GENERATIVE AI LAB PROJECT

Study on

AI-POWERED STORYTELLER

Submitted by

GUDIPATI RAJESH NAIDU
MEKALA RAHUL RAJ
BALAPANURU KOUSTUBH

Registration No:

2021BCSE07AED146
2021BCSE07AED125
2021BCSE07AED070

In partial fulfillment of the GENERATIVE AI LAB in Semester VII of the CSL708 Bachelor of Technology

(2021-25)

Branch: COMPUTER SCIENCE AND ENGINEERING

Specialization: AIML

of Alliance University

NOVEMBER 2024

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ALLIANCE COLLEGE OF ENGINEERING AND DESIGN

ALLIANCE UNIVERSITY, BENGALURU

ALLIANCE COLLEGE OF ENGINEERING AND DESIGN

(ALLIANCE UNIVERSITY, BENGALURU)

"AI-POWERED STORYTELLER "

Bona fide record of work done by

GUDIPATI RAJESH NAIDU (2021BCSE07AED146)

MEKALA RAHUL RAJ(2021BCSE07AED125)

BALAPANURU KOUSTUBH (2021BCSE07AED070)

A mini project report submitted in partial fulfillment of the requirements for the degree

of

BACHELOR OF TECHNOLOGY

Branch: COMPUTER SCIENCE AND ENGINEERING

Specialization: AIML

Of Alliance University

November 2024

…….…..…. …….……

Dr. Mousumi Karmakar **Dr. Tinka Singh**

Faculty guide Faculty guide

Department of Computer Science and Engineering Department of Computer Science and Engineering

Alliance College of Engineering and Design Alliance College of Engineering and Design

TABLE OF CONTENTS

SL.NO	CONTENT	PAGE NO
1	Abstract	1
2	Introduction	2
3	Methodology	3-4
4	Code	4-7
5	Conclusion	8
6	Results	9

ABSTRACT

This project uses generative AI to automate video storytelling for the first time. It explores how to create engaging narratives from simple text prompts using a new pipeline. This system employs large language models, text-to-speech technology, and diffusion models to collectively empower users to create unique video stories without the need for technical or creative skills, making video story content creation more democratic.

Initial work starts with a pre-trained GPT2 model which is a powerful language model that writes a short story on the user's input. The narrative of this video then acts as the foundation of the story, split up into its own individual sentences, which are then used as a blueprint to generate its visual and auditory content. With this segmentation, the part of the story is visualized and aural for each part of the story.

Each sentence is transformed into unique, captivating images, utilizing the cutting-edge diffusion model Stable Diffusion 2.1. The process adds color to a viewer's storytelling experience while giving the narrative a visual element to go with it. Google Text to Speech (gTTS) technology also simultaneously converts each sentence into spoken audio to create an additional immersive auditory layer making listening through the device feel even more engaging.

Finally, these generated components—images, audio, and text—are finally seamlessly integrated into a movie video format. Each image follows its audio narration and overlaid text subtitles creating a dynamic and accessible telling. The generative AI that we showcase here serves as a vision of how the transformative potential of AI in creating content is possible, showcasing a future in which AI and human creativity exist in seamless harmony, opening whole new doors for communicative and creative expression. Automated video storytelling is democratizing access to creative tools and streamlining the storytelling process, allowing everyone to bring their envisioned world to life in dynamic ways.

INTRODUCTION

Generative models are a subfield of generative AIs, which have been a big development and application of artificial intelligence. These models have a special ability to produce new, text, or images, or even audio and sometimes video, paring some of the limits of imagination. This project begins an exploration into the transformational power of generative AI by developing a novel video storytelling system that uses generative AI to automate this process. This project employs novel approaches that do not compromise on the seamless mapping of user's textual thoughts into compelling and expressive video stories, from large language models, diffusion models, and text-to-speech systems.

