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

Arham Chouradiya
chouradi@usc.edu

Jehil Vora
jehilyog@usc.edu

Pranav Parnerkar
parnerka@usc.edu

Saad Shaikh
shaikhm@usc.edu

Sahil Mondal
sahilmon@usc.edu

Abstract

Audio-visual (AV) content creation is a task that is hard to automate. This project focuses on a key aspect of AV content creation: Storytelling. We explore two distinct methods for generating AV stories by leveraging state-of-the-art text-to-video and text-to-audio models. Moreover, we validate our approaches by incorporating human feedback through a survey to gain confidence in producing desired outputs. To conclude, our survey results indicate that out of 5 generated AV stories, the average rating by 21 individuals is 3.69/5.0, demonstrating promising progress in bridging the gap between AI-generated and human-produced content.

1 Introduction

Content creation involves technical supervision by individuals with hard skills like cinematography, editing, music composition, script writing, etc. Previously, AI-generated content was fairly distinguishable from human-made productions due to inconsistencies resulting from a lack of technical supervision and compelling narratives. However, advancements in generative AI with deep learning diminish these disparities by the day.

Recent strides in text-to-video synthesis have witnessed the emergence of highly sophisticated algorithms capable of seamlessly translating textual input into dynamic visual representations by leveraging advanced neural network architectures. These models utilize a combination of natural language understanding and computer vision methodologies to process textual information and translate it into coherent video sequences.

The advancements in text-to-music models mirror the progress seen in text-to-video, harnessing sophisticated neural network architectures to interpret textual information and translate it into intricate musical compositions enriched by attention to context, semantics, and the vibrant nuances of sound.

We propose two distinct approaches to this problem in Section 2

2 Methods

2.1 Current Implementation

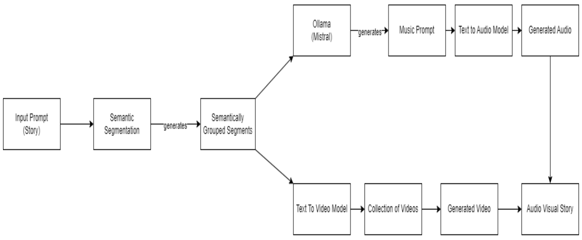

Figure 1: Current Flow

Our innovative approach commences with a carefully crafted input prompt. This prompt undergoes the crucial phase of semantic segmentation, leading to the creation of semantically grouped segments. These distinct segments then follow dual pathways, culminating in a harmonious synthesis of visual and auditory elements.

2.1.1 Semantic Segmentation

Semantic topic segmentation is a text analysis technique designed to automatically identify and group sentences or paragraphs within a document based on their semantic meaning. This method uses advanced natural language processing (NLP) models to capture the underlying context and relationships between sentences. This allows for creating meaningful segments where content with similar themes is grouped. Semantic topic segmentation is valuable for summarizing large volumes of text, organizing information, and uncovering key topics within a document, making it an essential tool in text analysis and information retrieval.

We employed a BERT-based sentence embedding and K-means clustering to categorize sentences into distinct thematic clusters for the topic segmentation task. Using the paraphrase-distilroberta-base-v1 model for accurate sentence embedding, the algorithm identifies semantic similarities among sentences, allowing it to group them into cohesive clusters. The number of clusters can be adjusted to control the granularity of the segmentation. We input the prompt to this model, fetch semantic segments and feed it into our pipeline to generate video and audio.

2.1.2 Collating the results

In the first pathway, the segmented text embarks on a transformative journey through a 'Text to Video' model, culminating in the creation of a diverse collection of videos. These videos, each representing a unique facet of the input text, are subsequently harmonized to produce a visually cohesive representation that captures the essence of the underlying semantic structure.

Concurrently, in the second pathway, the segmented text is the foundation for generating a musical prompt, accomplished using Mistral (Jiang et al., 2023) on Ollama (oll, 2023). This musical prompt plays a pivotal role in guiding the subsequent 'Text to Audio' model, steering it towards crafting an audio representation that complements the narrative envisioned by the input text. A detailed explanation of this musical prompt is provided in the next section, highlighting its significance in facilitating the model's ability to align with the emotional and thematic nuances of the given story.

The audio representation and video collection seamlessly merge to create an output that integrates visual and auditory components, offering a comprehensive and immersive storytelling experience, translating the richness of language into a captivating synthesis of sight and sound. You can find the implementation of this pipeline here [1]

Their fusion remains unexplored despite flourishing individual advancements in text-to-music and text-to-video technologies. To solve this task, this paper pioneers this exploration that combines technologies in sections 3.1.3 and 3.2.3.

