Text-to-Video Generation through Advanced Diffusion Models

Arnab Halder (ash186), Ashwin Patil (aap327), Harvish Jariwala (hj389), Jay Patil (jsp255), Sanchay Kanade (sk2656)

Abstract

This report examines the advancement of text-to-video (T2V) creation using the "Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation" approach. The major goal is to research and refine this unique approach by customizing it to certain data domains in order to improve the quality and relevance of created films. T2V generating systems' performance is assessed and enhanced through rigorous experimentation using domain-specific datasets. Furthermore, the potential for current text-to-image generating breakthroughs is investigated to improve the outcomes. The main difficulty is the complexity of directly generating movies from textual descriptions, despite tremendous advances in text-to-image conversion utilizing AI models such as Stable Diffusion. The emphasis is on short-form video generation using simple textual explanations, with an extra sophisticated function that allows for the impact of video style via an additional reference video. The scarcity of large-scale datasets with high-quality text-video pairs, the complexities of modeling higher-dimensional video data, and the uncertainties surrounding video captioning have all been cited as major hurdles in T2V production. This study discusses the insights gained from tackling these problems, as well as recommendations for future research in the dynamic field of T2V creation.

1. Introduction

Generative AI has progressed from creating static images to the more complex task of generating video content. This report examines the cutting-edge “Tune-A-Video” framework for text-to-video(T2V) generation. Our main objective is to research and improvise the novel approach by customizing it for particular data domains, aiming to elevate the quality and relevance of generated videos. We assess the performance of T2V generating systems and enhance them through rigorous experimentation with domain-specific datasets. Our study also explores how recent advancements in text-to-image generation can be leveraged to augment T2V outcomes. We address the high-dimensional complexity of video data, the challenges in modeling these data and the uncertainties of video captioning as significant hurdles in T2V production. This introduction sets the stage for discussing our methodologies, the framework used, and the insights gained from our experiments.

2. Related Work

Several significant contributions in the realms of diffusion models, text-to-image, and text-to-video generation have laid the foundation for the "Tune-A-Video" project and its subsequent advancements. Below are summaries of these pivotal works.

- **Evolution of Text-to-Image Models: The First Wave -** The first wave of text-to-image models revolutionized the field by enabling the generation of realistic images from textual descriptions. VQGAN-CLIP introduced a novel approach combining vector quantization and contrastive learning to produce high-quality images guided by textual prompts. XMC-GAN employed cross-modal alignment and attention mechanisms to generate visually coherent images based on textual inputs. GauGAN2 extended the capabilities by enabling users to interactively edit and manipulate generated images, enhancing the creativity and versatility of the models.
- **DALL-E Redefining Text-to-Image Generation -** DALL-E, introduced by OpenAI, represents a significant milestone in text-to-image generation. By leveraging a large-scale transformer architecture, DALL-E demonstrated the ability to generate diverse and contextually relevant images from textual prompts. Its innovative approach allowed for the creation of novel visual concepts and scenarios, ranging from surreal landscapes to abstract compositions. DALL-E showcased the potential of transformer-based models in understanding and translating textual descriptions into rich visual representations, opening new avenues for creative expression and content generation.
- **Advancements in Diffusion Models Pioneered by Stable Diffusion with DALL-E2 -** The emergence of diffusion models ushered in a new era of text-to-image generation, characterized by enhanced stability and fidelity in image generation. Pioneered by Stable Diffusion with DALL-E2 in 2022, these models utilize probabilistic diffusion processes to generate high-resolution images from textual descriptions. By iteratively refining a noise signal, diffusion models achieve impressive results in producing realistic and diverse visual content. The integration of diffusion techniques with DALL-E2's transformer architecture further improves the coherence and quality of generated images, marking a significant advancement in the field.
- **From GANs to Diffusion: A Paradigm Shift in Text-to-Image Generation -** There has been a notable shift in the text-to-image generation

paradigm from traditional GAN-based approaches to diffusion models. While GANs have been successful in generating visually compelling images, they often suffer from mode collapse and instability issues. Diffusion models address these challenges by adopting a probabilistic framework that iteratively refines noise signals to generate images. This shift represents a paradigmatic change in the field, offering improved stability, scalability, and control over the image generation process. Diffusion models have demonstrated superior performance in generating high-fidelity images from textual prompts, highlighting their potential as the next frontier in text-to-image generation.

