A. Webpage Demo

The videos in the main paper and appendix can be viewed with our demo webpage by opening the `webpage/index.html` in the supplementary material using a web browser.

B. Additional Results from Human Evaluation

As mentioned in Section 6.2, we conduct A/B comparison between Fairy with baselines. We ask annotators to compare our generated video with a baseline method's result, and decide which one is better.Each video pair is evaluated by three independent annotators, and the majority vote is considered as the final rating. We ask raters to evaluate by four attributes: frame quality – visual quality of single frame; temporal consistency – whether the frames are coherent or flickering; prompt faithfulness – whether the output followed the editing instruction or target prompt; input faithfulness – whether the output video followed the contents of the original video. We reported the overall rating in Section 6.2, Figure 10. Here, we report a more detailed comparison along each attributes in Figure 11, 12, 13. Compared with Rerender (Figure 11), Fairy loses in terms of single frame visual quality. This is mainly due to the limitation of the foundational image editing model, while Rerender utilizes LoRA to enhance frame quality. Yet, Fairy significantly outperforms Rerender in terms of temporal consistency, achieves better prompt faithfulness, and performs similarly in terms of input faithfulness. Compared with TokenFlow, Fairy outperforms significantly in terms of frame quality, temporal consistency, and prompt faithfulness. They performs similarly in terms of input faithfulness. Compared with Gen-1, Fairy significantly outperforms in all attributes.

In addition to the A/B comparison, in which we ask human annotators to compare our method with a baseline, we also conduct a standalone evaluation to examine output video's quality. Each time we show an annotator the original video, an editing instruction, and the result video. We then ask the annotator to rate the output as good or bad by the same four attributes. We ask 3 annotators to rate each video, and the decision is determined by their majority vote. We report the success rate by each attributes in Figure 14.

C. More Results

C.1. Character Swap

In Figure 15, we demonstrate more results of character swap, where Fairy is able to interchange individuals with various characters. Note that our model can adapt to different input aspect ratios without need for re-training.

