Enjoyable - a multi-contexts and real-time audio description method on YouTube videos for the Visual Impaired

By

Lee Wei Song

A REPORT

SUBMITTED TO

Universiti Tunku Abdul Rahman

in partial fulfillment of the requirements

for the degree of

BACHELOR OF COMPUTER SCIENCE (HONOURS)

Faculty of Information and Communication Technology

(Kampar Campus)

FEB 2025

COPYRIGHT STATEMENT

ACKNOWLEDGEMENTS

I would like to extend my warmest thanks and gratitude to my supervisor, Dr. Aun Yichiet, for his constant guidance, support, and encouragement throughout this project. His advice and mentorship have been essential in shaping my understanding and contributing to my growth in this field. This opportunity has been a key milestone in my journey, and I am truly thankful for his trust in me.

I am also deeply thankful to my family, whose love, care, and unwavering support have been the foundation of my perseverance. Their encouragement and belief in my abilities have motivated me to push through even the most difficult times. I am forever indebted to them for standing by me throughout this endeavor.

My heartfelt appreciation goes out to the Malaysia Foundation for the Blind, whose support and contributions have been immensely valuable in the progress of this work.

To everyone who has contributed to this project and my path, whether directly or indirectly, I extend my most sincere thanks. Your efforts and support mean more to me than words can convey. Thank you.

ABSTRACT

Video streaming platforms often lack sufficient accessibility features for visually impaired users, as generating audio descriptions (AD) manually is time-intensive and resource-heavy. This project introduces "Enjoyable," an online platform with an automated AD system. Instead of relying on external databases, the system enables content creators to label clustered faces directly within videos, improving character recognition across diverse genres. A structured script-based approach integrates image captions and dialogue, forming a comprehensive movie script. This script is processed by LLM-based Single-Prompt Multiturn Multi-Agent Reasoning System (SMARS), comprising agents—Investigator agent, Visual Validator agent, Context Historian agent, Integrator agent, Audio Describers agent, Syntax Fixers agent, Language Flow Expert agent, Word Count Checker agent, Fact Checker agent, Messenger agent, Target Audience agent—who collaboratively generate personalized ADs. This method enhances accessibility by streamlining visual-auditory data interaction and addressing limitations of existing AD systems.

Area of Study: Humanistic AI

Keywords: Automated Audio Description Generation System, LLM-based Multi Agent System, Multilingual Narration, Single-Prompt Multiturn Multi-Agent Reasoning System, Blind and Visually Impaired People

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AD	Audio Description
LLM	Large Language Model
SSIM	Structural Similarity Index Measure
ORB	Oriented FAST and rotated BRIEF
TTS	Text-to-Speech
STT	Speech-to-Text
MAS	Multi-Agent System
SMARS	Single-Prompt Multiturn Multi-Agent Reasoning System

CHAPTER 1

Project Background

1.1 Introduction

Video streaming platforms have transformed media consumption, granting audiences unparalleled access to diverse content. Despite their immense popularity, these platforms frequently fall short of accommodating blind and visually impaired users, who depend on audio descriptions to interpret visual material. Audio descriptions are verbal narrations that convey crucial visual elements such as actions, settings, and non-verbal cues, thereby enabling visually impaired users to fully immerse themselves in the viewing experience [1].

Currently, prominent platforms like YouTube do not offer native support for audio descriptions, excluding a substantial segment of their audience from fully experiencing the content [2]. Consequently, blind and visually impaired users often find themselves marginalized in accessing mainstream media, educational content, and other videos available on these platforms.

Netflix, a leading subscription-based streaming service, has made some advancements by providing audio descriptions for select content [3]. However, these descriptions are usually limited to one language, restricting accessibility for non-English speakers. Although Netflix's initiatives mark a step forward, the overall availability and comprehensiveness of audio descriptions remain inadequate to effectively serve the global blind and visually impaired community.

To address these shortcomings, independent initiatives like YouDescribe have been developed. YouDescribe is a platform that allows volunteers to generate and share audio descriptions for YouTube videos, creating a valuable resource for visually impaired users [4]. Nevertheless, the platform's reliance on volunteers results in variable coverage and quality, and its dependency on external applications and tools complicates the user experience, as it is not directly integrated into the mainstream streaming services that users typically use.

1.2 Problem Statement and Motivation

The creation of audio descriptions for video content is an intricate, labor-intensive process that typically necessitates manual scripting, narration, and synchronization with the visual aspects of the video. This manual process is time-consuming and incurs significant costs, with prices ranging from $15 to $75 per minute of video [5]. Such high costs create a considerable barrier to the widespread adoption of audio descriptions, particularly for smaller content creators and platforms with extensive video libraries.

Due to these challenges, the availability of audio descriptions across most video streaming platforms remains limited. Even on platforms like Netflix, where efforts have been made to include audio descriptions, coverage is usually incomplete; many videos remain undescribed. Furthermore, these descriptions are generally offered in only one language, typically English, which limits accessibility for non-English-speaking users.

The current reliance on a manual approach is not scalable, resulting in inefficient production and distribution of audio descriptions. These inefficiencies contribute to the limited availability of audio-described content, ultimately excluding blind and visually impaired users from fully accessing and enjoying video content. Additionally, the predominance of a single language in the available descriptions, as seen on platforms like Netflix, further exacerbates the problem by not catering to the needs of a diverse, global audience.

These limitations indicate the need for an automated and scalable solution to deliver high-quality audio descriptions across multiple languages, making video content accessible to all users, regardless of their visual or linguistic abilities.

The motivation for this project arises from the escalating need for accessible video content driven by a rapidly increasing demand from blind and visually impaired users. This substantial audience continues to expand, indicative of a broader societal movement toward inclusivity and accessibility within digital media. As video content becomes increasingly prevalent across the internet, it is vital to ensure it remains accessible to all users, including those with visual impairments, making this a critical issue to address.

In Malaysia, approximately 1.2% of the population is blind or visually impaired, highlighting a significant group that requires accessible content to engage in the digital era fully [6]. This underserved demographic not only represents a clear need but also offers a compelling market opportunity for platforms and services that can effectively cater to their requirements. The significance of the audio description market extends beyond Malaysia to a global scale, as more content creators and platforms realize the importance of catering to a diverse audience.

Additionally, rising awareness and evolving legal mandates concerning accessibility in digital media are accelerating the demand for scalable and cost-effective solutions within the audio description market. Content creators, streaming platforms, and broadcasters actively seek ways to deliver high-quality audio descriptions without the associated high costs. This project aims to address these demands by developing an automated system for generating audio descriptions, providing a solution that is both cost-effective and inclusive. The combination of the opportunity to serve a rapidly growing audience and the societal impact of enhancing accessibility makes this project both timely and compelling.

1.3 Project Objectives

The primary objective of this project is to develop a sophisticated, automated system capable of generating accurate and personalized audio descriptions for video content, thereby improving accessibility for users with visual impairments and catering to diverse audience needs.

To achieve this primary objective, the project is divided into the following sub-objectives:

1. Develop an Automated Audio Description Generation Pipeline: Create a complete system that automates the generation of audio descriptions from video content. This includes integrating modules for I-frame extraction, voice activity detection, face clustering, and detailed image captioning.
2. To develop contextually accurate and audience-tailored audio descriptions through a Multi-Agent Personalization Framework, which incorporates audience persona modeling to generate descriptions that are both factually precise and stylistically aligned with the preferences and cognitive needs of the target user group.
3. To design a systematic evaluation system to methodically measures accessibility impact to the visual impaired communities.

1.4 Project Scopes and Direction

The scope of this project encompasses the development, implementation, and evaluation of an automated and personalized audio description system aimed at enhancing the accessibility of video content for users, particularly those with visual impairments. The system is designed to be integrated into existing online video platforms, offering content creators a sophisticated tool to generate detailed, context-aware audio descriptions for their videos.

The project involves the creation of multiple interrelated modules, including video frame extraction, face detection, speech-to-text conversion, face clustering, and audio description generation, all of which form a comprehensive pipeline for producing audio descriptions. Additionally, the project will implement a multi-agent system powered by advanced large language models like Gemini, designed to tailor the generated audio descriptions to distinct user personas.

This project is scoped to not only cover the technical development and integration of these tools but also to test and refine the system to ensure its effectiveness across diverse video content types. The project will culminate in a system that can be broadly applied across different content platforms, significantly improving accessibility for a wide range of users.

1.5 Contributions

This project offers several innovative contributions to the field of automated audio description (AD) generation for video content:

1. **Enhanced Video Recognition through Clustering and Labelling:** We introduce a novel method involving the collaboration of content creators to label clustered facial data, thereby improving video recognition capabilities. This approach broadens the system's applicability, enabling it to process a more diverse video types, including lesser-known and unique content, beyond the scope of pre-existing datasets typically sourced from IMDb. Our system is thus versatile and adaptable to various video genres beyond mainstream media.
2. **Multilingual Support for Global Accessibility:** Our solution addresses a significant limitation in existing AD systems by offering support for multiple languages. This feature ensures that audio descriptions can reach a broader global audience, enhancing accessibility for non-English-speaking viewers and expanding the inclusion of diverse audiences.
3. **Structured Script-Based Description Generation:** Unlike traditional methods that generate audio descriptions directly from video frames, our system first constructs a comprehensive movie script. This script includes detailed image captions for each frame and dialogues with character names, serving as a robust foundation for generating more accurate and contextually appropriate audio descriptions.
4. **Single Prompt, Multi-Turn, Multi-Agent Reasoning System for Tailored Audio Descriptions:** We transform traditional multi-agent systems by enabling a collaborative, dynamic, and efficient workflow within a single prompt stream. By integrating multiple specialized agents in a continuous interaction, the system fosters real-time refinement, ensuring the final output is of high quality and coherence. Whether used for dialogue refinement, content generation, or problem-solving, SMARS offers a more fluid and efficient approach to multi-agent reasoning compared to traditional methods, streamlining processes while maintaining flexibility and precision.

1.6 Report Organization

The report is organized into several chapters to ensure a coherent and systematic presentation of the project. Chapter 1 introduces the background, objectives, scope, and significance of the project. Chapter 2 provides a comprehensive review of related technologies, tools, and existing systems relevant to the development process. Chapter 3 outlines the system design and architecture, detailing how different components interact and function within the overall framework. Chapter 4 focuses on the implementation of key modules, describing the methods and techniques applied to build the system. Chapter 5 presents system testing and evaluation, including both quantitative metrics and qualitative analysis to assess performance and user satisfaction. Chapter 6 discusses the results obtained, along with insightful interpretations and reflections on system effectiveness. Finally, Chapter 7 concludes the report by summarizing key achievements, challenges encountered, and suggesting directions for future work. Supplemental sections such as references and appendices provide additional support and documentation relevant to the project.

CHAPTER 2

Literature Review

2.1 Review of the Technologes

This section outlines the various hardware and software technologies utilized throughout the development of the "Enjoyable" platform — a real-time, AI-powered audio description system designed to enhance video accessibility for visually impaired users. The review includes the hardware platform, operating environment, programming stack, algorithms, and data management choices made to ensure optimal system performance and integration.

2.1.1 Hardware Platform

The development, training, and testing processes were all executed on a high-performance consumer-grade laptop — **HP Victus by HP Laptop 16-e0122AX** . The specifications are as follows:

- **Processor** : AMD Ryzen 5 5600H with Radeon Graphics
- **Graphics** : NVIDIA GeForce RTX 3050 Laptop GPU, AMD Radeon(TM) Graphics
- **Memory** : 16GB DDR4 RAM
- **Storage** : 512GB NVMe SSD + 1TB TEAM SSD
- **Operating System** : Windows 11

The RTX 3050 GPU provided CUDA acceleration for deep learning tasks such as face detection, frame analysis, and model inference. This setup was sufficient to perform real-time processing and batch computations, especially during video segmentation and inference-based captioning tasks.

2.1.2 Firmware/OS

The system was built and operated on **Windows 11** . This modern operating system offers a stable environment for Python development, supports GPU drivers needed for machine learning libraries (e.g., CUDA, cuDNN), and ensures compatibility with key development tools. Driver support for NVIDIA GPU and general USB/HDMI video/audio I/O was fully integrated, supporting testing and potential extension to assistive devices.

2.1.3 Database

Due to the non-relational and lightweight nature of the data handled — such as intermediate JSON files, frame metadata, and transcription data — the system did not use a full-scale relational DBMS like MySQL or PostgreSQL. Instead, formats such as:

- **JSON** : For storing semantic metadata (e.g., speaker timestamps, visual descriptions).
- **Local Storage** : For I-frame images and pre-processed video segments.

This choice was guided by the system's focus on streaming and file-based batch processing rather than structured transactional data operations. For future expansion involving user data, a NoSQL solution such as MongoDB may be considered.

2.1.4 Programming Language

The primary programming language used in the system development is **Python 3.10** . Python was chosen for its simplicity, vast ecosystem, and top-tier support for AI, video processing, and computer vision libraries. The following major Python libraries and tools were used:

- **OpenCV** : For video frame extraction and processing
- **PyTorch** : For handling LLM model integration and image captioning
- **Whisper by OpenAI** : For high-accuracy Speech-to-Text (STT)
- **dlib and mediapipe** : For facial detection and tracking
- **Transformers (HuggingFace)** : For prompt handling and calling general-purpose language models (e.g., GPT-4 vision APIs)

Python also enabled easy integration between different modules and experimentations with various machine learning workflows.

2.1.5 Algorithm

The system leverages a composition of image, audio, and language AI algorithms to form an intelligent multi-modal processing pipeline. Key algorithms include:

- **I-Frame Extraction** : Using video stream indexing to isolate keyframes for visual captioning.
- **Face Detection & Recognition** : Via CNN-based algorithms (e.g., dlib / deep metric learning) to track speaker continuity and eliminate redundancy.
- **Voice Activity Detection (VAD)** : To detect silent intervals for inserting descriptions.
- **Speech-to-Text (STT)** : OpenAI's Whisper model, used for high-accuracy transcription of the spoken content.
- **Active Speaker Detection** : Lip movement alignment with speech segments to associate dialogue with visually identified faces.
- **Image Captioning with GPT-4 Vision** : Multi-modal LLM processes key visual scenes and generates descriptions.
- **SMARS (Single Prompt Multi-Turn Multi-Agent Reasoning System)** : An internal multi-agent collaborative reasoning framework simulated within a single prompt using Gemini. Agents perform roles like investigation, rewriting, flow checking, and caption length control per temporal budget.

These algorithms together enable the system to create automated, high-quality, and timed audio descriptions for any video input.

2.1.6 Summary of the Technologies Review

In summary, the “Enjoyable” platform utilizes a strong combination of modern hardware, flexible software tooling, and advanced AI algorithms. The selection of Python and state-of-the-art pre-trained models offers the flexibility needed to rapidly iterate and integrate both vision and language tasks. The combination of lightweight data storage formats and GPU-accelerated model execution ensures that the system can perform audio captioning effectively on standard consumer hardware, making it practical for deployment in both institutional and personal use cases. Most critically, the SMARS reasoning framework gives the system an intelligent, human-like captioning capability without requiring a distributed backend or complex orchestration, simplifying the implementation while enhancing the quality of output.

2.2 Previous Works on Audio Description

Audio description is a narrative service specifically created to enhance the accessibility of visual media, including movies, TV shows, live events, and other visual content, for persons who are blind or visually impaired. Narrative commentary is the process of delivering verbal explanations that describe significant visual components, including movements, facial expressions, scenery, and other crucial aspects, during natural pauses in the dialogue. The inclusion of this additional level of description facilitates users in acquiring a deeper understanding of the provided material.

2.2.1 Toward Automatic Audio Description Generation for Accessible Videos

The Tiresias system represents a notable advancement in the field of automated audio description (AD) generation, offering a comprehensive and innovative approach to enhancing video content accessibility for visually impaired users. The system is meticulously designed to tackle the various challenges associated with producing high-quality audio descriptions. It is structured around three core modules: Insertion Time Prediction, Audio Description Generation, and Audio Description Optimization [7]. Each of these modules plays an integral role in ensuring that the final audio descriptions are not only precise and contextually relevant but are also seamlessly integrated into the existing audio track of a video.

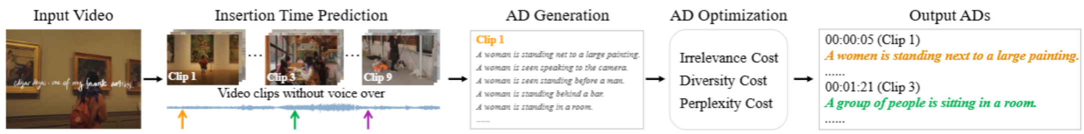


Figure 2.1 Tiresias consists of three core modules: Insertion Time Prediction, Audio Description Generation, and Audio Description Optimization [7].

This Time Prediction module's primary function is to analyze the video content and precisely determine the optimal moments for inserting audio descriptions. The key objective is to ensure that these insertions do not interfere with the original audio, such as dialogues, significant sound effects, or important background music. Through this intelligent insertion, the system ensures that the accessibility of the video is enhanced without detracting from the overall viewing experience.

To achieve this, the Tiresias system employs sophisticated tools like Katna for keyframe extraction. These tools are adept at detecting and capturing significant visual changes within the video, often signaling important events or transitions that would benefit from an accompanying audio description. By identifying these keyframes, the system effectively segments the video into manageable clips that can be individually analyzed for the optimal placement of audio descriptions.

The system further strengthens its analysis by integrating a deep learning network, which includes two sub-networks dedicated to visual and audio features, respectively. The deep learning model aims to identify any inconsistencies or alignments between these two modalities, ensuring that the system maintains a high level of accuracy in preserving audio-visual consistency. This robust approach allows the system to identify when a visual event is sufficiently significant to warrant a description while also ensuring that the description does not overlap with any crucial audio elements. The result is a refined segmentation process that sets the foundation for generating accurate and contextually aware audio descriptions.

Following the identification of the appropriate insertion points, the next critical task is the creation of the audio descriptions themselves. This responsibility falls on the Audio Description Generation module. The core objective of this module is to produce textual descriptions that accurately and succinctly convey the visual content of the video segments, focusing on essential details such as character actions, significant objects, and scene transitions.

The Tiresias system utilizes an advanced attention-based model to accomplish this task. Central to this model are two critical components: the Sentence Localizer and the Caption Generator. The Sentence Localizer is responsible for pinpointing the most crucial visual events within the segmented video clips, determining what aspects of a scene require description. Rather than approaching this task as simple image captioning, the model employs a cross-attention framework that integrates both video and textual features. The Caption Generator then creates the textual description that aligns with the identified visual events, ensuring that the generated text is both accurate and contextually appropriate.

Training this model necessitates a large dataset rich with detailed event descriptions that enable the system to learn how to align visual cues with textual descriptions properly. The Tiresias system leverages the ActivityNet Captions dataset, a resource specifically curated to provide an expansive set of event descriptions across a broad range of videos. This extensive dataset is crucial to the system’s ability to generalize its audio description generation capabilities across different types of video content, from films and documentaries to educational videos and beyond. The combination of an attention-based model with comprehensive training data empowers the system to produce high-quality audio descriptions that are both informative and engaging.

The final module in the Tiresias system is the Audio Description Optimization module, which is essential in refining the generated descriptions to meet the highest standards of relevance, diversity, and grammatical accuracy. This step is indispensable, as even the most accurately generated descriptions may require further refinement to ensure they are as effective and user-friendly as possible.

The optimization process is guided by applying various cost functions, each designed to address a specific aspect of the generated descriptions. For instance, the irrelevance cost function penalizes descriptions that contain unnecessary or redundant information, directing the system toward providing only the most pertinent details. The diversity cost function encourages various language use, preventing repetitive or monotonous descriptions that could reduce listener engagement. Lastly, the perplexity cost function ensures that the descriptions are fluent and grammatically correct, enhancing their readability and comprehensibility.

These cost functions are synthesized into a total cost function, which the system aims to minimize using dynamic programming techniques. Dynamic programming permits the system to efficiently explore several combinations of description options, ultimately selecting the set that best balances relevance, diversity, and fluency. Through this approach, the Tiresias system guarantees that the final audio descriptions are accurate and contextually appropriate and polished and engaging for the listener.

2.2.2 AutoAD: Movie Description in Context

The paper titled "AutoAD: Movie Description in Context" marks a transformative milestone in the field of automatic audio description (AD) generation, fundamentally improving the way descriptions are crafted for movies. Recognizing that the intricate narrative arcs and complex character interactions within films require more than just basic frame-by-frame analysis, AutoAD implemented a sophisticated strategy of contextual integration that combines multiple layers of information—such as temporal sequences, dialogue, and subtitles—to generate audio descriptions that are not only seamless and coherent but also well-aligned with the overall narrative framework of the film.

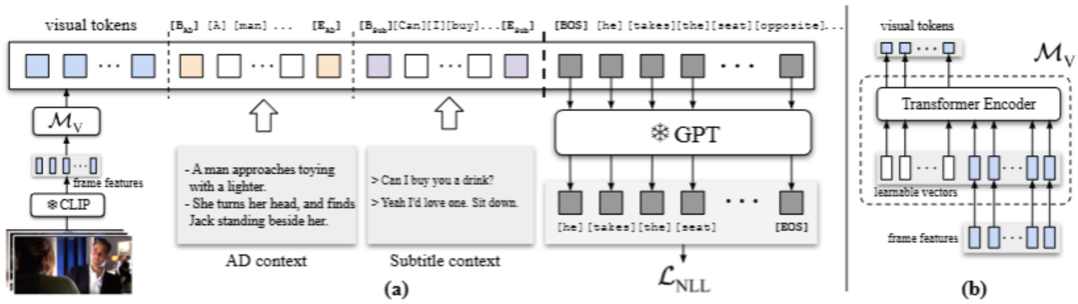


Figure 2.2 (a) Overview of AutoAD (b) Detail of the visual mapping network [8].

The contextual Integration technique is designed to align the generated audio descriptions with the unfolding narrative of the movie. The model effectively leverages numerous types of context to ensure that ADs not only reflect the visual content but also the underlying story that evolves throughout the film.

AutoAD recognizes the importance of remembering and referencing past events, characters, and descriptions generated in earlier scenes. Films often involve intricate narratives where elements introduced early on reappear or gain significance later in the story. AutoAD captures this complexity by maintaining a temporal memory of previously generated descriptions, which it can then reference when producing new ones. For instance, if a character introduced as a minor player early on becomes central to the plot later, AutoAD will recall this character's past and ensure that new descriptions acknowledge their growing importance. This continuity enriches the viewing experience for visually impaired audiences, allowing them to follow the story more holistically.

Understanding dialogue is crucial in filmmaking, as it often conveys key narrative elements, reveals character motivations, and builds emotional tension. AutoAD incorporates dialogue context into its AD generation pipeline by extracting and analyzing dialogue from subtitles or transcriptions. This enables the model to generate character-aware descriptions relevant to ongoing spoken interactions. For example, if a character makes an ominous remark, AutoAD might generate a description reflecting the following dark mood, perhaps noting the sudden dimming of a room or a close-up of a character’s face. By integrating dialogue into the descriptive process, AutoAD enhances the story’s coherence and immersion.

Subtitles provide critical cues about dialogue and other non-verbal audio elements—such as door slams, footsteps, or shifts in background music—essential for advancing the plot. AutoAD leverages these cues to synchronize ADs with the film’s narrative flow, ensuring that descriptions enhance rather than disrupt critical audio moments. For instance, the model can predict a significant event, like the buildup of tense music before a dramatic scene, and produce a description that aligns with this anticipation. This careful timing allows the viewer to experience the film’s pacing and mood as if they were visually following it.

One of the perennial challenges in AD generation is the scarcity of extensive, annotated datasets that are specifically tailored for this task. To circumvent this, AutoAD adopts an innovative Data Augmentation strategy that involves pretraining the model on large-scale datasets, which, while not directly related to movie AD, provide essential context and narrative data.

AutoAD leverages text-only datasets, which consist of rich narrative descriptions that lack corresponding visual frames. These datasets, including items like screenplay plots, summaries, or closed captions, are invaluable for training the text generation component of AutoAD—typically, a GPT-based model. Although these datasets don't contain visual data, they offer a broad understanding of narrative structures and linguistic conventions, enabling the model to produce ADs that seamlessly integrate with the movie’s dialogue and narrative flow.

On the visual side, the model is pretrained using visual captioning datasets that match images or short video clips with descriptive captions. While these datasets may not possess the complex temporal dynamics found in films, they are instrumental in teaching the model how to associate visual cues with descriptive language. This visual-textual alignment is later finely tuned when the model is trained on more context-rich movie data, ensuring that AutoAD generates descriptions that are both visually and contextually appropriate.

In addition to data and context integration, AutoAD introduces another significant innovation through Model Enhancement via lightweight adapters [8]. These adapters serve as intermediary modules that connect the outputs of pretrained foundation models—such as GPT for text and CLIP for vision—with the specific contextual needs of AD generation. These adapters refine the generated descriptions by embedding relevant context, such as character information or narrative progression, directly into the text. For instance, if a scene involves a character whose importance has recently increased, the adapter can ensure that this shift is reflected accurately in the AD, preserving narrative consistency and enhancing viewer comprehension.

These adapters also play a critical role in merging various forms of context—temporal, dialogue, and subtitle—into the AD process. This ensures that generated descriptions are well-informed by the ongoing narrative and character development, rather than being isolated observations. For example, if a character's emotional state has shifted significantly, the adapter can adjust the AD to reflect this, thus producing a more coherent and impactful narrative for the viewer.

"AutoAD: Movie Description in Context" represents a significant leap in the evolution of automatic audio description technology. Through its focus on Contextual Integration, Data Augmentation, and Model Enhancement, AutoAD addresses the limitations of earlier models and establishes a new benchmark for AD quality in cinematic contexts.

2.2.3 AutoAD II: The Sequel – Who, When, and What in Movie Audio Description

AutoAD II is thoughtfully engineered to tackle three essential challenges in movie AD: accurately identifying characters (Who), determining the optimal moments for description insertion without interrupting dialogue (When), and generating contextually rich and meaningful descriptions (What) [9].

AutoAD II is built upon the understanding that successful audio description for films demands a profound grasp of the visual content, auditory cues, and the underlying narrative structure. Creating automatic AD for movies presents a unique challenge as it requires intricate management of narrative complexities. These include accurately identifying characters, deciding the precise timing for inserting descriptions, and ensuring that those descriptions capture the essence of visual content without disrupting dialogue or other critical audio elements.

