

# Human Motion Aware Text-to-Video Generation with Explicit Camera Control

## Supplementary Materials

Taehoon Kim<sup>\*1</sup>   ChanHee Kang<sup>\*2</sup>   JaeHyuk Park<sup>\*1</sup>  
Daun Jeong<sup>\*1</sup>   ChangHee Yang<sup>\*2</sup>   Suk-Ju Kang<sup>†2</sup>   Kyeongbo Kong<sup>‡3</sup>  
Pukyong National University<sup>1</sup>,   Sogang University<sup>2</sup>,   Pusan National University<sup>3</sup>

kimth52001@pukeyong.ac.kr, jasperai@sogang.ac.kr, hyeok0831@naver.com

ekdnsdl15@naver.com, yangchanghee2251@gmail.com, sjkang@sogang.ac.kr, kbkong@pusan.ac.kr

This is a supplementary material for the paper, Human Motion Aware Text-to-Video Generation with Explicit Camera Control.

### 1. Details of User Study

The user study was conducted in the two form of surveys, with a total of 100 participants. The majority of respondents (82.0%) were in their 20s and interested in the latest generative AI algorithms. This was followed by people in their 30s (9.0%) and teens (4.0%). The user study consisted of 57 questions, and 73 videos were used in the survey. The survey was conducted online in the format shown in Fig.1 and Fig. 2

The first survey compared videos using the same prompts on four T2V models, T2V-Zero [1], Zeroscope [2], and Runway-Gen2 [3], including our algorithm. Each model generated videos using 16 identical prompts, for a total of 64 videos used in the first survey. The videos were rated by users on three criteria: semantic relevance of prompt and video, realism of human motion, and overall quality, and users were asked to select the video they would rate the highest on each criterion among the four T2V output videos generated with the same prompt.

As a result of the survey, 56.8% of the evaluators judged that the video applied to our T2V model was better in terms of semantic relevance of prompt and video, 52.1% of the evaluators preferred our results in terms of realism of human motion, And 50.3% of respondents said that in terms of overall quality, the video that went through our network was better overall. Considering that the importance of all three criteria for evaluating videos is the same, 53.0% of all evaluators rated videos that passed through our network better.

In the second survey, users predicted whether the subject in the video was Zooming in, Rotating, Moving, etc., based on what they saw. We also included videos with default camera pose control in the questionnaire to distinguish and

predict whether the camera pose was controlled or not. The predictions were organized into nine categories: Side View, Top View, Rotation[X], Translation[Y], etc. On average, 49.8% of users per prompt correctly predicted the motion-to-skeleton projection module applied to the video. Percentage numbers are rounded to the second decimal place.

### 2. Additional Results: Motion Ambiguity, Scale, and Temporal Consistency Problems


In this section additionally shows the results for motion ambiguity, scale, and temporal consistency problems. Looking at the top Fig . 3, it can be seen that the scale problem and the temporary consistency problem occurred in combination. We can see that can't make good quality videos without pose guidance because the text prompt is written very hard. Looking at the middle Fig . 3, it can be seen that the results of the scale problem and the temporary consistency problem occurred in combination. Likewise, it can be seen that high-quality videos cannot be made without pose guidance, and the last Fig . 3 shows that motion amplitude and temporal consistency problems occurred in combination. Likewise, you can't make good quality videos without pose guidance.


### 3. Additional Results: Camera Movement


In this section, we shows the validity of whether camera movement shows a good representation of a person's motion. we will visualize our results by manipulating a camera matrix. We applied translation, Zoom In, Zoom Out, and Rotation to the subject as diverse and complex as possible, and the results are shown in Fig. 4, 5, 6, 7, 8 and 9.. Rotation showed the best effect among most camera techniques because it was possible to get a more prominent effect of person's motion by making the visible direction diverse. That doesn't mean that other camera techniques have a bad effect. For example, the top of Fig. 5 shows the effect of run-


Please select the video (A), (B), (C), or (D) that you feel is the best for the criteria \* given below.