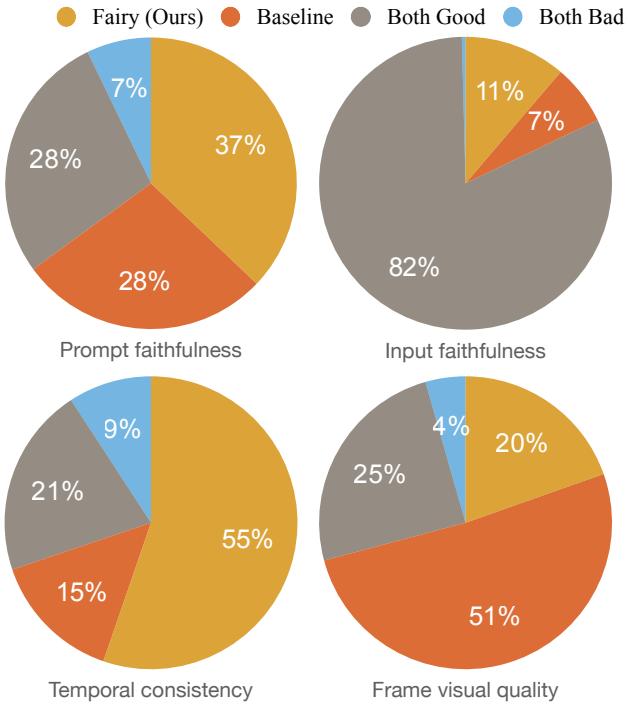

Figure 11. **Comparison with Rerender.**

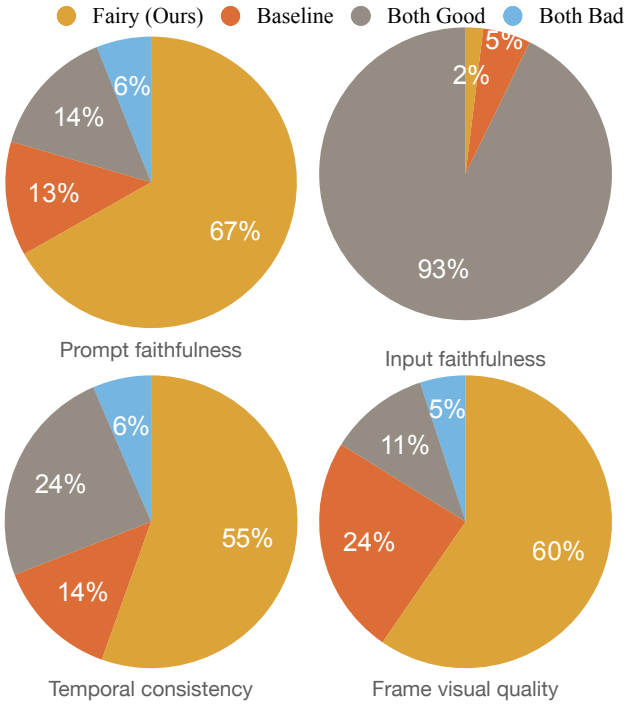

Figure 12. **Comparison with TokenFlow.**

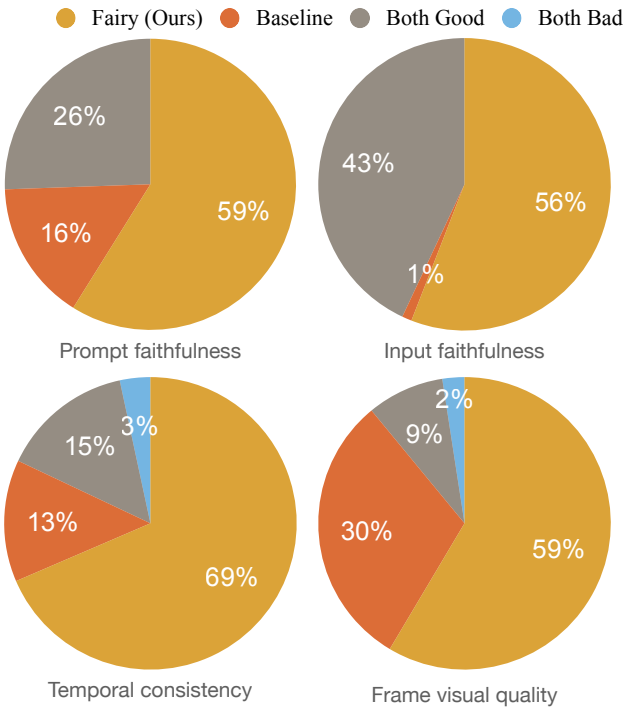

Figure 13. **Comparison with Gen-1.**

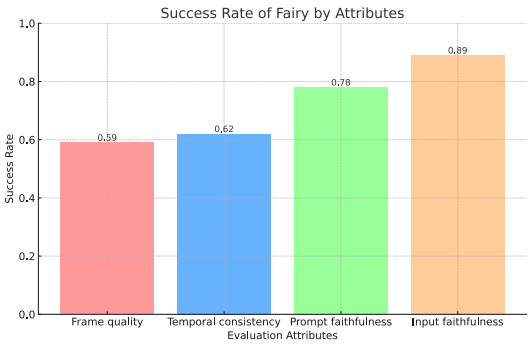

Figure 14. **Standalone success rate by attributes.** We report Fairy's success rate in terms of frame quality, temporal consistency, prompt faithfulness, and input faithfulness.

C.2. Stylization

Figure 16 demonstrates more stylization results of Fairy. In particular, our model is able to recognize various styles, while perform high quality and temporal consistent edit based on the stylization instructions.

C.3. Arbitrary Long Videos

Fairy is able to scale to arbitrary long video without memory issue due to the proposed anchor-based attention. In Figure 17, we show that our model is able to generate a 27 second long video with high quality, while the latency is less than 71.89 seconds via 6 A100 GPUs. In particular, the Fairy manage to retain decent temporal consistency even number of frames (664 frames) is way more than the number of anchor frames (3 frames).

C.4. Ablation Study

Figure 18 shows more ablation results by removing equivariant finetuning and anchor-based attention. We can see that without equivariant finetuning, the model is sensitive to local motion and movement of the subject and therefore degenerate the frame quality and temporal consistency. For instance, in the first video, the tail of the cat becomes the head of the lion in some of the frames, and the face of the cat in second video vary significantly between frames. Without anchor-based attention, the edit of each frame is completely independent, rendering in significant worse temporal consistency.