2.1.3 Enhancing current approach

With semantic segmentation, we faced two major issues:

- Deterministic timing: While semantically segmented and constructing coherent splits, the prompts are still exactly that: splits. There is a jagged transition between the scenes and an absence of camera angle panning.
- Length of generated prompts: As semantic segmentation divides a large corpus into smaller semantically coherent chunks, a prompt segment may be too long for the given model. Most text-to-video models have low token limits and thus make it challenging to construct descriptive prompts.

These ultimately enhanced our methodology as described in section 2.2.

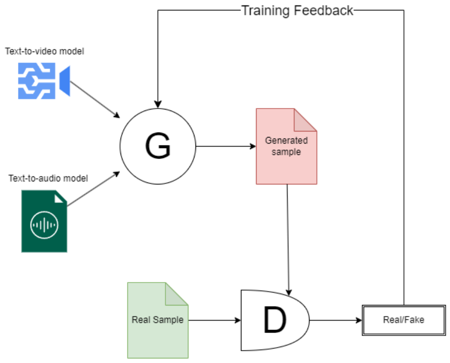

Figure 2: Audio-Video pipeline using a GAN

2.2 Generative Adversarial Networks

As the previous approach provided improved results over the naive output initially, we now seek to improve further on our solution. As a result of discussions with our advisor, we aimed to experiment with GAN's (Goodfellow et al., 2014). The idea is to use a GAN where the generator creates audio-video samples from the existing pipelines, and the discriminator compares them to actual audio-video examples. On every iteration, where the discriminator successfully discriminates between the two, we use this feedback to update the text-to-video and text-to-audio models as described in Figure 2.

3 Experiments

This section unveils our exploration into diverse Text-to-Video and Text-to-Audio models, shedding light on the rationale behind certain deliberate choices made while crafting and prototyping our pipeline, as delineated in section 2.

[1] https://github.com/jehilvora/CSCI-544-Project-Fall23-Group27

3.1 Text-to-Video Models

Text-to-video models are invaluable assets in creating a coherent sequence of images derived from textual input. As outlined in section 2.1.1, our approach involves iteratively supplying these models with text segments generated through the semantic segmentation process. This iterative process yields a collection of videos, subsequently compiled to produce the final generated video output. During our research, we identified three promising 'candidate' models suitable for integration into our pipeline. The subsequent subsections delve deeper into the intricacies and nuances of these models, elucidating their specific functionalities and strengths.

3.1.1 Phenaki

Phenaki (Villegas et al., 2022) presents an innovative video synthesis method using a two-part system: an encoder-decoder model compressing videos into tokens and a bi-directional masked transformer translating text embeddings into video tokens. Noteworthy is Phenaki's training approach, involving joint training on diverse image-text pairs and limited video-text examples, enhancing its generalization beyond existing datasets. Although Phenaki achieves remarkable results, surpassing established benchmarks, our decision not to adopt it stems from a lack of accessible open-source resources.

3.1.2 Text2Video-Zero

This model (Khachatryan et al., 2023) pioneers zero-shot text-to-video generation by offering a cost-effective solution without necessitating training or optimization. Leveraging existing text-to-image synthesis methods, notably Stable Diffusion adapted for videos, it enriches generated frame latent codes with motion dynamics for temporal consistency. Additionally, it reprograms frame-level self-attention to preserve context, appearance, and object identity across frames. This innovative approach yields high-quality and consistent videos despite lacking additional video dataset training.

Due to limited computational resources, generating variable-length videos became challenging for us. Consequently, we decided to discontinue the utilization of this particular model.

3.1.3 DAMO Vilab's Text-to-Video

This model is based on VideoFusion (Luo et al., 2023), a decomposed diffusion probabilistic model for high-quality video generation. The key idea is to decompose the noise added during diffusion into a shared "base noise" that captures standard content across frames and a per-frame "residual noise" for variability. Separate base and residual generator networks predict these noises, which are used to iteratively denoise the latent variables. Leveraging a pre-trained image diffusion model to predict the base noise allows efficient propagation of shared semantics. Experiments show that VideoFusion outperforms other methods, and ablation studies validate the benefits of the proposed noise decomposition. Overall, explicitly modeling shared vs. varying video components improves coherence while reducing the generation burden.

Considering the model's robustness and low compute overhead, we chose this model for our pipeline.

3.2 Text-to-Audio Models

Text-to-audio models play a pivotal role in transforming text into coherent auditory sequences. Our method involves integrating these models within our pipeline to craft soundtracks that complement the generated videos. This section explores three promising text-to-audio models, detailing their unique functionalities and strengths for seamless integration.