A man is running on the beach

(A) 

(B) 

(C) 


(D) 

	A	B	C	D
semantic relevance of prompt and video	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
realism of human motion	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
overall quality	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Figure 1. Type 1, questioning the coherence of the meaning with the text prompt and the realism of the behavior in the given video.

(4) What effect was applied to the subject in that video? \*

A person kicks a ball on forest



- Adjust the view of the subject (top-down view)
- Adjust the view of the subject (side view)
- Zoom In on the subject
- Zoom Out on a subject
- Rotate the subject (clockwise or counterclockwise, when the subject is viewed as a clock hands)
- Rotate the subject (rotate the subject to show its side view)
- Move Subject Position (X-axis direction)
- Move Subject Position (Y-axis direction)
- No effect seen

Figure 2. Type 2, asking whether an effect has been applied to the given video, and if so, what effect.

ning properly, and the third of Fig. 7 shows that the effect of translation  $x$  properly shows the expression of jumping. In addition, if you look at the top of Fig. 8, you can also see the effect of gradually moving away from the runner like the movie camera technique. In conclusion, adjusting the camera is a very promising technology and a technology that needs to be developed. We have shown the possibility of this through various verifications.

#### 4. Prompt Set

Our framework needs text description  $\mathcal{P}$  to generate video. Complex and diverse motion descriptions are required. Therefore, we took a test prompt set of HumanML3D [4] and randomly added location at the end of the prompts. Locations are randomly selected from these category: sea, forest, moon, beach, desert, and auditorium. We used these location specified prompts to Follow Your Pose [5] and in experiment with T2V-Zero [1], we simply modified the prompt into {person} – {verb} – {location} from the prompt that we use in FYP [5]. This is because if a complex prompt pass through T2V-Zero [1] then it outputs unrecognizable videos.

