

Latent-Reframe: Enabling Camera Control for Video Diffusion Models without Training

Supplementary Material

In the Supplementary Material, we provide additional results of Latent-Reframe (Appendix A), method details (Appendix B) and hyperparameter details (Appendix C).

A. More Results

In this section, we present additional results of Latent-Reframe to demonstrate its versatility across different settings, including basic camera poses (translational and rotational), diverse video styles, image-to-video generation, as well as higher resolution and longer videos.

Basic Pose Generation. Following MotionCtrl [46], the translational basic camera poses comprise six types: zoom in, zoom out, pan left, pan right, pan up, and pan down, as illustrated in Fig. 12. The rotational basic camera poses include four types: the camera’s own rotations (clockwise and counterclockwise) and the camera’s rotations around an object (rotate clockwise and counterclockwise), as shown in Fig. 11. These results indicate that Latent-Reframe can execute various basic camera controls without requiring any training.

Different Style Generation. For video generation in different styles, we follow AnimateDiff [12] and apply complex camera pose control across six different video generation styles: FilmVelvia, ToonYou, MajicMix, RcnzCartoon, Lyriel, and Tusun, as shown in Fig. 13. These results demonstrate that Latent-Reframe can effectively manage camera control for video generation across a wide range of styles, highlighting its versatility and suitability.

Camera Control Image-to-Video Generation. To demonstrate the versatility of our approach, we further present image-to-video generation results with camera control based on the EasyAnimate model [50] in Fig. 14. Our method maintains both visual quality and effective camera control while preserving content from the input image. Notably, as shown in the last row of Fig. 14, it also handles inputs lacking explicit 3D structure, such as abstract paintings, and still enables camera motion.

Higher Resolution and Longer Video Generation. Since our method builds on pretrained video diffusion models, it naturally inherits their capacity for higher resolution and longer video generation. As shown in Fig. 15 and Fig. 16, we evaluate our method using the EasyAnimate model [50], scaling up the resolution to 768×1024 and extending the video length to 49 frames. The results demonstrate that our method continues to support effective camera control under these challenging settings. As base video models advance, our approach is expected to yield even higher-quality con-

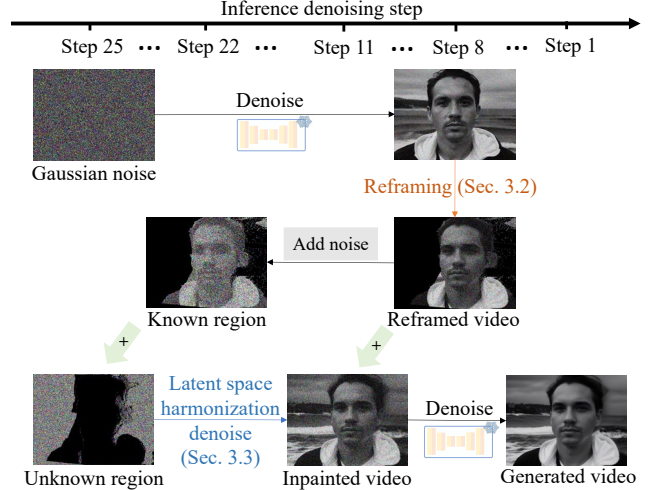


Figure 10. **Denoising time line of Latent-Reframe.** Here, the diffusion step for latent reframing is set to 8, and the noise reduction step is set to 3.

trolled video outputs.

B. Denoising Process Details

To provide a clearer explanation of the Latent-Reframe inference denoising process, we present the denoising timeline in Fig. 10.

1. The process begins with fully Gaussian noise at the 25th step. Denoising is conducted using the video diffusion model until reaching the predetermined latent reframing step, which is step 8.
2. Latent reframing, as described in Sec. 3.2, is then applied to generate the reframed video at the target camera pose. At this point, holes caused by occlusions are present, and the regions are differentiated into known and unknown regions.
3. The noise addition process starts. As outlined in Sec. 3.3, the noise level in the known region is set to be 3 steps lower than that in the unknown region. For instance, when the unknown region is at the 25th denoising step, the known region is at the 22th denoising step.
4. Denoising of the unknown region proceeds. At this stage, the input to the denoising network combines unknown and known regions, as described in Eq. 4 of the main paper. The output of the denoising network updates the unknown region for the next step, while the known region for the next step is obtained by adding corresponding noise to the reframed video.

