AutoWeather4D: Autonomous Driving Video Weather Conversion via G-Buffer Dual-Pass Editing

Tianyu Liu[1,2*] Weitao Xiong[1,3*] Kunming Luo[1,2]
Manyuan Zhang[4] Peng Li[1] Yuan Liu[1†] Ping Tan[1,2†]

[1]The Hong Kong University of Science and Technology
[2]Shenzhen Loop Area Institute
[3]Xiamen University [4]Meituan-M17, Hong Kong
[*]Equal Contribution [†]Equal Corresponding Author

Abstract. Generative video models have significantly advanced the photorealistic synthesis of adverse weather for autonomous driving; however, they consistently demand massive datasets to learn rare weather scenarios. While 3D-aware editing methods alleviate these data constraints by augmenting existing video footage, they are fundamentally bottlenecked by costly per-scene optimization and suffer from inherent geometric and illumination entanglement. In this work, we introduce AutoWeather4D, a feed-forward 3D-aware weather editing framework designed to explicitly decouple geometry and illumination. At the core of our approach is a G-buffer Dual-pass Editing mechanism. The Geometry Pass leverages explicit structural foundations to enable surface-anchored physical interactions, while the Light Pass analytically resolves light transport, accumulating the contributions of local illuminants into the global illumination to enable dynamic 3D local relighting. Extensive experiments demonstrate that AutoWeather4D achieves comparable photorealism and structural consistency to generative baselines while enabling fine-grained parametric physical control, serving as a practical data engine for autonomous driving.

Keywords: Autonomous Driving · Simulation · Vision + Graphics

1 Introduction

Recent advances in generative video models [1, 2, 29, 57, 83, 84] represent an important step towards the photorealistic synthesis of adverse weather conditions for autonomous driving. However, despite their impressive visual fidelity, these data-driven approaches consistently demand massive datasets to learn rare adverse weather patterns. Capturing such long-tail environmental data in the real world remains prohibitively expensive and logistically constrained.

To circumvent these data constraints, 3D-aware editing methods [10,27,44,51] offer a compelling alternative by augmenting existing video footage. By explicitly

Fig. 1: AutoWeather4D: Weather & Time-of-Day Control for Driving Videos. AutoWeather4D enables fine-grained control over weather (rain, snow, fog) and time-of-day (dawn, noon, night) in driving videos. The red zoom-in box showcases realistic snow accumulation, while the blue zoom-in box highlights rain-induced road wetness and ripples. See supplementary videos for dynamic visualizations.

grounding the synthesis process in 3D space, these approaches achieve high-fidelity and highly controllable weather effects without relying on massive, long-tail training datasets. They typically operate through a straightforward two-stage pipeline: first reconstructing a 3D representation of the captured scene, and subsequently applying weather-specific modifications to the underlying geometry and appearance. However, these methods are fundamentally bottlenecked by their reliance on painstakingly slow per-scene optimization. Requiring up to an hour of computation per video clip, this optimization paradigm is computationally prohibitive for large-scale data generation.

In this paper, we propose a 3D-aware editing method called AutoWeather4D, which brings the controllability and visual quality of 3D-aware editing to dynamic autonomous driving scenarios. By replacing the sluggish per-scene optimization with a novel feed-forward editing pipeline that explicitly decouples geometry and illumination, AutoWeather4D achieves rapid, high-quality, and physically plausible weather editing and lighting editing.

Designing such a framework is non-trivial, which first requires defining an editable and flexible 3D scene representation for the dynamic autonomous driving scenes. Existing 3D-aware editing pipelines heavily rely on scene representations like NeRF [36] or 3DGS [23]. However, a fundamental limitation of these frameworks is their inherent reliance on static scene assumptions for high-quality reconstruction. When confronted with the complex, highly dynamic environments typical of autonomous driving with moving vehicles and pedestrians, these optimized fields frequently fail to capture accurate underlying 3D geometry. Con-

Method	Env light/shadow control	Extra light source	Weather change	Feed-forward	4D dynamic scene	Tuning-free	Open-source
Cosmos-Transfer2.5 [1]	×	×	✓	✓	✓	✓	✓
WAN-FUN 2.2 [57]	×	×	✓	✓	✓	✓	✓
Ditto [2]	×	×	✓	✓	✓	✓	✓
WeatherWeaver [29]	×	×	✓	✓	✓	×	×
WeatherDiffusion [83]	×	×	✓	✓	×	×	×
SceneCrafter [84]	×	×	✓	✓	✓	×	×
RainyGS [10]	×	×	✓	×	✓	✓	×
WeatherEdit [44]	×	×	✓	×	✓	✓	✓
ClimateNeRF [27]	×	×	✓	×	×	✓	✓
DiffusionRenderer [28]	✓	×	×	✓	✓	✓	✓
AutoWeather4D(Ours)	✓	✓	✓	✓	✓	✓	✓

Table 1: Comparison of state-of-the-art weather and time-of-day synthesis models for autonomous driving. We evaluate existing paradigms across key capabilities: Env light/shadow control (arbitrarily adjusting environment light directions and correcting shadows), Extra light source (precisely adding and controlling local lights within driving scenes), and Weather change (synthesizing appearances under diverse weather conditions). Additionally, we compare their architectural and practical properties, including whether they are Feed-forward (requiring no per-scene optimization), applicable to 4D dynamic scenes, Tuning-free (requiring no extra datasets for weight fine-tuning).

sequently, spatially anchoring and consistently applying weather effects across dynamic elements becomes exceptionally difficult.

To address this, our method represents the dynamic scene by the extracted G-buffers of the videos with a feed-forward neural network [28]. By directly predicting dense, frame-wise geometric features (such as depth and normals) from the video stream, we entirely bypass the static-scene bottleneck of per-scene optimization. This explicit G-buffer formulation natively accommodates dynamic objects and provides a highly controllable, reliable structural foundation, making downstream weather editing both intuitive and geometrically precise.

Second, building upon our explicit geometric foundation, we address the severe illumination entanglement that plagues current weather editing paradigms. Existing 3D-aware methods typically assume a static, single global illumination setup, fundamentally baking the original scene's appearance and lighting directly into the optimized 3D representation. While this may suffice for static landscapes, it completely breaks down in dynamic autonomous driving environments. In these complex scenarios, realistic weather synthesis inherently requires modeling dynamic local lighting, such as moving vehicle headlights sweeping across wet surfaces or streetlights creating volumetric halos in the fog. To break this architectural barrier, AutoWeather4D introduces a fully decoupled Light Pass. By integrating physics-based lighting priors with our G-buffer-driven neural rendering, our framework thoroughly separates the global atmospheric conditions from localized, dynamic illuminants. This explicit decoupling unlocks the unprecedented ability to seamlessly insert, toggle, and physically relight 3D local sources under adverse weather conditions, ensuring that both static and dynamic elements react accurately to environmental changes (See Fig. 1).