### Motion Ambiguity, Scale, Temporal Consistency Problem

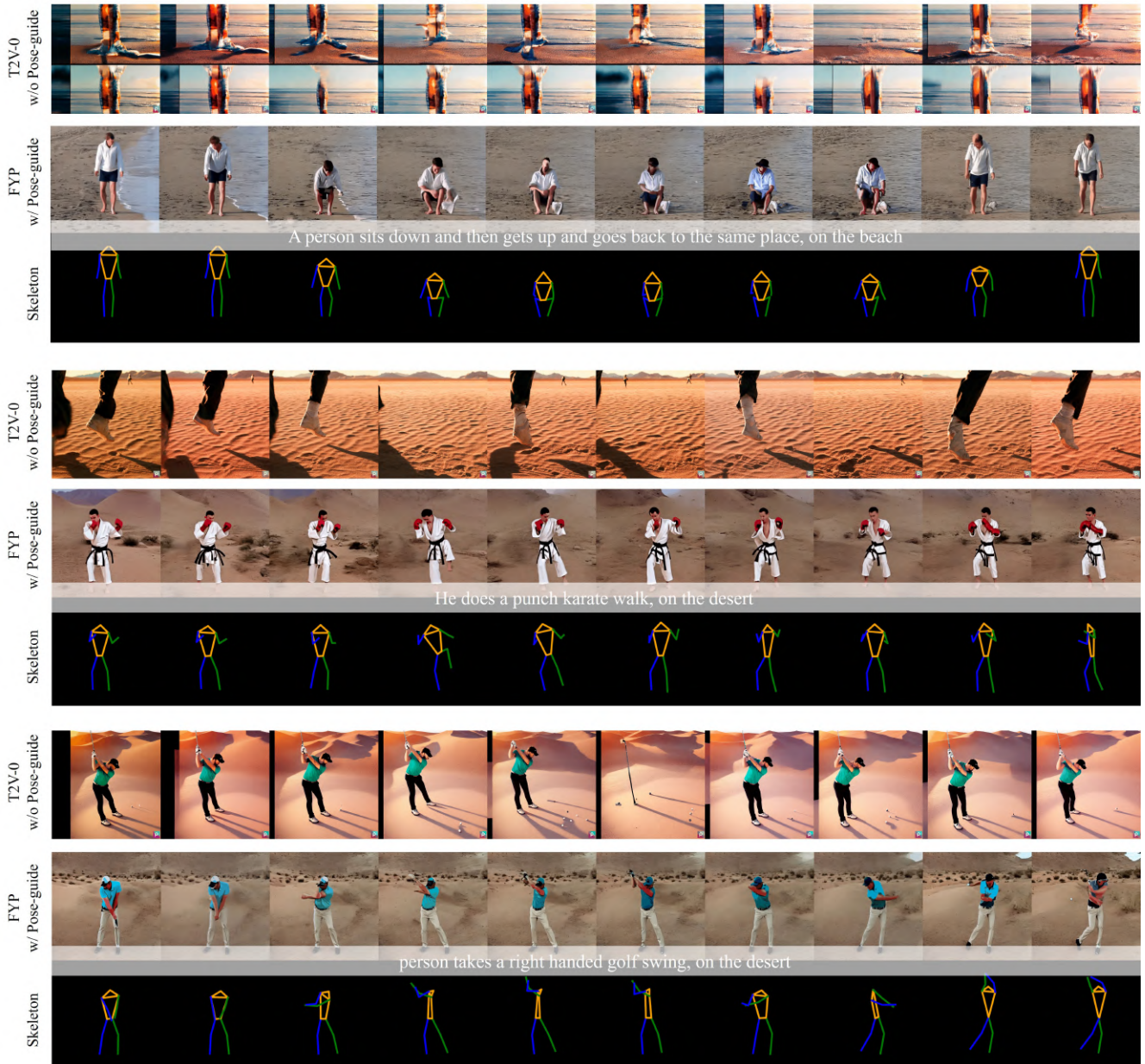


Figure 3. In this figure, we show results using FYP [5]. Motion Ambiguity, scale and inconsistency problems were in without pose guidance but not in with pose guidance. In figure on **Top** without pose guidance, The generated video is unnatural and even causes temporary consistency problems. However with pose guidance with fine scale generated. In figure on **Middle**, the output video generated only foot of human and frame inconsistency appeared. But with pose guidance consistency between consecutive frames with fine scale output generated. In figure on **Bottom**, the generated video without pose guidance looks fine but frame inconsistency and pose ambiguity occurred, but not in our method.