Figure 19 demonstrates results generated with different number of anchor frames. When number of anchor frames equals to 1, the global features model can leverage are too restricted, which lead to suboptimal edits. In contrast, we observe that when the number of anchor frames is greater than 7, the quality also gradually degrades, losing some visual details.

In Figure 20, we perform ablation study on the number of diffusion steps during generation. The model perform reasonably well when the number of diffusion step is above 10. We therefore set the diffusion step to 10 for all of our experiments to optimize the latency.

C.5. Limitations

Finally, Figure 21 demonstrates some limitations we point out in section 6.4. Since the model is never trained on video data, it does learn to generate concepts containing motion such as raining, lightning, or flames. Fairy also inherits the limit of the image editing model, where it is not able to follow the instructions that involve camera motion, such as zoom in or zoom out.

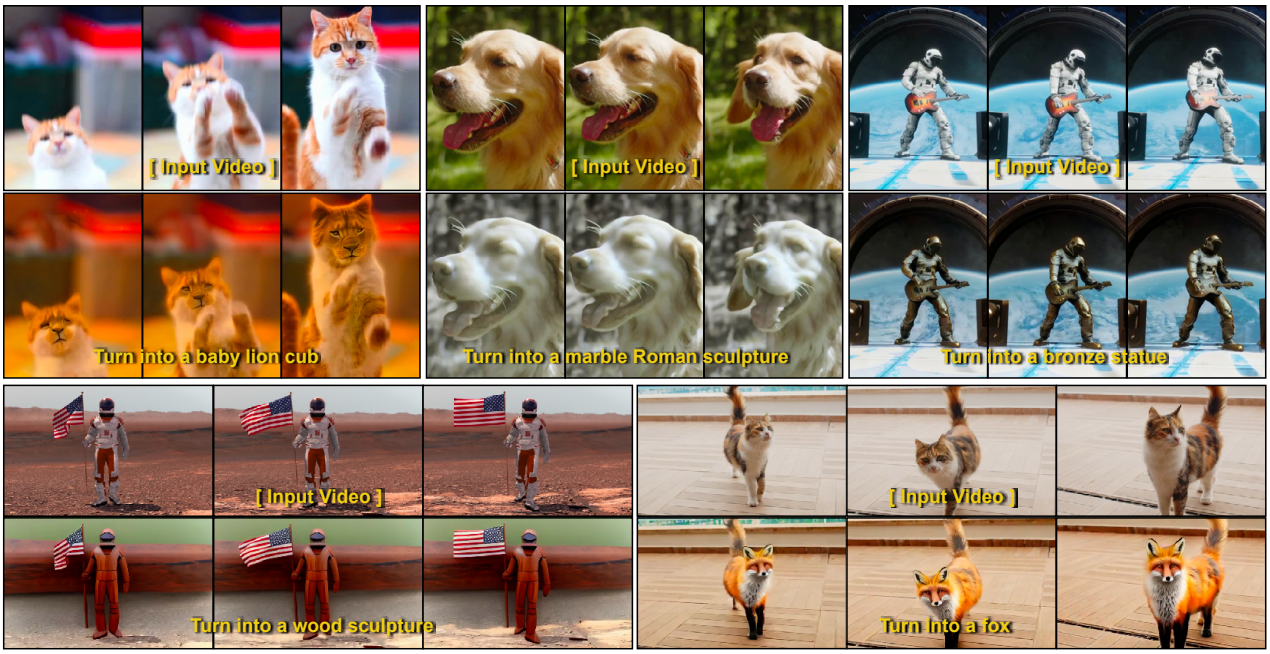

Figure 15. **Additional Results on Character Swap:** Fairy is able to interchange the characters for videos with arbitrary ratios.

