

Supplementary Material

Endless World: Real-Time 3D-Aware Long Video Generation

In this supplementary material, we first present our demo video in Sec. 1, followed by the user preference study in Sec. 2. We then discuss the limitations of our approach and potential future directions in Sec. 3. Next, we report detailed semantic and video quality metrics across different video durations using VBench in Sec. 4. Finally, we provide additional visualization examples in Sec. 5.

1. Demo video

We encourage readers to view the demo video included in the supplementary material. This video offers a more comprehensive and intuitive comparison of our method against existing approaches, showcasing qualitative results that cannot be fully captured through static figures or numerical tables alone. The demo video shows the comprehensive visualization of our approach in terms of visual quality, temporal consistency, and adherence to physical dynamics.

2. User preference study

To further assess the perceived quality of the generated videos, we conduct a user preference study focusing on 30 second video generation using the VBench dataset. From the full set of generated samples, we randomly select 50 videos and present them to human evaluators. Each rater is asked to score the visual quality of the generated videos on a 1–5 scale for video clips from 0-10s, 10-20s and 20-30s, considering factors such as clarity, temporal consistency, realism, and overall coherence. The evaluators compare our method directly with the self-forcing baseline, providing paired assessments for each video segment. The results indicate an evaluation of video quality across all time intervals, which reflecting our method’s stronger ability to maintain temporal stability and visual fidelity over different horizons. These findings align with our quantitative metrics and further demonstrate the effectiveness of our model in long-duration video generation.

3. Limitations

Despite the advantages of our approach, several limitations remain. Our method is designed to generate dynamic and

visually coherent long-duration videos—even approaching infinite video generation—by dynamically synthesizing scene elements based on the fused prompt. While this allows the model to maintain the appearance of key foreground subjects such as humans or animals, preserving the background consistently is considerably more challenging.

The primary difficulty arises from the fact that the background changes as the virtual camera moves. Because 3D reconstruction cannot fully recover or represent the complete surrounding environment, the model may struggle to maintain spatial consistency across extended time periods. As a result, background details may drift or evolve in unintended ways during very long sequences.

In contrast, the self-forcing baseline tends to keep the foreground subjects relatively static, which inadvertently stabilizes the background for the first several seconds. However, this lack of movement leads to a noticeable decline in overall visual quality, causing the generated videos to appear less realistic and more degraded over time.

Our method produces significantly higher-quality long videos, but background consistency remains a key challenge. Addressing this limitation is an important direction for future work. We aim to improve spatial coherence and develop a more robust solution for generating long or infinite videos with stable and realistic backgrounds.

4. Detailed dimension metrics

We provide the full dimension-wise evaluation results in Table 1 and Table 2. As shown in these tables, our approach demonstrates stable and robust performance on both 5-second and 30-second video clips. Across the quality-related dimensions—such as subject consistency, background consistency, temporal flickering, motion smoothness, aesthetic quality, imaging quality, and degree of dynamics—our method consistently outperforms or matches existing baselines. These metrics collectively capture how well the generated videos maintain coherent appearance, avoid artifacts, and preserve natural motion over time. In addition to quality metrics, we also evaluate semantic fidelity across multiple categories, including object class accuracy, multi-object handling, human action correctness, color consistency, spatial relationships, scene composition,

Method	Time	Consistency \uparrow		Smoothness \uparrow		Quality \uparrow		dynamic degree \uparrow
		Subject	Background	flickering	smoothness	aesthetic	imaging	
Endless World	5s	93.89	94.79	97.86	95.05	66.33	69.69	36.11
	30s	97.22	96.29	97.18	95.19	65.28	69.34	35.42

Table 1. Quality related video generation metrics across durations of 5s and 30s.

Method	Time	Object \uparrow			Relationship \uparrow			Style \uparrow		overall consistency \uparrow
		class	multiple	human action	color	spatial	scene	appearance	temporal	
Endless World	5s	95.68	89.84	94.0	84.72	79.03	69.27	71.24	67.2	74.41
	30s	95.89	90.55	94.0	84.66	79.84	67.62	70.99	67.11	74.42

Table 2. Semantic-related video generation metrics across durations of 5s and 30s.

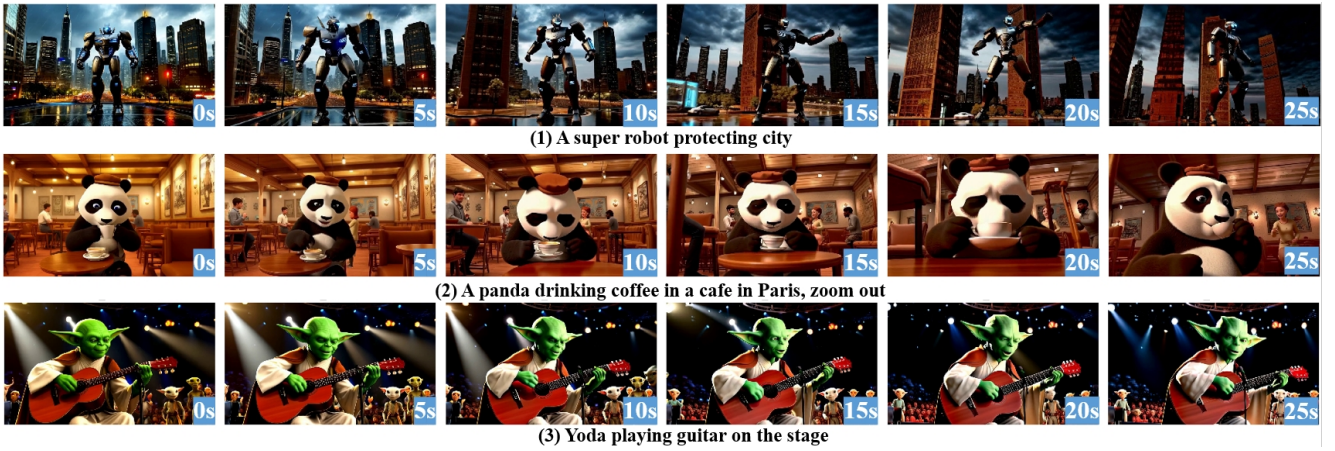


Figure 1. Additional visual results showcasing our model’s 30-second video generations.

Method	Video quality \uparrow			
	0–10s	10–20s	20–30s	Total
Self-forcing [1]	4.68	3.98	3.45	4.04
Endless World (Ours)	4.96	4.36	4.26	4.52

Table 3. Video quality comparison across different time intervals.

appearance style, temporal style, and overall semantic coherence. The results indicate that our model is able to maintain high semantic alignment with the input prompts, even for long-duration videos where semantic drift commonly occurs. Together, these findings highlight the effectiveness of our framework in generating long, coherent, and semantically faithful video sequences.

5. More visualization examples

Figure 1 presents additional examples of our 30-second video generations. We illustrate three representative cases:

(1) a giant robot protecting a city, (2) a panda drinking coffee in a Parisian café with a gradual zoom-out, and (3) Yoda playing guitar on a stage. These examples demonstrate that our model is capable of producing visually appealing, coherent, and engaging long-duration videos. The generated sequences maintain strong visual quality and compelling scene dynamics throughout the full 30-second duration, highlighting the effectiveness of our approach in long-horizon video generation.

References

- [1] Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *CoRR*, abs/2506.08009, 2025.