A. More Experimental Details

A.1. Implementation

Video Retrieval Algorithm. To construct the view frustum for a single camera pose, we fix the horizontal and vertical fields of view to 90° and 60°, respectively, and set the near and far clipping planes to 0 and 10. For mesh sampling, we uniformly sample 8 points along the width and 6 points along the height on the plane.

Choice of Context Number k. A larger value for k allows for the retrieval of more contextual information and reduces the number of required inference iterations, but it also increases computational overhead. but comes at the cost of increased computational overhead. To balance context visibility and computation, we adopt video consistency as the selection criterion, as shown in Tab. R1, and choose $k = 4$ as the appropriate setting.

Table R1. **Ablation on In-context Video Number** k. View Synchronization is evaluated on the Basic Benchmark.

N Shots/k videos	2	3	4	5	6
6	38.9	39.7	<u>40.3</u>	**40.8**	40.2
9	43.6	44.5	**45.6**	<u>45.4</u>	44.7
12	40.8	40.2	**41.2**	<u>40.9</u>	40.4

Long Video Generation. During training, we adopt 6 overlapped latent frames (corresponding to 21 decoded frames) and set the conditioning ratio to 0.45. Although the model is trained on 81-frame multi-view datasets, it generalizes effectively to long-form video generation. During inference, we produce a 93-frame initial chunk, followed by subsequent chunks of 71 frames.

Self-Conditioned Training. For the second training stage of the model evaluated on the Basic benchmark, we randomly sample 900 scenes from the MultiCamVideo dataset and 100 scenes from the SynCamVideo dataset. For each scene, we synthesize 1–5 videos, yielding roughly 3.5K training samples in total. In our current setup, the clean ground-truth video is used as the context for generating its noisy pseudo–GT counterpart. We did not observe clear performance gains when incorporating long-shot autoregressive generation into the synthetic video pipeline, as reported in Tab. R2. For each N-shot setting, we generate the same number of synthetic videos and post-train the model for 2K steps to ensure a fair comparison.

Table R2. **Ablation on N-Shots Synthetic Video Generation**. Evaluation is conducted on the Basic Benchmark.

	1 Shot	2 Shots	3 Shots	4 Shots
FVD↓	**425.8**	441.3	<u>436.4</u>	460.2
IQ↑	**58.5**	<u>57.6</u>	57.2	56.5

Long-range Error Accumulation. We study the error accumulation in long video generation with regarding to our strategy of memory bank and progressive generation. To evaluate this, we curate 20 test videos and measure the error using the drifting metric Δ^M_{drift} from FramePack [5]. As shown in Tab. R3, this issue is more pronounced with more views (>12), longer videos (>600 frames), and drastic camera motion. Potential strategies to mitigate this error include: (1) adopting the Diffusion Forcing or Self-Forcing training paradigm, and (2) transitioning from our current train-short-test-long to a train-long-test-long paradigm as used in LongLive [4].

Table R3. **Ablation Study of Memory Bank and Video Length on Long-Term Error Accumulation**.

	Camera Views				Video Length (frames)			
	9	12	15	18	400	600	800	1000
$\Delta^M_{drift}\downarrow$	**4.02**	<u>4.56</u>	5.13	7.23	**4.25**	<u>5.23</u>	7.83	9.72

Failure Case Analysis on OOD Cameras. We evaluate the failure case under challenging out-of-distribution camera trajectories. Unseen motions, such as elliptical, zigzag, or figure-eight paths, can induce significant visual distortions. To evaluate this, we curate 12 challenging camera trajectories and apply them to videos from the basic benchmark, resulting in FVD: 425.8→583.2, TransErr: 0.54→0.82, and RotErr: 0.21→0.34; even within the test camera distribution of the basic benchmark, scaling camera translation and rotation by 2× increases FVD from 425.8 to 503.4.

