Zero-Shot Video Deraining with Video Diffusion Models

Tuomas Varanka[1*], Juan Luis Gonzalez[2], Hyeongwoo Kim[2], Pablo Garrido[2], Xu Yao[2]

University of Oulu[1]   Flawless AI[2]

{juanluis.gonzalez, pablo.garrido, xu.yao}@flawlessai.com

tuomas.varanka@student.oulu.fi hyeongwoo.kim@imperial.ac.uk

Project Website: tvaranka.github.io/ZSVD

Abstract

Existing video deraining methods are often trained on paired datasets, either synthetic, which limits their ability to generalize to real-world rain, or captured by static cameras, which restricts their effectiveness in dynamic scenes with background and camera motion. Furthermore, recent works in fine-tuning diffusion models have shown promising results, but the fine-tuning tends to weaken the generative prior, limiting generalization to unseen cases. In this paper, we introduce the first zero-shot video deraining method for complex dynamic scenes that does not require synthetic data nor model fine-tuning, by leveraging a pretrained text-to-video diffusion model that demonstrates strong generalization capabilities. By inverting an input video into the latent space of diffusion models, its reconstruction process can be intervened and pushed away from the model's concept of rain using negative prompting. At the core of our approach is an attention switching mechanism that we found is crucial for maintaining dynamic backgrounds as well as structural consistency between the input and the derained video, mitigating artifacts introduced by naive negative prompting. Our approach is validated through extensive experiments on real-world rain datasets, demonstrating substantial improvements over prior methods and showcasing robust generalization without the need for supervised training.

1. Introduction

Rain can severely degrade the quality of captured video scenes due to occluding rain streaks. The degradation not only makes the video content more challenging to view but also adversely affects computer vision tasks, such as tracking and segmentation [12]. Most existing approaches [14, 28, 30, 62, 63, 66, 72] focus on supervised learning, which requires ground truth data of the same sample with and without rain. As collecting such data is extremely difficult, synthetic data has been frequently used. However, synthesizing realistic rain is challenging due to its complex inter-

(a) Input   (b) RainMamba [66]
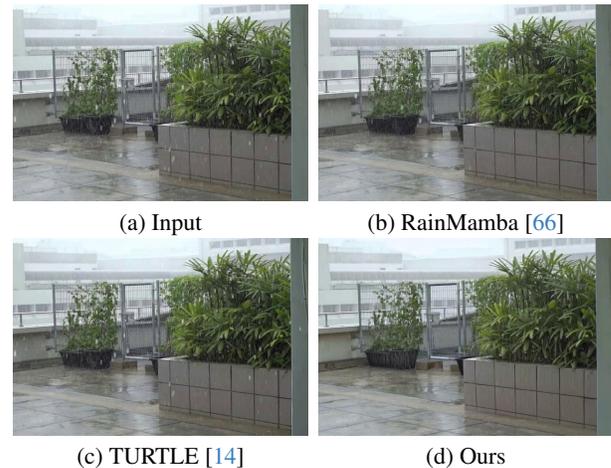
(c) TURTLE [14]   (d) Ours

Figure 1. **Visual results on video deraining on real rain video.** Compared with state-of-the-art methods such as RainMamba [66] and TURTLE [14], our zero-shot solution is more effective in removing rain streaks and generates temporally consistent results. Please refer to the supplementary for the corresponding video.

actions with the environment, depth variations, and translucency. As a result, methods trained on synthetic data often face a significant domain gap when applied to real-world scenarios. Fig. 2 shows examples of synthetic rain (top row) and real rain (middle row). To mitigate or completely avoid this domain gap, semi-supervised and unpaired learning approaches [62–64, 70] have been explored. These approaches, however, frequently fail in reconstructing background details or fully removing rain streaks.

Diffusion models have been used in various restoration tasks [29, 30, 32, 56] due to their powerful priors of various attributes and concepts learned from large-scale text-to-image training. To adapt the pre-trained diffusion models for restoration tasks, they must be modified to accept images as input, which calls for additional modules and training. However, introducing and training additional modules

---

*Work done during an internship at Flawless AI

(a) Synthetic examples from NTURain [6].

(b) Real-world examples.

(c) Generated examples using CogVideoX [69].

Figure 2. **Example cases of rain.** Notably, synthetic rain often exhibits unrealistic repetitive patterns not present in real-world scenes while lacking interactions with the scene objects and depth. Rainy scenes generated by large diffusion models show more realistic effects. Refer to the supplementary for the respective videos.

alters the original data flow, which can degrade the robust priors already embedded in the pre-trained models. Moreover, this process relies on synthetically generated data for fine-tuning, which creates a mismatch with real-world data, diminishing the effectiveness of the generative prior. And although rain exhibits distinct temporal patterns that could be leveraged, most deraining approaches [26, 30, 62, 63, 70] focus solely on the image domain. Real-world rain streaks, being transparent and often small, can easily blend into the background, making it highly ambiguous to discern them from stationary images alone. Another limitation of image-based diffusion models is that they do not ensure temporal stability and background structure consistency.

To combat the challenges of real-world data and the temporal aspect, we propose a *zero-shot* paradigm for restoration that only leverages a large-scale video diffusion model, eliminating the need for additional training or fine-tuning. Unlike previous U-Net-based methods that use Stable Diffusion [65] as their backbone, we opt for a transformer-based video generative model [69], which achieves superior video generation quality with high-fidelity motion. To fully exploit the capability of video diffusion models, we utilize inversion [47] to avoid training of new modules. Inversion finds the latent x_T from which the diffusion model gener-

ated a sample, allowing for the reconstruction of the sample x_0 from this latent x_T. By intervening in the reconstruction process [16, 35, 37], modifying the condition and applying classifier-free guidance [17], we can alter and guide the desired outcome. However, simply removing the rain-related condition during reconstruction is a naive approach and does not effectively remove the rain.

