A Review: Imagen Video: High Definition Video Generation with Diffusion Models

The paper "Imagen Video: High Definition Video Generation with Diffusion Models" provides a novel approach to producing high-quality video. The research uses diffusion models to tackle the issue of producing high-definition videos. The proposed model is based on the Cascade Diffusion technique, which is used to simulate video distribution. The approach works by modelling video distribution as a series of diffusion processes, with each process represented by a set of conditional distributions. The model then uses Progressive Distillation, a technique that employs numerous smaller models to make high-quality videos progressively.

The article is well motivated by the extensive success of text-to-image diffusion models, as well as recent huge advances in diffusion speed and training stability methods. The suggested method is efficient because it can handle high-dimensional issues while keeping each sub-model reasonably simple, and it reduces computation. The difficulty of creating high-quality videos from text is extremely difficult because to a number of reasons, including the necessity for temporal precision and coherence, a better knowledge of language and natural environments, and controllability. Diffusion models have shown to be an effective solution to this problem since they can represent data distribution while dealing with high-dimensional difficulties.

The paper's proposed model architecture is quite interesting. The authors employ Cascade Diffusion, which entails increasing the resolution of an image or video successively through a sequence of super-resolution diffusion models. Text embeddings are used in all models, not only the base model, to ensure that the generated images and videos are true to the input text. The authors additionally employ spatial resizing and frame skipping to reach the appropriate spatial resolution as well as more frames and temporal resolutions.

Cascade Diffusion Training is used to independently train the models, which are trained on both videos and images. The authors pack independent pictures into video sequences while masking off the temporal convolutional residual blocks' computation path. They also prevent cross-frame temporal attention by masking temporal attention maps. V-Parameterization is used for Progressive Distillation to the video models with classifier-free guidance for fast, high-quality sampling. V-Parameterization is intriguing because it helps to eliminate voids and colour-shifting artefacts that are known to impair high-resolution diffusion models, avoids temporal shifting, allows for faster convergence of sample quality measurements, and is beneficial for progressive distillation.

The paper uses an internal dataset consisting of 14 million video-text pairs and 60 million image-text pairs, as well as the publicly available LAION-400M image-text dataset. One of the challenges of working

with these datasets is that they contain offensive content, which the authors have mentioned as a limitation and ethical concern.

The paper's evaluation metrics do not measure generation diversity, which is a weakness of the evaluation. Moreover, the proposed model can only make short videos, and extending the length of attention quadratically increases computation. The study employs suboptimal approaches to achieve temporal coherence and fails to explain why some aspects of its architecture work. Additionally, even with progressive distillation, the suggested architecture is computationally costly, and this limits its applicability in real-world scenarios. The paper lacks insights into why certain parts of the architecture work, which is a limitation of the approach. Finally, the proposed approach models video distribution in the RGB space, whereas there could be a more efficient video generation method by synthesising videos in a low-dimensional latent space.

In summary, the research presents a novel way of producing high-definition videos utilizing diffusion models. The study is well-motivated and employs various innovative techniques, such as Cascade Diffusion and V-Parameterization for Progressive Distillation. Nevertheless, the work has limits in terms of ethical issues, assessment criteria, generation duration, and the suggested approach's processing cost. Future research should concentrate on overcoming these constraints in order to increase the efficiency and efficacy of video generation utilizing diffusion models.