LAMP: Learn A Motion Pattern for Few-Shot Video Generation

Ruiqi Wu[1,3*] Liangyu Chen[3] Tong Yang[3] Chunle Guo[2,1†] Chongyi Li[2,1] Xiangyu Zhang[3]
[1]VCIP, CS, Nankai University [2]NKIARI, Shenzhen Futian [3]MEGVII Technology
wuruiqi@mail.nankai.edu.cn, {chenliangyu, yangtong, zhangxiangyu}@megvii.com
{guochunle, lichongyi}@nankai.edu.cn

Abstract

Our supplementary materials give more details of our LAMP and more experiment results, which can be summarized as follows:

- *We introduce more details of our LAMP, especially the inference stage.*
- *We give more ablation experiments and more visual results.*
- *We provide the source code and video files.*

1. More Details of LAMP

In this section, we describe the inference process of our method more specifically. We first use a text-to-image model $\mathcal{M}_I$ *e.g.* SD-XL [2] to generate the first frame. Then, the latent features of the first frame are concatenated with the noise acquired by shared-noise sampling. At each step, only the features of the subsequent frames are updated by our video diffusion model $\mathcal{M}_V$ based on SD-v1.4 [3]. Besides, AdaIN [1] is used to ensure appearance consistency. Notice that AdaIN is only used in subsequent frames in the first 30 steps since forcing the latent features of subsequent frames, which are closer to the noise, to be consistent with the first frame can cause unwanted artifacts. Moreover, histogram matching is adopted on pixel space to remove the flicker. The pseudo code of the inference process is illustrated in Alg 1, where $\mathcal{P}_I$ and $\mathcal{P}_V$ are the prompts of the first frame and the whole video, t is the video length, and $T = 50$ denotes the total step of DDIM backward.

Algorithm 1 Pseudo code for inference stage

Require: $\mathcal{P}_I, \mathcal{P}_V, \mathcal{M}_I, \mathcal{M}_V, f \in \mathbb{N}, T \in \mathbb{N}$ and latent decoder $\mathcal{D}$

```
▷ First frame generation
x_T^1 ~ N(0, I)
x_0^1 = DDIM_Backward(x_T^1, T, P_I, M_I)
▷ Shared-noise sampling
x_b ~ N(0, I)
for i = {2, 3, ..., f} do
    x_T^i ~ N(0, I)
    x_T^i ← 0.8 x_T^i + 0.2 x_b
end for
▷ Video generation
x_T^{1:f} ← {x_0^1, x_T^2, x_T^3, ..., x_T^f}
for t = T − 1, ..., 0 do
    x_t^{2:f} ← M_V^{2:f}(x_{t+1}^{1:f}, t + 1, P_V)
    ▷ Post-processing on latent space
    if t > 20 then
        for i = 3, 4, ..., f do
            x_t^i ← AdaIN(x_t^i, x_t^2) ▷ Ensure consistency between predicted frames
        end for
    else
        for i = 2, 3, ..., f do
            x_t^i ← AdaIN(x_t^i, x_0^1) ▷ Ensure consistency with the first frame
        end for
    end if
end for
▷ Post-processing on pixel space
I^{1:f} = D(x_0^{1:f})
for i = 2, 3, ..., f do
    I^i ← Historgram_Matching(I^i, I^1)
end for
```

2. Experiments

2.1. More Ablation Study

In this section, we give more ablative results of inference processing as we can see in Fig 1. When we remove the AdaIN during the inference stage, the appearance consistency will be corrupted. *e.g.*, the horse is missing its front hooves in the last frame. Besides, the histogram matching can effectively restore the flicker between frames as illustrated in Fig 1(d). We empirically use AdaIN only between subsequent frames in the first 30 steps of DDIM backward, and in the last 20 steps, make the subsequent frames con-

(a) LAMP (full model)

(b) w/o post-processing

(c) w/o AdaIN

(d) w/o histogram matching

(e) AdaIN with the first frame in all steps

Figure 1. More ablative results.

sistent with the first frame passes through AdaIN. We try to employ AdaIN between subsequent frames and the first frame from the start. However, forcing subsequent frames that are close to noise during the early stage to be consistent with the first frame can produce unexpected artifacts as shown in Fig 1(e).

