

Zero-Shot Video Deraining with Video Diffusion Models

Supplementary Material

Video results for all corresponding figures are organized in `videos.html`. Benchmark results on NTURain [4], GT-Rain [1] and RealRain13 datasets can be found in `videos/benchmark_videos`. *For the best viewing experience, we strongly recommend opening `videos.html` alongside this document.*

1. Implementation Details

Our backbone model used in all of our experiments is CogVideoX-2b [20], a large-scale text-to-video generation model based on a diffusion transformer architecture. We remark that the existing 2B variant model is restricted to precisely 49 frames at a 480×720 resolution. However, such a limitation has already been addressed in the CogVideoX1.5-5B variant¹ and newer releases of diffusion models are likely to improve further. Due to the large size of the diffusion model, the inference time of our method is approximately 2 minutes and 50 seconds on one NVIDIA A100 GPU, with half of the time being allocated to video inversion and the other half to video reconstruction.

2. Rain prompt analysis

In Sec.3.2 of the main paper, we did an analysis on the rain conditions. Here we present the generated videos corresponding the different rain conditions and the corresponding deraining results in Fig. 9.

We first experiment with a simple prompt “rain”, which is able to remove some rain but generally performs poorly, see Fig. 9 (top right). To better understand the reason and analyze the failure, we generate a sample using the prompt, see Fig. 9 (top left). The prompt shows no visible rain pattern, indicating it has not been able to disentangle the rain concept properly.

Next, we conduct a large-scale analysis from 1K real-world rainy video crops. These videos are captioned using an automatic video captioner [10] and the text-embeddings are extracted using the T5 text-encoder [15]. We then extract the text-embeddings associated with the word “rain”, compute their mean, and use this as the condition for deraining. In Fig. 9 (middle), we show results when prompting the mean text-embedding, computed from the extracted text-embeddings. Such an approach improves over the base prompt “rain” and shows a clear rain footprint when used in generation. However, some rain streaks remain, and a faint background can still be observed in the generated sample.

¹<https://huggingface.co/THUDM/CogVideoX1.5-5B-SAT>

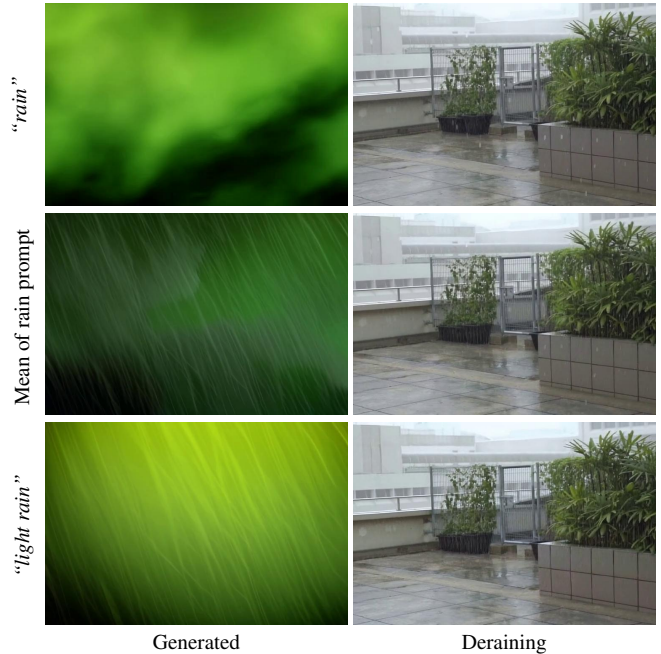


Figure 9. Different rain prompts and their respective results. *Left:* Using diffusion model to generate a video based on the prompt. *Right:* Deraining results with different prompts. The generated video from “rain” shows no rain. When using mean of rain prompts, the generated video shows a clear rain pattern and some background, indicating the prompt is not fully disentangled. “light rain” shows excellent rain-background disentanglement, and overall it performs the best for deraining.

We hypothesize that the prompt is not fully disentangled, causing limited deraining performance.