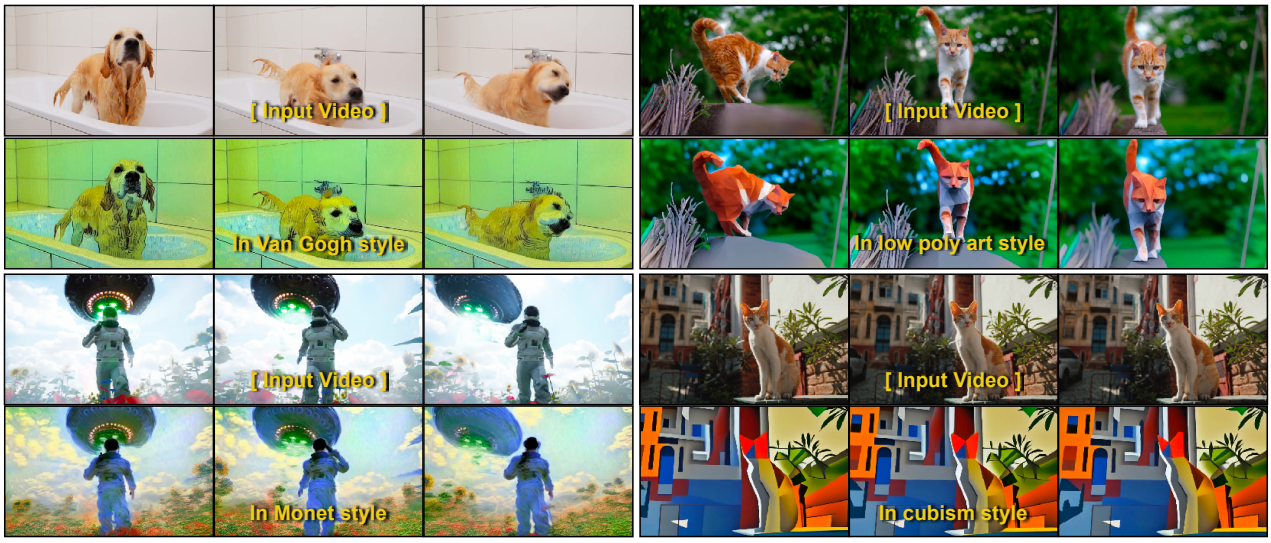

Figure 16. **Additional Results on Stylization:** Fairy enables a wide range of style editing.

Figure 17. **Any-length Video Editing.** Fairy is able to scale to arbitrary long video without memory issue.