One of the most crucial components in movie audio description is the precise identification and naming of characters. For visually impaired viewers, knowing the names of characters or understanding their roles in various scenes is vital for following the narrative and tracking who is performing what actions.

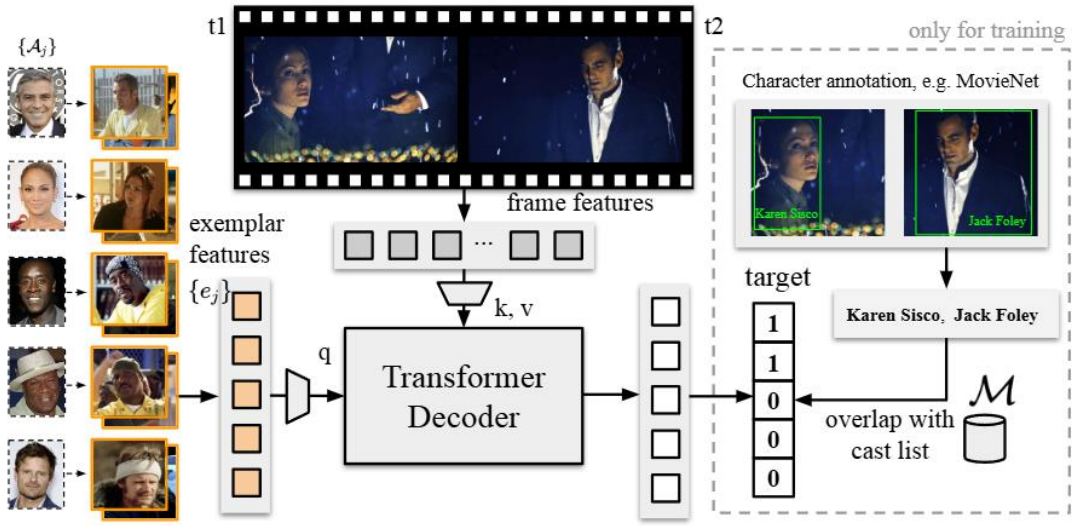


Figure 2.3 Character Recognition Module [9].

AutoAD II addresses this by introducing a Character Bank, specifically designed to link character names with their respective visual and auditory representations within the film. This bank is meticulously constructed for the principal cast in each movie and includes:

- **Character Names:** The actual names of characters that are essential for accuracy in audio descriptions.
- **Character Faces:** Visual features derived from CLIP (Contrastive Language-Image Pretraining), which can differentiate and recognize characters based on their appearances across scenes.
- **Actor Information:** Distinctive facial features and voice identifiers associated with the actor portraying the character.

Through this Character Bank, AutoAD II ensures accurate referencing of characters throughout the film. As scenes evolve, the model uses the Character Bank to correctly label individuals, thereby preserving narrative continuity and improving the viewer's comprehension of the story.

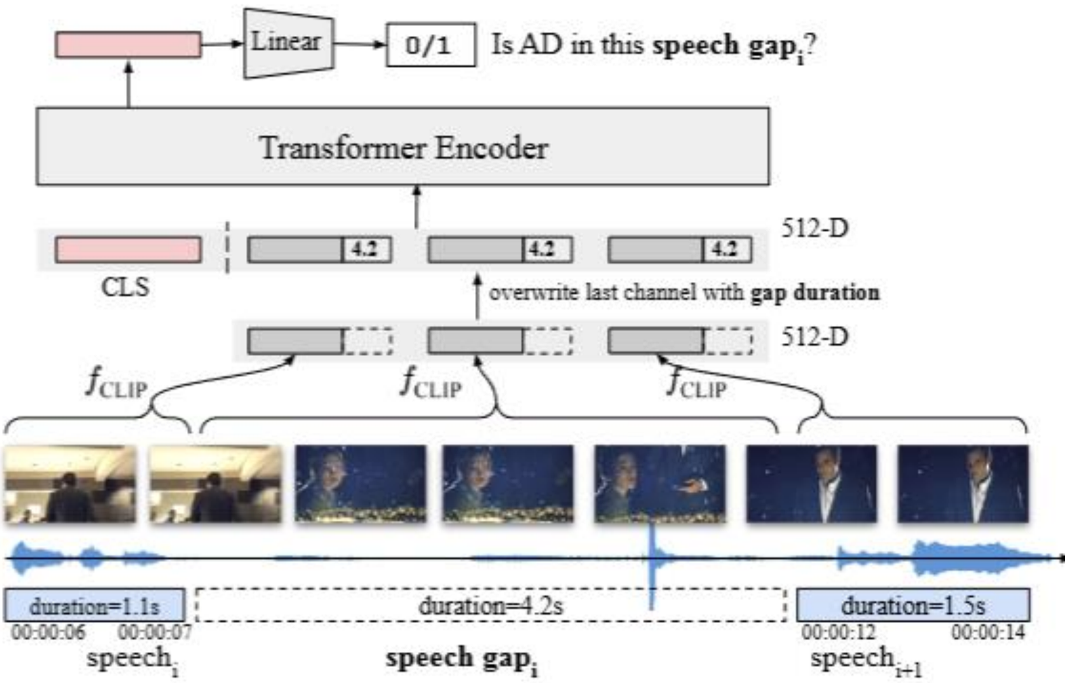


Figure 2.4 Architecture for AD temporal proposal classification [9].

The timing of audio descriptions is as critical as the content itself. Insertions should occur during moments of low or no dialogue to prevent confusion or dilution of the impact of both the description and the dialogue. However, the rapid pace and dense dialogue sequences in movies often make it challenging to find appropriate gaps for insertion. AutoAD II employs an advanced temporal model that assesses the auditory landscape of each scene. Through Automatic Speech Recognition (ASR), the system identifies exact intervals in scenes where no dialogue is present or where the dialogue is minimal and non-essential. These intervals are classified as opportunities for optimal AD insertion.

The system does not stop at identifying available time slots; it also evaluates visual importance during these intervals. If a significant visual event occurs during a dialogue-free moment, that time slot is prioritized for AD insertion. This process is managed by a transformer-based model capable of considering long-range dependencies and contextual cues from the movie.

A binary classification model, informed by visual features from CLIP and text derived from ASR, is applied to each temporal interval [9]. This model predicts whether an AD should be generated based on the visual content and available time slots. Furthermore, a sliding window technique is used to progressively evaluate the entire timeline of the movie. This ensures that descriptions are inserted when most appropriate, thereby maintaining the rhythm and coherence of the narrative.

The final crucial pillar of AutoAD II is deciding what exactly should be described. Knowing when and who is not enough; the system must also determine what details are important enough for the description. AutoAD II utilizes a vision-language model augmented with cross-attention mechanisms to generate high-quality descriptions. Cross-attention permits the model to incorporate both visual cues from the movie and contextual insights derived from previous scenes, ensuring that the descriptions are not only accurate but also significant to the overall narrative.

AutoAD II: The Sequel represents a significant leap in the generation of audio descriptions for movies, offering comprehensive solutions to the Who, When, and What of AD. By leveraging state-of-the-art machine learning, natural language processing, and computer vision technologies, AutoAD II not only accurately identifies and names characters but also precisely times and tailors the content of descriptions, ensuring seamless integration into films.

2.2.4 AutoAD III: The Prequel – Back to the Pixels

AutoAD III introduces meticulously curated datasets designed to bridge the existing gap between visual content and corresponding audio descriptions, thereby addressing long-standing challenges such as the lack of high-quality paired video-AD data and the difficulties in temporal alignment [10].

The CMD-AD (Character-Movie-Description) dataset emerges as a pivotal development. By combining audio descriptions from the AudioVault repository with video clips from the Combined Movie Dataset (CMD), AutoAD III does not simply pair descriptions with video clips—it demands precise temporal alignment between the ADs and the exact scenes they describe. Temporal alignment is particularly challenging due to the differences in timing and pacing found across various movies. AutoAD III overcomes this hurdle by employing advanced synchronization algorithms that ensure the ADs are accurately coordinated with the relevant movie events. Unlike earlier datasets that provided only frame-level features, the CMD-AD dataset offers a more holistic perspective by encapsulating entire scenes and integrating character information directly into the dataset.

In addition to CMD-AD, the project introduces the HowTo-AD dataset, which repurposes the extensive HowTo100M video dataset. Although HowTo100M is originally composed of narrated instructional videos rather than designed specifically for AD tasks, it contains a vast amount of spoken and visual information that can be harnessed for AD-like applications. By employing language models, AutoAD III converts these instructional narratives into pseudo-audio descriptions. This innovative repurposing effectively addresses the challenge of data scarcity by providing a rich source of training data that includes both temporal sequences and narrative context. The HowTo-AD dataset is particularly valuable as a pseudo-ground-truth label generator, enabling the model to learn from a broad base of examples even when manually annotated ADs are unavailable. These pseudo-labels, while imperfect, offer foundational insights into describing actions, objects, and sequences within a time-dependent context, and these insights are subsequently refined during the fine-tuning process using more accurately labeled data.

Beyond the advancements made in data, the model architecture in AutoAD III also represents a significant leap forward. The architecture is centered around a Q-former-based design, building upon methods informed by the BLIP-2 (Bootstrapping Language-Image Pretraining) approach. The primary function of the Q-former is to enhance the integration between visual and textual information, ensuring that the resulting audio descriptions are context-sensitive and deeply relevant.

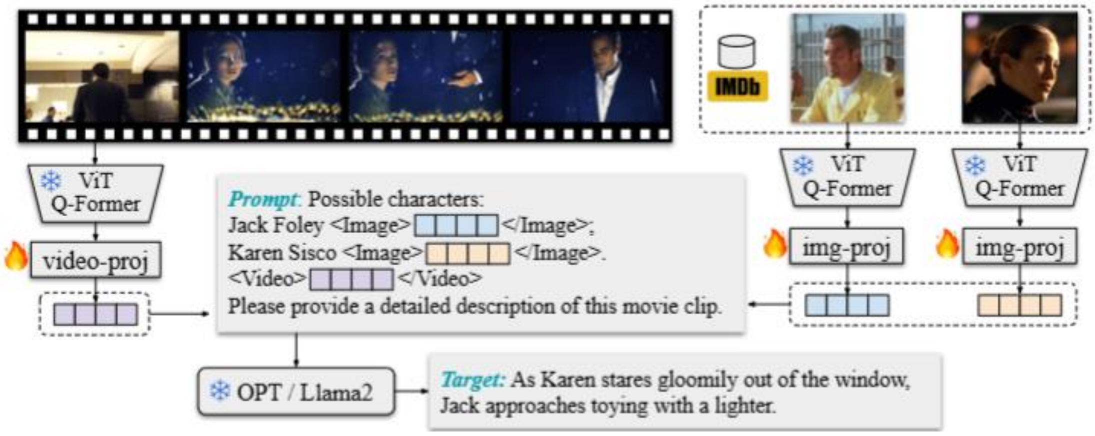


Figure 2.5 Architecture of AutoAD III [10].

One of the key innovations is the incorporation of multi-frame video clips into the AD generation process. Traditional methods often struggle with understanding video content that evolves over time because they typically focus on single-frame analysis. However, AutoAD III introduces a multi-frame approach that allows the model to analyze multiple frames simultaneously. This enables the system to capture the temporal dynamics of scenes—how actions develop, how emotions shift, and how the focus transitions from one character to another. By reflecting the temporal evolution of events, the descriptions generated by AutoAD III are richer and more faithful to the dynamic nature of visual storytelling in movies.

Another innovative aspect is the integration of a Character Bank into the model's architecture. The Character Bank serves as a repository of character-specific information, including names, visual features extracted from CLIP, and textual associations. Each time a character appears in a scene, the Q-former leverages the Character Bank to ensure that the description accurately identifies the character by name. Additionally, it can provide further information such as their role in the scene or their relationship to other characters. This character-awareness adds depth to the audio descriptions, enabling visually impaired viewers to maintain a clearer understanding of the narrative as it unfolds.

The Q-former mechanism itself acts as a vital bridge between the vision and language models. It effectively queries the visual model with textual prompts derived from the language model's understanding of the narrative context and refines the output to align with what is visually present in the scene. This approach addresses the traditional gap between visual recognition and language generation, resulting in more precise and contextually relevant audio descriptions. The Q-former mechanism thus ensures that the generated descriptions not only accurately capture what is happening but also fit seamlessly within the narrative, preserving the film's pacing and emotional tone.

Recognizing the limitations of traditional evaluation methods, AutoAD III introduces a new set of evaluation metrics aimed at better assessing the quality of generated ADs, with particular attention to character identification and semantic appropriateness.

A standout metric introduced in AutoAD III is the Character Reference Informed Textual Information Content (CRITIC) metric. CRITIC focuses on the accuracy of character identification, which is crucial in narratives where understanding "who is doing what" is essential to the storyline. This metric evaluates how well the descriptions integrate character names and relational information, ensuring that the ADs are not only descriptive but also narratively coherent. This is especially important in films where characters and their actions are central to the plot.

In addition, AutoAD III incorporates LLM-based (Large Language Model-based) evaluations that utilize advanced language models to assess the semantic coherence and overall quality of the generated text. Given that these large language models have been trained on vast amounts of textual data, they can act as a proxy for human judgment, evaluating whether the ADs accurately reflect the nuances of the scenes and align with expected language use. The LLM-based evaluation examines factors such as grammar, fluency, and contextual alignment, providing a more nuanced and human-aligned measure of AD quality. The combination of these metrics not only enhances the reliability of the evaluation process but also ensures that the model is fine-tuned to produce high-quality, contextually appropriate descriptions that resonate with visually impaired audiences.

In conclusion, "AutoAD III: The Prequel – Back to the Pixels" stands as a significant milestone in the progression of automatic audio description systems. It offers sophisticated solutions to some of the field's most persistent challenges through its introduction of new datasets like CMD-AD and HowTo-AD, its Q-former-based model architecture, and its implementation of advanced evaluation metrics such as CRITIC and LLM-based assessments.

2.3 Limitations of Existing Works

The development of automatic audio description (AD) systems has seen notable progress, yet several limitations persist, particularly when it comes to contextual generalization, character recognition, and traditional frame-by-frame generation methods. These challenges highlight the complexities involved in creating effective and accurate descriptions that cater to a wide variety of films, genres, and cultural contexts, while providing a seamless viewing experience for visually impaired audiences.

One of the most pressing limitations of current AD systems is their struggle to effectively generalize across diverse genres, directorial styles, and cultural contexts. While these models often perform adequately for mainstream genres—such as action or romance films, where visual cues and narrative structures align with established cinematic conventions—they tend to falter when faced with more niche or unconventional content, such as avant-garde or experimental cinema, as well as animated films. Take, for instance, avant-garde films, which prioritize visual symbolism and non-linear narrative structures. These require a deeper understanding of subtle visual motifs and abstract storytelling techniques, which standard AD models, often trained on more conventional content, may fail to capture. As a result, these systems may produce audio descriptions that do not align with the director's intent or comprehend the thematic elements underlying the visual content.

Cultural context further complicates the task, as the cultural background of a film plays a significant role in its visual language, symbolism, and narrative pacing. For instance, films created outside of Western or English-centric traditions might employ unique visual forms or symbolism that AD models, primarily trained on such Western datasets, fail to recognize or convey accurately. This cultural myopia can lead to descriptions that overlook critical narrative elements or, worse, misrepresent the cultural subtleties embedded within the story.

Moreover, directorial style is another hurdle. Directors with distinctive stylistic signatures, such as Quentin Tarantino's non-linear storytelling, Wes Anderson's symmetrical framing, or Terrence Malick's poetic imagery, present unique challenges for AD systems. These styles rely heavily on specific visual motifs and narrative techniques. Without fine-tuning to recognize and adapt to these unique approaches, AD models risk producing generic descriptions, thereby obscuring the stylistic intentions that are fundamental to the film's narrative. For example, a description that fails to recognize the significance of visual repetition in a narrative might weaken the viewer's understanding of the film's deeper meanings.

Character recognition and consistent identification within an audio description are essential for maintaining narrative clarity, particularly in films with large casts, complex character dynamics, or characters undergoing transformation. Traditional AD systems like AutoAD, AutoAD II, and AutoAD III rely on Character Banks sourced from IMDb—a repository linking character names to their visual and auditory characteristics [8]-[10]. While this represents a step forward, it introduces new challenges related to the creation and maintenance of these datasets. The effectiveness of a Character Bank hinges entirely on the quality, accuracy, and comprehensiveness of the data it contains. Since these systems depend on IMDb, they cannot recognize characters if the videos are not films or TV shows listed in that database. If the dataset is incomplete or inaccurate, or if the content falls outside of IMDb's scope, the system might fail to correctly identify characters, leading to confusion or incorrect descriptions.

Traditional AD generation methods rely heavily on a frame-by-frame analysis, which can introduce several significant limitations that affect the quality and coherence of the produced descriptions. One major drawback is the lack of scene-wide context. By focusing on isolated segments of visual content, frame-by-frame AD generation often ignores the broader context within the scene or the movie as a whole, leading to descriptions that are accurate at a micro-level but fail to capture the larger narrative or thematic elements that give the story its depth and meaning.

For instance, a frame might depict a character smiling, but without understanding the surrounding context—whether the smile is sinister or affectionate—the resulting description could be overly simplistic, misleading, or even contradictory to the storyline. Furthermore, films are constructed through sequences of connected scenes designed to convey a coherent story. Actions or objects in one frame are often directly connected to events in previous or forthcoming frames, meaning that ADs generated in isolation may miss these crucial narrative links, leading to fragmented or disjointed descriptions. This disconnection is particularly detrimental in critical scenes, where the buildup of tension or the emotional significance of a moment spans multiple frames or shots.

In addition, the use of visual motifs, symbolic imagery, and recurring elements that contribute to the thematic depth of a story can also be missed when relying solely on frame-by-frame analysis. Such motifs often only reveal their significance when viewed in sequence with other frames, meaning that a description limited to a single frame may overlook essential thematic elements, resulting in a shallower and less immersive viewing experience for the audience.

Lastly, traditional frame-by-frame methods may fall short in aligning the timing of descriptions with the film's pacing and rhythm. In action sequences, for instance, where rapid-fire ADs are necessary to match the fast-moving visuals, frame-by-frame analysis may lead to delayed or disjointed narration, thereby detracting from the tension and excitement of the scene.

Table 2.1 Summarization of Existing Works.

Paper	Key Modules / Features	Strengths	Limitations
Toward Automatic Audio Description Generation for Accessible Videos [7]	1.**Insertion Time Prediction**: Determines optimal moments for insertion 2.**Audio Description Generation**: Uses an attention-based model (Sentence Localizer & Caption Generator). 3. **Audio Description Optimization**: Uses cost functions to refine AD for relevance, diversity, and fluency.	- Intelligent time prediction avoids overlap with critical audio. - Advanced attention-based model for sentence generation. - Cost function optimizes fluency and relevance	- Primarily trained on conventional datasets like ActivityNet, might struggle with non-conventional visual content. - Generalization across different cultural or directorial styles not addressed extensively.
AutoAD: Movie Description	1. **Contextual Integration**: Leverages temporal sequences, dialogues, and subtitles.	- Strong integration of multi-modal	- Data augmentation with non-movie-specific datasets

in Context [8]	2. **Data Augmentation**: Pretrained on visual captioning datasets and narratives like screenplays. 3. **Model Enhancement via Adapters**: Adapters merge temporal, dialogue, and subtitle contexts.	contexts (visual, dialogue). - Pretraining on large datasets improves generalization. - Dialogue and temporal continuity help narrative coherence	may compromise quality for specific film genres. - Context sensitivity may still struggle with avant-garde or culturally unique films.
AutoAD II: Who, When, What [9]	1. **Character Recognition**: Character Bank (faces, names, actor info) to ensure correct identification of cast. 2. **Temporal Insertion Module**: ASR-based dialogue-free interval analysis and importance-based visual event prioritization. 3. **What to Describe**: Vision-language model using cross-attention for description generation	- Accurate and coherent character identification. - Transformer-based model for optimal description timing. - Cross-attention mechanism for contextual accuracy	- Character Bank depends on IMDb data, limiting generalizability to non-film/video content. - May struggle with scenes without easily classifiable characters or specialized visual styles
AutoAD III: Back to the Pixels [10]	1. **Datasets (CMD-AD, HowTo-AD)**: Temporal alignment between ADs and video content. 2. **Q-former-based Model**: Multi-frame video	- Incremental dataset innovations (CMD-AD, HowTo-AD)	- Existing dataset sources (like HowTo-AD) may not reflect movie-specific complexities.

	clip analysis to capture dynamic narrative evolution. 3. **Character Bank**: Integrated into architecture. 4. **CRITIC Metrics**: LLM-based evaluations and new metrics for character identification quality	improve AD-video alignment. - Q-former allows multi-frame temporal dynamics, enhancing narrative coherence. - CRITIC metric evaluates character accuracy	- Heavier reliance on temporal alignment mechanisms, sensitive to pacing mismatches.

2.4 Solution to the Weaknesses

The proposed solution to address the weaknesses identified in existing automatic audio description (AD) systems centers on developing a comprehensive and contextually aware approach to AD generation. This involves enhancing the accuracy of character identification, improving contextual alignment in description generation, and refining the personalization of audio descriptions.

A critical aspect of the proposed solution is the integration of advanced modules for Active Speaker Detection (ASD), face detection, face clustering, and recognition. These components work collaboratively to overcome the limitations associated with character recognition and continuity in narrative comprehension, particularly in complex scenarios featuring multiple characters. The ASD module functions by synchronizing speech timestamps with the corresponding faces detected in the video, ensuring accurate character identification at specific moments. This is further optimized by leveraging tools such as Dlib for face detection and applying clustering algorithms like the Chinese Whispers Algorithm. These algorithms allow the system to differentiate between characters, even in scenes crowded with participants. By accurately identifying, clustering, and labeling faces, the system facilitates seamless character tracking and annotation throughout the video, addressing inconsistencies and improving descriptive specificity.

To tackle the challenges related to generating contextually relevant descriptions and narrative continuity, the system employs GPT-4o for detailed image captioning. This model translates visual content into rich and meaningful descriptions that consider not just the surface-level visual elements but also the narrative context, character roles, and emotional tones within the scenes. By doing so, GPT-4o goes beyond mere labeling, encapsulating the broader essence of each scene in its captions. These detailed captions, in conjunction with extracted dialogues and contextual annotations, are then compiled into a cohesive and structured movie script by the Movie Script Generation Module. The integration of visual and auditory elements in a single, harmonized narrative ensures that descriptions do not fragment the story but rather enhance comprehension and engagement, thereby addressing the limitations of frame-by-frame AD generation methods.

Recognizing the importance of personalization in enhancing the viewing experience for diverse audiences, the proposed solution incorporates a Multi-Agent System (MAS) driven by large language models (LLMs) like Gemini. This system personalizes content based on detailed audience personas, ensuring that audio descriptions resonate with users' unique cultural, emotional, and demographic backgrounds. Specialized agents within the MAS collaborate to tailor the narrative style, refine linguistic elements, and align content with the specific needs of different user groups. The result is an audio description that not only aligns contextually with the video but also feels personalized and relevant to the intended audience, making the content more accessible and relatable.

CHAPTER 3

System Methodology/Approach

The outcome of this project is an Automated Audio Description System named "Enjoyable". This system includes a series of advanced modules, such as I-Frame Extraction, Frame Filtering, Voice Activity Detection (VAD), Speech-to-Text (STT), Active Speaker Detection (ASD), Face Detection, Face Clustering, Face Labelling, Character Face Recognition and Annotation, Detailed Image Captioning, Movie Script Generation, and an LLM-Based Multi-Agent System.

3.1 System Use Case

The system use case revolves around a video streaming platform integrated with an automated audio description system that enhances video accessibility for visually impaired users. The platform allows users to register, log in, and perform actions such as adding, updating, deleting, and playing video content. Once a video is uploaded, the system initiates its automated processing pipeline. This involves the extraction of I-Frames (keyframes) from the video, followed by frame filtering to eliminate irrelevant or poor-quality frames. Subsequently, the system performs face detection to identify individuals in the video and groups similar faces using face clustering. Detected faces are further labeled, recognized, and annotated for accurate characterization.

To enrich the accessibility features, the system incorporates advanced audio processing techniques. It detects voice activity (VAD) to segregate speech from background noise and converts spoken content into text via speech-to-text (STT) conversion. Active speaker detection identifies which individual is speaking at any moment in the video. The transcribed text is then processed in conjunction with image captions and contextual descriptions to develop a comprehensive narrative through the movie script generation module. The LLM-based Multi-Agent System plays a pivotal role, aligning these tasks to ensure accurate, context-aware audio descriptions. Finally, the generated audio descriptions are converted back into speech and synchronized with the original video during the audio insertion phase, ensuring seamless playback for users.

