Controllable Weather Synthesis and Removal with Video Diffusion Models

Supplementary Material

In the supplementary material, we provide additional implementation details (Sec. A) and further results (Sec. B). Please refer to the project website for more qualitative results and comparisons.

A. Implementation Details

Training Details Both weather removal and synthesis models are trained using AdamW optimizer with a learning rate of 3×10^{-5} for 20k iterations. The models are trained on 32 A100 GPUs with fp16 mixed-precision for around 2 days. During training, the video resolution and number of frames are randomized at multiple scales, making the model robust to various input resolutions and frame lengths. The resolutions include 384×576, 512×512, 1280×1920, and the frame lengths range from 1 to 16. After the full training stages, the models can precisely control six effects (benefited from simulation data), generalize to diverse content (benefited from generation data), and simulate realistic weather (benefited from real-world data), supported by the evaluation in main Sec. 5.

Weather Strength Definition We adopt standard definitions from Unreal Engine, which are grounded in physically meaningful quantities, e.g., cloud coverage (ratio of the sky), fog (density), raindrop or snowflake (count per unit volume per second), ground puddle (coverage ratio), and snow cover (height). During training, their intensity values are normalized to the range $[0, 1]$. This continuous representation enables fine-grained control and smooth transitions

B. Additional Results

In Fig. S5, both our WEATHER SYNTHESIS MODEL and WEATHER REMOVAL MODEL effectively edit the weather, preserve details (e.g., "STOP" on the road), and also maintain temporal consistency. In addition, the different weather conditions can be controlled precisely by changing the strength values of each effect, shown in Fig. S4.

In addition to video editing methods, we also compare the weather synthesis with 3D simulation method in Fig. S1. ClimateNeRF [7] relies on the high-quality geometry to integrate weather effects with the scene successfully and cannot perform well for regions that are not captured densely (e.g., rooftop). On the other hand, our weather synthesis model leverages the video diffusion model and synthesizes snowflakes, snow coverage covering the whole scene. Furthermore, we provide additional qualitative results of weather removal and weather synthesis in Fig. S6, S7, and S8, showing that our method generalize well to diverse video inputs.

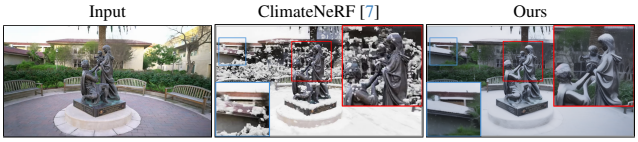


Figure S1. **Comparison with ClimateNeRF [7].** Our video model can coat delicate snow on the statue and rooftop surfaces, and also adjust the shading, which is hard for 3D simulation approaches [7].

User Study is a common approach for assessing perceptual realism. We conducted the user study mentioned in Sec. 5.1 on Amazon Mechanical Turk (MTurk) to compare our method with other baselines. Fig. S2 visualizes the example interface used for user study on the weather synthesis task. We asked users to make perceptual decisions on the pairwise comparison with the following criteria: 1) the integration of weather effects, 2) temporal consistency, and 3) content consistency. For weather removal, we used a similar user interface but asked users to choose videos with the least visible weather effects.

During the user study, we invited 11 users for each sample pair to perform binary preference selection. We used 40 videos for weather synthesis (4 baselines, 3 effects) and 55 for weather removal (6 baselines) evaluation. This results in $3 \times 40 \times 4 \times 11 \times 3 = 15840$ and $55 \times 6 \times 11 \times 3 = 10,890$ user selections for each evaluated task. For each evaluated scene video, we did majority voting from 11 users to determine which method is more preferred in this scene. The majority voting can efficiently filter the effects of random users. The full experiments are repeated 3 times to calculate the mean and standard deviation on the preference percentage.

Inspired by [10], we also used large vision-language models (VLM) as perceptual evaluators to perform similar perceptual preference selections. For each pair of methods to be compared, we randomly selected a frame of the video and fed these frames into VLM, then asked VLM to give a binary preference selection with the same criteria as we used in the human user study. We used Qwen2.5-VL-72B [2] as our local VLM perceptual evaluator. For each sample pair, we run VLM 7 times with different random seeds. The final VLM preference of a scene video is determined by the same majority voting process. Fig. S3 demonstrates two example preference outputs from VLM.

