Video Generation via Compressed Hand-Drawn Representations and Latent Diffusion Models

1st Tofara Moyo
Bulawayo , Zimbabwe
tofaramoyo@gmail.com, Mazusa AI

***Abstract*—We present a novel approach to video generation, leveraging compressed hand-drawn representations and latent diffusion models. Our methodology employs a unique two-stage process, wherein a variational auto encoder generates parameters based on input text, of a generic equation to be graphed into a frame, and a latent diffusion model refines these frames into photorealistic video content. These graphs are designed to look like hand drawn replicas of the frames in the dataset. By utilizing hand-drawn-like images as a compressed representation, we effectively reduce the dimensionality of the video generation problem, enabling tighter bottleneck architectures and improved efficiency. Our approach demonstrates significant potential for generating lenghty ,high-quality, text-conditioned videos, with applications in multimedia creation, robotics, and beyond.**

I. Introduction

The release of ChatGPT by OpenAI in 2022 marked a significant milestone in the development of generative models. Since then, substantial progress has been made in image generation, with models demonstrating exceptional capabilities in interpreting complex prompts and producing high-fidelity outputs.

Recent breakthroughs have also been achieved in video generation, with notable advancements in both duration and coherence. Specifically, video generation models have evolved from producing 5-second clips to generating 2-minute coherent videos, as exemplified by OpenAI's Sora and Google's VEO. These state-of-the-art video generation models are grounded in latent diffusion models, which employ a bottleneck mechanism to compress videos while preserving their essential characteristics.

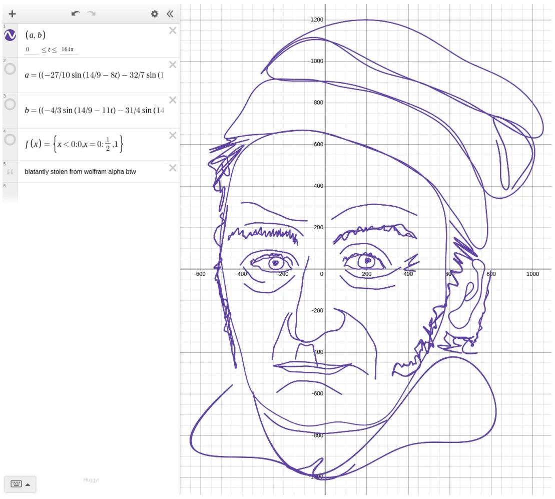
The bottleneck representation maintains the fundamental attributes of the original video, enabling the efficient generation of coherent and contextually relevant multimedia content. This development has profound implications for various applications, including multimedia creation, editing, and analysis.

This study investigates an alternative approach to representing the information encoded in the bottleneck of a latent diffusion model. We begin by acknowledging the necessity of a bottleneck mechanism due to the inherently high-dimensional nature of video data. Specifically, a typical high-definition (HD) video frame comprises approximately 1 billion parameters.

Furthermore, with 30 frames per second, video generation poses an extremely high-dimensional problem, characterized by an immense parameter space. This inherent complexity necessitates the use of a bottleneck or dimensionality reduction mechanism to facilitate efficient video generation and processing.

Notably, simple mathematical equations can be employed to generate handwritten-like drawings, akin to sketches. When considered as high-definition (HD) images, these drawings comprise a vast number of pixels, typically on the order of 10^9. However, despite this high pixel count, it is possible to accurately estimate the handwritten image using a remarkably compact representation, often requiring fewer than 300 parameters.

To illustrate this concept, Figure 1 presents a rendered image of the renowned mathematician Carl Friedrich Gauss, generated using equations of this form. This example highlights the impressive representational efficiency achievable with handwritten-like drawings, which can be effectively captured using a relatively small number of parameters.



1 (a, b)
$0 \le t \le 164\pi$

2 $a = ((-27/10 \sin(14/9 - 8t) - 32/7 \sin(1$

3 $b = ((-4/3 \sin(14/9 - 11t) - 31/4 \sin(14$

4 $f(x) = \left\{x < 0 : 0, x = 0 : \frac{1}{2}, 1\right\}$

5 blatantly stolen from wolfram alpha btw

6

As you can see this is an effective way of compressing an HD image from a billion parameters to fewer than 300.

II. Methodology

Our approach leverages a dataset of paired diagrams and their hand-drawn counterparts, obtained by crowd sourcing individuals to create simplified representations of key image components. Unlike conventional applications of such datasets, which focus on training algorithms to generate photorealistic images from simple drawings, our objective is to develop an algorithm capable of producing hand-drawn-like images from realistic inputs.

To achieve our goal, we employ a two-stage process. Initially, we utilize a video generation dataset, applying our hand-drawn image generation algorithm to each frame. This yields a compressed representation of the video content.

We then parameterize each frame using a generic equation comprising fewer than 300 parameters. By estimating these parameters for each frame and calculating a loss function based on the discrepancy between the hypothesized and actual frames, we establish a foundation for further processing.

Subsequently, we train a Variational Autoencoder (VAE) to generate batches of parameters that form the frames of novel videos. This process is conditioned on simultaneously input text.

In parallel, we train a Latent Diffusion Model (LDM) on the full video dataset, conditioning each video on its corresponding frames. Notably, this approach enables the utilization of a tighter bottleneck for long sequences, as the conditioning frames capture most of the video's information.

Furthermore, we condition the LDM on the same text used to generate frames in the VAE.Our final algorithm employs a two-stage generation process. Initially, the VAE generates parameters for video frames, optionally conditioned on text. These parameters are then converted into frames, which serve as conditioning inputs for the LDM in the final stage, producing the generated video.

III. Conclusion

This study introduces a novel approach to text-conditioned video generation, harnessing the power of compressed hand-drawn representations and latent diffusion models. By leveraging the unique characteristics of hand-drawn-like images, we effectively reduce the dimensionality of the video generation problem, enabling the development of more efficient and scalable architectures.

Our proposed methodology demonstrates significant potential for generating high-quality, text-conditioned videos of long lenghts, with applications in multimedia creation, robotics, and beyond. The use of compressed hand-drawn representations offers a promising direction for future research, particularly in the context of multimodal learning and generative models.

Ultimately, our study contributes to the growing body of research on text-conditioned video generation, offering a novel and innovative approach that pushes the boundaries of what is possible in this exciting and rapidly evolving field.

References

[1] Comparative Analysis of CHATGPT and the evolution of language models Oluwatosin Ogundare, Gustavo Quiros Araya

[2] Sora: A Review on Background, Technology, Limitations, and Opportunities of Large Vision -Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, Lifang He, Lichao Sun

[3] Imagen Video: High Definition Video Generation with Diffusion Models -Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, Tim Salimans

[4] High-Resolution Image Synthesis with Latent Diffusion Models-Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, Björn Ommer

[5] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," nature, vol. 521, no. 7553, pp. 436–444, 2015.

[6] SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis-Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, Robin Rombach