



Animatic Vision: AI Powered Story to Animation

Final Year Project Proposal
by

Muhammad Abdullah (2280214)
Atta-ur-Rehman (2280138)
Muhammad Humam Khan (2180125)

Supervised by: Ms. Tehreem Saboor

Faculty of Computing and Engineering Sciences
Shaheed Zulfikar Ali Bhutto Institute of Science and Technology
Islamabad, Pakistan

Spring 2025

Revision History

Compiled By	Checked By	Date	Reason for Change	Version
Atta-ur-Rehman	Ms. Tehreem Saboor	10th Sep 2025	Initial Version	1.0

Project Description

AnimaticVission is an AI-powered story-to-animation platform designed to transform the way digital content is created by converting written narratives into fully realized animated sequences. Leveraging large language models (LLMs) and LangGraph, the platform enables users to generate lifelike, lip-synced avatars that speak naturally, produce high-quality AI-generated images, and convert them into dynamic video sequences with minimal manual effort. By integrating natural language processing, image synthesis, and video generation into a unified workflow, AnimaticVission significantly reduces the time, complexity, and technical expertise typically required in traditional animation pipelines. The platform caters to a wide range of creators: filmmakers can visualize scripts and bring storyboards to life, educators can design engaging lessons, marketers can produce personalized content, and storytellers can experiment freely with narrative ideas. With a focus on precision, style, and creative freedom, AnimaticVission democratizes animation and empowers creators to bring imagination to life efficiently and effectively.

1 Introduction

Text-to-video (T2V) generation has made significant progress in recent years. However, generating videos that accurately depict multiple objects, attributes, and motions in complex and dynamic scenes based on fine-grained text prompts remains a challenging task. In this work, we aim to conduct a systematic study on compositional T2V [1].

WP3Compositional text-to-image (T2I) generation, which aims to compose multiple objects, attributes, and their relationships into complex scenes, has been widely studied in previous methods [2]. Benchmarks for compositional T2I generation have been accepted as an important evaluation dimension for T2I foundation models. However, most works on T2V generation focus on generating videos with simple text prompts, neglecting the significance of compositional T2V generation. Moreover, existing video generation benchmarks [3] primarily evaluate video quality, motion quality, and text-video alignment with single-object text prompts, and benchmarks for compositional T2V generation have not been systematically and extensively investigated in previous literatures.

AnimaticVission is an AI-powered story-to-animation platform that transforms written text into lifelike, lip-synced avatars and dynamic video sequences. By combining LLMs, LangGraph, natural language processing, and image synthesis, it simplifies traditional animation workflows. The platform enables filmmakers, educators, and marketers to quickly visualize scripts, create engaging lessons, and produce personalized content. AnimaticVission empowers creators to bring ideas to life efficiently, with precision, style, and creative freedom.

2 Application/Literature Review

Wav2Lip-HD system combines Wav2Lip for precise lip-syncing with Real-ESRGAN for super-resolution. It synchronizes lip movements with audio while enhancing video quality, processing frames individually and recombining them with the original audio.

Features:

- High-precision lip-syncing that accurately matches audio to video.
- Super-resolution of frames using Real-ESRGAN for sharper visuals.
- Supports various video resolutions and formats.
- Can process pre-recorded videos or generated animations.
- Python-based workflow compatible with GPU acceleration (CUDA).

Limitations:

- Requires significant GPU resources for real-time or high-resolution videos.
- Longer videos require more processing time due to frame-by-frame enhancement.
- Lip-sync quality may degrade if input audio is noisy or unclear.
- No automated scene generation; works only on provided video clips.
- Limited integration with end-to-end animation pipelines without additional scripting.

MuseTalk is a real-time, high-fidelity lip-syncing model capable of generating accurate lip movements from audio at 30+ FPS on an NVIDIA Tesla V100. It modifies the face region of a video (256×256) to match input audio, supporting multiple languages. MuseTalk uses spatio-temporal sampling, GAN, perceptual, and sync losses to maintain visual quality, identity consistency, and precise lip-sync accuracy. It can be applied to generated or real videos as part of a virtual human solution.