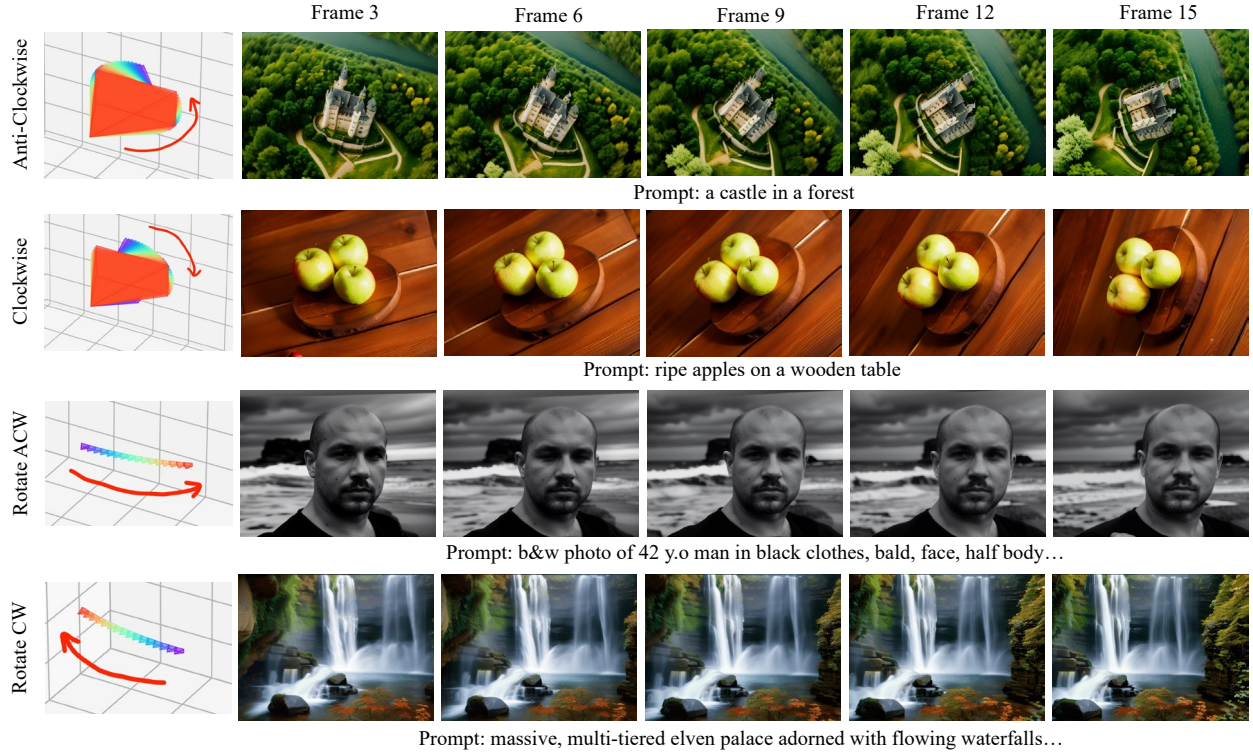


Figure 11. Results of Latent-Reframe camera control video generation based on rotational basic poses.

5. When the unknown region reaches the 11th denoising step, the known region has already reached the 8th step, which is the step for latent reframing. After this, the known region is no longer merged. The denoising process continues normally, updating the entire video until the final step.

C. Hyperparameter Settings

Tab. 2 further details the parameters employed in Latent-Reframe.

Table 2. Detail hyperparameter settings.