This innovative system relies on a carefully orchestrated pipeline that focuses on the collaborative powers of many generative AI models. At this stage, the process begins with a GPT 2, a pre-trained language model that can produce fluently human-like text with good coherence. In this video, the user is given a text prompt which it then takes and turns into a short story, which then becomes the narrative foundation of the video. This generated story is meticulously segmented into single sentences and then blueprinted from sentence to sentence into pieces of visual and auditory elements for a granular expressive storytelling project. We use Stable Diffusion 2.1, which is a state-of-the-art diffusion model, to give the narrative visual life and help breathe it. The story is translated into each sentence being converted to a photo using Stable Diffusion 2.1, where you convert textual descriptions into dynamic visualizations. Yet it improves storytelling in a way, overloading the visual aspect of the story with engaging visuals that further the narrative, adding meaning and emotionalism.

At the same time, Google Text to Speech (gTTS) technology is used to create an immersive auditory layer on the video, emphasizing the storytelling experience, and making this experience more accessible to a larger audience, as gTTS speaks the sentence and reads the text aloud, providing the narration of which tells readers where to go next, and enhance the story feeling throughout the journey. The system integrates seamlessly audio with visuals to provide an immersive, more holistic storytelling experience. Finally, Movie, a powerful video editing library, effortlessly stitches together the components that Movie generates—images, audio, and text—into a nice, polished