



# Movie Descriptor

**Course:** 2024S-T4 AIP Group A

**Project Coordinator:** Stanley Chor

**Project Advisor:** Bhavik Gandhi

**Date:** 1<sup>st</sup> Aug 2024

**TABLE OF CONTENTS**

	Page
ABSTRACT .....	02
ACKNOWLEDGMENT .....	03
INTRODUCTION .....	04
LITERATURE REVIEW .....	05
PROJECT GOALS .....	07
VALUE TO THE INDUSTRY .....	08
METHODS .....	09
TEAM STRUCTURE .....	11
FINDINGS .....	12
PROJECT WORK SCHEDULE .....	14
RISK MANAGEMENT PLAN .....	16
DISCUSSIONS .....	17
SCREENSHOTS .....	19
CONCLUSION .....	26
REFERENCES .....	27
RECOMMENDATION .....	28

## **Abstract**

This project addresses the significant gap in movie experiences for visually impaired audiences by developing an innovative Movie Descriptor system. The primary goal is to enhance the cinematic experience for users with visual impairments through the integration of detailed audio descriptions directly into movie scenes. These descriptions aim to provide a rich, immersive experience that captures the visual nuances of the film, allowing visually impaired individuals to enjoy the story, characters, and setting as fully as sighted viewers do.

To achieve this, the project leverages cutting-edge machine learning technologies, including BLIP (Bootstrapped Language Image Pretraining) and BART (Bidirectional and Auto-Regressive Transformers). These models are used to analyse video frames and generate precise, contextually relevant descriptions. Additionally, the system employs Google Text-to-Speech (GTTS) for converting the generated text descriptions into high-quality audio, ensuring clarity and natural-sounding speech.

The system is designed with user accessibility in mind, featuring a straightforward interface built with Streamlit, making it easy for users to navigate and enjoy movies. Furthermore, Hugging Face's platform supports the deployment, enhancing the system's performance and scalability. This approach not only broadens the cinematic experience for visually impaired individuals but also promotes inclusivity in media consumption, ensuring that everyone can enjoy the art of filmmaking without barriers.

## **ACKNOWLEDGMENT**

We would like to sincerely thank everyone who helped make the "Movie Descriptor" project a reality. A varied and brilliant team's combined work, commitment, and knowledge have made this project feasible. We would like to express our sincere gratitude to all the people on our team, project managers, stakeholders, and the community at large who helped us along the way.

We are grateful to Stanley Chor, the project coordinator, and Bhavik Gandhi, our project adviser, for realising the importance of this project and giving us the responsibility of reinventing the HR and recruitment industry. Your advice, comments, and steadfast support have been crucial in helping to shape this game-changing answer. Our goal is to surpass your expectations and fulfil our commitments.

We also thank the many experts from many industries who have contributed to this project, including data scientists, designers, developers, and professionals from numerous fields, for their significant efforts. The "Movie Descriptor" is the result of your unwavering commitment to invention and laborious effort. We are excited to carry on this cooperative journey, establish new benchmarks in the talent acquisition space, and improve hiring procedures all around the world in terms of efficacy and efficiency.

## **Introduction**

People who are visually impaired encounter many obstacles in their quest for a complete cinematic experience. Actions, facial expressions, and scenes are only a few examples of the visual components that make up the cinematic experience. Audiences with visual impairments lose out on these essential elements in the absence of thorough audio descriptions, which makes for a less immersive and interesting experience. By creating an extensive Movie Descriptor system that incorporates excellent audio descriptions into movie scenes, this project fills this gap. The intention is to offer visually impaired viewers a more authentic, immersive cinematic experience that faithfully captures the visual elements of films.

Movies are an effective storytelling tool that use both visual and aural cues to portray stories, feelings, and creative expression. It can be difficult for people who are blind or visually impaired to follow the storyline and completely understand the subtleties of the narrative due to the lack of visual information. While there are traditional audio description services, their range and calibre are frequently constrained. They could leave out important facts or delicate emotional undertones, leaving the event feeling unfinished.

With the help of these developments, our project seeks to create a dependable, expandable, and user-friendly system that can offer excellent audio descriptions for a variety of films.

## Literature Review

### Competing Applications

#### 1. Traditional Audio Description Services:

Traditional audio description services involve human narrators who provide descriptive commentary about the visual elements of a movie. These services are typically pre-recorded and synchronized with the movie to offer visually impaired audiences a richer experience. However, they are limited by the narrator's interpretation and may not capture all the subtle details or emotional nuances of the scenes.

#### 2. Automated Video Description Systems:

Automated video description systems utilize machine learning and natural language processing to generate descriptions of video content. Some notable research and applications in this area include:

- **Microsoft's Seeing:** A Seeing AI uses computer vision to describe the world around users, including reading text and recognizing objects and people. Although primarily designed for everyday tasks, it has features that could be adapted for describing visual content like movies. However, it lacks the focused and contextual description needed for comprehensive movie descriptions (Microsoft, 2024).
- **Video-to-Text (VTT) Methods:**
  1. Single Frame Models : Process each frame independently, lacking temporal context.
  2. CNN + RNN Models: Combine CNNs for feature extraction with RNNs for temporal sequences.
  3. 3D Convolutions: Capture spatiotemporal features directly from video chunks.
  4. Dynamic Encoders: Use RNNs to better retain temporal information.

#### Evaluation Metrics

- Common metrics include BLEU, ROUGE, METEOR, and CIDEr.
- Challenges remain in achieving robust and consistent evaluations compared to human assessments.

#### Recent Advances

- Spatial attention mechanisms and syntactic integration improve description accuracy.

## Conclusion

-Significant advancements in deep learning and large-scale datasets have improved VTT, but evaluation challenges and robust temporal feature extraction need further research.

- **Video-to-Text (VTT) Problem:** which involves generating natural language descriptions from video content. This is a key task in bridging vision and language, and it has significant applications in multimedia retrieval, surveillance, robotics, sign language translation, and assistance for the visually impaired.

## Key Areas Reviewed:

1. Methods for VTT: Matching-and-Ranking-Based Techniques: These techniques focus on retrieving the most relevant text descriptions from a predefined set based on the input video.

Natural Language Generation (NLG)-Based Techniques: These generate descriptions directly from video data using models that learn to produce text.

2. Evaluation Metrics: Common metrics used to evaluate these techniques include BLEU, ROUGE, METEOR, and CIDEr, which are borrowed from related fields like machine translation and image captioning.