3.2.1 MusicLM

MusicLM (Agostinelli et al., 2023) pioneers text-conditioned music generation with a hierarchical framework, merging distinct audio models for robustness. Leveraging MuLan's contrastive training aids scalability and noise resilience. Its two-stage Transformer architecture combines audio tokens for comprehensive, high-fidelity music aligned with diverse text prompts, outperforming previous quality and adherence.

However, despite the impressive outcomes demonstrated by MusicLM, the decision to refrain from its methodology stemmed primarily from the absence of accessible open-source materials and resources.

3.2.2 Make-An-Audio

Make-An-Audio (Huang et al., 2023) pioneers multimodal audio generation, utilizing CLAP and spectrogram autoencoders for text-to-music synthesis. It adapts prompts for personalized audio, innovates in audio inpainting with sophisticated masking, and extends capabilities to visual-to-audio translation.

This versatile framework minimizes dataset dependencies, empowering seamless text-to-audio synthesis.

The decision not to employ this approach was driven by its specialization in generating audio from prompts, diverging from our primary focus on generating music/melody specifically tailored for accompanying generated videos.

3.2.3 Audiocraft - MusicGen

MusicGen (Copet et al., 2023) revolutionizes conditional music generation with a single-stage transformer LM, enabling direct operation on compressed discrete music tokens. Unlike previous models, it eliminates the need for complex cascading, offering efficient token interleaving. This innovative approach empowers MusicGen to produce high-quality mono and stereo samples, responding adeptly to text or melodic feature conditioning for precise control over the generated musical output.

Given MusicGen's streamlined architecture and computational efficiency, we opt for its integration due to robustness and minimal computational demands in our pipeline.

4 Results and Discussions

As this project relied heavily on human objectivity for critical evaluation, we collected feedback on the generated samples through a Google form [2] from our peers to verify the validity of the samples against the provided prompts. We received encouraging results documented in Table 1, which shows great promise in our methodology and the outputs.

Sample	Average Rating (out of 5)
Generated Video 1	4.095238095
Generated Video 2	3.571428571
Generated Video 3	3.904761905
Generated Video 4	3.238095238
Generated Video 5	3.619047619
Average Rating	3.685714286

Table 1: Results of survey

4.1 Open source contributions

From our initial proposal, we aimed to use Phenaki (Villegas et al., 2022) as the text-to-video model. It boasts variable-length video generation, smooth scene transitions with realistic camera panning, and input lengths of up to 2,500 tokens. As this work is highly recent, open-source implementations of the paper and sufficient supporting work were absent. To rectify this issue, we are collaborating with a French research lab, Obvious Research (obv, 2023a), to build upon their current implementation of CViViT (obv, 2023b). We plan to work on two significant tasks with them.

1. Enhance the CViViT to include images and videos: The current implementation of CViViT uses only images to produce results that can validate the findings in the paper. It does not reach the performance seen in the paper, and thus, the pipeline has to be enhanced to accommodate both images and videos in variable splits, as described in the paper. As the lab at Obvious has computational resources to tackle this challenge, we will contribute to the training pipeline in the hopes of replicating the results from the paper.

2. Improving evaluation for the Phenaki text-to-video generation: The current model utilizes human inspection to validate the outputs. We will create an evaluation script that quantitatively evaluates the model using Fréchet Video Distance for the video datasets and Fréchet Inception Distance on the image dataset.

5 Conclusions

In conclusion, our project has successfully implemented a robust and innovative workflow, utilizing semantic segmentation, text-to-video, text-to-audio models, and Mistral for musical guidance. Despite our aspirations to enhance the quality of generated output through the integration of Generative Adversarial Networks (GAN), the limitation of computational resources and the lack of a suitably annotated dataset posed a challenge, preventing us from fully realizing this aspect of our vision. Acknowledging the potential of our work and the room for further improvement, we have strategically decided to contribute our findings and project framework to an open-source initiative. In collaboration with a research lab, we aim to extend this project's scope and explore avenues for refinement.

[2] https://github.com/jehilvora/CSCI-544-Project-Fall23-Group27/blob/main/Results/SurveyForm.pdf

6 Individual Contribution

The creation of the text-to-video pipeline, research on the GAN approach and the experimentation with music context models is performed by Pranav and Jehil. Saad, Arham and Sahil implemented the text-to-audio pipeline, semantic segmentation and the collation of the existing audio and video pipelines. Additionally, they managed the feedback process from sample generation to result aggregation.

References

2023a. Obvious research team.

2023. The ollama project.

2023b. Phenaki implementation from obvious research.

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.

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

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

Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. 2023. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. 2023. Text2video-zero: Text-to-image diffusion models are zero-shot video generators.

Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. 2023. Videofusion: Decomposed diffusion models for high-quality video generation.

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.