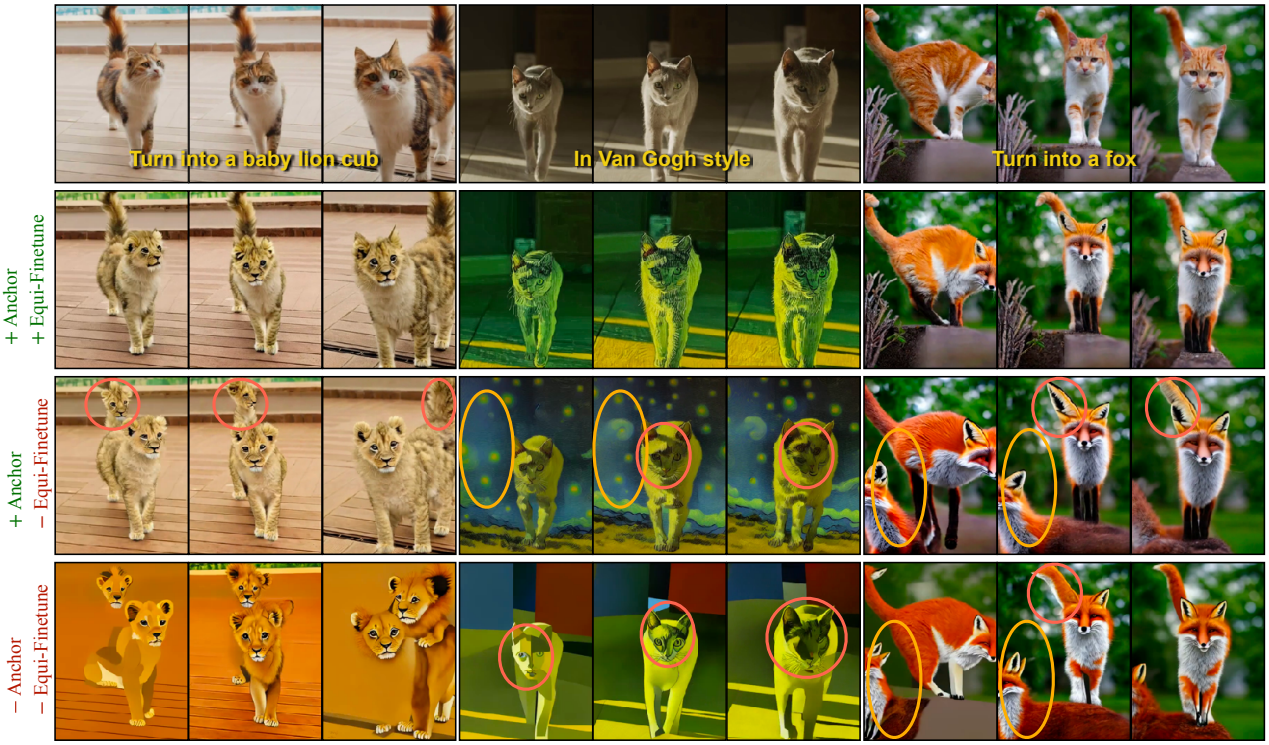

Figure 18. **Additional Results on Ablation Study.** We demonstrate that both equivariant finetuning and anchor-based attention are crucial to Fairy .

Figure 19. **Ablation study on number of anchor frames.** We found that setting number of anchor frames to 3 yields the best results.

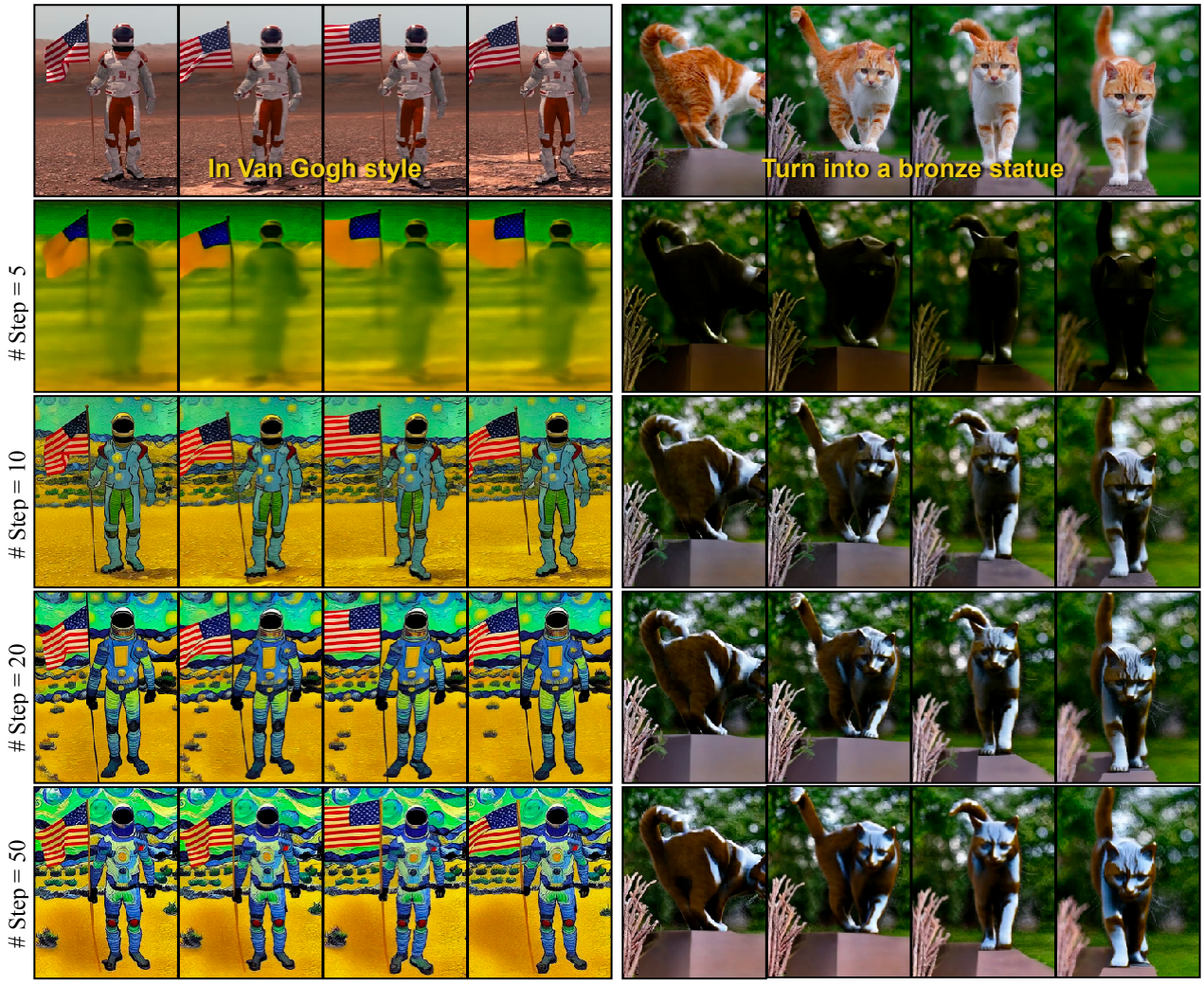

Figure 20. **Ablation study on number of diffusion steps.** We found that diffusion steps above 5 generally yield good results.