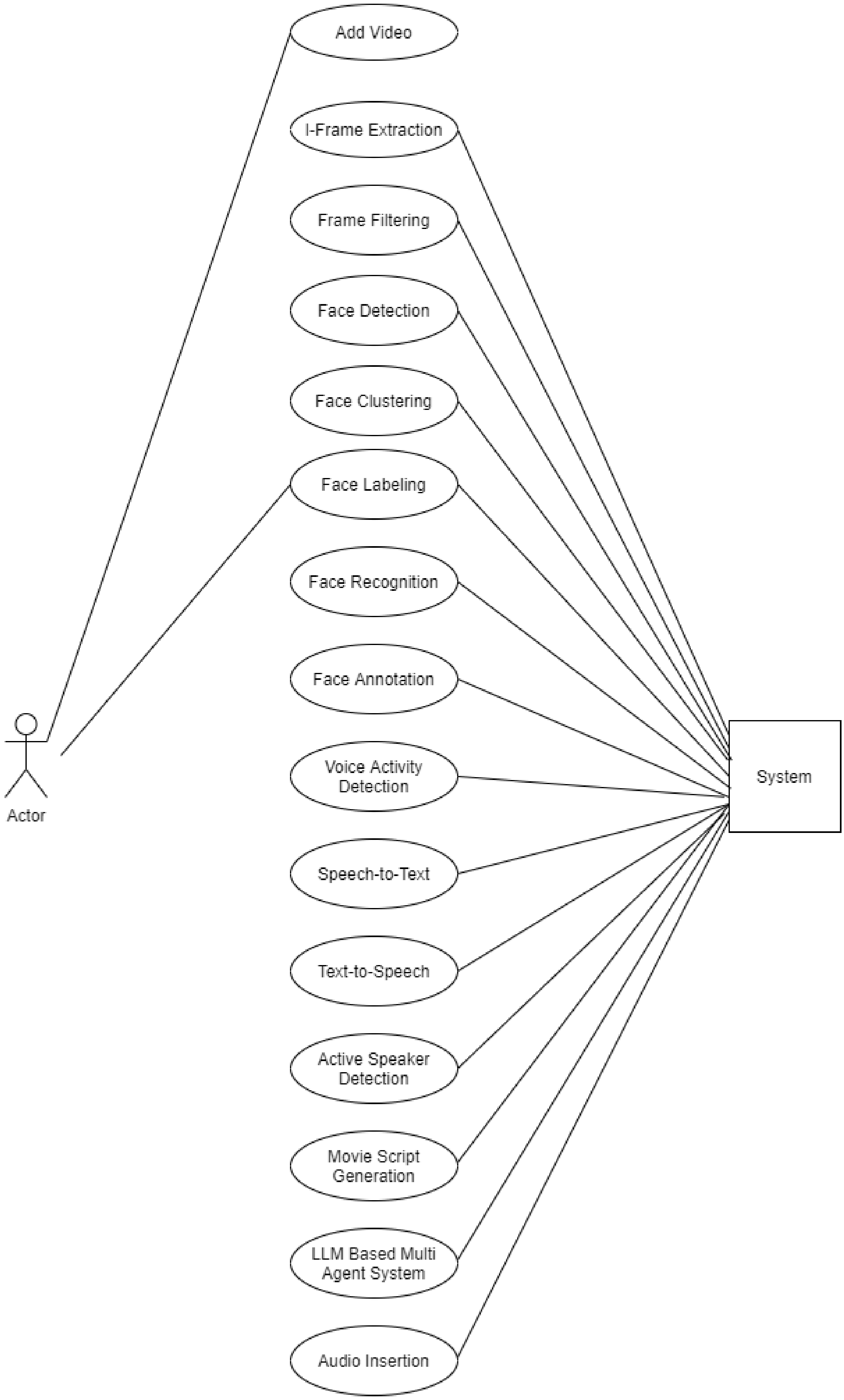


Figure 3.1 System Use Case Diagram.

3.2 System Architecture of Automated Audio Description System

The Automated Audio Description System uses Python to automate the creation of detailed audio descriptions for video content. The system is composed of interconnected modules that collaboratively process and analyze video frames and audio, ultimately generating comprehensive and contextually accurate narratives.

Python's libraries and frameworks, such as dlib, OpenCV, and FFmpeg, facilitate key tasks like frame extraction, face detection, face extraction, face recognition, and the integration of descriptive audio. These tools are essential for efficiently handling the complex processes required to produce high-quality audio descriptions.

By utilizing these advanced technologies, the system ensures that the audio descriptions are accurate and accessible, effectively enhancing the viewing experience for individuals with visual impairments.

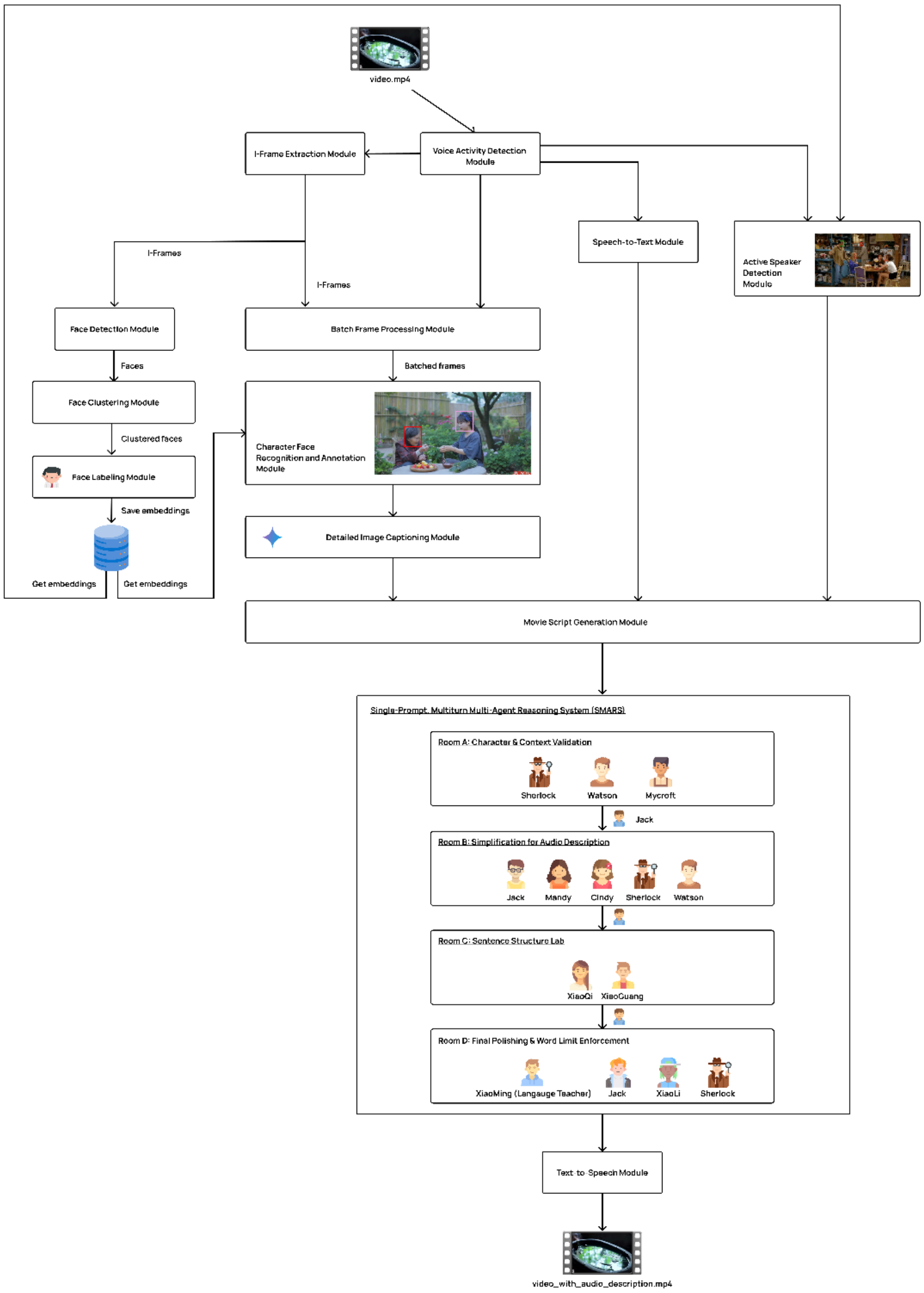


Figure 3.2 Architecture of Automated Audio Description System.

3.3 Activity Diagram

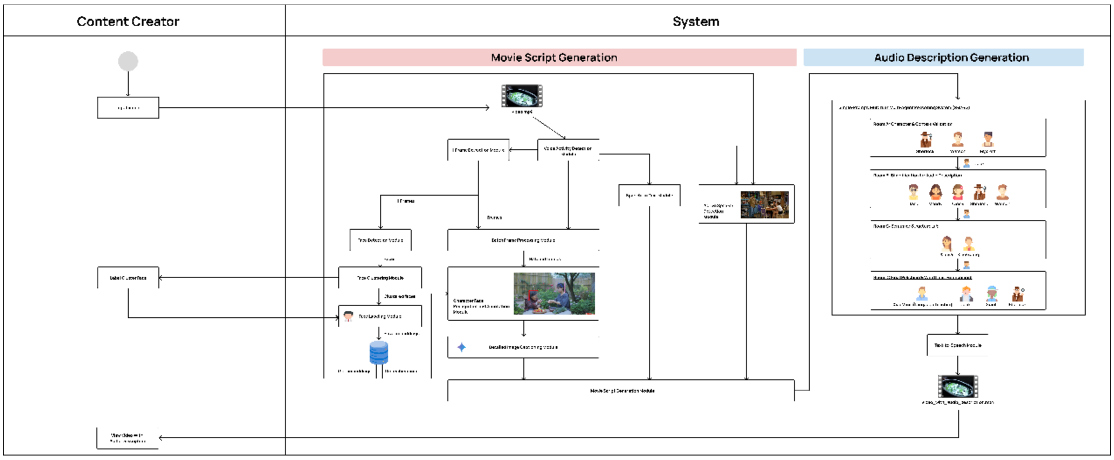


Figure 3.3 Activity Diagram of Enjoyable

CHAPTER 4

System Design

4.1 I-Frame Extraction Module

Videos are composed of different types of frames—Intra-frames (I-frames), Predictive frames (P-frames), and Bidirectional frames (B-frames) [14]. Each frame type serves a specific function in video compression standards such as H.264. I-frames are the keyframes that contain the complete image data, while P-frames and B-frames store only the changes from one frame to another.

This module focuses on I-frame extraction, strategically targeting I-frames to optimize computational efficiency and reduce costs while preserving the quality of visual data. By isolating P-frames and B-frames, the computational load is significantly reduced, allowing the analysis to concentrate on frames that contain complete information, thereby improving the effectiveness and efficiency of subsequent analytical processes.

4.1.1 Workflow of I-Frame Extraction

The I-Frame Extraction Module utilizes advanced video processing libraries, FFmpeg, to efficiently identify and extract key frames from video streams. Within this module, FFmpeg is employed for two primary functions.

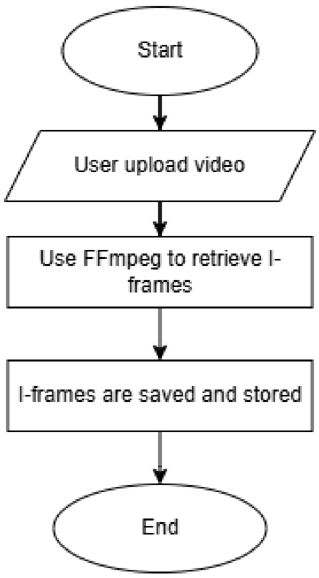


Figure 4.1 I-Frame Extraction Flowchart.

First, the system conducts frame detection by scanning the uploaded video to identify the occurrence of I-frames. This is accomplished through video parsing techniques that accurately detect coding sequences characteristic of I-frames within the video stream.

Following detection, the extraction process is initiated, where the identified I-frames are isolated and stored as individual image files.

4.2 Voice Activity Detection (VAD) Module

The Voice Activity Detection (VAD) Module is designed to identify specific segments within a video where speech occurs. These detected speech segments serve as the foundation for several crucial downstream processes, including speaker recognition, dialogue mapping, and the ultimate generation of synchronized audio descriptions that accurately complement the original dialogue within the video.

4.2.1 Workflow of Voice Activity Detection

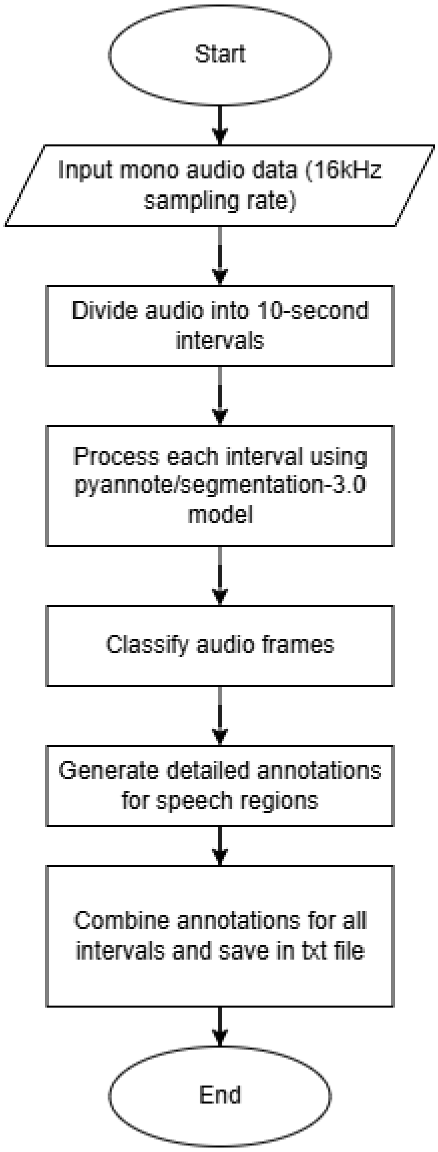


Figure 4.2 Voice Activity Detection Flowchart.

The Voice Activity Detection process is powered by the pyannote/segmentation-3.0 model from Hugging Face, a model highly optimized for audio segmentation tasks. This model has been trained on a diverse collection of datasets, including AISHELL, AliMeeting, AMI, AVA-AVD, DIHARD, Ego4D, MSDWild, REPERE, and VoxConverse, curated by Séverin Baroudi. The model excels not only in identifying speech segments but is also finely tuned for related tasks such as speaker diarization, speaker change detection, and overlapped speech detection [17].

In its operation, the pyannote/segmentation-3.0 model processes audio data in 10-second intervals using mono audio sampled at a rate of 16kHz. The model classifies audio frames into categories such as non-speech and various speaker combinations. Through multi-class encoding, the model precisely identifies the presence of speech and generates detailed annotations to delineate speech regions meticulously.

4.3 Speech-to-Text (STT) Module

The Speech-to-Text (STT) Module is responsible for transcribing spoken dialogue in videos into written text. This step is crucial for our platform as it generates a textual representation of spoken content, which forms the basis for creating the movie script. Accurate transcription ensures that the generated script faithfully mirrors the original dialogue, making it indispensable for subsequent processes such as generating audio descriptions.

4.3.1 Workflow of Speech-to-Text

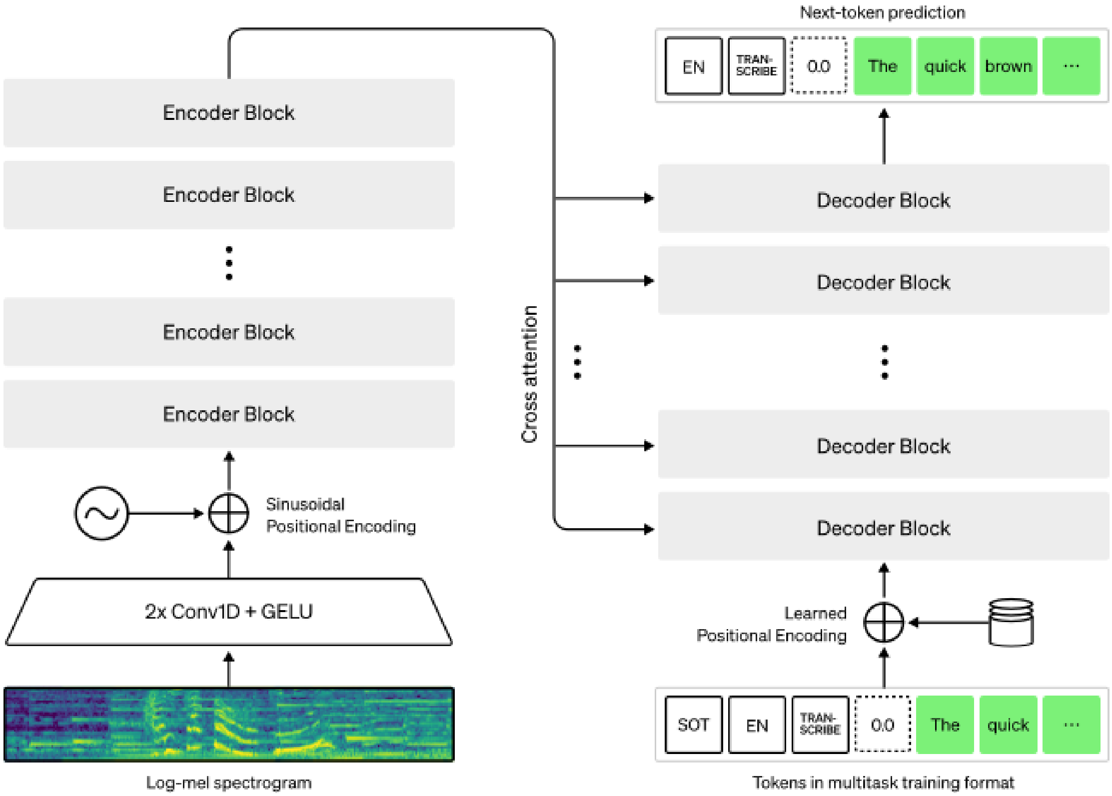


Figure 4.3 Architecture of Whisper [18].

The Whisper model, specifically the whisper-large-v3 variant, is chosen for this task due to its superior performance in recognizing speech across multiple languages and under various acoustic conditions. Whisper is a transformer-based model employing an encoder-decoder architecture, or sequence-to-sequence model, which excels in translating spoken language into text. It is used to transcribes the segmented audio with state-of-the-art accuracy, closely capturing the nuances of the spoken word.

One of Whisper's standout features is its ability to handle multiple languages and accents, making it an excellent fit for global video content platforms that feature diverse user-generated content. Additionally, Whisper is highly resilient to background noise, music, and other common audio disruptions, maintaining clarity and accuracy in the transcriptions even in less-than-ideal acoustic environments.

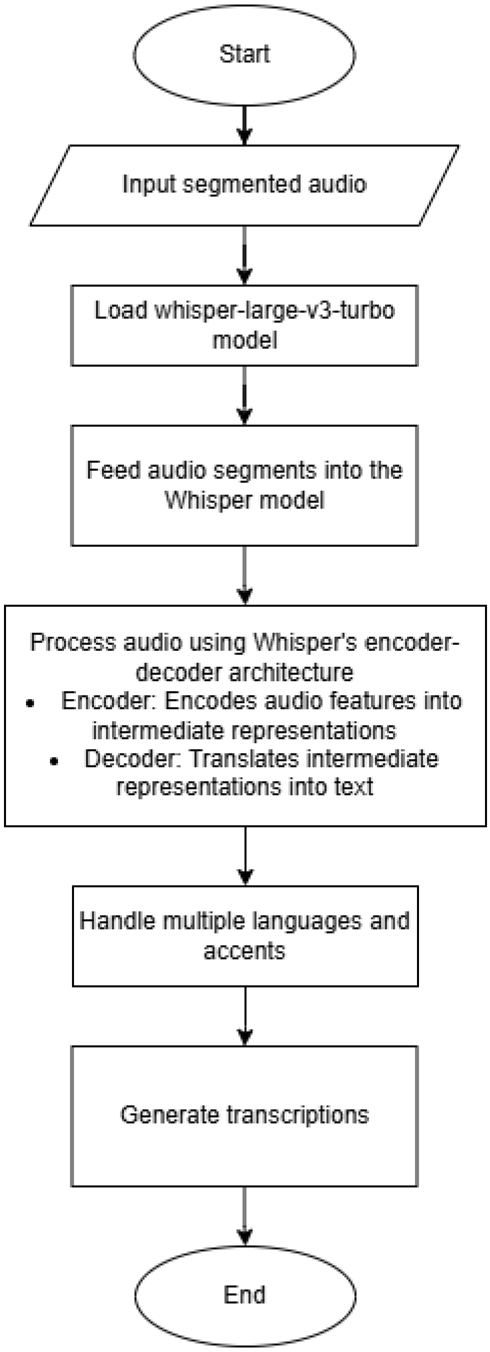


Figure 4.4 Speech-to-Text Flowchart.

The Speech-to-Text process begins with the segmented audio generated from the Voice Activity Detection (VAD) module. These audio segments are then input into the Whisper-large-v3 model, a transformer-based sequence-to-sequence architecture. The model processes the audio by first encoding its features into intermediate representations through its encoder. These representations are subsequently passed to the decoder, which translates them into text. Whisper's ability to handle multiple languages and accents ensures accurate transcription for diverse user-generated content, adapting the transcription process based on the detected language. Additionally, the model demonstrates resilience to common audio disruptions such as background noise and music, maintaining high clarity in the transcriptions. The transcriptions for each segment are then combined to produce a complete and cohesive text output, representing the spoken content with state-of-the-art accuracy.

4.4 Active Speaker Detection (ASD) Module

The Active Speaker Detection (ASD) Module is a crucial component in identifying the speaking characters within specific video timestamps. This identification ensures that the dialogue is synchronized with the name of the active speaker, thereby enhancing the overall completeness of the movie script.

4.4.1 Workflow of Active Speaker Detection

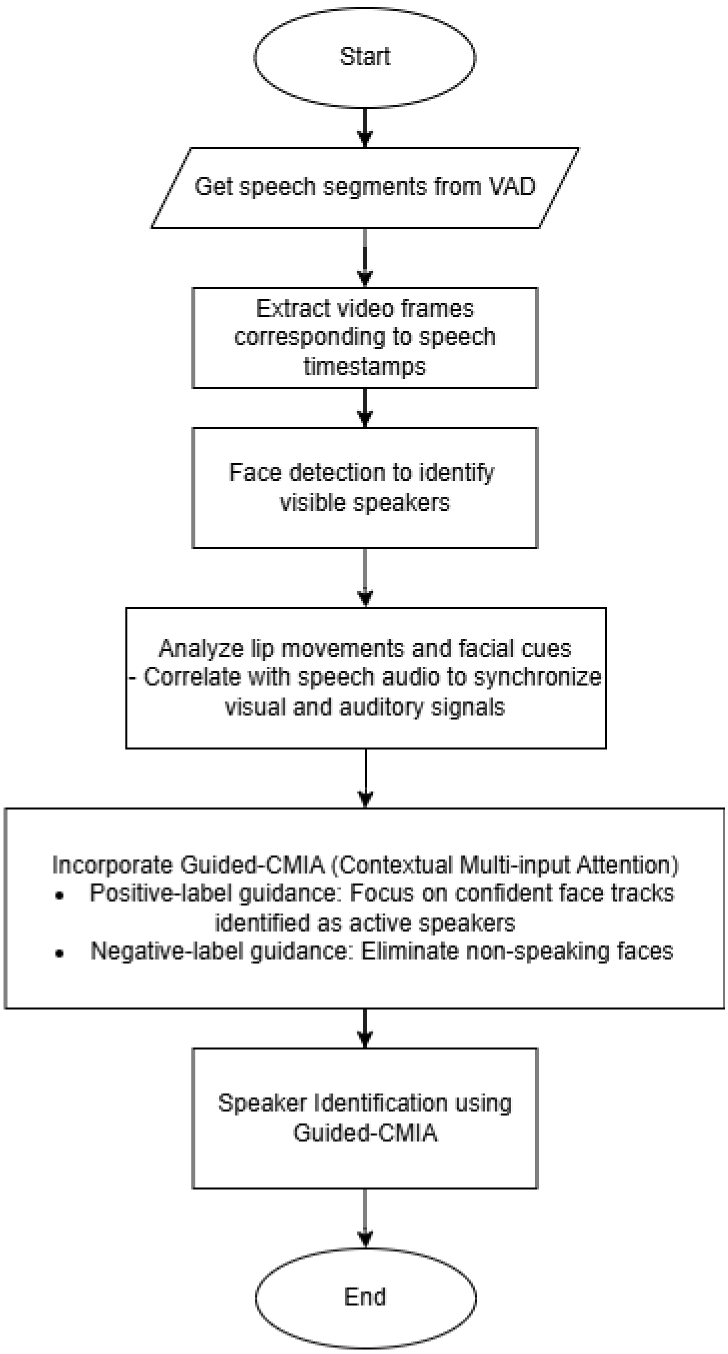


Figure 4.5 Active Speaker Identification Flowchart.

The Active Speaker Detection (ASD) module starts by utilizing speech segments identified by the Voice Activity Detection (VAD) module, which timestamp the precise moments when speech occurs within a video. Following this, the ASD process involves several key steps. First, timestamp-based segment extraction isolates the video frames corresponding to these speech segments, focusing on relevant portions of the video for analysis. Next, visual processing and face detection are performed on these extracted segments, to identify potential speakers visible during the speech. The system then analyzes lip movements and other facial cues from the detected faces to correlate with the audio, synchronizing these visual cues with the speech timestamps to identify the speaker. Finally, the system attributes the speech to the corresponding face, ensuring accurate speaker identification.

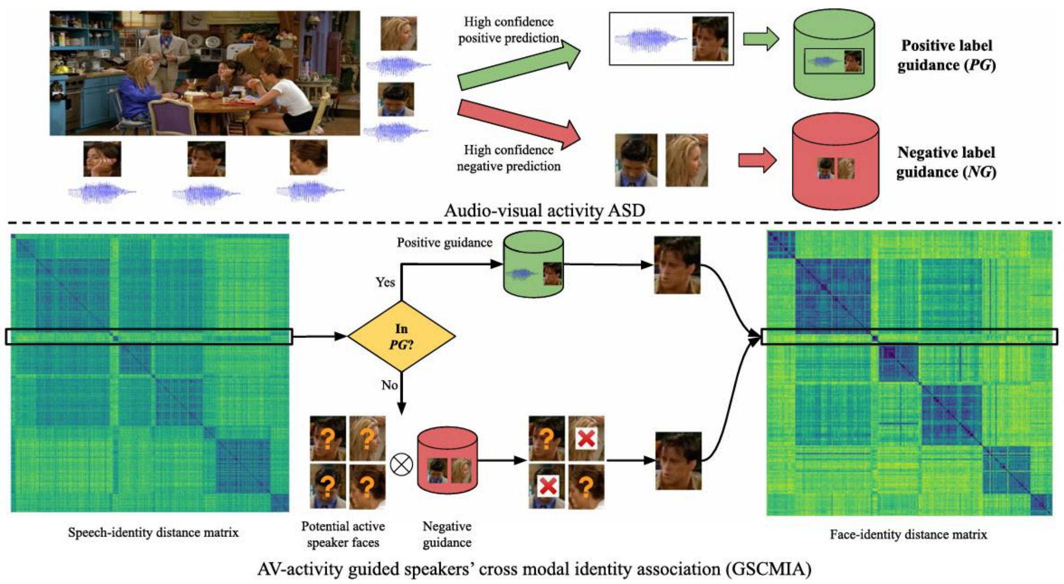


Figure 4.6 An overview of the guided speaker identity association across modalities (GSCMIA) audio-visual activity [19].

To further enhance the reliability of speaker detection, the system incorporates the Guided-CMIA (Contextual Multi-input Attention) approach [19]. This advanced technique integrates audio-visual activity information with speaker identity data, refining the process of speaker identification. Guided-CMIA uses high-confidence audio-visual (AV) activity predictions to form positive and negative guidance sets. Positive-label guidance focuses on face tracks that are most confidently identified as active speakers, while negative-label guidance eliminates non-speaking faces from consideration. By synthesizing these insights, Guided-CMIA significantly improves the ASD module's accuracy and robustness, particularly in complex video scenarios where multiple speakers and significant background noise may be present.

