Real-World Dual-Pixel Raindrop Removal: A New Benchmark and Degradation-Adaptive RWKV Baseline

Hao Yang[1,2] **Ruikun Zhang**[1] **Liyuan Pan**[1*]
[1]Beijing Institute of Technology, Beijing, China
[2]State Key Laboratory of General Artificial Intelligence, BIGAI, Beijing, China
{hao.yang, ruikun.zhang, liyuan.pan}@bit.edu.cn

Abstract

Raindrop removal is challenging because raindrops exhibit strong spatial non-uniformity and scene dependency. Although previous works have demonstrated that dual-pixel (DP) sensors can facilitate raindrop removal, their effectiveness in complex real-world scenarios remains limited due to local modeling paradigms and insufficient coverage of real data. In this work, we collect a large-scale real-world DP raindrop dataset that covers day/night scenes, diverse light conditions, and varying raindrop intensities. Based on this dataset, we establish a challenging benchmark for DP raindrop removal in real-world scenarios. Building upon this dataset, we develop DRWKV, which models raindrop removal as a degradation-conditioned state evolution process. DRWKV explicitly injects raindrop degradation information derived from DP sensors into the recurrent state update process, enabling the model to adaptively balance information preservation and content restoration across spatial regions while maintaining linear computational complexity. This design facilitates effective global modeling of spatially non-uniform raindrop degradations. Extensive experiments demonstrate that the proposed DRWKV outperforms prior methods on the proposed benchmark as well as existing datasets.

1 Introduction

Raindrops on lenses or windscreens represent a common yet complex optical degradation. They introduce refraction, occlusion, and severe structural distortion during image formation, significantly degrading overall visual quality. For intelligent systems dependent on stable visual input (e.g., autonomous driving [1], UAV inspection [2], outdoor surveillance [3], and multimodal visual understanding [4, 5]), these degradations severely impair the performance of downstream perception algorithms [6, 7]. Consequently, raindrop removal has emerged as a critical research direction in computer vision. However, despite substantial progress, highly diverse raindrop shapes, complex optical refraction, and strong coupling between raindrops and background content render it extremely challenging in real-world scenarios.

With the rapid development of deep learning, raindrop removal has achieved remarkable progress, *e.g.*, CNNs [8], neural architecture search (NAS) [9, 10], attention mechanisms [11], diffusion models [12], language priors [13], and generative priors [14]. Nevertheless, these methods still underperform in complex real-world scenarios, primarily due to the highly irregular spatial distribution, diverse geometric shapes, and complex refractive behaviors of real raindrops—factors that hinder single-image raindrop removal. While several works [15, 16, 17] have attempted to estimate raindrop masks

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

as priors to guide restoration in degraded regions, obtaining accurate masks from a single frame in real-world settings remains highly challenging.

Recently, dual-pixel sensors [19, 13, 20, 21, 18] have been widely adopted in modern imaging systems, including smartphones [22] and DSLR cameras [20]. A DP sensor captures two-view images (left and right views) in a single exposure, forming a stereo pair with a tiny baseline that provides informative stereo cues (Fig. 2). Such cues have proven valuable for various vision tasks, including depth estimation [19], defocus deblurring [20], and reflection removal [23]. More importantly, they hold significant potential for raindrop removal: DP images exhibit distinct DP disparity in raindrop regions, while in-focus background regions typically show near-zero disparity. This enables raindrop localization via disparity cues, rather than sole reliance on appearance.

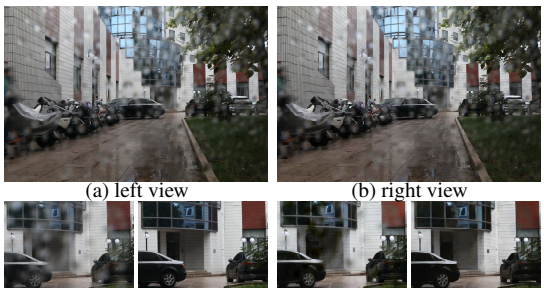


Figure 1: Illustration of DP raindrop removal. (a) and (b) show a pair of raindrop DP inputs, i.e., the left and right views. For simplicity, only the left-view image is shown as the input in the following figures. (c–f) show the zoomed-in degraded region, the ground-truth (GT) image, the result produced by DPRRN [18], and the result of our proposed method, respectively.

DPRRN [21, 18] first leveraged dual-pixel (DP) sensors for raindrop removal, using DP disparity as a prior to estimate raindrop masks that were then concatenated with the input to a CNN-based restoration network. They later extended the model to handle raindrops on tilted glass. While this opened a promising direction, several limitations remain: (i) CNNs struggle with large or strongly refractive raindrops (see Fig. 1), due to limited long-range dependency modeling; (ii) the raindrop prior is treated as an auxiliary cue rather than being integrated into the core state transition of restoration; and (iii) the dataset is restricted to grayscale inputs from the green channel of the DP sensor and relies heavily on synthetic data, resulting in substantial domain gaps and poor generalization in real-world scenarios.

To address these challenges, we propose a new raindrop removal framework that fully leverages the complementary advantages of raindrop map priors and the RWKV architecture [24]. To overcome (i), the lack of global modeling capability, we adopt the RWKV structure, which provides global receptive fields with linear complexity. RWKV performs recursive receptance-weighted key-value integration with linear complexity, capturing long-range dependencies across raindrop and background regions while maintaining linear inference cost. To address (ii), the insufficient utilization of DP masks, we introduce a Degradation-Adaptive State Modulation mechanism that adapts RWKV's recurrent state updates: the model preserves long-term memory in heavily degraded regions to enhance reconstruction while rapidly forgetting outdated states in clean background areas, achieving a more precise balance between "preservation" and "restoration." Finally, to address (iii), the scarcity of real DP raindrop data, we construct a large-scale real DP raindrop dataset containing full-color images captured across daytime and nighttime scenes, with diverse raindrop shapes, varying raindrop-to-lens distances, complex illumination, and richer real-world backgrounds, providing a significantly more challenging and comprehensive benchmark for future research.

Overall, our contributions can be summarized as follows:

- We propose DRWKV, which formulates dual-pixel raindrop removal as a degradation-conditioned state evolution problem with linear computational complexity.
- We introduce a degradation-adaptive state modulation mechanism that injects DP-derived raindrop priors into RWKV's recurrent state updates, enabling dynamic balancing between information preservation in clean regions and content restoration in degraded areas.
- We construct DNDP, the first large-scale real-world dual-pixel raindrop removal benchmark, featuring full-color images across day/night scenes, diverse lighting conditions, and complex raindrop patterns.

2 Related Work

Raindrop Removal. Adherent raindrop removal remains a challenging problem. With the advent of deep learning, the field has advanced rapidly, leading to numerous methods [13] and datasets [25]. AttGAN [8] pioneered attentive GAN-based raindrop removal and introduced the first large-scale real-world dataset. Subsequent works explored neural architecture search (NAS) [9, 10], Transformers [11, 26], diffusion models [15, 12], and high-level priors [13, 27]. However, single-frame methods still struggle in complex real-world scenes due to irregular raindrop distributions and diverse morphologies. To alleviate this issue, several studies exploit multi-view information, including stereo [28, 29] and light-field cameras [30].

Dual-Pixel Sensor. Recently, DP sensors have been adopted for raindrop removal. Compared with conventional stereo cameras, DP sensors offer a compact, single-unit design without requiring geometric calibration. Each pixel is split into two photodiodes, capturing left- and right-half views as slightly shifted observations (Fig. 2). Originally developed for phase-detection autofocus, DP sensing has been applied to depth estimation [19, 31], deblurring [20, 32], and reconstruction [33]. In real scenarios, cameras are typically focused on the background, making raindrops near the lens out of focus. As shown in Fig. 2, raindrop regions exhibit noticeable DP disparity, while in-focus background regions typically show nearly zero disparity. This indicates that DP disparity provides a strong prior for raindrop localization beyond appearance cues. DPRRN [21] first explored DP-based raindrop removal by estimating a raindrop mask from DP disparity and concatenating it with the input for CNN-based restoration. It was later extended to vertical DP sensor configurations [18]. However, these methods rely heavily on synthetic DP data, have limited real-world evaluation in scale and color diversity, and do not fully exploit the DP mask for degradation-aware restoration.

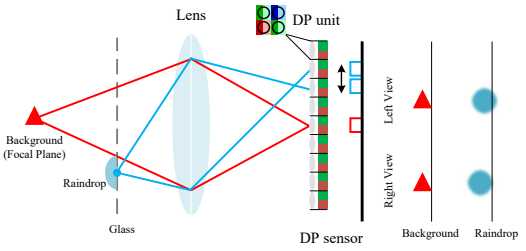


Figure 2: Illustration of the DP imaging model. Raindrops adhering to the glass lie outside the depth of field and induce measurable disparities between the left and right DP views, while in-focus background regions exhibit nearly zero disparity.

Receptance Weighted Key-Value (RWKV). RWKV was originally proposed for NLP [24], combining the parallel training efficiency of Transformers with the long-range memory of RNNs, and has recently been extended to vision, multimodal, and generative tasks, including Vision-RWKV [34], RWKV-SAM [35], Diffusion-RWKV [36], and RWKV-CLIP [37]. Despite these advances, its application to low-level image restoration remains underexplored [38, 39], with existing methods (e.g., Restore-RWKV [40], URWKV [41], Fourier-RWKV [42]) relying solely on feature recurrence without explicitly modeling degradation. In this paper, we present the first RWKV-based raindrop removal framework with degradation-adaptive state updates, enabling effective restoration under non-uniform real-world raindrop degradations.

3 Method

3.1 Overall Network Architecture

Fig. 3 illustrates the overall architecture of our method. Given a DP pair $(\mathbf{I}_{\mathrm{L}}, \mathbf{I}_{\mathrm{R}})$, our goal is to reconstruct a clean image $\mathbf{I}_{\mathrm{c}}$ that best approximates the ground-truth $\mathbf{I}_{\mathrm{gt}}$. Our framework consists of two branches: (i) an RWKV-based restoration backbone, which follows an encoder-decoder design to restore a clean image from the degraded DP inputs; and (ii) a degradation perception network that produces raindrop degradation maps from a DP pair. The detailed components are described in the following sections.

