Software Requirement Specification Document For WordReel

Mohammed Nasr, Ahmed Emad, Mohamed Said, Rana Ehab Dahshan
Supervised by: Dr.Fatma Helmy, Eng.Aly Mohammed

April 9, 2024

Table 1: Document version history

Version	Date	Reason for Change
1.0	13-jan-2024	SRS First version's specifications are defined.
2.0	05-apr-2024	Class diagram and usecase diagram are updated.

GitHub: https://github.com/MohhammedNasr/WordReel

Contents

Abstract

WordReel is built to narrow the gap between textual description and dynamic visuals. WordReel works by taking the long text input from the user and analysing it to extract keywords that are used as prompts to generate images from the dataset. Multiple images are used to generate the output video. or it will suggest scripts to the user, and the user will choose the one he likes, and the system will make a video of this script and voice over it. WordReel empowers content creators and helps them in various ways to create creative content without wasting time, the need for more resources, or any more efforts. WordReel also helps educators visualise the intended topic to make it easier to understand. WordReel serves as a bridge between technological innovation and creative expression by using AI, natural language processing (NLP), and computer vision to come up with new solutions.

1 Introduction

Artificial Intelligence has shaped the way media and content are perceived by users especially in the last few years. There have been great advancement in the development of a text-to-video AI system that's only increasing year after year. This research paper of WordReel explores how the implementation and functionalities of WordReel can enhance the user experience in content creation and visualization.

1.1 Purpose of this document

The goal of this document is to explain the relevance and applications of WordReel, an AI-driven text-to-video generator. The document not only emphasises the need of bridging the gap between verbal descriptions and dynamic graphics, but it also digs into the system's technical aspects, such as the class diagram. This detailed specification is intended to provide a clear overview of the WordReel project, including its objectives, scope, functional and non-functional requirements, design constraints, and data design. The inclusion of an initial class diagram and operational scenarios helps the document achieve its purpose of directing the development process from proposal to Software Design Document.

1.2 Scope of this document

This paper aims to provide a detailed description of the specifications and guidelines that will guide the creation of WordReel, an AI text-to-video generator. It provides developers, stakeholders, and consumers with a road map by outlining the project's goals, features, and limitations. The goal of the system, the corporate environment that influenced its development, and its possible uses in education and content creation are all covered in this document. It also sheds light on the non-functional requirements, data design considerations, design constraints, and preliminary object-oriented domain analysis of the project.

1.3 Business Context

The business demand for a text to video system can be seen in critical fields like content creation for generating creative solutions and leveling up the content being created. It can also be used in generating a Text to video systems are heavily needed in educational purposes to generate videos that can be used to teach people complex topics and visualizing the information.

2 Similar Systems

2.1 Academic

In [7]The authors designed Make-A-Video, which learns how the world moves from unsupervised video footage, with the goal of translating the incredible success made in Text-To-Image Models directly into Text-To-Video Models. Make-A-Video comes with three benefits: It accomplishes three things: (1) it expedites T2V model training (saving it from having to learn visual and multimodal representations from scratch); (2) it eliminates the necessity for paired text-video data; and (3) the resulting movies retain the breadth of current image generation models. The components that make up the Make-A-Video model are: a base T2I [4] model trained on text-image pairs; spatiotemporal convolution and attention layers that extend the temporal dimension of the networks' building blocks; and spatiotemporal networks that comprise both spatiotemporal layers and an additional essential component. a frame interpolation network for high frame rate generation intended for T2V generation The authors utilised a 2.3B subset of the English-text dataset from Schuhmann et al [6]. to train the image models. Examples of couples containing NSFW photos are eliminated. two harmful terms in the text, or pictures with a greater than 0.5 for the watermark probability. WebVid-10M [2]is what they utilitzed

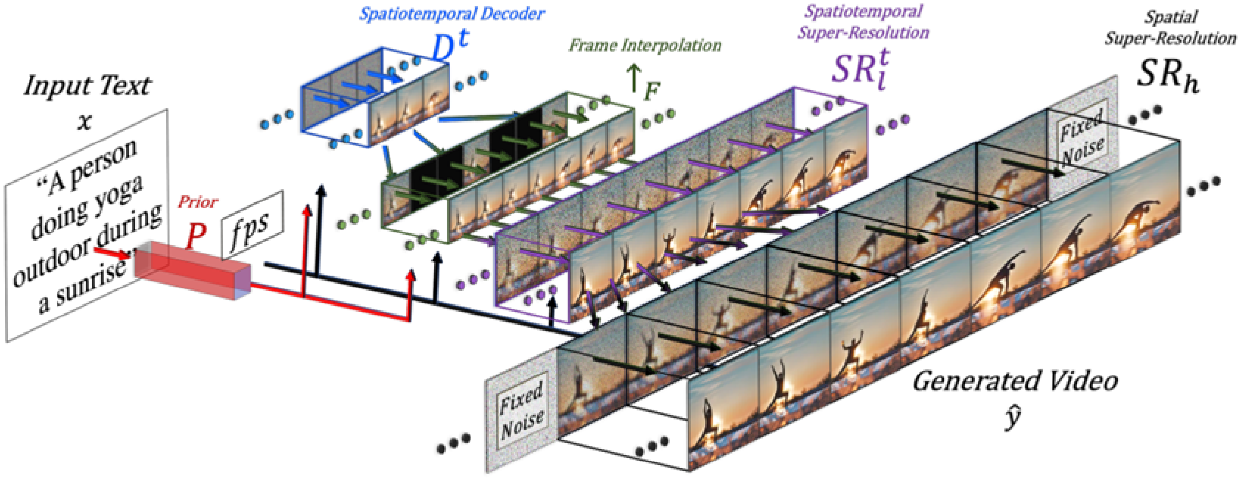


Figure 1: Make-A-Video Ai Model.

The authors of [10]discussed the difficulties faced by text-to-video models, including variable video lengths, limited high-quality text-video data, and computational costs. In order to overcome these problems, they developed a new model for learning video representation that reduces the video to

a compact illustration of distinct tokens. This tokenizer can handle variable-length films because it makes use of causal attention in time. They are utilizing a bidirectional masked transformer

[9] conditioned on pre-computed text tokens to produce video tokens from text. To construct the actual film, the created video tokens are de-tokenized. to deal with data problems. In contrast to earlier techniques for creating videos, Phenaki has the ability to create any length of video based on a series of prompts.