3. Methodology

a. Input

The T2V production process starts with a primary input in the form of a text prompt, like "A cat chasing a mouse through the garden." A short reference video can be provided to guide the style of the resulting output.

b. Output

The T2V creation procedure produces a brief video clip that visually correlates with the specified text prompt. Furthermore, if a reference video is provided, the resulting output may reflect its style.

c. Framework

The Tune-A-Video framework is the cornerstone of our text-to-video (T2V) creation process, utilizing advances in pre-trained Stable Diffusion models. This framework offers a systematic technique to fine-tuning these models for the specific goal of creating videos from textual descriptions.

d. Modifications

We endeavored to replicate the architecture of the author's implementation, yet we introduced several modifications due to limitations in memory and computational resources:

i. Reduced the number of frames in the decoder output from 16 to 12:
This adjustment was necessary to alleviate memory constraints and enhance computational efficiency. By reducing the number of frames in the decoder output, we aimed to optimize memory usage without compromising the overall performance of the model.

ii. Increased the frame skip in the frame interpolation network from 3 to 5:

The decision to augment the frame skip parameter was driven by the need to balance computational demands with model accuracy. By increasing the frame skip from 3 to 5, we aimed to reduce the computational burden of frame interpolation while still preserving the fidelity of the generated frames.

iii. Utilized our own dataset for fine-tuning the attention blocks:

To tailor the model to our specific requirements and improve its performance on our target tasks, we opted to fine-tune the attention blocks using our proprietary dataset. This approach allowed us to adapt the model's attention mechanisms to better capture relevant features and nuances present in our data, thereby enhancing its effectiveness for our intended applications.

e. Technique

i. **Model Fine-tuning (Training):** To train our T2V generating system, we use model fine-tuning. Initially, we use a pre-trained Stable Diffusion model and fine-tune it with a curated video dataset accompanied by appropriate written descriptions. This fine-tuning procedure allows the model to develop linkages between video frames and their accompanying verbal prompts, which improves its capacity to generate cohesive video sequences.

ii. **Inference (Generation):** After fine-tuning, the model is ready for inference, also known as video generation. At this point, we add new text prompts to the fine-tuned model. The model generates sequential frames that, when combined, constitute a complete video sequence that matches the textual description provided.

4. Data

We have only trained the temporal part of the dataset (attention block) and all the other layers were frozen.

1. **Unlabeled video data**
    - Animal Kingdom: We have used Animal Kingdom dataset for animal action recognition, which has about 50 hours of video data consisting of around 100 animal species and 140 action categories.

- UCF101: We have also used UCF101 that consists of about 13k videos belonging to around 100 action classes.

2. **Evaluation**
    - Kinetics: We used 20-25 samples from Kinetics dataset for generating video samples and compare the original videos with the generated videos.

5. Results

Some qualitative results from our implementation can be found below:

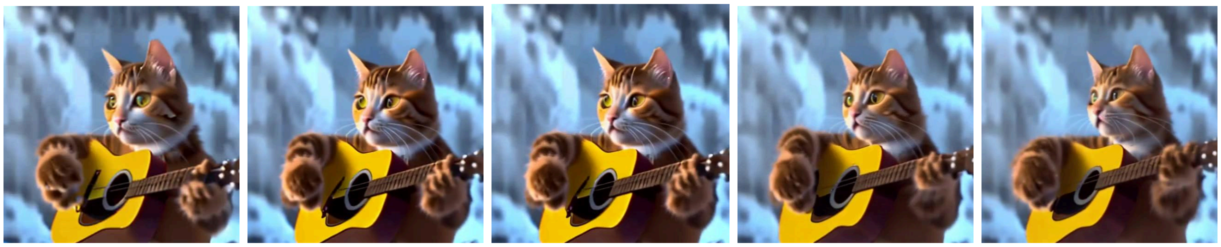

Fig. 1 Frames from the video generated when given the prompt “A cat is playing a guitar”

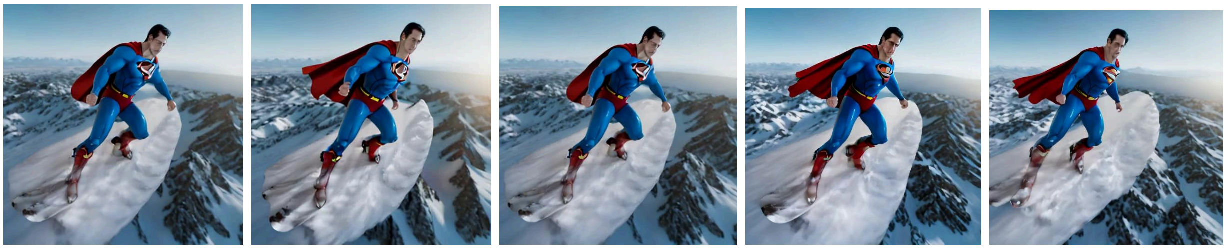

Fig. 2 Frames from the video generated when given the prompt “Superman is Skiing”