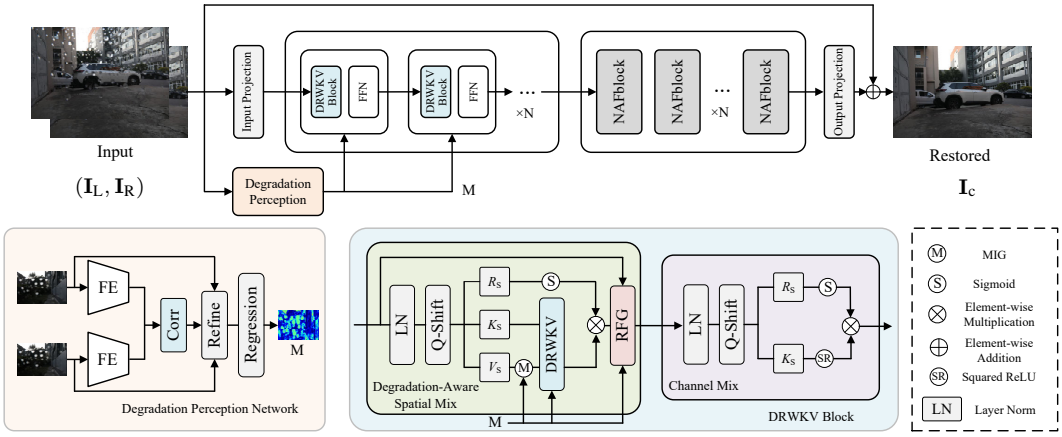


Figure 3: Overall architecture of the proposed framework. Given a 6-channel DP input pair $(\mathbf{I}_L, \mathbf{I}_R)$, the network aims to reconstruct a clean background image $\mathbf{I}_c$ that approximates the ground-truth background $\mathbf{I}_{gt}$. The DP pair is first processed by the degradation perception network to estimate raindrop degradation features, which are then used to modulate the RWKV-based restoration backbone. The backbone follows an encoder-decoder design and predicts a residual map that is added to the input to produce the final restored image $\mathbf{I}_c$.

3.2 Raindrop Degradation Prior

Accurate raindrop localization priors are crucial for effective image raindrop removal. In dual-pixel (DP) images, raindrop regions exhibit pronounced disparity, while in-focus background regions typically show near-zero disparity (see Fig. 2). Leveraging this geometric property, our Degradation Perception Network (DPN) extracts raindrop location priors from DP inputs to facilitate image restoration.

Degradation Perception Network. To obtain raindrop degradation priors from DP inputs, we design a disparity-based raindrop estimation module [43]. Given the left and right views $(\mathbf{I}_L, \mathbf{I}_R)$ captured by a DP sensor, the two images are first fed into convolutional encoders with shared weights to extract consistent feature representations: $\mathbf{F}_L = \mathrm{E}(\mathbf{I}_L), \mathbf{F}_R = \mathrm{E}(\mathbf{I}_R)$, where $E(\cdot)$ denotes the shared convolutional encoder. Next, we compute a correlation volume [44] between the two feature maps to capture disparity feature map $\mathbf{C} = \mathrm{Corr}(\mathbf{F}_L, \mathbf{F}_R)$. DP disparity provides a useful geometric cue, but defocused near-field background objects may also exhibit nonzero disparity. We therefore combine disparity features with image appearance and complementary degradation supervision, rather than using disparity alone as a raindrop decision rule.

We then introduce a lightweight Refine Network that takes the disparity features and raindrop images as input, performing spatial and channel-wise context aggregation to further refine the disparity representation. Finally, the refined features are regressed into a continuous-valued raindrop map $\mathbf{M}$, where each pixel in $\mathbf{M}$ represents the estimated raindrop degradation intensity.

Learning Objectives. Obtaining accurate raindrop map annotations in real-world is extremely challenging. Raindrops exhibit diverse shapes, thicknesses, and refraction patterns, making manual labeling costly and unstable. Therefore, we adopt a mask-free learning strategy to train our degradation perception network, allowing the model to learn a raindrop map without requiring ground-truth raindrop masks. Paired rainy/clean images are still used for supervision.

To this end, we design three complementary losses. First, let $\mathbf{I}_{\text{rain}} = (\mathbf{I}_L + \mathbf{I}_R)/2$ denote the observed rainy image. We treat the predicted raindrop map as a residual approximation between $\mathbf{I}_{\text{rain}}$ and the clean background $\mathbf{I}_{gt}$:

$$\mathcal{L}_{\text{res}} = \left\| \mathbf{M} - |\mathbf{I}_{\text{rain}} - \mathbf{I}_{gt}| \right\|_1 , \tag{1}$$

encouraging the output $\mathbf{M}$ to reflect the degradation intensity.

Second, we introduce a re-raining constraint: the clean image and the predicted raindrop map $\mathbf{M}$ are used to synthesize $\hat{\mathbf{I}}_{\text{rain}}$, which is encouraged to match the observed rainy image $\mathbf{I}_{\text{rain}}$:

$$\hat{\mathbf{I}}_{\text{rain}} = \text{RD}(\mathbf{I}_{\text{gt}}, \mathbf{M}), \qquad \mathcal{L}_{\text{syn}} = \|\hat{\mathbf{I}}_{\text{rain}} - \mathbf{I}_{\text{rain}}\|_1, \tag{2}$$

where RD is a distilled raindrop generation model [45] that takes a rain map and a clean image as input and outputs a rainy image.

Finally, leveraging the physical characteristics of DP sensors, we impose a DP background consistency constraint. For in-focus background regions, the left and right views are approximately consistent, whereas raindrops often produce noticeable shifts. We use the predicted background weight $\mathbf{1} - \mathbf{M}$ to impose the following soft geometric constraint:

$$\mathcal{L}_{\text{dp}} = \|(\mathbf{1} - \mathbf{M}) \cdot (\mathbf{I}_{\text{L}} - \mathbf{I}_{\text{R}})\|_1. \tag{3}$$

The final mask-free supervised optimization objective is $\mathcal{L}_{\text{mask}} = \mathcal{L}_{\text{res}} + \lambda_1 \mathcal{L}_{\text{syn}} + \lambda_2 \mathcal{L}_{\text{dp}}$, where $\lambda_1 = 0.1$ and $\lambda_2 = 0.1$ balance the contributions of different loss terms. The low-weight $\mathcal{L}_{\text{dp}}$ term is complemented by residual and re-synthesis supervision, which tie the estimated map to raindrop-induced degradation.

3.3 RWKV-based restoration backbone

Our raindrop removal backbone adopts RWKV as the core computation unit within a four-stage U-shaped multi-scale architecture. Given a degraded DP pair $(\mathbf{I}_{\text{L}}, \mathbf{I}_{\text{R}})$, shallow features $\mathbf{F}_0$ are first extracted via a convolutional projection and then processed by an encoder–decoder hierarchy. Each encoder stage consists of a DRWKV block and a CNN-based FFN [46], where raindrop degradation priors are injected into RWKV to guide restoration, while progressively reducing spatial resolution and increasing channel width. The decoder takes the low-resolution latent representation $\mathbf{F}_l$ and gradually reconstructs high-resolution features using NAFBlocks [47]. Finally, a convolutional projection generates the raindrop-free output image. The key modules are introduced below.

3.3.1 DRWKV Block

In the original RWKV model, sequence modeling relies on a single recurrent state update, where token interactions are determined solely by the current input and the previous hidden state. When applied to raindrop removal, this design has two key limitations. First, all spatial tokens are processed uniformly, preventing the model from adapting to spatially varying degradations. Second, the fixed decay behavior in RWKV may cause long-range information to diminish during recurrence, making early structural cues difficult to preserve in deeper restoration stages. Unlike conventional mask-guided restoration that uses the degradation map only as an auxiliary input, our formulation directly modulates the recurrent state transition, allowing the degradation prior to affect how historical information is retained, updated, and fused during long-range spatial propagation. To this end, we introduce a degradation-conditioned state control mechanism consisting of three components: the decay modulation gate (DMG), memory injection gate (MIG), and restoration fusion gate (RFG). Specifically, DMG enables spatially adaptive forgetting and preserves crucial cues in heavily degraded regions; MIG continuously strengthens corrupted-region representations during recurrent updates, compensating for insufficient information flow in fixed state transitions; and RFG adaptively fuses original and restored features at the output stage to prevent over-processing.

Degradation-Aware Spatial Mix. The spatial mix builds long-range dependencies along the spatial dimension. Given an input flattened to $\mathbf{F} \in \mathbb{R}^{T \times C}$ with $T = H \times W$, we first apply layer normalization (LN) and the Q-Shift [34], $\mathbf{F}_{\text{s}} = \text{Q-Shift}(\text{LN}(\mathbf{F}))$. We then project $\mathbf{F}_{\text{s}}$ into the receptance, key, and value:

$$\mathbf{R}_{\text{s}} = \mathbf{F}_{\text{s}} \mathbf{W}_{\mathbf{R}_{\text{s}}}, \quad \mathbf{K}_{\text{s}} = \mathbf{F}_{\text{s}} \mathbf{W}_{\mathbf{K}_{\text{s}}}, \quad \mathbf{V}_{\text{s}} = \mathbf{F}_{\text{s}} \mathbf{W}_{\mathbf{V}_{\text{s}}}, \tag{4}$$

where $\mathbf{W}_{\mathbf{R}_{\text{s}}}$, $\mathbf{W}_{\mathbf{K}_{\text{s}}}$, and $\mathbf{W}_{\mathbf{V}_{\text{s}}}$ denote the linear projection layers. To adaptively emphasize degraded tokens, we introduce a degradation-guided input modulation named MIG on the value tokens:

$$\tilde{\mathbf{V}}_{\text{s}} = \beta(m) \odot \mathbf{V}_{\text{s}}, \tag{5}$$

where $m \in [0, 1]^T$ denotes the token-wise degradation intensity derived from the raindrop map $\mathbf{M}$. In practice, $\mathbf{M}$ is predicted by DPN, resized to the current feature resolution, and flattened into m. The modulation function $\beta(\cdot)$ is implemented as a lightweight learnable linear layer that maps

Table 1: Quantitative comparison on synthetic and real raindrop datasets DPRRN [18] (Vertical and Tilted). The best results are highlighted in bold. The upper section presents single-image restoration methods, while the lower shows DP-based restoration methods.

