Qeios

Peer Review

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

Ondřej Texler[1]

1. Independent researcher

Paper focuses on a niche topic of generating animated, highly-stylized, expressive videos with abstract physics and object interactions. Paper provides a practical end-to-end framework to solve this slightly underexplored task: dataset collection and curation, generation model and training details, and an evaluation scheme.

"We build our animation dataset according to the observation that high quality text-video pairs are the cornerstone of video generation, which is proved by recent researches[18]. In this section, we give a detailed description of the construction of our animation dataset and the evaluation benchmark.""
This paragraph feels like signposting. The 1st sentence states something obvious that can easily be mentioned somewhere in the text of the following paragraphs, the 2nd sentence carries no information value; I suggest removing the paragraph.

"divide raw animation videos into clips" How many clips are there? Based on he numbers later in the text I am guessing it is 100mil, but if you can please specify?

Minor formating thing: there is a big empty blank space at the end of page 7.

I am missing an explanation for why did you decide to use [6] "Scalable diffusion models with transformers" as your base model. There is huge amount of models out there, I would expect, at least some of them, to be discussed in the Related Work with a clear resaon for choosing one over the others -- this portion of the related works is completely missing.

The Method section mainly talks about [6] "Scalable diffusion models with transformers", and it is not really clear to me what is this paper's contribution there. I assume ti is the "3D Full Attention" which is only mentioned, but never really explained in details. I mean, the contribution of the paper is

 doi.org/10.32388/I69OPV

certainly the dataset creation/curation and evaluation, but

it is not really clear to me what is technical contribution beyond that the network was trained on a novel-dataset obtained thanks to a novel curation pipeline.

Training scheme has a lot of steps and they are split intp multiple sections which might be confusing. Consider organizing it better, maybe number the steps. I am assuming that this is the outline of the training scheme:

- starting with the cogvideox
- train on 480p at 8fps
- train on 480p at 16fps
- train on 720p at 16 fps
- "Additionally, we applied stricter filtering as in section3, producing a 1M ultra high-quality dataset for final-stage fine-tuning" -- here I am not sure if this is another "train on 720p at 16 fps" step or if this is talking about the steps discussed in the enxt sections
- train 720p at 16 fps ... on the subtitles clean dataset

Overall the paper is well written, easy to follow for the most parts, and results look convincing.

Declarations

Potential competing interests: No potential competing interests to declare.