Extensive experiments on standard autonomous driving datasets demonstrate that AutoWeather4D synthesizes adverse weather and illumination conditions from existing footage, without requiring any auxiliary data. As summarized in Tab. 1, compared to existing paradigms, our framework uniquely achieves

decoupled control over geometric weather elements and global/local light transport without the need for per-scene optimization.

In summary, our main contributions are:

- We introduce AutoWeather4D, a feed-forward 3D-aware weather editing framework. It synthesizes adverse weather conditions from real-world driving videos while eliminating the need for per-scene optimization.
- We propose a G-buffer Dual-pass Editing mechanism. The Geometry Pass enables surface-anchored interactions (e.g., snow accumulation), while the Light Pass analytically accumulates local illuminants for 3D relighting.
- Experiments demonstrate that AutoWeather4D achieves comparable photorealism and structural consistency to generative baselines while enabling parametric physical control, serving as a practical data engine for autonomous driving.

2 Related Works

In this section, we review two related topics. First, we discuss advances in Climate simulation. Next, we discuss video simulator for autonomous driving.

2.1 Climate simulation

Physical based simulator: Classical computer graphics have long established the physical foundations for rendering weather effects like rain [46,61,62], snow [39, 41,63], and volumetric fog [4,22,42] using particle systems and scattering equations. However, these classical methods fundamentally rely on explicit 3D meshes or voxel grids and cannot be directly applied to monocular videos. Our method seamlessly integrates these classical physical priors into real-world video footage, preserving physical guarantees while enabling flexible video-based editing.

Network-based simulator: The advent of deep learning has enabled data-driven approaches to climate and weather simulation. In the image domain, early work applied CycleGAN [82] for weather transfer [53], while diffusion models such as Prompt-to-Prompt [18] and SDEdit [35] enabled weather modification via text prompts or sketch-based guidance. Recent methods target illumination control specifically: LightIt [24] conditions generation on one-bounce shadow maps; Retinex-Diffusion [74] reformulates the energy function of diffusion models to fulfill illumination alteration; DiLightNet [78] and IntrinsicAnything [7] decompose images into BRDF components for relighting; and IC-Light [80] adjusts illumination based on reference backgrounds. Video extensions include fine-tuning-based editors (WeatherWeaver [29], WeatherDiffusion [83], SceneCrafter [84], Ditto [2]), ControlNet-style conditioning methods (WAN-FUN 2.2 [57], Cosmos-Transfer2.5 [1]), and G-buffer decomposition approaches (DiffusionRender [28]). However, existing methods address either weather conversion or physically-based illumination control, but not both simultaneously—a critical limitation that our work resolves through unified physical rendering and diffusion-based synthesis.

Hybrid physics-and-learning simulators. Recent methods integrate classical graphics with deep learning across diverse 3D representations. NeRF-based

approaches [27] embed physical weather models or text-guided editing into neural radiance fields for high-fidelity rendering of atmospheric effects (fog, snow, flooding), though limited to static scenes. Mesh-based techniques [71, 85] convert NeRF [36] reconstructions into interactive meshes with rigid-body physics for real-time interaction. Gaussian Splatting methods leverage 3DGS [23] for efficient rendering: GaussianEditor [59]) enable cross-view 2D-to-3DGS manipulation, RainyGS [10] models physical raindrops, Weather-Magician [51] incorporates depth/normal supervision for multi-weather synthesis, DRAWER [72] combine 3DGS with mesh for articulated things, and WeatherEdit [44] extends to 4DGS for temporal control. Unlike these per-scene optimization reconstruction approaches, our method employs feed-forward 4D reconstruction [58, 64, 66], which—despite producing sparser outputs that pose additional challenges—eliminates scene-specific tuning and drastically reduces deployment time.

2.2 Autonomous driving video simulator

Autonomous driving world model simulators [15, 16, 21, 25, 26, 31, 32, 40, 43, 48, 60, 67, 68, 76, 81, 84] play a crucial role in generating complex traffic scenarios that are challenging to capture in real-world conditions, substantially reducing data collection costs for training self-driving systems—particularly benefiting end-to-end autonomous driving. Unlike prior approaches that rely on iterative neural optimization or latent-space manipulation, our method leverages classical graphics techniques by directly operating on the G-buffer, enabling explicit geometric and illumination control for efficient scene modification, which is absent in existing simulators.

3 Method

Overview. As shown in Fig. 2, we formulate weather editing as an efficient *analysis-and-synthesis* pipeline. The *analysis* stage decomposes the input video into explicit intrinsic G-buffers (Sec. 3.1) in a feed-forward manner, bypassing the prohibitive cost of per-scene optimization. The subsequent *synthesis* stage manipulates the decoupled scene geometry and illumination via a Dual-pass Editing mechanism (Sec. 3.2). Finally, the VidRefiner performs terminal refinement on the rendered sequence, incorporating sensor nuances while conditioning the generative process on the resolved physical dynamics (Sec. 3.3). All stages in this pipeline are fully automated.

3.1 Feed-Forward G-Buffer Extraction

Feed-forward Intrinsic Parsing. To bypass the static-scene assumptions of implicit optimization and ensure accurate geometric anchoring in dynamic environments, the monocular sequence is parsed into a unified G-buffer through a multi-source feed-forward extraction scheme. Initially, spatiotemporally coherent relative depth is instantiated by deploying Pi3 [66], a feed-forward 4D reconstruction backbone. Alongside this geometric extraction, intrinsic material properties (albedo, normal, metallic, roughness) are decoupled via a zero-shot diffusion-based inverse renderer [28]. Consolidating these multi-source extractions yields