In analyzing the extracted text-embeddings and their respective prompts further, we observe that context from neighboring words in the text-encoder [15] plays a crucial part in the disentanglement of a concept. In Fig. 9 (bottom), we utilize a simple prompt “light rain”, which is able to disentangle rain from the background in the generation and performs best in deraining.

3. Results on Synthetic Data

We test our method on the synthetic test set of NTURain [4], and compare against supervised methods in Table 1. Note that S2VD [21] and RainMamba [18] were trained on NTURain, while ours is training-free, which explains their unfair higher metrics. Histoformer [17] and Diff-Plugin [12] were trained on different synthetic rain datasets and thus generalize less well to NTURain. Our method, de-

Method	S2VD	RainMamba	Histoformer	Diff-Plugin	Ours
Trained on NTU	✓	✓			
PSNR \uparrow	37.37	37.87	29.96	24.91	27.66
SSIM \uparrow	0.9683	0.9738	0.9112	0.7683	0.8492

Table 1. Quantitative comparisons on NTURain [4] synthetic set.

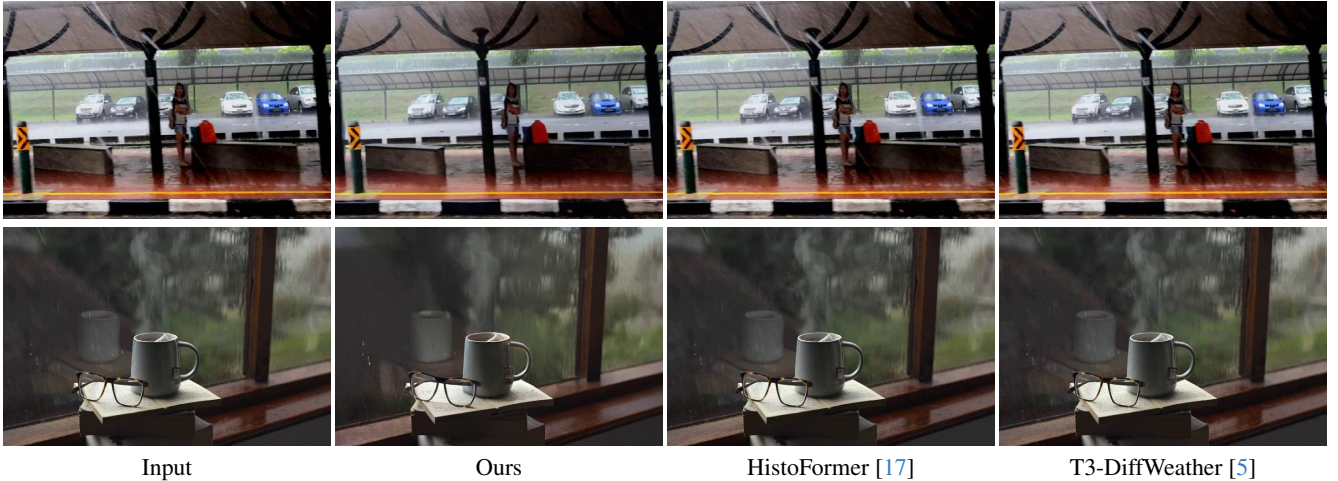


Figure 10. Selected frames from derained real-world videos. For best viewing experience see the supplementary video.

spite training-free, performs comparable to other supervised method under cross dataset validation set up.

4. Additional Results on Deraining

Fig. 10 demonstrates additional results on deraining tasks. Unlike HistoFormer [17] and T3-DiffWeather [5], the proposed method is able to remove rain from both scenes. Fig. 11 shows a qualitative comparison between our method and other baselines on GT-Rain [1]. The supplementary videos illustrate the difference more clearly. Please open [videos.html](#) to view the video results.

