Story Generation with State-of-the-Art Text-to-Video and Text-to-Audio Models

Arham Chouradia
chouradi@usc.edu

Jehil Vora
jehilyog@usc.edu

Pranav Parnerkar
parnerka@usc.edu

Saad Shaikh
shaikhm@usc.edu

Sahil Mondal
sahilmon@usc.edu

1 Project Domain and Goals

We propose the development of a cutting-edge story generation system that leverages state-of-the-art Multi-Modal Natural Language Processing (NLP) models (Text-to-Video and Text-to-Audio) to convert textual prompts into immersive audio-visual narratives. The objectives for this project are mentioned below:

1. Develop a state-of-the-art story generation system by enabling users to input prompts and receive audio-visual stories in response.

2. Incorporate cutting-edge NLP models for text-to-video and text-to-audio synthesis.

3. Solve the audio-video synchronization problem between the generated outputs.

4. Reduce hallucinations within the NLP model with prompt engineering.

1.1 Problem Description

1. **Content Creation Bottleneck:** In the digital era, generating content for audiovisual storytelling is crucial across various industries like marketing, entertainment, and education. However, this process is often time-consuming, resource-heavy, and requires specialized skills. Our project aims to automate this bottleneck with NLP-powered story generation.

2. **Lack of Personalized Content:** Many individuals and businesses struggle to create personalized and engaging content due to limited resources and expertise. Our project seeks to democratize content creation by offering an accessible system for generating customized stories, enhancing user engagement and brand visibility.

1.2 Role of NLP in the solution

In recent years, the technology sector has witnessed a remarkable surge in Natural Language Processing (NLP) advancements, largely driven by breakthroughs like the release of ChatGPT and similar technologies. Developers have been quick to embrace these innovations, utilizing NLP's language understanding capabilities to create a diverse range of applications. This surge in NLP technology has profoundly impacted content creation, offering several key benefits. NLP seamlessly integrates with visual and audio elements, elevating storytelling to an immersive level while streamlining the content generation process. It further enhances the experience by providing personalized narratives and exhibits scalability to handle a wide array of user-generated prompts, ensuring that content easily reaches a broader audience.

1.3 Larger Impact

Our project offers a multifaceted solution. Firstly, it empowers individuals and businesses by removing technical barriers in content creation, allowing them to focus on their creative ideas. Secondly, it brings economic benefits by automating production and making high-quality audio-visual storytelling accessible to small businesses and creators, thus fostering competitiveness and economic growth. Lastly, it adds educational and entertainment value, with the system generating engaging content across various industries, including education, entertainment, and marketing.

2 Related Work

Text-to-audio and text-to-video systems have been thoroughly investigated individually but not as a unit. In text-to-audio synthesis, the Contrastive Language-Audio Pretraining (CLAP) approach learns to connect language and audio using two encoders and contrastive learning to bring audio and text descriptions into a joint multimodal space

(Elizalde et al., 2022). A computationally cost-efficient approach, AudioLDM, uses pre-trained CLAP models to train latent diffusion models (LDMs) with audio embedding while providing text embedding as a condition during sampling (Liu et al., 2023) and achieves state-of-the-art performance with a single GPU.

In text-to-video synthesis, a popular approach taken is to use a conditional generative adversarial network that generates frame-by-frame images that ultimately develop a video from text input (Raja et al., 2023). Another approach, Tune-A-Video, uses a text-to-image diffusion model to create multiple images and a tailored spatiotemporal attention mechanism to provide continuous motion in video (Wu et al., 2023).

This project builds upon the established framework, MusicLM (Agostinelli et al., 2023), a model generating high-fidelity audio from text descriptions. We also utilize Phenaki (Villegas et al., 2022), a bidirectional masked transformer, for video generation. It is conditioned on pre-computed text tokens which are subsequently detokenized to get actual video.

3 Dataset

The MusicCaps dataset (Agostinelli et al., 2023), comprises 5.5k high-quality music clips from AudioSet, each accompanied by expertly crafted English textual descriptions. These descriptions were meticulously curated by ten skilled professional musicians. The dataset required no additional preprocessing as its creators had already thoroughly cleaned and prepared it.

The WebVid dataset (Bain et al., 2021) is a comprehensive collection of short videos, each accompanied by corresponding textual descriptions. These videos were sourced from stock footage sites, ensuring diversity and representativeness. The authors incorporated additional training data from a 1.8B parameter Phenaki model, which was trained on a vast corpus of approximately 15M text-video pairs at 8 FPS. The dataset also included approximately 50M text-images and about 400M pairs of LAION-400M. Notably, 80% of the training data came from the video dataset, with each image dataset contributing 10%. While we would not explicitly train on this data, we plan on utilizing its extrapolated features in our project.

4 Technical Challenge

With this implementation, we aim to target 3 main issues with story generation.

1. **Synchronization and Coherence**: Achieving synchronization and coherence between the audio and video generated by two separate models is a significant challenge. The generated audio and video components need to align seamlessly to create a compelling and coherent short story. Ensuring that the generated audio corresponds to the actions and scenes in the video is non-trivial.

2. **Realism and Quality**: Generating high-quality audio and video that look and sound realistic is another significant challenge. Both the text-to-audio system for generating audio and the video generation model need to produce content that is visually and auditorily convincing to the audience.

3. **Storytelling and Creativity**: Generating engaging and creative short stories require models to understand narrative structure, character development, and emotional expression. Achieving this level of understanding and creativity in both audio and video generation is a complex challenge.

In our current coursework, we do not work with multi-modal NLP models. Moreover, the task of creating synchrony between two NLP tasks is going to be all the more challenging. We push the envelope of what the class teaches us and even use techniques that use a different class of models with variations in training methods, data heterogeneity and evaluation criteria.

5 Evaluation

Because the task of evaluating the quality of AI-generated content is very subjective and we are using out-of-the-box models to solve the tasks, we plan to use human feedback as a measure of the quality of the final outputs. The use of surveys to create a crowdsourced opinion will best inform us of the effectiveness of the end result.

Name	Task
Arham Chouradiya	Feasibility of text-to-video models
Jehil Yogesh Vora	Feasibility of text-to-audio models
Pranav Parnerkar	Integration of Vision and Audio Synthesis Pipeline
Saad Shaikh	Setting up evaluation criteria and surveys
Sahil Mondal	Prompt engineering with the synthesis pipeline

Table 1: Work division

References

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank. 2023. Musiclm: Generating music from text.

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

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2023. Simple and controllable music generation. *arXiv preprint arXiv:2306.05284*.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2022. Clap: Learning audio concepts from natural language supervision.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D. Plumbley. 2023. Audioldm: Text-to-audio generation with latent diffusion models.

Sivakami Raja, Mierudhula S, and Potheeswari J. 2023. Text to video generation using deep learning. In *2023 Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)*, pages 1–7.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2022. Fastspeech 2: Fast and high-quality end-to-end text to speech.

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

Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. 2023. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7623–7633.