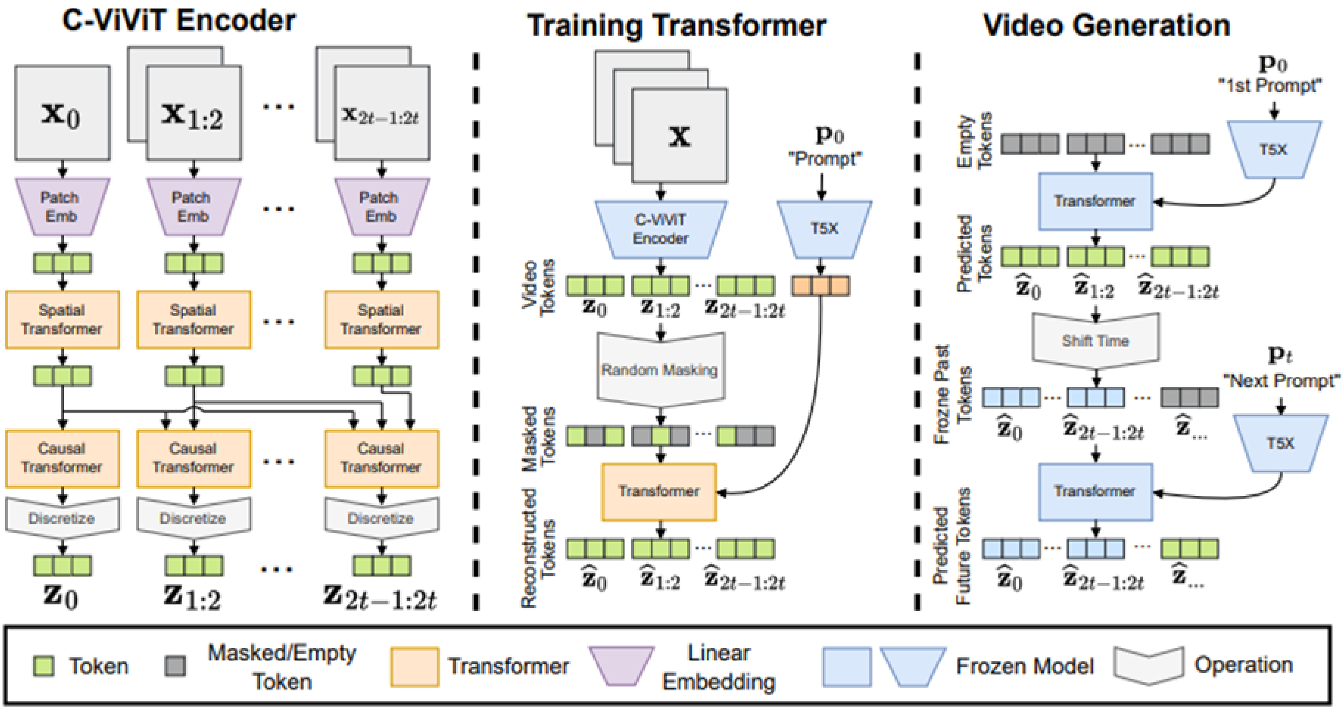


Figure 2: Phenaki Ai Model.

The frame discontinuity issue and text-free creation strategies make it difficult to adapt current video generation techniques to effectively complete this mission. In order to address these concerns The generative adversarial network with recurrent deconvolution was proposed by the authors. The generator (RDN) and discriminator (3D-CNN) sub-networks make up the generative adversarial network (GAN)[3]known as RD-GAN [11]. Gaussian distribution noise and text embedding are used by the generator to create realistic movies, and the discriminator divides videos into "real" and "fake" classes. Through a non-cooperative game, both sub-networks receive training. Until the generator improves to the point where it can produce realistic videos The utilized dataset is Because of its unusual data distribution, the UCF-101 dataset [8]—which consists of 13,320 recordings representing 101 human actions—is utilized to train a model.Videos are split into 16 consecutive frames in order to increase the dataset's size and provide a more thorough depiction.

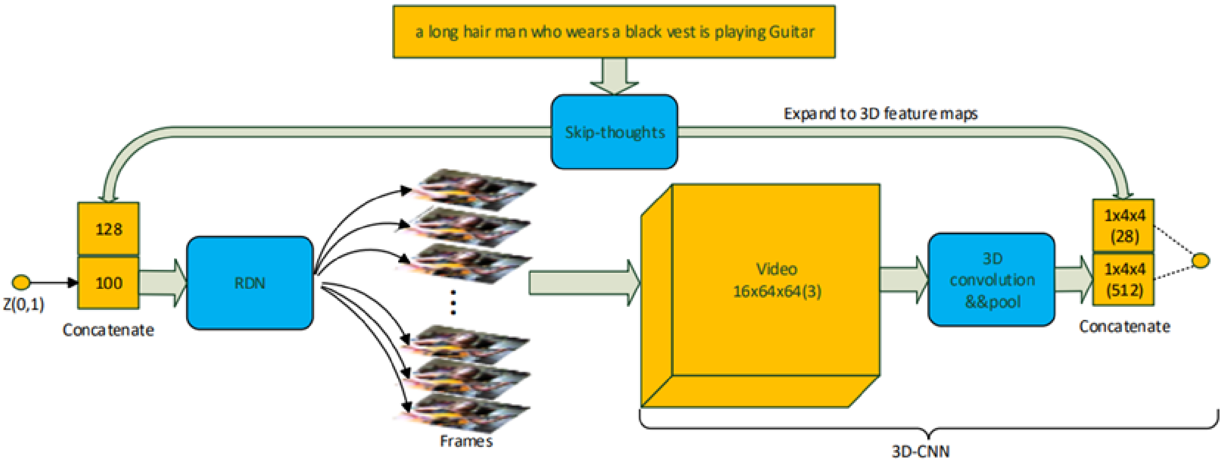
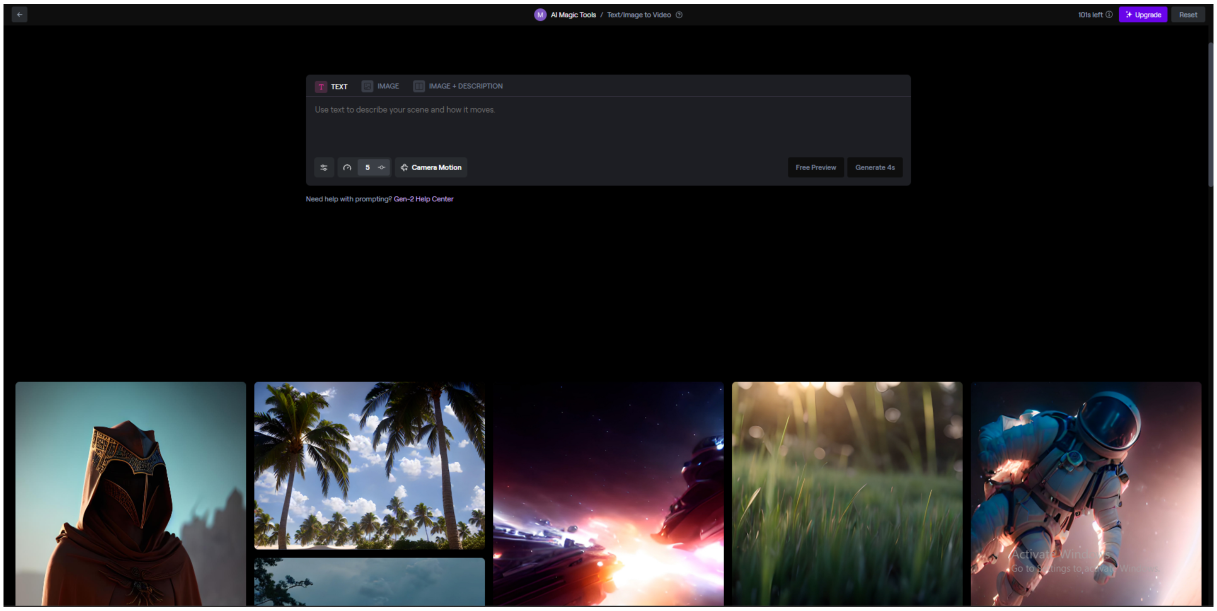


Figure 3: RD-GAN Ai Model.