5. Additional Results on Attention Switching

In Sec. 3.3 of the main paper, we propose using a subset of blocks \mathcal{B} for attention switching to enhance structural preservation. The selection of \mathcal{B} is based on the statistical analysis of high-frequency information in different blocks. We provide an attention map visualization in Fig. 12, showing the attention maps between the prompt “dog” and the first frame of the generated latent frame. Note that for the first four blocks, the attention is not localized but focuses more on the global features, while for the last fifteen blocks, the features show redundancy in spatial locality. We perform an ablation study on attention switching across the different blocks in Fig. 13. When attention switching is not applied in any of the transformer blocks, the result is

distorted. In utilizing attention switching in both the initial (the first four blocks) and the last fifteen blocks, the optimal result is obtained. Fig. 13 (bottom row) shows that using only the initial or the latter blocks results in less optimal results.

6. Inversion Techniques

To the best of our knowledge, no previous work has attempted to invert a video using video diffusion models, where the video is represented as a block instead of a set of frames. Frame-based inversion [2, 6–8] methods often lack temporal consistency and hence propose extended attention modules between frames, rely on depth maps and structured noise maps. By using a video diffusion model, the complexity in models can be significantly reduced with regard to temporal consistency.

We experimented with SDEdit [13], DDIM inversion [16], Null-text inversion [14] and DDPM inversion [11]. As Null-text inversion requires optimization at every time step, the method becomes impractical due to a long runtime of 50 minutes for a single video. Hence, we have not included it in this comparison. The results are shown in Fig. 14. Both video SDEdit and video DDIM inversion struggle to retain details for full inversion. An inversion that starts from t_s is more practical since these values can be used for modifying the content. At $t_s = 25$, both

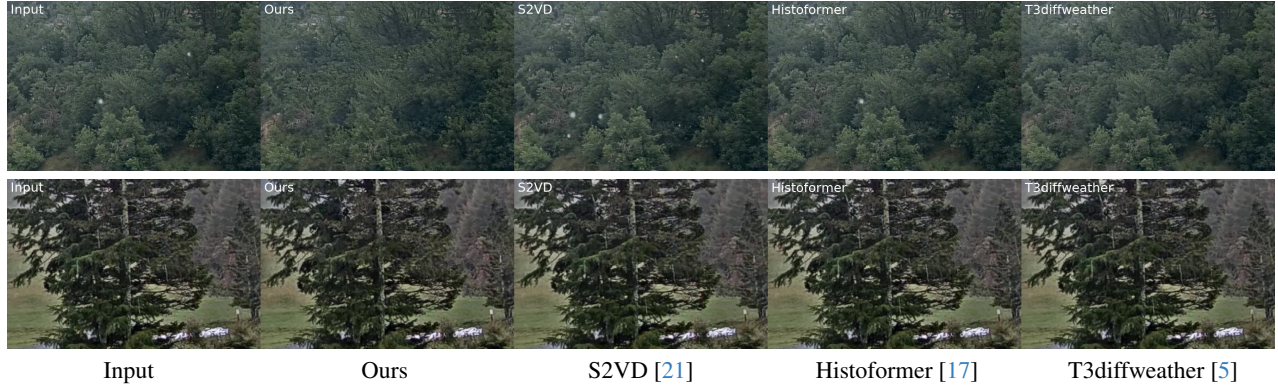


Figure 11. **Qualitative comparison on real rain videos from GT-Rain [1].** Please refer to the supplementary for best viewing experience.