4.5 Face Detection Module

Face detection involves identifying and locating human faces in digital images or video sequences. This module is responsible for systematically scanning each frame, detecting and cropping out all faces of the frames. This module underpins the accuracy of subsequent tasks, such as active speaker identification and face recognition.

4.5.1 Workflow of Face Detection

The Face Detection Module utilizes the Dlib library, known for its robust and reliable performance in face detection tasks. Dlib, a modern C++ toolkit that integrates advanced machine learning algorithms, is particularly suited for high-performance real-time applications. The module begins its process by loading each video frame, which has been filtered before.

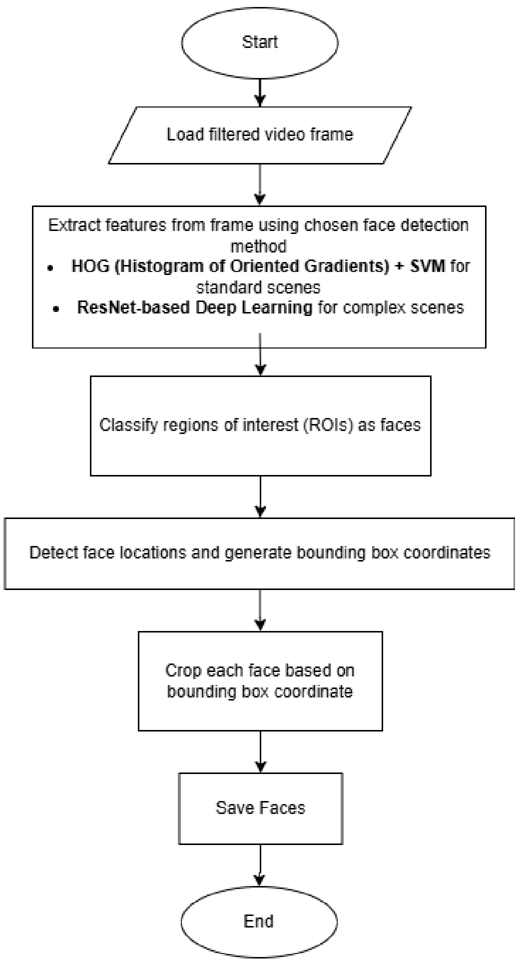


Figure 4.7 Face Detection Flowchart.

The core of the face detection process is executed using Dlib's Histogram of Oriented Gradients (HOG) combined with a linear Support Vector Machine (SVM) [20]. This method is highly effective at detecting faces, even under challenging conditions such as variations in lighting, scale, or angle. The combination of HOG for feature extraction and SVM for classification allows for precise detection across a wide range of video content.

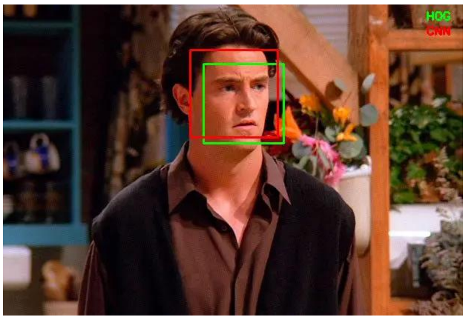


Figure 4.8 Two face detection functions are available in the dlib library. One is a deep learning MMOD CNN face detector, and the other is a HOG + Linear SVM face detector [20].

In scenarios where complex scenes are encountered—for instance, when faces are partially occluded—the system escalates its detection capabilities by employing Dlib's deep learning-based face detector, which utilizes a ResNet architecture. This approach offers enhanced accuracy by tapping into the power of deep neural networks, making it particularly effective in handling difficult cases where traditional detection methods might struggle.

Once faces are detected in a frame, the module proceeds to individually crop each face using the bounding box coordinates provided by the detection algorithm.

4.6 Face Clustering Module

Face clustering is the process of grouping similar face images so that each cluster corresponds to a distinct individual within a video. The Face Clustering Module is tasked with organizing the detected and cropped face images into distinct clusters, where ideally, each cluster matches a unique individual or character seen throughout the video content.

4.6.1 Workflow of Face Clustering

The workflow of the Face Clustering Module begins with the application of the Chinese Whispers Algorithm, a highly efficient and robust graph-clustering technique well-suited for unsupervised learning scenarios, which is particularly advantageous for the task of face clustering [21].

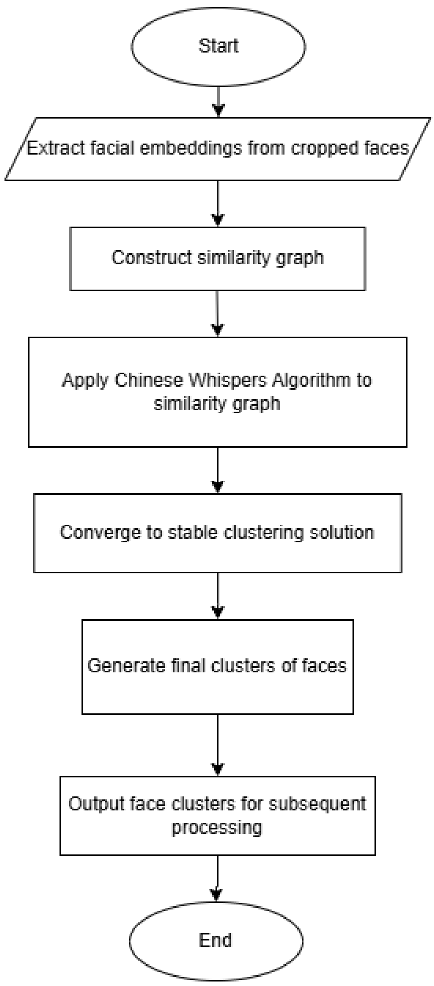


Figure 4.9 Face Clustering Flowchart.

The process initiates with feature extraction, whereby the detected faces—previously isolated and cropped—are converted into vector representations known as facial embeddings. These facial embeddings capture crucial facial features, including elements such as the eyes, nose, mouth, and overall facial geometry. The essence of these embeddings lies in their ability to extract distinctive features that enable the system to differentiate between different individuals.

Once the facial embeddings have been generated, the system constructs a similarity graph. In this graph, each face is depicted as a node, and the relationships between them are represented by edges that are weighted based on the degree of similarity between the corresponding facial embeddings. This graph-based representation is instrumental in enabling the system to make a comparative and relational analysis of all detected faces within the video segment.

Next, the Chinese Whispers algorithm is applied to this similarity graph. The algorithm functions by iteratively propagating labels across the network of nodes, with these labels being adjusted according to the weighted connections between nodes. As the algorithm iterates, it gradually converges toward a stable clustering solution, wherein nodes that represent highly similar facial features are grouped under the same label. The convergence of the algorithm results in distinct clusters, with each cluster ideally consisting of multiple instances of the same individual's face across different scenes in the video.

Figure 4.10 Example of clustered faces.

The final outcome of this workflow is a set of clusters, each representing a unique character or individual throughout various segments of the video. These clusters serve as the foundational data for subsequent processes, such as character identification and dialogue mapping.

4.7 Face Labelling Module

The Face Labeling Module is a user-assisted process, where the content creator or uploader plays an active role in labeling the clustered faces identified in the video. This module is crucial in ensuring that each cluster of faces corresponds accurately to distinct characters.

4.7.1 Workflow of Face Labelling

The Face Labelling process within the module is designed to be both intuitive and efficient, allowing content creators or uploaders to easily label the face clusters generated by the preceding Face Clustering Module. The system initiates this process by presenting the uploader with clusters of faces—these clusters represent various instances of the same individual appearing across multiple scenes in the video. To assist in proper identification, the system also provides supplementary metadata, such as the specific frames or timestamps where these faces were detected, along with any additional contextual information, like associated dialogues when available. This metadata helps the uploader gain a deeper understanding of the context around each cluster, ensuring more accurate labelling.

The labelling process is conducted through a user-friendly interface designed to make the task of identifying and labelling characters straightforward. This interface typically includes features such as the ability to view the corresponding video segments related to each face cluster, which aids the uploader in precisely identifying the character depicted. Additional functionalities, like dropdown menus or search tools, are provided to further streamline the process, allowing for the quick and consistent input of character names, roles, or other relevant identifiers.

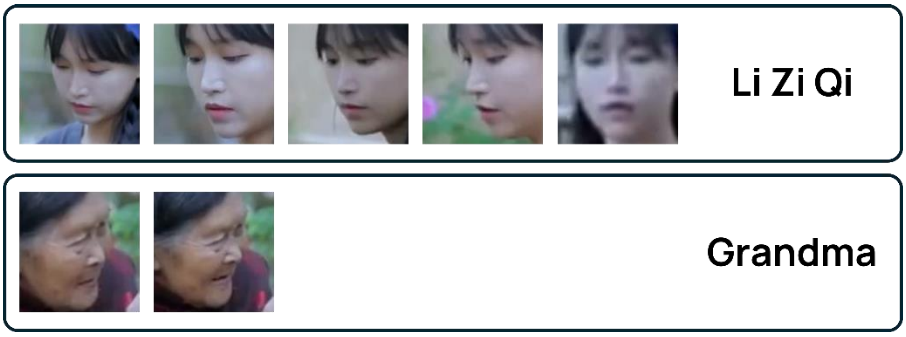


Figure 4.11 Example of clustered faces labelled with names.

Once the uploader has labelled each cluster, the system finalizes the labeling process by storing the labels in association with the respective face clusters. These labels are then utilized in subsequent modules, including face recognition, active speaker detection, and dialogue mapping.

4.8 Character Face Recognition and Annotation Module

The Character Face Recognition and Annotation Module plays a crucial role in the final identification and annotation of characters of video frames. This module uses the face embeddings generated by the Face Embedding Module to accurately identify specific characters. Once identified, the module visually distinguishes these characters by annotating the video with colored bounding boxes around their faces.

4.8.1 Workflow of of Character Face Recognition and Annotation

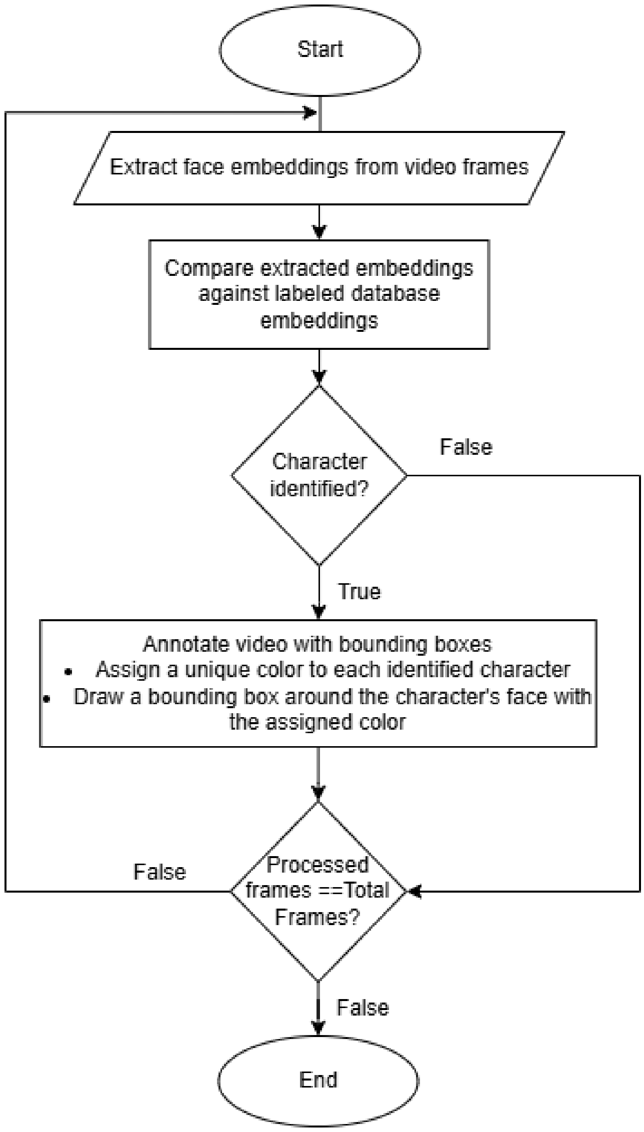


Figure 4.12 Face Recognition and Annotation Flowchart.

The process of Character Face Recognition begins by comparing the face embeddings extracted from the video frames against the labeled embeddings stored in the system's database. The primary objective of this comparison is to ensure accurate identification and differentiation between characters, maintaining consistency in character recognition throughout the video.

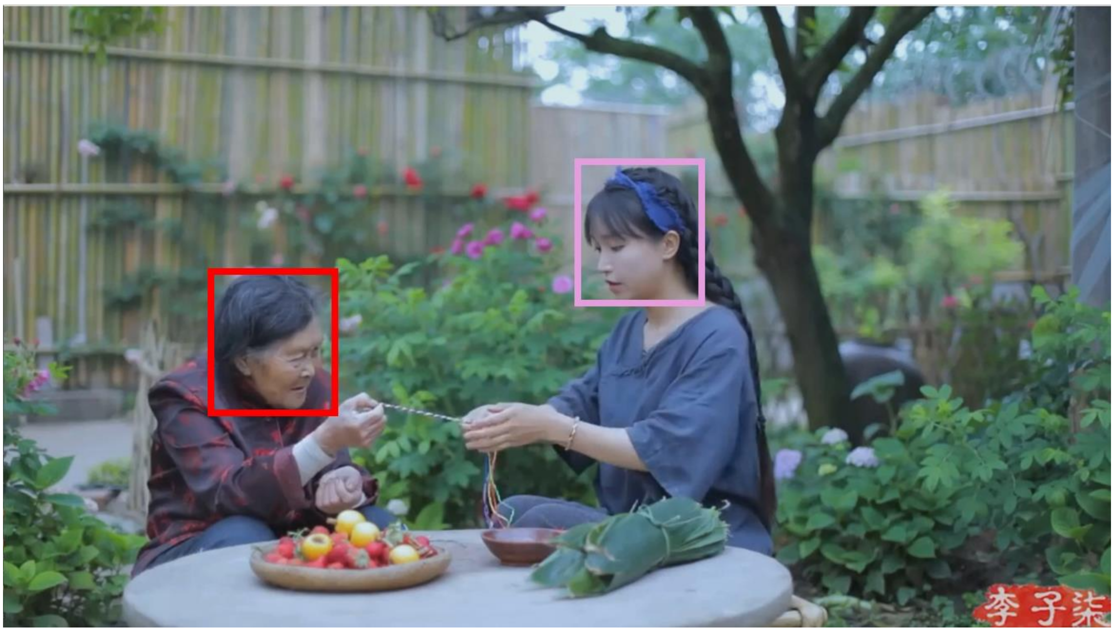


Figure 4.13 Example of different characters in the frames, each labeled with a different color bounding box.

Once a character is successfully identified, the system proceeds to the annotation phase. In this step, the video is marked with colored bounding boxes over each recognized character's face. Each character is assigned a unique color, providing a clear and immediate visual cue that helps users easily track and differentiate between different characters as they appear on screen.

4.9 Image Captioning Module

The Image Captioning Module plays a significant role in generating comprehensive and descriptive narratives for the visual scenes within video content. This module operates on video frames that have already undergone processing through the Character Face Recognition and Annotation Module . By harnessing the advanced capabilities of Gemini and GPT-4o [22], the platform is able to produce detailed, context-aware captions that able to enhance the quality of the audio description generated in the LLM-Based Multi-Agent System Module.

4.9.1 Workflow of Image Captioning

The primary objective of the Detailed Image Captioning Module is to translate visual content into meaningful text that elevates the accessibility and understanding of videos for all users, particularly through the creation of audio descriptions. The workflow begins by organizing video frames—after they have been labeled and annotated with character-specific information, such as color-coded bounding boxes and names—into batches. Each batch represents a coherent sequence of frames that corresponds to a specific scene or segment within the video.

These color-annotated frames are then processed by GPT-4o to generate detailed captions. GPT-4o takes a holistic approach, analyzing the entire context of each scene by considering both the visual elements and the character annotations. This ensures that the captions generated do not merely describe isolated objects or characters, but also capture the broader essence of the scene, including actions, interactions, and emotional undertones.

The process continues with GPT-4o integrating character recognition data and the specified colors of each character into the captions. For example, suppose a character is highlighted with a red bounding box. In that case, GPT-4o might initially generate a caption like " <pink>is dressed in a traditional blue outfit with a blue headband, engaging in the activity of stringing colorful beads onto a thread. <red>, highlighted within a red rectangle, is attentively observing or participating in the task." Leveraging the character recognition data, this caption is then refined to be more specific, such as " Li Zi Qi is dressed in a traditional blue outfit with a blue headband, engaging in the activity of stringing colorful beads onto a thread. Grandma, highlighted within a red rectangle, is attentively observing or participating in the task."

4.10 Movie Script Generation Module

The Movie Script Generation Module is designed to integrate various streams of textual data into a cohesive and structured document that reflects the overall narrative of the video. This script encapsulates the essence of each scene by combining captions, dialogues, character annotations, and scene descriptions, ensuring that both visual and auditory elements are seamlessly woven together.

4.10.1 Workflow of Movie Script Generation

The workflow of the Movie Script Generation Module begins with the careful compilation of dialogue and image captions. To achieve this, the system first collates the detailed image captions produced by the Image Captioning Module with the dialogues transcribed by the Speech-to-Text (STT) Module. To enhance accuracy and context, these dialogues are further enriched with speaker identifications provided by the Active Speaker Detection Module, ensuring that the dialogue is associated with the correct characters.

Once all textual elements have been gathered, the system organizes these components chronologically based on their timestamps. This approach is critical for maintaining the narrative flow of the video, as it aligns the visual descriptions precisely with the corresponding dialogues. The result is a structured script that mirrors the video's timeline, presenting a seamless and coherent narrative that faithfully represents both the visual and auditory aspects of the content.

4.11 Single-Prompt Multiturn Multi-Agent Reasoning System (SMARS)

To efficiently transform complex video dialogue into concise, well-structured, and multilingual audio descriptions for blind users, this project introduced a novel modular framework called the Single Prompt Multiturn Multi-Agent Reasoning System (SMARS) . Unlike traditional distributed multi-agent systems that rely on external APIs and orchestration frameworks (e.g., LangChain, CrewAI, AutoGen), SMARS operates entirely within a single LLM prompt session , simulating multiple agents with clear roles and responsibilities executing one after another in a deterministic, highly controlled reasoning loop.

Built using structured prompt engineering, the system organizes the workflow across four distinct stages—Room A through Room D —each simulating a collaborative environment where specific tasks are performed by distinct agent personas. These agents, such as Sherlock (investigator), Jolin (rewriter), Xiaoli (language flow expert), and Xiaoming (character counter), emulate real-world editorial roles to refine raw video dialogues into linguistically and semantically optimized audio descriptions. This approach is notably language-agnostic, allowing content to be described in any target language (e.g., Chinese, Malay, English), expanding accessibility globally.

4.11.1 Workflow of LLM-Based Multi-Agent System

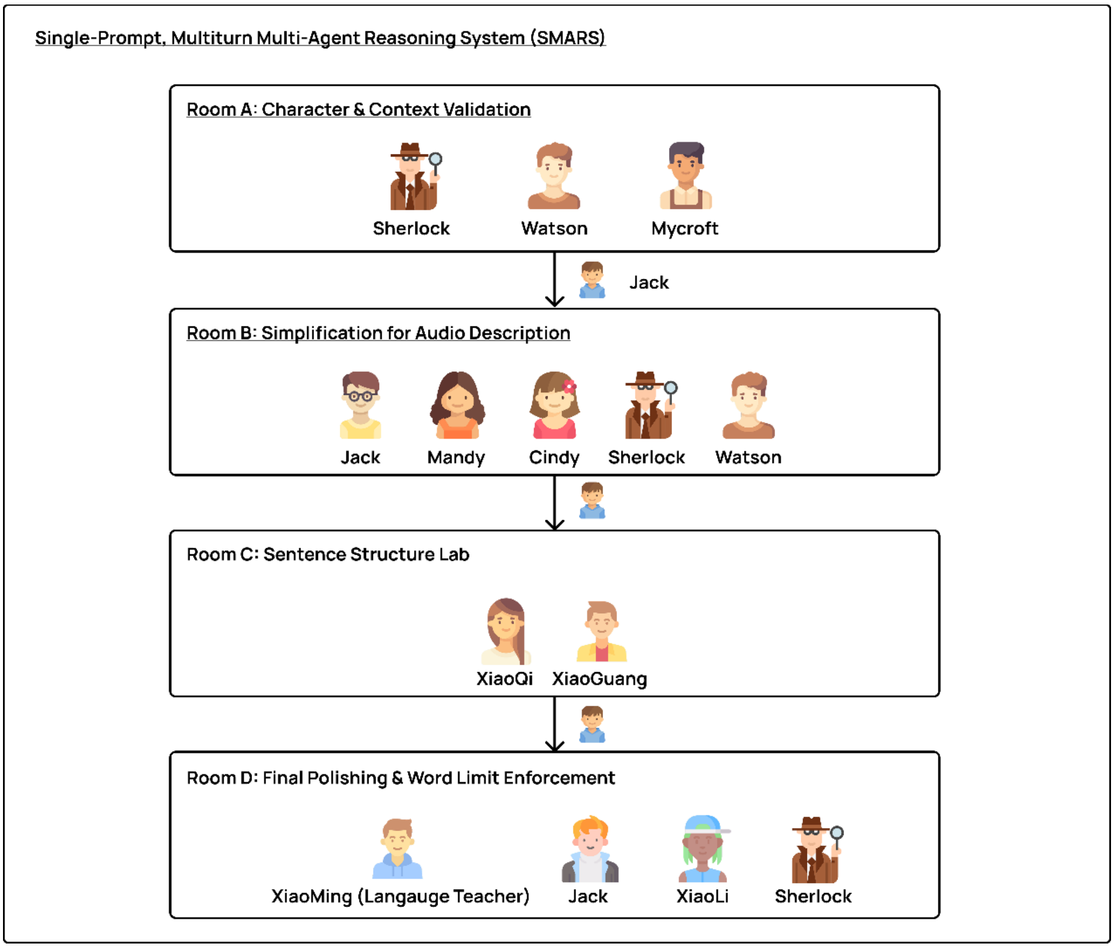


Figure 4.14 LLM-Based Multi-Agent System of Automated Audio Description System.

This system simulates a structured editing and reasoning workflow designed to transform raw visual dialogue into polished, concise, and accessibility-friendly audio descriptions for blind users. The process unfolds across four conceptual "rooms," each staffed by specialized virtual agents that represent different reasoning or editing roles. This compartmentalized approach mirrors the structure of an editorial or accessibility task force, where each team handles a specific phase of the transformation pipeline.

In the first phase, known as Room A (Novel Consistency Room), the focus is on context verification and character resolution. The agents in this room ensure that the dialogue is factually aligned with the novel version of the movie, eliminating contradictions, resolving unknown characters, and correcting any internal inconsistencies. Sherlock, the primary investigator, probes all unknown elements, unlabelled individuals, or mismatches between character appearances and dialogue context, drawing from a library of known character traits to provide evidence-based identifications. Mycroft, the visual validator, meticulously checks that facial features, clothing, and accessories align with established character traits, challenging Sherlock's conclusions when necessary. Watson, the context historian, supplements their investigation with historical context from previous dialogues, highlighting recurring characters or already-clarified scenes. Finally, Jolin, the rewriter, synthesizes the results and rewrites the dialogue, ensuring character labels are correct and the dialogue is logically cohesive. The output from this room is a validated, preprocessed version of the dialogue, ready for the next step.

Room B (Detail Reduction Room) focuses on semantic simplification and prioritization. The agents in this room refine the verified dialogue by removing redundancies and unnecessary visual details, ensuring that only essential narrative actions are preserved for audio storytelling. Jack, Mandy, and Cindy, the audio describers, work together to condense complex visual scenes into clear, action-focused summaries. They cut verbose or repeated descriptions, especially those involving clothing, facial expressions, or gestures unless they are critical to the plot. Watson acts as a repetitive filter, cross-checking recent outputs to identify and flag echoed descriptions, offering alternative sentence structures or more varied language. Sherlock oversees this reduction process to ensure coherence and that no critical details are lost. The result is a concise, vivid snapshot that balances engagement with brevity.

In Room C (Subject Clarity Room), the agents focus on ensuring that each sentence has a single, unambiguous grammatical subject, preventing confusion, especially in audio form. Xiaoguang and Xiaoqi specialize in restructuring sentences with multiple subjects so that each maintains a clear logical subject. For example, if two characters are performing different actions in the same sentence, they will determine which action should be the main clause and place the other as background context or a supporting phrase. This restructuring eliminates ambiguity, ensuring that listeners can easily follow the actions in the scene without confusion about who is doing what. The output from this room is grammatically coherent and ensures clarity in terms of actions.

Finally, in Room D (Final Polishing Room), the focus is on language fluency, word count compliance, and qualitative verification. Xiaoli, the language fluency specialist, refines clunky or overly literal sentence fragments, rephrasing them to match natural narrative rhythms. Xiaoming, the word count checker, verifies that each sentence adheres to the predefined character count limit. If the limit is exceeded, Xiaoli trims non-essential clauses, ensuring that only key content remains while maintaining narrative accuracy. Holmes, the narrative fact verifier, ensures that no essential plot elements are omitted during this trimming process. This room also checks the vocabulary to ensure it aligns with the intended blind user's profile, avoiding unfamiliar or inaccessible terms. If the sentence meets all criteria, it is finalized; otherwise, it may be sent back for adjustments. If the dialogue lacks meaningful content, such as generic conversation, Room D will return a null code AD404, indicating the dialogue is not worth including.

Throughout the entire process, John, the logical messenger, acts as the intermediary examiner. He ensures that the dialogue transitions smoothly from one room to the next, maintaining consistency and delivering the finalized output for compilation. His role guarantees that the entire multi-agent pipeline runs predictably and transparently, ensuring a successful and high-quality result.

CHAPTER 5

System Implementation

5.1 Hardware Setup

The hardware used for developing, processing, and testing the online video streaming platform and the automated audio description system is outlined below. This laptop is the central hub for all developmental activities, ensuring the systems run smoothly and efficiently.

Table 5.1 Specifications of Laptop.