2.2 Business Applications

Runway[5] was founded by artists on a mission to bring the unlimited creative potential of AI to everyone, they provide Generate videos using text, images or video clips. Generate compelling images with nothing but your words. Endlessly expand any image with simple text prompts. Instantly remix the style and com- position of any image.



Use text to describe your scene and how it moves.

5

Need help with prompting? Gen-2 Help Center

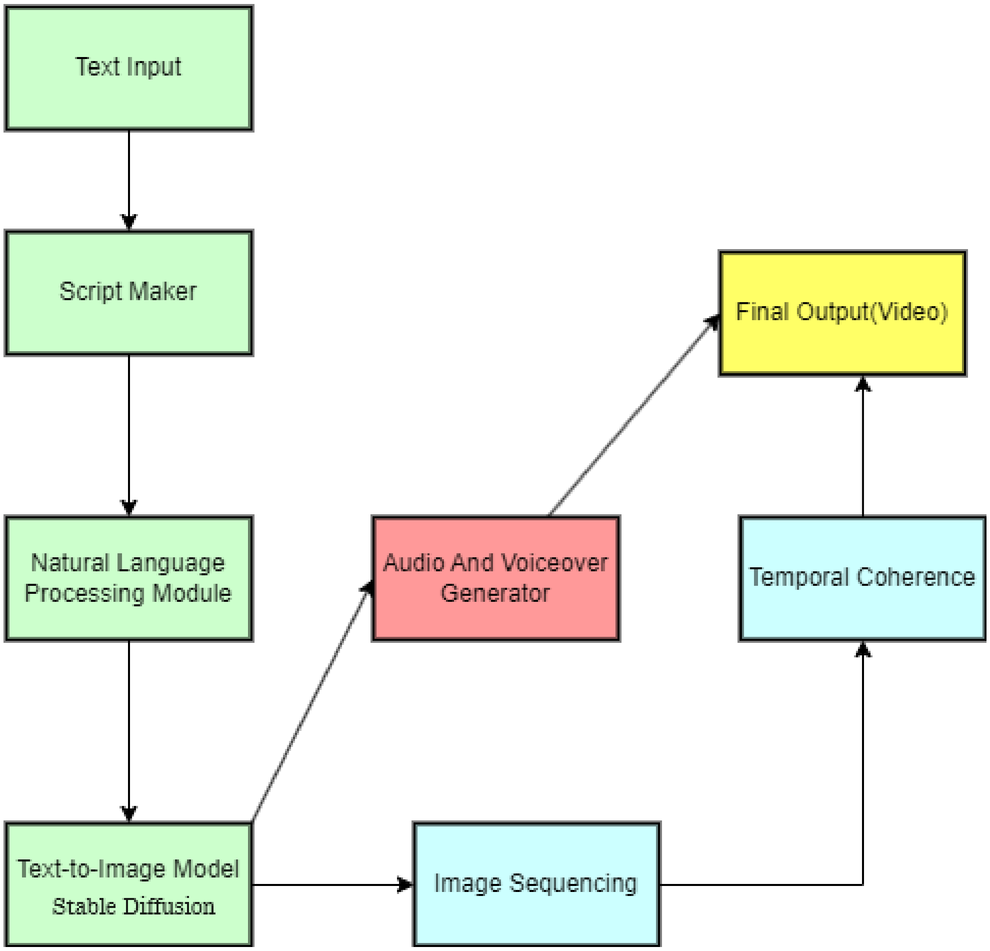
Figure 4: Runway website.

3 System Description

3.1 Problem Statement

Our problem when increasing video duration is to keep the quality of the video constant and produce it fast, as making 3 minutes and more generated videos is a really high GPU-intensive process that can cost a lot of time and money, and the current models only produce 2 minutes of generated videos.

3.2 System Overview



3.3 System Scope

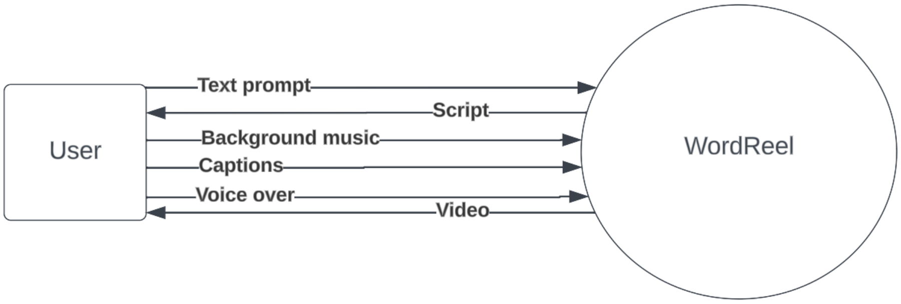
Our platform will contain the following features:

- Script Generator: Transforms user input into video scripts.
- Video Generator: Creates videos, including long, short videos and stories based on user provided text or scripts.
- Script Video Editor: Enables users to edit video content through text inputs.

- Voice Over: Allowing users to have a voice over on the videos.
- Background Music: Library of free music/sounds to be used on the videos.
- Captions: Automatically generates on video captions.

3.4 System Context

The WordReel system operates in a dynamic environment, bridging the gap between textual descriptions and dynamic graphics using AI-powered text-to-video production capabilities. WordReel, which sits at the crossroads of technological innovation and artistic expression, provides a one-of-a-kind solution through the use of AI, natural language processing (NLP), and computer vision. It enables content creators to translate text into entertaining videos, making content creation more efficient and accessible. WordReel also caters to educational purposes, helping to visualise complicated ideas. The system's functionality includes script production, video creation, script editing, voice-over, background music integration, and automatic caption development, giving customers a versatile toolkit.



3.5 Objectives

3.5.1 Increase the Length of the Video:

- **Objective:** We will increase the duration of the video from the current duration that exists.
- **Specific:** Increase video length from the current average to a minimum of three minutes.
- **Measurable:** Measure the duration of the video.
- **Assignable:** Each developer is responsible for implementing features to extend the video.
- **Realistic:** Extending video length aligns with current technological capabilities.
- **Time-related:** Achieve the goal by the end of the development cycle.