Parameter	Value
Video frames	16
Spatial resolution	512×384
DDIM denoising steps	25
Classifier-free guidance scale	7.5
Diffusion step for Latent-Reframe	8
Noise reduction step	3
Point map prediction model	MonST3R [56]
Sliding window size	3
Globally aligned points optimization steps	300

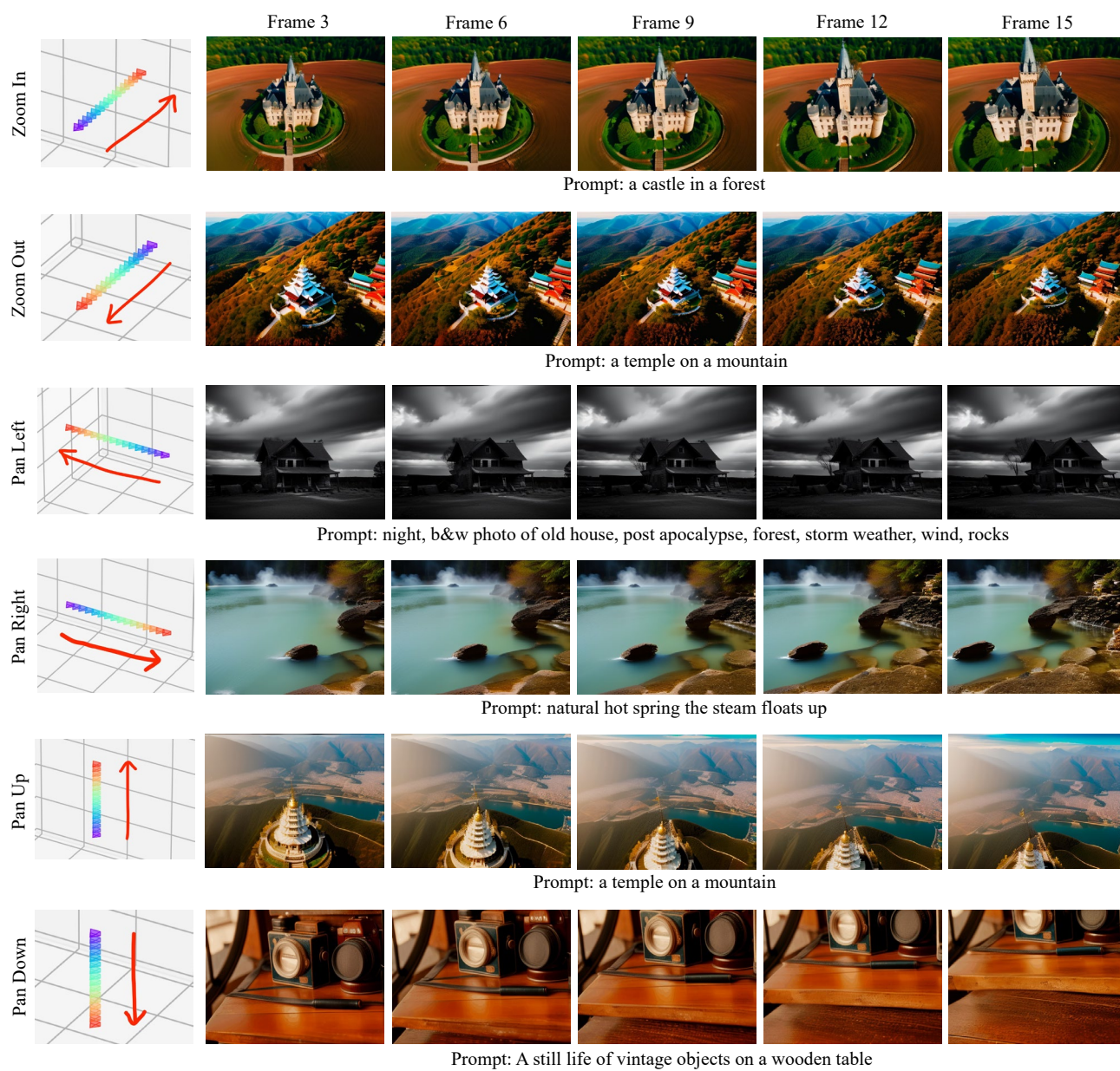


Figure 12. Results of Latent-Reframe camera control video generation based on translational basic poses.

