Animating the Uncaptured: Humanoid Mesh Animation with Video Diffusion Models

Supplementary Material

A. Full Prompt

We provide the prompt used to generate our results, where we replace the *ACTION* placeholder with the action we want to generate. The prompt emphasizes dynamic motions, human body realism, static camera and bright lighting. These are to avoid static videos, body morphing, camera zooms and dark lighting respectively. The prompt is as follows:

"An award winning documentary about a person ACTION. Energetically ACTION. Dynamic movement. Light grey person. Realistic movement. Realistic motion. Realistic human body. Wide angle shot showcasing the man, dynamic movement, this video is incredibly detailed and high resolution, the uniform light is impressive, a masterpiece. Clear illumination. Bright light. No dark light. Fixed camera. No zoom in."

B. Perceptual Study

We conduct a perceptual study to evaluate our method against MDM [4]. A total of 30 users participated in the study and gave consent for their data to be used for research purposes. We include a screenshot of the user study interface 1.

B.1. List of Prompts

We list the prompts used in the perceptual study:

- *"The person is taking cover and shooting"*
- *"The person is practicing with nunchaku"*
- *"The person is playing tennis*
- *"The person is practicing karate moves"*
- *"The person is taking cover"*
- *"The person is adjusting her glasses"*
- *"The person is stopping the traffic"*
- *"The person is diging a hole with a shovel"*
- *"The person is dancing disco fox"*
- *"The person is opening a door"*
- *"The person is tying her shoes"*
- *"The person is brushing her hair"*
- *"The person is mining with a pickaxe"*
- *"The person is smoking"*
- *"The person is getting ready to fight"*
- *"The person is bouncing a ball"*
- *"The person is practicing kickboxing"*

C. Justification of the choice of VDMs

The results of our method are generated using two closed-source models: RunwayML [2] and Kling AI [3]. While we believe that academic research should publish code and models to ensure reproducibility, we chose to use these models due to their superior performance on our task. However, we want to emphasize that our method is agnostic to the choice of VDMs and can be used with any model that generates videos from text prompts, threfore, with the fast pace of research in this area, we expect that our method can be easily adapted to future models such as VideoJAM [1]. We show an example of the generated video obtained from CogVideoX [5] in figure Fig. 2.

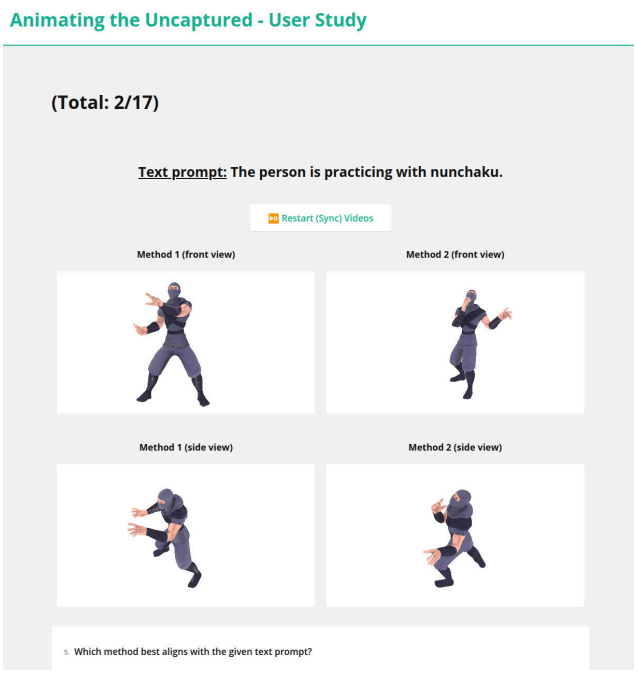

Figure 1. **Perceptual Study Visualization**. A screenshot of the user study interface. The user is presented with a text prompt and two methods to compare. Each method has two videos, one from the front and one from the side. The user can play the videos and compare them side by side. At the bottom of the page, the user has to answer a set of questions.

D. Ethical Considerations

We present a method for generating mesh animations from text prompts by leveraging video diffusion models. As a result, our approach inherits the ethical considerations associated with these models, particularly the potential for generating deepfakes of people performing actions. However, since our method requires a 3D mesh as input— which is costly and challenging to obtain for arbitrary individuals—

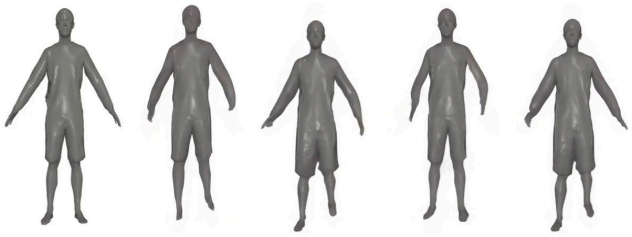

Figure 2. Example of a video generated with CogVideoX [5] conditioned on the rendering of a mesh and the prompt "The person is running".

we do not anticipate its use for creating deepfakes. Nevertheless, we acknowledge the importance of considering potential misuse and emphasize the need to raise awareness of the broader ethical implications.

E. Additional Results

We provide additional results of our method in figures Figs. 3 and 4, and additional comparisons against MDM in figure Fig. 5.

References

[1] Hila Chefer, Uriel Singer, Amit Zohar, Yuval Kirstain, Adam Polyak, Yaniv Taigman, Lior Wolf, and Shelly Sheynin. VideoJAM: Joint Appearance-Motion Representations for Enhanced Motion Generation in Video Models, 2025. arXiv:2502.02492 [cs]. 1

[2] Anastasis Germanidis. Introducing gen-3 alpha: A new frontier for video generation, 2024. 1

[3] Kling AI: AI-Powered Video Generation Platform. Kling ai, 2024. 1

[4] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human Motion Diffusion Model, 2022. 294 citations (Semantic Scholar/arXiv) [2024-03-17] 294 citations (Semantic Scholar/DOI) [2024-03-17] arXiv:2209.14916 [cs]. 1, 5

[5] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Xiaotao Gu, Yuxuan Zhang, Weihan Wang, Yean Cheng, Ting Liu, Bin Xu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer, 2024. arXiv:2408.06072 [cs]. 1, 2

Figure 3. **Additional qualitative results.** Visualization of generated mesh animations with our method. Each row shows: the prompt, the input mesh, and the generated mesh animations. For all generations, we visualize the mesh from the front and side views.

Figure 4. **Additional qualitative results.** Visualization of generated mesh animations with our method. Each row shows: the prompt, the input mesh, and the generated mesh animations. For all generations, we visualize the mesh from the front and side views.

Figure 5. **Additional comparisons with MDM [4].** We show two views (front and side) of the generated motions for multiple frames.