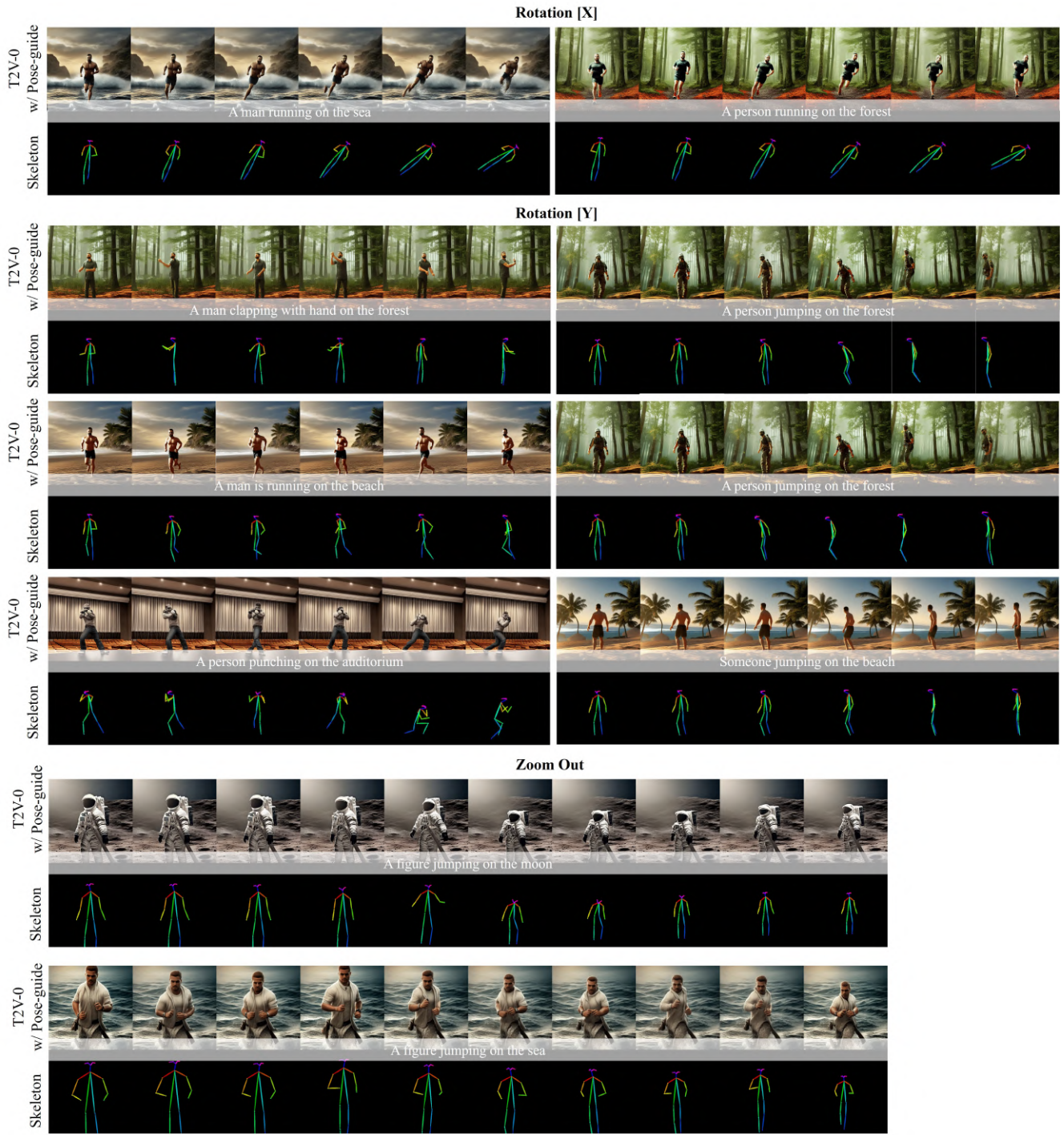


Figure 4. The above figure shows the results of camera control for Rotation [X], Rotation [Y], and Zoom Out. If Rotation [X] is applied, the runner can see the effect of running slightly to the side, and if Rotation [Y] is applied, the behavior of motion can be seen from various angles. Finally, if you Zoom Out, it shows that you can control the scale while gradually reducing the scale of the person.