A.2. Evaluation

FVD (Fréchet Video Distance). We use StyleGAN-V [2] as the feature extractor backbone, sample 49 frames per video interval, resize each frame to 432×768, and include all frames for evaluation. For videos with substantial camera motion (low FOV overlap) relative to the input video, we select alternative video with similar camera trajectories for evaluation. Thus this metric can provide an indirect measure of video similarity.

TransErr (Camera Translation Error).

$$\text{TransErr} = \sum_{i=1}^{n} \left\| \mathbf{T}^i_{gt} - \mathbf{T}^i_{pred} \right\|_2^2 \tag{1}$$

where $\mathbf{T}^i_{gt}$ and $\mathbf{T}^i_{pred}$ are the ground-truth and predicted translation vectors for the i-th frame. In our experiments, we first align the translation scale of cameras estimated by ViPE [1] or VGGT [3] with the input cameras before computing TransErr.

RotErr (Camera Rotation Error).

$$\text{RotErr} = \sum_{i=1}^{n} \arccos \frac{\text{tr}\left(\mathbf{R}^i_{gt}\mathbf{R}^{iT}_{pred}\right) - 1}{2} \tag{2}$$

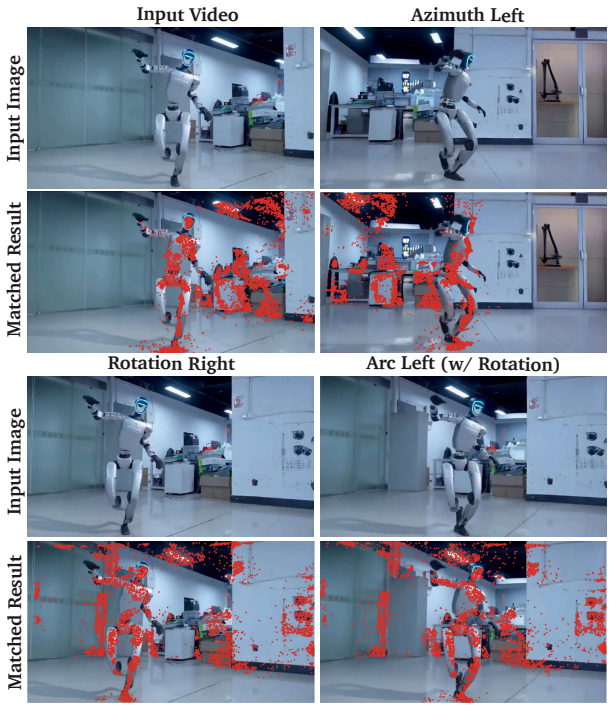

Figure S1. **Image Matching Result.** The red points indicate the matched pixel correspondences across the input images..

where $\mathbf{R}^i_{gt}$ and $\mathbf{R}^i_{pred}$ are the ground-truth and predicted rotation matrices for the i-th frame, and $\mathrm{tr}(\cdot)$ denotes the matrix trace. We report this metric in radians.

Mat. Pix. (Matched Pixels in Video Synchronization).

$$\text{Mat. Pix.} = \sum_{i=1}^{K} \mathbf{1}\big(C_i \geq \tau\big) \tag{3}$$

where K is the total number of pixels, C_i is the confidence score of the i-th pixel, τ is the confidence threshold, and $\mathbf{1}(\cdot)$ is the indicator function. Mat. Pix. counts pixels with confidence above the threshold. The qualitative matching results are presented in Fig. S1.

In our experiments, we set $\tau = 0.5$, resize frames to 432×768, and average all frames. As illustrated in Fig. S2, the sequential 12-shot trajectory is: (1) Rotation Left $\rightarrow$ (2) Arc Right (w/ Rot.) $\rightarrow$ (3) Azimuth Right $\rightarrow$ (4) Rotation Right $\rightarrow$ (5) Arc Left (w/ Rot.) $\rightarrow$ (6) Azimuth Left $\rightarrow$ (7) Tilt Up $\rightarrow$ (8) Translate Down (w/ Rot.) $\rightarrow$ (9) Tilt Down $\rightarrow$ (10) Translate Up (w/ Rot.) $\rightarrow$ (11) Elevation Up $\rightarrow$ (12) Zoom Out. View synchronization is computed for the video pairs shown in Tab. R4.