To push the reconstruction path away from a rainy video, we explicitly compute the difference between a reconstruction step with a null condition and a rain condition. This difference is then amplified and added to the reconstructed latent. However, amplifying differences may introduce artifacts, such as background distortions. To address this, we introduce a novel *attention-switching* technique that leverages the KV matrices associated with the null text features in specific transformer blocks. Prior works [5, 13, 16] show that cross- and self-attention layers of Stable Diffusion [42] encode rich semantic information. However, manipulating attention in diffusion transformer (DiT) based models [38] is challenging, as the Q, K, and V matrices are computed from a concatenation of vision and text tokens, resulting in joint-attention and entangled features. This differs from U-Net-based models [42, 65], which allow explicit access to cross-attention between text and image. So far, only a few works [22, 54] have explored the attention layers in DiT-based models. To handle the entangled features from joint attention, we employ a cross-condition hidden feature that combines text and video features from different conditions. This ensures alignment between the features when modifying the attention process. By manipulating the attention process from a subset of blocks we balance background preservation with deraining ability.

We conduct extensive evaluation experiments across multiple real-world rain datasets. Our approach surpasses state-of-the-art video deraining baselines in both quantitative metrics and qualitative results, achieving superior rain removal and temporal consistency. The contributions of our work can be summarized as follows:

- We propose a new *training-free* diffusion paradigm for restoration tasks that circumvents the difficulties in generating realistic synthetic data. To the best of our knowledge, our approach is the first zero-shot method for video deraining.
- We introduce an attention-switching mechanism for transformer-based video diffusion models, designed to preserve scene content and improve video fidelity.
- We analyze the effectiveness of different prompts and find that disentanglement via negative prompts plays a key role in achieving high-quality video deraining.

2. Related Work

Rain Removal   Initial approaches attempt to impose handcrafted priors for deraining [15, 24, 27, 31, 60], but of-

ten fail in more complex scenes. Paired learning approaches use synthetic rain data with ground truth [7, 8, 30, 49, 50, 66, 67]. For video deraining, optical flow has been utilized to capture temporal correlations [52, 67, 68]. Another line of work [10, 43, 59, 74] leverages event cameras to capture high temporal resolution information.

Mamba-based models have been explored in [43, 51, 66] for capturing temporal cues of rain. Diff-Plugin [30] utilizes a pre-trained diffusion model with task-specific modules. TURTLE [14] utilizes a causal history model to retain information from previous frames for various video restoration tasks, including de-raining and de-snowing. A depth prior and a pool of learnable prompts are used by [7] for adverse weather restoration problems. Unpaired approaches use a set of rainy and non-rainy samples [62–64, 70]. DerainCycleGAN [62] modifies the CycleGAN [75] for rain removal by using an unsupervised attention-guided rain streak extractor. NLCL [70] employs contrastive learning to perform deraining with non-local patches. NSB [63] presents a neural Schrödinger bridge [26] for deraining. More recently, WeatherWeaver [28] also leverages a video diffusion model for weather synthesis and removal by training SVD [2] on a curated dataset containing both synthetic and real data. In contrast to methods that rely on synthetic training data with limited generalization to real-world or learning a mapping from scratch via unpaired data, we adopt a *training-free* approach that leverages a large pretrained diffusion prior.

Diffusion-based restoration   Several approaches [4, 29, 30, 32, 61] use the strong generative prior from diffusion models to improve results for restoration. DiffBIR [29] adds a ControlNet [73] to inject image features into a pre-trained T2I diffusion model. Diff-Plugin [30] trains small modules to extract degradation-specific features. DA-CLIP [32] trains a module that optimizes the CLIP [40] image embedding, which is integrated into a diffusion model with cross-attention blocks. To use diffusion models, these methods add extra branches that encode image information, requiring model re-training, which results in deteriorating the existing model bias due to the change of input data. To avoid this, zero-shot approaches bypass training altogether. For instance, DDNM [61] uses the null-space to perform zero-shot diffusion restoration. We extend the zero-shot capabilities of diffusion models by improving structural preservation, including text as a prior, and extending the approach to video.

Diffusion-based editing   Recent advances in diffusion models [18, 48] have shown exceptional progress in text-to-image [39, 42, 44] and text-to-video generation [2, 19, 69]. Previous studies have explored the editing capacity of the generative diffusion priors [16, 35, 73]. Many of them adopt classifier-free guidance [17], a simple technique to boost the sampling quality. Woolf *et al.* [65] further discover that the generative process could be guided better with nega-

tive prompts that the model should exclude. Inspired by that, we adopt a negative prompt-based editing approach for rain removal. To edit real scenes within diffusion models, there exist several inversion methods to invert images into the latent space of diffusion models [23, 37, 46]. The edit-friendly DDPM inversion [23] achieves state-of-the-art results in terms of reconstruction and editing capacity. As such, we adopt this approach for video inversion.

3. Method

In this section, we present our method for zero-shot video deraining. Rather than training with synthetic data that generalizes poorly to real-world data, we leverage the powerful generative prior of diffusion models. Our method begins by inverting a given video into the diffusion latent space using an inversion technique. While some earlier inversion methods [37, 47] produce semantic or structural inconsistencies during inversion, recent advancements like DDPM inversion [23] can invert images with faithful semantics preservation, making zero-shot restoration possible. By intervening in the reconstruction process, it is then possible to modify the video content, such as degradation effects (e.g., rain), while preserving the underlying structural information and semantics. More specifically, we distill the model's existing knowledge of degradation effect to remove it from the reconstruction process. To further preserve structural information, we propose an *attention-switching* technique that leverages structural information from the KV matrices associated with the joint-attention. The overall pipeline of our proposed method is shown in Fig. 3.

3.1. Preliminary

Video Diffusion   Text-to-Video diffusion models [2, 19, 69] use text input to generate videos from Gaussian noise. Earlier works [2] adapt text-to-image diffusion models by injecting temporal blocks and fine-tuning on video datasets. These approaches, however, often struggle with temporal consistency and larger motion generation due to the use of 2D VAEs and short-range temporal attention.