3.5.2 Implement Advanced Edits for Scripts:

- **Objective:** Integrate advanced script editing features into WordReel, enabling users to add or delete words easily.
- **Specific:** Implement user-friendly script editing functionalities.
- **Measurable:** Track user satisfaction through feedback or surveys.
- **Assignable:** Developers will implement advanced script editing features.
- **Realistic:** Implementing advanced script editing features is realistic with current technology.
- **Time-related:** Achieve a 100% satisfaction rate with script editing.

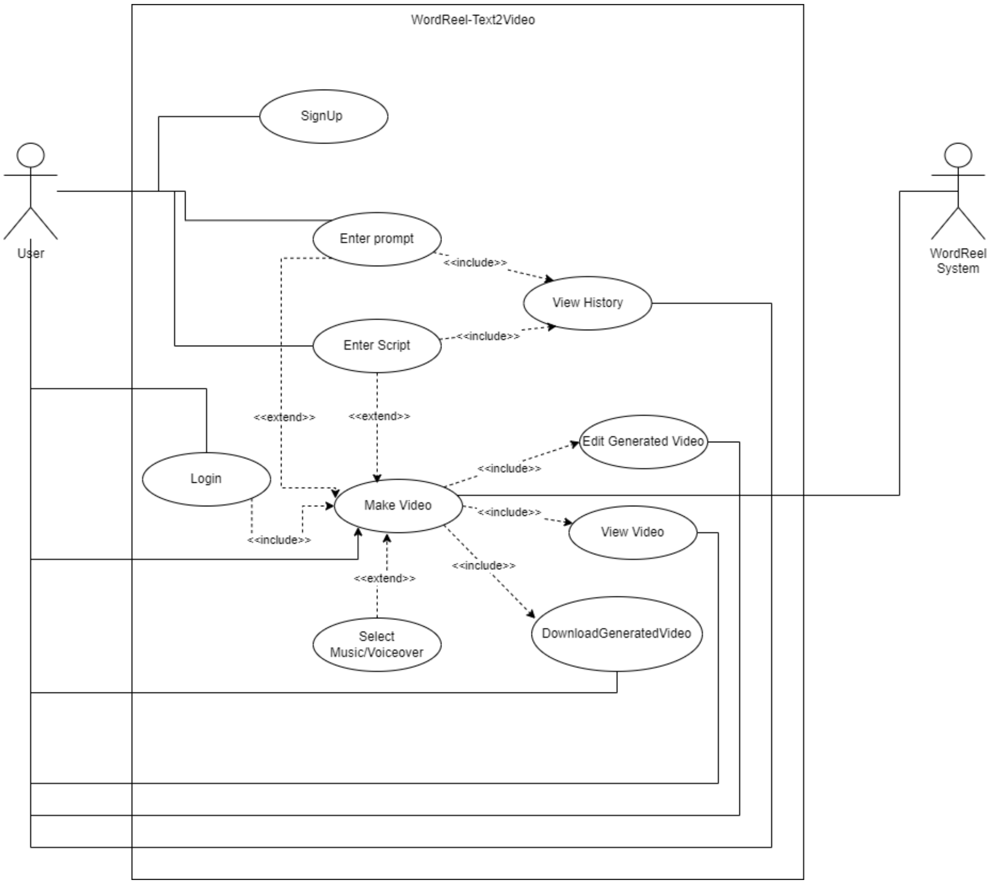
3.6 User Characteristics

WordReel is expected to be used by schools and content developers looking for a more efficient way to turn written material into captivating films. Social media influencers and marketing experts alike can profit from the system's capacity to generate aesthetically pleasing material. Alternatively, teachers can use **WordReel** to improve their lesson plans by turning textual information into visual aids that will engage students and increase learning.

- It is anticipated that users would possess differing levels of technical expertise,
- the system is intended to be easily navigable for a wide range of users with different age.

4 Functional Requirements

4.1 System Functions



1. The user shall be able to Sign Up on the system.
2. the user shall be able to login to the system using his account.
3. the user shall be able to enter a simple prompt to the system to create a small video .
4. the user shall be able to enter a script to the system to generate a long video or film.
5. the user shall be able to view history of the input prompt or script.
6. the user shall be able to create a video from the history of the input prompt or script.
7. the user shall be able to select the voice over of the characters and the music of the video.
8. the user shall be able to view the generated video
9. the user shall be able to edit on the generated video.
10. the user shall be able to edit his account.
11. the user shall be able to remove his account from the system.

12. the user shall be able to logout of the system.

4.2 Detailed Functional Specification

Name	User Authentication and Account Management
Code	UAM
Priority	Critical
Description	Handles user authentication and account management functionalities.
Input	User-provided username, password, recovery information
Output	User account status, authentication tokens
Pre-condition	User registration
Post-condition	Successful user authentication
Dependency	None
Risk	Unauthorized access, security vulnerabilities

Name	User Dashboard
Code	UD
Priority	High
Description	User-friendly dashboard displaying created videos, drafts, settings, and activity log.
Input	User interaction with the dashboard
Output	Display of created videos, drafts, settings, and activity log
Pre-condition	User authentication
Post-condition	Updated user dashboard reflecting recent activities
Dependency	UAM
Risk	Data inconsistency, unauthorized access to the dashboard

Name	Output Options
Code	OO
Priority	High
Description	Provides users with options to download the generated video in common formats.
Input	User selection of download format
Output	Downloaded video in the chosen format
Pre-condition	Successfully generated video
Post-condition	User downloads the selected video
Dependency	Real-time Preview and Editing
Risk	Compatibility issues with download formats

5 Design Constraints

The Design restrictions of WordReel exist in the demand of real-time processing. Computational resources are heavily requested for intricate visualizations. Complexity and Efficiency are crucial.

Name	Real-time Preview and Editing
Code	RTE
Priority	High
Description	Provides a real-time preview of the generated video for users to review and edit.
Input	User-requested edits, modifications
Output	Updated video preview after edits
Pre-condition	Successfully generated video
Post-condition	User saves edited video
Dependency	OO
Risk	Editing errors, system performance issues

Name	Text Input Handling
Code	TIH
Priority	High
Description	System accepts plain text input in multiple languages and formats.
Input	Text input in different languages and formats
Output	Processed and standardized text input
Pre-condition	User provides text input
Post-condition	Text input is accepted and ready for analysis
Dependency	TAU
Risk	Language compatibility issues, input format inconsistencies

Name	Text Analysis and Understanding
Code	TAU
Priority	Critical
Description	AI model comprehends the semantics and context of input text, extracts entities, sentiment, and keywords.
Input	Processed text input from TIH
Output	Extracted entities, sentiment analysis, key keywords
Pre-condition	Valid and processed text input
Post-condition	Successful extraction of entities, sentiment, and keywords
Dependency	None
Risk	Misinterpretation of context, inaccurate analysis

The integration of diverse media elements necessitates managing large datasets and intricate algorithms, emphasizing the need for robust memory handling are also a mandatory. The user input variability demands adaptability from the system to interpret diverse writing styles and intents. The System user interface UI must be intuitive to be adequate for users of different technical backgrounds and proficiency. The system must follow ethical constraints and privacy standards through the implementation of data handling practices.