Methods	Venue	Syn. (Vertical)		Real (Vertical)		Syn. (Tilted)		Real (Tilted)	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
AttGAN [8]	CVPR 18	30.22	0.9038	29.55	0.9014	31.85	0.9372	30.99	0.9120
CCN [10]	CVPR 21	35.07	0.9515	32.13	0.9273	35.70	0.9668	31.03	0.9170
MPRNet [48]	CVPR 21	36.76	0.9615	33.16	0.9393	38.27	0.9749	34.69	0.9372
Restormer [46]	CVPR 22	37.53	0.9638	32.11	0.9304	38.95	0.9761	33.89	0.9348
DGUNet [49]	CVPR 22	37.19	0.9633	31.99	0.9304	38.89	0.9764	34.42	0.9371
StereoWaterdrop [29]	IROS 21	34.55	0.9410	32.52	0.9260	34.97	0.9545	32.52	0.9187
EPRRNet [28]	IJCV 22	36.36	0.9583	33.02	0.9363	37.81	0.9734	32.36	0.9286
Histoformer [50]	ECCV 24	37.39	0.9608	32.30	0.9345	38.95	0.9749	33.88	0.9367
MoCE-IR [51]	CVPR 25	37.55	0.9642	32.22	0.9325	38.90	0.9752	33.85	0.9345
MambaIRv2 [52]	CVPR 25	37.65	0.9655	32.51	0.9291	38.99	0.9781	33.95	0.9306
RDPD+ [53]	ICCV 21	35.71	0.9555	34.18	0.9419	36.89	0.9688	33.98	0.9308
DPRRN [18]	TPAMI 24	37.96	0.9693	34.84	0.9521	39.46	0.9793	35.26	0.9393
Ours	–	**38.75**	**0.9701**	**35.35**	**0.9585**	**40.65**	**0.9808**	**36.84**	**0.9563**

m to token-wise modulation coefficients. In this way, tokens in heavily degraded regions receive stronger memory injection, while cleaner tokens are weakly modulated. The global interaction is then computed by linear-complexity RWKV as $wkv = \mathrm{DRWKV}(\mathbf{K}_\mathrm{s}, \tilde{\mathbf{V}}_\mathrm{s})$.

Then, a sigmoid-gated receptance modulates the attention output: $\mathbf{Y}_\mathrm{s} = \big(\sigma(\mathbf{R}_\mathrm{s}) \odot wkv\big)\mathbf{W}_{\mathbf{Y}_\mathrm{s}}$, where W_{Y_s} denote the linear layers and $\sigma(\cdot)$ denotes the sigmoid function. To further adapt restoration strength to degradation levels, we fuse the attention output with the original feature via a degradation-aware gate: $\mathbf{O}_\mathrm{s} = \gamma(m) \odot \mathbf{Y}_\mathrm{s} + \big(1 - \gamma(m)\big) \odot \mathbf{X}_\mathrm{s}$, where $\gamma(m)$ is a restoration fusion gate (RFG), which controls whether a token should rely more on the restored features or the original signal.

Channel Mix. The channel mix performs fusion along the channel dimension. Given $\mathbf{F} \in \mathbb{R}^{T\times C}$, we first apply LN and Q-Shift, $\mathbf{F}_\mathrm{c} = \text{Q-Shift}(\mathrm{LN}(\mathbf{F}))$. We then project $\mathbf{F}_\mathrm{c}$ into the receptance, key, and value:

$$\mathbf{R}_\mathrm{c} = \mathbf{F}_\mathrm{c}\mathbf{W}_{\mathbf{R}_\mathrm{c}}, \mathbf{K}_\mathrm{c} = \mathbf{F}_\mathrm{c}\mathbf{W}_{\mathbf{K}_\mathrm{c}}, \mathbf{V}_\mathrm{c} = \mathrm{SR}(K_\mathrm{c})\mathbf{W}_{\mathbf{V}_\mathrm{c}}, \tag{6}$$

where $\mathbf{W}_{\mathbf{R}_\mathrm{c}}$, $\mathbf{W}_{\mathbf{K}_\mathrm{c}}$, and $\mathbf{W}_{\mathbf{V}_\mathrm{c}}$ are linear projections, and $\mathrm{SR}(\cdot)$ is the squared-ReLU activation. $\mathbf{V}_\mathrm{c}$ is regressed from $\mathbf{K}_\mathrm{c}$ rather than directly from $\mathbf{F}_\mathrm{c}$, which can be viewed as an MLP transformation along channels. The output is $\mathbf{O}_\mathrm{c} = \big(\sigma(\mathbf{R}_\mathrm{c}) \odot \mathbf{V}_\mathrm{c}\big)\mathbf{W}_{\mathbf{O}_\mathrm{c}}$.

3.3.2 Recurrent WKV Attention

The WKV attention enables long-range dependency modeling with linear complexity, avoiding the quadratic cost of standard self-attention. Following prior work [40], we adopt Bi-WKV [34] as the basis of our WKV formulation to better adapt the recurrent computation to image structures.

Bidirectional attention. Given $\mathbf{K}_\mathrm{s}$ and $\mathbf{V}_\mathrm{s}$, the attention result for the t-th token, $wkv_t \in \mathbb{R}^C$, is

$$\begin{aligned} wkv_t &= \text{Bi-WKV}(\mathbf{K}_\mathrm{s}, \mathbf{V}_\mathrm{s})_t \\ &= \frac{\sum_{i=1,\, i\neq t}^{T} e^{-(|t-i|-1)/T\cdot w + k_i}\, v_i \;+\; e^{u+k_t}\, v_t}{\sum_{i=1,\, i\neq t}^{T} e^{-(|t-i|-1)/T\cdot w + k_i} \;+\; e^{u+k_t}}, \end{aligned} \tag{7}$$

where T is the number of tokens, and $k_i, v_i \in \mathbb{R}^C$ denote the i-th tokens of $\mathbf{K}_\mathrm{s}$ and $\mathbf{V}_\mathrm{s}$. The relative bias $-(|t-i|-1)/T$ is scaled by learnable $w \in \mathbb{R}^C$, while $u \in \mathbb{R}^C$ boosts the current token. Bi-WKV achieves global receptive fields with linear complexity while avoiding query-key multiplication.

Degradation-Modulated Recurrent Attention. To better preserve heavily degraded regions while suppressing background redundancy, we modulate the recurrent state with a degradation-aware decay term: $wkv^{(j)} = \text{Bi-WKV}^{(j)}\left(\Delta_\mathrm{dir}(\mathbf{K}_\mathrm{s}), \Delta_\mathrm{dir}(\alpha(m) \odot wkv^{(j-1)})\right)$, where $\Delta_\mathrm{dir}(\cdot)$ represents the



(a) (b) (c) (d)

Figure 4: Samples from our DNDP dataset and DPRRN [18] are shown in (a)–(b) and (c)–(d), respectively.



(a) Input (b) GT (c) DPRRN (d) Ours

Figure 5: Qualitative comparison on the synthetic dataset: (a) Input, (b) GT, (c) DPRRN, and (d) Ours.

changing direction operation and $\alpha(m)$ controls token-wise forgetting or memory retention named decay modulation gate (DMG), allowing degraded regions to retain memory with slower decay. After P recurrent passes, we have $wkv = \mathrm{DRWKV}(\mathbf{K}_\mathrm{s}, \mathbf{V}_\mathrm{s}) = wkv^{(P)}$. We use $P = 2$ with horizontal followed by vertical scanning.

3.4 Loss

For training, we employ image restoration loss (Input $\mathbf{I}_\mathrm{c}$ and $\mathbf{I}_\mathrm{gt}$) $\mathcal{L}_\mathrm{char}$ [46] and the raindrop mask estimation loss $\mathcal{L}_\mathrm{mask}$ (see Sec. 3.2). The overall training loss is formulated as $\mathcal{L} = \mathcal{L}_\mathrm{char} + \lambda\mathcal{L}_\mathrm{mask}$, where $\lambda = 0.2$ is a balancing weight.

4 Proposed Dataset

Data Acquisition. To construct real-world raindrop dataset, we use a Canon 5D Mark IV Camera equipped with a dual-pixel sensor. The camera is fixed on a tripod, and a transparent glass sheet is placed in front of the lens. The fixed-camera, static-scene protocol is used to obtain aligned rainy/clean supervision and is not a requirement at inference. A controlled spraying device produces water droplets of different densities and sizes on the glass substrate. For each scene, we first capture a clean background image without any obstruction as the reference. To obtain raindrops with diverse sizes, shapes, and spatial distributions, we vary the spraying pattern, water amount, and glass-to-lens distance while keeping camera parameters fixed within each paired acquisition. We repeat data acquisition under diverse lighting conditions, including varying daylight angles and nighttime artificial illumination, to capture diverse raindrop blur patterns.

Image Post-Processing. We store all RAW sensor data and later decode it into standard JPG format using Canon Digital Photo Professional 4 to ensure consistent image processing. We then perform a strict quality control process to remove all low-quality samples. Compared with the dataset of DPRRN [18], our dataset provides richer color information, higher resolution, more diverse degradations, and a significantly larger scale. While their DP raindrop dataset contains only single-channel images and about 130 real samples (with the rest being synthetic), our dataset expands to thousands of high-quality real images, offering a more comprehensive representation of real-world raindrop degradations (see Fig. 4). Our dataset is publicly released on our project page, and more details can be found in Appendix B.