CogVideoX [69] addresses these issues by using a 3D VAE to compress video data, avoiding temporal inconsistency of individual frame encoding.

The learning objective remains the same as conditional diffusion [65]:

$$L = \mathbb{E}_{x_0, \epsilon \sim \mathcal{N}(0,1), t, c} \left(||\epsilon - \epsilon_\theta(x_t, t, c)||_2^2 \right), \quad (1)$$

where t is a timestep, c is the conditional input (text), ϵ_θ is the diffusion model, and x_t represents a noisy latent video. To better align the text and image features, the MM-DiT (multimodal diffusion transformer [9]) blocks are used. The hidden features $h = h_{text}|h_{img}$ consist of concatenated text and image features, and the attention is performed simul-
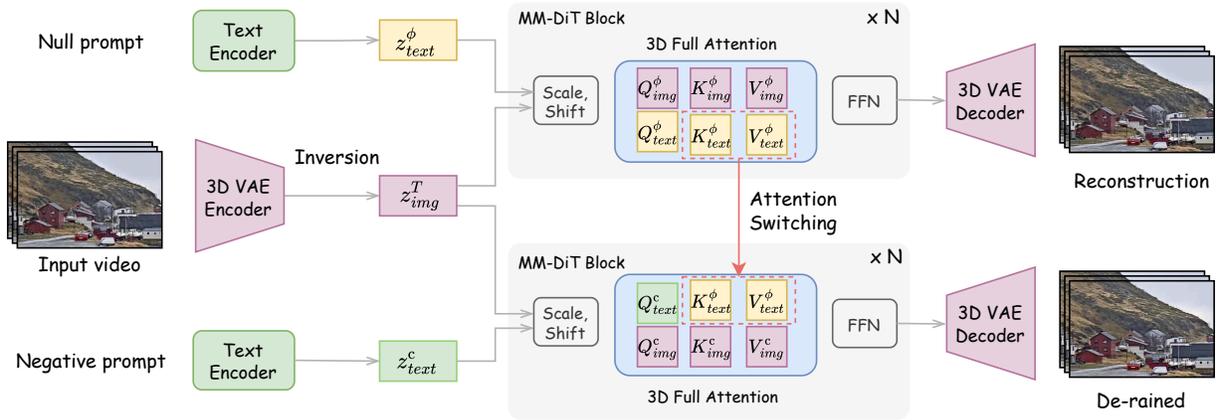
Figure 3. **Architecture of the proposed zero-shot video deraining approach.** First, video inversion is performed to extract the noise latent z_{img}^T. Next, starting from timestep t_s the model performs a reconstruction step with the null prompt and a rain condition step with the negative prompt. The two paths are then combined following Eq. (4). Attention switching is applied for blocks $\mathcal{B}$, where the $K_{text}^{\emptyset}$ and $V_{text}^{\emptyset}$ are extracted from the null condition and are used to replace their conditional equivalents K_{text}^c and V_{text}^c.

Figure 4. **Comparison between DDIM [47] and DDPM [23] inversion on video data.** Video DDIM inversion struggles with fully reconstructing the video and misses not only high-frequency details but also larger objects. The PSNR drop in video DDPM inversion is mostly caused by the VideoVAE (PSNR = 31.80) and numerical precision.

taneously across both modalities, for better alignment between the modalities.

Diffusion Image Inversion Inversion refers to find the latent noise x_T that generates a given sample x_0, enabling image editing by manipulating the reconstruction process. SDEdit [35] uses the simplest form of inversion by simply adding Gaussian noise, which limits the preservation of structural information. DDIM [47] further adopts a sampling process that is deterministic, compared to the stochastic DDPM [18]. The DDIM process can hence be inverted by re-arranging the sampling process with respect to the current step x_t and computing the score estimate $\hat{\epsilon}_\theta$ from x_{t-1} instead of x_t. Several approaches have been proposed to improve diffusion inversion [11, 20, 23, 36, 37, 57], with some of them learning additional information like text embeddings [37] or added noise [23].

Diffusion Video Inversion As no prior work exists on inverting latent videos, we experiment with several options

[23, 35, 37, 47]. See Fig. 4 and the supplementary material for results. Based on these results we utilize DDPM inversion [23], which first gathers the noisy latents x_t from

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \tilde{\epsilon}_t, \qquad (2)$$

associated with each timestep and then computes the noise required for cleaning x_t from the previous x_{t-1} from

$$z_t = \frac{x_{t-1} - \mu_t(x_t, c)}{\sigma_t}. \qquad (3)$$

Each of the z_t contains the information required for reconstructing the video at timestep t. The video can then be reconstructed from the set $x_t, z_t, \ldots, z_1$ by starting from x_t, predicting x_{t-1} and adding the extracted noise map z_t progressively, until reaching x_0.

3.2. Training-free Rain Removal

Instead of fully reconstructing the video from the inverted latent x_T, it is possible to modify the outcome by intersecting the reconstruction process. Here, we aim to modify the video by removing any unwanted degradation effect, i.e., rain, while retaining the background information. A naive approach would simply perform inversion with a rain condition and then remove that condition in reconstruction. However, this does not significantly alter the diffusion path and results in mostly reconstructing the original sample. We find that for sufficient results, the path needs to be explicitly separated from the reconstruction path. We achieve the above via negative prompt editing, akin to [36].

Pushing the score estimate $\hat{\epsilon}_\theta(x_t)$ away from the degradation effect is done by using an additional term $\epsilon_\theta(x_t) - \epsilon_\theta(x_t, c)$, where c denotes the degradation effect, i.e., rain.

Amplifying the difference term with a scalar λ and adding the difference to the original score estimate allows us to separate the reconstruction from the degradation effect. This corresponds to the classifier-free guidance term [17] used to guide image generation. Mathematically, the above is performed as follows:

$$\hat{\epsilon}_\theta(x_t) = \epsilon_\theta(x_t) + \lambda(\epsilon_\theta(x_t) - \epsilon_\theta(x_t, c)). \qquad (4)$$

Note that we can target different levels of rain by varying the trade-off factor λ. To avoid deviating too much from the reconstruction path, which would result in lower fidelity, a skipping timestep t_s is used. For initial steps, the sample fully follows the reconstruction path; it deviates from it only after t_s.