5.1 Standards Compliance

WordReel respects industry guidelines and especially usability and accessibility. WordReel meets performance benchmarks for its real-time data processing and embracing user interoperability standards to allow seamless integration with external tools, APIs and platforms. Accessibility standards are adhered to maintain industry best practices through ensuring inclusive experience and catering to diverse user needs. The system supports regular updates and undergoes maintenance to adhere guarantee sustained functionality. Prevailing AI ethics is applied to promote trust, reliability of the system and security. The system also promotes responsible AI use, transparency and fairness in content representation and creation. WordReel aligns with data protection regulations to ensure user privacy, consent and data protection.

5.2 Hardware Limitations

WordReel hardware limitations are tied up to computationalintensity. Text to video generations requires substantial processing power, advanced graphic processing units GPUs or even tensor processing units TPUs. Memory constraints might arise due to handling large datasets and complex algorithms. All of this contributes to the responsiveness of the system. Low-end devices users might encounter some serious system lags. In addition to all of that, storage capacity is crucial to handle extensive data libraries. These limitations highlight the importance of functionality balancing and hardware capabilities to ensure optimal performance.

5.3 Other Constraints as appropriate

Other constraints of WordReel include bandwidth limitations that might arise when accessing external APIs or cloud services. Copyrights legal constraints related to licensing agreements in music, images or other media incorporated in the outputted video require meticulous adherence. Cross Platform testing should be considered to make sure the system is compatible with different browsers and devices. Ongoing AI advancements encourage constant periodic updates so that the system is up to date and sustainable.

6 Non-functional Requirements

Non-functional Requirements	Quality Attribute
The system should generate a 5-Response time for user interactions with the UI should be less than 1 second	Performance
The system must support 100 concurrent users without significant performance degradation, and its architecture must be scalable to handle increased text prompts and user requests.	Scalability
The system should have a 99.9% uptime and be capable of recovering within 5 minutes in case of failure.	Reliability
The system should be available 24/7, except during scheduled maintenance windows. Backup and disaster recovery mechanisms should be in place to ensure data integrity	Availability
User data, including text prompts and videos, should be encrypted both in transit and at rest, and access to sensitive services should be restricted based on user roles and permissions.	Security
The system should facilitate user access, facilitate easy updates and maintenance, ensure well-documented code, and establish a knowledge transfer plan for developers and support staff.	Maintainability
The user interface should be intuitive and user-friendly, requiring minimal training for users to operate. The system should comply with accessibility standards to ensure usability for users with disabilities	Usability
The system should comply with relevant data protection regulations (e.g., GDPR) and industry standards. Licensing and usage of external APIs should adhere to legal and ethical standards	Compliance
The system should have robust monitoring tools in place to track resource utilization, identify performance bottlenecks, and generate logs for auditing purposes. Alerts should be set up to notify administrators of abnormal system behavior or performance issues	Performance Monitoring
Adaptability focuses on the system's capability to evolve and accommodate changes in technology, user requirements, or business needs without major disruptions. This ensures the system's longevity and relevance over time	Adaptability

Table 2: Non-functional Requirements and Quality Attributes

7 Data Design

Dataset

We are using Stable Diffusion model that is trained on pairs of images and captions taken from LAION-5B [6], where 5 billion image-text pairs were classified based on language and filtered into separate datasets by resolution

Database

We are going to save user's scripts as form of chats so users can access it later and modify it if needed for a modified output, for saving scripts and having them connected with user's google account we are going to use Firebase.

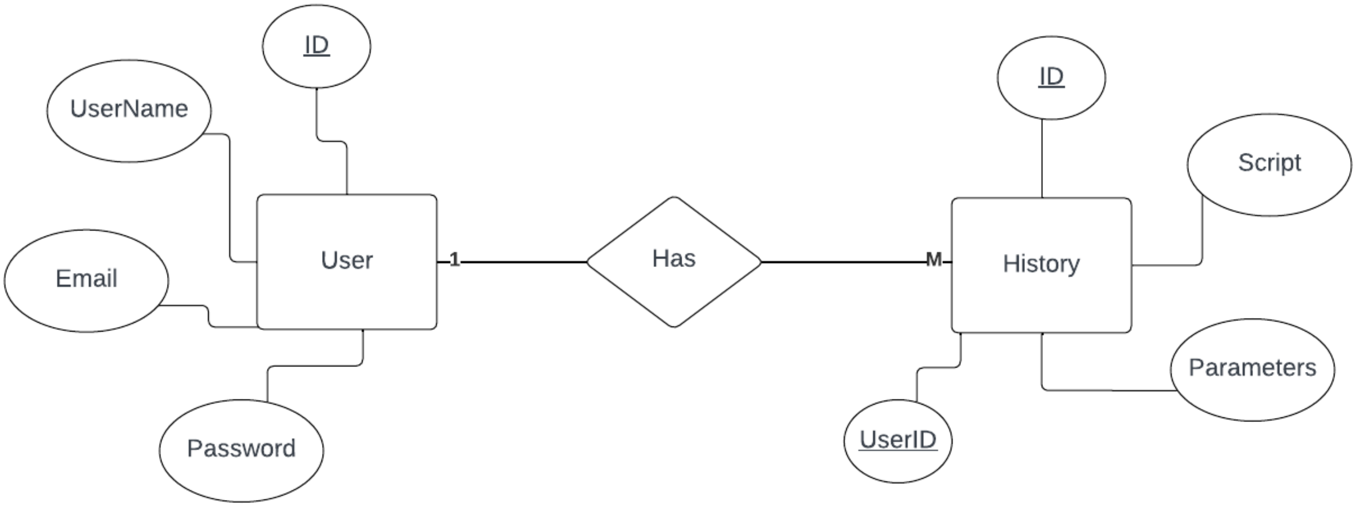
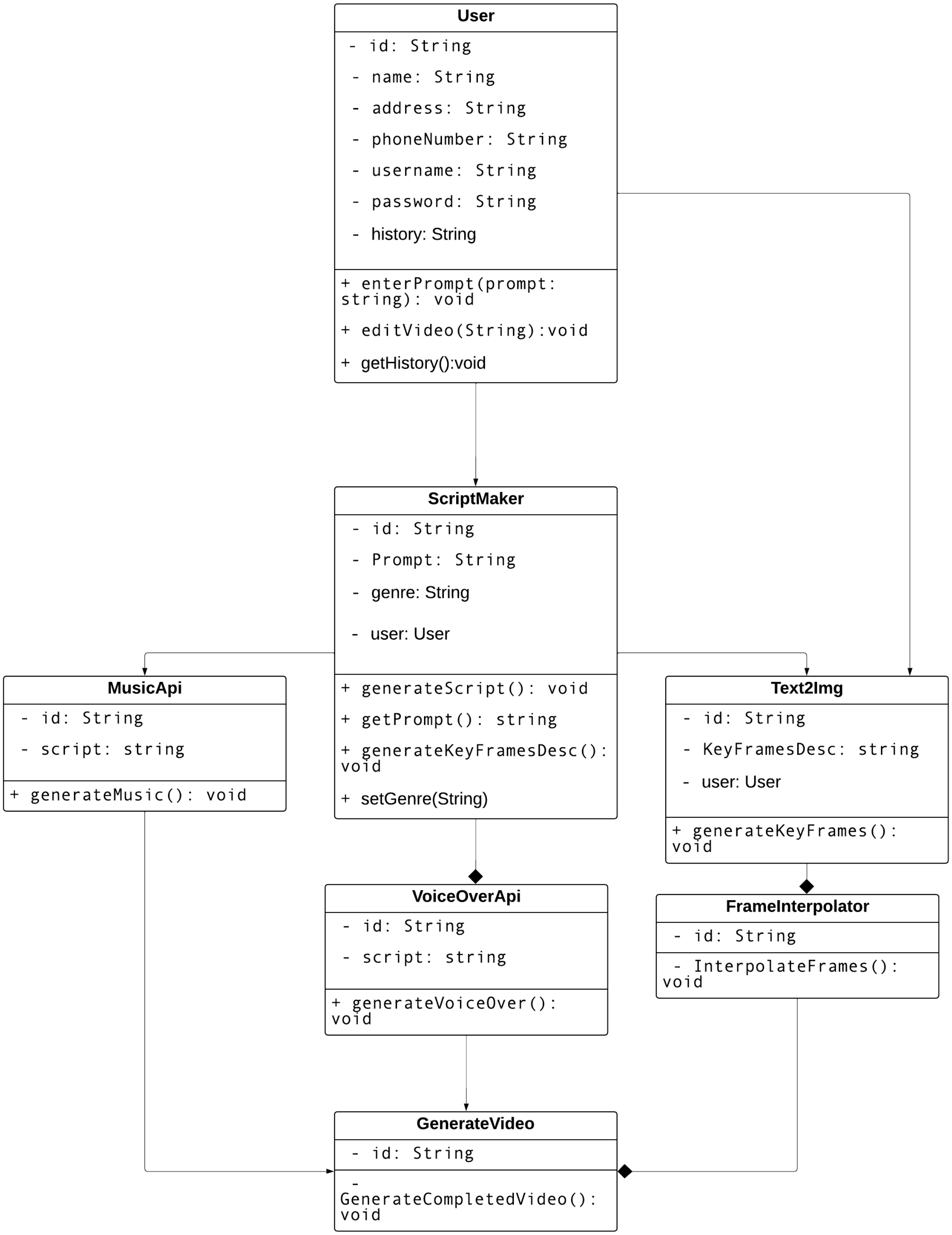


