[Samsung PRISM] Preliminary Discussion

Develop a Quality Assurance framework for Video Generation Models

Team

1. College Professor(s):
   1. Prof. Joshva Devadas T/ joshvadevadas.t@vit.ac.in
   2. Prof. Balaji G N/ balaji.gn@vit.ac.in
2. Students:
   1. Parv Aggarwal / parv.aggarwal2021@vitstudent.ac.in
   2. Alisa Banerjee / alisa.banerjee2021@vitstudent.ac.in
   3. Arya Mukhopadhyay / arya.mukhopadhyay2021@vitstudent.ac.in
3. Department:

   SCOPE,SCORE

Date: 18 Dec 2023

Develop a Quality Assurance framework for Video Generation Models

Problem Statement

Context

Develop a comprehensive framework for conducting equitable and standardized comparisons of AI Generated Videos, ensuring 'apple-to-apple' assessments. This framework should establish consistent evaluation metrics, benchmark datasets, and evaluation protocols to facilitate a fair and meaningful comparison of various video generation techniques. The goal is to advance the field of computer vision by providing a reliable means of model assessment and selection.

Statement

Currently Video Generation Models' QA is mainly subjective in Nature, Building a Framework consisting of SOTA metrics will enable easier comparison of Image Generation Models

Worklet Details

Duration (Months)	Members Count	Mentors
6	4	Pranal Prasad Dongare ☎ +91-7022250561 ✉ pranal.p@samsung.com Tushar Madaan ☎ +91-9205301569 ✉ tushar.m2@samsung.com

Pre-Requisite

- Familiar with Coding predominantly Python
- Familiar with Video Generation Space & Evaluation Metrics

Expectations

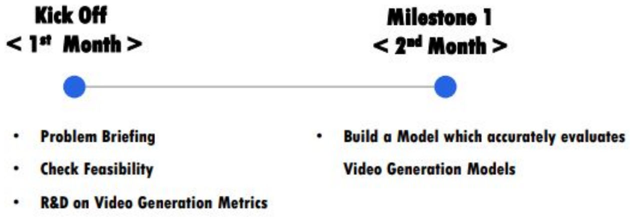
Undertaken Tasks

- Exhaustive Literature Survey of Image Generation Metrics [VideoCLIP]
- Develop a AI backed model for Quality Assurance & Identify SOTA Benchmark Levels

KPI

- Model should consider all possibilities & parameters of the Video Generation Space

Timeline



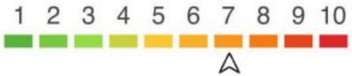
Kick Off
< 1st Month >

- Problem Briefing
- Check Feasibility
- R&D on Video Generation Metrics

Milestone 1
< 2nd Month >

- Build a Model which accurately evaluates Video Generation Models

Complexity



Literature survey and study

- **Major Observations / Conclusions:**

Title: PHENAKI: VARIABLE LENGTH VIDEO GENERATION FROM OPEN DOMAIN TEXTUAL DESCRIPTIONS
Date of Publication: 5 Oct 2022 (Google Research)
Observations: Phenaki is a model that uses tokenization to compress videos into discrete tokens. It employs causal attention for variable-length videos and generates video tokens from text using a bidirectional masked transformer. Joint training on image-text pairs and a small set of video-text examples enhances generalization. Phenaki stands out by generating arbitrary long videos based on text prompts in an open domain, surpassing previous methods in flexibility and performance.
Link: https://sites.research.google/phenaki/

Title: Generative Rendering: Controllable 4D-Guided Video Generation with 2D Diffusion Models
Date of Publication: 3 Dec 2023 (Featured on Huggingface)
Observations: This study merges dynamic 3D meshes' controllability with emerging diffusion models for realistic animated content. Using a low-fidelity animated mesh, ground truth information is injected into a text-to-image model, yielding high-quality, consistent frames. Effective for motion scenarios with rigged assets or varied camera paths, this method enhances controllability and expressivity.
Link: https://huggingface.co/papers/2312.01409

Title: VideoBooth: Diffusion-based Video Generation with Image Prompts
Date of Publication: 1 Dec 2023
Observations: VideoBooth, a proposed model, embeds image prompts coarsely for high-level information and finely for detailed multi-scale encoding. The attention injection module refines details in the first frame, ensuring temporal consistency. VideoBooth excels in generating high-quality videos for specified subjects in image prompts, showcasing state-of-the-art performance. It is a versatile framework that works across various image prompts with a single model and a feed-forward pass.
Link: https://doi.org/10.48550/arXiv.2312.00777

