

Semantic-Aware, Physics-Informed, Geometry-Grounded Weather Synthesis

Chenghao Qian¹, Nedko Savov², Lingdong Kong³,
Yeying Jin³, Rui Song⁵, Wenjing Li^{1,4}, Zhun Zhong⁴,
Jiaqi Ma⁵, Gustav Markkula¹, and Luc Van Gool²

¹ University of Leeds, UK

² INSAIT, Sofia University "St. Kliment Ohridski"

³ National University of Singapore, Singapore

⁴ Hefei University of Technology, China

⁵ University of California, Los Angeles, USA



Fig. 1: Weather Synthesis Examples. Given an original input video, our method can synthesize diverse weather effects, including the *type* of appearance (snowy or rainy), the *event* condition (static or dynamic), the *duration* (how long the weather lasts), the *direction* and *strength* of wind influence on particles, and the overall *severity*. This enables precise and interpretable control over atmospheric appearance and particle dynamics, yielding diverse and realistic weather effects.

Abstract. Weather synthesis aims to add weather effects to input videos while preserving scene identity, structure, and motion. The key limitation of existing methods is the lack of diversity in weather appearance and effective control over weather dynamics (e.g., temporal evolution and particle motion). Most approaches rely on text prompts, which are inherently underspecified and often fail to produce detailed weather characteristics. Additionally, general-purpose video editors optimized for clean and aesthetic outputs tend to suppress heavy weather phenomena, making dense particle effects difficult to generate. To address these, we propose a *Semantic-Aware, Physics-Informed, and Geometry-Grounded* framework that steers an off-the-shelf video editor to synthesize diverse global appearances and detailed particle dynamics. We factorize the synthesis into three conditional signals, so that each provides a distinct

 Corresponding authors.

and stable source of guidance: *semantics* specifies what the weather should look like, *dynamics* governs how it evolves over time, and *geometry* determines where it should appear in the scene. Specifically, we introduce (1) semantic-aware appearance anchoring to establish the target appearance from scene semantics and user input; (2) physics-informed dynamic simulation to generate particle effects by simulating a Gaussian-represented particle field under gravity, wind, and turbulence; and (3) geometry-grounded video synthesis to align the simulated particles with target scene geometry and synthesize the final video. Experiments demonstrate that our method produces diverse, physically and visually realistic weather effects. Furthermore, we show that our synthesized data significantly improves the robustness of autonomous driving semantic segmentation under adverse weather conditions. See our project page at <https://jumponthemoon.github.io/w-crafter/>.

Keywords: Weather Synthesis · Video Editing · Particle Simulation

1 Introduction

What would the Colosseum look like blanketed in heavy snowfall, or a desert pyramid drenched by a torrential downpour? Such conditions are rarely captured in the real world, yet can be envisioned through weather synthesis, where weather effects are synthesized and added to existing videos. This capability is valuable for creative applications, where one wishes to control specific weather conditions on real footage [12, 66]. Moreover, it can serve as a data augmentation tool for training perception systems [2, 54], enhancing the robustness of autonomous driving and robotic systems under diverse weather conditions [31, 56].

Existing methods of weather video synthesis mainly follow two directions. Reconstruction-based pipelines [40, 55] simulate weather effects in a 3D scene using NeRF [49] or 3DGS [29], then re-render the scene under synthesized conditions. While they offer interpretable control, the handcrafted effects lack diversity and are sensitive to reconstruction errors, resulting in unrealistic results. Data-driven approach [42] trains diffusion models on curated weather datasets to directly synthesize effects conditioned on the input video. Although these methods enable style control, the resulting appearance is often biased toward the training distribution, which limits diversity. Furthermore, synthesizing novel weather effects typically requires extensive additional data and costly re-training.

