
WorldWeaver: Generating Long-Horizon Video Worlds via Rich Perception – *Supplementary Material* –

Anonymous Author(s)

Affiliation

Address

email

1 The content of this supplementary PDF is organized as follows:

- 2 • Provided more visualizations and comparisons.
- 3 • Systematically conducted a user study to compare our method with open-source models.
- 4 • Additional ablation and analysis experiments.
- 5 • Discussion on limitations and broader impact.

6 Contents

7	A Additional Visual Results	1
8	A.1 More visualizations	1
9	B User Study	1
10	C More Discussions	3
11	C.1 Drift resistance experiment details	3
12	C.2 Ablation on number of groups and length of memory bank	4
13	D Broader Impacts and Limitations	5
14	D.1 Limitations and future works	5
15	D.2 Broader impacts	5

16 A Additional Visual Results

17 A.1 More visualizations

18 As shown in Fig. [S1](#), we present additional videos generated by WorldWeaver. Additionally, we
19 recommend accessing the *viewer.html* file for a visual comparison between our approach and current
20 state-of-the-art methods

21 B User Study

22 To conduct a comprehensive comparison, we evaluate our model (based on Wan2.1 1.3B [\[3\]](#)) against
23 recently released state-of-the-art long-video models of similar scale, including SkyReels-V2 1.3B [\[1\]](#)



Figure S1: More visualizations.

24 and MAGI 4.5B [2]. Specifically, we use 48 prompts, each containing 4–6 actions, to generate videos
 25 lasting 20–30 seconds. All prompts are provided in the `prompts.txt` file. We engage 15 annotators
 26 to complete a questionnaire, with the UI screenshot presented in Fig. S2. The questionnaire comprises
 27 five questions: (1) Which video has the highest overall image quality? (2) Which video exhibits
 28 better consistency (considering both subject and background)? (3) Which video shows the smallest
 29 difference between the first 5 seconds and the last 5 seconds (in terms of quality and consistency)?

Username (enter to start)

worldweaver

Start

Please answer the five questions

Prompt: A boy skips stones by the riverbank, he cheers when one skips three times, he searches for a flatter stone, then tries again with determination.

01: 1

02: 2

03: 3

Quality Answer

Consistency Answer

Difference Answer

Motion Answer

Alignment Answer

Which video has the highest overall image quality?

Vote 1

Vote 2

Vote 3

Tie

Which video has better consistency, primarily considering ID consistency (e.g., maintaining consistent subject and background appearance throughout)?

Vote 1

Vote 2

Vote 3

Tie

Which video has the smallest difference between the first 5 seconds and the last 5 seconds, comprehensively considering color, video style, ID consistency, and video quality? Note: Focus only on the difference, not the overall quality.

Vote 1

Vote 2

Vote 3

Tie

Which video has smoother motion with minimal action distortion and sudden twitches?

Vote 1

Vote 2

Vote 3

Tie

Which video aligns best with the overall action described in the prompt?

Vote 1

Vote 2

Vote 3

Tie

Figure S2: Visualization of the user study interface.

30 (4) Which video has smoother motion? (5) Which video aligns best with the overall action described
 31 in the prompt.

32 Results are shown in Fig. S3. Our method demonstrates performance comparable to MAGI in motion
 33 smoothness and consistency metrics, while achieving superior results in quality drift. However, a
 34 slight gap in image quality remains, with overall performance generally surpassing SkyReels-V2.
 35 These findings are generally consistent with our measurements on the VBench benchmark.

36 C More Discussions

37 C.1 Drift resistance experiment details

38 To investigate the drift resistance of perceptual conditions and color information, we conduct the
 39 following experiment after model training. For the depth-only and RGB-only variants, we retain
 40 only the weights corresponding to the relevant input and output channels, while loading all other
 41 parameters from the fully trained model. We then select a single image or depth map and replicate
 42 it across 81 frames (to match the model’s input sequence length), appending "static" to the training
 43 caption. This process generates 10,000 such samples, which we fine-tune for 1,000 steps each,
 44 enabling the model to perform a simple task: consistently outputting static images in a streaming
 45 manner. Notably, since optical flow represents motion between frames and lacks temporal continuity

3

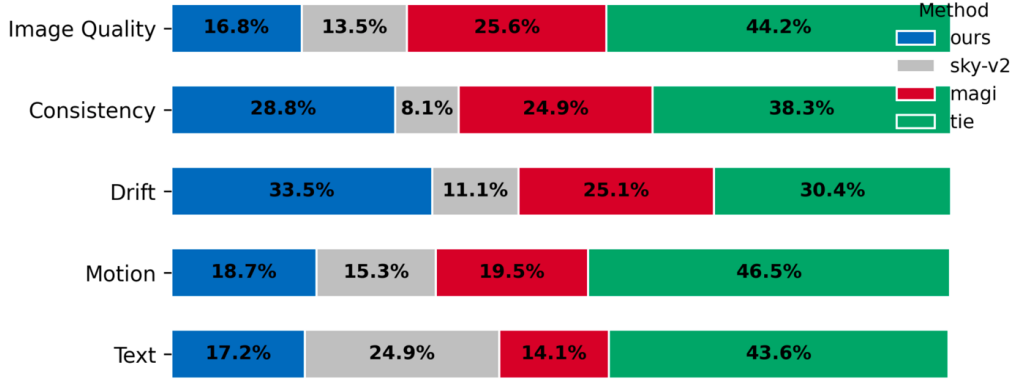


Figure S3: Results of user study.

as a signal, its isolated output lacks meaningful consistency, so we exclude it from this experiment. We assess the drift resistance by measuring the normalized Mean Squared Error (MSE) between the subsequently generated frames and the first generated frame. As shown in Fig. S4, depth output alone demonstrates superior drift resistance. Jointly outputting both RGB and depth mitigates the drift phenomenon in RGB.