B. More Qualitative Results

We present additional qualitative results on the Basic Benchmark in Fig. S3, along with comparative visualizations in Fig. S4. Further results on the Agibot Benchmark are shown in Fig. S5, with corresponding comparisons in Fig. S6. We also include extended long video generation results in Fig. S7 and illustrate the focal-length effect in Fig. S8.

References

[1] Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, et al. Vipe: Video pose engine for 3d geometric perception. *arXiv preprint arXiv:2508.10934*, 2025. 1

[2] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *CVPR*, 2022. 1

[3] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *CVPR*, 2025. 1

[4] Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, et al. Longlive: Real-time interactive long video generation. In *ICLR*, 2026. 1

[5] Lvmin Zhang and Maneesh Agrawala. Packing input frame context in next-frame prediction models for video generation. *arXiv preprint arXiv:2504.12626*, 2025. 1

Table R4. Video Pairs for Multi-shot Video Synchronization Calculation on the Basic Benchmark.

N-Shots	Calculated Video Pairs
3 Shots	(Rotation Left, Arc Right (w/ Rot.)) (Rotation Left, Azimuth Right)
6 Shots	(Rotation Left, Arc Right (w/ Rot.)) (Rotation Left, Azimuth Right) (Rotation Right, Arc Left (w/ Rot.)) (Rotation Right, Azimuth Left)
9 Shots	(Rotation Left, Arc Right (w/ Rot.)) (Rotation Left, Azimuth Right) (Rotation Right, Arc Left (w/ Rot.)) (Rotation Right, Azimuth Left) (Tilt Up, Translate Down (w/ Rot.)) (Tilt Down, Translate Up (w/ Rot.))
12 Shots	(Rotation Left, Arc Right (w/ Rot.)) (Rotation Left, Azimuth Right) (Rotation Right, Arc Left (w/ Rot.)) (Rotation Right, Azimuth Left) (Tilt Up, Translate Down (w/ Rot.)) (Tilt Down, Translate Up (w/ Rot.)) (Translate Up (w/ Rot.), Elevation Up) (Translate Up (w/ Rot.), Zoom Out)

Figure S2. **Full Camera Trajectories on the Basic Benchmark.** The sequence proceeds as: (1) Rotation Left→(2) Arc Right (w/ Rot.)→(3) Azimuth Right→(4) Rotation Right→(5) Arc Left (w/ Rot.)→(6) Azimuth Left→(7) Tilt Up→(8) Translate Down (w/ Rot.)→(9) Tilt Down→(10) Translate Up (w/ Rot.)→(11) Elevation Up→(12) Zoom Out.