2.2. More Visual Results

In our supplementary materials, we provide more visual results of 8 motion patterns as shown in Fig. 2-9. Our LAMP can generate diverse and high-quality results with proper motion. Besides, the video files are also provided.

3. Limitation and Future Works

In our experiments, we observed that the occurrence of failure cases increased as our method attempted to learn complex motions. More effective modules for motion learning are potential solutions to this issue. Besides, we found that the motion of the foreground object sometimes influences the background's stability. We believe that learning the foreground and background movements independently might be an effective solution. We leave these improvements in our future work.

4. Broader Impacts

Because our work is based on existing text-to-image techniques, it may carry flaws in the pre-trained models themselves. Like other generative models, our model can generate unsafe or biased videos, which may cause harm without a safety checker. Before the model is actually deployed, it is important to thoroughly evaluate the potential risks of the model and filter for harmful content.

References

[1] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 1

[2] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 1

[3] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

Figure 2. Visual results on ‘horse run’ motion. **Video files can be found in supplementary materials**.

Figure 3. Visual results on ‘birds fly’ motion. **Video files can be found in supplementary materials**.

Figure 4. Visual results on ‘helicopter’ motion. **Video files can be found in supplementary materials**.

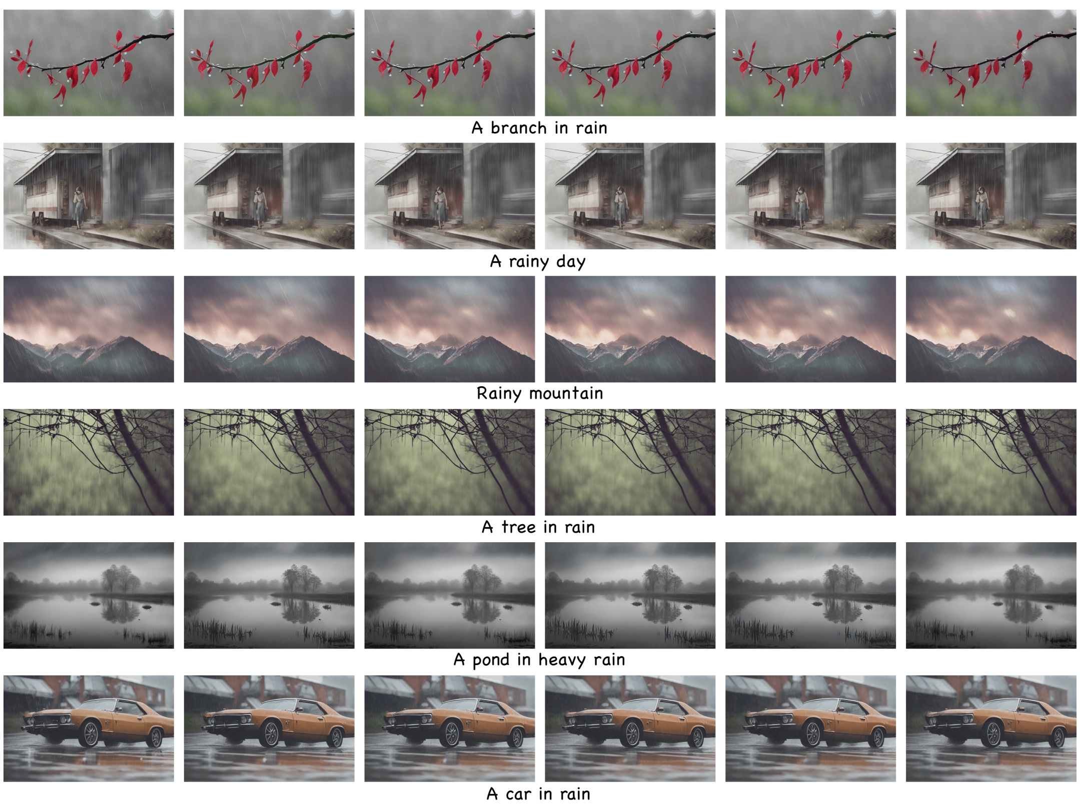

Figure 5. Visual results on ‘rain’ motion. **Video files can be found in supplementary materials**.