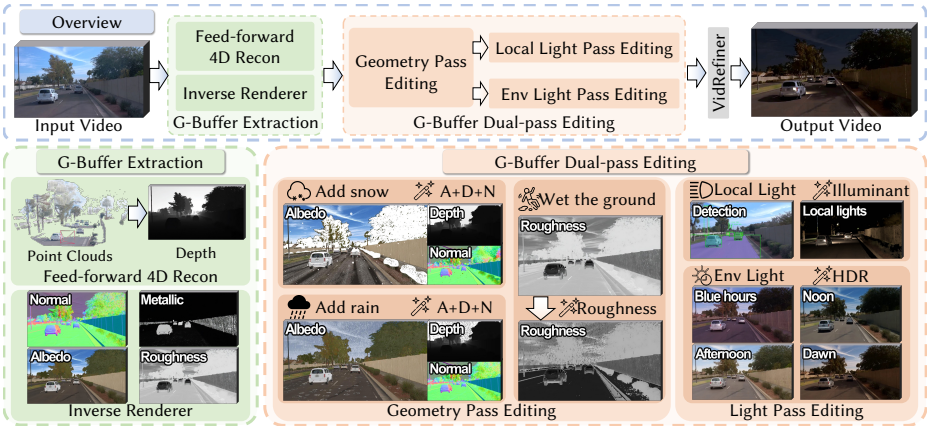

Fig. 2: Overview of our framework. The pipeline formulates physically-grounded video editing for multi-weather and time-of-day synthesis. We first extract explicit G-buffers from the input video: metric depth **D** via feed-forward 4D reconstruction, alongside intrinsic material properties (normal **N**, metallic **M**, albedo **A**, roughness **R**) via an inverse renderer. The scene modifications are analytically resolved through the **G-Buffer Dual-Pass Editing**: (1) The *Geometry Pass* physically modulates $\mathbf{A}, \mathbf{N}, \mathbf{R}$ to instantiate explicit weather mechanics (e.g., snow, rain, ground wetness); (2) The *Light Pass* executes parametric illumination control, independently synthesizing detected local light sources and global environmental lighting (e.g., dawn, noon, blue hours) to reflect atmospheric and temporal shifts. Finally, the deterministic rendered sequence is processed by the **VidRefiner**. This terminal refiner synthesizes real-world sensor nuances while preserving the classical shading cues and explicit scene dynamics resolved in the dual-pass stages.

a preliminary state for downstream editing processing, necessitating absolute metric scale depth and spatial bounding to ensure physical validity.

Relative Depth Alignment and Metric G-buffer Construction. The relative scale of the reconstructed geometry conflicts with the metric requirements of physical light transport. To obtain a physically meaningful scale, we align the relative depth predicted by the feed-forward 4D reconstruction backbone to sparse LiDAR point clouds sampled from non-sky and non-occluded regions, solving for a global scale and bias. For strictly monocular configurations without LiDAR, we instead recover the scale from a camera-height prior by fitting a road plane on the predicted 3D points [11]. After calibration, each frame t is represented as a pixel-aligned G-buffer $\mathcal{G}_t = \{D_t, N_t, A_t, R_t, M_t, S_t\}$, where D_t is the metric depth, N_t the surface normal, A_t the albedo, R_t the roughness, M_t the metallic map, and S_t the valid-scene/sky mask. This metric G-buffer provides the shared physical substrate for distance-dependent effects, including local light attenuation, volumetric fog, and depth-tested rain/snow particles.

Sky-Aware Material Extraction. Sky regions often produce unstable depth estimates due to their weak texture and effectively infinite distance. We therefore estimate a sky mask S_t for each frame and clamp sky depth to the 99th percentile of the non-sky depth distribution over the sequence [73]. This

prevents artificial illumination from being incorrectly projected onto the sky while preserving a stable depth range for downstream light transport. The inverse renderer is applied to valid scene regions to estimate (A_t, N_t, R_t, M_t), ensuring pixel-level correspondence between material maps and metric geometry.

The implementation details of alignment and sky-aware material extraction are provided in (Sec. 6 in supplementary materials).

3.2 G-Buffer Dual-pass Editing

To extend high-fidelity 3D-aware editing to dynamic driving scenarios, we propose a Dual-Pass Editing mechanism. This pipeline systematically decouples structural scene modifications from illumination transport: the **Geometry Pass** first updates the intrinsic state of the scene, which then serves as the physical foundation for the **Light Pass** to analytically resolve radiance. By operating on explicit G-buffers, this mechanism ensures that all synthesized environmental changes remain anchored to the underlying 3D structure.

Geometry Pass: Surface-Anchored Interaction The Geometry Pass transforms the intrinsic albedo, normal, and roughness to incorporate the physical presence of weather elements. These updated surface descriptors parameterize the subsequent Light Pass, ensuring all illumination transport is analytically resolved over the modified scene structure. Specifically, we instantiate these surface-anchored modifications through explicit physical models for two representative weather conditions:

Multi-Representation Snow Synthesis. To bridge the scale gap between individual snowflakes and terrain-scale coverage, we employ a hybrid simulation: (1) *Metaball-based Surface Buildup* iteratively evaluates an SPH Poly6 kernel [38] over the extracted normal maps, restricting accumulation to upward-facing structures to maintain geometric rational; (2) *Grid-based Ground Modeling* utilizes procedural patterns for varied snow density alongside a physically-based wetness model that darkens albedo and reduces roughness to simulate thawing transitions; and (3) *Kinematic Falling Particles* are rendered via temporally-persistent screen-space rasterization to ensure inter-frame kinematic continuity. Implementation details are provided in (Sec. 7 of the supplementary material).

Physically-Grounded Rain Dynamics. We decouple rain synthesis into kinematic streaks and standing water. Falling drops are modeled as kinematic particles governed by a vector summation of Gunn–Kinzer terminal velocities [69] (vertical gravity-drag equilibrium) and parametric wind fields (horizontal displacement). We parameterize these trajectories as volumetric Signed Distance Fields (SDFs), explicitly depth-testing against the extracted depth to enforce precise spatial occlusion. For ground interactions, puddle masks generated via Fractional Brownian Motion (FBM) physically modulate the local albedo and roughness. Concurrently, surface normals within these masked regions are perturbed using procedural ripple maps to approximate dynamic impact responses. Implementation details are provided in (Sec. 8 of the supplementary material).

Light Pass: Decoupled Illumination Control Given the updated G-buffers from the Geometry Pass, the Light Pass computes the final scene illumination. By operating directly on these explicit material properties, we can independently synthesize local light sources and global atmospheric scattering, enabling direct parametric relighting. Specifically, we instantiate this parametric relighting through tailored physical models for three representative illumination scenarios:

Nocturnal Local Relighting. We explicitly model artificial sources (e.g., streetlights, headlights) as 3D spotlights, estimating their spatial positions via semantic masks and the metric depth. Surface radiance is then analytically evaluated using the Cook-Torrance BRDF [8], which is directly parameterized by the edited G-buffers to enforce physically consistent material responses. For non-illuminated regions, a parametric Look-Up Table (LUT) shifts ambient color temperatures toward warm nocturnal tones to maintain minimal visibility. Light sources estimation details are provided in (Sec. 9 of the supplementary materials), while the BRDF and LUT implementation details are provided in (Sec. 10 of the supplementary materials).