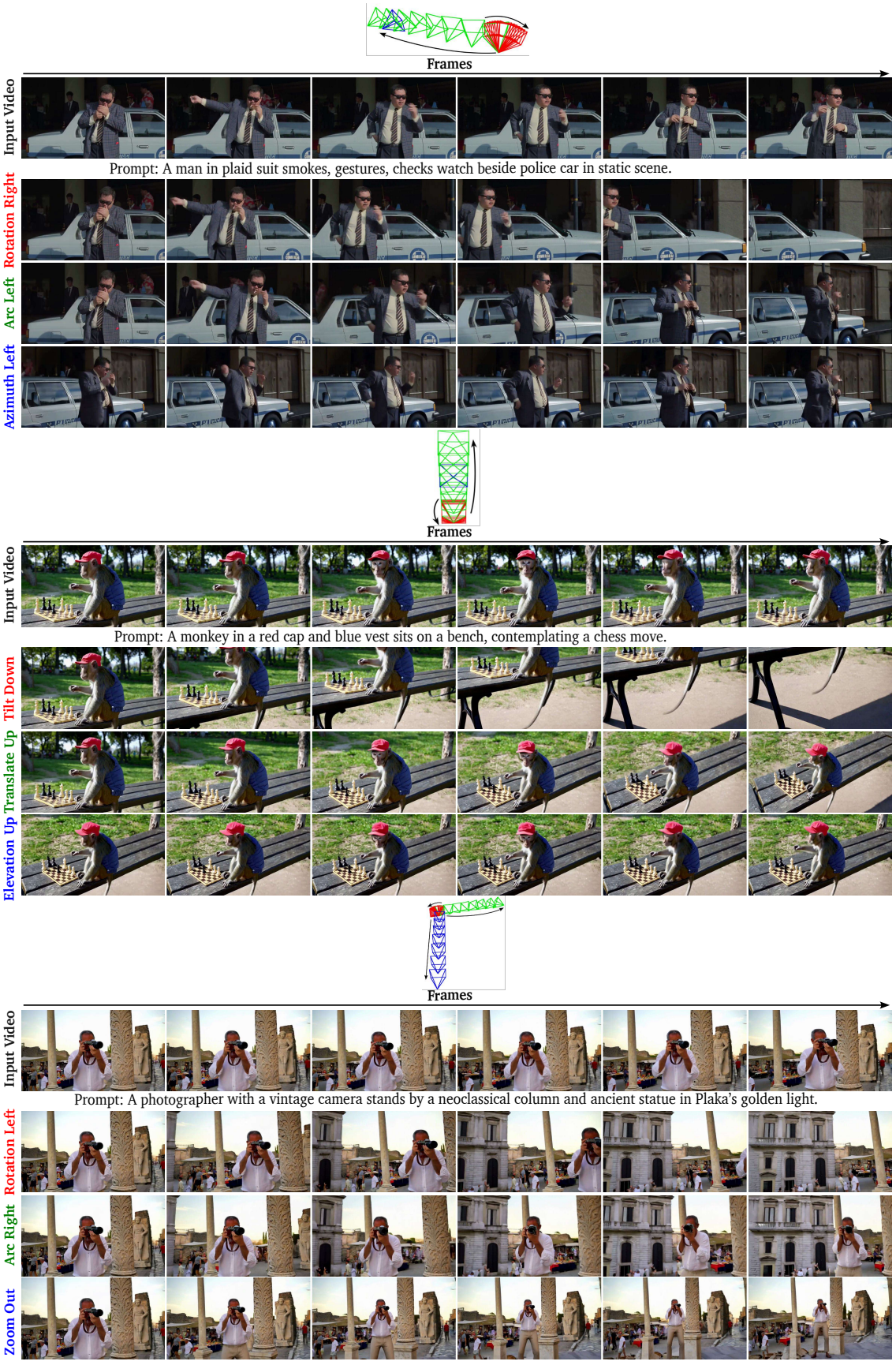

Figure S3. **More Visual Results on the Basic Benchmark.** Our method generates consistent hallucinated context in unseen region.

Figure S4. **More Qualitative Comparison on the Basic Benchmark.** The figures above and below correspond to frames 54 and 88, respectively. Please check full videos on the website provided in the website.

