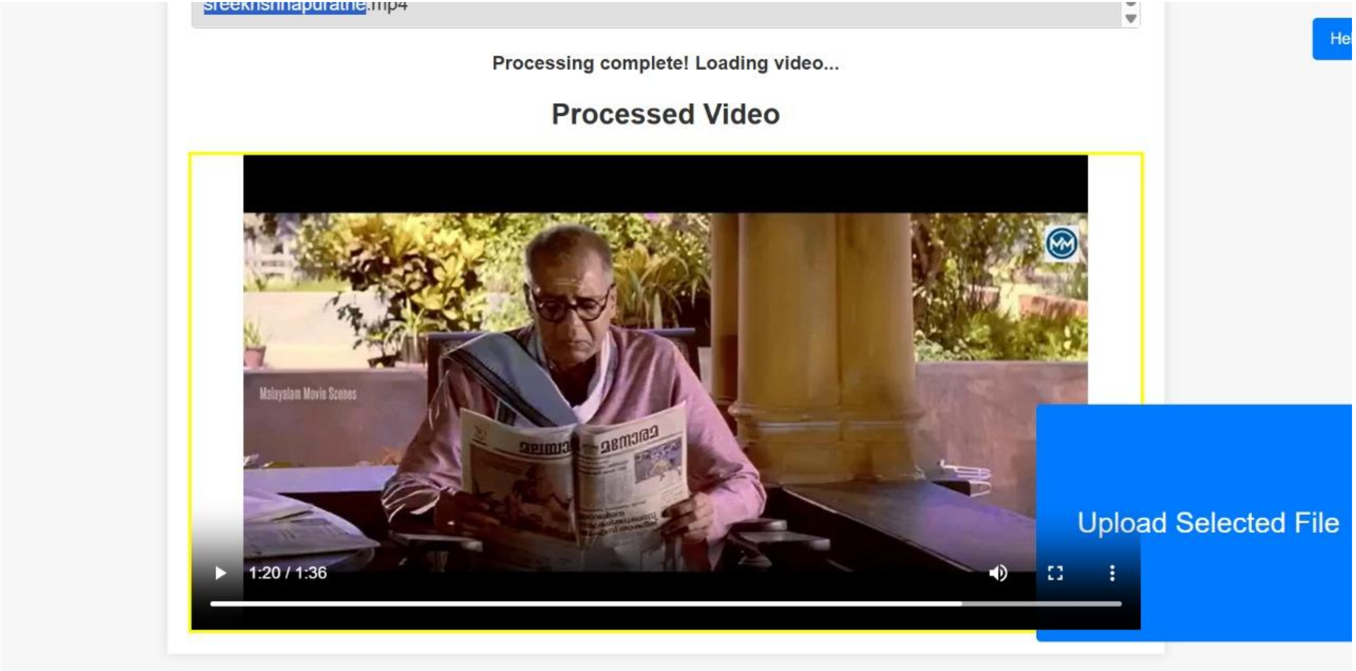
RESULTS

- BLIND FRIENDLY GUI
- AUDIO INTEGRATED OUTPUT VIDEO
- .srt FILE WITH CAPTIONS
- LANGUAGE DETECTED

BLIND FRIENDLY GUI



AUDIO INTEGRATED OUTPUT VIDEO



.srt FILE WITH CAPTIONS

```
1
00:00:00,000 --> 00:00:09,000
The video shows a man and woman playing with a baby. In this scene, you can see a person
is in the bottom-right part of the scene.

2
00:00:27,000 --> 00:00:36,000
The video shows a woman sitting on a couch eating food. In this scene, you can see a
person is in the middle-center part of the scene.
```