3. Benchmark Datasets: The review discusses various datasets that have been developed to facilitate VTT research, including multilingual datasets like VaTeX which support English and Chinese.

## In-Text Citations

- Microsoft. (2024). *Seeing AI: The app that narrates the world around you*. Retrieved from [Microsoft Seeing AI website](#)
- Perez-Martin J, Bustos B, Guimarães SJF, Sipiran I, Pérez J, Said GC. A comprehensive review of the Video-to-Text problem. arXiv.org. <https://arxiv.org/abs/2103.14785>. Published March 27, 2021.
- Li J, Li D, Xiong C, Hoi S. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. arXiv.org. <https://arxiv.org/abs/2201.12086>. Published January 28, 2022.

## **Project Goals**

The Accessible Multimedia Project aims to transform video accessibility by offering all-inclusive and automated methods for producing textual and audio descriptions for videos. Our goal is to improve the viewing experience for text and audio-only users as well as those with visual impairments. Using text-to-speech technology and sophisticated AI models like VisionEncoderDecoder (e.g., ViT-GPT2), we want to establish an inclusive environment where all users can easily access and comprehend video information.

Features of the project include seamless video downloading, extraction of video metadata, and an easy-to-use online interface for interaction. We also prioritise guaranteeing the accessibility of our product by making it available to the public on Hugging Face. We hope that this initiative will empower those who are visually impaired, encourage inclusivity, and establish new benchmarks for multimedia accessibility.



## **Value to Industry**

### **Enhanced User Engagement**

The Accessible Multimedia Project improves user experience by providing automated text and audio descriptions for video content. This feature ensures that visually impaired users and those who prefer audio content can engage more fully with multimedia, leading to higher user satisfaction and retention.

### **Compliance with Accessibility Standards**

By adopting this technology, companies can ensure compliance with important accessibility standards such as the ADA (Americans with Disabilities Act) and WCAG (Web Content Accessibility Guidelines). Meeting these standards not only avoids potential legal issues but also enhances the company's reputation as an inclusive and socially responsible entity.

### **Broadening Audience Base**

The project opens up content to a wider audience, including millions of visually impaired individuals globally. This expansion can lead to a significant increase in viewership and customer base, tapping into an underserved market segment.

### **Cost-Effective and Scalable Solution**

The web interface, deployed via platforms like Google Colab and Streamlit, with ngrok support for public access, makes the solution easy to use and deploy. This scalability ensures that businesses can implement the technology without significant infrastructure changes, keeping costs low while maximizing reach and impact.

### **Improved Market Position**

Adopting this innovative solution demonstrates a commitment to inclusivity and technological advancement. This can enhance brand reputation, making the company more attractive to a broader audience and strengthening its position in the competitive digital marketplace.

## Methods

### Technologies Used in Detail

#### Computer Vision

- **Frame Extraction:** We utilized computer vision techniques to extract frames from videos. This step is crucial for analyzing the visual content and generating descriptions for each frame. Frame extraction allows for a detailed and comprehensive analysis of the video, ensuring that all important visual elements are captured.

#### Machine Learning/Natural Language Processing

- **VIT Model:** Initially, we used the Vision Transformer (VIT) model to generate descriptions from video frames. However, it was found to be inefficient and produced irrelevant outputs, leading us to replace it with a more suitable model.
- **Salesforce BLIP Model:** The BLIP (Bootstrapping Language-Image Pre-training) model was employed to generate text descriptions from video frames. This model leverages large-scale pre-training on various visual and textual tasks, making it capable of producing detailed and accurate descriptions.
- **BART Model:** We used the BART (Bidirectional and Auto-Regressive Transformers) model for summarizing text descriptions generated by the BLIP model. The BART model helps in creating coherent and concise narratives by summarizing the frame-level descriptions into scene-level descriptions.
- **Regex:** Regular expressions (Regex) were utilized to clean and refine the summarized descriptions from the BART model. This step ensures that the final descriptions are free from any extraneous information and are clear and concise.

#### Audio/Video Processing

- **gTTS (Google Text-to-Speech):** The Google Text-to-Speech (gTTS) tool was used to convert the text descriptions into audio form. This conversion is essential for providing an immersive audio experience for visually impaired users.
- **ffmpeg:** We used ffmpeg to merge the audio descriptions with the video. This tool allows for seamless integration of audio and video, ensuring that the audio descriptions are perfectly synchronized with the visual content.

#### Frontend

- **HTML/CSS/JavaScript:** These technologies were used to create the frontend of our application. The frontend provides a user-friendly interface where users can input video links and receive audio-described outputs.

## Deployment Platforms

- **Streamlit:** Streamlit was used for creating the user interface of the model. It provides an interactive and intuitive interface for users to interact with the system.
- **Hugging Face:** Hugging Face was utilized for deploying the models and managing the workflow. This platform ensures that the models are easily accessible and can be efficiently managed.

## Methods Used to Gather Data

- **Surveys:** We conducted surveys among visually impaired individuals to gather feedback on their experience with traditional audio description services and their expectations from an automated system. This feedback was crucial in understanding the user needs and refining our solution.
- **Research:** Extensive research was conducted to identify the best models and techniques for generating and summarizing video descriptions. This included reviewing academic papers, studying existing solutions, and experimenting with different models.

## Reasons for Using Methods Listed

- **Frame Extraction:** By extracting frames, we ensure a detailed analysis of the video content, capturing all significant visual elements that need to be described.
- **Model Selection (BLIP and BART):** The choice of the BLIP model for generating descriptions and the BART model for summarization was based on their proven capabilities in handling complex visual and textual tasks. These models are well-suited for creating detailed and coherent descriptions.
- **gTTS and ffmpeg:** These tools were chosen for their high-quality output and ease of integration. gTTS provides natural-sounding speech, while ffmpeg ensures seamless merging of audio and video.
- **Surveys:** Gathering direct feedback from visually impaired individuals was essential to understand their specific needs and preferences. This user-centered approach ensures that the final product is tailored to meet their expectations.
- **Research:** Conducting thorough research helped us identify the most effective models and techniques, ensuring the robustness and efficiency of our solution.

## Team Structure

The team "Movie Descriptor" is organised to take advantage of a range of specialisations and guarantee productive cooperation. The team members are arranged into important roles, each of which contributes a unique set of talents necessary for the project's successful completion.