Failure Cases We show failure cases in Fig. S9. High-frequency details such as human faces are sometimes lost. This issue is primarily due to the limited capacity of our base model Stable Video Diffusion [3]. The VAE of Stable Video Diffusion has 8x spatial compression, leading to causes significant degradation and altering of image details. In contrast,

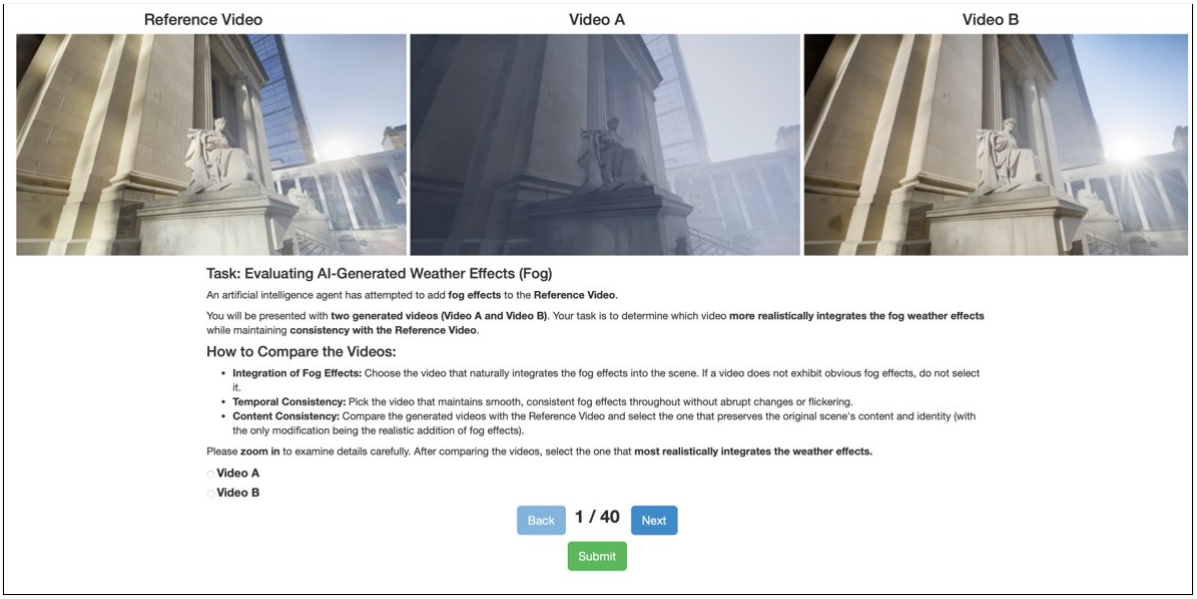
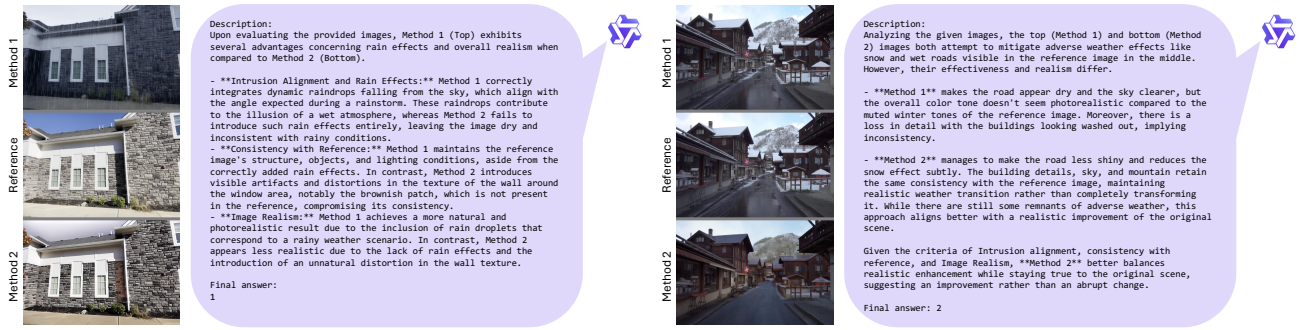


Figure S2. **Example of user study interface for comparing two generated videos for weather synthesis.**



(a) Weather Synthesis (Rain) Example: Ours vs. AnyV2V

(b) Weather Removal Example: HistoFormer vs. Ours

Figure S3. **Examples on perceptual preference evaluation with VLM.** We instructed VLM to first briefly describe the observation, then give the reason why it makes this decision.

recent tokenizers offer significantly improved fidelity [1, 13]. Our results appear to have reached Stable Video Diffusion's quality limit. Upgrading to a more powerful video model could significantly improve the overall quality.