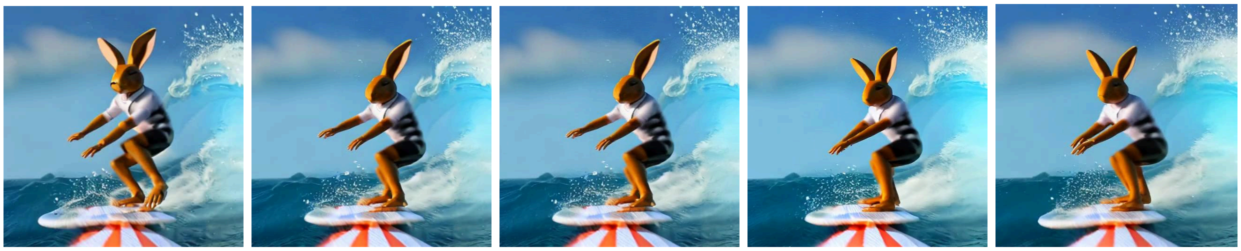

Fig. 3 Frames from the video generated when given the prompt “Rabbit surfing, cartoon style”

Some quantitative results from our implementation can be found as below:

Method	**Frame Consistency (CLIP Score)**	**Textual Alignment (CLIP Score)**
Our Method (Implementation)	88.1	25.3
One Shot Tuning	92.4	27.58

Table 1: CLIP scores for frame consistency and textual alignment for One Shot Tuning and our method

Because of the restricted data and resources utilized during training, our model exhibits a slight decrease in performance compared to One Shot Tuning. This slight decline in performance can also be attributed to alterations made by us in the architecture to that from the author's initial implementation.

6. Future Work

- **Architecture changes -** Expanding the number of frames produced by the decoder could result in videos with finer detail and smoother motion. By boosting the temporal resolution of the generated videos, models can capture subtler nuances in movement, leading to more lifelike and immersive visual sequences.
- **Increase the number of frames in the decoder -** Expanding the number of frames produced by the decoder could result in videos with finer detail and smoother motion. By boosting the temporal resolution of the generated videos, models can capture subtler nuances in movement, leading to more lifelike and immersive visual sequences.
- **Use Spatio-temporal super-resolution from Tune-a-video -** Incorporating spatio-temporal super-resolution techniques from Tune-a-video can enhance the visual fidelity and resolution of generated videos. Leveraging advanced algorithms for upscaling both spatial and temporal dimensions enables models to produce videos with sharper detail and higher overall quality, improving the viewing experience.
- **Scaling Up -** Future investigations might focus on scaling up text-to-video models by leveraging larger datasets and employing more advanced model architectures. Scaling up could lead to enhancements in both the quality and diversity of generated videos, as larger datasets provide models with a richer variety of training examples, while more powerful architectures enable better capture of complex patterns and relationships.

- **Complex Actions -** Addressing the challenge of generating longer videos and depicting more intricate interactions described in the prompts is a critical area for further exploration. Models need to maintain coherence and consistency over extended video sequences, accurately portraying the progression of actions and events described in the text prompts to deliver compelling and immersive videos.
- **Style Control -** Refining techniques for more precise control over the stylistic attributes of generated videos is crucial for enabling users to customize the visual output according to their preferences. Future research could focus on developing methods for fine-tuning style manipulation, allowing users to specify desired aesthetic characteristics such as color palettes, lighting effects, and artistic styles with greater accuracy and flexibility.

7. References

1. Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, Mike Zheng Shou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp
2. Rajat Arora and Yong Jae Lee. Singan-gif: Learning a generative video model from a single gif. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 1310–1319, 2021. 3
3. Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. arXiv preprint arXiv:2204.14217, 2022. Chen, Haoxin, et al. "Videocrafter2: Overcoming data limitations for high-quality video diffusion models." arXiv preprint arXiv:2401.09047 (2024).
4. Khachatryan, Levon, et al. "Text2video-zero: Text-to-image diffusion models are zero-shot video generators." arXiv preprint arXiv:2303.13439 (2023).
5. Uriel Singer et al. “Make-A-video: Text-to-video generation without text-video data.”, Sept. 2022
6. A beginner’s guide to language models. URL: https://builtin.com/data-science/beginners-guide-language-models.
7. Xiangyu Chen et al. Activating more pixels in image Super-Resolution Transformer. Mar. 2023.URL: https://arxiv.org/abs/2205.04437.