5 Experiments and Analysis

5.1 Experimental Setup

Implementation Details. All experiments are conducted on a workstation equipped with four NVIDIA RTX 4090 GPUs, and the models are implemented in PyTorch. We use Adam as the optimizer with an initial learning rate of 2×10^{-4}, which is cosine-annealed to 1×10^{-6}. We train for 600k iterations with a batch size of 2. During training, input images are cropped to 256×256, with random rotation and horizontal flipping for data augmentation. The code, datasets, and additional details are available on our project page.

Table 2: Quantitative comparison on our DNDP under daytime, nighttime, and overall scenes. The best results are highlighted in bold.

Methods	Venue	Daytime			Nighttime			Overall		
		PSNR	SSIM	CLIP-IQA+	PSNR	SSIM	CLIP-IQA+	PSNR	SSIM	CLIP-IQA+
MPRNet [48]	CVPR 21	21.10	0.8068	0.330	24.36	0.8199	0.266	22.49	0.8124	0.302
DGUNet [49]	CVPR 22	22.25	0.8061	0.415	23.97	0.8196	0.321	22.98	0.8119	0.375
Restormer [46]	CVPR 22	23.72	0.8214	0.400	23.99	0.8194	0.311	23.84	0.8206	0.362
Histoformer [50]	ECCV 24	23.54	0.8134	0.468	22.98	0.7955	0.332	23.31	0.8057	0.410
MoCE-IR [51]	CVPR 25	24.49	0.8355	0.467	24.17	0.8181	0.336	24.35	0.8281	0.411
MambaIRv2 [52]	CVPR 25	24.60	0.8192	0.434	24.49	0.8240	0.315	24.55	0.8212	0.383
DPRRN [18]	TPAMI 24	24.93	0.8446	0.474	23.59	0.8106	0.319	24.36	0.8300	0.408
Ours	–	**25.85**	**0.8543**	**0.489**	**25.36**	**0.8396**	**0.322**	**25.64**	**0.8480**	**0.417**



Figure 6: Qualitative evaluation of different image restoration methods. The first column represents rain images, the second column shows the ground truth (GT), and the subsequent columns display the restoration results of SOTA methods and our method.

Datasets. We use the DP raindrop removal dataset proposed by DPRRN [18], which was captured using a Google Pixel camera. The dataset consists of Syn. (Vertical), Syn. (Tilted), Real (Vertical), and Real (Tilted) subsets. Among them, Syn. (Vertical) and Syn. (Tilted) contain 1,960/492 images for training/testing, while Real (Vertical) and Real (Tilted) include 125 and 130 real-world images for testing. We also evaluate our method on our collected DNDP dataset, captured using a Canon 5D Mark IV equipped with a DP sensor. The DNDP dataset comprises 5,500 training images (57% daytime and 43% nighttime) and 1,300 testing images (55% daytime and 45% nighttime, providing a large-scale and realistic benchmark for DP raindrop removal.

Evaluation Metrics. We employ distortion-oriented metrics, PSNR and SSIM [54], to evaluate the pixel-level fidelity and structural consistency. In addition, we adopt the perception-oriented metric CLIP-IQA+ [55] to assess perceptual image quality.

Baseline. We compare with classic raindrop removal methods, including AttGAN[8], CCN[10], MPRNet [48], Restormer [46], DGUNet [49] StereoWaterdrop [29], EPRRNet[28], Histoformer [50], MoCE-IR [51], MambaIRv2[52], RDPD+[53], and DPRRN [18]. For evaluation on our DNDP dataset, we retrain the top seven methods ranked on Syn. (Vertical). For single-image baselines, we adapt only the input projection from three to six channels by concatenating the left and right views. The remaining network components are unchanged.

5.2 Results

Quantitative evaluation. We first evaluate the proposed method on the synthetic DP raindrop datasets, including Syn. (Vertical) and Syn. (Tilted). As shown in Tab. 1, our method achieves the

Table 3: Ablations of the proposed modules.

Components				PSNR↑	SSIM↑
DP	DPN	DRWKV	FFN		
✗	✗	✗	✗	23.25	0.8219
✓	✗	✗	✗	23.58	0.8285
✓	✓	✗	✗	23.95	0.8301
✓	✓	✓	✗	24.50	0.8407
✓	✓	✓	✓	**25.64**	**0.8480**

Table 4: Ablations of the three components of DRWKV.

Components			PSNR↑	SSIM↑
DMG	MIG	RFG		
✗	✗	✗	25.12	0.8424
✓	✗	✗	25.33	0.8432
✓	✓	✗	25.50	0.8462
✓	✓	✓	**25.64**	**0.8480**

best PSNR of 38.75 dB and 40.65 dB on the two datasets, outperforming the second-best method by 0.79 dB and 1.19 dB, respectively. On the real-world Real (Vertical) and Real (Tilted) datasets, our method also delivers superior performance. On our proposed DNDP dataset (Tab. 2), our method achieves the best results on both daytime and nighttime subsets, surpassing DPRRN by 1.28 dB in overall PSNR. It also obtains the highest CLIP-IQA+ score, indicating better perceptual quality. In addition, our method runs in 0.41 s on an RTX 4090. The expanded computational comparison is reported in the appendix (Tab. 11).

Qualitative evaluation. Comparisons on the synthetic dataset and the real-world DNDP dataset, are shown in Figs. 5 and 6. The results indicate that our method produces sharper and cleaner restorations than existing approaches. Additional results are provided in Appendix E.

Nighttime Performance. The PSNR gain over DPRRN increases from 0.92 dB in daytime to 1.77 dB at night. The smaller nighttime gain is in CLIP-IQA+ (0.322 versus 0.319), rather than PSNR. Residual underexposure and sensor noise may limit perceived image quality even when raindrop removal improves. CLIP-IQA+ also reflects overall appearance, including brightness, contrast, and naturalness.

Downstream Perception. We further evaluate detection and segmentation on 90 annotated road/outdoor test images. DRWKV obtains the best scores among the compared methods in this evaluation (Tab. 14). In particular, restored inputs improve Box mAP and Mask mAP by 9.99 and 10.15 percentage points over rainy inputs with the same frozen SAM 3 [56]. Full protocols and results are provided in Appendix C.

5.3 Ablation Studies and Discussions

We conduct ablation studies on the DNDP dataset to validate the effectiveness of each module.

Model Architecture. Tab. 3 presents the ablation studies conducted on different network components. DP denotes the dual-pixel input; removing DP means that we instead feed a single image obtained by averaging the left and right views. DPN, DRWKV, and FFN correspond to the three modules illustrated in Fig. 3. By progressively disabling each component, we assess its individual contribution to the overall performance. The full model achieves the best results, demonstrating the effectiveness of all proposed modules.

Effectiveness of DRWKV Design. We further analyze the effectiveness of the degradation-aware state update in the proposed DRWKV, as reported in Tab. 4. When all three gating modules are enabled, the model can adaptively control memory decay (DMG), dynamic input gating (MIG), and restoration fusion gating (RFG) along the temporal dimension, forming a degradation-driven memory evolution mechanism that achieves the best overall performance.

Robustness to Near-field Objects. We evaluate 128 DNDP test pairs, manually annotating near-field regions N on clean images and raindrop regions R on rainy images for evaluation only. Tab. 5 reports full-image and near-field PSNR/SSIM and mask IoU/Dice within N. Adding residual and re-synthesis supervision to DP-only mask learning improves near-field PSNR by 0.37 dB and mask IoU by 0.1176, supporting complementary supervision for ambiguous near-field disparity. Extremely defocused close objects remain challenging.

Table 5: Near-field robustness on 128 test pairs. PSNR/SSIM and mask IoU/Dice in near-field regions.

Method	PSNR / SSIM↑		Mask↑	
	Overall	Near-field	IoU	Dice
DPRRN	27.38/0.904	31.68/0.930	0.1463	0.253
DRWKV ($\mathcal{L}_{\text{dp}}$-only)	27.67/0.905	32.42/0.931	0.3856	0.489
DRWKV	**28.16/0.908**	**32.79/0.933**	**0.5032**	**0.669**

Table 6: Parameter-matched results with a 24-hour training budget.

Method	Params (M)	PSNR↑	SSIM↑
MoCE-IR	17.24	23.78	0.824
MambaIRv2	17.19	23.51	0.815
DRWKV	17.11	**24.67**	**0.832**

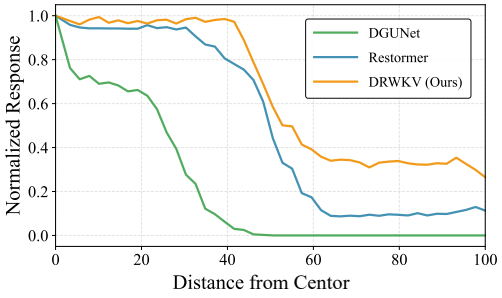
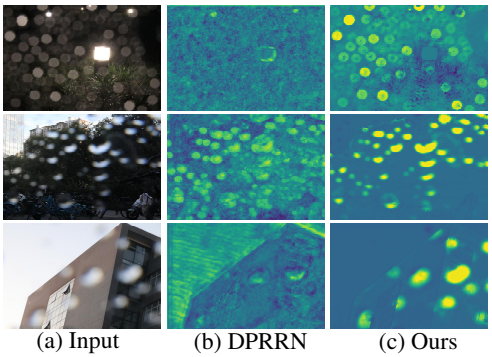


Figure 7: Long-range dependency visualization. The horizontal axis denotes pixel distance from the center, and the vertical axis shows the averaged gradient response. RWKV maintains stronger responses at larger distances, indicating superior long-range modeling capability.



(a) Input (b) DPRRN (c) Ours

Figure 8: Raindrop mask visualization. (a) shows the input image, while (b) and (c) present the raindrop masks estimated by DPRRN and our method, respectively. Our method produces the most accurate and reliable raindrop masks across all scenarios.

Comparisons under Matched Training Budgets. We resize MoCE-IR [51] and MambaIRv2 [52] to 17.24M and 17.19M parameters, respectively, close to DRWKV (17.11M), and train all three for approximately 24 hours under the same hardware and training setting. As shown in Tab. 6, DRWKV improves PSNR by 0.89 dB and 1.16 dB over the matched MoCE-IR and MambaIRv2.