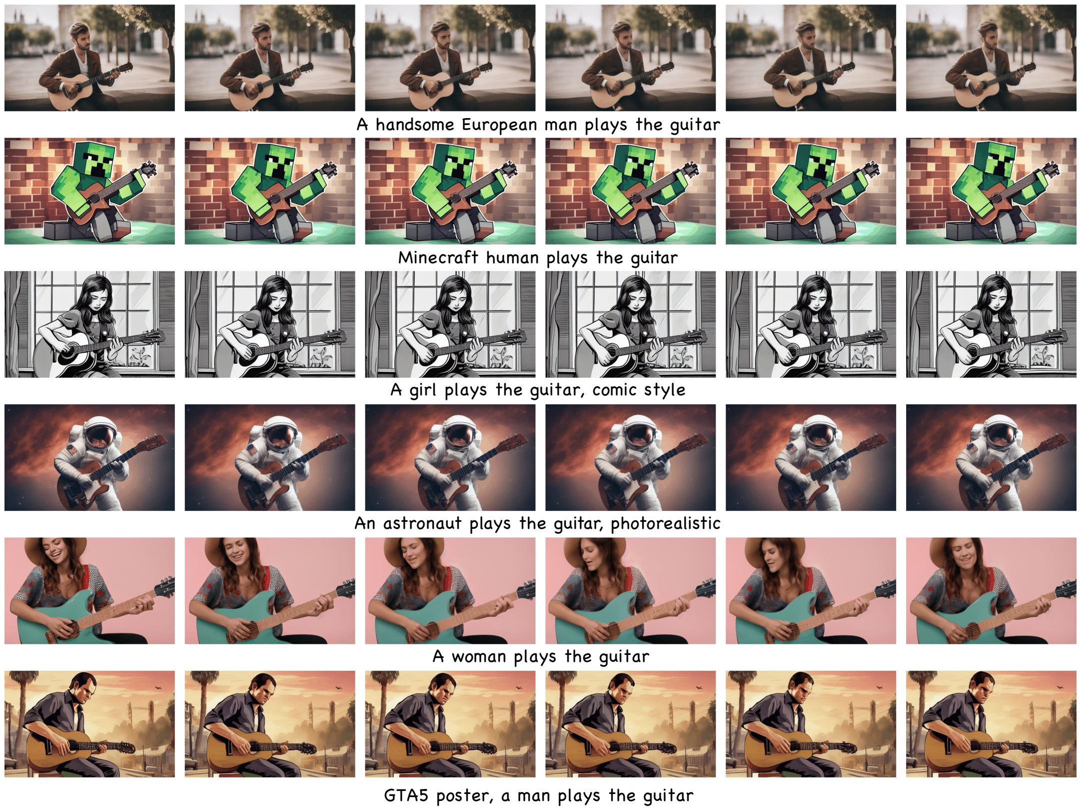

Figure 6. Visual results on 'play the guitar' motion. **Video files can be found in supplementary materials**.