finished video. The text subtitles superimposed with the audio narrations play the audio narrations with the original text simultaneously synchronized and the sequence of images is synchronized with the original text. Generative AI, aside from leading to many AI satire jokes, can have incredible power when applied to content creation. It will democratize access to creative tools that so many would not otherwise be able to afford and streamline the storytelling process.

METHODOLOGY

The methodology for this project Code we have taken a methodological approach to the automation of video storytelling using generative AI. There are several steps that serve as the methodology in its own right using a new set of AI models and techniques to reach this final result. Within the first stage of a generation pipeline, an initially generated story is first focused on story generation through a pre-trained GPT-2 model. This large language model is an interactive creative seed for the story that takes user-defined text and prompts as input. The video content is based on what GPT-2, a generation that is set upon by the prompt, creates: a short story. After breaking this story apart into individual sentences, each is a part of an extended story.

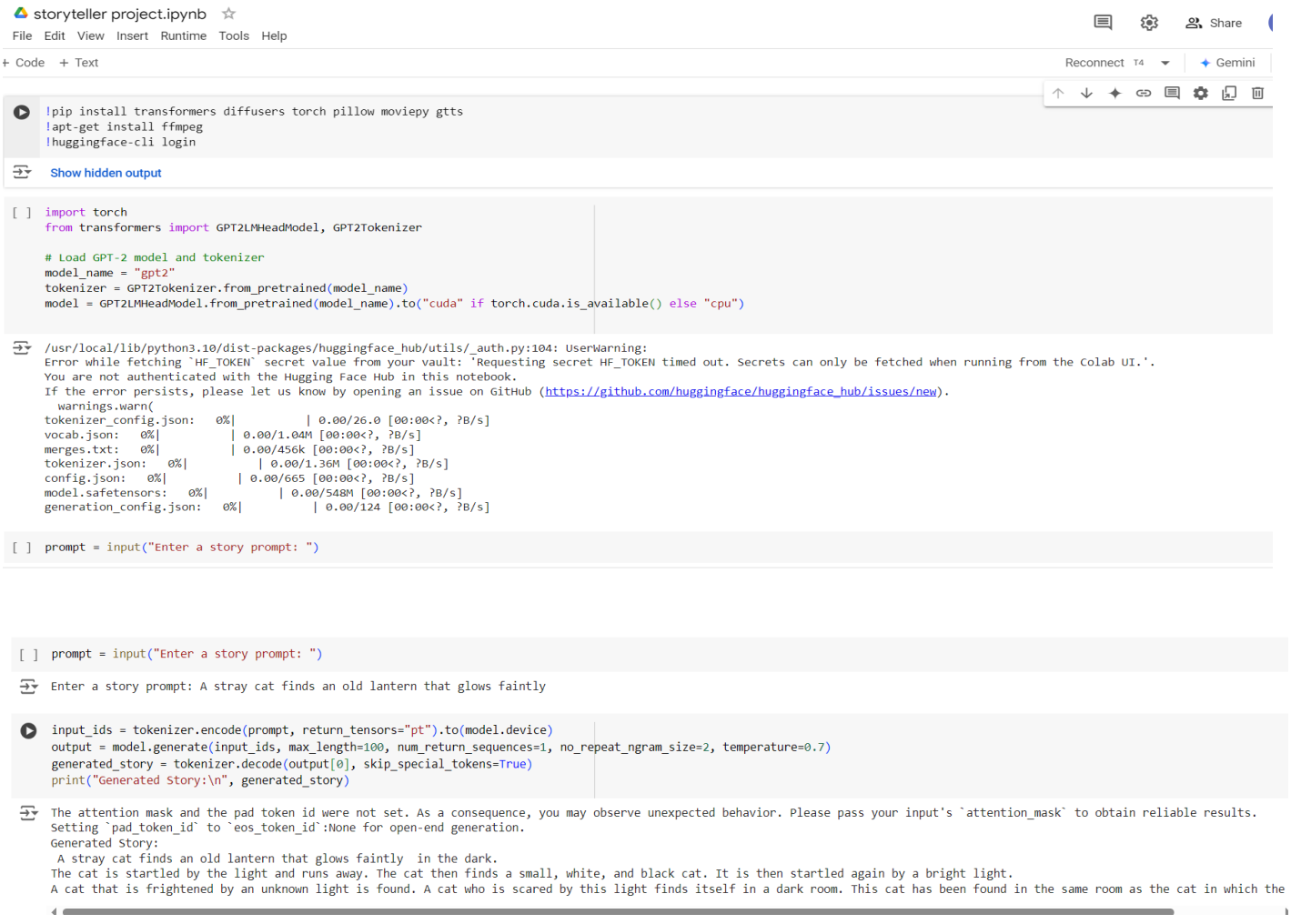
The second stage consists of producing visual content for each sentence in the story. That relies on the use of the most powerful diffusion model in the market for now: Stable Diffusion 2.1, through which we can generate very high-quality images from text. Each sentence supplied to Stable Diffusion 2.1 is used to create a unique image describing the contents of that sentence. When the textual narrative is turned into a series of images, the visual backbone of the video sticks.