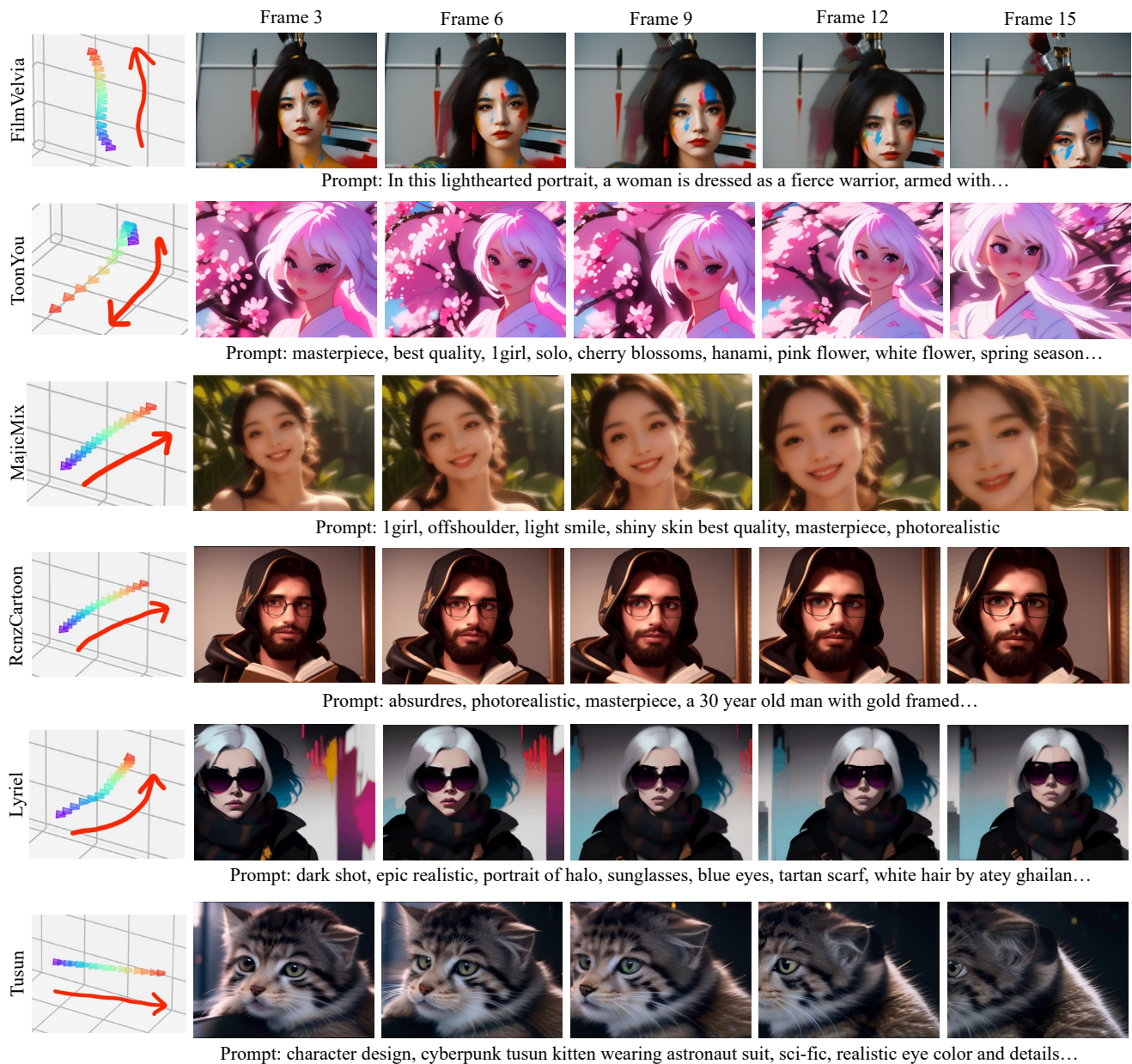


Figure 13. Results of Latent-Reframe complex camera control in various video styles.