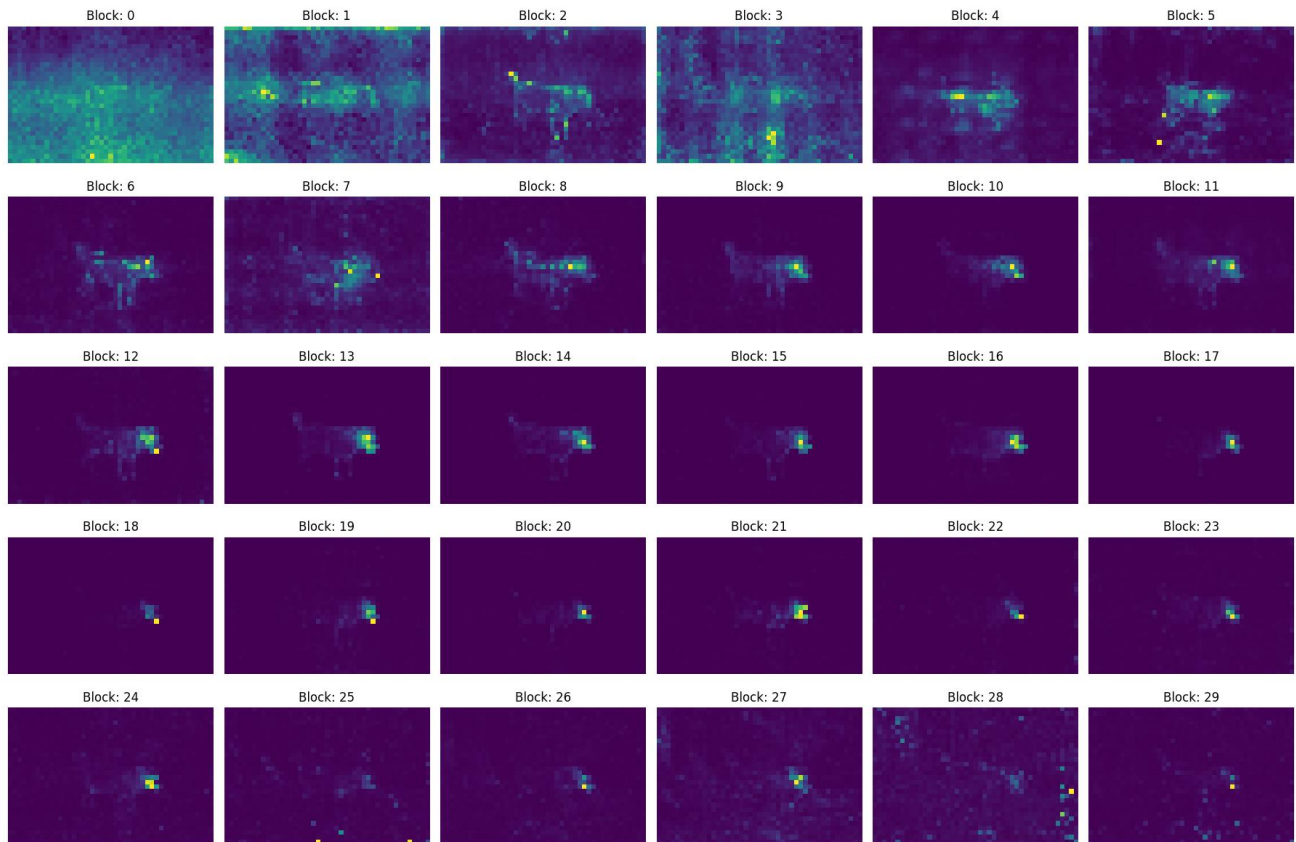


Figure 12. Attention map from the prompt *A dog walking in a rainy forest*. The maps are constructed from the query Q corresponding to the word *dog*, and keys K are from the first latent frame. Note that the first four blocks mainly contain global information, while the last ~ 15 blocks contain mostly redundant spatial information.

approaches can get the global scene structure but are still unable to produce scene details. Video DDPM inversion is capable of reconstructing the scene with fine-grained details even from the initial timestep.

7. Additional Results on Desnowing

Desnowing [3] is the task of removing snow, akin to deraining. The problem has been less studied, likely due to less available data and posing issues less frequently for applications. We experimented with our approach on snow in