Local light detection, localization, and activation. We instantiate vehicle lights in a state-aware manner rather than globally brightening the scene. Vehicles are first detected and tracked with an off-the-shelf 3D detector [13] and tracker [65] using the calibrated metric depth. For each tracked 3D cuboid, headlights and taillights are localized at fixed relative positions on the front and rear faces of the cuboid, respectively. The light direction is aligned with the vehicle heading, and its spatial extent is rendered as a depth-tested 3D spotlight cone. To avoid activating implausible lights, we only turn on headlights for moving vehicles under night or low-visibility fog conditions, while stationary parked vehicles remain unlit. Streetlights are localized from semantic masks and lifted to 3D using the metric depth, after which their radiance is accumulated with vehicle lights in the same Light Pass.

Volumetric Atmospheric Scattering. We formulate foggy environments by analytically resolving volumetric scattering via a single-scattering Radiative Transfer Equation (RTE) model equipped with the Henyey-Greenstein phase function [17]. Evaluated directly against the calibrated metric depth $\mathbf{D}$, this explicit formulation yields distance-dependent visibility attenuation and localized light halos. The implementation details are provided in (Sec. 11 of the supplementary materials).

Environment Harmonization. To synthesize global ambient illumination for regions with sparse 3D geometry, we employ a neural forward renderer conditioned on an HDR environment map. The synthesized ambient radiance is linearly blended with the local light pass, effectively completing the deferred shading cycle. Implementation and fusion details are provided in (Sec.12 of the supplementary materials).

3.3 VidRefiner

VidRefiner is implemented as an SDEdit-style [35] refinement module adapted from WAN-FUN 2.2 [57]. Its goal is not to re-edit the scene semantics or physical

	S.A.	Night	Fog	Rain	Snow
Time (s)	128.1	167.1	170.9	2.2	67.6

Table 2: Running time (seconds) per video for different tasks and weather conditions on a NVIDIA V100 GPU. S.A.: Semantic Annotation, which is computed once and shared by all four weather.

attributes, but to improve sensor realism after the deterministic G-buffer rendering. To keep the refinement stage anchored to the deterministic G-buffer rendering, we condition the denoising trajectory with two complementary constraints:

Latent Initialization. The rendered sequence serves as a comprehensive structural and spectral anchor, injecting low-frequency priors into the generative process. By perturbing VAE-encoded latents to a pivot timestep t_s, the reverse diffusion trajectory inherits the global layout, color distribution, and coarse lighting resolved in the physical simulation. This initialization restricts the generative process to high-frequency textural refinement, preventing unconstrained global synthesis while preserving the deterministic scene structure.

Boundary Conditioning. To complement the low-frequency priors, high-frequency spatial constraints are enforced via spatiotemporally coherent boundaries extracted from the rendered output. This integration utilizes a lightweight backbone [57] pre-aligned for multi-channel conditioning, facilitating direct channel-wise concatenation without secondary fine-tuning. Unlike cross-attention mechanisms providing latent-level semantic guidance, this input-level formulation imposes an explicit spatial bias. The architectural choice of a lightweight model further restricts the synthesis of fine-grained textures to the resolved geometric limits, ensuring the structural integrity of edited elements remains invariant during photorealistic refinement.

The implementation details about the VidRefiner are provided in (Sec. 14 of the supplementary materials)

4 Experiment

4.1 Experimental Setting

To validate our weather and time-of-day conversion method, we conduct experiments using PyTorch on NVIDIA GPUs (V100 for our method and most baselines; A100 for resource-intensive baselines like Cosmos-Transfer2.5 [1] and Ditto [2]). We evaluate on 120 scenes from the Waymo Open Dataset [34], specifically using NOTR — a versatile subset of Waymo encompassing diverse driving scenarios, as surveyed in [75]. The time consumption of the core component (G-buffer Dual-pass Editing) is reported in Tab. 2.

4.2 Baselines

Baseline Selection. Existing 3D-aware weather synthesis frameworks [10,27] are fundamentally constrained to static scenes, rendering them structurally incompatible with the highly dynamic environments of autonomous driving. Consequently,

to evaluate temporal consistency and semantic fidelity in dynamic sequences, the framework is benchmarked against state-of-the-art video editing and foundation models: Video-P2P [30], Ditto [2], Cosmos-Transfer2.5 [1], and WAN-FUN 2.2 [57]. The comparative analysis is further extended to include domain-specific architectures: the inverse-rendering framework DiffusionRenderer [28] (constrained to translations conditioned on HDR environment maps) and the concurrent work WeatherEdit [44] (bounded to fog, snow, and rain synthesis).

Evaluation Protocol. The evaluation covers four primary weather and lighting conditions for autonomous driving: fog, midnight, rain, and snow. For baselines that require text inputs, prompts are constructed by directly concatenating the original scene description with the target weather condition. This standardized text input provides consistent guidance across all generative models, ensuring that performance differences are not caused by manual prompt engineering (Prompt details in Supplementary materials Sec. 13).

Evaluation Metrics. The synthesis fidelity and physical consistency of the generated sequences are systematically quantified across three dimensions. (1) *Editing Instruction Adherence* utilizes the CLIP score [19] to evaluate the precise execution of the targeted weather/time-of-day translation. (2) *Structural Consistency* assesses geometric preservation through a bounding-box Intersection-over-Union (IoU) protocol. By comparing projected 2D ground-truth LiDAR boxes against extractions from a pre-trained monocular 3D detector [77], the relative IoU serves as a strict indicator of structural rigidity across all baselines (AABB projection formulations in Supp. Sec. 15). (3) *Identity Stability* enforces the semantic invariance of foreground subjects. Computed via patch-level CLIP feature similarity before and after editing, this metric rigorously penalizes generative hallucinations while accommodating valid physical material alterations, such as snow accumulation or wet surface reflections.

4.3 User Study

A Two-Alternative Forced Choice (2AFC) study is conducted with 12 independent raters. To prevent scene selection bias, evaluation sequences are randomly sampled across diverse weather conditions and strictly standardized in resolution and duration (512×512, 30 frames). The framework is benchmarked against the four baselines (10 paired comparisons per baseline). To negate cognitive bias, each trial enforces strict double-blind randomization for both presentation order and left-right layout. Evaluation is governed by two explicit criteria: (1) *Spatial Fidelity*, assessing photorealism and editing instruction adherence; and (2) *Temporal Coherence*, penalizing background flickering and evaluating the motion continuity of dynamic weather. Final win rates are aggregated from 1,440 independent responses.

4.4 Results

Qualitative Results. We present qualitative results of our editing framework in Fig. 1. As illustrated, given an autonomous driving video, our approach

Fig. 3: Qualitative Comparisons of AutoWeather4D on Waymo Weather/ Time-of-day Conversions: Validating Physically Plausible and Fine-Grained Control for Autonomous Driving.