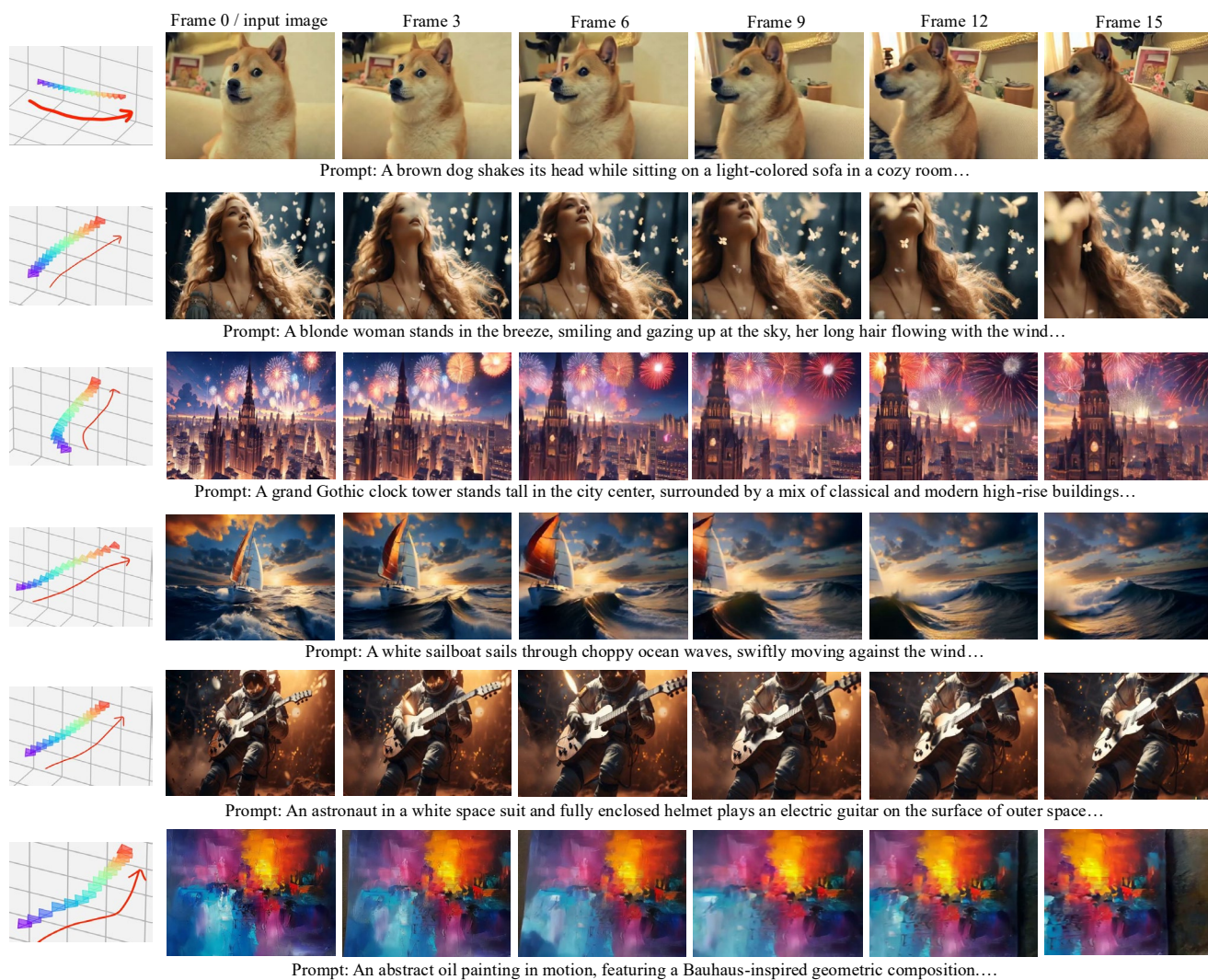


Figure 14. **Results of Latent-Reframe complex camera control in image-to-video generation. The input image also serves as the first frame of the generated video.**