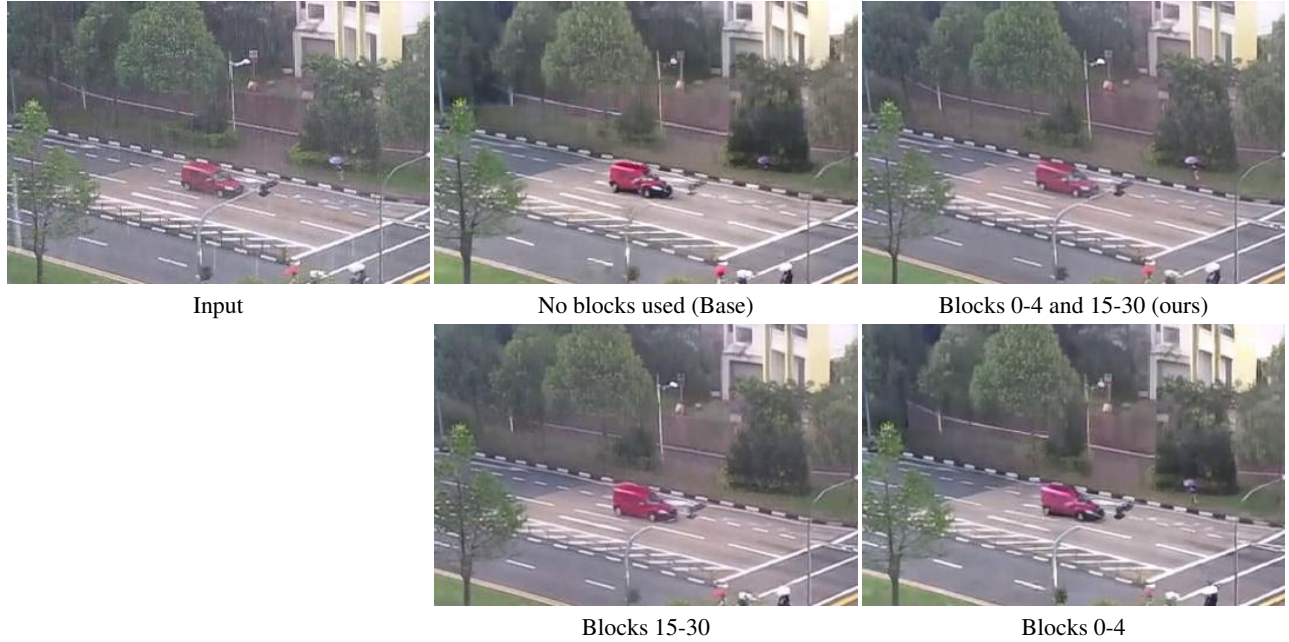


Figure 13. Ablation study on different selection of blocks \mathcal{B} for attention switching. Using both the initial four blocks and the later fifteen blocks for attention switching obtains the best results. This can be observed by analyzing the distortions caused by other settings when compared to the input image.

Fig. 15 with samples collected from the internet and RealSnow85 [19]. The proposed method is able to better remove snow compared to the state-of-the-art method TURTLE [9]. However, compared to rainy cases, the proposed method is less effective at removing all of the snow and sometimes struggles with structural preservation.

After performing a similar analysis to that in Fig. 9, we find that the base model CogVideoX has not properly disentangled the concept of snow. Fig. 16 shows how the snow prompt generates a forest background, whereas the rain prompt in Fig. 5 generates no background. We hypothesize that the forest background is due to the training data, where snowy scenes mainly contain a forest in the background. This property harms the desnowing process, as the score estimate $\hat{\epsilon}_\theta(x_t)$ is not only pushed away from the snowy concept but also the forest concept, leading to worse structure preservation. Using larger video diffusion models in the future would likely disentangle the concepts better, potentially improving different restoration tasks, e.g., desnowing.