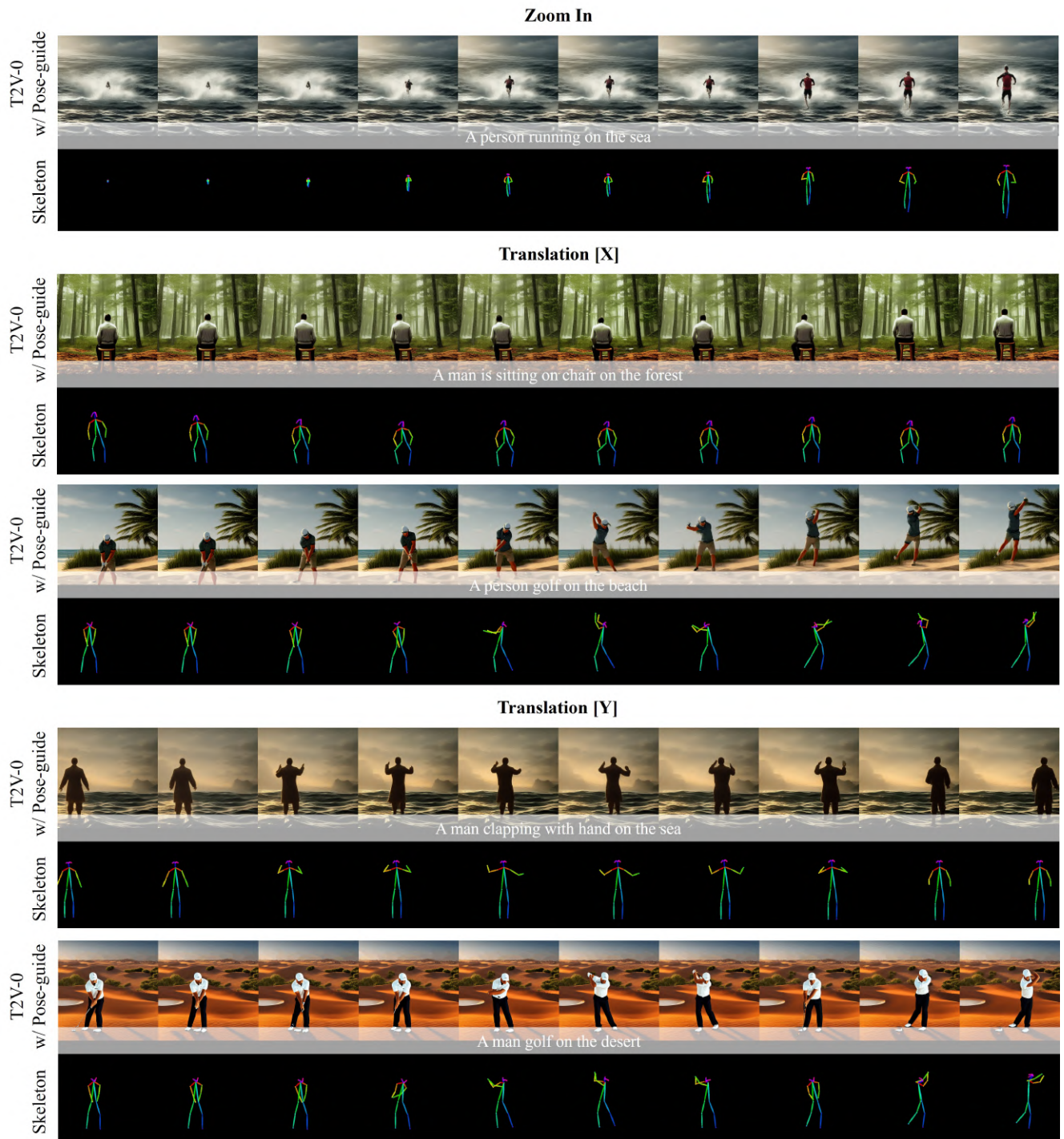


Figure 5. The above figure shows the results of camera control for Zoom In, Translation [X], and Translation [Y]. When Zoom In is applied, it is practically possible to create a video like a person running closer and closer. When Translation [X] is applied, the subject can be moved sideways, and when Translation [Y] is applied, the subject can be gradually raised.

### Zoom Out, Translation [X], Rotation [Y]



Figure 6. The above figure shows when Zoom Out, Translation [X], and Rotation [Y] are applied at once. The effect we expected was to move left or right, showing accurate motion through rotation as a person moves away. In most cases, there is no particular effect, but in the case of the second figure result, we obtained the desired result. This shows that camera control is very difficult but not impossible.



Figure 7. The top figure shows when Zoom In, Translation [Y], and Rotation [Y] are applied at once. The effect we expected was that as a person got closer and closer, he or she showed accurate motion through rotation and expected the effect of going up. It can be seen that a video that meets our expectations has been created, but due to the limitations of T2V, the background has not changed properly, creating a little awkward video. The Bottom figure shows when Zoom In, Translation [X], and Rotation [Y] are applied at once. The effect we expected was that as people got closer and closer, they showed accurate motion through rotation and expected the effect of going up sideways. In the case of run and jumping, it can be seen that we obtained the desired image.





Figure 8. The above figures show the results of camera control for Zoom out + Rotation [X], Translation [Y] + Rotation [Y]. It can be seen that most behavioral descriptions of run or jump are properly video-generated as intended.