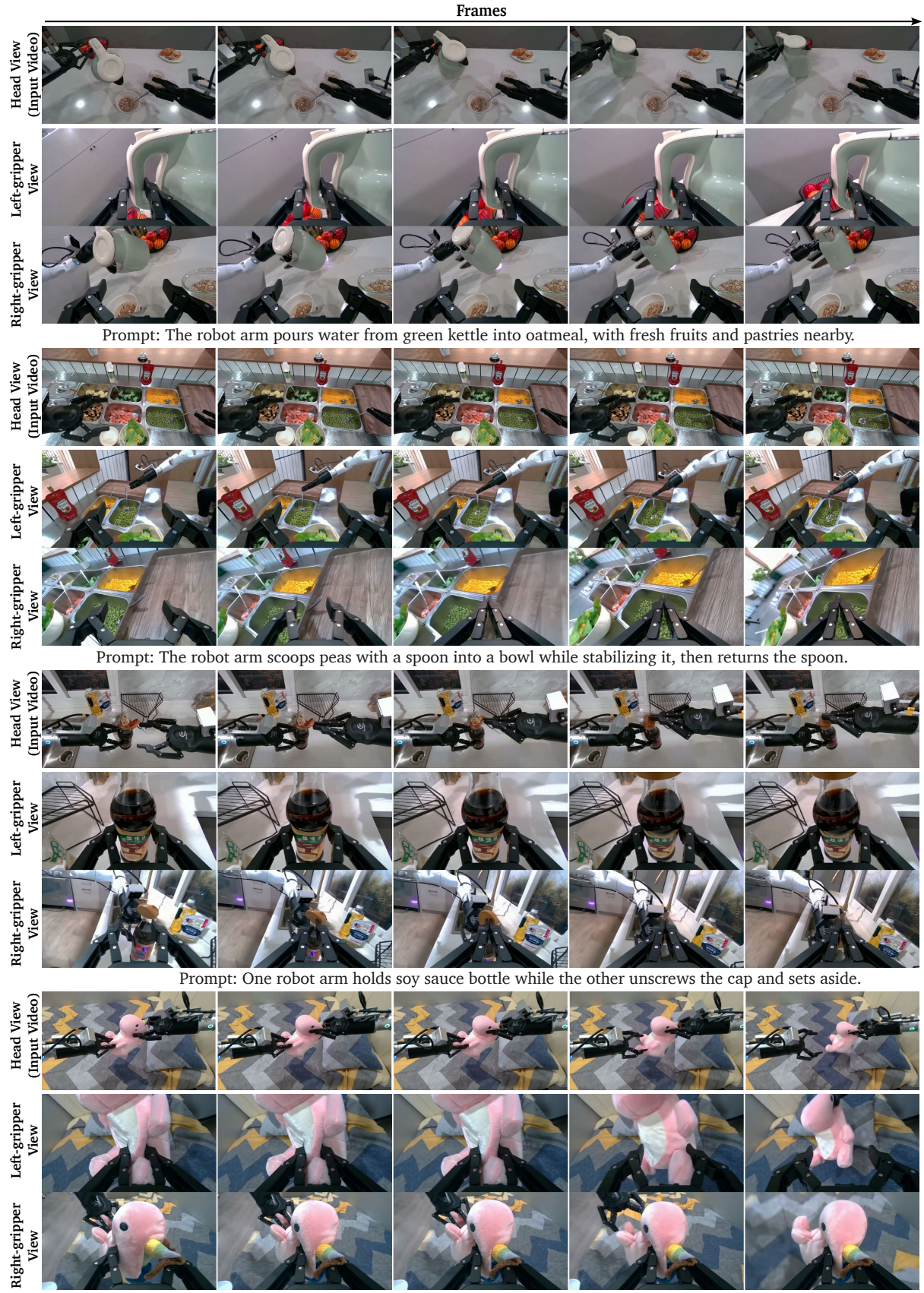

Figure S5. **More Visual Results on the Agibot Benchmark.** The sequence proceeds as: (1) Left-gripper View→(2) Right-gripper View.