Features:

- Real-time inference at 30+ FPS.
- Audio-driven lip-sync for multiple languages (Chinese, English, Japanese).
- High-quality output with enhanced clarity and identity preservation.
- Supports face region modification to improve generation results.

- Open-source training and inference code with pre-trained checkpoints.

Limitations:

- Requires a high-performance GPU (e.g., NVIDIA Tesla V100).
- Limited to face region resolution (256×256).
- Complex setups for training new models from scratch.
- Real-time performance may degrade on longer videos or lower-end hardware.
- Works primarily for single-face regions, not multi-person scenes.

FaceFusion is an AI-powered system for face swapping and blending, enabling realistic merging of facial features between source and target images or videos. It uses deep learning models to align, warp, and blend facial regions while preserving expressions, skin tone, and identity features. FaceFusion is widely applied in video editing, virtual avatars, entertainment, and creative content generation.

Features:

- Realistic face swapping with preserved expressions and identity.
- Works on both images and video frames.
- High-quality blending to maintain skin tone and lighting consistency.
- Supports multiple faces in a scene with selective swapping.
- Compatible with GPU acceleration for faster processing.

Limitations:

- Requires good quality source and target images/videos for realistic results.
- Processing high-resolution videos can be time-consuming.
- May fail under extreme facial poses or occlusions.
- Limited support for dynamic lighting changes across frames.
- Ethical and privacy considerations when used for deepfakes.

Table 1: Applications Comparison

<i>Features</i>	Applications			
	Wav2Lip-HD	MuseTalk	FaceFusion	Proposed System
Text / Story Input	✗	✗	✗	✓
Scene Breakdown	✓	✓	✓	✓
Character Generation	✗	✗	✗	✗
Background GenerationAudio Narration (TTS)	✗	✗	✗	✓
Audio Narration (TTS)	✓	✓	✓	✓
Web Interface	✗	✗	✗	✓
Face Swapping / Lip-Sync	✓	✓	✓	✓

3 Problem Statement

Traditional animation is a highly time-consuming and resource-intensive process that requires significant technical expertise, specialized tools, and manual effort. Creating lifelike animations, especially those involving lip-synced avatars and dynamic visuals, often requires large teams and extended production timelines, making it inaccessible to many creators. Existing solutions for animation and video generation are limited in creative flexibility, lack integration of natural language processing, or fail to provide a seamless pipeline from text to animated content. This gap creates a barrier for story maker, educators, and marketers who need efficient, scalable, and cost-effective ways to transform written narratives into engaging visual stories. Therefore, there is a pressing need for an AI-powered platform that simplifies the animation process by unifying text interpretation, avatar generation, lip synchronization, and video synthesis into a single, user-friendly solution.

4 Project Aim and Objectives

4.1 Project Aim

The aim of this project is to design and develop **AnimaticVission**, an AI-powered story-to-animation platform that transforms written text into fully animated sequences. By integrating advanced language models, lip-synced avatars, and dynamic video synthesis, the system seeks to simplify traditional animation workflows while enabling creators to produce engaging, visually rich content with greater speed, accuracy, and creative free-

dom. Unlike existing tools, **AnimaticVission** provides full scene animation, customizable avatars, and real-time lip-syncing, offering superior creative flexibility.

4.2 Project Objectives

- To integrate Large Language Models (LLMs) and LangGraph for analyzing and structuring textual input into animation-ready scripts.
- To implement lip-synchronization technology for realistic speech and avatar animations.
- To establish a streamlined pipeline that unifies text processing, image generation, and video synthesis.
- To design a user-friendly interface accessible to filmmakers, educators, and marketers without requiring technical expertise.
- To ensure scalability, customization, and creative freedom for diverse storytelling applications.

4.3 Expected Output

A web-based application allowing users to input text and generate fully animated video sequences, supported by cloud-based AI models, GPU servers for inference, and a user-friendly interface for multi-platform access.