Which rain conditions are effective? The condition c in Eq. (4) is a textual prompt, consistent with the use of T2V models. It should only be related to rain to prevent the model from modifying other parts of the sample. We find that the model's ability to disentangle a concept plays a crucial role in the deraining process.

We first experiment with a simple prompt, *"rain"*, which removes some rain but generally performs poorly. To better understand the reason, we generate a sample using the prompt *"rain"* but did not observe clear rain pattern, indicating it has not been able to disentangle the rain concept.

Next, we conduct a large-scale analysis from $1K$ real-world rainy video crops. These videos are captioned using a video captioner [21] and the text-embeddings are extracted using the T5 text-encoder [41]. We then extract the text-embeddings associated with the word *"rain"*, compute their mean, and use this as the condition for deraining. When prompting the mean text-embedding, we observe improved performance over the base prompt *rain"*, with a clearer rain footprint in the generated samples. However, some rain streaks and faint background artifacts remain, suggesting that the prompt is not fully disentangled.

By analyzing the text-embeddings and their respective prompts further, we observe that context from neighboring words in the text-encoder [41] plays a crucial part in the disentanglement of a concept. For instance, a simple prompt like *"light rain"* demonstrates better disentanglement of rain from the background and performs better in deraining.

3.3. Attention Switching in DiTs

When analyzing the cross-attention maps of the diffusion model, we empirically observe that not all blocks contribute to high-frequency information. To determine whether a block is responsible for low- or high-frequency information, we conduct a study by skipping blocks one-by-one. For each block, we replace the conditional features with null features during generation. We then compare the result to the case with all blocks active in Fig. 5. This experiment is
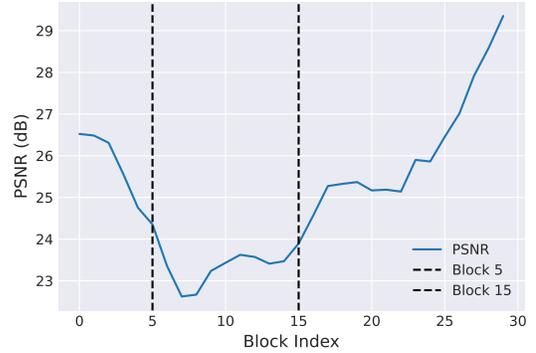

Figure 5. **Impact of skipping an individual block on the PSNR.** Losing high-frequency information, which is captured by blocks 0-5 and blocks 15-30, has a lower impact on PSNR.

performed using 10 prompts, and we compute the average PSNR impact of each block by averaging its impact across these prompts. It can be observed that the high PSNR values are mainly in the initial blocks 0-5 and later blocks 15-30, indicating they contain high-frequency information.

Attention Switching We refer to the selected blocks containing high-frequency information as $\mathcal{B}$. A similar observation on the impacts of different blocks in DiTs is also found in [33]. By using another observation made with image diffusion models, where the key K and value V matrices mostly contain the structure of the image [5], we are able to perform stronger deraining with a larger λ on the blocks that mostly contain high-frequency information, while simultaneously retaining video fidelity by limiting the impact from blocks $\mathcal{B}$.

We modify the attention block

$$Attn_b = \begin{cases} Attn(Q^c, K_{text}^{\phi}|K_{img}^c, V_{text}^{\phi}|V_{img}^c) & \text{if } b \in \mathcal{B} \\ Attn(Q^c, K^c, V^c) & \text{else,} \end{cases}$$
$$(5)$$

by replacing the K_{text}^c and V_{text}^c with their respective ones from the null condition. By using K_{text}^{ϕ} and V_{text}^{ϕ} the structural information from the reconstruction latent can be used to better retain content in the video. As only the blocks $\mathcal{B}$ are modified that do not contain high-frequency information, the model's deraining ability is not affected.

However, the use of MM-DiT [9] blocks in CogVideoX [69] presents challenges that restrict the use of Eq. (5). These issues stem from the joint-attention that limits passing text-embeddings only to the first transformer block instead of all individual blocks as with previous UNet architectures like Stable Diffusion [65]. The conditional text-embeddings serve as input only to the first block and the remaining blocks receive processed hidden states from the previous blocks. Due to the joint-attention the text features are updated and influenced by the visual features, ef-

fectively entangling them in the process as information is stored in the text features [55].

Cross-condition hidden features   When applying Eq. (5) directly on MM-DiT blocks without any modifications the output will result in a significant number of artifacts as the newly switched null conditioned text features $h_{text}^{\emptyset}$ do not align with the visual features h_{img}^c. This is because the image and text features get entangled during the denoising process as hidden hidden features are passed from one block to the next one. Instead of using the null hidden feature $h^{\emptyset} = h_{text}^{\emptyset} | h_{img}^{\emptyset}$, the cross-condition hidden feature

$$h^{\emptyset,c} = h_{text}^{\emptyset} | h_{img}^c, \qquad (6)$$

is used to obtain the attention inputs $K = P_K(h^{\emptyset,c})$ and $V = P_V(h^{\emptyset,c})$ to be used in place of the null text inputs in Eq. (5). This way, the null conditioned matrices $(KV)^{\emptyset}$ for the text will be aligned with the prompt conditioned $(KV)^c$ visual features.