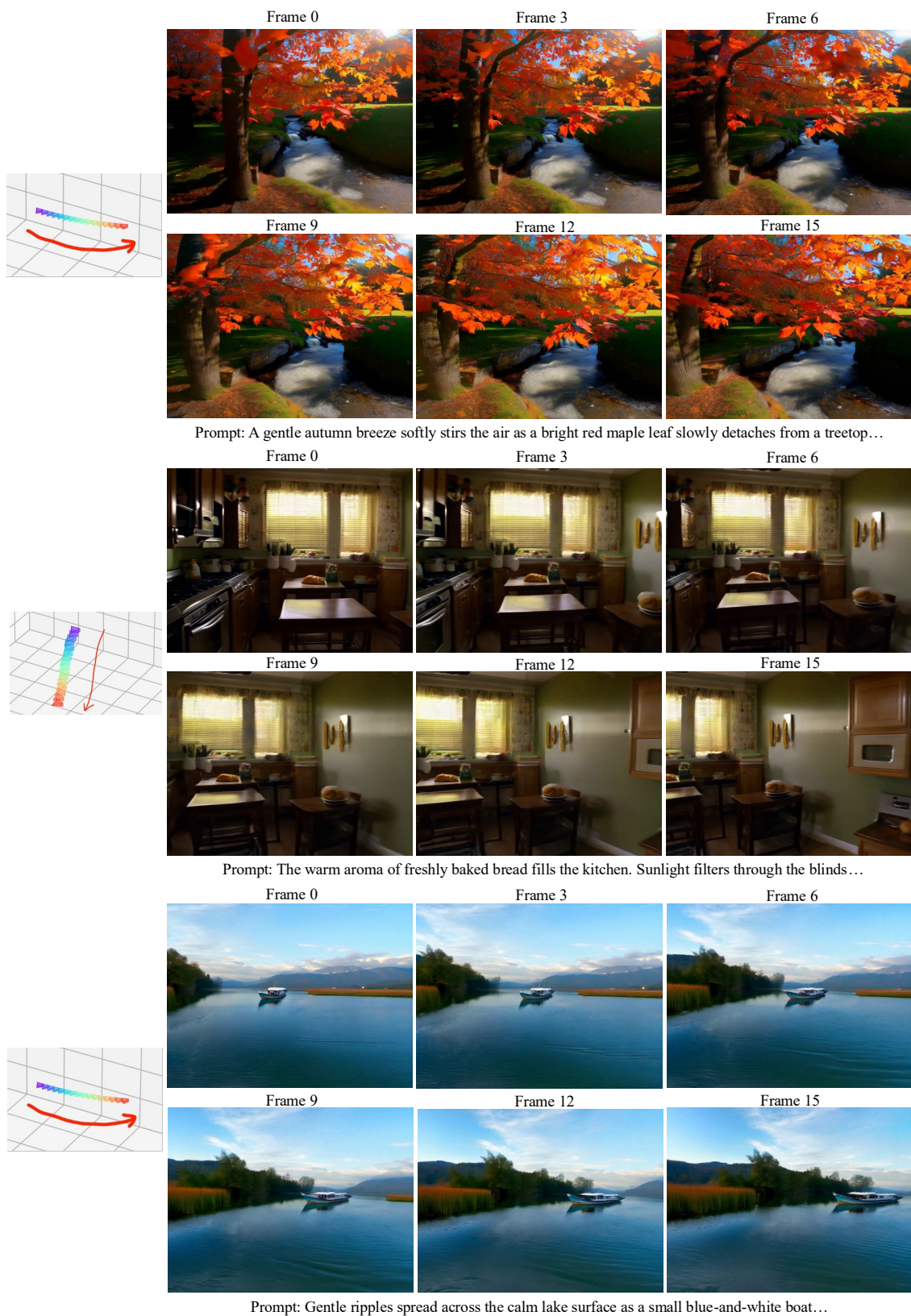


Figure 15. Results of Latent-Reframe complex camera control in higher resolution (768×1024) generation.