5 Scope and Significance

5.1 Project Scope

Write down the scope and modules of the proposed project.

5.2 Project Significance

The scope of **AnimaticVission** covers the design and development of an AI-powered text-to-animation platform that leverages large language models (LLMs), LangGraph, and multimodal generation techniques. The system enables users to create high-quality, lifelike animations from simple text prompts. The project is designed with the following modules:

1. Text Understanding and Prompt Processing

Utilizes LLMs for semantic understanding of user prompts and extracts objects, attributes, emotions, and scene descriptions for animation generation.

2. Avatar and Character Generation

Creates lifelike avatars with customizable appearances and generates lip-synced speech aligned with the provided script.

3. Image and Scene Composition

Uses AI-driven image generation to create backgrounds, environments, and props, while supporting compositional arrangements of multiple objects and attributes.

4. Video Synthesis Module

Integrates avatars and scenes into coherent, dynamic video sequences, handling motion, transitions, and temporal consistency.

5. User Interface and Workflow Management

Provides an intuitive interface for script input, customization, and preview, with options to export animations in various formats for film, education, and marketing.

6 Project Significance

The proposed system carries significant advantages and benefits, both technically and practically:

- **Efficiency in Content Creation**

Reduces the time and complexity of traditional animation pipelines and enables rapid prototyping and iteration of animated content.

- **Accessibility for Non-Experts**

Lowers the barrier to entry by allowing creators with no animation skills to generate professional-quality animations.

- **Enhanced Storytelling**

Supports multi-object, multi-attribute scenes, enabling complex and creative narratives. Provides natural lip-sync and expressive avatars for engaging communication.

- **Cross-Domain Applications**

- *Educators*: Creation of interactive and engaging lessons.
- *Marketers*: Personalized and scalable promotional content.

- **Scalability and Adaptability**

Can be extended to support different languages, cultural contexts, and creative styles. Flexible enough to integrate with existing digital media pipelines.

7 Project Development Methodology

Distribute the project goals into smaller objectives/modules and highlight deliverables for each objective.

Explain the modules of project through a system level block diagram. Students may also mention tools, technologies and suitability of the method(s) to be employed with justification.

In case of a research problem, show the few approaches that will be investigated in the project?

8 Tools and Technologies

The development of **AnimaticVision** follows a structured methodology that ensures the platform effectively bridges the gap between text-based storytelling and dynamic animation. The methodology is divided into the following key phases:

1. Research Phase

- **Literature Review:** Study existing animation pipelines, AI-powered content generation tools, and story-to-animation frameworks to identify gaps and opportunities.
- **Technology Exploration:** Analyze Large Language Models (LLMs), LangGraph, AI-driven image generation, lip-syncing models, and video synthesis techniques to determine the most suitable technologies for integration.
- **User Needs Assessment:** Conduct surveys and interviews with filmmakers, educators, and marketers to understand the challenges in current animation workflows and the desired features for a streamlined platform.

2. Design Phase

- **System Architecture Design:** Define the platform's modular architecture, integrating natural language processing, image generation, and video synthesis components.
- **Workflow Design:** Map the end-to-end user experience, from script input to final animated output, ensuring usability and efficiency.
- **Prototyping:** Develop wireframes and mockups for the user interface, focusing on intuitive controls for text input, style selection, and video preview.
- **Model Selection & Integration:** Select appropriate AI models for avatar creation, lip-syncing, image generation, and video synthesis, and design a workflow for seamless interaction between them.

3. Development Phase

- **Backend Implementation:** Build APIs and server-side logic to process text inputs, generate images, and synthesize video sequences efficiently.
- **Frontend Implementation:** Develop an interactive interface for users to input scripts, customize animations, and preview results in real-time.
- **Integration & Testing:** Integrate all AI modules, ensuring synchronization between lip-synced avatars and generated video frames, followed by rigorous testing for accuracy, performance, and user experience.
- **Optimization:** Enhance the system for speed, scalability, and resource efficiency, enabling creators to generate high-quality animations quickly.