This third stage teaches how to make audio narration onto the story. Google Text to Speech (gTTS) technology is executed on each sentence to convert it into spoken audio. The addition of such offers us a layer of audio to offer as a companion to the visual bits, and it adds to the narrative. Presented is a method to generate all audio clips in perfect synchronization with their respective image (audio clip A corresponding to image 1), which smoothly advances the audio clips as a part of a story.

The fourth stage integrates generated components into a final video format using images, audio, and text. A versatile video editing library Movie is being used for this purpose. These are synchronized audio clips but with a sequence of images arranged chronologically. And text subtitles, displaying the original sentences produced by the generated story, are also overlaid on the video. With it, a viewer can have textual content for the visual and auditory content that they like or require.

After that, you will export the video in a format such as mp4, and then you'll view the video, or you'll share it. In this project we use a novel approach to automated video storytelling utilizing a new methodology - by doing so, we generate interesting and dynamic stories purely through a series of simple text prompts using generative AI models. The systematic workflow ensures textual, visual, and auditory elements are coherent and cohesive and the viewers get compelling multimedia. What we've shown here is that AI now can do this kind of rethinking of content generation a new type of writing the story the story told by the human being is now possible.

CODE



```
[ ] import re

    # Function to split story into sentences
    def split_story_into_sentences(story):
        # Use regex to split by punctuation followed by whitespace or end of text
        sentences = re.split(r'(?<=[.,!?])\s+', story)
        return [sentence.strip() for sentence in sentences if sentence]

    # Split the story into individual sentences
    story_sentences = split_story_into_sentences(generated_story)
```

```
from diffusers import StableDiffusionPipeline
from PIL import Image

# Load Stable Diffusion 2.1 model
pipe = StableDiffusionPipeline.from_pretrained("stabilityai/stable-diffusion-2-1").to("cuda" if torch.cuda.is_available() else "cpu")

# Generate an image for each sentence in the story
```

```
model_index.json: 100% 537/537 [00:00<00:00, 27.2kB/s]
Fetching 13 files: 100% 13/13 [00:25<00:00, 2.28s/it]
model.safetensors: 100% 1.36G/1.36G [00:14<00:00, 141MB/s]
tokenizer/vocab.json: 100% 1.06M/1.06M [00:01<00:00, 1.01MB/s]
scheduler/scheduler_config.json: 100% 345/345 [00:00<00:00, 3.47kB/s]
tokenizer/special_tokens_map.json: 100% 460/460 [00:00<00:00, 4.15kB/s]
```

```
images = []
for sentence in story_sentences:
    if sentence.strip():
        print(f"Generating image for sentence: '{sentence}'")
        image = pipe(sentence).images[0]
        images.append(image)
```

```
Generating image for sentence: 'A stray cat finds an old lantern that glows faintly  in the dark.'
100% 50/50 [01:06<00:00, 1.29s/it]
Generating image for sentence: 'The cat is startled by the light and runs away.'
100% 50/50 [01:04<00:00, 1.31s/it]
Generating image for sentence: 'The cat then finds a small,'
100% 50/50 [01:05<00:00, 1.30s/it]
Generating image for sentence: 'white,'
100% 50/50 [01:05<00:00, 1.31s/it]
Generating image for sentence: 'and black cat.'
100% 50/50 [01:05<00:00, 1.30s/it]
Generating image for sentence: 'It is then startled again by a bright light.'
100% 50/50 [01:05<00:00, 1.30s/it]
Generating image for sentence: 'A cat that is frightened by an unknown light is found.'
100% 50/50 [01:05<00:00, 1.30s/it]
Generating image for sentence: 'A cat who is scared by this light finds itself in a dark room.'
100% 50/50 [01:05<00:00, 1.30s/it]
```

```
[ ] from gtts import gTTS

    audio_clips = []
    for i, part in enumerate(story_sentences):
        if part.strip():
            # Generate TTS audio for each segment
            tts = gTTS(text=part, lang='en')
            audio_path = f"audio_clip_{i}.mp3"
            tts.save(audio_path)
            audio_clips.append(audio_path)
```

```
from moviepy.editor import ImageClip, AudioFileClip, concatenate_videoclips

# Create a video clip for each image and its corresponding audio
clips = []
for i, image in enumerate(images):
    # Save the image temporarily
    image_path = f"image_{i}.png"
    image.save(image_path)

    # Load the audio and get its duration
    audio_clip = AudioFileClip(audio_clips[i])
    audio_duration = audio_clip.duration  # Get the duration of the audio file

    # Load the image and set its duration to match the audio duration
    img_clip = ImageClip(image_path).set_duration(audio_duration)

    # Combine image and audio into a single clip
    img_clip = img_clip.set_audio(audio_clip)
    clips.append(img_clip)

# Concatenate all clips into a single video
video = concatenate_videoclips(clips, method="compose")

# Export the final video
output_video_path = "story_video.mp4"
video.write_videofile(output_video_path, fps=24, codec="libx264", audio_codec="aac")

print(f"Video saved at {output_video_path}")
```

