

Chhaya: Photorealistic Contextual Animation Generation

Md Zunayedul Islam
Department of Computer Science
BRAC University
Dhaka, Bangladesh
md.zunayedul.islam@g.bracu.ac.bd

Abstract—We introduce Chhaya, a derived process that allows to build a deep learning model that is able to create photorealistic frames of contextual animation. This model should be able to go through the textual description given by the user and generate photorealistic animation using deep understanding. This will allow people’s creativity and research taken to another level.

Index Terms—animation, image, description, network, frames

I. INTRODUCTION

After the introduction of photorealistic text-to-image diffusion models [1], the research on this topic has become so popular that it led to industrial commercialization and constant creative improvements as per usage. Although text-to-image generation is still under research, our imagination does not stop here. If we can generate images from textual descriptions, can we animate them as well? Generating animation from scratch can be extremely difficult for a computer given the current technology we have. But can we make several frames of images from textual description so that it behaves like an animation? In this paper, we have generated images of same description several times and converted it into an animation using GLIDE method [2].

II. RELATED PREVIOUS WORKS

Ho et al. (2022) proposed and described Imagen Video, a text-conditional video generation system that uses a cascade of video diffusion models to generate high-definition videos. The authors claimed their system is designed to be scalable and to transfer findings from previous work on diffusion-based image generation to the video generation setting. It is capable of generating high-fidelity videos with a high degree of controllability and world knowledge, including the ability to generate diverse videos with text animations and 3D object understanding. The authors apply progressive distillation to their video models for fast, high-quality sampling. Overall, the authors presented a promising approach for generating videos from text. [3]

Singer et al. (2022) introduced Make-A-Video, an approach for generating videos from text that takes advantage of recent progress in text-to-image (T2I) generation. This approach uses paired text-image data to learn what the world looks like and how it is described, and unsupervised video footage to learn

how the world moves. The approach has several advantages, including the ability to accelerate training of the T2V model, not requiring paired text-video data, and generating videos with the diversity of today’s image generation models. The authors designed a spatial-temporal pipeline to generate high-resolution, high-frame-rate videos. The authors built 2 new datasets based on MNIST and CATER for their model. They claims that Make-A-Video sets the new state-of-the-art in text-to-video generation. The research paper was written professionally and in a simple standard. The work is done in details and seems effective. [4]

Hu et al. (2022) proposed a novel video generation task called Text-Image-to-Video generation (TI2V) that aims to generate videos from a static image and a text description. According to them, the key challenges of this task lie in aligning appearance and motion from different modalities and in handling uncertainty in text descriptions. To address these challenges, the authors propose a Motion Anchor-based video GEnerator (MAGE) that uses an innovative motion anchor structure to store appearance-motion aligned representations and allows for the injection of explicit conditions and implicit randomness. The authors also build two new datasets for evaluating their method. The authors claim their experiments to show that MAGE is effective and that the TI2V task has potential. Their approach is generic and but the fact that they created datasets for their own was itself a dedicated approach. [5]

Fu et al. (2022) introduced a novel task called text-guided video completion (TVC), which involves generating a video from partial frames guided by a natural language instruction. To address this task, the authors propose a model called Multimodal Masked Video Generation (MMVG) that discretizes the video frames into visual tokens and masks most of them during training to perform video completion from any time point. At inference time, a single MMVG model can address all three cases of TVC (video prediction, rewind, and infilling) by applying the appropriate masking conditions. The authors evaluated MMVG on various video scenarios and show that it is effective in generating high-quality visual appearances with text guidance for TVC. their approach is quite different than most other from the year of publication. Although the quality of the output videos seem questionable. [6]

Villegas et al. (2022) presented Phenaki, a model that can

generate realistic videos from a sequence of textual prompts. To address the challenges of generating videos from text, such as the computational cost and limited quantities of high-quality text-video data, the authors introduce a new model for learning video representation that compresses the video into a small set of discrete tokens. This tokenizer uses causal attention in time, which allows it to work with variable-length videos. To generate video tokens from text, the authors use a bidirectional masked transformer conditioned on precomputed text tokens. The generated video tokens are then detokenized to create the actual video. The authors demonstrate that joint training on a large corpus of image-text pairs and a smaller number of video-text examples can result in generalization beyond what is available in the video datasets. The authors claim that compared to previous methods, Phenaki can generate arbitrary-length videos from time-variable prompts in the open domain and produces better spatiotemporal consistency than per-frame baselines. Overall, the authors presented a promising approach for generating videos from text. [7]

Kim et al. (2020) proposed a novel training framework called Text-to-Image-to-Video Generative Adversarial Network (TiVGAN) for generating videos based on text descriptions. The authors trained their model gradually on more and more consecutive frames, starting with a single video frame, and this step-by-step learning process helps stabilize the training and enables the creation of high-resolution video based on the given text descriptions. The authors claim that experiments on various datasets show that the proposed method is effective. Although their use in proposed model was GAN, the quality of the videos is not high. [8]