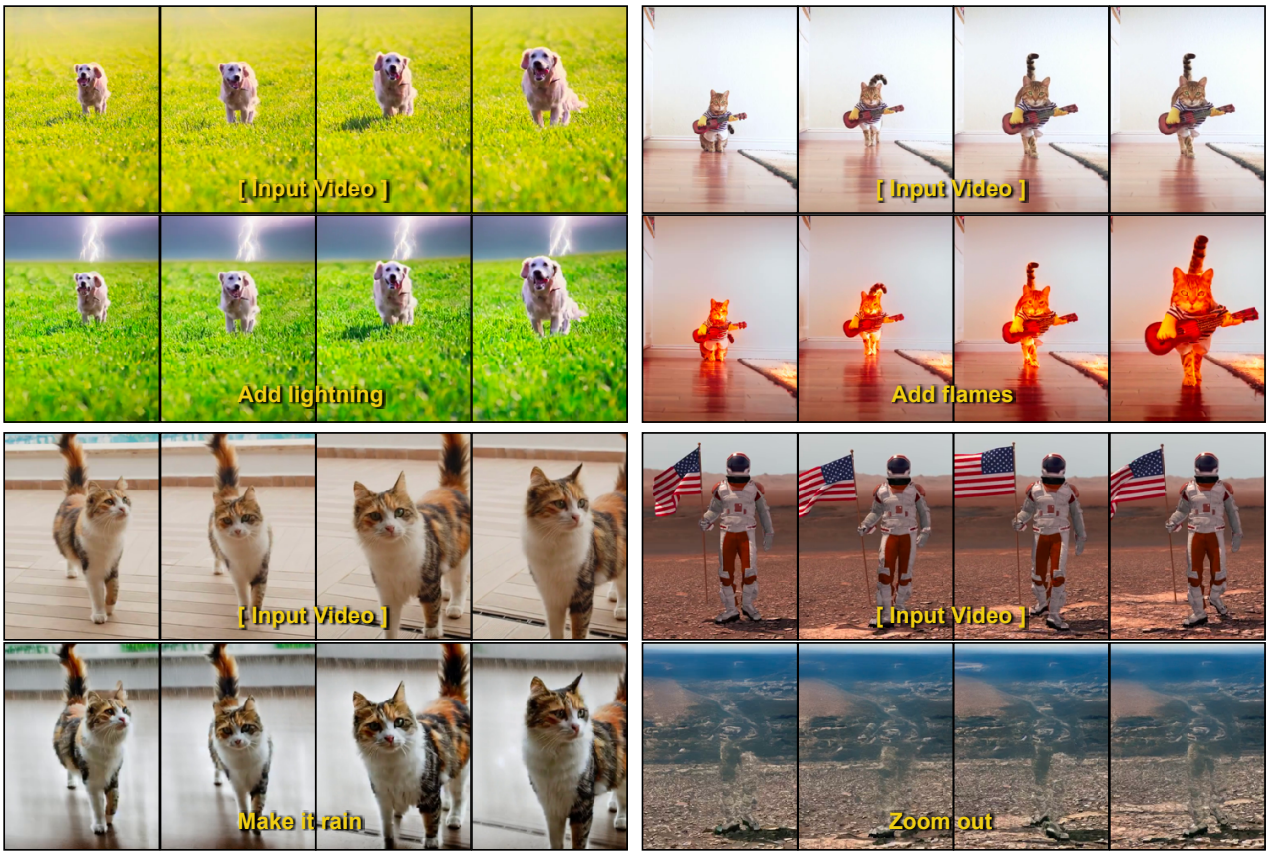

Figure 21. **Limitations of Fairy.** Our model cannot accurately render dynamic visual effects, such as lightning, flames, or rain.