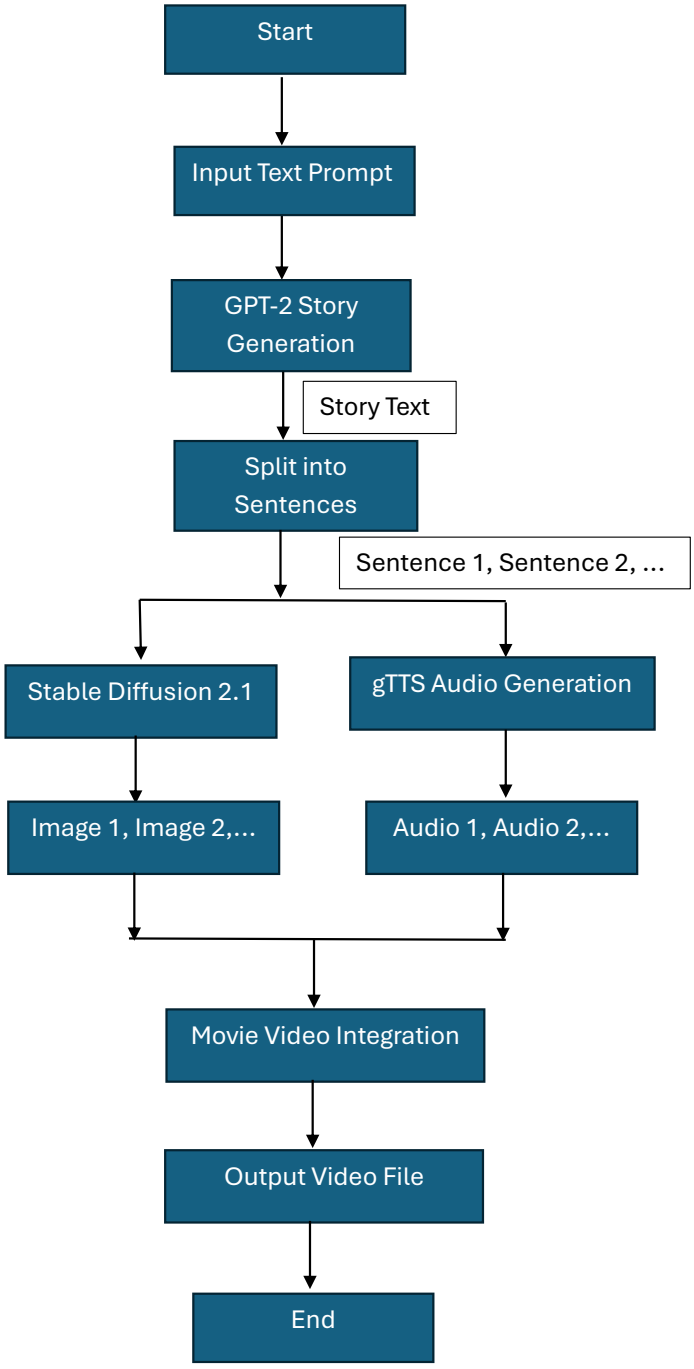
```
WARNING:py.warnings:/usr/local/lib/python3.10/dist-packages/moviepy/video/io/sliders.py:61: SyntaxWarning: "is" with a literal. Did you mean "=="?
  if event.key is 'enter':

Moviepy - Building video story_video.mp4.
MoviePy - Writing audio in story_videoTEMP_MPY_wvf_snd.mp4
MoviePy - Done.
Moviepy - Writing video story_video.mp4

Moviepy - Done !
Moviepy - video ready story_video.mp4
Video saved at story_video.mp4
```

-

ARCHITECTURE



CONCLUSION

In this project we successfully demonstrated generative AI's potential for video storytelling automation, transforming simple text prompts into engaging multimedia narratives. Through integrating large language models, diffusion models, and text-to-speech technology, we have built a system that allows users to easily craft novels without requiring hours of technical expertise or creative skills. However, one of the project's innovative details is demonstrating the synergistic potential of using different AI models for creative content generation, giving rise to the opportunity for new personalized storytelling experiences.

The results of this project underscored the efficacy of the approach employed to generate coherent and visually appealing video narratives. Stable Diffusion 2.1 and GPT2 have successfully achieved prompt-based imaginative story generation and visually captivating images. immersive audio narration was added through the integration of gTTS technology. Additionally, Movie helped blend all these elements into a coherent video format, which in turn showed the possibility of automating the whole video creation process.

By showcasing this, we are taking a great leap towards a future in which AI will become indistinguishable from human creativity, providing all the means to tell stories dynamically and engagingly to those who want to tell them. Further refinements and explorations are possible, but this project lays a strong foundation for future research and development in the use of AI to propel video storytelling in exciting new directions, leading to more personalized narrative experiences. With this technology's potential applications ranging from entertainment and education to marketing and communication, content creators and storytellers alike will be able to discover exciting possibilities for creative storytelling.