9 Work Plan

9.1 Team Structure

Define roles of each team member in your group. See example Table 2.

Table 2: Team Structure

Sr. No.	Team Members	Role
1	Mr. Atta-ur-Rehman	Project Manager
2	Mr. Atta-ur-Rehman	MERN Stack Developer
3	Mr. Muhammad Abdullah	QA Engineer, Automated Testing
4	Mr. Muhammad Humam Khan	Research and Development, AI / ML LLM
5	Mr. Muhammad Humam Khan	LLM Engineer, Natural Language Processing
6	Mr. Muhammad Abdullah	Frontend

9.2 Work Distribution

Clear work division among group members in a table. See example Table 3.

Table 3: Work Distribution

Sr. No.	Team Member	Work Assignment
1	Mr. Muhammad Abdullah	Frontend Development
2	Mr. Atta-ur-Rehman	Backend Development
3	Mr. Muhammad Human Khan	Research and Development of LLM Models
4	Mr. Muhammad Abdullah	QA Tester

9.3 Gantt Chart

Fig. 1 shows gives the visual representatio of the overall development process on the basis of the tasks, timeline, dateline, milestones and dependencies.

Note: Clear milestones should be defined at the start of the project which includes a Gantt chart AND Figure should be clear and readable.

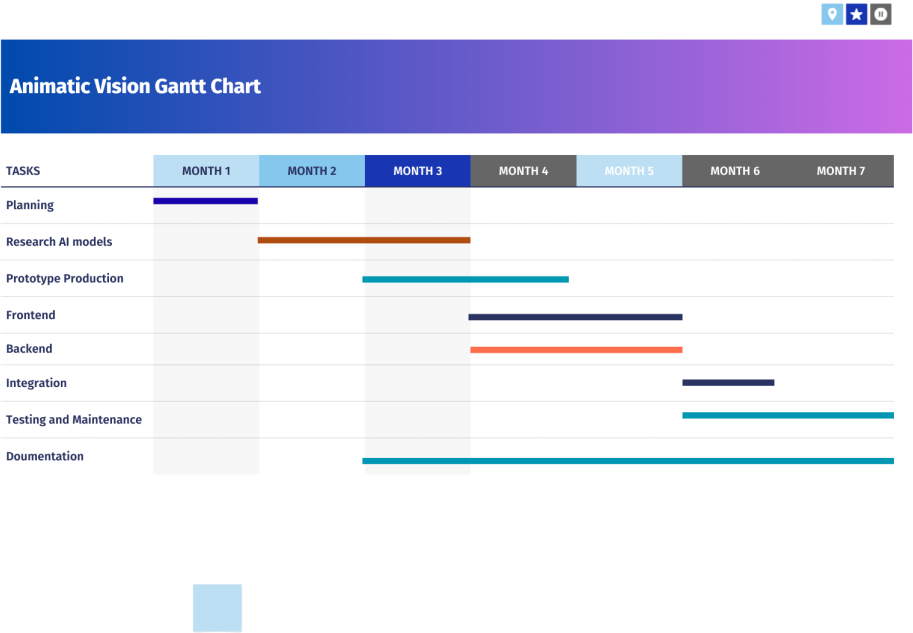


Figure 1: Gantt Chart

References

- [1] Sun, Kaiyue, et al. "T2v-compbench: A comprehensive benchmark for compositional text-to-video generation." Proceedings of the Computer Vision and Pattern Recognition Conference. 2025.
- [2] Fakhabi, Erfan Esmaeili. Text-Image Alignment in Diffusion Models: The Role of Attention Sink. Diss. Purdue University Graduate School, 2025.
- [3] Huang, Yifei, Jilan Xu, Baoqi Pei, Yuping He, Guo Chen, Mingfang Zhang, Lijin Yang et al. "An egocentric vision-language model based portable real-time smart assistant." arXiv preprint arXiv:2503.04250 (2025).