Name	Student id	Role
Prabhakaran Sankareswaran	500221867	Project Manager
Jaivin Jacob	500219980	Technical Lead
Bhanwar Preet Singh	500220399	AI Researcher
Doris Kadiri	500205244	ML Engineer
Kuldip Hareshbhai Mangrola	500219171	Data Analyst
Abhisek Singh	500218488	AI Engineer
Neha Mehmi	500219717	UI Developer

### Labour hours-

Team Member Name	Labour Hours
Prabhakaran Sankareswaran	420 Hrs
Neha	420 Hrs
Jaivin Jacob	420 Hrs
Doris Kadiri	420 Hrs
Bhanwar Preet Singh	420 Hrs
Kuldip Hareshbhai Mangrola	420 Hrs
Abhishek Singh	420 Hrs

## Findings

### Theoretically implementation vs Execution

**Initial Planning** The project began with thorough research into video captioning processes and the best approaches for our needs. We divided tasks, coordinated with stakeholders to gather requirements, and set milestones for the project's progression.

**Tool Selection and Integration** For the frontend, we selected Streamlit due to its ease of use and rapid prototyping capabilities. For the backend, we chose the BLIP model for its robust frame-by-frame description capabilities and BART for efficient text summarization. Additionally, we utilized Regex for cleaning summarized text and gTTS for high-quality text-to-audio conversion. FFmpeg was chosen to merge the generated audio with the original video, ensuring synchronization and audio quality.

**Development and Testing** In the development phase, we successfully integrated the BLIP, BART, and gTTS models into the pipeline, ensuring seamless data flow between models. We synchronized the Streamlit frontend with Hugging Face backend processes, implementing real-time progress indicators for users. Extensive testing was conducted, including unit tests, integration tests, and user acceptance testing (UAT), with iterations based on feedback to enhance performance and user experience.

**Theoretical vs. Practical Execution** The theoretical foundation of the project—focused on robust model selection, integration strategies, and ensuring accessibility—was practically implemented through the development of a user-friendly application. The combination of Streamlit, BLIP, BART, Regex, gTTS, and ffmpeg translated theoretical concepts into a functional tool that delivers accessible multimedia content effectively.

### Findings from Surveying Audience: User Feedback and Usability Testing

**Accessibility and Ease of Use:** Users appreciated the straightforward interface and real-time feedback provided by Streamlit. Suggested minor UI/UX improvements which were promptly addressed.

**Quality of Descriptions:** Initial feedback indicated a high satisfaction rate with the accuracy of frame-by-frame descriptions. However, users highlighted the need for more natural and contextually aware summaries.

**Audio Quality and Synchronization:** Positive feedback on the audio quality and seamless integration with original video audio. Some users suggested options for different voice styles and speeds.

### Findings from Research and Technology Advancement: Technological Insights

**Model Performance:** The BLIP model demonstrated high accuracy in generating frame-by-frame descriptions but required substantial computational resources. Optimization techniques were explored to improve efficiency.

**Text Summarization:** The BART model effectively summarized text but occasionally missed contextual nuances. Incorporating additional NLP techniques helped address these gaps.

**Audio Conversion:** gTTS provided clear and natural-sounding audio, though exploring alternative TTS models could offer more customization options.

### **Innovations and Improvements**

**Real-time Processing:** Achieved near real-time processing by optimizing model performance and streamlining the data pipeline.

**Customization Features:** Introduced options for users to select different voice types and adjust speech rates, enhancing user experience.

**Enhanced Summarization:** Implemented additional NLP techniques to improve the contextual accuracy of summaries.

### **How These Findings Set Us Apart from Other Tools: Unique Selling Points**

**Comprehensive Integration:** The seamless integration of multiple advanced models (BLIP, BART, gTTS) ensures high-quality descriptions and audio output, distinguishing our tool from simpler, less sophisticated alternatives.

**User-centric Design:** Streamlit's user-friendly interface, combined with customizable audio options, makes our tool accessible and appealing to a broad user base.

**Real-time Performance:** Optimized processing allows for near real-time description generation, a significant advantage over slower, batch-processing tools.

**Continuous Improvement:** Regular user feedback and research-driven updates ensure that our tool evolves with technological advancements, maintaining a competitive edge.

Overall, the combination of advanced AI models, a focus on user experience, and continuous innovation positions our video describer tool as a leading solution in the market.

### Project Work Schedule

Schedule	Deliverable Description
Week 3	<ol style="list-style-type: none"> <li>1. Research on project description.</li> <li>2. Create a project board and assign tasks.</li> <li>3. Set up a git repository according to guidelines.</li> <li>4. Set up a Slack for communication</li> </ol>
Week 4	<ol style="list-style-type: none"> <li>1. Started with the real dataset</li> <li>2. Perform data Extraction and preprocessing to ensure quality.</li> <li>3. Creating CSV file with text description.</li> </ol>
Week 5	<ol style="list-style-type: none"> <li>1. Develop initial prototypes for audio description generation models.</li> <li>2. Implement object detection and scene recognition algorithms.</li> <li>3. Train models using collected datasets.</li> <li>4. Evaluate models using various metrics.</li> </ol>
Week 6	<ol style="list-style-type: none"> <li>1. Research and implement techniques for activity recognition in videos.</li> <li>2. Compare different modeling techniques for activity and scene analysis.</li> <li>3. Integrate various models into a system.</li> </ol>
Week 7	<ol style="list-style-type: none"> <li>1. Implement models for enhanced video analysis.</li> <li>2. Apply transfer learning techniques to improve model performance.</li> <li>3. Develop models for scene description.</li> </ol>
Week 8	<ol style="list-style-type: none"> <li>1. Develop models for the description alignment.</li> <li>2. Developed model (VIT model) to process the video and able to create text and audio description.</li> <li>3. Working on narration process.</li> <li>4. Created slides for the progress</li> </ol>
Week 9	<ol style="list-style-type: none"> <li>1. Compile and organize project progress and results for the presentation.</li> <li>2. Create slides highlighting key milestones, challenges, and achievements.</li> </ol>
Week 10	<ol style="list-style-type: none"> <li>1. Working on Spacetime model and fine tuning on it for better performance.</li> <li>2. Implementing the deployment process of VIT model.</li> <li>3. Working on the marketing video of the product.</li> </ol>