Figure 5: Entity Relationship Diagram.

8 Preliminary Object-Oriented Domain Analysis



9 Operational Scenarios

1. **Scenario 1:** Users can create their animated stories or videos using a text-based interface, by entering the video prompt or using our script generator.
2. **Scenario 2:** Users can use our script maker to generate a script that can be used to make a long story video, the script also can be saved to be used later.
3. **Scenario 3:** The user can select to add voice over for characters dialogue, captions and background music/sounds to make the story more entertaining and give it some life.
4. **Scenario 4:** Users edit their story using the script tool. They change dialogues, scenes, and how the story goes. The video updates as they edit the script.

10 Project Plan

Task	**Start date**	**End date**	**Duration**
Ideas and Supervisor	20/08/2023	15/09/2023	26 Days
Information collection and research	15/09/2023	15/10/2023	30 Days
Survey and proposal preparation	1/11/2023	15/11/2023	15 Days
SRS preparation	15/11/2023	20/12/2023	35 Days
SRS presentation	20/12/2023	25/12/2023	5 Days
SDD preparation	15/01/2024	20/02/2024	35 Days
SDD presentation	20/02/2024	25/02/2024	5 Days
Website development	1/12/2023	1/01/2024	31 Days
Platform development	1/12/2023	10/04/2024	131 Days
Prototype	1/12/2023	10/01/2024	40 Days
Testing and validating	1/03/2023	1/04/2024	31 Days
Thesis	25/06/2024	28/06/2024	3 Days

