Qeios

Peer Review

Review of: "AVID: Adapting Video Diffusion Models to World Models"

Mineui Hong[1]

1. Seoul National University, Korea, Republic of

Summary:

This paper proposes a method to implement an action-conditioned world model by leveraging pretrained video diffusion models. Since the parameters of most large-scale video diffusion models are not publicly accessible, fine-tuning these models for specific tasks is challenging. To address this, the authors propose a novel approach where an adapter is trained and applied during inference. This adapter guides the denoising process to condition video generation on robot actions without requiring direct fine-tuning of the pretrained model. Specifically, the adapter takes as input the generated video at each denoising step and the corresponding action, modifying the denoised video to align with the action. Experimental results demonstrate that the proposed method outperforms approaches that either require access to pretrained model parameters or do not require access, in action-conditioned video generation tasks.

Strength:

The proposed method is theoretically solid while appearing straightforward in concept. It has significant potential for broader applications in leveraging large-scale pretrained models across various domains without requiring parameter access.

Suggestions:

1. From my understanding, unlike other baselines that do not require access to pretrained model parameters, the proposed method uses the output of the pretrained model during training. This could result in significantly higher computational costs. It would be valuable for the authors to address this point explicitly in the paper and compare the training time computational costs with those of other baselines. Additionally, including results on how the performance of each model changes when training time is further restricted (e.g., 3 days of training → 1 day of training) would enhance the paper.

 doi.org/10.32388/DJB4A0

2. In Tables 1 and 2, the paper states, *"Shading indicates method requires access to the model parameters,"* but this shading is not visible on the article's webpage. It would improve clarity if the authors provided separate tables for comparisons between methods that require access to pretrained model parameters and those that do not. This separation would make the results easier to interpret and highlight the contributions of the proposed approach.

3. I kindly ask the authors to consider citing the paper titled "Diffused Task-Agnostic Milestone Planner," if they find it relevant to their work.

Declarations

Potential competing interests: No potential competing interests to declare.