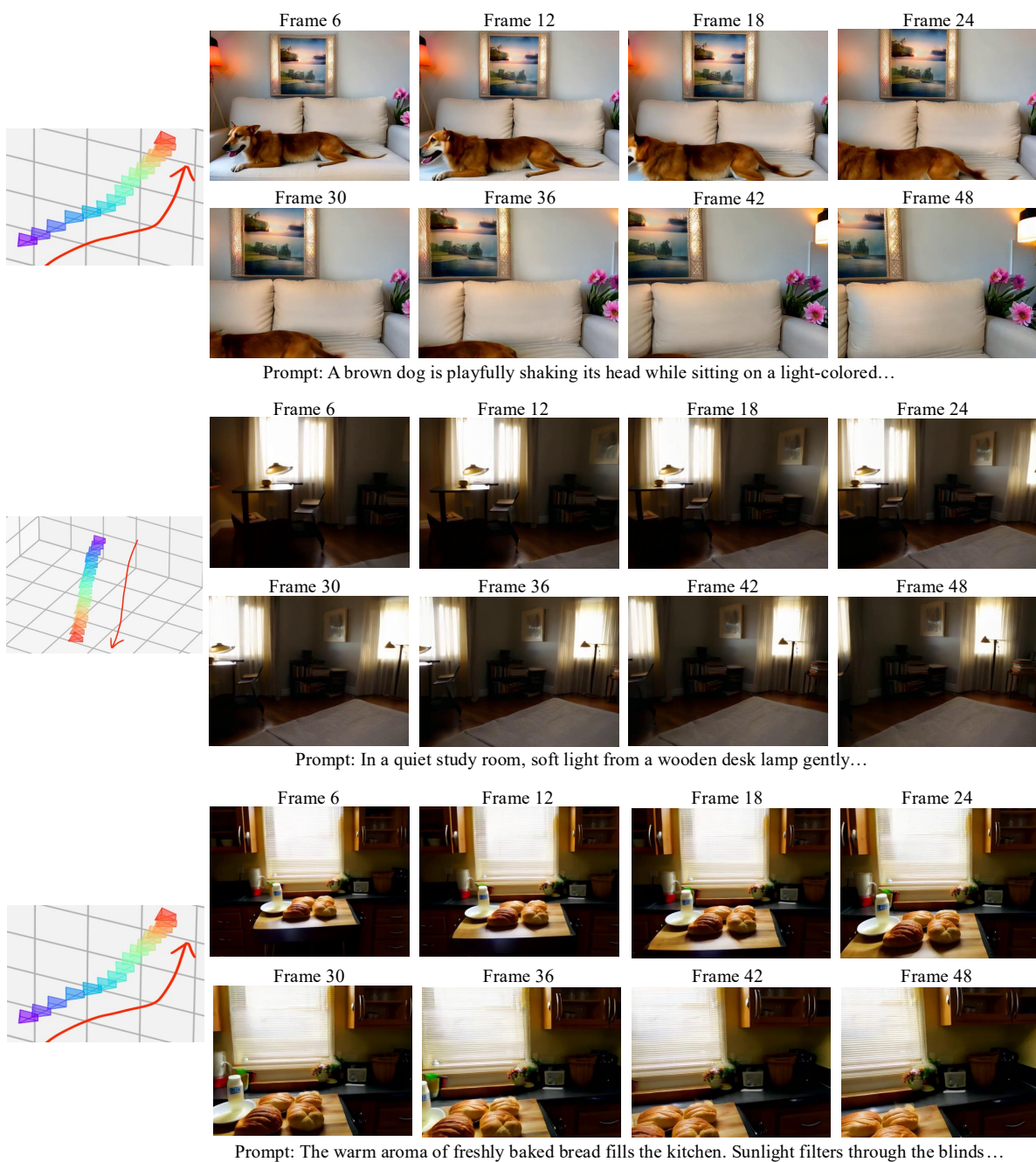


Figure 16. Results of Latent-Reframe complex camera control in longer video (49 frames) generation.