Split attention for retaining conditional text features Since the attention switching is only performed for blocks $\mathcal{B}$ this causes a problem due to the forward processing of hidden features. If the key K and value V matrices are changed from conditional to null for the first block following Eq. (5), it will remain changed for the remaining blocks as the text features are obtained from the previous block. However, we only want this to be the case for blocks $\mathcal{B}$, not all of them. To avoid this, we first compute the hidden features associated with the conditional prompt and discard the visual features using standard attention.

$$
\begin{aligned}
h_{text,b}^c &= block_i(h_{text,b-1}^c, h_{img,b-1}^c) \\
h_{img,b}^c &= block_i(h_{text,b-1}^{\emptyset,c}, h_{img,b-1}^c)
\end{aligned} \Big|_{, b \in \mathcal{B}_{\text{initial}}} \quad (7)
$$

Then, we compute the visual features using the attention switching with the cross-condition hidden features. This process ensures that not only are all the features aligned but that the conditional text features are retained through all blocks. This attention split is only necessary for the early blocks $\mathcal{B}_{\text{initial}}$ to ensure that the conditional text features h_{text}^c can be used in the later blocks.

4. Experiments and Results

Experimental settings   In our experiments, We employ the 2B-variant of CogVideoX [69]. Regarding the hyperparameter λ, we observe that the optimal value varies depending on the rain intensity. For simplicity, we set $\lambda = 15$, which performs well for most cases. For the skipping timestep t_s, we use $t_s = 40$, although this can be further optimized for individual videos. For the blocks $\mathcal{B}$ used for attention switching, we select the initial blocks $[0 - 4]$ and later blocks $[15 - 29]$, based on our block analysis of high frequency information. All the input videos are cropped and

resized to a resolution of 720×480. The number of inference steps is set up to 100, and the entire deraining process takes approximately 3 minutes on one NVIDIA H100 GPU.

Datasets.   Most existing rain datasets are synthetic, as capturing real-world rain data with corresponding deraining ground truth is challenging. NTURain [6] includes a real-world rain testing dataset without ground truth. GT-Rain [1] is a large-scale dataset of real-world rainy video and clean image pairs captured using webcams. However, many low-resolution videos exhibit significant compression artifacts. To address this, we exclude all videos with resolutions below 450×300. To address the need for high-quality real-world rain videos with dynamic scenes, we collected an additional dataset of 13 videos, denoted as *RealRain13*. For experimental validation, we utilize the real-world testing data from NTURain [6], the filtered subset of GT-Rain [1], and our collected *RealRain13* dataset. This combination ensures a robust evaluation across diverse real-world scenarios.

Evaluation Metrics.   Evaluating our approach and baselines on real-world rain videos presents some challenges as no ground truth data is generally available. To evaluate the effectiveness of rain removal and the quality of the video, we utilize three metrics adapted to our particular problem: *MUSIQ* [25], *CLIP-IQA* [58], and *Warp Error*. For video quality assessment, we use a non-reference-based metric, MUSIQ [25]. As MUSIQ is a per-frame metric, we average out results across video frames to obtain the final score. For measuring rain removal, we utilize CLIP-IQA, which computes the cosine similarity to text prompts. We modify the metric for rain removal by using a matching prompt. Warp error measures temporal consistency by the difference between a frame and its previous frame warped using the estimated optical flow [53]. Inconsistent rain removal often leads to flickering in the output video, resulting in higher warp error.

Comparison baselines.   We compare our method with several state-of-the-art methods for deraining, including RainMamba [66], TURTLE [14], Diff-Plugin [30], T3-DiffWeather [7], S2VD [71] and HistoFormer [50]. For all the methods, we use their official implementations.

4.1. Qualitative Results

Fig. 6 presents qualitative results on RealRain13 datasets. From Fig. 6, we observe that TURTLE [14] often fails to remove rain and can introduce artifacts (middle row). Diff-Plugin [30] demonstrates improved rain removal but struggles with retaining details and temporal consistency, which is likely caused by randomness in the diffusion and applying it frame-by-frame. Furthermore, some of the perceived rain removal results from the loss of high-frequency details rather than effective deraining. While RainMamba [66] can remove larger rain streaks, it often fails to remove smaller

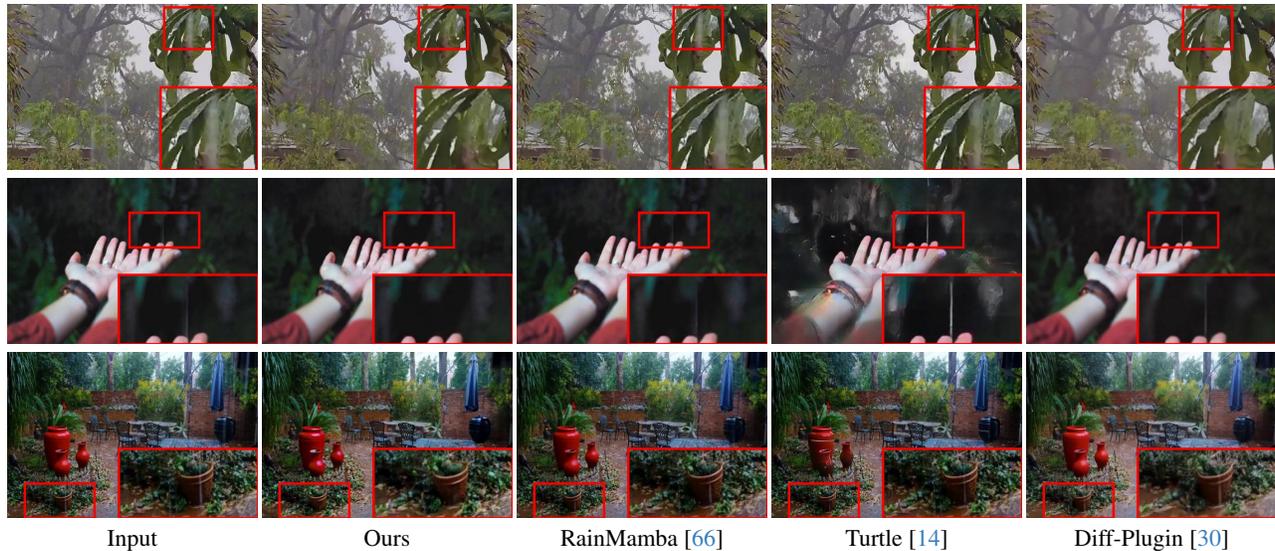|				
Input	Ours	RainMamba [66]	Turtle [14]	Diff-Plugin [30]