.srt FILE WITH CAPTIONS

captions.srt

File Edit View

```
1
00:00:00,000 --> 00:00:10,000
वीडियो एक महिला को काली शर्ट में दिखाता है।इस दृश्य में,  आप देख सकते हैं कि एक व्यक्ति दृश्य के मध्य-सही हिस्से में है।

2
00:00:28,000 --> 00:00:41,000
वीडियो से पता चलता है कि एक महिला बाथरूम में अपने बाल धो रही है।इस दृश्य में,  आप देख सकते हैं कि एक व्यक्ति दृश्य के
मध्य-केंद्र भाग में है।
```

.srt FILE WITH CAPTIONS

captions.srt

File Edit View

```
1
00:00:00,000 --> 00:00:10,000
ഒരു വീടിന് ചുറ്റും നിൽക്കുന്ന ഒരു കൂട്ടം ആളുകൾ കാണിക്കുന്നു.ഈ രംഗത്തിൽ,  ഒരു
വ്യക്തി രംഗത്തിന്റെ മധ്യ കേന്ദ്രത്തിൽ ഭാഗത്താണെന്ന് നിങ്ങൾക്ക് കാണാൻ കഴിയും.

2
00:00:31,000 --> 00:00:43,000
മറ്റൊരു മനുഷ്യനുമായി സംസാരിക്കുന്ന ടാങ്ക് ടോപ്പിൽ ഒരു മനുഷ്യനെ വീഡിയോ
കാണിക്കുന്നു.ഈ രംഗത്തിൽ,  ഒരു വ്യക്തി രംഗത്തിന്റെ മധ്യ കേന്ദ്രത്തിൽ
ഭാഗത്താണെന്ന് നിങ്ങൾക്ക് കാണാൻ കഴിയും.

3
00:01:01,000 --> 00:01:11,000
ഒരു മനുഷ്യനെ മറ്റൊരു മനുഷ്യൻ വഞ്ചിക്കപ്പെടുന്നതായി കാണിക്കുന്നു.ഈ രംഗത്തിൽ,
ഒരു വ്യക്തി രംഗത്തിന്റെ മധ്യ-വലത് ഭാഗത്താണെന്ന് നിങ്ങൾക്ക് കാണാൻ കഴിയും.
```


```
1
00:00:00,000 --> 00:00:10,000
The video shows a group of people standing in front of a house. In this
scene, you can see A car is in the middle-left part of the scene.

2
```


FANDANGO
MOVIECLIPS
00:00:10
00:00:26
processed_video

Language Detected

```
1
00:00:00,000 --> 00:00:10,000
The video shows a group of people standing in front of a house. In this scene, you can
see A car is in the middle-left part of the scene.

1
00:00:00,000 --> 00:00:10,000
ഒരു വീടിന് ചുറ്റും നിൽക്കുന്ന ഒരു കൂട്ടം ആളുകൾ കാണിക്കുന്നു.ഈ രംഗത്തിൽ,  ഒരു
വ്യക്തി രംഗത്തിന്റെ മധ്യ കേന്ദ്രത്തിൽ ഭാഗത്താണെന്ന് നിങ്ങൾക്ക് കാണാൻ കഴിയും.


1
00:00:00,000 --> 00:00:10,000
वीडियो एक महिला को काली शर्ट में दिखाता है।इस दृश्य में,  आप देख सकते हैं कि एक व्यक्ति दृश्य के मध्य-सही हिस्से में है।
```

Conclusion

Our proposed system **advances audio description generation** by automating the process and generating accurate, synchronized descriptions directly from video content. It **enhances accessibility for blind and visually impaired** users by providing efficient and user-friendly audio descriptions, making visual media more inclusive compared to existing methods.