Our data collection includes limited night-time videos, leading to potential imperfect simulation in these scenarios. Future work could improve visual quality by collecting additional specialized data.

References

[1] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025. 2

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1

[3] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 1

[4] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *ICLR*, 2024. 3, 4, 5

[5] Yun Guo et al. From sky to the ground: A large-scale benchmark and simple baseline towards real rain removal. In *ICCV*, 2023. 4

[6] Max Ku, Cong Wei, Weiming Ren, Harry Yang, and Wenhu

Figure S4. **Controlling the strength of weather effects.**

Figure S5. **Temporally-Consistent Synthesis and Removal.** Left: weather synthesis. Right: weather removal.

Chen. Anyv2v: A tuning-free framework for any video-to-video editing tasks. *arXiv preprint arXiv:2403.14468*, 2024. 5

[7] Yuan Li, Zhi-Hao Lin, David Forsyth, Jia-Bin Huang, and Shenlong Wang. Climatenerf: Extreme weather synthesis in neural radiance field. In *ICCV*, 2023. 1

[8] Ozan Özdenizci and Robert Legenstein. Restoring vision in adverse weather conditions with patch-based denoising diffusion models. *IEEE TPAMI*, 2023. 4

[9] Shangquan Sun, Wenqi Ren, Xinwei Gao, Rui Wang, and Xiaochun Cao. Restoring images in adverse weather conditions via histogram transformer. *ECCV*, 2024. 3, 4

[10] Tong Wu, Guandao Yang, Zhibing Li, Kai Zhang, Ziwei Liu, Leonidas Guibas, Dahua Lin, and Gordon Wetzstein. Gpt-4v(ision) is a human-aligned evaluator for text-to-3d generation. In *CVPR*, 2024. 1

[11] Hongtao Wu et al. Rainmamba: Enhanced locality learning with state space models for video deraining. In *ACM MM*, 2024. 4

[12] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Fresco: Spatial-temporal correspondence for zero-shot video translation. In *CVPR*, 2024. 5