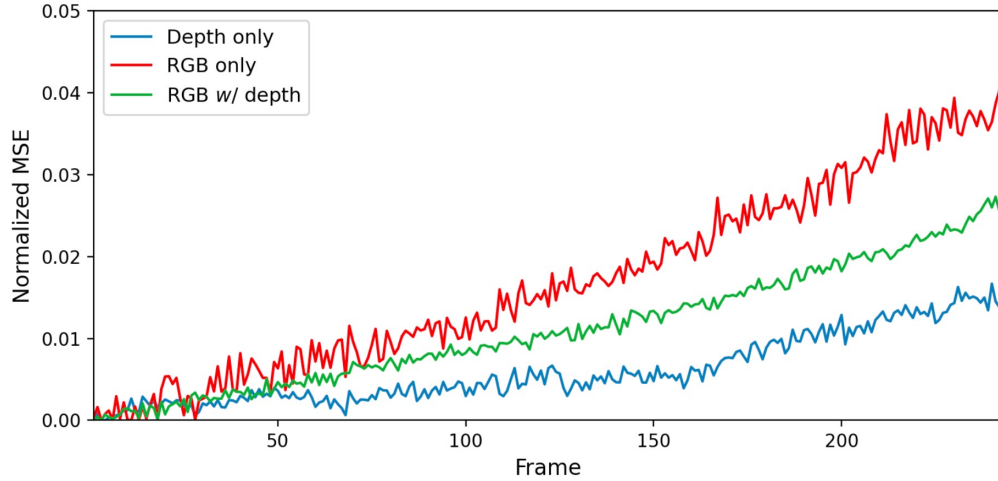


Figure S4: Normalized mse for drift resistance.

C.2 Ablation on number of groups and length of memory bank

Table S1: **Ablation study.** Notations such as m1_g4_n5 indicate a memory bank length $M = 1$, prediction groups $G = 4$, and frames per group $N = 5$, and so forth, while Frameoutput refers to the number of video frames produced by a single complete denoising pass.

Setting	Subject consistency	Background consistency	Image quality	Motion smoothness	$\Delta_{\text{drift}}^{\text{Quality}}$	frame output
m1_g4_n5	88.85	91.30	0.61	0.75	0.05	20
m5_g4_n4	90.92	92.39	0.60	0.75	0.07	16
m13_g4_n2	91.04	92.51	0.58	0.74	0.09	8
m5_g8_n2	90.64	91.87	0.62	0.73	0.06	16
m1_g2_n10	90.55	91.23	0.56	0.68	0.12	20

To further analyze the impact of memory bank length and the number of prediction groups, we conduct an ablation study on these hyperparameters. Let F denote the total number of input latent frames, G the number of prediction groups, N the number of frames per group, and M the memory bank length. These parameters are related by $N \cdot G + M = F$. We fix $F = 21$ for all settings, in order to fix the computational cost. Within the memory bank, one frame is reserved for short-term memory (fully denoised frame), while the remaining $M - 1$ frames are dedicated to long-term memory. As shown in Tab. S1, considering the overall performance across all metrics as well as the number of frames generated under the same computation, we set $M = 5$ and $N = G = 4$ in our final training configuration. This experiment is conducted on the robotic manipulation dataset using 16 A100 GPUs, consistent with the ablation studies in the main paper, including the discussion on methods for long videos and the analysis of noise levels in the memory bank. For these analyses, we extract the first frame from videos where DROID operations fail, using these frames as input images. This subset of data is entirely unseen during training. Based on these first frames, we employ ChatGPT o3 to generate 30 prompts for 15–20 second robotic operations, each containing 3–5 action commands (e.g., move, grasp, switch, push/pull). For the ablation study on the contribution of perceptual conditions, we use 60 short 5-second videos to isolate the effects from long video frameworks. All prompts are provided in `prompt_robotic.txt`.

D Broader Impacts and Limitations

D.1 Limitations and future works

Despite its strengths, our work has several limitations. First, learning stable physical dynamics from the complexity of the real world remains a long way off, and our model is far from perfect. Operations involving very small objects can still exhibit sudden disappearances, since even depth-based cues cannot fully capture these small objects. Second, although fine-tuning existing models allows us to generate 20~30s videos, error accumulation persists over longer horizons, which we leave as an important direction for future work. Finally, while video depth is shown to be the most effective perceptual condition in our experiments, investigating additional or complementary cues to further exploit the rich information in real-world data also represents a promising avenue for further study.

D.2 Broader impacts

As video generation technology continues to advance, the development of robust authentication and forgery detection methods becomes increasingly critical. The ability to create highly realistic videos, such as those generated by our model, underscores the need for parallel advancements in counterfeit identification to mitigate potential misuse, ensuring the integrity of digital content in an era of rapid technological evolution.

85 **References**

- 86 [1] Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, Mingyuan Fan, Hao Zhang, Sheng
87 Chen, Zheng Chen, Chengcheng Ma, Weiming Xiong, Wei Wang, Nuo Pang, Kang Kang, Zhiheng Xu,
88 Yuzhe Jin, Yupeng Liang, Yubing Song, Peng Zhao, Boyuan Xu, Di Qiu, Debang Li, Zhengcong Fei, Yang
89 Li, and Yahui Zhou. Skyreels-v2: Infinite-length film generative model, 2025.
- 90 [2] Sand-AI. Magi-1: Autoregressive video generation at scale, 2025.
- 91 [3] Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao
92 Yang, Jianyuan Zeng, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint*
93 *arXiv:2503.20314*, 2025.