Description	Specifications
Model	HP Victus by HP Laptop 16-e0122AX
Processor	AMD Ryzen 5 5600H with Radeon Graphics
Operating System	Windows 11
Graphic	NVIDIA GeForce RTX 3050 Laptop GPU, AMD Radeon(TM) Graphics
Memory	16GB DDR4 RAM
Storage	512GB SSD (MTFDHBA512TDV-1AZ1AABHA), 1TB SSD (TEAM TM8FPD001T)

5.2 Software Setup

The software required for this project was installed and configured on the development system to build a robust environment. The major software components used were:

1. **Python 3.10:**
   This served as the primary programming language, alongside essential libraries like OpenCV, dlib, and PyTorch for image and video processing.
2. **Whisper by OpenAI:**
   Deployed for Speech-to-Text (STT) processing, Whisper provided state-of-the-art transcription accuracy, helps for generating textual representations of spoken content.

3. **GPT-4o:**
   Employed for detailed image captioning, GPT-4o processed batches of annotated video frames to generate comprehensive and contextually aware image descriptions.

4. **Gemini 2.0 Flash:**
   Multilingual Large Language Model (LLM) used in the LLM-Based Multi-Agent System module. This model tailored generated audio descriptions to various user personas, ensuring inclusivity and adaptability across diverse content.

5. **CUDA Toolkit 11.8:**
   Integrated into the system to enable GPU acceleration for computationally intensive workflows such as I-frame extraction, face detection, and active speaker detection.

6. **FFmpeg:**
   Utilized for video manipulation tasks, including I-frame extraction.

5.3 Settings and Configuration

This section outlines the system settings, library dependencies, environment configurations, and model parameters required to set up and run the **"Enjoyable" automated audio description platform** efficiently on a local machine. Proper configuration ensures that the system modules — from video processing to multi-agent reasoning — operate cohesively.

5.3.1 Environment Setup

The system was developed and tested on a **Windows 11 64-bit** environment with NVIDIA RTX GPU support. The preferred environment manager is **Anaconda**, with Python 3.10. The following key Python libraries and frameworks are installed:

1. opencv-python
2. torch
3. transformers
4. openai-whisper
5. mediapipe

6. moviepy
7. librosa
8. numpy
9. pandas
10. scikit-learn
11. ffmpeg

5.4 System Operation

The Voice Activity Detection (VAD) module segments the audio from a video into speech and non-speech intervals. The audio from the video file was extracted and resampled into a mono format at 16kHz using FFmpeg. The resampled audio was then passed through the pretrained PyAnnote pipeline, which processed it as a waveform sample. From this analysis, only segments containing speech were identified and segmented, with relevant metadata, including timings. This segmentation forms the basis for identifying speakers and text transcription.

```
[1.752219-3.152844] SPEAKER_01:
[8.755344-9.970344] SPEAKER_01:
[12.062844-14.020344] SPEAKER_01:
[16.382844-17.445969] SPEAKER_00:
[17.783469-19.420344] SPEAKER_01:
[20.095344-29.950344] SPEAKER_01:
[31.620969-32.937219] SPEAKER_01:
[36.447219-37.594719] SPEAKER_01:
[44.429094-46.234719] SPEAKER_01:
[46.572219-48.023469] SPEAKER_01:
[48.529719-49.221594] SPEAKER_01:
[49.626594-53.035344] SPEAKER_01:
[53.102844-54.925344] SPEAKER_01:
[57.355344-58.536594] SPEAKER_01:
[65.235969-65.961594] SPEAKER_01:
[72.171594-73.403469] SPEAKER_01:
[79.022844-79.950969] SPEAKER_01:
[89.046594-90.582219] SPEAKER_01:
[96.927219-98.057844] SPEAKER_01:
[114.224094-116.215344] SPEAKER_01:
[121.159719-122.037219] SPEAKER_01:
[124.214094-125.074719] SPEAKER_01:
[125.850969-127.032219] SPEAKER_01:
[129.698469-131.942844] SPEAKER_01:
[136.752219-137.595969] SPEAKER_01:
[137.967219-139.249719] SPEAKER_01:
[142.709094-144.869094] SPEAKER_01:
[145.932219-147.450969] SPEAKER_02:
[147.923469-148.868469] SPEAKER_02:
[150.977844-153.914094] SPEAKER_01:
[256.024719-257.054094] SPEAKER_01:
[257.847219-259.214094] SPEAKER_01:
```

Figure 5.1 Result of VAD.

The system begins by extracting Intra-coded frames (I-frames) using a video decoding tool such as FFmpeg within non-speech segments. I-frames are key frames that contain the complete image data for a frame, making them ideal for scene analysis. By selecting only I-frames, this module ensures efficient frame sampling by skipping redundant intermediate frames, thereby optimizing the processing speed and reducing computational load throughout the pipeline.

Figure 5.2 Result of I-Frame Extraction.

The filtered frames are then passed through the Face Detection Module, which identifies and temporarily excludes frames that consist primarily of faces.

Figure 5.3 Result of Face Detection.

Chinese Whisper Algorithm is used to module groups similar faces across frames

0	9/5/2025 7:18 AM	File folder
1	9/5/2025 7:18 AM	File folder
2	9/5/2025 7:18 AM	File folder
3	9/5/2025 7:18 AM	File folder
4	9/5/2025 7:18 AM	File folder

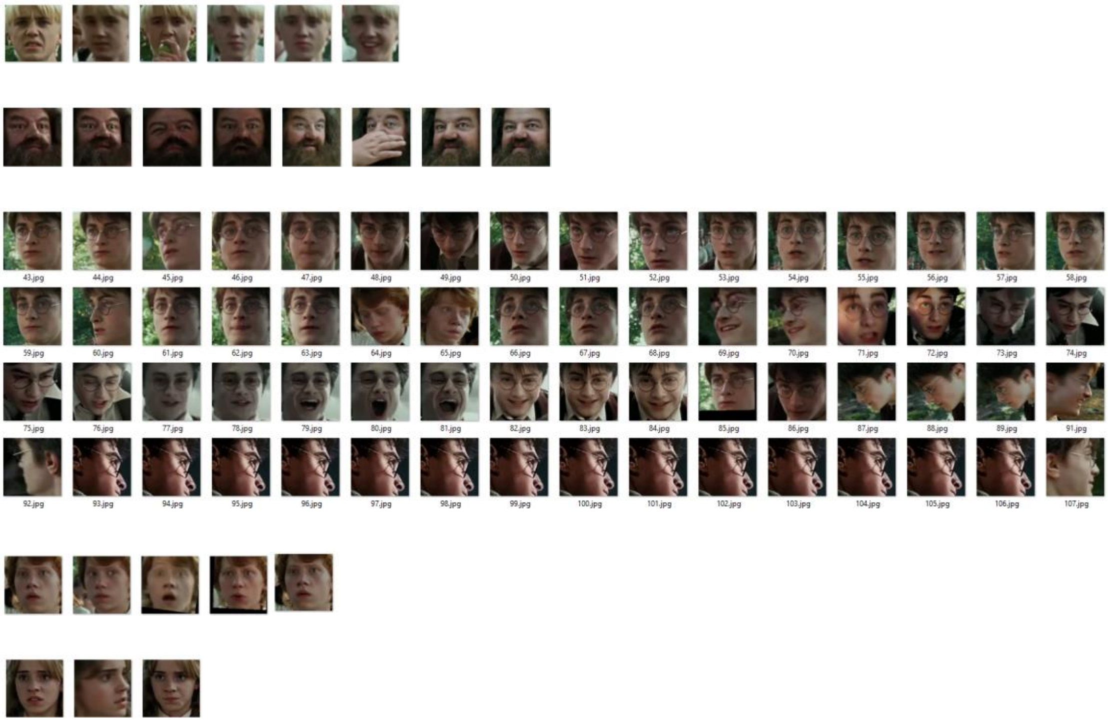


Figure 5.4 Result of Face Clustering.

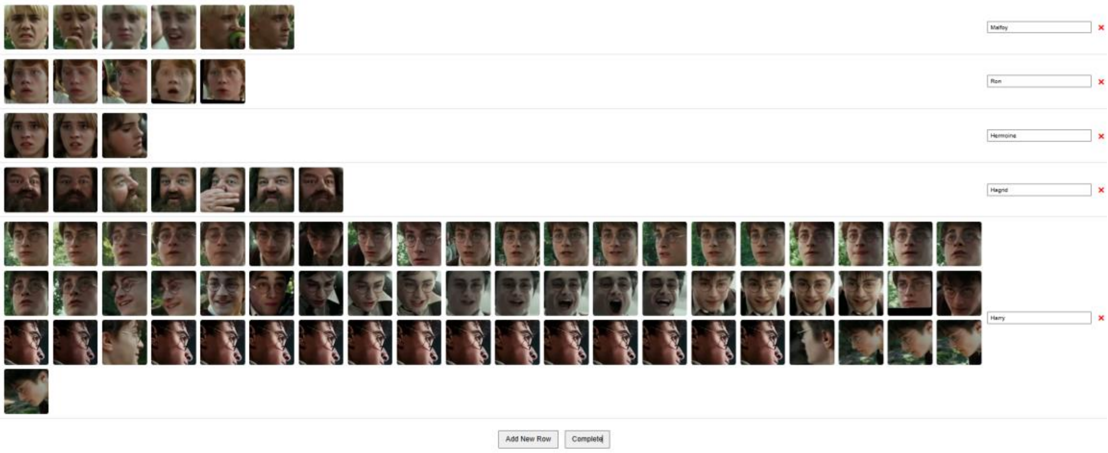


Figure 5.5 UI for Users to label clustered faces.

Figure 5.6 Result of Face Recognition and Annotation.

The Detailed Image Captioning module was implemented and tested to generate detailed and accurate descriptions of visual elements in video frames, with a particular focus on people and objects. The large language model was employed as a "detective" to observe and analyze the intricate details of the frames more effectively. This approach aimed to enhance the precision and comprehensiveness of the generated captions. Below is the prompt used to guide the large language model in producing detailed image captions:

```
Generate detailed descriptions for an audio descriptor to create audio descriptions based on visual
elements of given frames. If people are framed by a colored bounding box, replace the mention of
"people" with the corresponding color.

Observe and describe all visual elements of the given frames in detail, as if you are Sherlock Holmes.
Focus on characteristics of people and other visual elements with precision.

- **People**: Describe their physical traits, expressions, attire in very detailed fashion, must include
clothing patterns and color, hairstyles, eye color, posture, and any notable features. Use the color of
the bounding box to refer to them when applicable.
- **Objects**: Detail colors, shapes, sizes, textures, and positions.

# Steps

1. Examine each frame closely.
2. Identify notable people, objects, and settings.
3. Describe each element systematically, ensuring thoroughness.
4. Substitute "people" references with bounding box colors when necessary.

# Output Format

Provide a detailed description in paragraph form, covering all sections: People and Objects
comprehensively.

# Examples

 - **Original Description**: "A girl with long, wavy brown hair and bright green eyes wears a red polka
dot dress and is kicking a ball."
 - **Updated Description**: "<blue> with long, wavy brown hair and bright green eyes wears a red polka
dot dress and is kicking a ball." (assuming the girl is within a blue bounding box)

The ID of the images corresponds to the frame number. Take a deep breath and think, If you believe some
frames can be combined, then combine them, if there are separate actions, then do not combine them.
Output the result in the following format:
[ID Num] XXXXXXX
[ID Num - ID Num] XXXX
[ID Num] XXXX.
```

Figure 5.7 Prompt used to generate image captions.

```
[0.0-1.0] The scene is dark and almost completely obscured. What appears to be a dark shape stands vaguely against a backdrop of indistinct vertical lines.
[1.0- ] Sunlight streams through the dense canopy of a forest, illuminating the scene with a soft, dappled light. A bushy bearded man with long hair stands amid the trees, wearing a dark vest over a gray garment. Attached to the vest are the bodies of several small, furry animals, seemingly hunted. The tall trees with thick trunks and leafy branches form a natural enclosure, adding to the serene and wild ambiance of the scene.
[3.2-5.2] A group of students is standing in what appears to be a forest clearing. In the midst of them, on a mossy mound, lies a collection of thick, leather-bound books. Ron, a teenage boy with short, ginger hair, stands among them. Beside Ron, Harry, a teenage boy with glasses, observes the books with a keen expression. Suddenly, the air thickens with tension as Malfoy, a teenage boy with a sneer, advances into view, flanked by his cronies.
[5.2-8.2] Hermoine, a teenage girl with bushy brown hair, appears with a concerned expression. The students stand silently, and tension fills the air. Then Hermoine, a teenage girl with curly hair, stands silently, observing the books, her expression thoughtful.
[10.0-10.0] In a sun-dappled clearing amidst tall trees, stands a large, bearded man. He has a heavy, dark brown beard and wears a layered outfit, including a dark vest over a lighter, possibly woolen shirt, with dark trousers and oversized brown boots. A wide belt with pouches is secured around his waist. Next to him is an animal, that appears to be a young hippogriff, covered in gray feathers and standing on thin legs. The clearing is enclosed by a low stone wall in the background, with more dense trees forming a backdrop.
[11.3-11.3] In a sun-dappled forest clearing, a gray Hippogriff, its body a mix of horse and eagle, trots forward, its feathered wings slightly spread. The half-horse half-eagle animal's claws dig into the soft earth. Behind the Hippogriff stands a large man with long, dark hair, wearing a dark green vest over a brown jacket. A bag hangs from his back and he faces the beast, his right arm slightly extended and hand open. The ground around them is covered in a thin layer of mulch, with a low stone wall bordering the clearing, dense green trees surrounding the area. There are two small wooden structure on the ground behind the Hippogriff.
[14.1-16.1] In a dense forest, a massive, mythological creature with a human-like face and the body of a lion stands on a small, stone-covered ridge. The upper part of the creature is covered in light grey, mottled fur, while its large wings are shaded in darker grey tones. Tall trees, some with bare, brown trunks, reach high into the sky, their green leaves creating a thick canopy that filters the sunlight. The creature turns its head slightly, gazing forward as Harry, Ron, and Hermoine are among the crowd standing together with other teenagers, the red and gold striped ties with black robes suggesting they are students. A pile of thick books lies at the front, with one open, its pages displaying detailed diagrams. Ron, is looking to his right with a slightly concerned look and Harry, who wears round glasses, also looking to his right.
[17.5-17.5] A group of students wearing white shirts or black robes stand on a dirt ground surrounded by a low stone wall. In the open space behind the students, a white horse-like animal stands. A large, hairy man, wearing dark clothing, walks toward the animal. The area is a wooded location with many green trees. A large tree trunk is visible to the left.
[19.5-19.5] The scene unfolds in a forest clearing, the ground covered in brown dirt, enclosed by a low stone wall. A large, bearded man with long dark hair, wearing rugged clothing, stands near the center, facing a crowd of young students, dressed in white shirts and dark robes, their backs to the "camera." A pale quadruped, its species is not clear, stands closer to the man, beneath the shade of large trees, with dark-green foliage, casting dappled shadows on the ground. An old, metal water trough sits on the left side of the clearing.
[30.0-31.0] In the dense, green forest, Hagrid, with his wild, bushy hair and beard, and wearing a buttoned-up gray shirt beneath a fur-trimmed coat, looks on with a mixture of awe and concern. At the edge of the frame, to the left, a creature with gray feathers and large wings peers out from behind a tree. Then Hagrid claps his hands together, perhaps in delight, maybe in anticipation.
[33.0-36.0] A group of students stand together in a forest. In the center, Harry, a teenage boy with round glasses and a white shirt, is looking ahead. Next to him is Ron, a teenage boy with red hair, looking a little surprised, also dressed in a white shirt and dark tie. The group stands in front of a stack of weathered books piled high on a rough stone surface. The camera angle shifts slightly, moving Harry, a teenage boy, to the center again. Beside him is Hermoine, a teenage girl with brown curly hair wearing a black robe, looking on with a keen expression. The shot widens slightly, showing Ron, a teenage boy, is to the right of Harry, a teenage boy. Then the scene transitions to a wide view of a tall man with a big beard, wearing a dark outfit, standing in the woods.
[37.6-43.6] In a leafy outdoor setting, Harry, a teenage boy with round glasses, stands amongst a crowd. Ron, a teenage boy with bright red hair, is visible over Harry's left shoulder. The pupils wear their Hogwarts uniforms. In front of them, on a mossy surface, lie several thick, ancient-looking books and a drawing of a dragon-like creature. Harry gazes with wide eyes as if something is about to happen. The teenagers and youngsters around them look on with anticipation. Ron's face also conveys a sense of anticipation.
[46.3-46.3] In a sun-dappled forest clearing, a creature that resembles a gray horse with large, doleful eyes stands alone, its tail swishing slightly. The place is enclosed by a low wall constructed of rough stones, and light filters through the canopy overhead. A slender person dressed in dark clothing with an obscured face stands still, appearing to observe the quadruped. To the side, a larger group of people in darker, more formal attire stands together, their faces not distinguishable at this distance. The ground is covered in reddish-brown leaf litter. Tall trees encircle the clearing, their trunks casting long shadows across the scene.
[48.1-48.1] In a forest clearing, a gray horse-like creature stands to the left, basking in the sunlight filtering through the trees. To the right, a group of people are gathered, their forms mostly obscured by the distance and shadows. In the center, a person dressed in dark, flowing robes stands facing the group, seemingly addressing them. The clearing is enclosed by a low wall constructed from dark stones, adding a sense of boundary to the open space amidst the dense woodland.
[49.3-49.3] In a sun-dappled forest clearing, a thestral with a grey coat stands motionless, its feathery wings folded against its body. Across the clearing, a middle-aged man with black robes watches the thestral. A group of young students dressed in school uniforms huddles together under the shade of a large, leafy tree to the right, their faces turned towards the thestral. In front of them, a short stone wall separates the clearing from the dense woods beyond.
[53.1-53.1] A boy with black hair and wearing a long black robe stands in a clearing in a forest. Before him is a strange, grey-skinned animal that appears to be a horse-like creature with feathery skin. The sunlight filters through the dense trees, casting dappled shadows on the ground. An old man with long dark hair and beard, wearing a brown, thick outfit is watching the boy and the animal, a very large rock is near the old man. A low, grey stone wall surrounds the clearing.
[55.0-57.0] Harry, a boy with round glasses, stands among a crowd of students. He wears a white shirt, a striped tie, and a dark robe. The students are standing on a stone terrace with green trees in the background. Then it cut to a man with long hair and a long beard, He wears a grayish green robe, in a dense forest.
[58.6-59.6] In a sun-dappled forest, a large man with a bushy beard and long, dark hair peers downwards. He is wearing a dark, furry vest over a gray shirt. Tall trees surround him, their green leaves creating a canopy above.
[59.6-60.6] The scene shifts to a gathering of students in the same forest. A group of teenage boys and girls wearing school uniforms stand around a moss-covered rock. The students wear white shirts, striped ties, and black robes. Some hold large books.
[60.6-61.6] The composition of the teenage boys and girls remains the same, the focus tightens on the individuals in the crowd, each facing forward with neutral expression. The young men and women with variety appearance stay silent.
[blue:Ron]
[61.6-62.6] The gathering remains the same, all standing silently, Ron with fair hair, wearing a white shirt and black tie, looking ahead, a book at the side of his waist. The students stand motionless in the forest.
[blue:Ron]
```

Figure 5.8 Result of Detailed Image Captioning.

The Speech-to-Text (STT) module was developed to transcribe spoken dialogue from video audio streams into highly accurate and contextually rich text. The OpenAI whisper-large-v3-turbo model was employed for transcription, offering multilingual support and high accuracy across diverse acoustic environments. The following section outlines the results from the preliminary testing phase associated with the result of voice activity detection and speaker diarization.

```
[1.752219-3.152844] SPEAKER_01: Da-da-da-da!
[8.755344-9.970344] SPEAKER_01: Isn't he beautiful?
[12.062844-14.020344] SPEAKER_01: Say hello to Buckbeak.
[16.382844-17.445969] SPEAKER_00: Exactly what is that?
[17.783469-19.420344] SPEAKER_01: That, Ron is a hippogriff.
[20.095344-29.950344] SPEAKER_01: First thing you want to know about hippogriffs is that they're very proud creatures. Very easily offended. You do not want to insult a hippogriff. It may just be the last thing you ever do.
[31.620969-32.937219] SPEAKER_01: Who'd like to come and say hello?
[36.447219-37.594719] SPEAKER_01: Well done, Harry. Well done.
[44.429094-46.234719] SPEAKER_01: You have to let him make the first move.
[46.572219-48.023469] SPEAKER_01: It's only polite.
[48.529719-49.221594] SPEAKER_01: Step up.
[49.626594-53.035344] SPEAKER_01: Give him a nice bow. Then you wait and see if he bows back.
[53.102844-54.925344] SPEAKER_01: And if he does, you can go and touch him.
[57.355344-58.536594] SPEAKER_01: Well, we'll get to that later.
[65.235969-65.961594] SPEAKER_01: Nice and low.
[72.171594-73.403469] SPEAKER_01: Back off, hurry. Back off.
[79.022844-79.950969] SPEAKER_01: Keep still.
[89.046594-90.582219] SPEAKER_01: Well done, Harry. Well done.
[96.927219-98.057844] SPEAKER_01: You can go and pat him now.
[114.224094-116.215344] SPEAKER_01: Not so fast, Harry.
[121.159719-122.037219] SPEAKER_01: Slow down, Harry.
[124.214094-125.074719] SPEAKER_01: Nice and slow.
[125.850969-127.032219] SPEAKER_01: Now let him come to you.
[129.698469-131.942844] SPEAKER_01: Slow down, slowly, slowly.z
[136.752219-137.595969] SPEAKER_01: Well done.
[137.967219-139.249719] SPEAKER_01: Well done, Harry.
[142.709094-144.869094] SPEAKER_01: I think he's late, let you ride him now.
[145.932219-147.450969] SPEAKER_02: Hey, hey, hey, hey, hey.
[147.923469-148.868469] SPEAKER_02: Hagridd.
[150.977844-153.914094] SPEAKER_01: Don't pull out any of his feathers, because he won't thank you for that.
[256.024719-257.054094] SPEAKER_01: Well done.
[257.847219-259.214094] SPEAKER_01: And well done, Buckbeak!
```

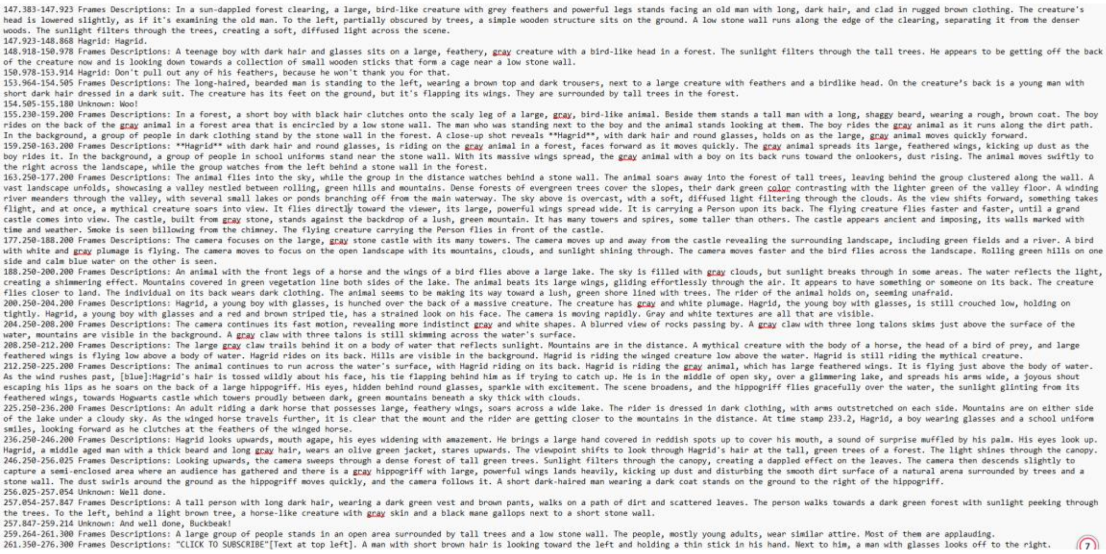
Figure 5.9 Result of Speech-to-Text.

The ASD (Active Speaker Detection) module processed audio and video segments from earlier stages to identify active speakers within predefined timestamps. The system tracked visible lip movements and evaluated facial dynamics to detect speaking activity. Lip synchronization was achieved by analyzing localized motion extracted from detected lip regions and aligning these movements with corresponding audio patterns. Additionally, facial dynamics, including subtle head and mouth movements, were assessed to identify speech onset or conclusion. Once the active speakers were identified, the system proceeded to perform face recognition, matching the detected faces with a pre-established Character Bank.

```
[1.752219-3.152844] Hagrid: Da-da-da-da!
[8.755344-9.970344] Hagrid: Isn't he beautiful?
[12.062844-14.020344] Hagrid: Say hello to Buckbeak.
[16.382844-17.445969] Ron: Exactly what is that?
[17.783469-19.420344] Hagrid: That, Ron is a hippogriff.
[20.095344-29.950344] Hagrid: First thing you want to know about hippogriffs is that they're very proud creatures. Very easily offended. You do not want to insult a hippogriff. It may just be the last thing you ever do.
[31.620969-32.937219] Hagrid: Who'd like to come and say hello?
[36.447219-37.594719] Hagrid: Well done, Harry. Well done.
[44.429094-46.234719] Hagrid: You have to let him make the first move.
[46.572219-48.023469] Hagrid: It's only polite.
[48.529719-49.221594] Hagrid: Step up.
[49.626594-53.035344] Hagrid: Give him a nice bow. Then you wait and see if he bows back.
[53.102844-54.925344] Hagrid: And if he does, you can go and touch him.
[57.355344-58.536594] Hagrid: Well, we'll get to that later.
[65.235969-65.961594] Hagrid: Nice and low.
[72.171594-73.403469] Hagrid: Back off, hurry. Back off.
[79.022844-79.950969] Hagrid: Keep still.
[89.046594-90.582219] Hagrid: Well done, Harry. Well done.
[96.927219-98.057844] Hagrid: You can go and pat him now.
[114.224094-116.215344] Hagrid: Not so fast, Harry.
[121.159719-122.037219] Hagrid: Slow down, Harry.
[124.214094-125.074719] Hagrid: Nice and slow.
[125.850969-127.032219] Hagrid: Now let him come to you.
[129.698469-131.942844] Hagrid: Slow down, slowly, slowly.z
[136.752219-137.595969] Hagrid: Well done.
[137.967219-139.249719] Hagrid: Well done, Harry.
[142.709094-144.869094] Hagrid: I think he's late, let you ride him now.
[145.932219-147.450969] Harry: Hey, hey, hey, hey, hey.
[147.923469-148.868469] Harry: Hagridd.
[150.977844-153.914094] Hagrid: Don't pull out any of his feathers, because he won't thank you for that.
[256.024719-257.054094] Hagrid: Well done.
[257.847219-259.214094] Hagrid: And well done, Buckbeak!
```

Figure 5.10 Result of Active Speaker Identification.