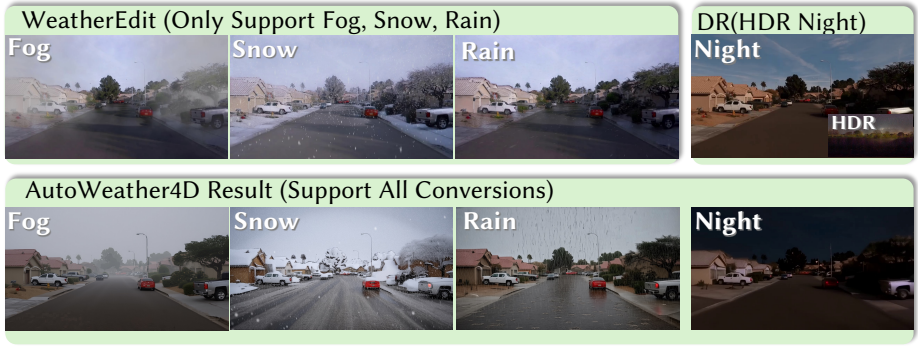

Fig. 4: Qualitative Comparisons with domain-specific architectures: Validating spatially anchoring weather effects. DR: DiffusionRenderer [28].

enables precise control over key scene attributes—including shadows, lighting, and geometry—to facilitate conversions between different weather conditions and times of day.

Comparisons Results. We provide qualitative examples in Fig. 3 to demonstrate the effectiveness of our method. Video-P2P struggles to follow instructions, while Ditto introduces background structural artifacts and generative hallucinations (e.g., extraneous architectural elements in row 4). In contrast to WAN-FUN and Cosmos-Transfer, our approach enforces physical plausibility and fine-grained spatial control. Specifically, row 1 illustrates how our explicit parameterization allows for localized manipulation of individual light sources and fog density, effectively resolving the ambiguity of pure text-conditioning. Furthermore, as shown in rows 2 and 3, the baselines fail to disentangle source lighting, resulting in biased road brightness or incompatible hard shadows. Our method, however, correctly models illumination transport. Finally, row 4 validates that our pipeline

strictly preserves foreground geometry. These structural and photometric consistencies are directly attributed to our explicit illumination decomposition and parameterized light control modules.

Comparison with Domain-Specific Architectures. As illustrated in Fig. 4, WeatherEdit retains hard shadows under highly scattering conditions (e.g., fog, rain, snow). In contrast, the decoupled G-buffer representation mitigates the shape-radiance ambiguity [79] by anchoring the editing process in explicit geometry. This structural prior further facilitates consistent diffuse shading and enables localized weather interactions, such as snow accumulation and surface ripples. Furthermore, compared to DiffusionRenderer [28], the 3D-aware formulation leverages spatial priors to isolate the sky region. This depth-based separation avoids unintended sky relighting and supports localized illumination driven by ego-vehicle headlights. Extended results are provided in (Supp. Sec. 17) and the supplementary videos.

Model	CLIP Score (↑)	Vehicle 3D Detection IoU (↑)	Vehicle CLIP cosine similarity (↑)	Human Evaluation (↑)
Video-P2P	0.2448	-	-	0
Ditto	0.2532	0.805	0.769	0.425
Cosmos-Transfer2.5	0.2558	0.913	0.837	0.580
WAN-FUN 2.2	0.2577	0.888	0.794	0.668
Ours	**0.2586**	**0.915**	**0.871**	**0.826**

Table 3: Quantitative evaluation of weather and time-of-day conversions on the Waymo dataset. All results are derived from 120 source videos, each converted into four common adverse conditions (rain, snow, fog, night) with 57 frames per video. This results in a total of 27,360 frames evaluated. "-" indicates omitted due to frame cropping. All metrics are averaged across all frames, where **red** indicates the best performance.

Quantitative Result Analysis. As shown in Tab. 3, AutoWeather4D achieves performance comparable to existing baselines. While numerical results are on par with massive data-driven models (e.g., Cosmos-Transfer2.5), they demonstrate that our framework maintains high generation fidelity. Crucially, our method complements the implicit generation process by introducing explicit physical control, offering a deterministic alternative for fine-grained editing. Extended comparative evaluations on general generative performance—encompassing FVD and distribution-level metrics distinct from core editing fidelity—are deferred to (Supplementary Materials Sec. 15).

4.5 Ablation Study

Comprehensive architectural ablations—encompassing isolated modules, combinatorial configurations, and granular physical parameterizations (rain, snow, local illumination, and VidRefiner conditioning strength)—are strictly deferred

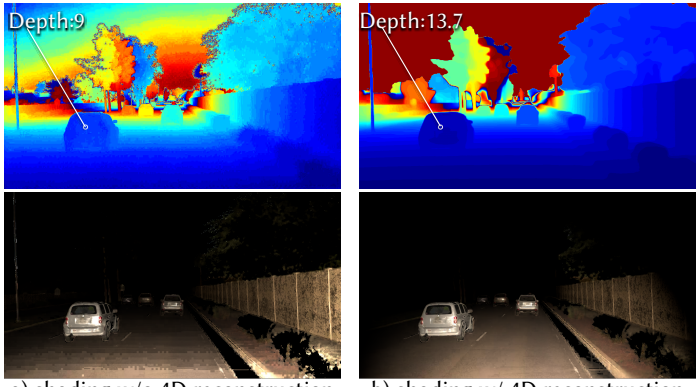

Fig. 5: Ablation of 4D reconstruction. (a) Integer-quantized depth priors [28] induce severe spatial discretization and aliasing during local relighting. (b) The deployed feed-forward 4D reconstruction establishes a continuous floating-point manifold, enforcing smooth, artifact-free illumination gradients.

to the Supplementary Materials (Sec. 16). The subsequent analysis explicitly isolates the structural necessity of the continuous 4D reconstruction.

Effect of 4D reconstruction. Distance-based light attenuation dictates a strict mathematical requirement for continuous spatial gradients. Utilizing standard inverse rendering alone extracts integer-quantized depth maps. Under explicit local illumination, this spatial discretization inherently provokes abrupt discontinuities in the light attenuation function, manifesting as severe jagged aliasing across the rendered surfaces (Fig. 5a). By integrating the feed-forward 4D reconstruction backbone, the pipeline recovers a continuous, floating-point geometric manifold. This non-discrete structural prior deterministically resolves spatial step artifacts, guaranteeing natural and continuous illumination gradients during dynamic relighting (Fig. 5b).

4.6 Applications

Efficacy in Perception Data Augmentation. Following established weather synthesis protocols [44], the utility of AutoWeather4D is assessed for downstream perception data augmentation. The synthesized sequences serve as supplementary training data to address domain gaps in semantic segmentation. Source-domain semantic annotations are directly transferred to the synthesized adverse videos. Subsequently, the HRDA segmentation model [20] is fine-tuned on 6,480 augmented frames for 20k iterations. The segmentation performance is evaluated via mIoU and mAcc using the Cityscapes [9] taxonomy across two standard adverse-condition datasets: ACDC [50] (snow, rain, fog, night) and Dark Zurich [49] (day, night).