Figure 6. **Qualitative comparison of deraining on real-world videos using different approaches.** Our method achieves effective rain removal, faithfully preserving the original content and temporal consistency. Please refer to the supplementary for best viewing experience.
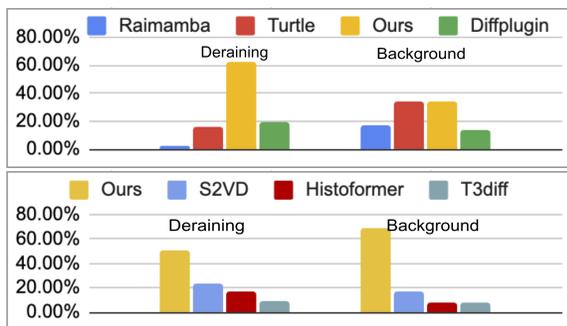

Figure 7. **User study.** Top: Comparison with RainMamba [66], Turtle [14], Diff-Plugin [30] and ours. Bottom: Comparison with S2VD [71], Histoformer [50], T3DiffWeather [7] and ours.

streaks further in depth. Our proposed method is able to effectively remove rain from all scenes with good temporal consistency and fidelity to the original video.

We conducted a user study to assess the perceptual quality of the deraining results, shown in Fig. 7. The study was conducted with 24 participants, who were asked to select their preferences among four different methods based on two criteria: deraining quality and background preservation. The first set compared our method with Rain-Mamba [66], Turtle [14] and Diff-Plugin [30], while the second set compared our method with S2VD [71], Histoformer [50], and T3diff [7]. Each participant evaluated both sets, covering 34 comparisons, totaling 136 videos and 1,632 answers. Our method was the most preferred for deraining quality across all comparisons, and for background preservation in the second set.

4.2. Quantitative Results

Tab. 1 presents the quantitative comparison results on NTU-Rain [6], RealRain13 and GT-Rain [1]. Among the evaluated methods, our approach achieves the best result for CLIP-IQA across all datasets. While other methods perform significantly worse, Diff-Plugin attains comparable results. Notably, the input video achieves the best MUSIQ score, likely because MUSIQ is trained on multi-scale image datasets that typically contain no rain as a degradation effect. This metric highlights a key limitation of Diff-Plugin: it frequently alters the background and introduces artifacts. Furthermore, the Warp Error highlights temporal inconsistencies in many compared methods, with our method achieving the lowest error across all datasets, demonstrating better temporal consistency.

4.3. Ablation Study

We ablate the core components of the approach in Tab. 2, including video inversion, rain condition and the use of attention-switching technique.

Video Inversion   We compare the used video DDPM inversion with video DDIM inversion. Since the DDIM inversion performs poorly in editing, set $t_s = 70$ to avoid large deviations from the scene. From Tab. 2 (top) we can see that The DDIM inversion approach obtains a reasonable performance in CLIP-IQA, but obtains poor result in MUSIQ, indicating poor reconstruction. The deraining ability is weakened by the high t_s, and despite it the method performs reconstruction poorly. Compared to image DDIM inversion, the video latent is significantly more complex and has more potential paths. Therefore, a more precise inversion such as

Method	Zero-shot	NTURain			RealRain13			GT-Rain		
		CLIP-IQA↓	Warp↓	MUSIQ↑	CLIP-IQA↓	Warp↓	MUSIQ↑	CLIP-IQA↓	Warp↓	MUSIQ↑
Input		0.879	0.096	<u>59.17</u>	0.683	0.035	**60.12**	0.133	0.0109	51.14
HistoFormer [50]		0.742	0.059	57.06	0.632	0.023	58.54	0.133	0.0082	51.37
T3-DiffWeather [7]		0.659	0.057	57.92	0.574	0.023	58.50	0.124	0.0094	<u>51.58</u>
S2VD [71]		0.700	0.058	57.06	0.612	<u>0.020</u>	58.31	0.129	<u>0.0080</u>	51.21
Diff-Plugin [30]		<u>0.430</u>	0.063	51.47	<u>0.467</u>	0.026	55.85	0.134	0.0113	48.59
RainMamba [66]		0.734	<u>0.057</u>	56.84	0.619	0.022	58.57	0.170	0.0097	**52.88**
Turtle [14]		0.829	0.064	**59.82**	0.595	0.027	<u>59.94</u>	<u>0.116</u>	0.0098	51.71
Ours	✓	**0.324**	**0.047**	55.43	**0.447**	**0.015**	57.00	**0.075**	**0.0064**	50.18

Table 1. **Quantitative comparisons with state-of-the-art video deraining methods** on NTURain [6], RealRain13, and GT-Rain [1].

Ours		NTURain		RealRain13	
		CLIP-IQA ↓	MUSIQ ↑	CLIP-IQA ↓	MUSIQ ↑
Input		0.842	**59.17**	0.652	**60.12**
DDIM inv. [47]		0.528	52.67	0.555	57.12
DDPM inv. [23]	✓	0.324	55.43	0.447	57.00
Implicit prompt		0.787	54.16	0.700	58.75
"Rain"		0.686	53.59	0.668	57.83
Mean prompts		0.525	55.13	0.601	58.74
"Light rain"	✓	0.324	55.43	0.447	57.00
w/o attn-switch		0.437	54.18	0.457	57.14
w/ attn-switch	✓	**0.324**	55.43	**0.447**	57.00

Table 2. **Ablation study** on NTURain [6] and RealRain13.

| Input | Base | Attention switching |

Figure 8. **Ablation study of attention switching.** Although both cases can remove the rain well, the base model struggles with artifacts, such as missing objects, color shifts, and saturation.

the video DDPM inversion is required.