Table 3: Task and Time Plan

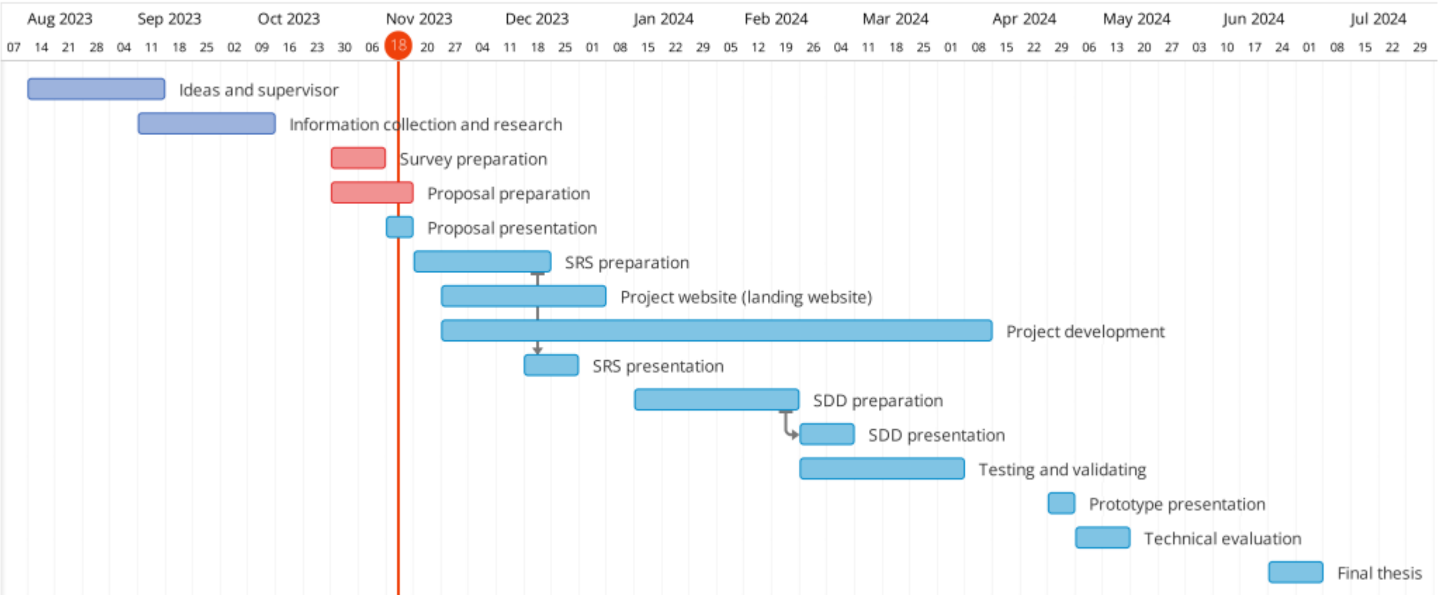


Figure 6 : Gantt chart

11 Appendices

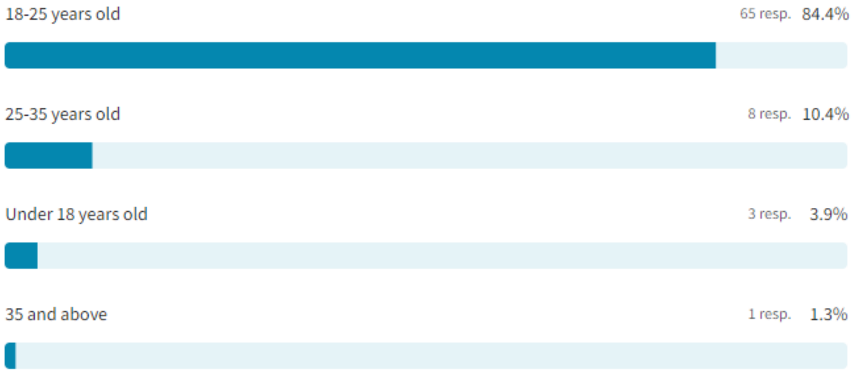
11.1 Definitions, Acronyms, Abbreviations

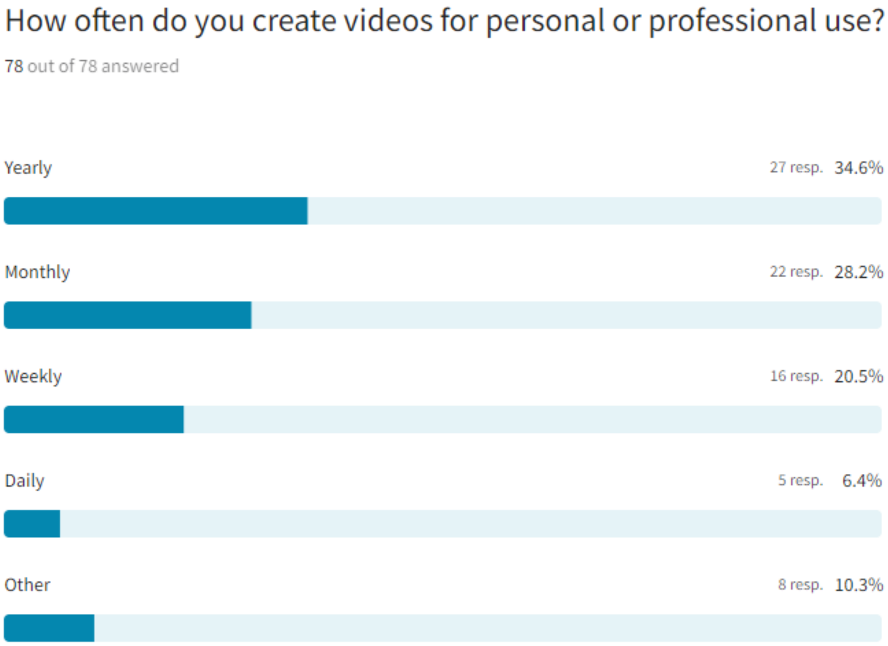
Term	Stands For
AI	Artificial Intelligence
NLP	Natural Language Processing
T2V Model	Text-To-Video Model
NSFW	Not Safe For Work
SVM	Support Vector Machine
GAN	Generative Adversarial Networks
Tokenizer	Tool/algorithm breaking down text into tokens
T2I Model	Text-To-Image Model
API	Application programming interface
GDPR	General Data Protection Regulation

11.2 Supportive Documents

Survey

We conducted a survey using Typeform[1]



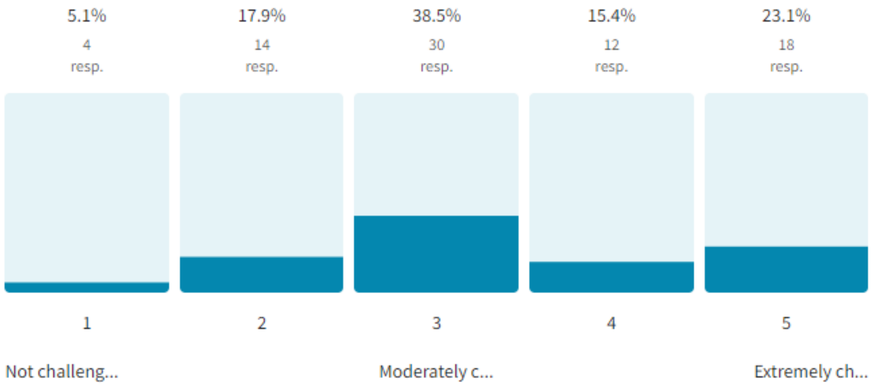

How often do you create videos for personal or professional use?

78 out of 78 answered

how challenging do you find the process of creating a video?