Week 11	<ol style="list-style-type: none"> <li>1. Working on a portfolio website.</li> <li>2. Working on UI and deployment</li> <li>3. Working on BLEU score for the model and planning to go with better perform one.</li> <li>4. Working on increasing model accuracy, also trying to speed up the process from analyzing the video, processing, and combining the generated audio caption with the actual video.</li> </ol>
Week 12	<ol style="list-style-type: none"> <li>1. Compared the T5 and Bart model summary</li> <li>2. Testing the other models for deployment</li> <li>3. Working on Final Presentation and Final Report</li> <li>4. Done with Expo booth</li> </ol>
Week 13	<ol style="list-style-type: none"> <li>1. Finished with WIX portfolio.</li> <li>2. Work done on login webpage.</li> <li>3. Need to work on deployment of BLIP model.</li> <li>4. Finished with final ppt.</li> <li>5. Progressing on final report.</li> </ol>
Week 14	Final Project Report + WIX project portfolio website Hosting project expo booth



## **Risk Management**

### **Technical Risks**

There could be problems with model correctness in the Movie Descriptor system, where BLIP and BART might not produce descriptions that are in-depth enough. Implement thorough validation using a variety of datasets to lessen this, and update the models often in response to user input. Detailed integration testing and modular component creation are crucial to prevent integration issues between GTTS, Streamlit, and description generation. Scalable cloud solutions and system optimisation can be used to alleviate performance difficulties under high demand.

### **Operational Risks**

It's possible that the user interface isn't completely user-friendly or accessible. This can be lessened by making sure accessibility guidelines are followed and by doing usability testing with visually impaired people. Streamlit or Hugging Face deployment challenges call for extensive testing in a staging environment together with a fallback strategy in case something goes wrong.

### **Content Risks**

It's important to handle sensitive or improper content, and using content filters and moderation tools will assist reduce the danger. Additionally, make sure that copyright rules are followed by acquiring the required licenses for the use of content.

### **Security Risks**

To prevent data privacy breaches, robust encryption and access controls should be in place, with regular security audits to identify vulnerabilities. System vulnerabilities can be mitigated through software updates and penetration testing.

### **Financial Risks**

Potential dangers include budget overruns and financing shortages, which can be avoided by making a thorough budget, keeping a tight eye on spending, and developing a long-term financial plan that may involve partnerships or grants.

## Discussion

### In-depth Discussion About Our Product

The Movie Descriptor for Visually Impaired People, a comprehensive solution designed to enhance the movie-watching experience for visually impaired individuals. Our product leverages advanced machine learning, natural language processing, and audio processing technologies to generate detailed audio descriptions that provide a rich understanding of the visual elements of a movie.

Our primary goal is to make movies more accessible and enjoyable for visually impaired audiences by providing accurate, coherent, and engaging audio descriptions that capture the essence of each scene. By doing so, we aim to bridge the gap between visual and non-visual experiences, offering a more inclusive entertainment experience.

The development process involved extensive research, user feedback, and iterative testing to ensure that our solution not only meets but exceeds user expectations. We have focused on creating a seamless and user-friendly experience, from the initial input of video links to the final delivery of synchronized audio descriptions.

Our approach combines the best of computer vision, machine learning, and audio processing technologies to deliver a product that is both innovative and practical. We are confident that our solution will set a new standard for accessibility in the entertainment industry, making movies more enjoyable for visually impaired individuals.

### Features

#### 1. High-Quality Descriptions:

- **Salesforce BLIP Model:** Generates detailed and accurate text descriptions for each video frame.
- **BART Model:** Summarizes frame-level descriptions into coherent scene-level narratives, ensuring clarity and engagement.

#### 2. User-Friendly Interface:

- **HTML/CSS/JavaScript:** Provides an intuitive and accessible frontend, allowing users to easily upload videos and access audio-described content.
- **Streamlit:** Facilitates an interactive user interface, making it easy for users to interact with the system.

#### 3. Audio-Video Synchronization:

- **ffmpeg Integration:** Ensures that audio descriptions are perfectly synchronized with the video, providing a seamless viewing experience.
- **gTTS (Google Text-to-Speech):** Converts text descriptions into natural-sounding audio, enhancing the overall quality of the output.

### **How Your Research Impacts Your Technology**

Our research played a crucial role in shaping the development and refinement of our product. By studying existing solutions, such as traditional audio description services and automated video description systems, we identified gaps and opportunities for improvement. Researching advanced models like BLIP and BART allowed us to leverage cutting-edge technology to generate and summarize descriptions effectively. Additionally, exploring different methods for text-to-speech conversion and audio-video synchronization ensured that our product delivers high-quality and seamless audio descriptions. Overall, our research informed the selection of technologies and methodologies, ensuring that our product is robust, efficient, and user-centric.

### **How Your Findings Impact Your Technology**

The findings from our research and development process significantly influenced the functionality and effectiveness of our product. Key findings include:

1. **Model Performance:** The initial use of the VIT model highlighted inefficiencies and irrelevant outputs, leading us to switch to the more effective BLIP model for generating descriptions.
2. **Summarization Effectiveness:** The BART model proved to be highly effective in summarizing detailed frame-level descriptions into coherent narratives, enhancing the quality of audio descriptions.
3. **Text-to-Speech Quality:** gTTS was chosen for its natural-sounding speech output, ensuring that the audio descriptions are engaging and easy to understand.
4. **Seamless Integration:** Using ffmpeg for merging audio with video ensured perfect synchronization, providing a seamless experience for users.

These findings informed our decisions and refinements throughout the development process, resulting in a product that meets the needs and expectations of visually impaired users.