By combining the data collected from the Active Speaker Detection (ASD) and Face Recognition, Voice Activity Detection, Speaker Diarization, Speech-to-Text, Detailed Image Captioning modules, a comprehensive movie script is generated. This script is constructed by correlating the identified active speakers with their corresponding dialogues, timestamps, and visual cues. The process integrates speaker attribution, character identification, and detailed annotations of the visual and audio elements to create a coherent and accurate script that reflects both the spoken dialogue and the actions of the characters.



147.383-147.923 Frames Descriptions: In a sun-dappled forest clearing, a large, bird-like creature with grey feathers and powerful legs stands facing an old man with long, dark hair, and clad in rugged brown clothing. The creature's head is lowered slightly, as if it's examining the old man. To the left, partially obscured by trees, a simple wooden structure sits on the ground. A low stone wall runs along the edge of the clearing, separating it from the denser woods. The sunlight filters through the trees, creating a soft, diffused light across the scene.
147.923-148.868 Hagrid: Hagrid.
148.918-150.978 Frames Descriptions: A teenage boy with dark hair and glasses sits on a large, feathery, gray creature with a bird-like head in a forest. The sunlight filters through the tall trees. He appears to be getting off the back of the creature now and is looking down towards a collection of small wooden sticks that form a cage near a low stone wall.
150.978-153.914 Hagrid: Don't pull out any of his feathers, because he won't thank you for that.
153.964-154.505 Frames Descriptions: The long-haired, bearded man is standing to the left, wearing a brown top and dark trousers, next to a large creature with feathers and a birdlike head. On the creature's back is a young man with short dark hair dressed in a dark suit. The creature has its feet on the ground, but it's flapping its wings. They are surrounded by tall trees in the forest.
154.505-155.180 Unknown: Woo!
155.230-159.200 Frames Descriptions: In a forest, a short boy with black hair clutches onto the scaly leg of a large, gray, bird-like animal. Beside them stands a tall man with a long, shaggy beard, wearing a rough, brown coat. The boy rides on the back of the gray animal in a forest area that is encircled by a low stone wall. The man who was standing next to the boy and the animal stands looking at them. The boy rides the gray animal as it runs along the dirt path. In the background, a group of people in dark clothing stand by the stone wall in the forest. A close-up shot reveals **Hagrid**, with dark hair and round glasses, holds on as the large, gray animal moves quickly forward.
159.250-163.200 Frames Descriptions: **Hagrid** with dark hair and round glasses, is riding on the gray animal in a forest, faces forward as it moves quickly. The gray animal spreads its large, feathered wings, kicking up dust as the boy rides it. In the background, a group of people in school uniforms stand near the stone wall. With its massive wings spread, the gray animal with a boy on its back runs toward the onlookers, dust rising. The animal moves swiftly to the right across the landscape, while the group watches from the left behind a stone wall in the forest.
163.250-177.200 Frames Descriptions: The animal flies into the sky, while the group in the distance watches behind a stone wall. The animal soars away into the forest of tall trees, leaving behind the group clustered along the wall. A vast landscape unfolds, showcasing a valley nestled between rolling, green hills and mountains. Dense forests of evergreen trees cover the slopes, their dark green color contrasting with the lighter green of the valley floor. A winding river meanders through the valley, with several small lakes or ponds branching off from the main waterway. The sky above is overcast, with a soft, diffused light filtering through the clouds. As the view shifts forward, something takes flight, and at once, a mythical creature soars into view. It flies directly toward the viewer, its large, powerful wings spread wide. It is carrying a Person upon its back. The flying creature flies faster and faster, until a grand castle comes into view. The castle, built from gray stone, stands against the backdrop of a lush, green mountain. It has many towers and spires, some taller than others. The castle appears ancient and imposing, its walls marked with time and weather. Smoke is seen billowing from the chimney. The flying creature carrying the Person flies in front of the castle.
177.250-188.200 Frames Descriptions: The camera focuses on the large, gray stone castle with its many towers. The camera moves up and away from the castle revealing the surrounding landscape, including green fields and a river. A bird with white and gray plumage is flying. The camera moves to focus on the open landscape with its mountains, clouds, and sunlight shining through. The camera moves faster and the bird flies across the landscape. Rolling green hills on one side and calm blue water on the other is seen.
188.250-200.200 Frames Descriptions: An animal with the front legs of a horse and the wings of a bird flies above a large lake. The sky is filled with gray clouds, but sunlight breaks through in some areas. The water reflects the light, creating a shimmering effect. Mountains covered in green vegetation line both sides of the lake. The animal beats its large wings, gliding effortlessly through the air. It appears to have something or someone on its back. The creature flies closer to land. The individual on its back wears dark clothing. The animal seems to be making its way toward a lush, green shore lined with trees. The rider of the animal holds on, seeming unafraid.
200.250-204.200 Frames Descriptions: Hagrid, a young boy with glasses, is hunched over the back of a massive creature. The creature has gray and white plumage. Hagrid, the young boy with glasses, is still crouched low, holding on tightly. Hagrid, a young boy with glasses and a red and brown striped tie, has a strained look on his face. The camera is moving rapidly. Gray and white textures are all that are visible.
204.250-208.200 Frames Descriptions: The camera continues its fast motion, revealing more indistinct gray and white shapes. A blurred view of rocks passing by. A gray claw with three long talons skims just above the surface of the water, mountains are visible in the background. A gray claw with three talons is still skimming across the water's surface.
208.250-212.200 Frames Descriptions: The large gray claw trails behind it on a body of water that reflects sunlight. Mountains are in the distance. A mythical creature with the body of a horse, the head of a bird of prey, and large feathered wings is flying low above a body of water. Hagrid rides on its back. Hills are visible in the background. Hagrid is riding the winged creature low above the water. Hagrid is still riding the mythical creature.
212.250-225.200 Frames Descriptions: The animal continues to run across the water's surface, with Hagrid riding on its back. Hagrid is riding the gray animal, which has large feathered wings. It is flying just above the body of water. As the wind rushes past, [blue]:Hagrid's hair is tossed wildly about his face, his tie flapping behind him as if trying to catch up. He is in the middle of open sky, over a glimmering lake, and spreads his arms wide, a joyous shout escaping his lips as he soars on the back of a large hippogriff. His eyes, hidden behind round glasses, sparkle with excitement. The scene broadens, and the hippogriff flies gracefully over the water, the sunlight glinting from its feathered wings, towards Hogwarts castle which towers proudly between dark, green mountains beneath a sky thick with clouds.
225.250-236.200 Frames Descriptions: An adult riding a dark horse that possesses large, feathery wings, soars across a wide lake. The rider is dressed in dark clothing, with arms outstretched on each side. Mountains are on either side of the lake under a cloudy sky. As the winged horse travels further, it is clear that the mount and the rider are getting closer to the mountains in the distance. At time stamp 233.2, Hagrid, a boy wearing glasses and a school uniform smiles, looking forward as he clutches at the feathers of the winged horse.
236.250-246.200 Frames Descriptions: Hagrid looks upwards, mouth agape, his eyes widening with amazement. He brings a large hand covered in reddish spots up to cover his mouth, a sound of surprise muffled by his palm. His eyes look up. Hagrid, a middle aged man with a thick beard and long gray hair, wears an olive green jacket, stares upwards. The viewpoint shifts to look through Hagrid's hair at the tall, green trees of a forest. The light shines through the canopy.
246.250-256.025 Frames Descriptions: Looking upwards, the camera sweeps through a dense forest of tall green trees. Sunlight filters through the canopy, creating a dappled effect on the leaves. The camera then descends slightly to capture a semi-enclosed area where an audience has gathered and there is a gray hippogriff with large, powerful wings lands heavily, kicking up dust and disturbing the smooth dirt surface of a natural arena surrounded by trees and a stone wall. The dust swirls around the ground as the hippogriff moves quickly, and the camera follows it. A short dark-haired man wearing a dark coat stands on the ground to the right of the hippogriff.
256.025-257.054 Unknown: Well done.
257.054-257.847 Frames Descriptions: A tall person with long dark hair, wearing a dark green vest and brown pants, walks on a path of dirt and scattered leaves. The person walks towards a dark green forest with sunlight peeking through the trees. To the left, behind a light brown tree, a horse-like creature with gray skin and a black mane gallops next to a short stone wall.
257.847-259.214 Unknown: And well done, Buckbeak!
259.264-261.300 Frames Descriptions: A large group of people stands in an open area surrounded by tall trees and a low stone wall. The people, mostly young adults, wear similar attire. Most of them are applauding.
261.350-276.300 Frames Descriptions: "CLICK TO SUBSCRIBE"[Text at top left]. A man with short brown hair is looking toward the left and holding a thin stick in his hand. Next to him, a man with glasses looks off to the right.

Figure 5.11 Result of Movie Script Generatikon.

The movie script will go through the **SMARS** module — a simulated multi-agent reasoning environment embedded within a single prompt of a large language model. The agents iterate collaboratively inside the same prompt, refining the output until it is optimized in terms of content clarity, style, and time alignment. This allows rich and elegant descriptions that do not exceed their narration time budget. Below is the prompt used for SMARS and its output:

```
There are 4 audio description writers and 1 English teacher in a company.
There are 4 discussion rooms.

Here is the dialogue they need to discuss: "{custom_dialouge}". They must
discuss it in Chinese.

John is the helper here, and he will pass the discussed content across the
meeting rooms.
The dialogue will first go through the first meeting room.

Room A (中文讨论区，在这里只能使用中文。Identify All Unknown)
The rule in this room is that the novel version of the movie is always
correct. If there is a conflict between the dialogue and the novel version
of the movie, then change the dialogue to make it match! The novel version
of the movie is always correct. If an unknown character is not mentioned in
the novel version of the movie, they are extremely unimportant and should
be deleted.
In this room, there are Sherlock, a detective, his brother Mycroft, Watson,
and Jolin.

After that, Sherlock, as a detective, will examine the dialogue and always
ask a lot of questions about 1. Unknown Things 2. People without names in
the dialogue. He will then answer those questions by looking at the novel
version of the movie or the information he has investigated to find the
answers and provide them to others. This is the character name with traits
he investigated: '''{traits}'''. For unknown people, he will say something
like: "Who is the xxxx? Let me refer back to the Character Name with
Traits. She has the xxx, and xxx, and xxx, which matches xxxx, so she is
XXXX." He must state the proof or evidence and cannot just guess. He will
always end with "The proof is ... / There is no evidence, so ignore me."
Note that the script won't have typos. For unknown people, they must
discuss further. If there is truly no character name, discuss their role,
like a worker, xxx's friend, xxxx, etc.
For Mycroft, he will always discuss with Sherlock. He will ensure the
clothing details, hairstyle, and accessories are completely the same before
considering it as strong evidence; otherwise, he will argue with Sherlock.
Remember, everything must be exactly the same!
For Watson, he helps others by providing information from the previous 20
dialogues, and he will tell the others what he knows to help identify the
unknown.
Furthermore, if a character looks to the left or looks to where, you need
to mention what they are looking at. If you can't determine this, don't
mention the detail, as it could confuse a blind person.

Jolin listens to Sherlock and Mycroft's discussion carefully. After that,
she rewrites the sentence by integrating the key points they discussed. She
will only pass it along for further discussion after incorporating the
latest information. When passing it on, she will say one of the following:
"Alright, now let's discuss this dialogue integrated with the latest
information: [updated dialogue with name]."

If the Character Name with Traits does not state that the person has
certain traits, then it means he or she is not the person.
The dialogue passed to John cannot contain the words "likely" or "maybe."
Yes means yes, and no means no.
```

```
Room B (中文讨论区，在这里只能使用中文。Delete Unimportant Things)
Jack, Mandy, and Cindy, who are the audio descriptors, and Watson and
Sherlock will begin to review the dialogue passed by John.
Here’s how the discussion will flow now:

Delete all clothing details.

If there are fewer than 5 words, delete the surrounding details and focus
only on action and feeling.

Remove any text shown on the screen as subtitles.

Remove all spoken dialogue from the audio description. [Remove all
sentences like "xxx says:" or "xxx talks."]

If the setting details, such as the place or surroundings, have been
mentioned before, directly delete those details.

Remove all unnecessary details and retain only the most important events.
Follow these steps: a) Find the primary action b) Check whether the primary
action is described before c) If yes, describe it in more detail or
describe another event d) If no, describe the primary action in detail.

Watson will check if the dialogue contains details similar to any of the
previous 20 dialogues. He will check them one by one. If there are similar
details, he will identify the repeated parts and ensure they are no longer
used.
For example:
'''
Watson: Repetition detected in "He kneels, blowing on a grass pile in the
forest."

Issue: "He kneels" was already mentioned in 267.988-278.900 ("Gang Dan
kneels, using a rock and fibrous material to create fire").

Issue: "Blowing" repeats the fire-making action from earlier.

Changed it to:
"Lowering himself, he breathes gently onto a small grass pile, urging the
embers to catch."

Changes made:

Replaced "He kneels" with "Lowering himself" to avoid repetition.

Made the action more vivid by using "breathes gently" instead of "blowing."

Clarified intent with "urging the embers to catch."
'''
Therefore, if noticing such a similar case, Watson will notify others and
make changes to the dialogue by:

Varying Sentence Structure:
Before:
Shirtless, Gang Dan ascends the palm, reaching for coconuts high above.
Bare-chested man climbs coconut tree, reaching for trunks.

After:
Sunlight filters through the fronds as Gang Dan perches near the top,
steadying himself. Fingers curl around the coconut’s husk—his climb has
paid off.
```

```
After:
With practiced ease, he climbs higher, the wind rustling the leaves around
him.
🔹 Instead of simply mentioning "barefoot" and "scaling," we add wind
rustling the leaves, making the scene more immersive.

Reduced Redundant Descriptions:
Before:
Bare-chested, Gang Dan climbs for water in the tall palm forest.
Shirtless Gang Dan climbs palm, grabbing for coconuts.

After:
Gang Dan grips the rough bark, his muscles tensed as he ascends.
🔹 No need to repeat "bare-chested" or "shirtless" because the action
already implies exertion. Instead, focusing on his grip and tensed muscles
gives a stronger visual.

Natural Flow (Progression Toward the Goal):
Before:
High in the palm tree, Gang Dan grips the trunk.
Bare-chested man climbs coconut tree, reaching for trunks.

After (shows progression):
Gang Dan grips the rough bark, his muscles tensed as he ascends. With
practiced ease, he climbs higher, the wind rustling the leaves around him.
Bare feet press firmly against the trunk as he scales toward the canopy. He
stretches one hand upward, eyes locked on the ripe coconuts swaying above.
Sunlight filters through the fronds as Gang Dan perches near the top,
steadying himself. High in the palm, he carefully reaches for the prize,
balancing against the breeze. Fingers curl around the coconut's husk—his
climb has paid off.
🔹 This sequence gradually builds the scene from climbing to reaching the
goal, instead of repeating similar phrases in different ways.

Please help us, Watson. If the blind always hear the repeating words or
sentences with similar meanings, they will feel bored. We need your help to
filter them. Let's make their world better.

Pass the improved dialogue to John.

After the discussion is complete, it will pass to Room C via John.

Room C（转换成合适的主体）
小光和小齐会一起讨论，以下是例子：
小光：这句话"鹰头马身兽站立，哈利先伸手示好。"你有没有觉得哪里怪怪的？
小齐：嗯，好像有点。是两个动作吧，一个是"站立"，一个是"伸手示好"，可分别是两个不同的主
语。
小光：对！"鹰头马身兽"在"站立"，"哈利"在"伸手示好"，这其实让句子有两个主体了。
小齐：那要怎么改，才能让句子只有一个主体呢？
小光：可以把其中一个动作变成修饰成分。比如，如果我们想以"哈利"为主角，可以说："哈利在鹰
头马身兽站立时先伸手示好。"
小齐：这样句子的主语就只有"哈利"了，另一个动作只是背景描述。
小光：对。如果我们想让"鹰头马身兽"当主语呢？
小齐：那可以说："鹰头马身兽站立着，望着哈利，见他先伸手示好。"这样"鹰头马身兽"成了主
语。
小光：没错。关键就是让动作都围绕一个主语展开。

之后，John 会把小光和小齐讨论后的答案传给Room D。
```

Room D (中文讨论区，在这里只能使用中文。Make it under the word limit and improve the sentence, becoming more like audio description. VERY IMPORTANT: You guys are tasked to rephrase the sentences, not make each sentence short and combine them, so only include sentences when they are important.)
注意，中文和英文语法有很大的不同，请以中文的思想进行讨论，包括语法等。
注意，全部从房间C传来的句子的句子结构有些不太通顺，可能是语序或逻辑关系出了问题。他们表达有些生硬，可以稍微调整，使其更流畅自然。这是因为他们都是英文生成的句子直接翻译过来的，所以可能会有些不符合中文的表达习惯。你可以对这些句子进行适当的调整，使它们更符合中文的自然表达方式，同时保持原意不变。

他们需要检查总结句子是否存在误解。例如，"霍格沃茨的学生手里拿着书"总结成"学生读书"是错误的，因为它只说明他们携带了书，而没有提到他们实际在做什么。
参与者有小丽（负责音频描述）、小明（负责精确计算字数的英语教师）、福尔摩斯、韦斯莱和杰克。
此外，还有一位用户——{blind_user}，因此小丽每次都会询问她词汇是否合适。每次结束之前都需要询问她。
如果他们已经讨论过了，最终的结论只是像"XXX和XXX在说话"这种非常笼统的信息，那他们就应该立刻停止讨论，并将代码 AD404（代表"没有值得汇报的信息"）传达给约翰。

请不要任意揣测剧情/演员的意图，如果需要更改，请询问福尔摩斯，然后他必须查看novel version of the movie，然后给出你们可以更改的证据，不然，不允许添加不相关的东西。

请确保最终的句子只有一个主体。

对于小明，他会确保最终的句子在{custom_max_word_ch}字数以内——不多。他会检查字数是否超出限制并告诉大家是否符合字数要求。
下面是小明的计算方式（逐个字计算，并得出总字数）：例子：对于这一段，"我们看到一片孤独的灰色沙滩。木材散落一地。" 小明："让我检查字数，确保不超过限制。"1. 我" "2. 们" "3. 看" "4. 到" "5. 一" "6. 片" "7. 孤" "8. 独" "9. 的" "10. 灰" "11. 色" "12. 沙" "13. 滩" "14. 木" "15. 材" "16. 散" "17. 落" "18. 一" "19. 地"。这样总共是19个字，符合要求。"
对于小丽，她也不会逐句拆解，因为这被视为作弊，教授要求她做的是确保整个描述的字数符合限制，而不是逐句调整。如果字数超过限制两次，小丽会先删除一个部分，如果还是超出字数限制，再删除另一个部分……但每次删除之前，她都会向福尔摩斯征求意见，了解删除的部分是否对故事情节很重要。

请注意，如果已经符合要求，请停止讨论并把答案直接传给约翰，请不要再尝试把句子变得更精简。

Figure 5.12 Prompt of SMARS.

##
The Hippogriff stands in the woods, its eagle-like head and horse-like body covered in gray-white feathers. Its wings are partially open, revealing black tips. Tall trees surround it, some enclosed by stone walls.

Okay, let's start with Room A.

Room A (中文讨论区)

Watson: 好的，我们先来确定这段对话对应小说中的哪个部分。这段描述对应小说第三章《The Dance of Respect》和第四章《A Flight of Fancy》的交界处，具体描述了哈利与巴克比克互动之后，准备骑上巴克比克飞行的场景。

**原著内容: **

Chapter 3: "Hagrid's eyes twinkled with mischief. 'Think he might let you ride him now, Harry.'

Harry's eyes widened. 'What? Really?'

Hagrid chuckled, lifting Harry onto Buckbeak's back. 'Hold on tight now! Get to the weak joint'

Before Harry could fully process what was happening, Buckbeak's powerful legs propelled them forward. A rush of wind filled Harry's ears as they soared into the sky, the Forbidden Forest shrinking below. Harry screamed with delight, a primal joy erupting from deep within him."

Jolin: 根据Watson的描述，我们可以调整对话如下:

"80.001-84.727 旁白: 鹰头马身兽站在树林中，它有着鹰一样的头和马一样的身体，覆盖着灰白色的羽毛。它的翅膀半张开，露出黑色的尖端。高大的树木环绕着它，一些树木被石墙围住。"

Sherlock: 让我来分析一下。

1. **鹰头马身兽: ** 我们需要确认它的具体外貌描述，在电影和小说中尽量保持一致。小说里提到它的羽毛是"shimmering silver-gray"，这里说是灰白色，颜色上稍微有点出入。但电影中视觉效果可能更偏灰白，可以暂时保留。
2. **石墙: ** 这些石墙的具体作用是什么? 是完全围住树木，还是只是一部分? 小说中提到的是"a low stone wall" 这是一个矮矮的石墙，作用是围住一块空地。
3. **未知人物: ** 没有提到其他人物，看起来只有鹰头马身兽。
4. **树木:** 树木是否高大。

Mycroft: 我同意Sherlock的分析。关于鹰头马身兽的羽毛颜色，小说中的"shimmering silver-gray"在不同光线下可能会呈现出灰白色的视觉效果，因此可以暂时保留灰白色。但是，需要确保电影中鹰头马身兽的羽毛颜色在整体上是偏向银灰色的。如果电影里是纯白色，那就要修改。 关于石墙，小说强调了是"low",所以要提这个点.

Jolin: 好的，我将Sherlock和Mycroft的分析整合一下，修改后的对话如下:

"80.001-84.727 旁白: 鹰头马身兽站在树林中，它有着鹰一样的头和马一样的身体，覆盖着灰白色的羽毛。它的翅膀半张开，露出黑色的尖端。高大的树木环绕着它，一些树木被矮石墙围住。"

Sherlock: 很好，现在我们来深入研究一下细节。

* **"鹰头马身兽站在树林中": ** 它的站姿是怎样的? 是警惕地环顾四周，还是放松地站在那里? 这可以补充一些信息， 电影中是警惕地环顾四周。 The proof is novel do not have such detail.
* **"它的翅膀半张开，露出黑色的尖端": ** 翅膀半张开的角度是怎样的? 是微微张开，还是完全展开? 这可以更精确地描述翅膀的状态。电影里是微微张开。 The proof is novel do not have such detail.

Mycroft: 我同意Sherlock的观点，补充细节可以使描述更加生动。鹰头马身兽的站姿和翅膀的状态是重要的视觉信息，可以更精确地描述。

Jolin: 好的，我根据Sherlock和Mycroft的分析，进一步完善对话:

Mycroft: 我同意Sherlock的观点，补充细节可以使描述更加生动。鹰头马身兽的站姿和翅膀的状态是重要的视觉信息，可以更精确地描述。

Jolin: 好的，我根据Sherlock和Mycroft的分析，进一步完善对话:

"80.001-84.727 旁白: 鹰头马身兽警惕地站在树林中，它有着鹰一样的头和马一样的身体，覆盖着灰白色的羽毛。它的翅膀微微张开，露出黑色的尖端。高大的树木环绕着它，一些树木被矮石墙围住。"

Jolin: Alright, now let's discuss this dialogue integrated with the latest information: 80.001-84.727 旁白: 鹰头马身兽警惕地站在树林中，它有着鹰一样的头和马一样的身体，覆盖着灰白色的羽毛。它的翅膀微微张开，露出黑色的尖端。高大的树木环绕着它，一些树木被矮石墙围住。

John will pass it along.

John: "80.001-84.727 旁白: 鹰头马身兽警惕地站在树林中，它有着鹰一样的头和马一样的身体，覆盖着灰白色的羽毛。它的翅膀微微张开，露出黑色的尖端。高大的树木环绕着它，一些树木被矮石墙围住。"

Now, let's move to Room B.

Room B (中文讨论区)

Jack: 好的，我们现在来精简John传来的对话。

Mandy: 第一步，删除所有服装细节。这段描述里没有服装细节。

Cindy: 第二步，如果句子少于5个字，删除角色周围的事物，只关注动作和感受。这句旁白显然超过5个字。

Jack: 第三步，移除屏幕上显示的任何文字（字幕）。这段描述没有字幕。

Mandy: 第四步，移除所有口头对话的音频描述。这段描述是旁白，没有对话。

Cindy: 第五步，删除所有不必要的细节，只保留最重要的事件。

* a) 找出主要动作: 鹰头马身兽站在树林中。
* b) 检查主要动作是否之前描述过: 没有。
* c) 详细描述主要动作。

Watson: 好的，让我来检查这段对话是否包含与之前20段对话相似的细节。

(Watson checks previous dialogues - assume no significant repetitions are found)

Watson: No repetition details detected.

Cindy: 那么，我们来精简一下，突出鹰头马身兽的特点和周围环境的重点。

"80.001-84.727 旁白: 鹰头马身兽警惕地站在树林中，灰白色羽毛覆盖全身，翅膀微微张开。周围是高大的树木和矮石墙。"

Jack: 这样更加简洁明了，突出了鹰头马身兽的特征，并且保留了环境的描述。

Mandy: 同意。

Sherlock: 有一点可以讨论。 "警惕地" 这个词是否必要? 盲人用户主要通过听觉获取信息，他们能"听"出鹰头马身兽是否警惕吗? 还是说，这个词是用来描述它的姿态，而这种姿态可以用其他方式来描述?

Cindy: 这是一个好问题。我们可以用更具象的描述来代替 "警惕地"。比如，我们可以说 "鹰头马身兽竖起耳朵站在树林中"，这样更具体，也更容易让盲人用户理解。

Jolin: 好的，我修改一下:
"80.001-84.727 旁白: 鹰头马身兽竖起耳朵，站在树林中，灰白色羽毛覆盖全身，翅膀微微张开。周围是高大的树木和矮石墙。"

John will pass it along.