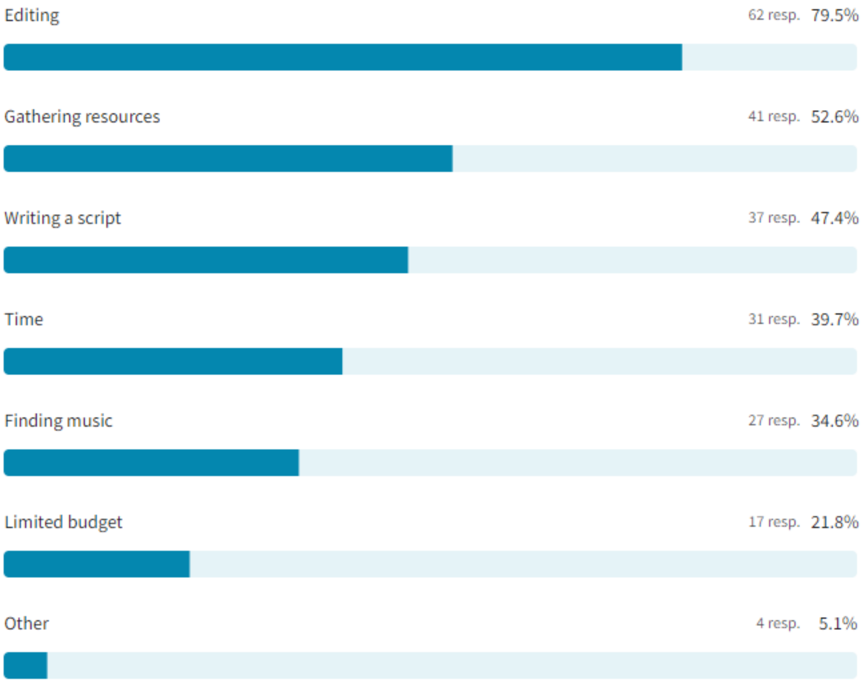
78 out of 78 answered

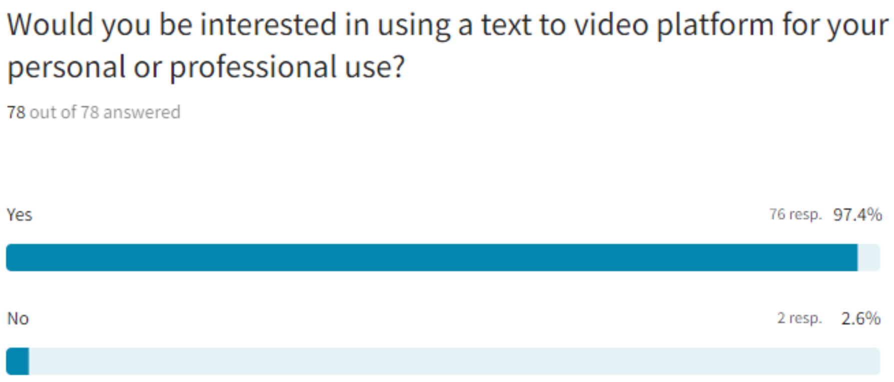
3.3 Average rating



What are the challenges you face when creating a video

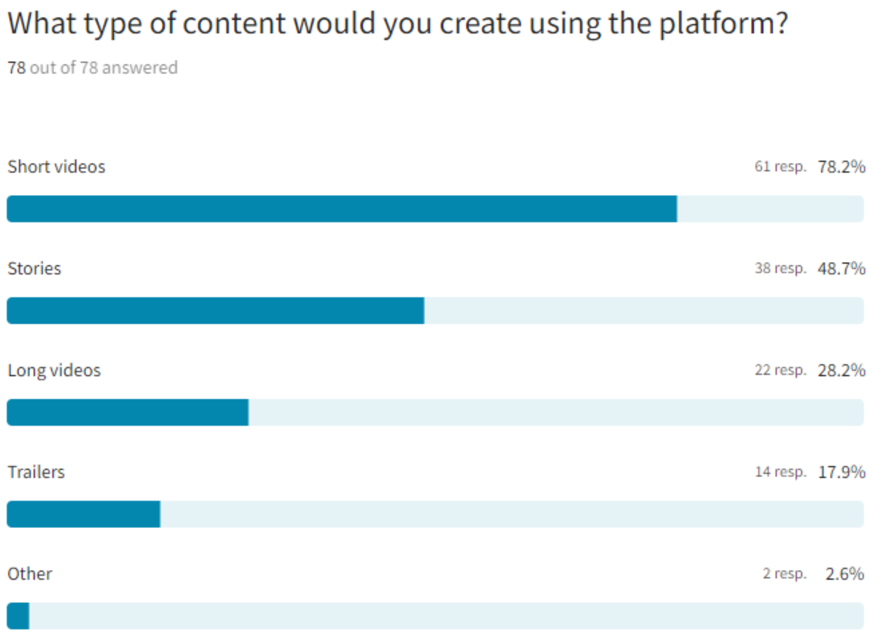
78 out of 78 answered





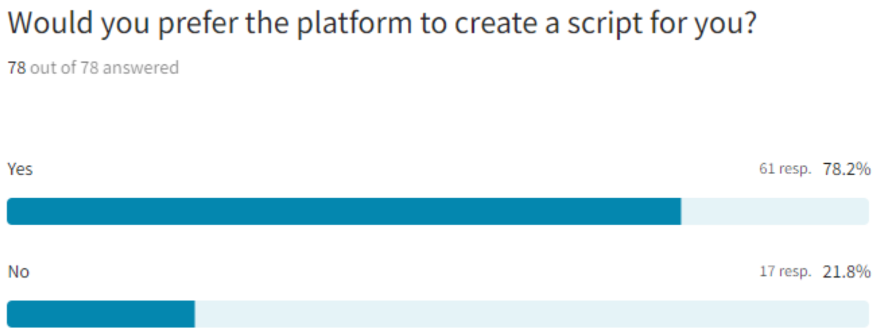
Would you be interested in using a text to video platform for your personal or professional use?

78 out of 78 answered



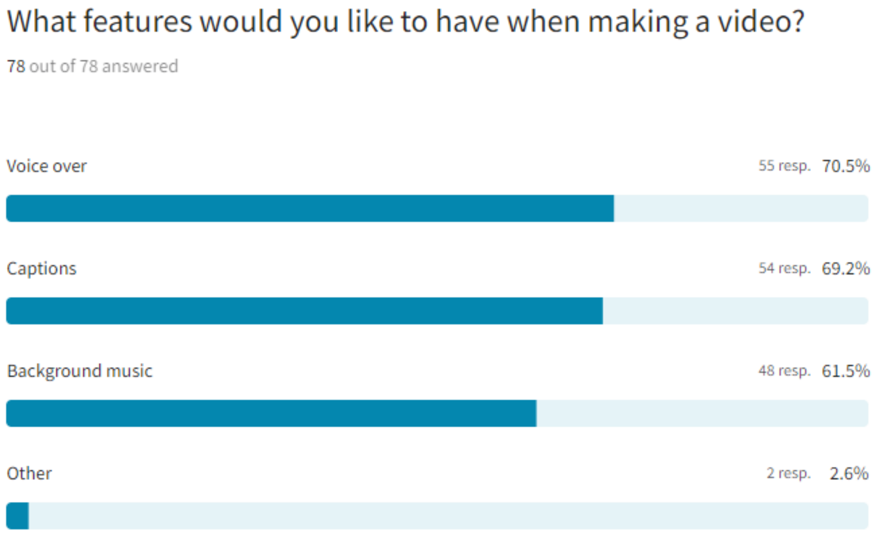
What type of content would you create using the platform?

78 out of 78 answered



Would you prefer the platform to create a script for you?

78 out of 78 answered



What features would you like to have when making a video?

78 out of 78 answered

References

[1] Typeform. [Online; accessed on 14/11/2023].

[2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*, 2021.

[3] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[4] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR, 2016.

[5] RunwayML Team. RunwayML Video Tools, 2023. [Online; accessed on 14/11/2023].

[6] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022.

[7] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

[8] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild, 2012.

[9] Richard E. Turner. An introduction to transformers, 2023.

[10] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description, 2022.

[11] Hongyuan Yu, Yan Huang, Lihong Pi, and Liang Wang. Recurrent deconvolutional generative adversarial networks with application to text guided video generation. *arXiv preprint arXiv:2008.05856*, 2020.