

---

# Supplementary Material of PlayerOne

---

Anonymous Author(s)

Affiliation

Address

email

## 1 More Experimental Results

2 **Framework comparisons with prior works.** Fig. 2 illustrates the framework comparison between  
3 our PlayerOne and other competitors. Prior studies [4, 16, 2] have mainly concentrated on simulations  
4 within game-like environments, yet they often fail to accurately replicate real-world scenarios. Within  
5 these simulated environments, users are typically restricted to performing predefined actions, such  
6 as directional movements. This limitation confines user interactions to a constructed world, thereby  
7 restricting the execution of freeform movements akin to those in real-world settings. Existing realistic  
8 world simulators [8, 10, 1] often focus solely on world-consistent generation, lacking mechanisms  
9 for human movement control. As a result, users are relegated to passive observers, rather than active  
10 participants, within the environment. This significantly impacts the user experience by hindering the  
11 formation of a genuine connection with the simulated world. In contrast to these limitations, our  
12 approach enables freeform motion control for users, enhancing their interactive experience.

13 **More visualization results.** Here we provide more simulated results in Fig. 3 and Fig. 4. In terms  
14 of first-person action alignment and world consistency, we have achieved outstanding results in both  
15 game and real-world scenarios. Additionally, we selected highly dynamic settings, such as driving  
16 scenes, where our method successfully models the world with high accuracy while maintaining  
17 excellent video fluidity. More visualization results can be referred in the submitted video.

18 **Visualization of the scene reconstruction.** This section presents a comprehensive visualization  
19 of the reconstructed scenes using our PlayerOne. As illustrated in Fig. 1, our approach adeptly  
20 reconstructs both scenes and video frames through a progressive methodology. This ensures not only  
21 inter-frame coherence but also overarching scene consistency across a diverse array of scenarios.  
22 Specifically, the method seamlessly integrates temporal and spatial elements to maintain visual  
23 congruity, even in complex environments. The robustness of our technique is further reflected  
24 in its capacity to adapt to varying scene dynamics and compositions, thereby offering a reliable  
25 framework for generating high-quality, consistent video outputs. Through these visualizations,  
26 the effectiveness of our PlayerOne in achieving smooth and coherent reconstructions is clearly  
27 demonstrated, highlighting its potential applications in advanced graphical simulations and interactive  
28 environments.

29 **Investigation on the impact of rendering.** In addition to analyzing DuST3R [14] within the  
30 manuscript, we extend our comparison by substituting the point map rendering technique with several  
31 alternative methods, including MonST3R [17] and MAST3R [6]. As detailed in Table 1, our approach  
32 exhibits remarkable generalization capabilities across the spectrum of rendering strategies. This robust  
33 performance underscores the versatility and adaptability of our method, making it highly effective  
34 in accommodating diverse rendering paradigms. By integrating various rendering methodologies,  
35 we showcase the method’s extensive applicability and resilience in maintaining high-quality outputs,  
36 regardless of the specific techniques employed. The comparative analysis further reflects our method’s  
37 potential for broad applicability in dynamic, real-world settings, demonstrating consistent, optimal  
38 performance in diverse operational contexts.

Table 1: **Investigation on the impact of different rendering methods.** Our PlayerOne shows great robustness against different rendering methods.

	DINO-Score (↑)	CLIP-Score (↑)	MPJPE (↓)	MRRPE (↓)	PSNR(↑)	FVD (↓)	LPIPS(↓)
CUT3R [13]	67.8	88.2	127.16	163.62	50.3	236.12	0.0663
MonST3R [17]	67.1	88.4	127.68	164.90	49.8	235.09	0.0771
MASt3R [6]	67.4	87.8	127.35	163.06	50.1	240.10	0.0724