Long-Range Dependency Analysis. We adopt a gradient response to analyze long-range dependency modeling [57] across architectures. A central local output region is used as the target signal, and gradient magnitudes with respect to input pixels are aggregated by their spatial distance to the target center. As shown in Fig. 7, DGUNet (CNN-based) exhibits a rapid response decay, reflecting its local receptive field, while Restormer (Transformer-based) maintains responses over a broader range. In contrast, our RWKV preserves stronger responses at larger distances, indicating more effective long-range dependency modeling ability.

Raindrop Map Visualization. To demonstrate the effectiveness of the proposed DPN in estimating raindrop map, we visualize the raindrop maps predicted by DPN. As shown in Fig. 8, the raindrop estimation module in DPRRN [18] completely fails under nighttime conditions, producing noisy maps. In contrast, our method achieves the most accurate and consistent raindrop estimations in both day and night scenes, clearly demonstrating its robustness.

6 Conclusion

In this paper, we present DRWKV, a degradation-aware RWKV framework for dual-pixel (DP) raindrop removal. By explicitly injecting DP-derived raindrop degradation priors into the recurrent state update, DRWKV effectively models spatially non-uniform degradations while preserving linear computational complexity. In addition, we construct a large-scale real-world DP raindrop dataset covering diverse day/night scenes, establishing a challenging benchmark for real-world evaluation. Extensive experiments demonstrate that DRWKV consistently outperforms existing methods.

Acknowledgments and Disclosure of Funding

This work is supported by the National Natural Science Foundation of China (62302045), the Fundamental Research Funds for the Central Universities, and in part by the Opening Project of the State Key Laboratory of General Artificial Intelligence, BIGAI/Peking University, Beijing, China (Project No. SKLAGI20250P05).

References

[1] Yasin Almalioglu, Mehmet Turan, Niki Trigoni, and Andrew Markham. Deep learning-based robust positioning for all-weather autonomous driving. *Nature Machine Intelligence*, 4(9):749–760, 2022.

[2] Wenhui Chang, Hongming Chen, Xin He, Xiang Chen, and Liangduo Shen. UAV-Rain1k: A benchmark for raindrop removal from UAV aerial imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 15–22, 2024.

[3] Yan-Feng Lu, Jing-Wen Gao, Qian Yu, Yi Li, Yi-Sheng Lv, and Hong Qiao. A cross-scale and illumination invariance-based model for robust object detection in traffic surveillance scenarios. *IEEE Transactions on Intelligent Transportation Systems*, 24(7):6989–6999, 2023.

[4] Yudong Han, Qingpei Guo, Liyuan Pan, Liu Liu, Yu Guan, and Ming Yang. DynFocus: Dynamic cooperative network empowers LLMs with video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8512–8522, June 2025.

[5] Yudong Han, Yong Wang, Zaiquan Yang, Zhen Qu, Liyuan Pan, and Xiangxiang Chu. Visual enhanced depth scaling for multimodal latent reasoning. *arXiv preprint arXiv:2604.10500*, 2026.

[6] Xiaorui Sun, Jun Liu, Hengtao Shen, Xiaofeng Zhu, and Ping Hu. On efficient variants of segment anything model: A survey. *International Journal of Computer Vision*, 133(10):7406–7436, 2025.

[7] Junke Wang, Zuxuan Wu, Dongdong Chen, Chong Luo, Xiyang Dai, Lu Yuan, and Yu-Gang Jiang. OmniTracker: Unifying visual object tracking by tracking-with-detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4):3159–3174, 2025.

[8] Rui Qian, Robby T Tan, Wenhan Yang, Jiajun Su, and Jiaying Liu. Attentive generative adversarial network for raindrop removal from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2482–2491, 2018.

[9] Ruoteng Li, Robby T Tan, and Loong-Fah Cheong. All in one bad weather removal using architectural search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3175–3185, 2020.

[10] Ruijie Quan, Xin Yu, Yuanzhi Liang, and Yi Yang. Removing raindrops and rain streaks in one go. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9147–9156, 2021.

[11] Jeya Maria Jose Valanarasu, Rajeev Yasarla, and Vishal M. Patel. TransWeather: Transformer-based restoration of images degraded by adverse weather conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2353–2363, 2022.

[12] Ozan Özdenizci and Robert Legenstein. Restoring vision in adverse weather conditions with patch-based denoising diffusion models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10346–10357, 2023.

[13] Hao Yang, Liyuan Pan, Yan Yang, and Wei Liang. Language-driven all-in-one adverse weather removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24902–24912, 2024.

[14] Hao Yang, Yan Yang, Ruikun Zhang, and Liyuan Pan. A preliminary study on GPT-Image generation model for image restoration. *arXiv preprint arXiv:2505.05621*, 2025.

[15] Lhuqita Fazry and Valentino Vito. Unsupervised raindrop removal from a single image using conditional diffusion models. *arXiv preprint arXiv:2505.08190*, 2025.

[16] Zhixiang Hao, Shaodi You, Yu Li, Kunming Li, and Feng Lu. Learning from synthetic photorealistic raindrop for single image raindrop removal. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 4340–4349, 2019.

[17] Huijiao Wang, Shenghao Zhao, Lei Yu, and Xulei Yang. Unsupervised network for single image raindrop removal. *arXiv preprint arXiv:2412.03019*, 2024.

[18] Yizhou Li, Yusuke Monno, and Masatoshi Okutomi. Dual-pixel raindrop removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10748–10762, 2024.

[19] Liyuan Pan, Richard Hartley, Liu Liu, Zhiwei Xu, Shah Chowdhury, Yan Yang, Hongguang Zhang, Hongdong Li, and Miaomiao Liu. Weakly-supervised depth estimation and image deblurring via dual-pixel sensors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):11314–11330, 2024.

[20] Abdullah Abuolaim and Michael S Brown. Defocus deblurring using dual-pixel data. In *European Conference on Computer Vision (ECCV)*, pages 111–126. Springer, 2020.

[21] Yizhou Li, Yusuke Monno, and Masatoshi Okutomi. Dual-pixel raindrop removal. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press, 2022.

[22] Shumian Xin, Neal Wadhwa, Tianfan Xue, Jonathan T Barron, Pratul P Srinivasan, Jiawen Chen, Ioannis Gkioulekas, and Rahul Garg. Defocus map estimation and deblurring from a single dual-pixel image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2228–2238, 2021.

[23] Abhijith Punnappurath and Michael S Brown. Reflection removal using a dual-pixel sensor. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1556–1565, 2019.

[24] Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Leon Derczynski, et al. RWKV: Reinventing RNNs for the transformer era. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14048–14077, 2023.

[25] Yeying Jin, Xin Li, Jiadong Wang, Yan Zhang, and Malu Zhang. Raindrop clarity: A dual-focused dataset for day and night raindrop removal. In *European Conference on Computer Vision (ECCV)*, pages 1–17. Springer, 2024.

[26] Sixiang Chen, Tian Ye, Jinbin Bai, Erkang Chen, Jun Shi, and Lei Zhu. Sparse sampling transformer with uncertainty-driven ranking for unified removal of raindrops and rain streaks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13106–13117, 2023.

[27] Chao Wang, Zhedong Zheng, Ruijie Quan, and Yi Yang. Depth-aware blind image decomposition for real-world adverse weather recovery. In *European Conference on Computer Vision (ECCV)*, pages 379–397. Springer, 2024.

[28] Kaihao Zhang, Wenhan Luo, Yanjiang Yu, Wenqi Ren, Fang Zhao, Changsheng Li, Lin Ma, Wei Liu, and Hongdong Li. Beyond monocular deraining: Parallel stereo deraining network via semantic prior. *International Journal of Computer Vision*, 130(7):1754–1769, 2022.

[29] Zifan Shi, Na Fan, Dit-Yan Yeung, and Qifeng Chen. Stereo waterdrop removal with row-wise dilated attention. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3829–3836. IEEE, 2021.

[30] Tao Yang, Xiaofei Chang, Hang Su, Nathan Crombez, Yassine Ruichek, Tomas Krajnik, and Zhi Yan. Raindrop removal with light field image using image inpainting. *IEEE Access*, 8:58416–58426, 2020.

[31] Teppei Kurita, Yuhi Kondo, Legong Sun, Takayuki Sasaki, Sho Nitta, Yasuhiro Hashimoto, Yoshinori Muramatsu, and Yusuke Moriuchi. Revisiting disparity from dual-pixel images: Physics-informed lightweight depth estimation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 8389–8399. IEEE, 2025.

[32] Hao Yang, Liyuan Pan, Yan Yang, Richard Hartley, and Miaomiao Liu. LDP: Language-driven dual-pixel image defocus deblurring network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24078–24087, 2024.

[33] Kohei Ashida, Hiroaki Santo, Fumio Okura, and Yasuyuki Matsushita. Resolving scale ambiguity in multi-view 3D reconstruction using dual-pixel sensors. In *European Conference on Computer Vision (ECCV)*, pages 162–178. Springer, 2024.

[34] Yuchen Duan, Weiyun Wang, Zhe Chen, Xizhou Zhu, Lewei Lu, Tong Lu, Yu Qiao, Hongsheng Li, Jifeng Dai, and Wenhai Wang. Vision-RWKV: Efficient and scalable visual perception with RWKV-like architectures. In *International Conference on Learning Representations (ICLR)*, pages 83166–83182, 2025.

[35] Haobo Yuan, Xiangtai Li, Lu Qi, Tao Zhang, Ming-Hsuan Yang, Shuicheng Yan, and Chen Change Loy. Mamba or RWKV: Exploring high-quality and high-efficiency segment anything model. *arXiv preprint arXiv:2406.19369*, 2024.

[36] Zhengcong Fei, Mingyuan Fan, Changqian Yu, Debang Li, and Junshi Huang. Diffusion-RWKV: Scaling RWKV-like architectures for diffusion models. *arXiv preprint arXiv:2404.04478*, 2024.