References

- Campos, V.P., Gonçalves, L.M., Ribeiro, W.L., Araújo, T.M., Do Rego, T.G., Figueiredo, P.H., Vieira, S.F., Costa, T.F., Moraes, C.C., Cruz, A.C. and Araújo, F.A., 2023. Machine generation of audio description for blind and visually impaired people. *ACM Transactions on Accessible Computing*, *16*(2), pp.1-28.
- Nandini, H.M., Chethan, H.K. and Rashmi, B.S., 2022. Shot based keyframe extraction using edge-LBP approach. *Journal of King Saud University-Computer and Information Sciences*, *34*(7), pp.4537-4545.
- Han, T., Bain, M., Nagrani, A., Varol, G., Xie, W. and Zisserman, A., 2023. Autoad ii: The sequel-who, when, and what in movie audio description. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 13645-13655).
- Ren, S., Yao, L., Li, S., Sun, X. and Hou, L., 2024. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 14313-14323).
- O. Ye, X. Wei, Z. Yu, Y. Fu, and Y. Yang "A Video Captioning Method by Semantic Topic-Guided Generation," *Comput. Mater. Contin.*, vol. 78, no. 1, pp. 1071-1093. 2024. https://doi.org/10.32604/cmc.2023.046418

References (Contd.)

- Song, E., Chai, W., Wang, G., Zhang, Y., Zhou, H., Wu, F., Chi, H., Guo, X., Ye, T., Zhang, Y. and Lu, Y., 2024. Moviechat: From dense token to sparse memory for long video understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 18221-18232).
- Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P. and Wang, L., 2024. Mvbench: A comprehensive multi-modal video understanding benchmark. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 22195-22206).
- Huang, B., Wang, X., Chen, H., Song, Z. and Zhu, W., 2024. Vtimellm: Empower llm to grasp video moments. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 14271-14280).
- Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L. and Qiao, Y., 2023. Videochat: Chat-centric video understanding. arXiv preprint arXiv:2305.06355.

References (Contd.)

- Wu, J., Chen, T., Wu, H., Yang, Z., Luo, G. and Lin, L., 2020. Fine-grained image captioning with global-local discriminative objective. IEEE Transactions on Multimedia, 23, pp.2413-2427.
- Luo, Y., Ji, J., Sun, X., Cao, L., Wu, Y., Huang, F., Lin, C.W. and Ji, R., 2021, May. Dual-level collaborative transformer for image captioning. In Proceedings of the AAAI conference on artificial intelligence (Vol. 35, No. 3,pp. 2286-2293).
- Fei, Z., Fan, M., Zhu, L., Huang, J., Wei, X. and Wei, X., 2023, June. Uncertainty-aware image captioning. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 1, pp. 614-622).
- Cornia, M., Stefanini, M., Baraldi, L. and Cucchiara, R., 2020. Meshed-memory transformer for image captioning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp.10578-10587).
- Fei, Z., 2022, October. Efficient modeling of future context for image captioning. In Proceedings of the 30th ACM International Conference on Multimedia (pp. 5026-5035).

References(Contd.)

- Lisena, P., Laaksonen, J. and Troncy, R., 2021, June. FaceRec: an interactive framework for face recognition in video archives. In DataTV 2021, 2nd International Workshop on Data-driven Personalisation of Television.
- Singh, G. and Goel, A.K., 2020, March. Face detection and recognition system using digital image processing. In 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA) (pp. 348-352). IEEE.
- Im, D.H., Seo, Y.S., Kim, H., Hwang, E. and Park, J., 2020, October. Person re-identification in movies/dramas. In 2020 International Conference on Information and Communication Technology Convergence (ICTC) (pp. 1596-1598). IEEE.
- Kim, H., Lee, E.C., Seo, Y., Im, D.H. and Lee, I.K., 2020. Character detection in animated movies using multi-style adaptation and visual attention. IEEE Transactions on Multimedia, 23, pp.1990-2004.

FUTURE WORK

- Implement real-time processing so that live video streams (e.g., Zoom, YouTube Live) can be described on the go.
- A predefined character bank can be integrated into the system to recognize and consistently refer to recurring individuals by name.
- Allow users to choose the tone or type of narrator voice (e.g., calm, energetic, robotic, etc.). Accessibility meets personalization

THANK YOU