Prompts   We evaluate the method with four different prompts. The implicit prompt refers to first inverting the video using *"rain"* and then using an empty prompt during deraining. *"Rain"*, mean of rain prompts and *"light rain"* correspond to the ones used in Sec. 3.2. We can observe the CLIP-IQA results in the worst performance with the implicit prompt out of the four prompts. As expected, an explicit push from the rainy concept is required, which can be seen from the improved CLIP-IQA, while MUSIQ remains relatively stable with minor drop for *"rain"* due to the poor disentanglement.

Attention Switching   When switching the KV matrices in the attention blocks a larger $\lambda = 25$ can be used which results in better deraining, which can be observed from the CLIP-IQA. Simultaneously, the fidelity of the original video is preserved, as observed from the MUSIQ score for NTURain and the visual comparison in Fig. 8.

4.4. Discussion and Limitations

The proposed method relies on the performance of existing video diffusion models. The used CogVideoX [69] is the first open-sourced video diffusion model to use latent videos, instead of latent frames, significantly contributing to the temporal consistency. With the rapid advancement of video diffusion [2, 3, 34, 45, 69], the performance of the proposed approach is expected to improve further. The method struggles with heavy rain, where the rain is constant, as it may be misinterpreted as part of the scene. In this work we focus on deraining, but show in the supplementary that the approach can be generalized to other restoration tasks such as snowing, however, currently it is limited by the performance of the underlying video diffusion model.

5. Conclusion

We present a new training-free paradigm for video restoration, focusing on rain degradations, which are hard to synthesize. To circumvent the use of synthetic rain data, which fails to generalize to real-world cases, we leverage a large generative prior from diffusion models. We avoid costly training of additional modules for specific input samples via inversion. We analyze what makes a prompt effective and propose a method for improving sample fidelity by smartly steering the attention process. By utilizing attention features from the reconstruction process and integrating them with attention switching the method improves background preservation. We showcase the method's ability to remove rain in challenging real-world rain scenarios. This work lays the groundwork for addressing different video restoration problems by leveraging large video diffusion models.

References

[1] Yunhao Ba, Howard Zhang, Ethan Yang, Akira Suzuki, Arnold Pfahnl, Chethan Chinder Chandrappa, Celso M De Melo, Suya 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. 6, 7, 8

[2] 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. 3, 8

[3] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. 8

[4] Cong Cao, Huanjing Yue, Xin Liu, and Jingyu Yang. Zero-shot video restoration and enhancement using pre-trained image diffusion model. *arXiv preprint arXiv:2407.01960*, 2024. 3

[5] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22560–22570, 2023. 2, 5

[6] 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. 2, 6, 7, 8

[7] 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. 3, 6, 7, 8

[8] Xiang Chen, Jinshan Pan, and Jiangxin Dong. Bidirectional multi-scale implicit neural representations for image deraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25627–25636, 2024. 3

[9] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis, march 2024. *URL http://arxiv. org/abs/2403.03206*. 3, 5

[10] Xueyang Fu, Chengzhi Cao, Senyan Xu, Fanrui Zhang, Kunyu Wang, and Zheng-Jun Zha. Event-driven heterogeneous network for video deraining. *International Journal of Computer Vision*, 132(12):5841–5861, 2024. 3

[11] Daniel Garibi, Or Patashnik, Andrey Voynov, Hadar Averbuch-Elor, and Daniel Cohen-Or. Renoise: Real image inversion through iterative noising. *arXiv preprint arXiv:2403.14602*, 2024. 4

[12] Blake Gella, Howard Zhang, Rishi Upadhyay, Tiffany Chang, Matthew Waliman, Yunhao Ba, Alex Wong, and Achuta Kadambi. Weatherproof: A paired-dataset approach to semantic segmentation in adverse weather. *CoRR*, abs/2312.09534, 2023. 1

[13] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023. 2

[14] 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. 1, 3, 6, 7, 8

[15] Shuhang Gu, Deyu Meng, Wangmeng Zuo, and Lei Zhang. Joint convolutional analysis and synthesis sparse representation for single image layer separation. In *Proceedings of the IEEE international conference on computer vision*, pages 1708–1716, 2017. 2

[16] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *The Eleventh International Conference on Learning Representations*, 2022. 2, 3

[17] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 2, 3, 5

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3, 4

[19] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022. 3

[20] Seongmin Hong, Kyeonghyun Lee, Suh Yoon Jeon, Hyewon Bae, and Se Young Chun. On exact inversion of dpm-solvers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7069–7078, 2024. 4

[21] Wenyi Hong, Weihan 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. 5

[22] Chao Huang, Susan Liang, Yunlong Tang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Scaling concept with text-guided diffusion models. *arXiv preprint arXiv:2410.24151*, 2024. 2

[23] Inbar Huberman-Spiegelglas, Vladimir Kulikov, and Tomer 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. 3, 4, 8

[24] Li-Wei Kang, Chia-Wen Lin, and Yu-Hsiang Fu. Automatic single-image-based rain streaks removal via image decomposition. *IEEE transactions on image processing*, 21(4):1742–1755, 2011. 2

[25] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5148–5157, 2021. 6

[26] Beomsu Kim, Gihyun Kwon, Kwanyoung Kim, and Jong Chul Ye. Unpaired image-to-image translation via neural schrödinger bridge. In *ICLR*, 2024. 2, 3

[27] Yu Li, Robby T Tan, Xiaojie Guo, Jiangbo Lu, and Michael S Brown. Rain streak removal using layer priors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2736–2744, 2016. 2

[28] Chih-Hao Lin, Zian Wang, Ruofan Liang, Yuxuan Zhang, Sanja Fidler, Shenlong Wang, and Zan Gojcic. Controllable weather synthesis and removal with video diffusion models. *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 1, 3

[29] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Wanli Ouyang, Yu Qiao, and Chao Dong. Diff-bir: Towards blind image restoration with generative diffusion prior. *arXiv preprint arXiv:2308.15070*, 2023. 1, 3