Table 2: **Investigation on the impact of different filtering ratios.**

Ratio	DINO-Score (↑)	CLIP-Score (↑)	MPJPE (↓)	MRRPE (↓)	PSNR(↑)	FVD (↓)	LPIPS(↓)
0	64.2	85.0	123.50	158.10	47.5	228.50	0.0587
5	65.4	86.5	125.00	160.20	48.7	230.75	0.0600
10	<b>67.8</b>	<b>88.2</b>	<b>127.16</b>	<b>163.62</b>	<b>50.3</b>	<b>236.12</b>	<b>0.0663</b>
15	66.0	87.2	126.00	162.00	49.8	234.00	0.0640

**Investigation on the impact of filtering.** In this study, we examine how the filtering ratio affects model performance. As demonstrated in Table 2, increasing the filtering ratio leads to a decline in model performance, which can be attributed to insufficient data. Conversely, the absence of filtering also causes performance deterioration, primarily because noisy data is introduced during training, negatively impacting the accuracy of action alignment. Therefore, we have determined that a filtering ratio of 10% optimizes performance.

## 2 Broader Impact

The proposed PlayerOne for video generation, designed to facilitate freeform human motion control within environments created from user-provided images while producing world-consistent videos, demonstrates considerable potential across diverse domains. It is particularly adept at generating engaging and dynamic educational content, thereby fostering experiential learning through interactive simulations. Moreover, the model optimizes the production of high-quality, consistent visual content for films, television, and online media, dramatically reducing both production time and costs. It also enables the creation of interactive narratives, allowing users to influence the storyline through their interactions within the generated environments, thus enhancing user engagement and narrative immersion. Beyond these applications, the world model serves as a valuable tool for research on human behavior and interactions within controlled virtual settings, offering insights for fields such as psychology, sociology, and human-computer interaction. By integrating these capabilities, the proposed world model not only amplifies existing applications but also paves the way for novel research and development, significantly contributing to technological progress and societal advancement.

## 3 Limitations & Discussion

While significant strides have been made in egocentric interaction and coherent world modeling, certain limitations persist. Despite the compelling outcomes, the performance in game scenarios is somewhat diminished compared to realistic scenarios, likely due to the disproportionate amount of realistic training data available. Moreover, in highly dynamic scenes, predictions may falter, reflecting the constraints inherent in the current base model. Future research endeavors could potentially overcome these challenges by investigating novel action representations, incorporating an expanded dataset for game scenarios, and adopting a more robust base model.

## 4 License of assets

**Datasets** (Apache 2.0 License) Nymeria [9]/FT-HID [5]/EgoExo-Fitness [7] (Creative Commons Attribution 4.0 International), EgoExo4D [3]/Egovid-5M [15](MIT License).

**Codes** The official repository of Aether [11] (MIT License), the official repository of Cosmos [1] (Apache 2.0 License), the official repository of Wan2.1 [12] (Apache 2.0 License).



Figure 1: **The reconstructed scene of our PlayerOne.** Our PlayerOne can achieve relatively precise scene reconstruction by jointly modeling of video frames and scenes.

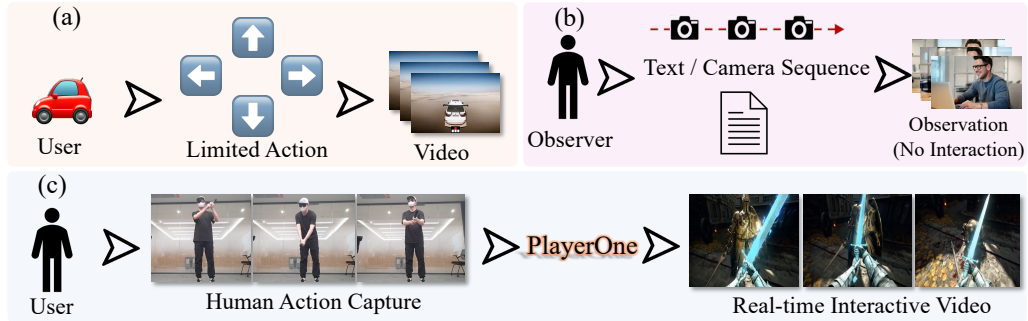


Figure 2: **Difference between our PlayerOne and prior works.** Our PlayerOne can enable freeform movements in the simulated world and achieve great world consistency across diverse scenarios.