RESULTS

```
print("Generated Story:\n", generated_story)
```

```
The attention mask and the pad token id were not set. As a consequence, you may observe unexpected behavior. Please pass your input's `attention_mask` to obtain reliable results.
Setting `pad_token_id` to `eos_token_id`:None for open-end generation.
Generated Story:
 A stray cat finds an old lantern that glows faintly  in the dark.
The cat is startled by the light and runs away. The cat then finds a small, white, and black cat. It is then startled again by a bright light.
A cat that is frightened by an unknown light is found. A cat who is scared by this light finds itself in a dark room. This cat has been found in the same room as the cat in which the lantern was found,
```

Fig 1: Story generated by the GPT2

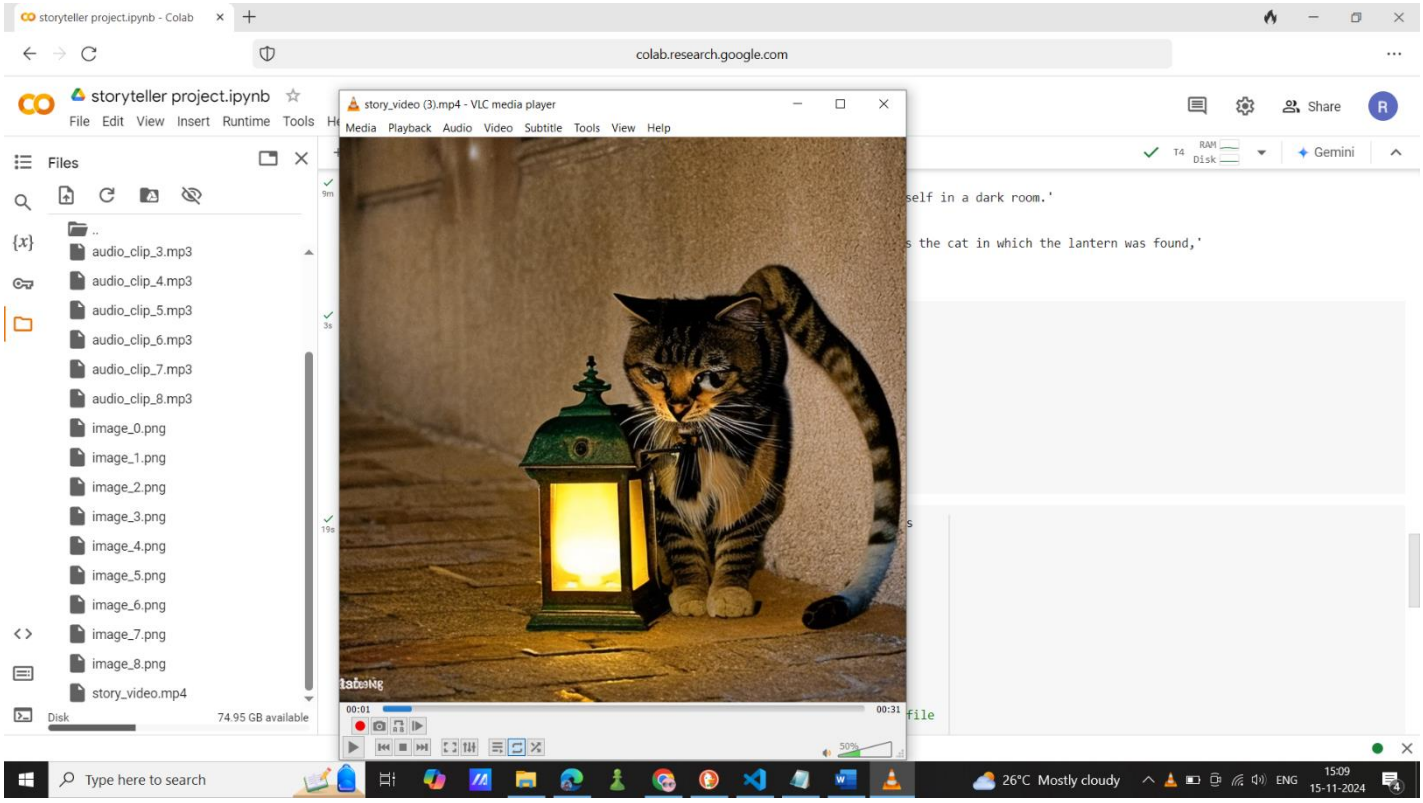


Fig 2: Final output (video) of the generated story