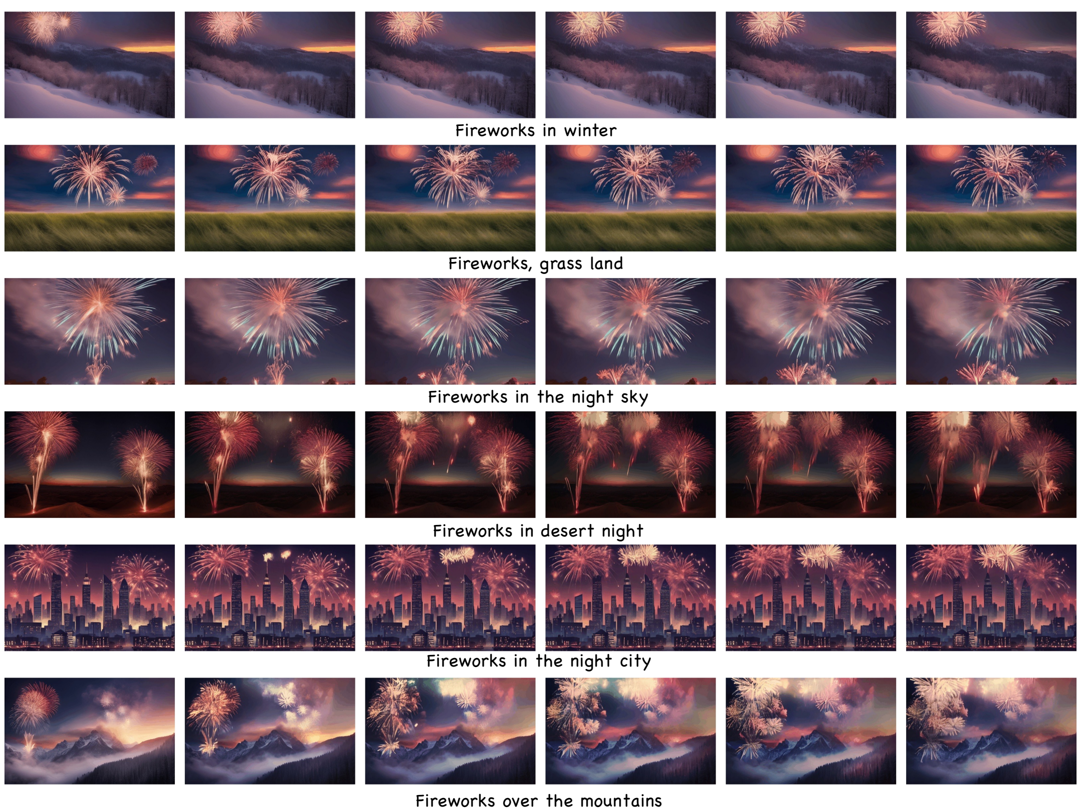

Figure 7. Visual results on 'firework' motion. **Video files can be found in supplementary materials**.

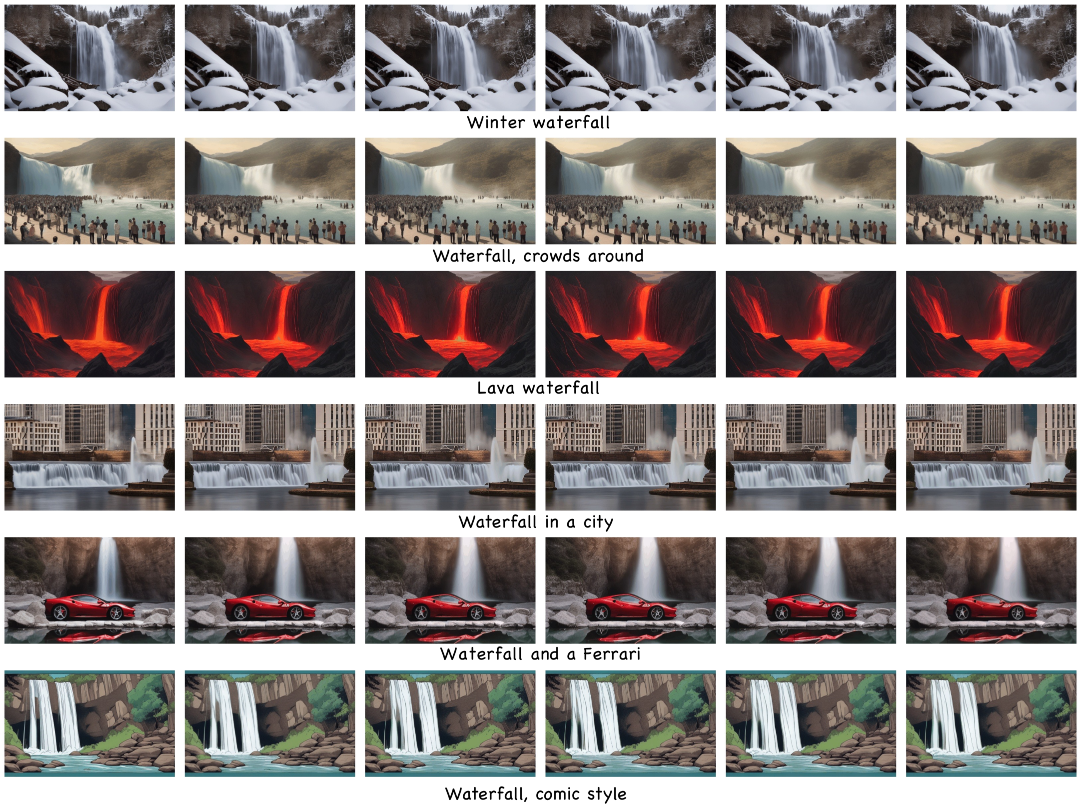

Figure 8. Visual results on 'waterfall' motion. **Video files can be found in supplementary materials**.

Figure 9. Visual results on 'turn to smile' motion. **Video files can be found in supplementary materials**.