### **How Your Surveys Impact Your Technology**

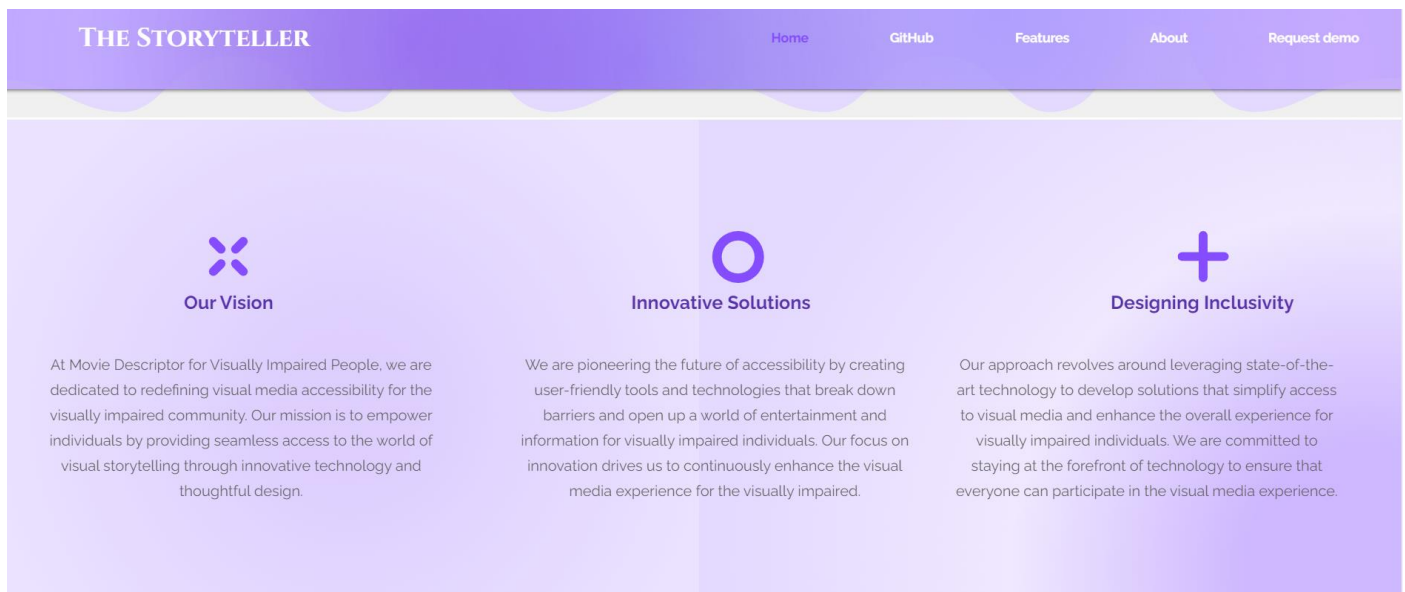
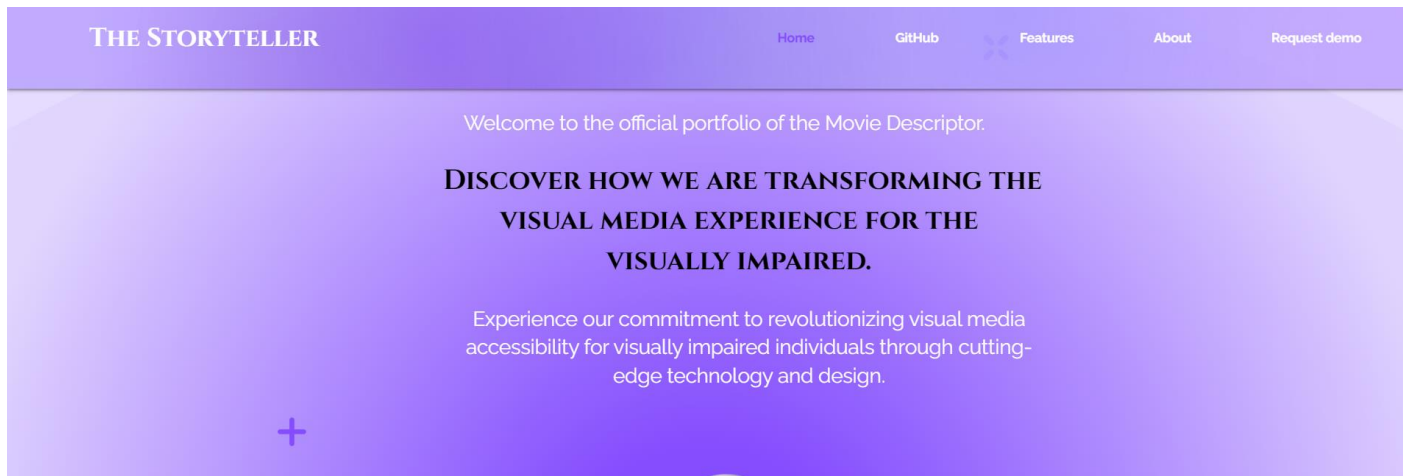
Surveys conducted among visually impaired individuals provided valuable insights into their experiences and preferences. Key impacts of survey feedback include:

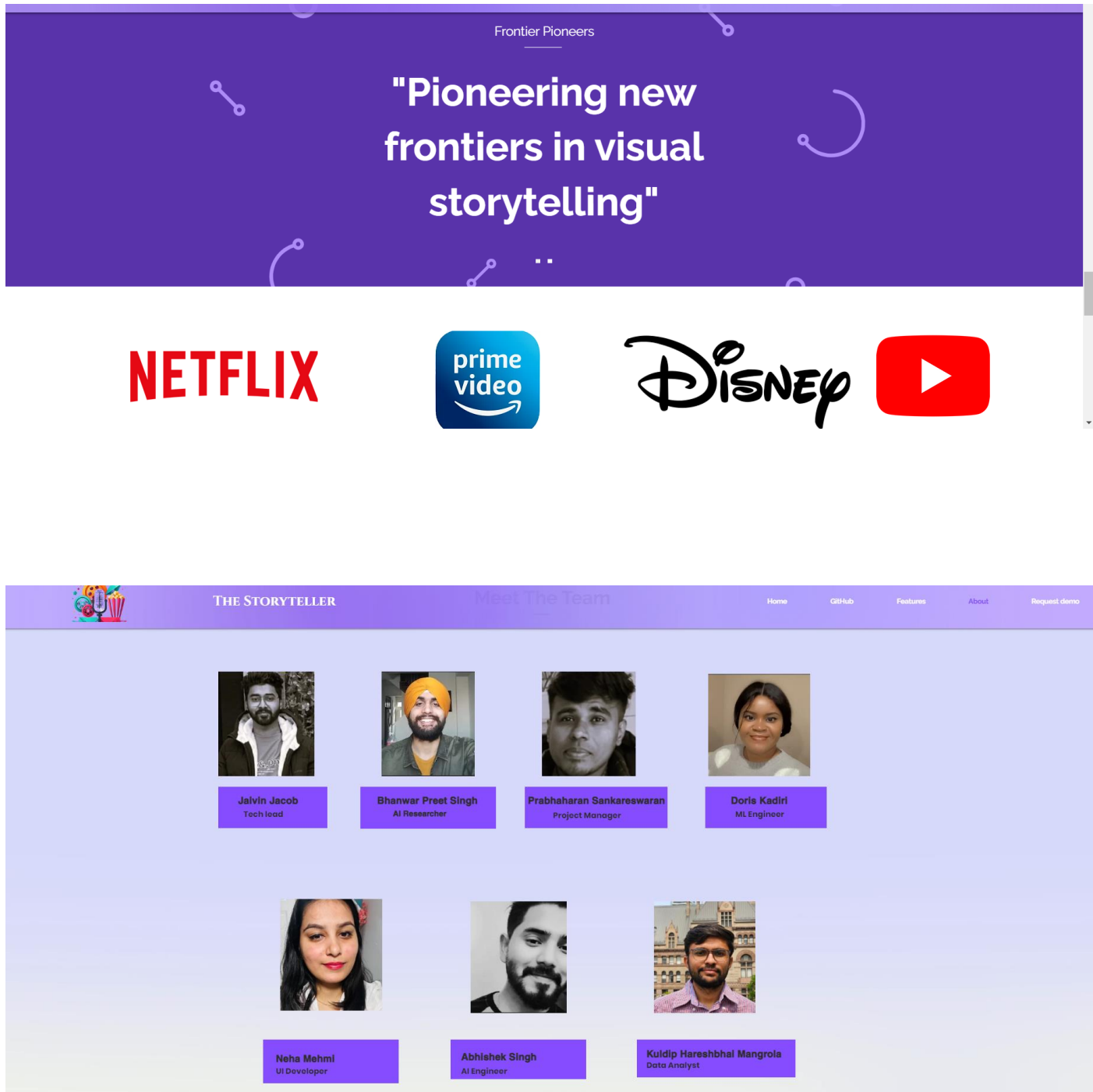
1. **User-Centric Design:** Feedback from surveys emphasized the need for clear and detailed audio descriptions, guiding the development of our descriptive models and summarization techniques.
2. **Audio Quality:** Users expressed a preference for natural-sounding audio, leading us to select gTTS for text-to-speech conversion.
3. **Interface Usability:** Survey responses highlighted the importance of a user-friendly interface, informing the design of our frontend with HTML, CSS, and JavaScript to ensure ease of use.
4. **Feature Preferences:** Users indicated a desire for synchronized audio-video playback, reinforcing our decision to use ffmpeg for merging audio descriptions with video content.

By incorporating survey feedback, we ensured that our product is tailored to meet the specific needs and preferences of visually impaired users, resulting in a more effective and user-friendly solution.

## Screenshots

### WIX Portfolio:





## Website

### The StoryTeller

[Home](#) [Features](#) [About](#) [Team](#) [Contact](#) [LOGIN](#)

## Experience Movies Like Never Before

Discover how we are transforming  
the Movie experience for the visually  
impaired.



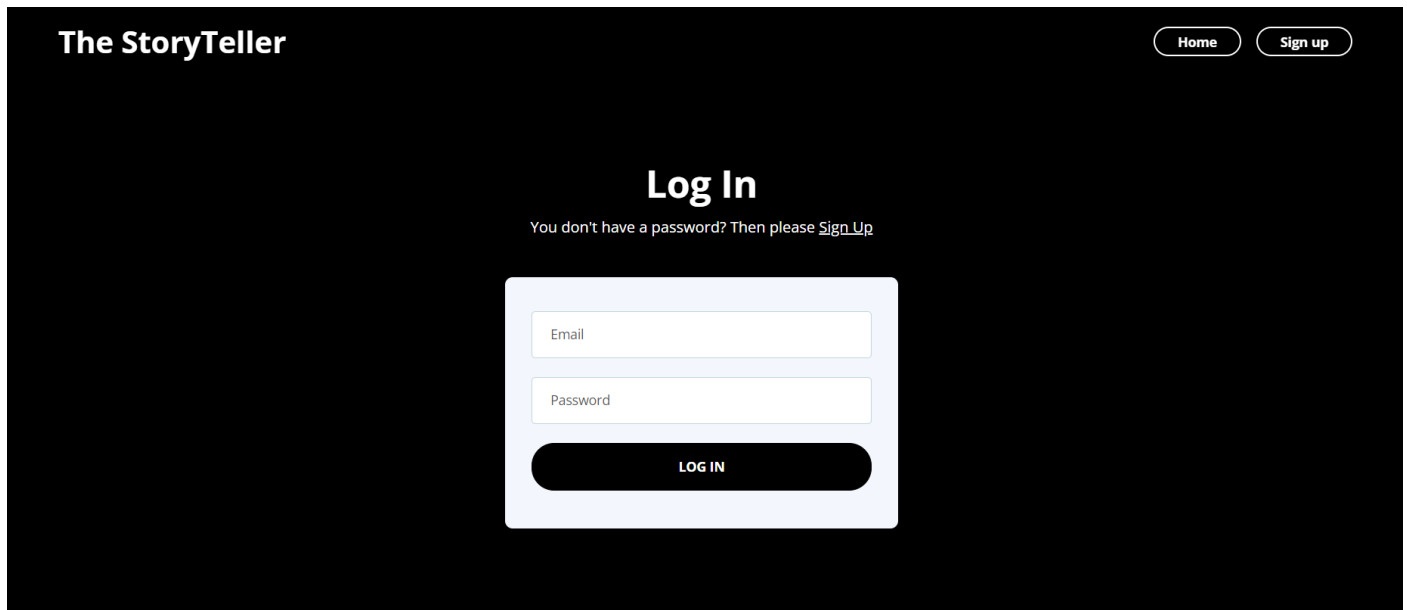
### The StoryTeller

[Home](#) [Features](#) [About](#) [Team](#) [Contact](#) [LOGIN](#)

#### TECHNOLOGIES



## Login Page:



The StoryTeller

Home Sign up

## Log In

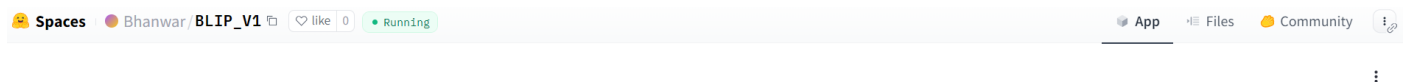
You don't have a password? Then please [Sign Up](#)