[30] 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, 2, 3, 6, 7, 8

[31] Yu Luo, Yong Xu, and Hui Ji. Removing rain from a single image via discriminative sparse coding. In *Proceedings of the IEEE international conference on computer vision*, pages 3397–3405, 2015. 2

[32] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Controlling vision-language models for multi-task image restoration. In *The Twelfth International Conference on Learning Representations*. 1, 3

[33] Zhengyao Lv, Chenyang Si, Junhao Song, Zhenyu Yang, Yu Qiao, Ziwei Liu, and Kwan-Yee K Wong. Fastercache: Training-free video diffusion model acceleration with high quality. *arXiv preprint arXiv:2410.19355*, 2024. 5

[34] Willi Menapace, Aliaksandr Siarohin, Ivan Skorokhodov, Ekaterina Deyneka, Tsai-Shien Chen, Anil Kag, Yuwei Fang, Aleksei Stoliar, Elisa Ricci, Jian Ren, et al. Snap video: Scaled spatiotemporal transformers for text-to-video synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7038–7048, 2024. 8

[35] 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, 3, 4

[36] Daiki Miyake, Akihiro Iohara, Yu Saito, and Toshiyuki Tanaka. Negative-prompt inversion: Fast image inversion for editing with text-guided diffusion models. *arXiv preprint arXiv:2305.16807*, 2023. 4

[37] 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, 3, 4

[38] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 2

[39] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 3

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[41] 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. 5

[42] 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. 2, 3

[43] Ciyu Ruan, Ruishan Guo, Zihang Gong, Jingao Xu, Wenhan Yang, and Xinlei Chen. Pre-mamba: A 4d state space model for ultra-high-frequent event camera deraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 3

[44] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 3

[45] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 8

[46] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3

[47] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022. 2, 3, 4, 8

[48] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 3

[49] Shangquan Sun, Wenqi Ren, Jingzhi Li, Kaihao Zhang, Meiyu Liang, and Xiaochun Cao. Event-aware video deraining via multi-patch progressive learning. *IEEE Transactions on Image Processing*, 32:3040–3053, 2023. 3

[50] 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. 3, 6, 7, 8

[51] Shangquan Sun, Wenqi Ren, Juxiang Zhou, Jianhou Gan, Rui Wang, and Xiaochun Cao. A hybrid transformer-mamba network for single image deraining. *arXiv preprint arXiv:2409.00410*, 2024. 3

[52] Shangquan Sun, Wenqi Ren, Juxiang Zhou, Shu Wang, Jian-hou Gan, and Xiaochun Cao. Semi-supervised state-space model with dynamic stacking filter for real-world video deraining. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26114–26124, 2025. 3

[53] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 6

[54] Yoad Tewel, Rinon Gal, Dvir Samuel, Yuval Atzmon, Lior Wolf, and Gal Chechik. Add-it: Training-free object insertion in images with pretrained diffusion models. *arXiv preprint arXiv:2411.07232*, 2024. 2

[55] Michael Toker, Ido Galil, Hadas Orgad, Rinon Gal, Yoad Tewel, Gal Chechik, and Yonatan Belinkov. Padding tone: A mechanistic analysis of padding tokens in t2i models, 2025. 6

[56] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023. 1

[57] Bram Wallace, Akash Gokul, and Nikhil Naik. Edict: Exact diffusion inversion via coupled transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22532–22541, 2023. 4

[58] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2555–2563, 2023. 6

[59] Jin Wang, Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. Unsupervised video deraining with an event camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10831–10840, 2023. 3

[60] Yinglong Wang, Shuaicheng Liu, Chen Chen, and Bing Zeng. A hierarchical approach for rain or snow removing in a single color image. *IEEE Transactions on Image Processing*, 26(8):3936–3950, 2017. 2

[61] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. *arXiv preprint arXiv:2212.00490*, 2022. 3

[62] Yanyan Wei, Zhao Zhang, Yang Wang, Mingliang Xu, Yi Yang, Shuicheng Yan, and Meng Wang. Deraincyclegan: Rain attentive cyclegan for single image deraining and rainmaking. *IEEE Transactions on Image Processing*, 30:4788–4801, 2021. 1, 2, 3

[63] Yuanbo Wen, Tao Gao, and Ting Chen. Neural schrödinger bridge for unpaired real-world image deraining. *Information Sciences*, 682:121199, 2024. 1, 2, 3

[64] Yuanbo Wen, Tao Gao, and Ting Chen. Unpaired photo-realistic image deraining with energy-informed diffusion model. *arXiv preprint arXiv:2407.17193*, 2024. 1, 3

[65] Max Woolf. Stable diffusion 2.0 and the importance of negative prompts for good results, 2022. 2, 3, 5

[66] 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, 3, 6, 7, 8

[67] Wenhan Yang, Jiaying Liu, and Jiashi Feng. Frame-consistent recurrent video deraining with dual-level flow. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1661–1670, 2019. 3

[68] Wenhan Yang, Robby T Tan, Shiqi Wang, and Jiaying Liu. Self-learning video rain streak removal: When cyclic consistency meets temporal correspondence. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1720–1729, 2020. 3

[69] 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, 3, 5, 6, 8

[70] Yuntong Ye, Changfeng Yu, Yi Chang, Lin Zhu, Xi-Le Zhao, Luxin Yan, and Yonghong Tian. Unsupervised deraining: Where contrastive learning meets self-similarity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5821–5830, 2022. 1, 2, 3

[71] 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. 6, 7, 8

[72] Howard Zhang, Yunhao Ba, Ethan Yang, Varan Mehra, Blake Gella, Akira Suzuki, Arnold Pfahnl, Chethan Chinder Chandrappa, Alex Wong, and Achuta Kadambi. Weather-stream: Light transport automation of single image deweathering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1

[73] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 3

[74] Yueyi Zhang, Jin Wang, Wenming Weng, Xiaoyan Sun, and Zhiwei Xiong. Egvd: Event-guided video deraining. *IEEE Transactions on Neural Networks and Learning Systems*, 2025. 3

[75] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 3