[37] Tiancheng Gu, Kaicheng Yang, Xiang An, Ziyong Feng, Dongnan Liu, Weidong Cai, and Jiankang Deng. RWKV-CLIP: A robust vision-language representation learner. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4799–4812, 2024.

[38] Xijun Wang, Xin Zhou, Yi Wang, Songto Zeng, Xinyu Liu, Haobo Shen, Xianying Wang, Ping Li, and Lei Zhu. RainRWKV: a deep RWKV model for video deraining. *The Visual Computer*, 41:6645–6656, 2025.

[39] Wenjie Cai, Qingguo Meng, Zhenyu Wang, Xingbo Dong, and Zhe Jin. EvRWKV: A continuous interactive RWKV framework for effective event-guided low-light image enhancement. *arXiv preprint arXiv:2507.03184*, 2025.

[40] Zhiwen Yang, Jiayin Li, Hui Zhang, Dan Zhao, Bingzheng Wei, and Yan Xu. Restore-RWKV: Efficient and effective medical image restoration with RWKV. *IEEE Journal of Biomedical and Health Informatics*, 30(1):513–526, 2026.

[41] Rui Xu, Yuzhen Niu, Yuezhou Li, Huangbiao Xu, Wenxi Liu, and Yuzhong Chen. URWKV: Unified RWKV model with multi-state perspective for low-light image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21267–21276, 2025.

[42] Lirong Zheng, Yanshan Li, Rui Yu, and Kaihao Zhang. Fourier-RWKV: A multi-state perception network for efficient image dehazing. *arXiv preprint arXiv:2512.08161*, 2025.

[43] Yan Yang, Liyuan Pan, Liu Liu, and Miaomiao Liu. K3DN: Disparity-aware kernel estimation for dual-pixel defocus deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13263–13272, 2023.

[44] Haofei Xu and Juyong Zhang. AANet: Adaptive aggregation network for efficient stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1959–1968, 2020.

[45] Vera Soboleva and Oleg Shipitko. Raindrops on windshield: Dataset and lightweight gradient-based detection algorithm. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–7. IEEE, 2021.

[46] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5728–5739, 2022.

[47] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *European Conference on Computer Vision (ECCV)*, pages 17–33. Springer, 2022.

[48] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14821–14831, 2021.

[49] Chong Mou, Qian Wang, and Jian Zhang. Deep generalized unfolding networks for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17399–17410, 2022.

[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 (ECCV)*, pages 111–129. Springer, 2024.

[51] Eduard Zamfir, Zongwei Wu, Nancy Mehta, Yuedong Tan, Danda Pani Paudel, Yulun Zhang, and Radu Timofte. Complexity experts are task-discriminative learners for any image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12753–12763, 2025.

[52] Hang Guo, Yong Guo, Yaohua Zha, Yulun Zhang, Wenbo Li, Tao Dai, Shu-Tao Xia, and Yawei Li. MambaIRv2: Attentive state space restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28124–28133, 2025.

[53] Abdullah Abuolaim, Mauricio Delbracio, Damien Kelly, Michael S Brown, and Peyman Milanfar. Learning to reduce defocus blur by realistically modeling dual-pixel data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2289–2298, 2021.

[54] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[55] Jianyi Wang, Kelvin C. K. Chan, and Chen Change Loy. Exploring CLIP for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2555–2563, 2023.

[56] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, et al. SAM 3: Segment anything with concepts. In *International Conference on Learning Representations (ICLR)*, pages 138846–138923, 2026.

[57] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

[58] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[59] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[60] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1874–1883, 2016.

[61] Horia Porav, Tom Bruls, and Paul Newman. I can see clearly now: Image restoration via de-raining. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 7087–7093. IEEE, 2019.

[62] Yudong Han, Yong Wang, Zaiquan Yang, Liang Lin, Chongyang Tao, Xiangxiang Chu, and Liyuan Pan. Beyond entropy: Self-diagnostic multi-role token optimization for video reasoning. *arXiv preprint arXiv:2610.03400*, 2026.

Appendix

In this Appendix, we include the content that the main paper does not show due to space limitations. We provide additional implementation details, more settings for dataset collection, extended ablation studies, and extra experimental results.

A Implementation Details

Degradation Perception Network. Our degradation perception network (DPN) consists of two weight-shared encoders, each implemented as a three-level U-Net [58], where every level contains two residual blocks [59]. We then aggregate the extracted features using a correlation volume [44], computed as:

$$\mathbf{C}(x, y, d) = \langle \mathbf{F}_{\mathrm{L}}(x, y),\, \mathbf{F}_{\mathrm{R}}(x - d, y) \rangle, \tag{8}$$

where $\langle \cdot \rangle$ denotes channel-wise dot-product similarity. Next, the left and right images, together with the disparity features, are fed into a NAFBlock [47] for refinement. Finally, a 3×3 convolution produces the predicted raindrop map $\mathbf{M}$.

Raindrop Synthesis Model. For Eq. (2) in the main paper, we require a lightweight and differentiable raindrop synthesis model. To this end, we first randomly generate a raindrop mask for each clean image and feed the clean image together with the sampled mask into the physical synthesis model [45] to obtain paired raindrop-degraded results. We then train a small U-Net [58] to approximate this physical process by taking the clean image and mask as inputs and using the synthesized rainy image as the supervision signal. A lightweight U-Net [58] is optimized with an $\mathcal{L}_1$ loss so that its synthesized raindrop output closely matches the physical model. After training, this U-Net serves as a differentiable surrogate for the otherwise non-differentiable raindrop synthesis process.

Restoration backbone. Our raindrop removal backbone adopts RWKV as the key unit and follows a 4-level U-shaped multi-scale architecture. Given a degraded input image, the restoration network first applies a 3×3 convolutional projection to obtain shallow feature representations F_0. Starting from F_0, the encoder progressively reduces the spatial resolution while increasing the channel width. The decoder takes the low-resolution latent representation $F_l \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times 8C}$ as input and gradually reconstructs high-resolution features, where each decoder stage is composed of NAFBlocks. For downsampling and upsampling operations, we adopt convolution combined with pixel-shuffle and pixel-unshuffle operations [60]. To further enhance the restoration capability, encoder features are fused with the corresponding decoder features through cross-level skip connections [58]. After concatenation, a convolution layer is applied to halve the channel dimension. Finally, a convolutional projection generates the residual map $R \in \mathbb{R}^{H \times W \times 3}$ from the refined deep representation F_d, which is added to the degraded input image to produce the restored output.

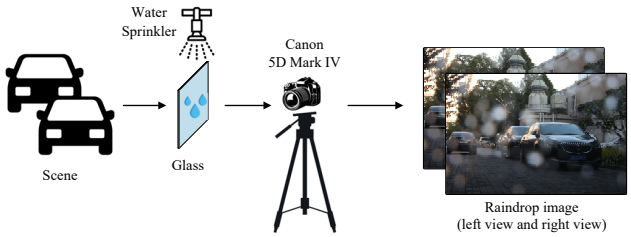


Figure 9: We build a scalable image acquisition pipeline to capture high-quality paired images, consisting of a tripod equipped with a gimbal, an electric water sprayer, and a Canon 5D Mark IV DSLR camera. The camera shutter and all shooting parameters are remotely controlled via a smartphone.

Table 7: Summary of existing raindrop removal datasets. *Multi-distance* indicates whether the dataset accounts for varying raindrop-to-camera distances during capture; *Real* specifies whether the training set contains real-world images; and *Day/Night* denotes the time of day at which the scenes were captured.

Datasets	Scales	Year	Dual-Pixel	3-channel color	Multi-distance	Real?	Day/Night
Raindrop [8]	919	2018	×	✓	×	✓	Day
RainDS-Real [10]	248	2021	×	✓	×	✓	Day
RainDS-Syn [10]	1,200	2021	×	✓	×	×	Day
RobotCar Porav [61]	4,818	2019	×	✓	×	✓	Day
Windshield [45]	3,390	2020	×	✓	×	✓	Day
RD Clarity [25]	15,186	2024	×	✓	×	✓	Day&Night
DPRRN-V [21]	2,452 + 125	2022	✓	×	×	×	Day
DPRRN-T [18]	2,452 + 130	2024	✓	×	×	×	Day
DNDP (Our)	6,800	2025	✓	✓	✓	✓	Day&Night

Table 8: Ablations of the loss functions. The best results are highlighted in bold.

Loss Components				PSNR↑	SSIM↑
$\mathcal{L}_{\text{char}}$	$\mathcal{L}_{\text{res}}$	$\mathcal{L}_{\text{dp}}$	$\mathcal{L}_{\text{syn}}$		
✓	✗	✗	✗	24.97	0.8366
✓	✓	✗	✗	25.38	0.8451
✓	✓	✓	✗	25.49	0.8468
✓	✓	✓	✓	**25.64**	**0.8480**

B Dataset Collection Setting

Equipment. Our image acquisition system is built upon a Canon EOS 5D Mark IV camera equipped with a dual-pixel sensor and paired with a Canon EF 24–70mm f/2.8L II USM lens. We use an electric water sprayer and adjust its power to produce water droplets of varying sizes and densities. The camera is mounted on a tripod with a gimbal and controlled via a smart remote shutter to ensure precise and stable shooting. To simulate different visual system configurations, we employ glass plates made from various materials, including quartz glass and acrylic panels. All image pairs were captured with a fixed tripod, fixed camera parameters, and static scenes. We further performed pair-level quality control to remove samples with visible misalignment, illumination inconsistency, or scene motion. For each candidate pair, we inspected both global alignment and local background consistency in non-raindrop regions.

Parameters. To cover a wide range of real-world application scenarios, our data acquisition process employs variable focal lengths and apertures, spanning a focal range of 24 mm to 70 mm and an aperture range of f/2.8 to f/22. The ISO is fixed at 100, and the exposure compensation is set to 0, while the shutter speed is determined by the camera's built in metering system. In addition, we simulate diverse raindrop conditions by capturing images under varying illumination from morning to evening and with different raindrop densities. The aperture and camera-to-glass distance were continuously varied during data collection to produce diverse raindrop shapes and appearances. The training and testing sets consist of nonoverlapping scenes. All raindrop variations from a given scene and viewpoint are assigned to the same split. For compatibility with most DP raindrop removal methods, we downsample the image resolution from 6720×4480 to 1680×1120. Nevertheless, we retain the original high resolution images for optional use.