Email

Password

LOG IN

## Lobby Page:



Spaces | Bhanwar / BLIP\_V1 | like 0 | Running

App Files Community

## Video Processing

Upload a video

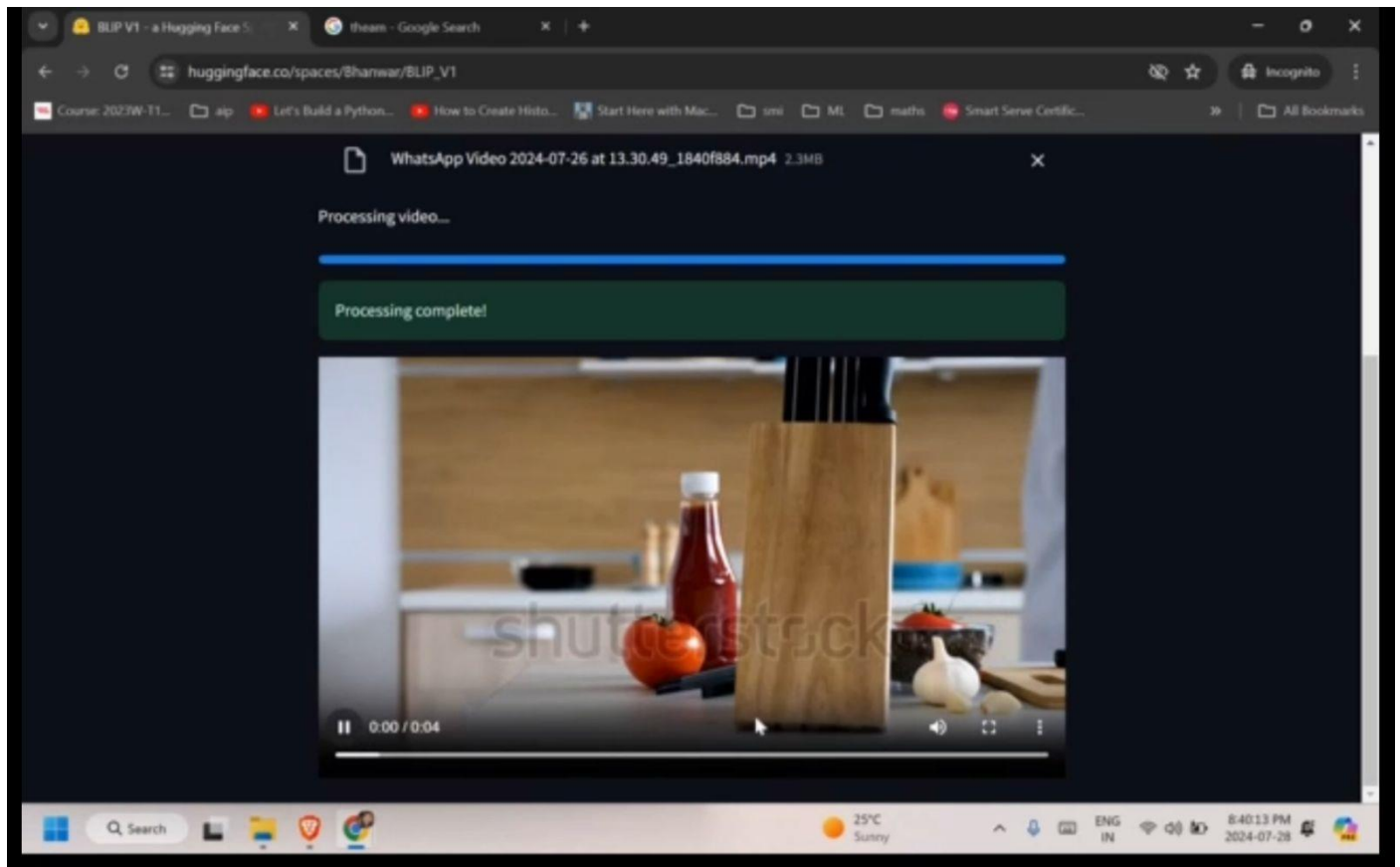


Drag and drop file here

Limit 200MB per file • MP4, AVI, MOV, MPEG4

Browse files

Output of video with audio description:





## Output of the generated descriptions and performance score

Out[39]:

	video_id	generated_description	description
0	10046243	ants are eating on the ground ants are eating ...	Ants eating dead insect
1	1005626710	a large orange tori tori tori tori tori tori t...	Kyoto,japan-sep 4,2017: timelapse of the visit...
2	1006641379	a man is seen in the middle of a flooded stree...	Circa 1940s - a film about copper mining and s...
3	1006733308	a bed with a pair of black pants and a white s...	Black kit classic menswear. men's accessories ...
4	1007094259	a woman doing yoga in the park a woman doing y...	Young asian woman yoga outdoors keep calm and ...
...	...	...	...
994	9643991	a close up of a curtain with a white backgroun...	White linen cloth on the wind
995	9709337	a person is laying in a hammol a person is lay...	Pan from person's legs resting on a hammock to...
996	9769340	aerial view of the hong skyline and the hong r...	Bridge to haeundae, south korea, wide shot, la...
997	9940922	a small dog is playing with a stick a small do...	Dog looking at camera turning head chewing bon...
998	9944846	a large white building with a large white roof...	Three grouped beach basket in the rising sun f...

Epoch	Training Loss	Validation Loss
1	0.891500	0.886428
2	0.775200	0.798386
3	0.854600	0.784755
4	0.650600	0.776618
5	0.583500	0.774487
6	0.707700	0.770232
7	0.744600	0.766309
8	0.810300	0.767289
9	0.738600	0.765439
10	0.739100	0.765402

```
In [16]: # Function to generate caption for a specific video
def generate_caption_for_video(video_path, model, feature_extractor, tokenizer):
    frames = extract_frames(video_path)
    caption = generate_caption(model, feature_extractor, tokenizer, frames)
    return caption

# Example usage:
video_path = 'videos_1000/1063125190.mp4' # Specify the path to the video
caption = generate_caption_for_video(video_path, model, feature_extractor, tokenizer)
print(f"Generated Caption: {caption}")
```