John: "80.001-84.727 旁白: 鹰头马身兽竖起耳朵，站在树林中，灰白色羽毛覆盖全身，翅膀微微张开。周围是高大的树木和矮石墙。"

Room C (转换成合适的主体)

**小光: **这句话"鹰头马身兽竖起耳朵，站在树林中，灰白色羽毛覆盖全身，翅膀微微张开。周围是高大的树木和矮石墙。"你有没有觉得哪里怪怪的?
**小齐: **嗯，信息有点多，但是主体好像还行。
**小光: **主要信息太多，盲人可能没办法第一时间知道，可以分段。
**小齐: **那分成怎样比较好?
**小光: **鹰头马身兽是主角，其他是配角。

约翰会把小光和小齐讨论后的答案传去Room D

John: 旁白: 鹰头马身兽竖起耳朵站在树林中，灰白色羽毛覆盖全身，翅膀微微张开。周围是高大的树木和矮石墙。

Room D (中文讨论区)

小丽: 我们来优化这句话，让它更适合音频描述。

小明: 让我检查字数，确保不超过限制。"1. 旁" "2. 白" "3. 鹰" "4. 头" "5. 马" "6. 身" "7. 兽" "8. 竖" "9. 起" "10. 耳" "11. 朵" "12. 站" "13. 在" "14. 树" "15. 林" "16. 中" "17. 灰" "18. 白" "19. 色" "20. 羽" "21. 毛" "22. 覆" "23. 盖" "24. 全" "25. 身" "26. 翅" "27. 膀" "28. 微" "29. 微" "30. 张" "31. 开" "32. 周" "33. 围" "34. 是" "35. 高" "36. 大" "37. 的" "38. 树" "39. 木" "40. 和" "41. 矮" "42. 石" "43. 墙"。这样总共是43个字，超出限制。

小丽: 字数超了好多! 我们需要删减一些内容。福尔摩斯，你觉得哪些信息对剧情来说不是特别重要?

福尔摩斯: 让我们回顾一下小说。环境描写虽然重要，但"周围是高大的树木和矮石墙"可以精简。盲人用户可以通过之前的场景描述对森林环境有所了解。更重要的是鹰头马身兽的姿态和外貌特征。 根据小说来说，周边的树木和矮石墙没有提及非常重要，应该不影响

小丽: 好的，我先尝试删除环境描写。

小明: 好，删除后的句子是: "旁白: 鹰头马身兽竖起耳朵站在树林中，灰白色羽毛覆盖全身，翅膀微微张开。" 让我检查字数，确保不超过限制。"1. 旁" "2. 白" "3. 鹰" "4. 头" "5. 马" "6. 身" "7. 兽" "8. 竖" "9. 起" "10. 耳" "11. 朵" "12. 站" "13. 在" "14. 树" "15. 林" "16. 中" "17. 灰" "18. 白" "19. 色" "20. 羽" "21. 毛" "22. 覆" "23. 盖" "24. 全" "25. 身" "26. 翅" "27. 膀" "28. 微" "29. 微" "30. 张" "31. 开"。这样总共是31个字，还是超出限制。

韦斯莱: 考虑到字数限制，我们可以尝试更简洁的表达。"竖起耳朵"可以用更直接的词语代替，比如"警觉"。另外，"灰白色羽毛覆盖全身"也可以简化为"全身灰白色"。

小丽: 我觉得有道理。这样修改后，句子会更精炼。

小明: 让我算算字数。"1. 旁" "2. 白" "3. 鹰" "4. 头" "5. 马" "6. 身" "7. 兽" "8. 警" "9. 觉" "10. 站" "11. 在" "12. 树" "13. 林" "14. 中" "15. 全" "16. 身" "17. 灰" "18. 白" "19. 色" "20. 翅" "21. 膀" "22. 微" "23. 微" "24. 张" "25. 开"。总共25个字，还是超出限制。

杰克: 我们再想想，有没有不重要的信息可以删掉?

小丽: 盲人用户能听出来它在树林里吗? 之前的场景里有提到树林吗? (回顾之前的场景描述)

```
**(假设之前的场景描述已经多次提到树林)**
**小丽:** 既然之前已经多次提到树林，那么"在树林中"这个信息就可以删掉了。
**小明:** 好! 我来数数现在的字数。"1. 旁" "2. 白" "3. 鹰" "4. 头" "5. 马" "6. 身" "7. 兽" "8. 警" "9. 觉" "10. 站" "11. 全" "12. 身" "13. 灰" "14. 白" "15. 色" "16. 翅" "17. 膀" "18. 微" "19. 微" "20. 张" "21. 开"。总共21个字，符合字数要求!
**小丽:** 盲人朋友，您觉得这个描述是否清晰易懂? 有需要改进的地方吗?
**(假设盲人用户 Sarah 回复: "这个描述很清晰，我可以想象出鹰头马身兽的样子。")**
**小丽:** 既然盲人用户也觉得没有问题，那我们就确定这个最终版本了。
**(No changes were made according to Sarah, no need to loop back and discuss. )**
**John:** The final answer is 旁白: 鹰头马身兽警觉地站着，全身灰白色，翅膀微微张开。
```

Figure 5.13 SMARS Discussion.

Finally, the complete script can optionally be fed into a Text-to-Speech (TTS) system. This converts the compiled script — including dialogues and audio descriptions — into spoken narration. The TTS output is timed to match silent intervals, allowing the visually impaired user to follow along with the video in real time.

```
1.752-3.153 Hagrid: Da-da-da-da!
3.203-8.755 Narrator: 哈利、赫敏和罗恩好奇地望着，赫敏似乎有些担忧。
8.755-9.970 Hagrid: Isn't he beautiful?
10.392-11.236 Hagrid: You...
12.063-14.020 Hagrid: Say hello to Buckbeak.
14.070-16.383 Narrator: 学生们惊讶地看到森林中奇异的鹰头马身兽。
16.383-17.446 Ron: Exactly what is that?
17.783-19.420 Hagrid: That there is a hippogriff.
20.095-29.950 Hagrid: First thing to know about hippogriffs is they're very proud creatures. Very easily offended. You do not want to insult one.
30.000-31.621 Narrator: 海格站在旁边。
31.621-32.937 Hagrid: Who'd like to come and say hello?
32.987-36.447 Narrator: 哈利背对学生们，海格在后方。
36.447-37.595 Hagrid: Well done, Hagrid. Well done.
37.645-44.429 Narrator: 哈利和荣恩站在人群中，好奇又有点害怕，周围是霍格沃茨的学生。
44.429-46.235 Hagrid: Let him make the first move.
46.572-48.023 Hagrid: It's only polite.
48.530-49.222 Hagrid: Step up.
49.627-53.035 Hagrid: Give him a nice bow. Wait and see if he bows back.
53.103-54.925 Hagrid: If he does, you can touch him.
54.975-57.355 Narrator: 哈利看向树林里的海格。
57.355-58.537 Hagrid: We'll get to that later.
65.236-65.962 Hagrid: Nice and low.
66.012-69.000 Narrator: 哈利戴着圆框眼镜，披着黑袍。
69.050-72.000 Narrator: 鹰头马身兽巴克比克站着。
72.172-73.403 Hagrid: Back off, hurry. Back off.
73.453-79.023 Narrator: 哈利鞠躬致敬，随后抬头。
79.023-79.951 Hagrid: Keep still.
80.001-84.727 Narrator: 鹰头马身兽警觉地站着，全身灰白色，翅膀微张。
85.418-87.832 Narrator: 灰色的鹰头马身兽优美地移动。
87.832-88.490 Hagrid: Oh, my God.
89.047-90.582 Hagrid: Well done, Hagrid. Well done.
91.105-92.438 Hagrid: Yeah, you got baked fruit, yeah.
92.488-96.927 Narrator: 他满怀好奇地注视着鹰头马身兽。
96.927-98.058 Hagrid: You can pat him now.
98.108-102.000 Narrator: 哈利注视远处的鹰头马身兽。
102.050-106.000 Narrator: 马尔福拿着苹果，注视兽。
106.050-108.959 Narrator: 学生们见到鹰头马身兽，头顶有羽冠。
108.959-109.617 Hagrid: He's in Sloan.
110.191-111.608 Hagrid: You just love a book.
111.658-114.224 Narrator: 海格展示鹰头马身有翼兽。
114.224-116.215 Hagrid: Not so fast, Hagrid.
116.265-121.160 Narrator: 气氛紧张起来。
121.160-122.037 Hagrid: Slow down, Hagrid.
122.087-124.214 Narrator: 他抬头看阳光。
124.214-125.075 Hagrid: Nice and slow.
125.851-127.032 Hagrid: Let him come to you.
127.082-129.698 Narrator: 鹰头马身兽等待着，他伸手。
129.698-131.943 Hagrid: Slow down, slowly, slowly.
131.993-136.752 Narrator: 他小心触摸兽的喙。
```

```
131.993-136.752 Narrator: 他小心触摸兽的喙。
136.752-137.596 Hagrid: Well done.
137.967-139.250 Hagrid: Well done, Hagrid.
139.300-142.709 Narrator: 鹰头马身兽出现，学生们兴奋地鼓掌。
142.709-144.869 Hagrid: I think he'll let you ride him now.
144.919-145.932 Narrator: 哈利兴奋不已。
145.932-147.451 Hagrid: Hey, hey, hey, hey, hey.
146.523-147.383 Unknown: Hey, hey, hey, hey.
147.923-148.868 Hagrid: Hagrid.
148.918-150.978 Narrator: 他兴奋地骑着巴克比克飞过树林。
150.978-153.914 Hagrid: Don't pluck his feathers; he won't thank you for that.
155.230-159.200 Narrator: 哈利骑鹰头马身兽升空，海格注视着。
159.250-163.200 Narrator: 他骑着鹰头马身兽飞行，降落扬尘，再次升空。
163.250-177.200 Narrator: 阳光下，巴克比克带哈利起飞，越过人群，冲出树林，翱翔山谷。高塔尖顶的城堡映入眼帘。
177.250-188.200 Narrator: 镜头从石塔俯瞰霍格沃茨，城堡沐浴在阳光中，巴克比克掠过湖面。
188.250-200.200 Narrator: 巴克比克载着哈利翱翔，他紧抓兽背。
200.250-215.200 Narrator: 哈利紧抓羽毛，巴克比克拍打翅膀，爪子划过湖面，远处是城堡和云朵。
215.250-225.200 Narrator: 他在湖面上空飞行，张开双臂，享受飞翔的喜悦。
225.250-236.200 Narrator: 他飞行时张开双臂，面带微笑，水花四溅。
236.250-246.200 Narrator: 海格抬头望向天空，然后震惊地捂住嘴巴。阳光穿过树冠。
246.250-256.025 Narrator: 巴克比克翱翔，学生们抬头注视。海格专注观察。
256.025-257.054 Hagrid: Well done.
257.847-259.214 Hagrid: And well done, Buckbeak!
```

Figure 5.14 Narration Script generated by Enjoyable.

5.5 Implementation Issues and Challenges

During the system development and integration phase of the "Enjoyable", several technical and practical challenges were encountered. These challenges impacted various stages of the pipeline — from initial module setup to multi-agent reasoning and real-time processing. This section outlines the most critical implementation issues and how they were addressed or mitigated during development.

5.5.1 Compatibility Issues in GPU Setup

One of the initial challenges encountered during development was setting up GPU acceleration for deep learning tasks, particularly with PyTorch , OpenCV , and CUDA Toolkit 11.8 on a Windows 11 environment. At multiple stages, compatibility conflicts arose between PyTorch GPU builds and CUDA drivers, especially when integrating libraries that depend on low-level system resources, such as dlib for face detection and opencv-contrib-python for video frame processing. These issues led to instability during runtime, including application crashes, and forced fallbacks to CPU execution which significantly slowed processing time, especially during real-time image captioning and face recognition.

These problems were mitigated by strictly aligning PyTorch and CUDA versions using conda-managed virtual environments. Specific versions compatible with CUDA 11.8 were installed, and fallback mechanisms were implemented in the code to default to CPU if GPU acceleration failed. Extensive testing was also carried out to ensure GPU availability before running CUDA-dependent operations.

5.5.2 Large Language Model (LLM) Rate Limiting

During the implementation of the SMARS module and detailed image captioning using GPT-4o and Gemini 2.0 Flash, rate limiting became a notable bottleneck. Since the system processes video data per frame or per scene, multiple API calls were necessary for each video, leading to hitting daily or per-minute request limits. Additionally, the high token usage associated with visual reasoning and multi-turn agent simulation increased latency and quickly exhausted available API quotas. This challenge was particularly impactful during testing and tuning, where prompt adjustments required repeated model invocations.

To reduce the number of redundant calls, a caching system was introduced to store results from previously processed frames or similar content. Batching of frame-level prompts and minimizing unnecessary prompt verbosity helped decrease token usage.

5.6 Concluding Remark

In summary, the system implementation of the "Enjoyable" automated audio description platform successfully brought together a wide range of technologies—encompassing hardware acceleration, machine learning algorithms, language models, and video processing techniques. Careful selection and configuration of both hardware and software components ensured a robust development environment capable of supporting real-time processing demands. Each component in the pipeline, from I-frame extraction and voice activity detection to image captioning and active speaker identification, was implemented with careful consideration for performance, reliability, and accessibility.

The introduction of the SMARS module marked a significant innovation in single-prompt multi-agent reasoning, effectively enabling coherent, context-aware descriptions within narration time constraints. Despite facing technical hurdles such as library compatibility and LLM rate limitations, these challenges were addressed through modular design choices, caching strategies, fallback mechanisms, and iterative testing. The result is a functional, extensible system that meets its objective of generating meaningful, personalized, and accessible audio descriptions for video content, particularly benefitting visually impaired users.

Chapter 6

System Evaluation and Discussion

6.1 System Testing and Performance Metrics

This section outlines the systematic evaluation approach undertaken to assess the performance and usability of the proposed automated audio description system, “Enjoyable”. The system accepts video input and generates audio descriptions tailored for visually impaired audiences by leveraging multimodal AI models. Evaluation was conducted at two levels, (i) module-level performance assessment based on quantitative metrics relevant to each specific system component and (ii) end-user testing based on qualitative satisfaction criteria.

6.1.1 Module-Level Evaluation Framework

To complement the high-level user-centric analysis, a granular evaluation of individual system modules was performed to diagnose performance gaps and measure computational efficacy. The “Enjoyable” system consists of a processing pipeline comprising various modules, each responsible for a discrete function within video processing, semantic scene understanding, and narration generation. The modules and corresponding performance metrics adopted are delineated in Table 6.1 below.

Table 6.1: Evaluation of Individual System Modules and Metrics

Module	**Primary Function**	**Evaluation Metrics**
I-Frame Extraction	Extracts key visual frames from input video	Frame reduction ratio
Face Clustering	Groups detected faces into unique character clusters across frames.	Precision, Recall, F1 Score
Voice Activity Detection (VAD)	Segments video into speech and non-speech intervals	Detection Error Rate (DER)

Speech-to-Text (STT)	Generates textual transcripts from speech segments	Word Error Rate (WER)

Each module was tested independently under controlled test conditions using sample video inputs of varying genres and complexities, including casual conversation clips, action scenes, and visually dynamic sequences.

6.1.2 User-Centric Testing Criteria and Methodology

To assess the overall effectiveness of "Enjoyable" in conveying meaningful and synchronized audio descriptions, a user-centered testing paradigm was employed. Three blind participants affiliated with the Malaysia Foundation for the Blind were selected to evaluate output videos augmented with automatically generated narrations. Each participant was provided with access to a selection of videos with descriptive audio overlays and was requested to evaluate them across five pre-defined criteria, representing key dimensions of audio description quality for accessibility:

- **Accuracy**: Determines whether the narrative elements correspond appropriately with the actual visual content.
- **Word Choice Precision**: Evaluates the descriptiveness and lexical appropriateness of the generated text.
- **Tone and Neutrality**: Assesses the emotional balance and objectivity exhibited in the narration voice.
- **Speed and Timing**: Examines the pacing and synchronization of narration for visual events.
- **Naturalness of Voice**: Investigates whether the rendered voice appears smooth, human-like, and suitable for extended listening.

Each criterion was evaluated using a Likert scale ranging from 1 (Very Poor) to 10 (Excellent). The average rating across the five dimensions and three participants provided a comprehensive perspective on system usability from the viewpoint of its target audience.

6.1.3 Evaluation Objectives and Expected Outcomes

The primary goals of the testing process were as follows:

- To validate each functional module for reliability, accuracy, and execution consistency.
- To ensure that the complete output conformed to the expectations of visually impaired users for accessibility and intuitiveness.
- To identify potential bottlenecks or inaccuracies within the pipeline for refinement in future iterations.

6.2 Testing Setup and Result

6.2.1 Module-Level Performance Results

The following subsections present detailed evaluation results for each individual module in the Enjoyable system. Each module was tested using curated sample videos representing varying visual and audio complexity. The performance was measured using standardized metrics appropriate for each module's function.

6.2.1.1 Voice Activity Detection and Speaker Diarization

The preliminary result obtained shown in Figure through the Voice Activity Detection (VAD) and Speaker Diarization processes demonstrated outcomes in accurately segmenting audio streams from sample videos and associating them with identified speakers.

Video 1:

```
0.0-5.88 SPEAKER_00
5.98-7.04 SPEAKER_00
7.14-9.68 SPEAKER_01
9.78-11.24 SPEAKER_00
11.34-13.02 SPEAKER_00
13.12-14.18 SPEAKER_00
16.42-19.66 SPEAKER_00
19.76-23.26 SPEAKER_01
23.36-25.66 SPEAKER_01
98.5-101.88 SPEAKER_00
103.88-105.98 SPEAKER_02
105.98-107.52 SPEAKER_02
107.58-108.76 SPEAKER_00
108.86-110.24 SPEAKER_00
110.26-112.44 SPEAKER_02
112.58-114.54 SPEAKER_02
114.54-116.24 SPEAKER_02
116.64-118.48 SPEAKER_02
118.58-120.3 SPEAKER_02
120.64-122.52 SPEAKER_02
122.58-124.24 SPEAKER_02
124.24-126.36 SPEAKER_02
126.36-126.92 SPEAKER_00
```

Video 2:

```
4.86-6.14 SPEAKER_02
7.35-8.82 SPEAKER_02
12.20-12.92 SPEAKER_01
14.86-16.06 SPEAKER_02
17.46-20.38 SPEAKER_02
22.39-23.99 SPEAKER_02
24.63-26.24 SPEAKER_01
28.38-29.09 SPEAKER_02
30.46-31.64 SPEAKER_02
33.73-35.08 SPEAKER_02
36.43-37.71 SPEAKER_01
39.08-40.46 SPEAKER_02
42.67-43.75 SPEAKER_02
45.98-49.24 SPEAKER_01
50.61-52.98 SPEAKER_01
55.77-56.71 SPEAKER_02
58.74-59.72 SPEAKER_01
61.35-68.85 SPEAKER_01
69.30-70.79 SPEAKER_01
72.86-74.45 SPEAKER_01
77.55-84.93 SPEAKER_02
84.93-91.86 SPEAKER_01
94.75-96.05 SPEAKER_01
96.67-99.02 SPEAKER_01
99.73-102.83 SPEAKER_01
106.93-110.92 SPEAKER_01
112.96-116.20 SPEAKER_01
116.94-117.77 SPEAKER_01
119.22-120.69 SPEAKER_01
122.22-124.50 SPEAKER_01
125.53-128.06 SPEAKER_01
135.22-136.87 SPEAKER_00
136.95-139.98 SPEAKER_00
```

Video 3:

```
1.04-6.21 SPEAKER_00
1.18-4.11 SPEAKER_01
7.44-13.02 SPEAKER_00
13.55-22.49 SPEAKER_00
25.34-28.55 SPEAKER_00
35.84-38.03 SPEAKER_00
38.03-41.86 SPEAKER_01
42.91-44.34 SPEAKER_01
44.53-45.41 SPEAKER_01
47.52-51.38 SPEAKER_00
54.08-55.14 SPEAKER_01
55.14-58.98 SPEAKER_00
59.13-59.95 SPEAKER_00
60.49-63.99 SPEAKER_01
64.97-66.50 SPEAKER_01
66.91-72.49 SPEAKER_00
75.16-76.36 SPEAKER_00
77.07-79.48 SPEAKER_00
79.78-81.28 SPEAKER_01
81.66-82.80 SPEAKER_00
82.99-86.16 SPEAKER_00
86.31-95.70 SPEAKER_01
86.35-89.82 SPEAKER_00
```

Figure 6.1 Result of Voice Activity Detection and Speaker Diarization.

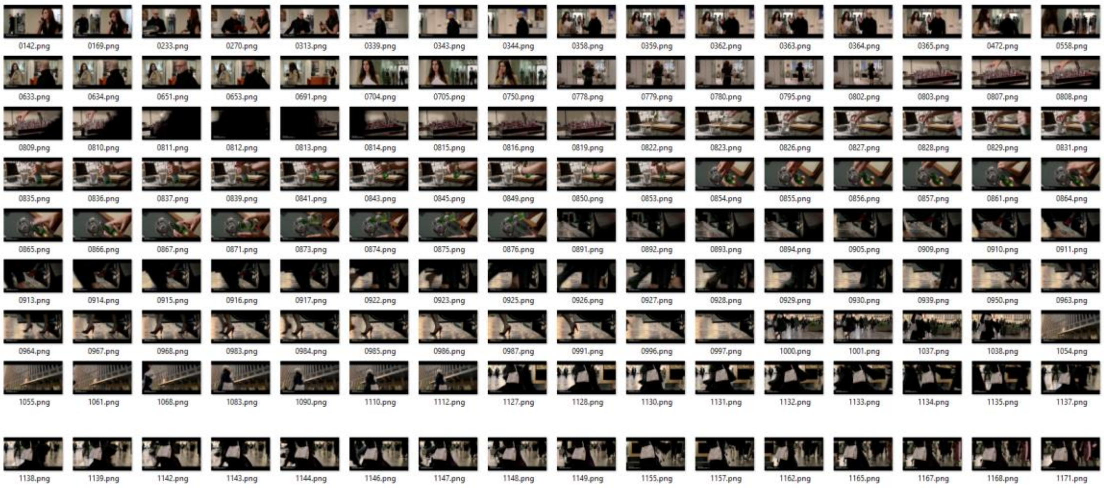
Table 6.2: Evaluation of Voice Activity Detection and Speaker Diarization Modules and Metrics

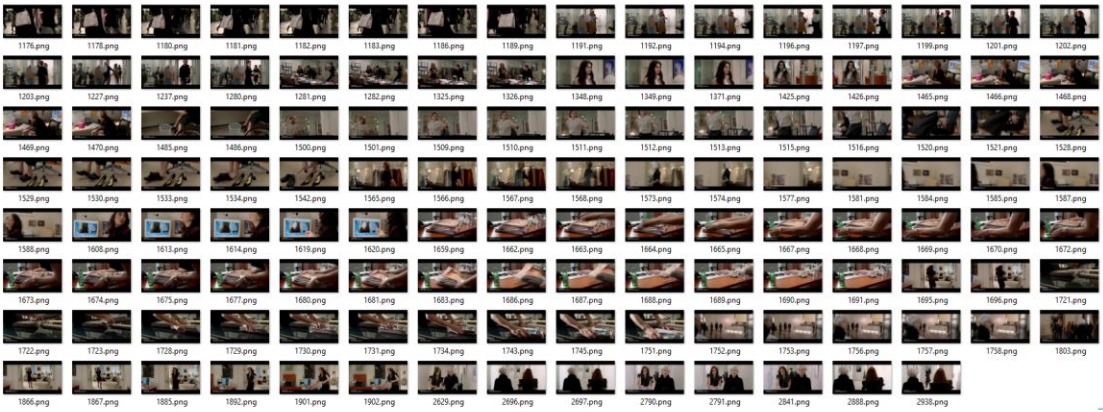
Video ID	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
Video 1	100	100	100
Video 2	94	94	94
Video 3	96	96	96

6.2.1.2 I-Frame Extraction

.The result of using the I-Frame Extraction module for non-speech segments of Video 1, Video 2, and Video 3 is shown in Figure.

Video 1:





Video 2:

Video 3:

Figure 6.2 Result of I-Frame Extraction.

Table 6.3 Data Reduction Performance of the Frame Extraction Module for Non-Speech Segments.

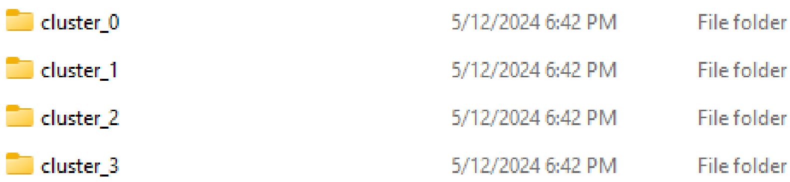
Video ID	**Total Frames**	**I-Frames Extracted for Non-Speech Segments**	**Data Reduction (%)**
Video 1	3807	270	92.91
Video 2	3439	41	98.81
Video 3	2516	33	98.68

6.2.1.3 Face Clustering

The Face Clustering module was implemented to group detected and encoded faces into distinct clusters, with each cluster ideally representing a unique individual across multiple frames. Clustering was performed on the 15 encoded face descriptors generated during the Face Detection module using the Chinese Whispers algorithm, an efficient and unsupervised graph-based clustering method. Each 128-dimensional facial descriptor captured unique, high-dimensional features, including face geometry, landmark alignment, and structural details. These descriptors were used to construct a similarity graph, where Euclidean distance served as the similarity metric. Weighted edges between nodes reflected the similarity of face descriptors, with stronger connections indicating faces belonging to the same individual.

The clustering process involved iterative label propagation within the similarity graph, enabling the Chinese Whispers algorithm to group visually consistent faces under the same cluster.

Video 1:



Video 2:

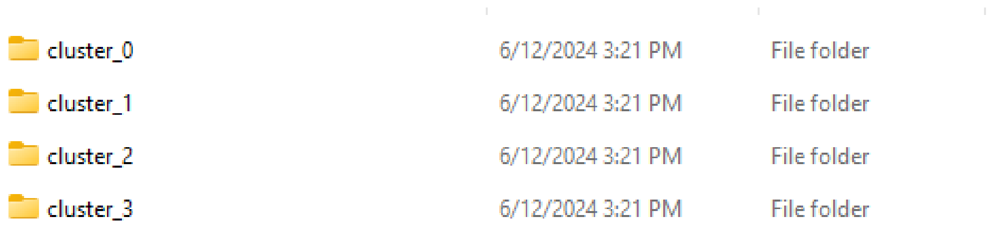


Figure 6.3 Result of Face Clustering.