[13] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 2

Figure S6. **Additional qualitative results of weather removal.**

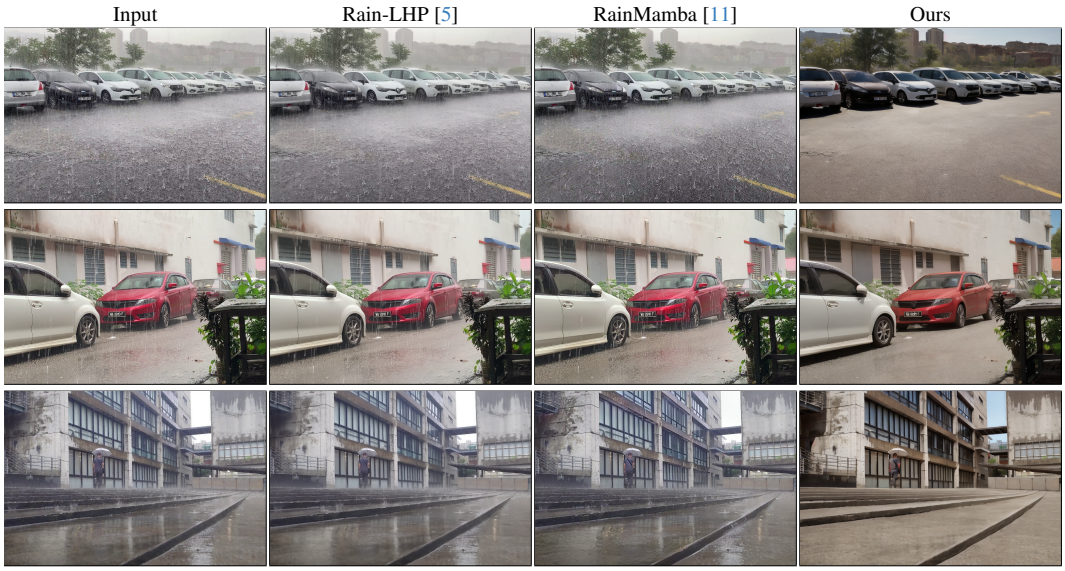


Figure S7. **Additional qualitative results of rain removal.** We compare our rain removal results with recent non-diffusion methods [5, 11].

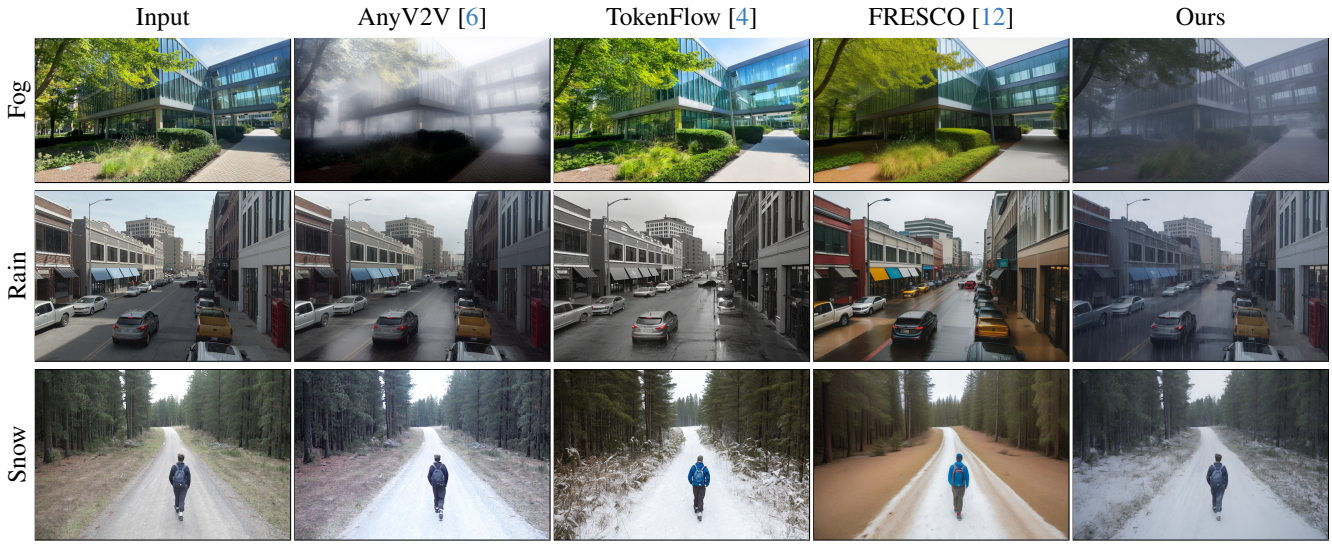


Figure S8. **Additional qualitative results of weather synthesis.**

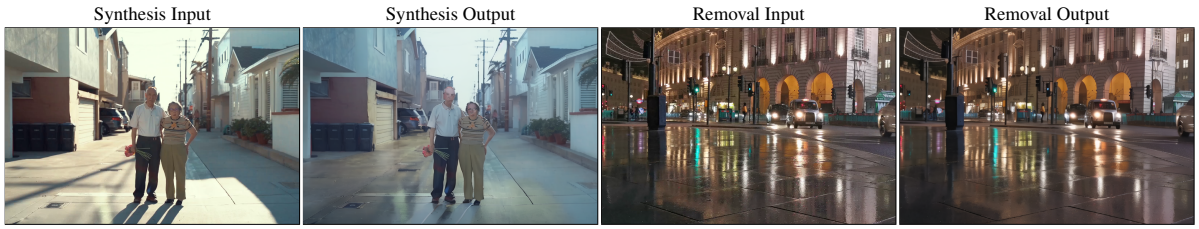


Figure S9. **Limitation.** Our method has a few failure cases, such as human facial details and night videos.