Chen et al. (2020) proposed a novel Bottom-up GAN (BoGAN) method for generating videos from a text description. To ensure the coherence of the generated frames and match the language descriptions semantically, the authors design a bottom-up optimization mechanism to train BoGAN. This mechanism includes a region-level loss via attention mechanism to preserve local semantic alignment and draw details in different sub-regions of the video, as well as a frame-level and video-level discriminator to maintain the fidelity of each frame and the coherence across frames. The authors evaluate the effectiveness of BoGAN on two synthetic datasets and two real-world datasets. Although they used GNA in their model, the model seems to be suitable for only comparatively smaller textual descriptions. [9]

Deng et al. (2019) proposed and described a novel approach called Introspective Recurrent Convolutional GAN (IRC-GAN) for generating videos from given text. The authors used a recurrent transconvolutional generator that integrates LSTM cells with 2D transconvolutional layers to take both the definition of each video frame and temporal coherence into account, resulting in videos with better visual quality. They also uses mutual information introspection to measure the semantic consistency between the generated videos and the corresponding text. They compiled experiments on three datasets show that IRC-GAN is effective at generating plausible videos from given text and compares favorably with state-

of-the-art methods. At the time of publication, the approach of GAN described in this paper was a success. But it appears that latest approaches on GAN performs a lot better than this one. [10]

III. METHODOLOGY

In this project, we have used the GLIDE approach for generating frames for our desired animation. GLIDE (Guided Language-to-Image Diffusion for Generation and Editing), a diffusion model achieves performance comparable to DALL-E despite utilizing only one-third of the parameters. In addition to producing images from text, GLIDE may be used to change existing images by using natural language text prompts to insert new objects, add shadows and reflections, conduct image inpainting, and so on. It can also convert basic line drawings into photorealistic photos, and it has powerful zero-sample production and repair capabilities for complicated circumstances.

GLIDE uses diffusion models [11] with deep neural network in it called *Imagen*. Imagen is trained on pre-trained BERT for text embedding. It is trained on the correlation between text and images. After GLIDE trained itself, it takes a text as sample input and generates N number of images of the same context. Then we have taken all images into account and combined them to produce an animation into graphics interchange format.

IV. RESULT

For testing, we have given the following text as an input and generated 6 frames for instance in order to make the animation.

"A black bird flying freely in the sky"

Then our model produces the following frames for our animation.



Fig. 1. Frames of our animation

V. DISCUSSION

Although our approach could somehow generate frames for our animation, they are yet to be clearer and more contextual.

The generated animation appears like a slideshow of images instead of an actual animation. We had to derive Imagen diffusion models and pre-trained BERT for training GLIDE, due to the limitation of time and resources. But it is believed that with more time, resources and manpower, we will be able to improve this and make a better approach in text-to-animation synthesis.

VI. CONCLUSION

As the name suggests, this is a project that is aimed to build *Chhaya*, an approach to produce compositional animations from textual descriptions given by the users as per their imaginations. This project will help people enhancing their creativity and imaginations about what is possible and what not. It will also help in academic presentations for students and researchers. With further development of this project and meeting its limitations, this can be treated as an easy tool for commercial production in film industries as well. Although a study has been done on Guided Language-to-Image Diffusion for Generation and Editing, this project will take animation production to the next level by using photorealistic image frames and merging them together. The output of this project will be unique every time and generate varieties of beautiful contextual animations in real time that does not exist in the real world opening windows to people's creative imaginations.

REFERENCES

- [1] Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo J. Kim, and Sung-Hoon Yoon. Perception prioritized training of diffusion models. *ArXiv*, abs/2204.00227, 2022.
- [2] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models, 2021.
- [3] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.
- [4] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.
- [5] Yaosi Hu, Chong Luo, and Zhenzhong Chen. Make it move: controllable image-to-video generation with text descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18219–18228, 2022.
- [6] Tsu-Jui Fu, Licheng Yu, Ning Zhang, Cheng-Yang Fu, Jong-Chyi Su, William Yang Wang, and Sean Bell. Tell me what happened: Unifying text-guided video completion via multimodal masked video generation. *arXiv preprint arXiv:2211.12824*, 2022.
- [7] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399*, 2022.
- [8] Doyeon Kim, Donggyu Joo, and Junmo Kim. Tivgan: Text to image to video generation with step-by-step evolutionary generator. *IEEE Access*, 8:153113–153122, 2020.
- [9] Qi Chen, Qi Wu, Jian Chen, Qingyao Wu, Anton van den Hengel, and Minghui Tan. Scripted video generation with a bottom-up generative adversarial network. *IEEE Transactions on Image Processing*, 29:7454–7467, 2020.
- [10] Kangle Deng, Tianyi Fei, Xin Huang, and Yuxin Peng. Irc-gan: Introspective recurrent convolutional gan for text-to-video generation. In *IJCAI*, pages 2216–2222, 2019.
- [11] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, Seyedeh Sara Mahdavi, Raphael Gontijo Lopes, Tim Salimans, Jonathan Ho, David Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. 2022.