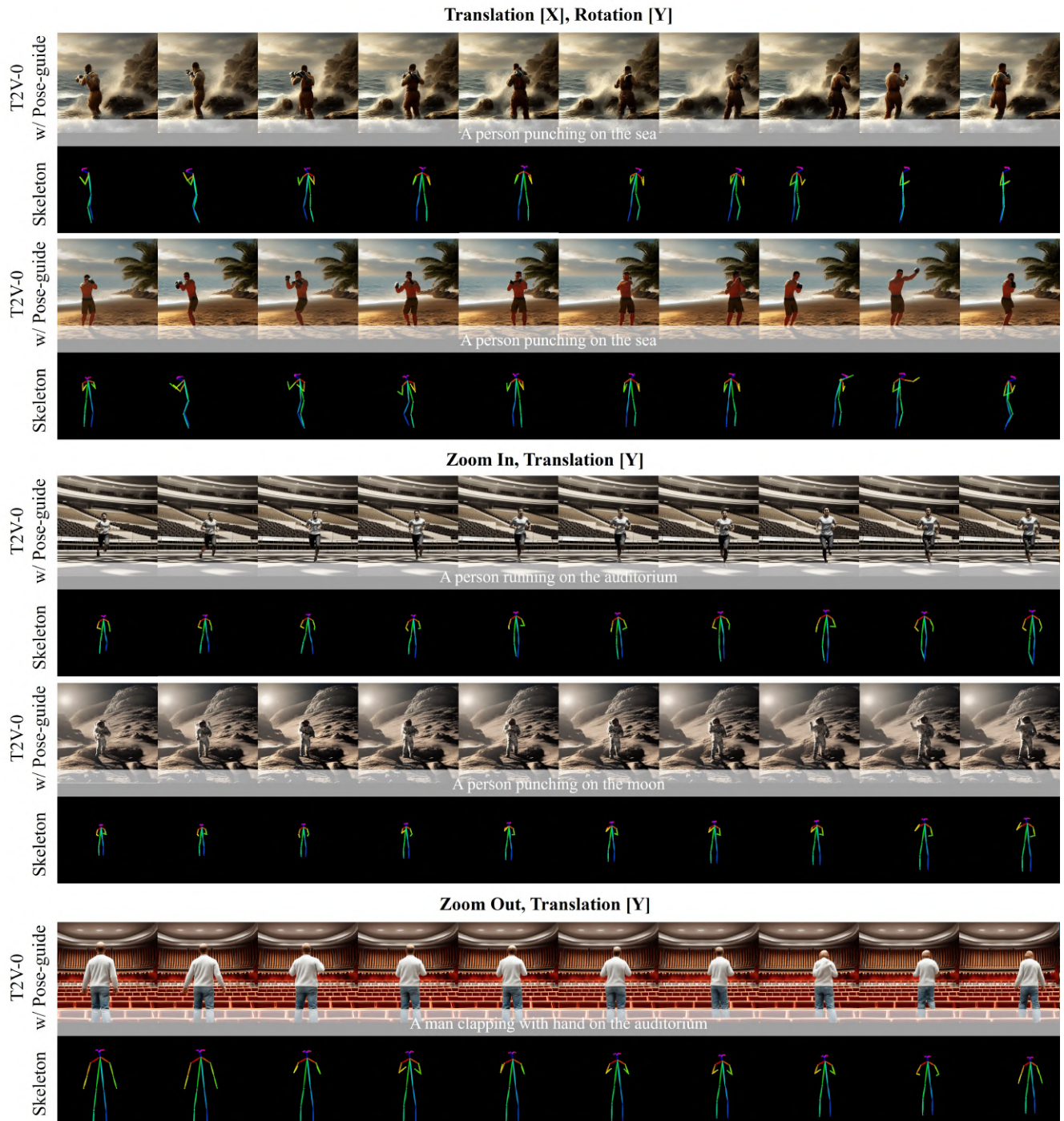


Figure 9. The above figures show the results of applying two camera controls at once, show that the rotation accurately shows the behavioral description of the punch, Zoom In can control the size of the subject, and Translation [Y] can create the effect of climbing the subject.

## References

- [1] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439*, 2023. 1, 3
- [2] cerspense/zeroscope\_v2\_576w. [https://huggingface.co/cerspense/zeroscope\\_v2\\_576w](https://huggingface.co/cerspense/zeroscope_v2_576w). Last Updated: 2023-07-01. 1
- [3] Gen-2 by Runway. <https://research.runwayml.com/gen2>. 1
- [4] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5142–5151, 2022. 3
- [5] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Ying Shan, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. *arXiv preprint arXiv:2304.01186*, 2023. 3, 4