## References

- [1] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv:2501.03575*, 2025.
- [2] Ruili Feng, Han Zhang, Zhantao Yang, Jie Xiao, Zhilei Shu, Zhiheng Liu, Andy Zheng, Yukun Huang, Yu Liu, and Hongyang Zhang. The matrix: Infinite-horizon world generation with

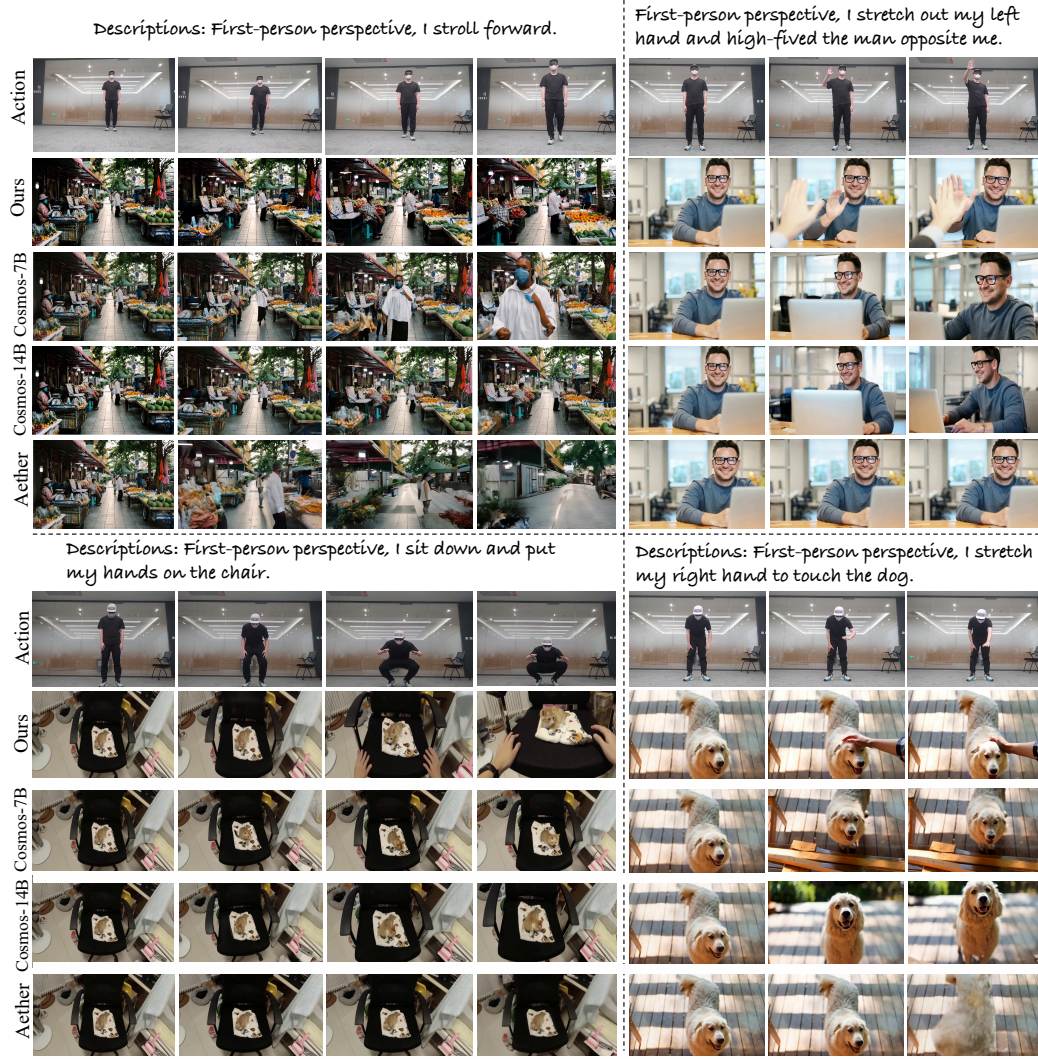


Figure 3: **More visualization results generated by our PlayerOne.** Our method demonstrates great superiority in both motion alignment and environmental interaction across different domains.

- 79 real-time moving control. *arXiv:2412.03568*, 2024.
- 80 [3] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyl-  
81 los Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d:  
82 Understanding skilled human activity from first-and third-person perspectives. In *CVPR*, 2024.
- 83 [4] Junliang Guo, Yang Ye, Tianyu He, Haoyu Wu, Yushu Jiang, Tim Pearce, and Jiang  
84 Bian. Mineworld: a real-time and open-source interactive world model on minecraft.  
85 *arXiv:2504.08388*, 2025.
- 86 [5] Zihui Guo, Yonghong Hou Hou, Pichao Wang, Zhimin Gao, Mingliang Xu, and Wanqing Li.  
87 Ft-hid: A large scale rgb-d dataset for first and third person human interaction analysis. *Neural*  
88 *Computing and Applications*, 2022.
- 89 [6] Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3d with  
90 mast3r, 2024.
- 91 [7] Yuan-Ming Li, Wei-Jin Huang, An-Lan Wang, Ling-An Zeng, Jing-Ke Meng, and Wei-Shi  
92 Zheng. Egoexo-fitness: towards egocentric and exocentric full-body action understanding. In  
93 *ECCV*, 2024.





Figure 4: **More visualization results generated by our PlayerOne.** Our method demonstrates great superiority in both motion alignment and environmental interaction across different domains.



