Investigate Video Generation through Machine Imagination

Current Limitations and Possible Future Directions

James Guo
Johns Hopkins University
Baltimore, BA 21218
sguo45@jhu.edu

Abstract

Machine imagination is a new field within machine learning, as it assists on generating texts/images/videos that do not exist based on a database. While scientists are thinking of enhancing and using this technology, we shall evaluate how it is right now and how it will progress in the future. This essay investigates the progress and experiences in this realm and their perspectives on the future of video generation in terms of its application sides.

1 Introduction

As artificial intelligence progresses, language models were trained to generate contents from a specific prompt, which develops into a branch of *Machine Imagination*. Certain initial tasks were to develop textual imaginations, to static figures, and eventually to higher dimensional forms of picture, *i.e.*, 3D worlds or videos. certain language models have been used to generate videos.

While the models progressed, it has gradually been capable of generating (multiple) images that align with the prompt. In particular, in the process of producing images with the *realistic* prompts, the images are capable of obeying the perspective and physical laws. However, when such processes were extended to video generation, which is generating a series of "*continuous*" pictures, these language models are sometimes are not reliable.

Specifically, given the probabilistic noises (ξ) throughout the model, generating a series of picture *often* cannot be a consistent video. Moreover, in (3), the authors mentioned used an example of generation of human figures, where some prevalent issues are with "occlusion issue, body deformation, appearance inconsistency, temporal in-alignment, and unnatural pose." These limitation would allow audiences to identify that the video is *not* real, which diverges from some intentions of generating realistic imaginations.

While given all these limitations, the field of video generation is still contained with large potentials. Also in (3), this technology is believed to be able to generate videos of longer length, more details, and control of interact-ability and controllability for the users when providing a prompt.

2 Analysis

As we explore into video generation as an intersection of machine learning, language model, computer vision, and computer graphics, we shall carefully evaluate its limitations and the potentials, of how it is like currently from the past and how it can be developed as a more powerful tool in the future.

Final project for EN.601.208 Machine Imagination at *Johns Hopkins University*.

2.1 Methodologies of Video Generation

From (4), the basis of video generation is on **Generative Adversarial Networks**, which exploits the uses a generator network (G) and a distinguishing network (D). According to (4), the G network extends a low-dimensional input (text) into a high-dimensional output (video) that preserves translations in space and time, while abiding to laws of how objects shall move; the D network aims to classify the real and generated scenes, which could well evaluate the motions and the step-by-step images. Through this process, it forms a sort of reenforcement learning where the generator is trained to the best of its extent in producing videos that cannot be identified. By (4), some current strategies are to separate between the foreground and background, using encoders, and various approaches.

2.2 Major Limitations of the Models

In (2), the **diffusion model** is described as encoder, adding noise, and the decoder to transform a low dimensional text to a high dimension image/video. Through (2), an experiment was conducted with respect to different conditions (number of object and complexity of prompt) posed on different video generation models to evaluate the object formation and video consistency of the objects. From their experiment (2) concluded that the quality of productions decreases as complexity increases, while certain models generated deformed human bodies (*e.g.* distorted face and legs).

With the evaluations of (5), another major issue is with the length of the video: as the generators are asked to generate the videos, the videos contents and motions start to become uncontrolled beyond the training length. This reveals that the models become impotent as they are asked to generate longer videos and the contents start to deviate more.

2.3 Some Current Strategies

Given such limitations on video generation, there are various strategies right now to overcome the issues. As mentioned in (4), reenforcement learning as a self-supervised learning method is able to improve the video quality as well as the quality of the models. Mentioned in (1), another strategy is to separate the learning process of the video generation, *i.e.*, into motion learning and appearance disentangling stage, while this training pipeline is able to improve the video quality especially with little data involved. This strategy is connecting the spacial and temporal fields altogether when producing what is desired.

2.4 Applications to the Future

Speaking of video generation, the two general goals for the future are: 1. to develop models that can produce realistic, high-quality videos with limited for no data input, and 2. apply such innovation to different and diverse scenarios.

The first part, according to (2; 1), is more dependent on developing and improving different models (**GAN**, **Diffusion-based**, and **Transformer-based**), training better models (*e.g.* having low quality video draft combined with high quality images to train again), and innovations of building more structures to generate models that are more powerful.

The second part is more application-deviated. From (3), the goal is to add flexibility into the video generations, such as having longer, photorealistic videos that can be controlled concurrently by a human during interactions. Such expectations anticipates much higher levels of proficiency and efficiency of the model, leaving an "*unbounded*" upper cap for machine imagination.

3 Conclusion

As an emerging field, video generations through machine imagination is still developing with various new models and new algorithms. While in its early stage, we have to admit that it still embraces various limitations that constraints the model to develop its full potentials. As we obtain more training data and develop different techniques to render videos, it would have more applications to assist the efficient, real-time generation of personalized and high-quality output in the future.

References

[1] H. Chen, Y. Zhang, X. Cun, M. Xia, X. Wang, C. Weng, and Y. Shan, "Videocrafter2: Overcoming data limitations for high-quality video diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7310–7320, June 2024.

[2] A. Srivastava, R. Sista, P. P Chakrabarti, and D. Sheet, "Exploring the limits of vlms: A dataset for evaluating text-to-video generation," in *Proceedings of the Fifteenth Indian Conference on Computer Vision Graphics and Image Processing*, ICVGIP '24, (New York, NY, USA), Association for Computing Machinery, 2025.

[3] W. Lei, J. Wang, F. Ma, G. Huang, and L. Liu, "A comprehensive survey on human video generation: Challenges, methods, and insights," 2024.

[4] C. Vondrick, H. Pirsiavash, and A. Torralba, "Generating videos with scene dynamics," in *Advances in Neural Information Processing Systems* (D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, eds.), vol. 29, Curran Associates, Inc., 2016.

[5] S. Ge, T. Hayes, H. Yang, X. Yin, G. Pang, D. Jacobs, J.-B. Huang, and D. Parikh, "Long video generation withătime-agnostic vqgan andătime-sensitive transformer," in *Computer Vision – ECCV 2022* (S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, eds.), (Cham), pp. 102–118, Springer Nature Switzerland, 2022.