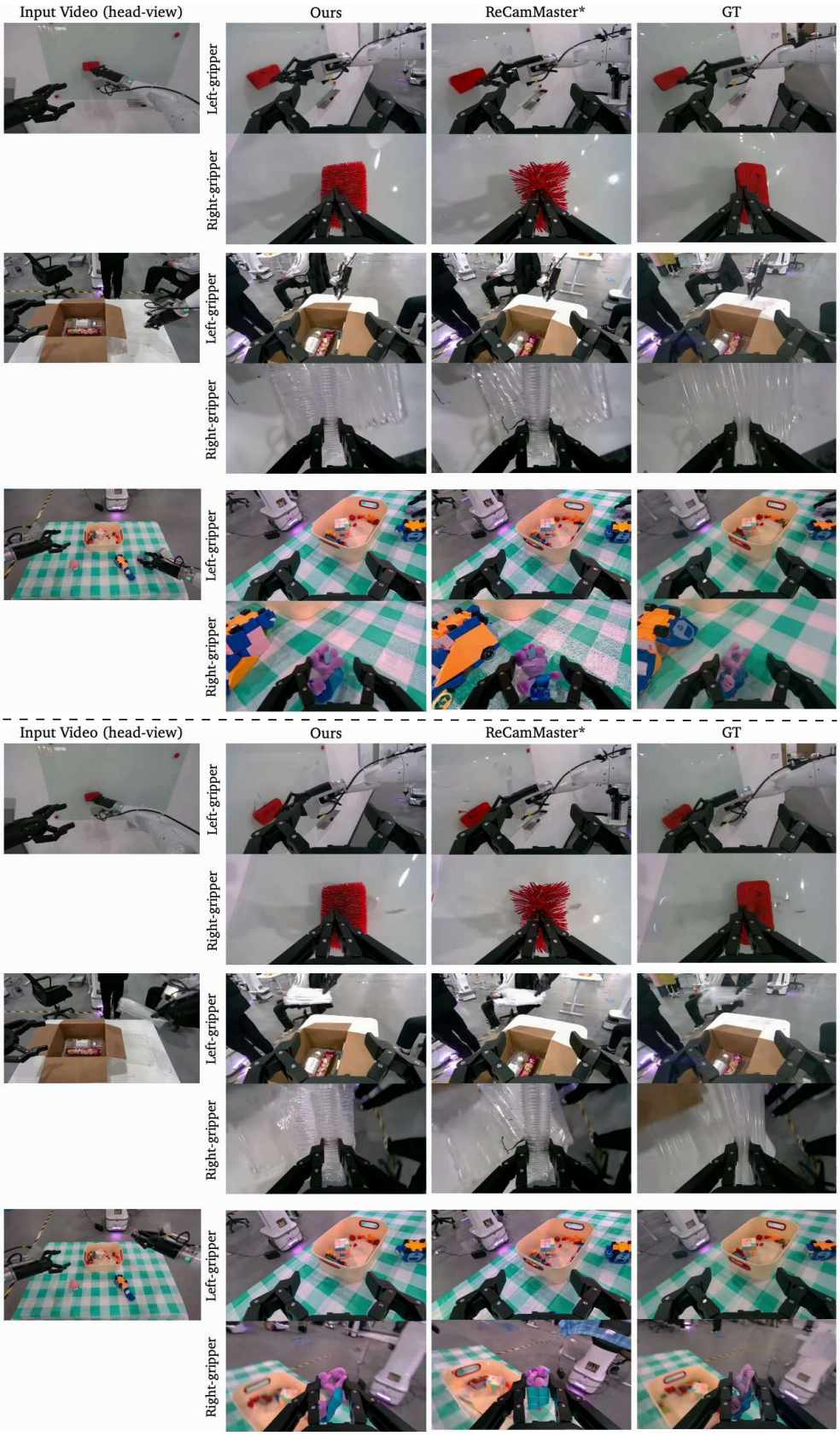

Figure S6. **More Qualitative Comparison on the Agibot Benchmark.** The figures above and below correspond to frames 24 and 93, respectively. Compared to our method, ReCamMaster* exhibits noticeably stronger object distortion and inconsistency.