Figure 5: **More visualization results generated by our PlayerOne.** Our method demonstrates great superiority in both motion alignment and environmental interaction across different domains.

- 94 [8] Hanwen Liang, Junli Cao, Vidit Goel, Guocheng Qian, Sergei Korolev, Demetri Terzopoulos,  
 95 Konstantinos N. Plataniotis, Sergey Tulyakov, and Jian Ren. Wonderland: Navigating 3d scenes  
 96 from a single image. *arXiv:2412.12091*, 2024.
- 97 [9] Lingni Ma, Yuting Ye, Fangzhou Hong, Vladimir Guzov, Yifeng Jiang, Rowan Postyeni, Luis  
 98 Pesqueira, Alexander Gamino, Vijay Baiyya, Hyo Jin Kim, et al. Nymeria: A massive collection  
 99 of multimodal egocentric daily motion in the wild. In *ECCV*, 2024.



Figure 6: **More visualization results generated by our PlayerOne.** Our method demonstrates great superiority in both motion alignment and environmental interaction across different domains.

- [10] Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. *arXiv:2503.03751*, 2025.
- [11] Aether Team, Haoyi Zhu, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Chunhua Shen, Jiangmiao Pang, and Tong He. Aether: Geometric-aware unified world modeling. *arXiv:2503.18945*, 2025.
- [12] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv:2503.20314*, 2025.
- [13] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. *arXiv:2501.12387*, 2025.
- [14] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024.
- [15] Xiaofeng Wang, Kang Zhao, Feng Liu, Jiayu Wang, Guosheng Zhao, Xiaoyi Bao, Zheng Zhu, Yingya Zhang, and Xingang Wang. Egovid-5m: A large-scale video-action dataset for egocentric video generation. *arXiv:2411.08380*, 2024.
- [16] Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. Worldmem: Long-term consistent world simulation with memory. *arXiv:2504.12369*, 2025.
- [17] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arxiv:2410.03825*, 2024.