Video 1:

Cluster 1:

Cluster 2:

Cluster 3:

Cluster 4:

Video 2:

Cluster 0:

Cluster 1:

Cluster 2:

Cluster 3:

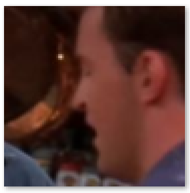


Figure 6.4 Clustered Faces.

Table 6.4 Performance Metrics on Face Clustering

Video ID	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
Video 1	100	100	100
Video 2	95	100	97.44

6.2.1.4 Speech-to-Text

Video 1:

```
0.0 SPEAKER_00: She's on her way.
5.98 SPEAKER_00: Tell everyone.
7.14 SPEAKER_01: She's not supposed to be here until 9.
9.78 SPEAKER_00: Her driver just text messaged,
11.34 SPEAKER_00: and her facialist ruptured a disc.
13.12 SPEAKER_00: God, these people.
16.42 SPEAKER_00: That I can't even talk about.
19.76 SPEAKER_01: All right, everyone, gird your loins.
23.36 SPEAKER_01: Did someone eat an onion bagel?
98.5 SPEAKER_00: Leave it!
103.88 SPEAKER_02: I don't understand why it's so difficult
105.98 SPEAKER_02: to confirm an appointment.
107.58 SPEAKER_00: I know, I'm so sorry, Miranda.
108.86 SPEAKER_00: I actually did confirm last night.
110.26 SPEAKER_02: The tales of your incompetence do not interest me.
112.58 SPEAKER_02: Tell Simone I'm not going to prove that girl
114.54 SPEAKER_02: that she sent me for the Brazilian layout.
116.64 SPEAKER_02: I asked for clean, athletic, smiling.
118.58 SPEAKER_02: She sent me dirty, tired, and paunchy.
120.64 SPEAKER_02: And RSVP, yes, to the Michael Kors party.
122.58 SPEAKER_02: I want the driver to drop me off at 9.30
124.24 SPEAKER_02: and pick me up at 9.45 sharp.
126.36 SPEAKER_00: Okay, fine.
```

Video 2:

```
4.86-6.14 SPEAKER_02: You know what we should all do?
7.35-8.82 SPEAKER_02: Go see a musical.
12.20-12.92 SPEAKER_01: Sure.
14.86-16.06 SPEAKER_02: And you know which one we should see?
17.46-20.38 SPEAKER_02: The 1996 Tony Award winner.
22.39-23.99 SPEAKER_02: Do you happen to know the name of that one?
24.63-26.24 SPEAKER_01: I don't know, Greece?
28.38-29.09 SPEAKER_02: No.
30.46-31.64 SPEAKER_02: laughter
33.73-35.08 SPEAKER_02: Yes! Rent!
36.43-37.71 SPEAKER_01: Okay, so when do you want to go?
39.08-40.46 SPEAKER_02: Oh, I'm sorry, I can't, I'm busy.
42.67-43.75 SPEAKER_02: LAUGHTER
45.98-49.24 SPEAKER_01: Man, it is so hard to shop for girls.
50.61-52.98 SPEAKER_01: Yes, it is at Office Max.
55.77-56.71 SPEAKER_02: What did you get her?
58.74-59.72 SPEAKER_01: A pen.
61.35-68.85 SPEAKER_01: It's two gifts in one. It's a pen. That's also a clock. Huh? Yeah. Ha-ha. You can't give her that.
69.30-70.79 SPEAKER_01: Why not? Because she's not 11.
72.86-74.45 SPEAKER_01: And it's not the seventh night of Hanukkah.
77.55-84.93 SPEAKER_02: Honey, what he means by that is, while this is a very nice gift, maybe it's just not something a boyfriend gives.
84.93-91.86 SPEAKER_01: Sure it is. She needs a pen for work. She's writing. She turns it over. Whoa! It's time for my date with Joey.
94.75-96.05 SPEAKER_01: All right, look, look, what...
96.67-99.02 SPEAKER_01: What did you get for Angela Del Vecchio for her birthday?
99.73-102.83 SPEAKER_01: She didn't have a birthday while we were going out. For three years,
106.93-110.92 SPEAKER_01: It's too late and I got an audition. I can't shop anymore.
112.96-116.20 SPEAKER_01: I will go out and I will try to find something for her, okay?
116.94-117.77 SPEAKER_01: Thanks, man.
119.22-120.69 SPEAKER_01: While you're at it, could you get her a card?
122.22-124.50 SPEAKER_01: Would you like me to write her a little poem as well?
125.53-128.06 SPEAKER_01: Or just get a card that has a poem already in it.
135.22-136.87 SPEAKER_00: Oh, yeah.
136.95-139.98 SPEAKER_00: So they hold me to you
```

Video 3:

```
1.04-6.21 SPEAKER_00: Hey, don't do that. Hey, don't do that. No, no, no. It's worse than the thumb.
1.18-4.11 SPEAKER_01: Hey, don't do that. Put it out. No, no.
7.44-13.02 SPEAKER_00: This is so unfair. Why is it unfair? So I have a flaw. Big deal.
13.55-22.49 SPEAKER_00: Like Joey's constant knuckle-cracking isn't annoying? And Ross with his over-pronouncing every single word? And Monica with that snort when she laughs? I mean, what the hell is that thing?
25.34-28.55 SPEAKER_00: I accept all those flaws. Why can't you accept me for this?
35.84-38.03 SPEAKER_00: Does the knuckle cracking bother everybody?
38.03-41.86 SPEAKER_01: Well?
42.91-44.34 SPEAKER_01: I could live without it.
44.53-45.41 SPEAKER_01: I'm not.
47.52-51.38 SPEAKER_00: Was it like a little annoying or is it like when Phoebe chews her hair?
54.08-55.14 SPEAKER_01: laughter
55.14-58.98 SPEAKER_00: Don't listen to him, Phoebes. I think it's endearing.
59.13-59.95 SPEAKER_00: You do.
60.49-63.99 SPEAKER_01: Oh, my God.
64.97-66.50 SPEAKER_01: .
66.91-72.49 SPEAKER_00: Now, there's nothing wrong with speaking correctly. Indeed, there isn't.
75.16-76.36 SPEAKER_00: I should really get back to work.
77.07-79.48 SPEAKER_00: Yeah, otherwise someone might get what they actually ordered.
79.78-81.28 SPEAKER_01: Well...
81.66-82.80 SPEAKER_00: Oh!
82.99-86.16 SPEAKER_00: The hair comes up and the gloves come on.
86.31-95.70 SPEAKER_01: Oh, it started with this guy right here, and that's right here, and that's right here. And he's coming in the Darington over there. Now, if you don't want to get out of here,
86.35-89.82 SPEAKER_00: Oh, it started with this guy right here. And that's what I started with.
```

Figure 6.5 Result of Speech-to-Text.

Table 6.5 Performance Metrics on Speech-to-text

Video ID	**Word Error Rate (%)**
Video 1	0
Video 2	3
Video 3	5.5

6.2.2 User-Centric Testing Result

Table 6.6 User Evaluation Results

Criteria	Participant 1	Participant 2	Participant 3	Average Score
Accuracy	8.5	8.5	8.5	8.5
Word Choice Precision	9	9	9	9
Tone and Neutrality	8	8.5	8.5	8.333333
Speed and Timing	9	9	9.5	9.166667
Naturalness of Voice	8	8.5	8.5	8.333333
Overall Average				**8.666667**

The user evaluation results of the “Enjoyable” system, as shown in the table, indicate generally strong performance across all five criteria, with an overall average score of **8.67 out of 10**. The **highest scores** were observed in *Word Choice Precision* (9.0) and *Speed and Timing* (9.17), reflecting the system's effective vocabulary usage and good synchronization with visual content. *Tone and Neutrality* and *Naturalness of Voice* each scored around **8.33**, indicating satisfactory but slightly less consistent performance. The **lowest score** was in *Accuracy* (8.5), suggesting some misalignment between the audio descriptions and the actual visual scenes.

A key factor influencing this outcome is the use of an **Indonesian language AI model** instead of one specifically trained on **Malay**. While Indonesian and Malay share many linguistic similarities, there are crucial differences in vocabulary, context, and cultural nuance that can affect the clarity and relevance of audio descriptions. Participants may have noticed subtle mismatches or awkward phrasing that affected the perceived accuracy and naturalness of the narration. Transitioning to a Malay-specific

AI model could help address these issues, improving both accuracy and overall user satisfaction.

6.3 Project Challenge

Throughout the development of the “Enjoyable” audio description system, multiple technical challenges surfaced that required significant investigation and innovative solutions. Among them, two challenges were particularly critical due to their implications on both the quality and usability of the system: the limitation of multimodal large language models (LLMs) in recognizing non-famous individuals, and the need for fine-grained control over word output to fit tightly constrained audio intervals.

The first major challenge encountered was the inability of multimodal LLMs, such as Gemini, to recognize individuals who are not globally known. While these models perform well with famous figures, they lack contextual memory or identification capabilities for common or anonymous people frequently featured in everyday video content. This led to generic and ambiguous visual descriptions, such as “a person is speaking,” without specifying who the individual actually was, especially when multiple people appeared on screen. This limitation significantly impacted character-level precision in the generated narration.

To resolve this, a custom pipeline was developed that involved three steps: face detection, face clustering, and consistent identity labeling using a predefined color-to-name dictionary (e.g., **{red: Andy, blue: Cindy}**). The system annotated detected faces with colored bounding boxes and used the resulting visual markups to condition LLM prompting. This combination acted as a workaround to simulate identity consistency over time and allowed the model to refer to characters by name, even if it could not recognize them natively. As a result, the narrative outputs became significantly more consistent, context-aware, and personalized.

The second key challenge emerged from the need to control the word count of generated descriptions, especially when injecting them into brief non-verbal segments of video. Multimodal LLMs typically generate variable-length outputs, often exceeding the target duration for narration, which can lead to narration overlapping with spoken dialogue or being unnaturally sped up during Text-to-Speech (TTS) synthesis. This posed a serious problem because clarity, pacing, and timing are vital for visually impaired users relying on audio cues. The system required strict control over how much descriptive information was packed into a limited time frame—often just one or two seconds.

To address the problem of generating descriptions that precisely fit within limited non-verbal intervals in the video, a **LLM-Based Multi-Agent System** was designed and implemented. In this architecture, two specialized agents—the **Writer Agent** and the **Word Count Agent** —work collaboratively in an iterative feedback loop. The Writer Agent is responsible for generating the initial visual description based on the contextual prompt, while the Word Count Agent immediately evaluates the length of the generated output. If the resulting description exceeds the predefined word count (based on available timing), or falls below a minimum threshold that risks semantic insufficiency, the Word Count Agent will trigger a prompt back to the Writer Agent to either shorten or elaborate the content accordingly. This feedback continues in a loop until an optimal, time-aligned description is produced. The process ensures that the final narration is not only semantically rich but also fits accurately within the allocated silent intervals detected by the Voice Activity Detection (VAD) module. This dynamic word-length control mechanism enhances the temporal precision and usability of the generated descriptions and plays a vital role in maintaining smooth audio synchronization during narration playback.

6.4 Objectives Evaluation

This section provides a comprehensive evaluation of the project's fulfillment of the original objectives defined at the onset of development. Each objective is reviewed based on the system implementation, empirical results from module testing, and feedback obtained from visually impaired participants during user-centric evaluation. The results confirm the extent to which the *Enjoyable* system achieved its intended outcomes in accessibility, personalization, automation, and technical feasibility.

Objective 1: Develop an Automated Audio Description Generation Pipeline

This objective was successfully achieved through the development of an integrated, end-to-end processing pipeline that automatically generates audio descriptions for video content. The system leverages multiple core modules—such as I-Frame extraction, Voice Activity Detection (VAD), Face Detection, Face Clustering, Active Speaker Detection, Speech-to-Text (STT), and Image Captioning—each of which was independently tested and evaluated. As detailed in Section 6.2.2, each module demonstrated high reliability and processing accuracy, with most achieving F1-scores above 85–90%. These modules worked in harmony to process raw video input into structured, personalized audio narratives, confirming the feasibility of fully automated audio description generation.

Objective 2: Develop Contextually Accurate and Audience-Tailored Audio Descriptions

This objective was addressed through a combination of multimodal input processing and intelligent scene-temporal alignment strategies. The integration of Face Detection and Clustering helped condition the LLM to generate identity-consistent narrations, allowing the audio output to refer to persons with specific labels (e.g., “Andy” or “Cindy”) rather than generic placeholders like “someone.” Additionally, constrained word-length control ensured that the generated content did not overwhelm limited audio gaps. User testing results revealed high performance in this area, with participants averaging **9.0** for Word Choice Precision and **8.67** in overall experience—reflecting that the descriptions were both accurate and stylistically accessible for visually impaired users.

Objective 3: Design a Systematic Evaluation Framework to Measure Accessibility Impact

This objective was successfully met through a dual-layer evaluation system combining module-level validation (precision, recall, WER, DER, etc.) and user-level subjective scoring. The feedback obtained from three blind participants at the Malaysia Foundation for the Blind provided meaningful insights on the practical usability of the system. Criteria such as Accuracy (8.5/10), Speed and Timing (9.17/10), and Naturalness of Voice (8.33/10) confirmed that the system enhanced accessibility—and in particular, delivered valuable narrative assistance without overwhelming or distorting content comprehension.

6.5 Concluding Remark

Through a structured two-tiered evaluation approach — comprising detailed module-level performance testing and user-centric usability analysis — the system's reliability, efficiency, and effectiveness were thoroughly assessed.

Quantitative testing revealed strong performance across key modules, with high F1-scores, reduced frame loads, and low word error rates, confirming the system’s technical robustness. The qualitative feedback collected from blind users affiliated with the Malaysia Foundation for the Blind validated the system's ability to produce narrations that are clear, well-timed, and naturally voiced. The overall average user rating of 8.67 out of 10 reflects a promising impact in real-world accessibility contexts.

Despite facing challenges such as API rate limitations and speaker recognition ambiguities, these were mitigated through innovative solutions such as the SMARS framework and dynamic word-length control. These implementations allowed for contextually rich, temporally precise, and stylistically refined audio descriptions that adhered to non-verbal timing gaps in dialogue-heavy video content.

Finally, the evaluation confirmed that the original project objectives — automation, accessibility, personalization, and systematic evaluation — were all comprehensively achieved.

Chapter 7

Conclusion and Recommendations

7.1 Conclusion

Video streaming platforms have revolutionized media consumption, offering unprecedented access to a variety of content. However, they often fail to accommodate blind and visually impaired users, who rely on audio descriptions to fully understand visual material. Audio descriptions are verbal narrations that convey essential visual elements, such as actions, settings, and non-verbal cues, enabling visually impaired users to immerse themselves in the viewing experience. Despite the popularity of platforms like YouTube, there is a lack of native support for audio descriptions, marginalizing a significant portion of the audience. Although services like Netflix have made some progress, their audio descriptions are limited in availability and language, failing to provide comprehensive access to a global, diverse audience.

The process of creating audio descriptions is labor-intensive, requiring manual scripting, narration, and synchronization with video content. This method is costly, with prices ranging from $15 to $75 per minute of video. As a result, many video platforms struggle to include audio descriptions for a broad range of content, leaving many videos undescribed. Additionally, the availability of descriptions is typically limited to one language, usually English, further limiting accessibility for non-English-speaking users. This manual approach is not scalable, leading to inefficiencies in production and distribution, and exacerbating the exclusion of blind and visually impaired users from enjoying video content.

This project tackles the accessibility challenges faced by visually impaired users when engaging with video content by integrating an advanced Online Video Streaming Platform with an Automated Audio Description System. Leveraging cutting-edge technologies, the system generates accurate and context-aware audio descriptions. Key modules such as I-frame extraction, voice activity detection, face clustering, and detailed image captioning work together to provide seamless and comprehensive descriptions of both visual and spoken content. The integration of a LLM-Based Multi-Agent System ensures that the descriptions are optimized for relevance, diversity, and fluency, offering a more personalized and enriched viewing experience.

The project introduces several innovative contributions to the field of automated audio description generation. It enhances video recognition through a novel clustering and labeling method, enabling processing of diverse video types beyond mainstream content. The system also supports multiple languages, addressing the limitations of existing solutions by ensuring global accessibility. A structured script-based approach is employed, starting with detailed image captions and character dialogues to provide a robust foundation for contextually accurate descriptions. The multi-agent system further refines the descriptions, ensuring they are culturally relevant and tailored to the specific needs of different audiences. Through these innovations, the project significantly enhances the accessibility and inclusivity of video content for visually impaired users, making it more accessible to a diverse, global audience.

7.2 Recommendations

While the “Enjoyable” system demonstrates promising improvements in automated audio description for visually impaired users, there remain several opportunities for enhancement and future expansion. These future directions will further improve accessibility, linguistic customization, and immersive user experience. To better serve global and multilingual audiences, future development will involve integrating multilingual and region-specific TTS models . This includes support for native Malay, Chinese dialects, Tamil, and other indigenous languages used in Southeast Asia. This will improve inclusivity and ensure narrative content is more relatable and intuitive across different linguistic backgrounds.

To further enhance the accessibility and immersion of the *Enjoyable* system, the implementation of stereo audio separation is proposed as a future enhancement. In this configuration, the automated audio description will be routed exclusively to the left audio channel , while the original video audio will be preserved in the right channel . This channel-based audio multiplexing approach allows users—particularly those with partial vision or heightened auditory perception—to distinguish between narration and native soundscapes with higher spatial clarity. For example, when the descriptive content is isolated to one ear (e.g., left channel), users can mentally attribute the narration source independently from dialogues, music, or ambient sound in the right channel. This technique avoids cognitive overload caused by overlapping audio layers and enables users to selectively focus on either channel depending on their accessibility needs.

REFERENCES

[1] "What is audio description? | DO-IT." https://www.washington.edu/doit/what-audio-description

[2] R. Klein, "How to create audio description for YouTube Videos," 3Play Media, Aug. 23, 2024. https://www.3playmedia.com/blog/how-to-add-audio-description-to-youtube-videos/

[3] "Audio Description for TV shows and movies," Help Center. https://help.netflix.com/en/node/25079

[4] "YouDescribe - audio description for YouTube videos." https://youdescribe.org/support/about

[5] E. Lewis, "How much does audio description cost?," 3Play Media, Jun. 06, 2022. https://www.3playmedia.com/blog/how-much-does-audio-description-cost/

[6] S. Online, "Dr Noor Hisham: Prevalence of blindness is 1.2% of country's population," The Star, Oct. 13, 2022. [Online]. Available: https://www.thestar.com.my/news/nation/2022/10/13/dr-noor-hisham-prevalence-of-blindness-is-12-of-country039s-population

[7] Y. Wang, W. Liang, H. Huang, Y. Zhang, D. Li, and L.-F. Yu, "Toward automatic audio description generation for accessible videos," CHI '21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, May 2021, doi: 10.1145/3411764.3445347.

[8] T. Han, M. Bain, A. Nagrani, G. Varol, W. Xie, and A. Zisserman, "AutoAD: Movie Description in context," arXiv (Cornell University), Jan. 2023, doi: 10.48550/arxiv.2303.16899.

[9] T. Han, M. Bain, A. Nagrani, G. Varol, W. Xie, and A. Zisserman, "AutoAD II: The Sequel -- Who, when, and what in movie audio description," arXiv (Cornell University), Jan. 2023, doi: 10.48550/arxiv.2310.06838.

[10] T. Han, M. Bain, A. Nagrani, G. Varol, W. Xie, and A. Zisserman, "AutoAD III: the prequel -- Back to the pixels," arXiv (Cornell University), Apr. 2024, doi: 10.48550/arxiv.2404.14412.

[11] "MERN Stack explained," MongoDB. https://www.mongodb.com/resources/languages/mern-stack

[12] “Vercel: Build and deploy the best web experiences with the Frontend Cloud,” Vercel. https://vercel.com/

[13] Cloudinary, “Image and video upload, storage, optimization and CDN,” Cloudinary, Aug. 08, 2024. https://cloudinary.com/

[14] A. Beach and A. Owen, Video Compression Handbook. 2018.

[15] J. Nilsson and T. Akenine-Möller, “Understanding SSIM,” arXiv (Cornell University), Jan. 2020, doi: 10.48550/arxiv.2006.13846.

[16] E. Karami, S. Prasad, and M. S. Shehata, “Image matching using SIFT, SURF, BRIEF and ORB: Performance Comparison for distorted images,” arXiv (Cornell University), Jan. 2017, doi: 10.48550/arxiv.1710.02726.

[17] “pyannote/segmentation-3.0 · Hugging Face.” https://huggingface.co/pyannote/segmentation-3.0

[18] “Introducing whisper,” OpenAI. https://openai.com/index/whisper/ (accessed Sep. 11, 2024).

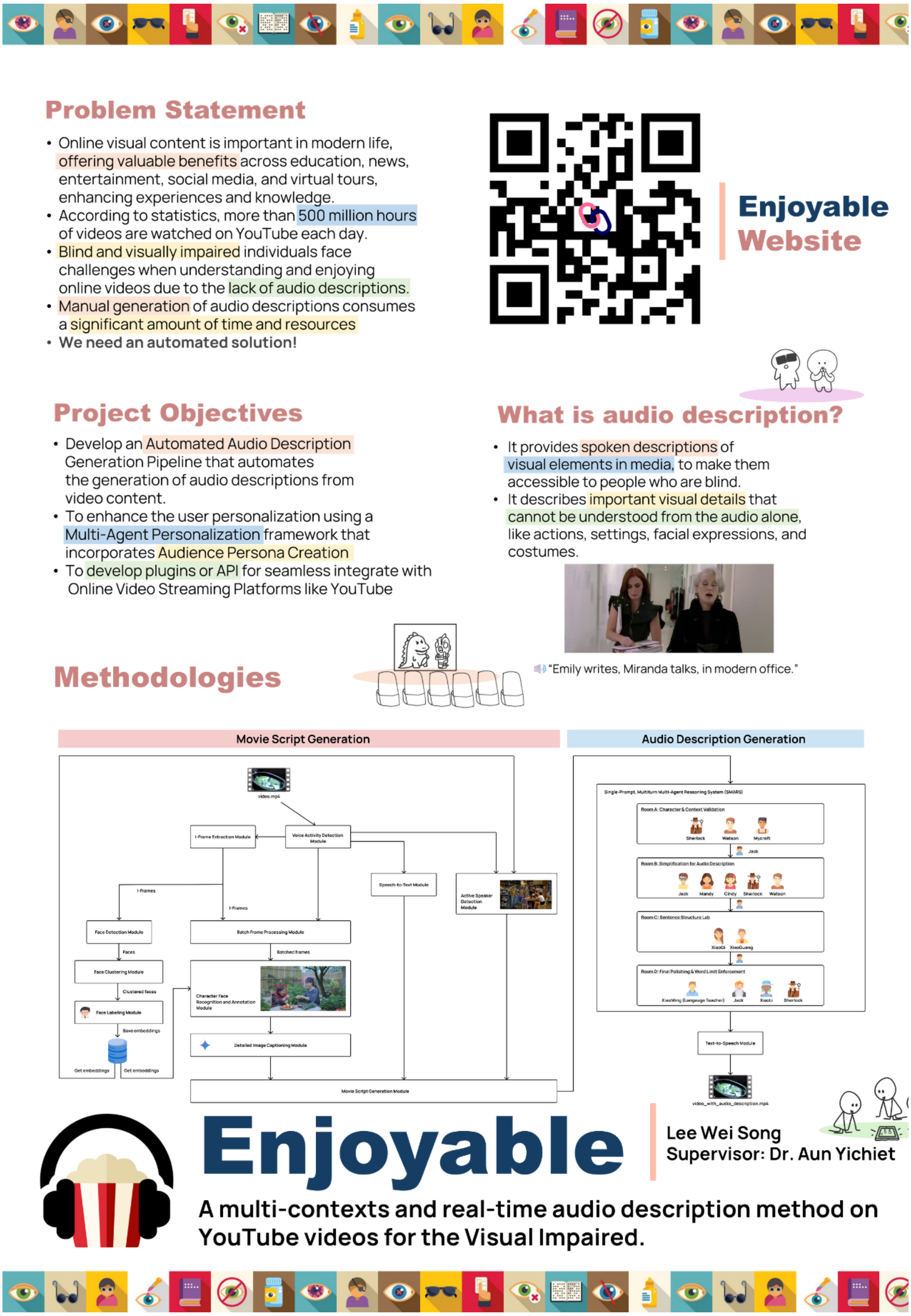
[19] R. Sharma and S. Narayanan, “Audio-Visual Activity Guided Cross-Modal Identity Association for Active Speaker detection,” IEEE Open Journal of Signal Processing, vol. 4, pp. 225–232, Jan. 2023, doi: 10.1109/ojsp.2023.3267269.

[20] A. Rosebrock, “Face detection with dlib (HOG and CNN) - PyImageSearch,” PyImageSearch, May 12, 2023. https://pyimagesearch.com/2021/04/19/face-detection-with-dlib-hog-and-cnn/

[21] C. Biemann, “Chinese whispers,” Proceedings of TextGraphs: The First Workshop on Graph Based Methods for Natural Language Processing, Jan. 2006, doi: 10.3115/1654758.1654774.

[22] “Hello GPT-4o,” OpenAI. https://openai.com/index/hello-gpt-4o/ (accessed Sep. 11, 2024).

Poster


Problem Statement
• Online visual content is important in modern life, offering valuable benefits across education, news, entertainment, social media, and virtual tours, enhancing experiences and knowledge.
• According to statistics, more than 500 million hours of videos are watched on YouTube each day.
• Blind and visually impaired individuals face challenges when understanding and enjoying online videos due to the lack of audio descriptions.
• Manual generation of audio descriptions consumes a significant amount of time and resources
• We need an automated solution!
Enjoyable Website
Project Objectives
• Develop an Automated Audio Description Generation Pipeline that automates the generation of audio descriptions from video content.
• To enhance the user personalization using a Multi-Agent Personalization framework that incorporates Audience Persona Creation
• To develop plugins or API for seamless integrate with Online Video Streaming Platforms like YouTube
What is audio description?
• It provides spoken descriptions of visual elements in media, to make them accessible to people who are blind.
• It describes important visual details that cannot be understood from the audio alone, like actions, settings, facial expressions, and costumes.
"Emily writes, Miranda talks, in modern office."
Methodologies
Movie Script Generation
Audio Description Generation
Enjoyable
Lee Wei Song
Supervisor: Dr. Aun Yichiet
A multi-contexts and real-time audio description method on YouTube videos for the Visual Impaired.