Comparison with Other raindrop Datasets. As shown in Tab. 7, most existing raindrop removal datasets are not captured using dual-pixel sensors. Only Li *et al*. [21, 18] utilize a Dual-Pixel-based raindrop dataset. However, their dataset is trained solely on synthetic data, lacks multi-channel color images, and does not include raindrop samples captured at different distances or with diverse raindrop patterns.

Table 9: Ablation studies on hyperparameter settings. The best results are highlighted in bold.

(a) Ablation on λ_1

λ_1	PSNR$_\uparrow$	SSIM$_\uparrow$
0.05	25.60	0.8472
0.1	**25.64**	**0.8480**
0.2	25.61	0.8474

(b) Ablation on λ_2

λ_2	PSNR$_\uparrow$	SSIM$_\uparrow$
0.05	25.57	0.8471
0.1	**25.64**	**0.8480**
0.2	25.51	0.8464

(c) Ablation on λ

λ	PSNR$_\uparrow$	SSIM$_\uparrow$
0.1	25.55	0.8475
0.2	**25.64**	**0.8480**
0.5	25.48	0.8460

Table 10: Effectiveness of Raindrop maps. The best results are highlighted in bold.

Variant	PSNR$_\uparrow$	SSIM$_\uparrow$
i) w/o Raindrop map	24.88	0.8324
ii) Learnable gate	24.97	0.8366
iii) Concat	25.21	0.8430
iv) Ours	**25.64**	**0.8480**

C Ablation Studies and Discussions

Effectiveness of Losses. We investigate the effectiveness of each loss component used in our framework. When only $\mathcal{L}_{\text{char}}$ is applied, the model degenerates into a baseline variant that removes the DPN and replaces our modulation mechanism with a learnable gate. As shown in Tab. 8, using the full set of losses leads to the highest restoration performance. This consistently superior result demonstrates that each loss term plays a meaningful and complementary role, jointly contributing to the overall effectiveness of our method.

Loss Hyperparameter Analysis. We conduct ablation studies to analyze the influence of loss weights λ_1, λ_2, and λ. As shown in Table 9, the best performance is achieved with $\lambda_1 = 0.1$, $\lambda_2 = 0.1$, and $\lambda = 0.2$, yielding 25.64 dB PSNR and 0.8480 SSIM. Smaller or larger values of λ_1 and λ_2 lead to slight performance drops, indicating that properly balancing different loss terms is important for stable optimization. Similarly, $\lambda = 0.2$ outperforms both $\lambda = 0.1$ and $\lambda = 0.5$, suggesting that the selected configuration provides a better trade-off among loss components.

Effectiveness of Raindrop Map. We investigate the effectiveness of using the raindrop map as a prior for raindrop removal in Tab. 10. To demonstrate the benefit of incorporating the raindrop map, we evaluate four baseline variants: (i) w/o Raindrop Map, which removes the raindrop-map input; (ii) Learnable gate, which replaces the RWKV gates with learnable parameters; (iii) Concat, which concatenates the raindrop map with the DP pair as input; (iv) Ours, which uses the raindrop map $\mathbf{M}$ to modulate the RWKV gates. Our model consistently outperforms all baselines, showing that incorporating the raindrop map as a prior is highly effective.

Cross-dataset and Cross-device Generalization. To evaluate transfer beyond DNDP, we train all models only on DNDP and directly test them on DPRRN Real-Vertical and DPRRN Real-Tilted, without fine-tuning. The training images are captured with an EOS 5D Mark IV, while the DPRRN test sets come from a Google Pixel camera. As shown in Tab. 12, DRWKV achieves the best PSNR and SSIM on both targets.

Model Complexity. Table 11 compares DPRRN, DRWKV, and MoCE-IR on DNDP. DRWKV achieves 25.64 dB with 17.11M parameters, 1,922G FLOPs, and 0.41 s runtime. Compared with MoCE-IR, it improves PSNR by 1.29 dB while using fewer parameters and lower FLOPs. The matched-budget results in Tabs. 6 and 17 provide a separate controlled comparison.

Table 11: Expanded computational comparison on DNDP. Runtime is measured on an RTX 4090. FLOPs are reported in G ($1\,\mathrm{T} = 1000\,\mathrm{G}$). These are the full-model comparisons rather than the 24-hour matched-budget experiments.

Method	Params (M)	FLOPs (G)	Runtime (s)	PSNR↑
DPRRN [18]	19.48	7,073	0.87	24.36
DRWKV	**17.110**	**1,922**	**0.41**	**25.64**
MoCE-IR [51]	25.354	5,684	2.35	24.35

Table 12: Zero-shot cross-dataset and cross-device evaluation. All models are trained on DNDP without target-domain fine-tuning. Entries are PSNR (dB) / SSIM.

Method	DPRRN Real-Vertical	DPRRN Real-Tilted
MambaIRv2 [52]	29.01 / 0.852	30.60 / 0.855
DPRRN [18]	29.05 / 0.886	30.85 / 0.863
DRWKV	**30.20 / 0.894**	**31.62 / 0.875**

Run-to-run Stability. We independently train DPRRN and DRWKV with three random seeds while keeping the data split and training settings fixed. Table 13 reports the mean and standard deviation of PSNR, SSIM, and CLIP-IQA+ across the three complete runs. DRWKV achieves 25.65 ± 0.04 dB PSNR, compared with 24.34 ± 0.13 dB for DPRRN. These aggregate results are distinguished from the single-run entries in Tab. 2. The variation reflects the random training runs, not uncertainty across individual test images. The reported standard deviations are not confidence intervals or formal significance tests.

Table 13: Mean $\pm$ standard deviation across three independently trained random seeds on DNDP. Higher is better for all metrics.

Method	PSNR	SSIM	CLIP-IQA+
DPRRN [18]	24.34 ± 0.13	0.8287 ± 0.0071	0.4070 ± 0.0056
DRWKV	$\mathbf{25.65 \pm 0.04}$	$\mathbf{0.8479 \pm 0.0007}$	$\mathbf{0.4173 \pm 0.0015}$

Downstream Detection and Segmentation. We select 90 road/outdoor images from the DNDP test set through stratified sampling and manually annotate bounding boxes and instance masks on their clean ground-truth images. The four categories are tree, bicycle, car, and motorcycle. We apply the same frozen SAM 3 [56] with category-name prompts to all rainy and restored inputs, without fine-tuning on DNDP. Table 14 reports detection and segmentation metrics on a percentage scale. DRWKV improves Box mAP from 66.50 to 76.49 and Mask mAP from 66.81 to 76.96, gains of 9.99 and 10.15 percentage points over the rainy inputs. These results support downstream utility on this annotated subset.

Table 14: Downstream evaluation on 90 annotated DNDP road/outdoor images using the same frozen SAM 3. All values are percentages and higher is better.

Input	Detection F1	Box AP50	Box mAP	Mask mAP	Mask mIoU	Mask Dice
Rainy input	79.26	71.38	66.50	66.81	76.33	85.98
MambaIRv2	81.67	73.81	67.21	68.32	77.49	87.06
DPRRN	80.83	70.78	64.90	66.05	79.04	88.27
DRWKV	**86.71**	**82.20**	**76.49**	**76.96**	**86.63**	**92.81**

Scanning Configuration Ablation. With recurrence depth fixed at $P = 2$, we compare horizontal (H) and vertical (V) scanning sequences in Tab. 15. The tested sequences span only 0.06 dB PSNR, indicating limited sensitivity within these settings. These small differences alone do not establish a statistically significant ordering among the variants. We also evaluate recurrence depths $P \in \{1, 2, 4\}$ in Tab. 16. Increasing P from 1 to 2 improves PSNR by 0.32 dB with a runtime increase from 0.36 to 0.41 s. Increasing P to 4 provides only a further 0.04 dB while increasing runtime to 1.03 s. We therefore use $P = 2$ with $H \rightarrow V$ scanning as the default configuration to balance restoration quality and runtime.

Table 15: Scanning-order ablation at fixed recurrence depth $P = 2$.

Scanning order	PSNR↑	SSIM↑
$H \to H$	25.61	0.8472
$V \to V$	25.58	0.8470
$V \to H$	25.63	0.8475
$H \to V$ (default)	**25.64**	**0.8480**

Table 16: Recurrence-depth ablation. All other model settings are unchanged. The default setting is $P = 2$.

P	Scanning sequence	PSNR↑	SSIM↑	Runtime (s)↓
1	H	25.32	0.8429	**0.36**
2	$H \to V$	25.64	0.8480	0.41
4	$H \to V \to H \to V$	**25.68**	**0.8487**	1.03

Inference-FLOP-matched Comparison. In a separate controlled experiment, we resize MambaIRv2 to approximately 1,922G inference FLOPs, matching DRWKV, and train both models for 24 hours using the same hardware and training settings. DRWKV achieves 24.67 dB, compared with 23.16 dB for the FLOP-matched MambaIRv2 (Tab. 17). This is a separate setting from the parameter-matched comparison in Tab. 6. Matching inference FLOPs does not imply matching parameter counts or training FLOPs.

Table 17: Inference-FLOP-matched comparison with a 24-hour training budget.

Architecture	Inference FLOPs (G)	Training (h)	PSNR↑
MambaIRv2 (FLOP-matched)	$\approx$ 1,922	24	23.16
DRWKV	$\approx$ 1,922	24	**24.67**

Future Work and Limitation. When raindrops densely accumulate on the lens or glass surface, they can occlude large portions of the underlying objects, causing severe loss of scene texture and structural information and making image restoration more difficult. Future work will explore joint optimization of image restoration and downstream perception tasks, extend applications to object detection, image segmentation, object tracking, and video reasoning [62], and investigate more challenging scenarios involving dense raindrop occlusion and complex optical degradations to improve the applicability of the method in real-world environments.