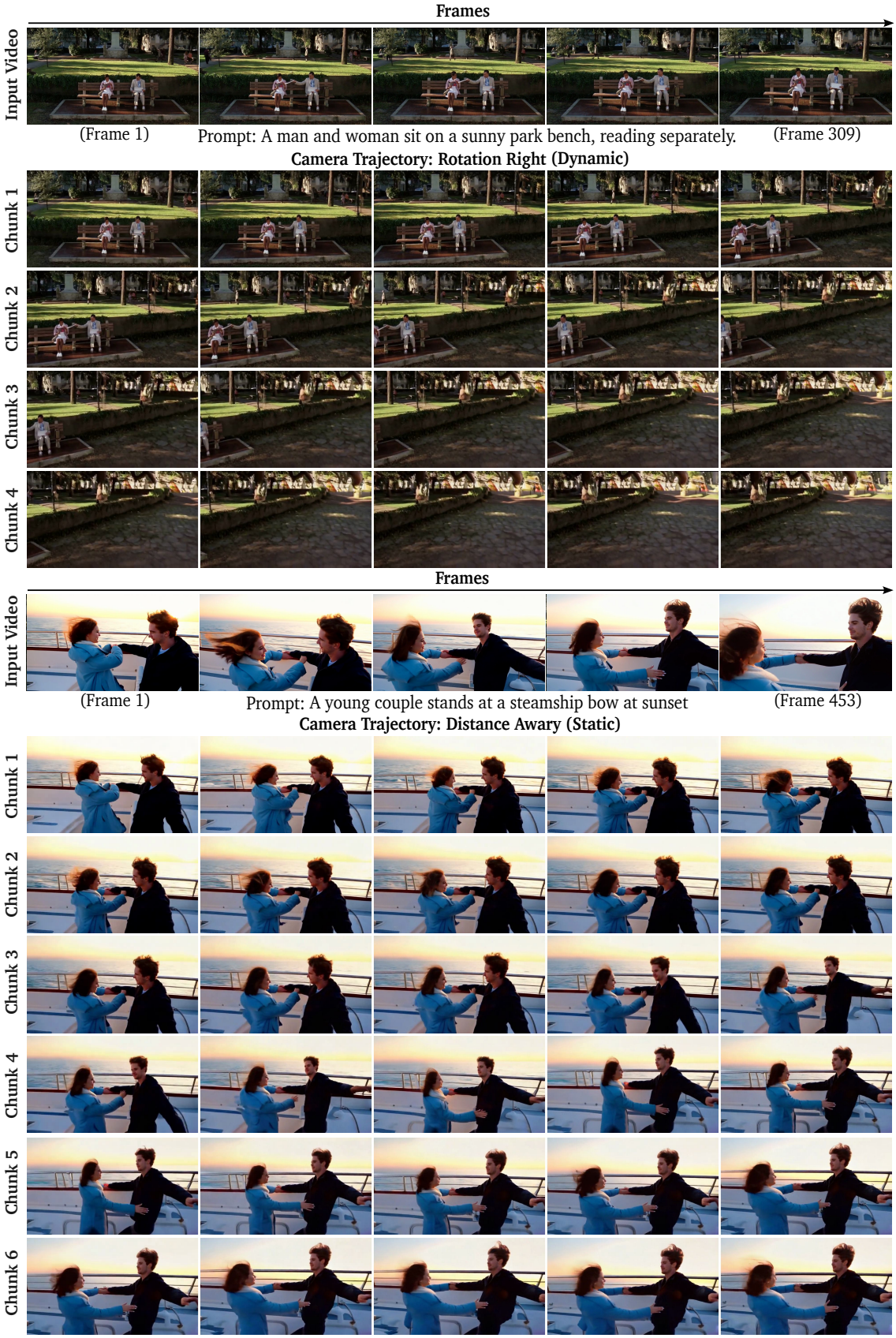

Figure S7. **More Long Video Generation Results under Dynamic and Static Novel-Camera Settings.**