Literature survey and study

Title: Alias-Free Generative Adversarial Networks
Year of Publication: 2021 (ACM Transactions on Graphics, Volume: 39 Issue: 4)
Observations: In this paper, a novel approach to generative adversarial networks (GANs) is presented to eliminate aliasing—an unwanted effect causing fine details in generated images to appear fixed on the screen rather than transforming naturally. By eliminating all positional references, the model achieves the capability to generate images with a more natural hierarchy of details. This results in internal representations that accurately connect details to underlying surfaces, enhancing the coherence and realism of the generated images.
Link: https://proceedings.neurips.cc/paper_files/paper/2021/hash/076ccd93ad68be51f23707988e934906-Abstract.html

Title: Learning Temporal Coherence via Self-Supervision for GAN-based Video Generation
Year of Publication: 2020 (Advances in Neural Information Processing Systems 34 (NeurIPS 2021))
Observations: The paper introduces a smart technique for computers to create videos seamlessly. It's like teaching the computer to make each part of the video smoothly transition to the next, all on its own. This is achieved through a special learning process called self-supervision, which allows the computer to figure out the best way to ensure a natural flow. They use a tool called GAN, where one part acts like an artist creating, and the other part acts like a critic checking for quality. This makes the computer adept at producing videos that not only look great but also flow naturally from one moment to the next.
Link: https://dl.acm.org/doi/abs/10.1145/3386569.3392457

Title: Stagemix video generation using face and body keypoints detection
Year of Publication:25 April 2022 (Multimedia Tools and Applications,Volume 81,Issue 27)
Observations: The proposed method for Stagemix video generation using face and body keypoints detection is a promising approach that can significantly reduce the time and effort required for video editing. The use of deep learning-based keypoints extraction to automatically create Stagemix videos without the need for extensive editing knowledge is a significant contribution to the field of video editing and multimedia applications. The experimental results presented in the paper demonstrate the effectiveness of the proposed method in generating natural and attractive Stagemix videos.

Literature survey and study

Title: Dynamic GAN for high-quality sign language video generation from skeletal poses using generative adversarial networks
Year of Publication:28 June 2022 (Soft Computing - A Fusion of Foundations, Methodologies and Applications, Volume: 26 Issue: 23)
Observations: The dynamic GAN model is a novel generative framework that uses deep generative models to generate high-quality sign language videos from skeletal poses and person images. The model employs a multi-step process that includes mapping of skeletal poses and ground truth images, generation of human-based sign gesture images, image classification and alignment, deblurring techniques, generation of intermediate frames, and a discriminator network that evaluates the generated images and sequences.
Link: https://dl.acm.org/doi/abs/10.1007/s00500-022-07014-x

Title: Swin-GAN: Adversarial Video Generation on Complex Datasets
Year of Publication: July 15, 2019
Observations: It introduces the Dual Video Discriminator GAN (DVD-GAN) model, which aims to generate high-quality video samples at resolutions up to 256 × 256 and lengths up to 48 frames. The authors address the challenges of modeling natural video and demonstrate the capabilities of their model on the Kinetics-600 dataset, achieving state-of-the-art results for video synthesis and prediction tasks. The paper provides insights into the dataset, model architecture, and results, positioning their work within the context of prior research in the field of video generation.
Link: https://arxiv.org/abs/1907.06571

Queries

- **Challenges :**

1. Since there are substantial differences between these methods it is hard to compare them on an equal footing.
2. If the training data isn't diverse or high-quality, the model might not perform well compared to models trained on better datasets.
3. There is simply not enough video data available to cover all the concepts present in text-image datasets
4. Using text prompts alone cannot fully capture the visual characteristics of the image prompt.
5. Generative rendering cannot achieve real-time animations due to the multi-step inference of current diffusion models
6. Model suffers more from temporal changes due to the lack of a correspondence injection module.
7. Aliasing in generative adversarial networks (GANs) lead to less realistic and coherent images.
8. The approach involving training multiple GANs and can be computationally expensive, especially for high-resolution videos.
9. The limitations of the proposed method for Stagemix video generation include the focus on natural frame intersection of a video, which only utilizes singer regions, movement, and choreography information, but does not incorporate other useful information such as music and audience speech
10. The model is not yet able to generate realistic videos in an unconstrained setting. Additionally, the model is computationally expensive, and training on larger datasets or longer videos may require significant computational resources.

Thank you