Our paired training data are currently collected primarily in controlled static scenes using a single primary camera. Future work will expand the range of capture devices, scene types, and lighting conditions to enhance data diversity and generalization. The current downstream evaluation covers only four object categories. We will include more categories and complex scenes, further investigate how similarities in disparity and appearance between extremely defocused near-field objects and raindrops affect degradation estimation, and conduct more comprehensive experiments to assess the stability and reliability of the method.

D Impact Statement

Our DRWKV framework and DNDP benchmark aim to improve image visibility under raindrop degradation. This can benefit safety-critical applications such as autonomous driving, UAV inspection, outdoor robotics, and traffic surveillance, where clearer visual inputs may improve downstream perception reliability. The proposed real-world dual-pixel benchmark may also support more realistic evaluation of weather-robust vision systems. However, enhanced image restoration could be misused in large-scale surveillance, raising potential privacy concerns without proper legal and ethical oversight. In addition, training deep restoration models requires computational resources and may contribute to energy consumption. Future work should explore more efficient models and responsible deployment protocols.



Figure 10: Our dataset contains various raindrop patterns captured in both daytime and nighttime conditions. From the first row to the fourth row, we present different raindrop appearance patterns, while the final row shows the corresponding ground-truth images.

E Additional Results

More Examples of the DNDP Dataset. Fig. 10 presents additional examples from the DNDP dataset. Our dataset comprises images with varying degrees of degradation across multiple indoor and outdoor scenes, where each degraded image has a corresponding ground truth image.

Results on Syn. (Tilted and Vertical) Dataset. We provide additional qualitative comparison results of SOTA DP raindrop removal methods, as shown in Fig. 11.

Results on Real (Tilted and Vertical) Dataset. We provide additional qualitative comparison results of SOTA DP raindrop removal methods, as shown in Fig. 12.

Results on DNDP Dataset. We provide additional qualitative comparison results of SOTA DP raindrop removal methods, as shown in Fig. 13.

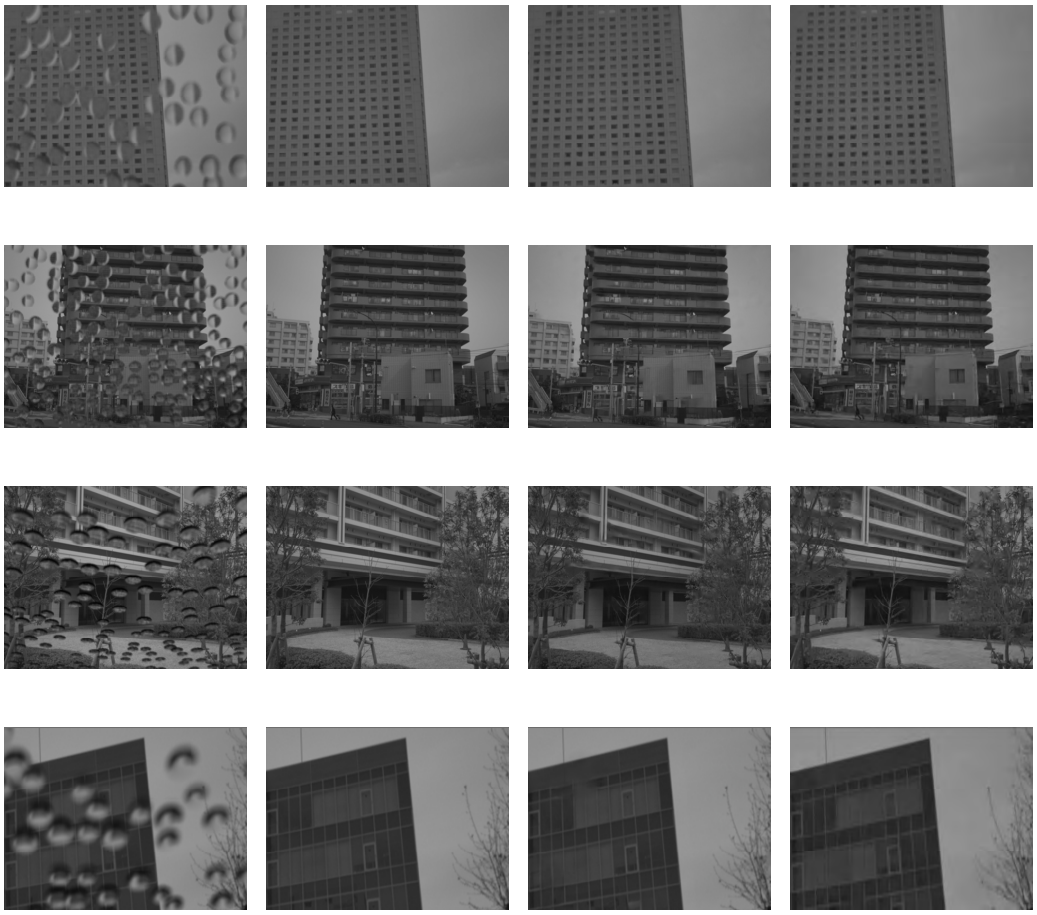


Figure 11: Qualitative evaluation on the Syn. (Vertical) and Syn. (Tilted) datasets proposed by Li *et al*. [18]. The first column shows the raindrop-degraded inputs, the second column presents the ground-truth clean images, and the subsequent columns display the restoration results of the current state-of-the-art DP-based raindrop removal method and our approach.

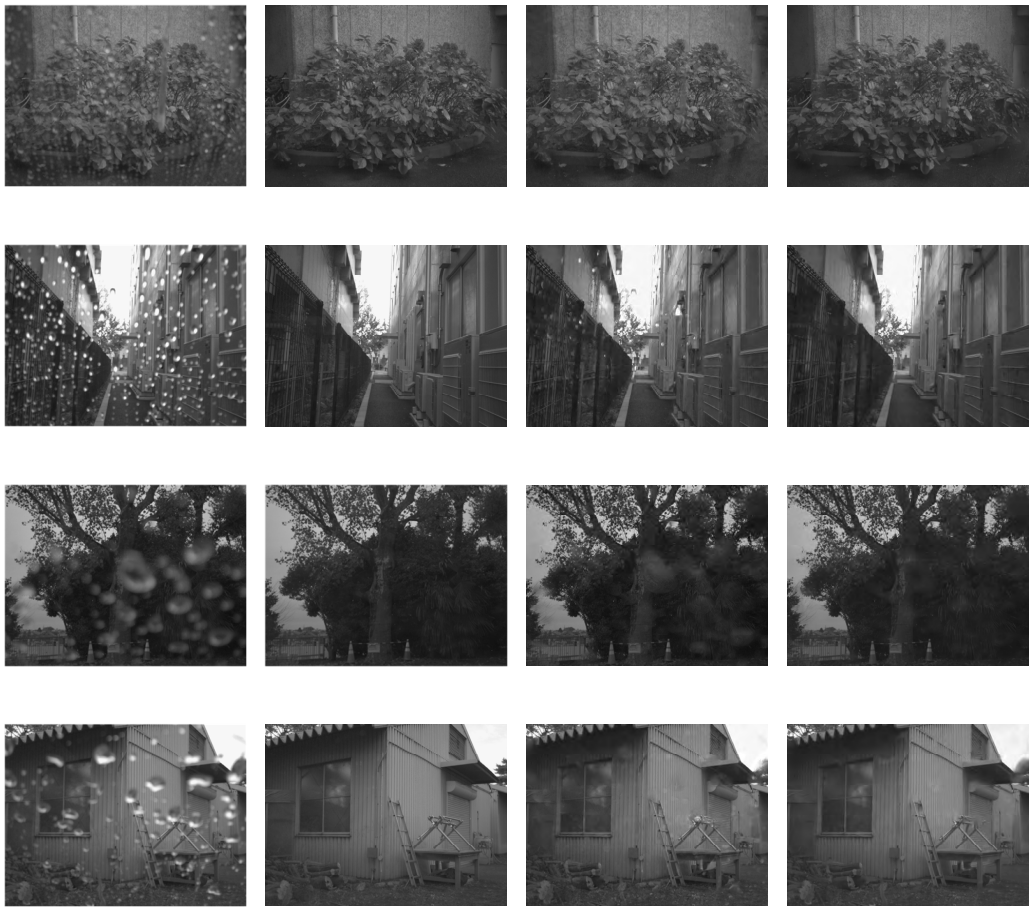


Figure 12: Qualitative evaluation on the Real (Vertical) and Real (Tilted) datasets proposed by Li *et al*. [18]. The first column shows the raindrop-degraded inputs, the second column presents the ground-truth clean images, and the subsequent columns display the restoration results of the current state-of-the-art DP-based raindrop removal method and our approach.



Figure 13: Qualitative evaluation on our DNDP dataset. The first column shows the raindrop-degraded inputs, the second column presents the ground-truth clean images, and the subsequent columns display the restoration results of the current state-of-the-art DP-based raindrop removal method and our approach.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Relevant information is included in Sec. 1

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Relevant information is included in Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Relevant information is included in Sec. 3, Sec. 4 and Sec. 5.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We will release our code after review, but we have already provided a detailed explanation of how to implement our algorithm and the specific implementation details in the paper.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Relevant information is included in Sec. 3 and Sec. 5

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Table 13 reports the mean and standard deviation of PSNR, SSIM, and CLIP-IQA+ across three independently trained random seeds for DPRRN and DRWKV under the same data split and training settings. These are run-to-run standard deviations, not confidence intervals.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Compute resources are reported in the experimental setup and appendix, including the separate 24-hour matched-budget comparisons.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in this paper complies with the NeurIPS Code of Ethics. in every respect.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The broader impacts are discussed in the Appendix D.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite all the work covered in this article and follow their license.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: This paper introduces DNDP, a new real-world dual-pixel raindrop removal dataset. The dataset will publicly released upon acceptance.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subject.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were used only for writing, editing, and formatting assistance, and did not contribute to the core methodology, data annotation, data filtering, experimental design, or scientific conclusions.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.