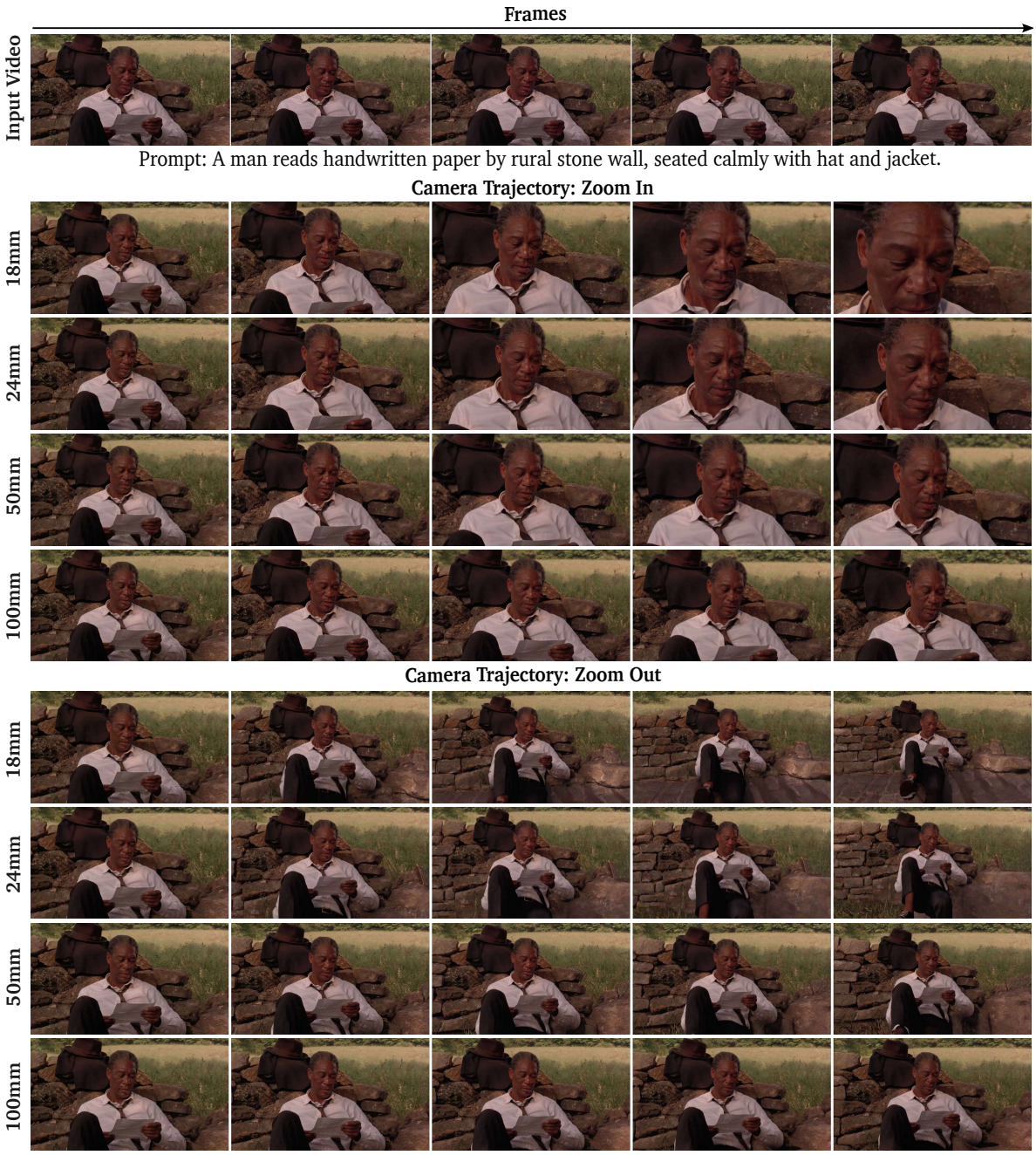

Figure S8. **More Focal Length Effect Results.** Our method synthesizes depth-of-field variations across focal lengths (18mm→100mm). Shorter focal lengths produce more greater changes in the resulting field-of-view (FOV).