References

- [1] Yunhao Ba, Howard Zhang, Ethan Yang, Akira Suzuki, Arnold Pfahnl, Chethan Chinder Chandrappa, Celso M De Melo, Suyu You, Stefano Soatto, Alex Wong, et al. Not just streaks: Towards ground truth for single image deraining. In *European Conference on Computer Vision*, pages 723–740. Springer, 2022. 1, 2, 3
- [2] Feng Chen, Zhen Yang, Bohan Zhuang, and Qi Wu. Streaming video diffusion: Online video editing with diffusion models. *arXiv preprint arXiv:2405.19726*, 2024. 2
- [3] Haoyu Chen, Jingjing Ren, Jinjin Gu, Hongtao Wu, Xuequan Lu, Haoming Cai, and Lei Zhu. Snow removal in video: A new dataset and a novel method. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13165–13176. IEEE, 2023. 3
- [4] Jie Chen, Cheen-Hau Tan, Junhui Hou, Lap-Pui Chau, and He Li. Robust video content alignment and compensation for rain removal in a cnn framework. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6286–6295, 2018. 1, 2
- [5] Sixiang Chen, Tian Ye, Kai Zhang, Zhaohu Xing, Yunlong Lin, and Lei Zhu. Teaching tailored to talent: Adverse weather restoration via prompt pool and depth-anything constraint. In *European Conference on Computer Vision*, pages 95–115. Springer, 2024. 2, 3
- [6] Nathaniel Cohen, Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. Slicedit: Zero-shot video editing with text-to-image diffusion models using spatio-temporal slices. *arXiv preprint arXiv:2405.12211*, 2024. 2
- [7] Ruoyu Feng, Wenming Weng, Yanhui Wang, Yuhui Yuan, Jianmin Bao, Chong Luo, Zhibo Chen, and Baining Guo. Ccredit: Creative and controllable video editing via diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6712–6722, 2024.
- [8] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel.



Figure 14. Comparison of video inversion results at different skip values t_s . Video SDEdit inversion loses the entire scene structure at $t_s = 0$, while video DDIM inversion can retain some structure from the camera motion and the cars. Video DDPM inversion retains the scene with only a minor loss in high-frequency details. Results improve for higher skip values t_s . Note that PSNR is averaged over all video frames.

Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023. 2

- [9] Amirhosein Ghasemabadi, Muhammad Janjua, Mohammad Salameh, and Di Niu. Learning truncated causal history model for video restoration. *Advances in Neural Information Processing Systems*, 37:27584–27615, 2024. 4, 6

- [10] Wenyi Hong, Weihang Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*, 2024. 1

- [11] Inbar Huberman-Spiegelglas, Vladimir Kulikov, and Tomer



Figure 15. Selected frames from desnowed real-world videos. Base refers to the case without attention switching. For the best viewing experience, see the supplementary material.



Figure 16. Visualization of different snow prompts. *Left*: The result generated with the prompt “snow” produces snow on the ground instead of a falling snow effect. *Middle*: In using the prompt “snowing”, the model generates falling snow. *Right*: Unlike the results in Fig. 9 for the prompt *light*, the same prompt affects snow generation differently. Note that the generated prompt is entangled with a snowy forest.

Michaeli. An edit friendly ddpm noise space: Inversion and manipulations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12469–12478, 2024. 2

[12] Yuhao Liu, Zhanghan Ke, Fang Liu, Nanxuan Zhao, and Rynson WH Lau. Diff-plugin: Revitalizing details for diffusion-based low-level tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern*

Recognition, pages 4197–4208, 2024. [1](#)

- [13] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. [2](#)
- [14] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023. [2](#)
- [15] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. [1](#)
- [16] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022. [2](#)
- [17] Shangquan Sun, Wenqi Ren, Xinwei Gao, Rui Wang, and Xiaochun Cao. Restoring images in adverse weather conditions via histogram transformer. In *European Conference on Computer Vision*, pages 111–129. Springer, 2024. [1](#), [2](#), [3](#)
- [18] Hongtao Wu, Yijun Yang, Huihui Xu, Weiming Wang, Jinni Zhou, and Lei Zhu. Rainmamba: Enhanced locality learning with state space models for video deraining. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7881–7890, 2024. [1](#)
- [19] Hongtao Wu, Yijun Yang, Angelica I Aviles-Rivero, Jingjing Ren, Sixiang Chen, Haoyu Chen, and Lei Zhu. Semi-supervised video desnowing network via temporal decoupling experts and distribution-driven contrastive regularization. In *European Conference on Computer Vision*, pages 70–89. Springer, 2025. [4](#)
- [20] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiao-han Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. [1](#)
- [21] Zongsheng Yue, Jianwen Xie, Qian Zhao, and Deyu Meng. Semi-supervised video deraining with dynamical rain generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 642–652, 2021. [1](#), [3](#)