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

1 Tasks that have been performed

Since the beginning of the project, much progress has been made in our goal of solving our original problem statement.

- **Text-to-video pipeline:** The first half of the generative task was to have a functioning text-to-video model that could generate video samples of reasonable quality given a prompt. We reviewed three architectures, VideoFusion (Luo et al., 2023), Make-a-video (Singer et al., 2022) and Phenaki (Villegas et al., 2023), and assessed their suitability for our current task. Given the constraint on the computational resources available to us, we were able to achieve impressive results (Chouradiya et al., 2023d) with VideoFusion (Luo et al., 2023) (Implementation snapshot: (Chouradiya et al., 2023c)).

- **Text-to-music pipeline:** The latter phase of the generative task was dedicated to the implementation of a robust text-to-music model, with the goal of producing audio samples of decent quality in direct response to textual cues. In response to this, we conducted a comprehensive investigation into various facets, assessing their alignment with the project's objectives. Two text-to-audio models, namely Make-an-Audio (Huang et al., 2023) and AudioGen (Kreuk et al., 2023), were examined in conjunction with two text-to-music models, MusicLM (Agostinelli et al., 2023) and MusicGen (Copet et al., 2023). Our current emphasis is directed towards generating suitable musical cues for the produced videos. The compelling results achieved with MusicGen, along with its demonstrated adaptability in generating diverse musical audio cues (Chouradiya et al., 2023a) for a wide spectrum of textual prompts, underpin the robustness of our approach in this undertaking. (Implementation snapshot: (Chouradiya et al., 2023b)).

2 Risks and challenges

- **Absence of prior baseline:** In our literature survey, the task of getting synchronized video and audio from text prompts has not been proposed. The absence of a reference point complicates the task of demonstrating the effectiveness and significance of the proposed approach. It can be challenging as this may raise questions about the novelty and relevance of the work.

- **Computational deficiency:** Computational deficiency poses a significant challenge for training models to synchronize audio and video generation tasks. The resource-intensive nature of these tasks demands substantial computing power, which may not be readily available or feasible for us. This deficiency hampers the ability to train and fine-tune these models effectively.

- **No well-defined evaluation metric:** Unavailability of a baseline deters how we can quantify our results for this task. The coherence of synchronized audio and video is inherently subjective. Different people may have varying perceptions of quality, realism, and aesthetics, making it challenging to establish a universally objective evaluation metric. Unlike text-based or tabular data, there is often no ground truth reference to compare against. In many cases, it's hard to define an absolute gold standard for this task. Hence, for our task, we must consider how well the modalities complement each other and whether they create a coherent experience.

3 Plan to mitigate the risks and address the challenges

- **Overcoming Baseline Absence:** Given the absence of a baseline, our group project addresses this challenge by exploring two promising avenues. Firstly, we're implementing Generative Adversarial Networks (GANs) to evaluate their potential in generating desired outputs. Secondly, we're leveraging semantic segmentation as prompts for the models, aiming to assess its effectiveness as an alternative approach. This dual-pronged strategy allows us to comprehensively evaluate and compare the performance of both methods, ultimately paving the way for a more informed and robust solution. Through this systematic approach, we aim not only to mitigate the challenges posed by the lack of a baseline but also to gain valuable insights into the most effective techniques for our specific task.

- **Strategic Focus:** Given the substantial computational resources required for video processing, our group project is adopting a strategic approach to mitigate this challenge. Instead of allocating extensive resources to train and fine-tune video data, we're prioritizing the synchronization of generated audio with video samples. By concentrating on this aspect, we aim to achieve a more efficient use of resources while still producing high-quality results.

- **Human-Centric Evaluation:** In light of the inherent subjectivity in assessing good synchronicity, our group project opts for a human-centric approach. We plan to conduct thorough evaluations through surveys, drawing on the collective insights of our team members to gauge the quality of synchronicity in our outputs. This methodology not only leverages human perception, which is crucial in tasks like these but also allows for a comprehensive and nuanced assessment. We plan to use Qualtrics or Google Forms as survey platforms.

4 Individual Contributions

The text-to-video pipeline will be created and maintained by Pranav and Jehil. They will also implement the GAN architecture as one approach to solving the problem. Saad, Arham and Sahil are responsible for the text-to-audio pipeline, implementing a semantic segmentation approach that integrates into the existing audio and video pipelines, and researching the available options for setting up surveys (Qualtrics, Google Forms, etc.) to collect feedback on the generated results.

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.

Arham Chouradiya, Jehil Vora, Pranav Parnerkar, Saad Shaikh, and Sahil Mondal. 2023a. Text to audio results.

Arham Chouradiya, Jehil Vora, Pranav Parnerkar, Saad Shaikh, and Sahil Mondal. 2023b. Text to music implementation screenshot.

Arham Chouradiya, Jehil Vora, Pranav Parnerkar, Saad Shaikh, and Sahil Mondal. 2023c. Text to video implementation screenshot.

Arham Chouradiya, Jehil Vora, Pranav Parnerkar, Saad Shaikh, and Sahil Mondal. 2023d. Text to video results.

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

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.

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. 2023. Audiogen: Textually guided audio generation.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. 2022. Make-a-video: Text-to-video generation without text-video data.

Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. 2023. Phenaki: Variable length video generation from open domain textual descriptions. In *International Conference on Learning Representations*.