Qeios

Peer Review

Review of: "AniSora: Exploring the Frontiers of Animation Video Generation in the Sora Era"

A. H. Bermano[1]

1. Tel Aviv University, Israel

The paper presents a full framework for developing an animation video model, from data collection through architecture, training, and evaluation. As the paper correctly points out, most models out there specialize in or focus on realistic video generation and demonstrate weaker capabilities when it comes to stylized or animated content.

Hence, this work sets out to fill this gap and describes a very intensive piece of engineering. The paper is written more as a tech report rather than a research paper: it describes the size of the collected data, the architecture of the trained model, and the procedure to evaluate it. These are all highly relevant and important details, but it is clear, at least to this reviewer, that much more experimentation and learning have happened before reaching this model.

The model seems to provide high-quality results, with satisfactory evaluation both on a standard benchmark (vbench) and through subjective human studies. Questions regarding design choices, however, arise at all stages of the description. Perhaps the most important part of the work is data collection. How would incorporating more data influence the results? Why start from cogX and not a different model? Why not incorporate static images as well for some of the processes and layers? What about fixed-length vs. varying-length clips? Why were the specific scores chosen for data filtering and not others? These are all questions I'm sure the developers of this model have thought about and could share insights on with the community, while the dataset section amounts to only 2.5 paragraphs.

The same goes for the description of the model itself. There are several very interesting questions regarding the design and training of the model, yet they are not discussed. A prominent example is where the paper states that video patchification "is a critical step." Indeed, that's probably true, but only a very brief description of the patchification strategy is discussed, with nothing regarding the experience of

 doi.org/10.32388/DXSGVY

choosing it. The description and use of 3D full attention are not clear, but again sound like an interesting design choice with much to learn from it. In 4.3, again, it seems there is an interesting frequency-related bias investigation under the surface that the community could learn a lot from.

The experiments are more detailed. Here, it seems that AniSora has the most advantage in character consistency and motion smoothness. Consistency is to be expected, as the whole point of data collection and training is for the model to understand exactly these appearances better. For smoothness, it is not clear that this is actually a desired feature. As the paper itself states, animations often portray aggregated non-physical motion. Perhaps it would be better to measure discontinuities or jitter. Indeed, it seems there already are newer versions of AniSora out there that demonstrate richer motions.

Overall, this is a thorough and significant effort that the community can definitely gain from. On the academic side, it seems that more discussions could help portray the lessons learned from this effort better.

Declarations

Potential competing interests: No potential competing interests to declare.