Tab. 4 reports the quantitative impact of the synthesized sequences on downstream perception robustness. While absolute mIoU increments on ACDC and

Setting	ACDC mIoU(↑)	ACDC mAcc(↑)	DarkZurich mIoU(↑)	DarkZurich mAcc(↑)
w/o augmentation	49.20	60.72	23.92	38.29
w/ cosmos	49.66 (+0.93%)	62.31 (+2.62%)	23.93 (+0.04%)	39.52 (+3.21%)
w/ ours	**49.81 (+1.24%)**	**62.52 (+2.96%)**	**24.09 (+0.71%)**	**39.73 (+3.76%)**

Table 4: AutoWeather4D for data augmentation in adverse weather semantic segmentation.

Fig. 6: Global Illumination Editing driven by HDR environment map.

Dark Zurich remain marginal, this reflects the inherent saturation of zero-shot cross-domain transfer within a highly optimized baseline (HRDA). Consequently, this proxy evaluation is formulated strictly to validate the geometric fidelity of the synthesized data, rather than to advance the segmentation state-of-the-art. Under severe atmospheric shifts, standard generative baselines exhibit structural degradation, resulting in negligible transfer benefits (e.g., Cosmos yielding a +0.04% mIoU increment on Dark Zurich). Conversely, the consistent gains provided by AutoWeather4D demonstrate that explicitly anchored synthesis reliably preserves the underlying scene geometry required for robust downstream training.

Illumination Control. The decoupled representation enables HDR-driven illumination. Parameters such as sun altitude and ambient color are adjustable without changing scene geometry (Fig 6). The G-buffer ensures consistent shading and light transport across different environment maps.

Furthermore, the architectural advantage of AutoWeather4D extends beyond static data augmentation. As demonstrated in Fig. 7, our explicit parameterized control (e.g., continuously scaling fog density or toggling local lights) enables the targeted synthesis of specific challenging scenarios where downstream algorithms commonly fail. While comprehensive downstream diagnostic benchmarking is left for future work, this fine-grained, deterministic controllability inherently provides the foundation for generating continuous perturbation sequences. This establishes a highly controllable capability for future system-level robustness evaluation, providing a complementary deterministic approach to purely data-driven generative baselines, avoiding the need for excessively large-scale specific data collection.

Fig. 7: Explicit Parameterized Control for Physically Consistent Weather and Time-of-Day Video Editing.

5 Conclusion

To mitigate data scarcity, AutoWeather4D aligns deterministic graphics with video diffusion to transform real-world footage into adverse scenarios. By anchoring the generative process to explicit G-buffer priors, we transition from synthesis with entangled geometry and illumination toward physically grounded, parametric simulation. This framework serves as a complementary data source, offering potential for future diagnostics of perception failure modes in adverse conditions.

Limitations. While our explicit-implicit bridging achieves physically grounded synthesis for primary weather conditions, capturing extreme long-tail dynamic interactions (e.g., complex fluid dynamics of vehicle splash) remains challenging for decoupled pipelines. Future work will explore integrating localized generative priors specifically for these unstructured phenomena, complementing our deterministic structural anchors. Additionally, balancing severe environmental perturbations (e.g., heavy fog occlusion) with the structural retention of distant background elements necessitates careful calibration. While our boundary conditioning effectively anchors foreground geometries, future iterations could incorporate semantic-aware attenuation masks to dynamically modulate the diffusion intensity across critical autonomous driving regions of interest.

6 Acknowledgements

This work is supported by the Shenzhen Loop Area Institute under grant FPF10120250006.

References

1. Ali, A., Bai, J., Bala, M., Balaji, Y., Blakeman, A., Cai, T., Cao, J., Cao, T., Cha, E., Chao, Y.W., et al.: World simulation with video foundation models for physical ai. arXiv preprint arXiv:2511.00062 (2025)
2. Bai, Q., Wang, Q., Hao, O., Yu, Y., Wang, H., Wang, W., Cheng, K.L., Ma, S., Zeng, Y., Liu, Z., Xu, Y., Shen, Y., Chen, Q.: Scaling instruction-based video editing with a high-quality synthetic dataset. arXiv preprint arXiv:2510.15742 (2025)
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)

4. Bruneton, E., Neyret, F.: Precomputed atmospheric scattering. In: Computer graphics forum. vol. 27, pp. 1079–1086. Wiley Online Library (2008)
5. Canny, J.: A computational approach to edge detection. IEEE Transactions on pattern analysis and machine intelligence (6), 679–698 (2009)
6. Chen, S., Guo, H., Zhu, S., Zhang, F., Huang, Z., Feng, J., Kang, B.: Video depth anything: Consistent depth estimation for super-long videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22831–22840 (2025)
7. Chen, X., Peng, S., Yang, D., Liu, Y., Pan, B., Lv, C., Zhou, X.: Intrinsicanything: Learning diffusion priors for inverse rendering under unknown illumination. In: European Conference on Computer Vision. pp. 450–467. Springer (2024)
8. Cook, R.L., Torrance, K.E.: A reflectance model for computer graphics. ACM Transactions on Graphics **1**(1), 7–24 (1982)
9. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3213–3223 (2016)
10. Dai, Q., Ni, X., Shen, Q., Chen, W., Chen, B., Chu, M.: Rainygs: Efficient rain synthesis with physically-based gaussian splatting. In: Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference. pp. 16153–16162 (2025)
11. Dijk, T.v., Croon, G.d.: How do neural networks see depth in single images? In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2183–2191 (2019)
12. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. Advances in neural information processing systems **27** (2014)
13. Fan, L., Yang, Y., Mao, Y., Wang, F., Chen, Y., Wang, N., Zhang, Z.: Once detected, never lost: Surpassing human performance in offline lidar based 3d object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19820–19829 (2023)
14. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM **24**(6), 381–395 (1981)
15. Gao, R., Chen, K., Xie, E., Hong, L., Li, Z., Yeung, D.Y., Xu, Q.: Magicdrive: Street view generation with diverse 3d geometry control. In: The Twelfth International Conference on Learning Representations (2024)
16. Gao, S., Yang, J., Chen, L., Chitta, K., Qiu, Y., Geiger, A., Zhang, J., Li, H.: Vista: A generalizable driving world model with high fidelity and versatile controllability. Advances in Neural Information Processing Systems **37**, 91560–91596 (2024)
17. Henyey, L.G., Greenstein, J.L.: Diffuse radiation in the galaxy. Astrophysical Journal, vol. 93, p. 70-83 (1941). **93**, 70–83 (1941)
18. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross-attention control. In: The Eleventh International Conference on Learning Representations (2023)
19. Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. In: Proceedings of the 2021 conference on empirical methods in natural language processing. pp. 7514–7528 (2021)
20. Hoyer, L., Dai, D., Van Gool, L.: Hrda: Context-aware high-resolution domain-adaptive semantic segmentation. In: European Conference on Computer Vision. pp. 372–391. Springer (2022)