Generated Caption: a tray of oranges and a piece of bread

2	1007829319	Small boy feeding pigeons on the street	Little boy in red jacket and red pants is feeding pigeons in the park.	Little boy feeding pigeons in the park
3	1008414787	Aerial view beachfront destination usa picnic island blvd, tampa, fl 33616. indian rocks beach is a city in pinellas county, florida, united states.	Aerial view of a small island in the mediterranean sea.	Aerial view of a small island in the middle of the sea
4	1010428679	Aerial view dubrovnik old town in dalmatia, croatia - prominent travel destination of croatia. dubrovnik old town was listed as unesco world heritage sites in 1979.	Aerial view of kosteli town, croatia	Aerial view of the old town of kosteler, greece.
5	1011212219	Bat-eared fox resting on ground.	Kangaroo with white eyes and brown fur	Kangaroo in the wild
6	1011773219	Side view of a giant sci-fi interplanetary spaceship flying on neptune background, 3d animation. texture of planet was created in graphic editor without photos.	Spacecraft flying over a planet	Space station in space
7	1012764110	Bartender is stirring cocktails on the bar	Bartender making a cocktail at the bar.	Bartender pouring a cocktail in a glass
8	1013564174	Business people working hd animation	Close up of young young man talking to classmates in a classroom	Businessman talking to a woman at the office
9	1017093517	Felodipine - male doctor with mobile phone opens and touches hologram active ingredient of medicine	Doctor using a smartphone and a tablet.	Doctor using a tablet in a hospital
10	1017514603	Tokyo, japan - october 8th, 2018. first person point of view hyperlapse clip of walking in the ginza street.	Tokyo, japan - june 20, 2019: people walking on the crosswalk in	Tokyo, japan - september 16, 2018: people walking on the
11	1018200826	Sexy blond female posing for a camera on the wild beach	Beautiful young woman leaning against a tree trunk	Beautiful young woman in a bikini posing in a tree
12	1020794734	Lassen volcanic national park sulphur works area	Aerial view of a road in the mountains.	Aerial view of a mountain road in the background

## Conclusion

Investigating the causes of the VIT model's initial subpar performance requires a lot of work. This entails examining a number of factors, including the architecture, training set, and hyperparameters of the model, in order to find any possible weaknesses. Were the training datasets sufficiently representative and diversified, for instance? Was there an overfitting or underfitting of the model? Are there any particular hyperparameters or features of the model design that may be changed to enhance performance?

To make sure the VIT model is receiving and processing data as intended, we should also examine the complete pipeline, including preprocessing stages and integration with other system components. This analysis will assist in determining whether the problems resulted from integration issues, data quality concerns, or model limits.

Our Movie Descriptor project successfully enhances the movie experience for visually impaired individuals by providing clear and accurate audio descriptions that match the video. We've used advanced technology to create high-quality, easy-to-use descriptions and ensure they are perfectly synchronized with the video.

With the help of our Movie Descriptor project, visually impaired people may now enjoy a more engaging moviegoing experience as we provide precise audio descriptions that match the video. We have created a system that uses cutting-edge technology to produce clear, comprehensible descriptions, enabling visually challenged viewers to participate completely in and enjoy films. Extensive machine learning models combined with an intuitive interface ensure that explanations are not only accurate but also perfectly synchronised with the action on screen. This accomplishment shows how dedicated we are to enhancing the cinematic experience for those who depend on audio descriptions by making films more inclusive and accessible.

The project's objective of improving movie accessibility for those with visual impairments has been accomplished. We have created a product that offers comprehensive and synchronised audio descriptions by carefully choosing the right technology and giving user demands first priority. As a result, people with visual impairments can now enjoy a more inclusive and engaging cinematic experience than they might have with a previous inaccessible film. By combining sophisticated machine learning models with an intuitive interface, we were able to achieve our goals and provide a useful tool that will increase the enjoyment of films.

## References

- Microsoft. (2024). *Seeing AI: The app that narrates the world around you*. Retrieved from [Microsoft Seeing AI website](#)
- Perez-Martin J, Bustos B, Guimarães SJF, Sipiran I, Pérez J, Said GC. A comprehensive review of the Video-to-Text problem. arXiv.org. <https://arxiv.org/abs/2103.14785>. Published March 27, 2021.
- Li J, Li D, Xiong C, Hoi S. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. arXiv.org. <https://arxiv.org/abs/2201.12086>. Published January 28, 2022.
- Hugging Face – The AI community building the future. <https://huggingface.co/>.
- Streamlit • A faster way to build and share data apps. <https://streamlit.io/>.
- gTTS — gTTS documentation. <https://gtts.readthedocs.io/en/latest/>.
- BLIP. [https://huggingface.co/docs/transformers/en/model\\_doc/blip](https://huggingface.co/docs/transformers/en/model_doc/blip).
- Vision Transformer (ViT). <https://arxiv.org/abs/2010.11929>.
- Bidirectional and Auto-Regressive Transformers (BART). <https://arxiv.org/abs/1910.13461>.
- Americans with Disabilities Act (ADA). <https://www.ada.gov/>.
- Web Content Accessibility Guidelines (WCAG). <https://www.w3.org/WAI/standards-guidelines/wcag/>.
- Salesforce BLIP Model. <https://huggingface.co/salesforce/blip>.
- Regex101: Online regex tester and debugger. <https://regex101.com/>.

## **Recommendation**

### **Real-Time Descriptions**

Expand the system's functionality to provide live, in-the-moment event descriptions. This would improve accessibility in real-time circumstances by enabling users to hear instantaneous audio descriptions for live broadcasts, such as sporting events or live TV shows.

### **Customization Options**

Provide user-customizable parameters so they can adjust the audio descriptions to suit their tastes. To provide a customised and cosy viewing experience, users should be able to change the explanation pace, voice options, and detail level.

### **Multi-Language Support**

Add further language support to the system. This will accommodate a range of user demands and preferences and make the product available to a larger, worldwide audience by offering audio descriptions in multiple languages.

### **Advanced Synchronization**

Improve the sync between the video and audio descriptions, especially in sequences with rapid motion. It is vital to utilise enhanced algorithms and methodologies to guarantee that explanations maintain a seamless alignment with the on-screen action, particularly during dynamic periods.

### **Better Feedback System**

Provide a more reliable feedback system that will make it simple for users to offer their opinions and for us to thoroughly consider and act upon their recommendations. Detailed reporting capabilities, user-friendly feedback channels, and a methodical process for integrating user input into continuous product enhancements should all be part of this system.

These updates and changes will help ensure our Movie Descriptor remains a top choice for enhancing movie accessibility for visually impaired users.