21. Jia, F., Mao, W., Liu, Y., Zhao, Y., Wen, Y., Zhang, C., Zhang, X., Wang, T.: Adriver-i: A general world model for autonomous driving. arXiv preprint arXiv:2311.13549 (2023)
22. Kajiya, J.T., Von Herzen, B.P.: Ray tracing volume densities. ACM SIGGRAPH computer graphics **18**(3), 165–174 (1984)
23. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (July 2023)
24. Kocsis, P., Philip, J., Sunkavalli, K., Nießner, M., Hold-Geoffroy, Y.: Lightit: Illumination modeling and control for diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9359–9369 (2024)
25. Kong, X., Watson, D., Strümpler, Y., Niemeyer, M., Tombari, F.: Causnvs: Autoregressive multi-view diffusion for flexible 3d novel view synthesis. arXiv preprint arXiv:2509.06579 (2025)
26. Li, X., Zhang, Y., Ye, X.: Drivingdiffusion: layout-guided multi-view driving scenarios video generation with latent diffusion model. In: European Conference on Computer Vision. pp. 469–485. Springer (2024)
27. Li, Y., Lin, Z.H., Forsyth, D., Huang, J.B., Wang, S.: Climatenerf: Extreme weather synthesis in neural radiance field. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3227–3238 (2023)
28. Liang, R., Gojcic, Z., Ling, H., Munkberg, J., Hasselgren, J., Lin, C.H., Gao, J., Keller, A., Vijaykumar, N., Fidler, S., et al.: Diffusion renderer: Neural inverse and forward rendering with video diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 26069–26080 (2025)
29. Lin, C.H., Wang, Z., Liang, R., Zhang, Y., Fidler, S., Wang, S., Gojcic, Z.: Controllable weather synthesis and removal with video diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13580–13591 (2025)
30. Liu, S., Zhang, Y., Li, W., Lin, Z., Jia, J.: Video-p2p: Video editing with cross-attention control. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8599–8608 (2024)
31. Ljungbergh, W., Taveira, B., Zheng, W., Tonderski, A., Peng, C., Kahl, F., Petersson, C., Felsberg, M., Keutzer, K., Tomizuka, M., et al.: R3d2: Realistic 3d asset insertion via diffusion for autonomous driving simulation. arXiv preprint arXiv:2506.07826 (2025)
32. Ma, E., Zhou, L., Tang, T., Zhang, Z., Han, D., Jiang, J., Zhan, K., Jia, P., Lang, X., Sun, H., et al.: Unleashing generalization of end-to-end autonomous driving with controllable long video generation. arXiv preprint arXiv:2406.01349 (2024)
33. Mandelbrot, B.B., Van Ness, J.W.: Fractional brownian motions, fractional noises and applications. SIAM review **10**(4), 422–437 (1968)
34. Mei, J., Zhu, A.Z., Yan, X., Yan, H., Qiao, S., Chen, L.C., Kretzschmar, H.: Waymo open dataset: Panoramic video panoptic segmentation. In: European Conference on Computer Vision. pp. 53–72. Springer (2022)
35. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. In: International Conference on Learning Representations (2022)
36. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: European Conference on Computer Vision. pp. 405–421. Springer (2020)

37. Minderer, M., Gritsenko, A., Stone, A., Neumann, M., Weissenborn, D., Dosovitskiy, A., Mahendran, A., Arnab, A., Dehghani, M., Shen, Z., et al.: Simple open-vocabulary object detection. In: European Conference on Computer Vision. pp. 728–755. Springer (2022)
38. Müller, M., Charypar, D., Gross, M.: Particle-based fluid simulation for interactive applications. In: Proceedings of the 2003 ACM SIGGRAPH/Eurographics symposium on Computer animation. pp. 154–159 (2003)
39. Muraoka, K., Chiba, N.: Visual simulation of snowfall, snow cover and snowmelt. In: Proceedings Seventh International Conference on Parallel and Distributed Systems: Workshops. pp. 187–194. IEEE (2000)
40. Ni, C., Zhao, G., Wang, X., Zhu, Z., Qin, W., Huang, G., Liu, C., Chen, Y., Wang, Y., Zhang, X., et al.: Recondreamer: Crafting world models for driving scene reconstruction via online restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1559–1569 (2025)
41. Nishita, T., Iwasaki, H., Dobashi, Y., Nakamae, E.: A modeling and rendering method for snow by using metaballs. In: Computer Graphics Forum. vol. 16, pp. C357–C364. Wiley Online Library (1997)
42. Nishita, T., Sirai, T., Tadamura, K., Nakamae, E.: Display of the earth taking into account atmospheric scattering. In: Proceedings of the 20th annual conference on Computer graphics and interactive techniques. pp. 175–182 (1993)
43. Pun, A., Sun, G., Wang, J., Chen, Y., Yang, Z., Manivasagam, S., Ma, W.C., Urtasun, R.: Neural lighting simulation for urban scenes. Advances in Neural Information Processing Systems **36**, 19291–19326 (2023)
44. Qian, C., Li, W., Guo, Y., Markkula, G.: Weatheredit: Controllable weather editing with 4d gaussian field. arXiv preprint arXiv:2505.20471 (2025)
45. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. In: The Thirteenth International Conference on Learning Representations (2025)
46. Reeves, W.T.: Particle system: A technique for modeling a class of fuzzy objects. ACM Transactions on Graphics **2**(2), 91–108 (1983)
47. Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., et al.: Grounded sam: Assembling open-world models for diverse visual tasks. arXiv preprint arXiv:2401.14159 (2024)
48. Russell, L., Hu, A., Bertoni, L., Fedoseev, G., Shotton, J., Arani, E., Corrado, G.: Gaia-2: A controllable multi-view generative world model for autonomous driving. arXiv preprint arXiv:2503.20523 (2025)
49. Sakaridis, C., Dai, D., Van Gool, L.: Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7374–7383 (2019)
50. Sakaridis, C., Dai, D., Van Gool, L.: Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10765–10775 (2021)
51. Sang, C., Qian, Y., Zhang, J., Wang, C., Yang, M.: Weather-magician: Reconstruction and rendering framework for 4d weather synthesis in real time. arXiv preprint arXiv:2505.19919 (2025)
52. Schlick, C.: An inexpensive brdf model for physically-based rendering. In: Computer graphics forum. vol. 13, pp. 233–246. Wiley Online Library (1994)
53. Schmidt, V., Luccioni, A., Teng, M., Zhang, T., Reynaud, A., Raghupathi, S., Cosne, G., Juraver, A., Vardanyan, V., Hernández-García, A., et al.: Climategan:

Raising climate change awareness by generating images of floods. In: International Conference on Learning Representations (2022)
54. Smith, B.: Geometrical shadowing of a random rough surface. IEEE transactions on antennas and propagation **15**(5), 668–671 (1967)
55. Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Fvd: A new metric for video generation. In: DGS@ICLR (2019)
56. Walter, B., Marschner, S.R., Li, H., Torrance, K.E.: Microfacet models for refraction through rough surfaces. Rendering techniques **2007**, 18th (2007)
57. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
58. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5294–5306 (2025)
59. Wang, J., Fang, J., Zhang, X., Xie, L., Tian, Q.: Gaussianeditor: Editing 3d gaussians delicately with text instructions. In: Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference. pp. 20902–20911 (2024)
60. Wang, L., Zheng, W., Ren, Y., Jiang, H., Cui, Z., Yu, H., Lu, J.: Occsora: 4d occupancy generation models as world simulators for autonomous driving. arXiv preprint arXiv:2405.20337 (2024)
61. Wang, L., Lin, Z., Fang, T., Yang, X., Yu, X., Kang, S.B.: Real-time rendering of realistic rain. ACM SIGGRAPH 2006 Sketches on-SIGGRAPH'06 p. 156 (2006)
62. Wang, N., Wade, B.: Rendering falling rain and snow. ACM SIGGRAPH 2004 Sketches on-SIGGRAPH'04 p. 14 (2004)
63. Wang, R.j., Tian, J.q., Ni, Z.: Real time simulation of rain and snow based on particle system. Journal of System Simulation **15**(4), 495–496 (2003)
64. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20697–20709 (2024)
65. Wang, X., Qi, S., Zhao, J., Zhou, H., Zhang, S., Wang, G., Tu, K., Guo, S., Zhao, J., Li, J., et al.: Mctrack: A unified 3d multi-object tracking framework for autonomous driving. In: 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 4551–4558. IEEE (2025)
66. Wang, Y., Zhou, J., Zhu, H., Chang, W., Zhou, Y., Li, Z., Chen, J., Pang, J., Shen, C., He, T.: π^3: Scalable permutation-equivariant visual geometry learning. arXiv e-prints pp. arXiv–2507 (2025)
67. Wei, Y., Wang, Z., Lu, Y., Xu, C., Liu, C., Zhao, H., Chen, S., Wang, Y.: Editable scene simulation for autonomous driving via collaborative llm-agents. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15077–15087 (2024)
68. Wen, Y., Zhao, Y., Liu, Y., Jia, F., Wang, Y., Luo, C., Zhang, C., Wang, T., Sun, X., Zhang, X.: Panacea: Panoramic and controllable video generation for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6902–6912 (2024)
69. Wobus, H.B., Murray, F.W., Koenig, L.R.: Calculation of the terminal velocity of water drops. Journal of Applied Meteorology (1962-1982) pp. 751–754 (1971)
70. Wu, H., Zhang, E., Liao, L., Chen, C., Hou, J., Wang, A., Sun, W., Yan, Q., Lin, W.: Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20144–20154 (2023)

71. Xia, H., Lin, Z.H., Ma, W.C., Wang, S.: Video2game: Real-time interactive realistic and browser-compatible environment from a single video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4578–4588 (2024)
72. Xia, H., Su, E., Memmel, M., Jain, A., Yu, R., Mbiziwo-Tiapo, N., Farhadi, A., Gupta, A., Wang, S., Ma, W.C.: Drawer: Digital reconstruction and articulation with environment realism. In: Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference. pp. 21771–21782 (2025)
73. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in neural information processing systems **34**, 12077–12090 (2021)
74. Xing, X., Hu, V.T., Metzen, J.H., Groh, K., Karaoglu, S., Gevers, T.: Retinex-diffusion: On controlling illumination conditions in diffusion models via retinex theory. arXiv preprint arXiv:2407.20785 (2024)
75. Yang, J., Ivanovic, B., Litany, O., Weng, X., Kim, S.W., Li, B., Che, T., Xu, D., Fidler, S., Pavone, M., et al.: Emernerf: Emergent spatial-temporal scene decomposition via self-supervision. In: The Twelfth International Conference on Learning Representations (2024)
76. Yang, Z., Chen, Y., Wang, J., Manivasagam, S., Ma, W.C., Yang, A.J., Urtasun, R.: Unisim: A neural closed-loop sensor simulator. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1389–1399 (2023)
77. Yao, J., Gu, H., Chen, X., Wang, J., Cheng, Z.: Open vocabulary monocular 3d object detection. In: Thirteenth International Conference on 3D Vision (2026)
78. Zeng, C., Dong, Y., Peers, P., Kong, Y., Wu, H., Tong, X.: Dilightnet: Fine-grained lighting control for diffusion-based image generation. In: ACM SIGGRAPH 2024 Conference Papers. pp. 1–12 (2024)
79. Zhang, K., Riegler, G., Snavely, N., Koltun, V.: Nerf++: Analyzing and improving neural radiance fields. arXiv preprint arXiv:2010.07492 (2020)
80. Zhang, L., Rao, A., Agrawala, M.: Scaling in-the-wild training for diffusion-based illumination harmonization and editing by imposing consistent light transport. In: The Thirteenth International Conference on Learning Representations (2025)
81. Zhao, G., Ni, C., Wang, X., Zhu, Z., Zhang, X., Wang, Y., Huang, G., Chen, X., Wang, B., Zhang, Y., et al.: Drivedreamer4d: World models are effective data machines for 4d driving scene representation. In: Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference. pp. 12015–12026 (2025)
82. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2223–2232 (2017)
83. Zhu, Y., Zhu, Z., Hašan, M., Yang, J., Xie, J., Wang, B.: Weatherdiffusion: Weather-guided diffusion model for forward and inverse rendering. arXiv preprint arXiv:2508.06982 (2025)
84. Zhu, Z., Zou, Y., Jiang, C.M., Sun, B., Casser, V., Huang, X., Wang, J., Yang, Z., Gao, R., Guibas, L., et al.: Scenecrafter: Controllable multi-view driving scene editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6812–6822 (2025)
85. Zhuang, J., Wang, C., Lin, L., Liu, L., Li, G.: Dreameditor: Text-driven 3d scene editing with neural fields. In: ACM SIGGRAPH Asia 2023 Conference Papers. pp. 1–10 (2023)