More recently, conditional video editors incorporate inputs such as depth [21, 27], optical flow [9, 45] and pose [17, 48] together with text prompts, allowing flexible controllability for video-to-video translation [34, 41]. However, these models are typically trained on clean, high-quality videos, where heavy-weather degradations are under-represented. As a result, prompting for “heavy snow” or “intense rain” often yields only mild effects, failing to produce dense particles and realistic dynamics. *Interestingly*, we observe that modern video editor models still encode non-trivial priors about particle dynamics (*e.g.*, falling snowflakes or rain streaks). In particular, injecting particle visual primitives and trajectories into



Fig. 2: Comparison between text prompt-only and particle-conditioned editing. Text prompts alone often yield mild precipitation with weak particle details. In contrast, particle-conditioned editing enables dense snowflakes and rain streaks synthesis, indicating it effectively activates latent weather priors in pretrained video editors.

structured conditioning (*e.g.*, RGB or depth) can activate particle-aware generation: once a particle-augmented input is provided, the pretrained model can synthesize raindrops and snowflakes without finetuning, as illustrated in Fig. 2. This suggests that the challenge is not whether such models can generate weather particles, but how to *activate* this behavior in an effective manner.

Our key insight is that a well-designed conditioning space is essential: structured signals that jointly encode appearance and dynamics provide explicit guidance beyond text-only prompts. Motivated by this, we propose a *Semantic-Aware, Physics-Informed, and Geometry-Grounded* framework. Specifically, we decompose weather synthesis into three complementary conditioning signals: semantics, dynamics, and geometry, instantiated as semantic-aware appearance anchoring, physics-informed dynamic simulation, and geometry-grounded video synthesis. For *semantics*, we leverage a Multimodal Large Language Model (MLLM) [1] to construct appearance anchors that define the target weather style based on user intent. For *dynamics*, we simulate a particle field represented by Gaussian primitives under physical priors, governing particle motion and temporal behavior. For *geometry*, we fuse scene and weather particles with geometry to form particle-augmented depths as conditioning for video synthesis. By combining these three conditioning inputs, we steer an off-the-shelf video diffusion model to synthesize diverse and dynamic weather effects with high controllability, as shown in Fig. 1. Our contributions are fourfold:

- A ***novel and controllable weather synthesis framework*** that enables diverse global appearances and fine-grained particle dynamics by conditioning on structured signals of semantics, dynamics, and geometry.
- A ***semantic-aware editing strategy*** that binds intended weather effects to scene semantics, improving controllability and reliability, and yielding diverse and realistic synthesis results.
- A ***physics-informed particle representation*** that effectively activates latent weather priors in pretrained diffusion models, producing temporally coherent and physically plausible dynamic effects.
- A ***geometry-grounded mechanism*** that couples particle dynamics with 3D scene geometry, ensuring temporally consistent trajectories and spatially accurate placement across video frames.

2 Related Work

Weather Effect Synthesis. This task has been extensively explored in both graphics and vision [30, 32, 67]. Conventional approaches model precipitation and atmospherics through particle-based simulation [19, 61, 62] and physical scattering models [22, 23, 50]. Learning-based methods [24, 33, 39, 52, 54, 60] further perform weather translation or augmentation from data, enabling flexible appearance changes without explicit simulation. With the emergence of neural rendering, several works [15, 40, 55, 59, 64, 68] inject weather into implicit or explicit 3D neural representations and re-render scenes under simulated conditions. These pipelines offer interpretable controls, but often require costly per-scene optimization and are sensitive to reconstruction errors. More recently, WeatherWeaver [42] trains a weather-specific video diffusion model using synthetic datasets curated by graphics engines [13], but the resulting appearance is tied to the simulator, and transferring to new domains or styles typically requires additional data.

Unlike reconstruction-based pipelines and simulator-curated editors, we neither require costly per-scene optimization nor rely on fixed training distributions. Instead, we introduce a semantic-aware anchoring strategy to define scene-specific atmospheric styles and leverage physics-simulated particles to govern realistic weather dynamics. By grounding these elements with 3D geometric cues, we enable a general video diffusion model to render diverse and consistent weather effects in a zero-shot manner.

Conditional Video Editing. Conditional video editing performs video-to-video translation under user control while preserving identity and motion [14, 18, 28, 36]. Existing methods achieve control by conditioning generative models on structured signals such as depth [8, 20, 69], optical flow [11, 45], camera trajectories [4, 5, 37], or pose [17, 48, 51]. Recent works further support multi-condition inputs for richer, more compositional control [2, 3, 27, 47]. Compared with prompt-only editing [9, 10, 35, 63, 70], such structured conditions better constrain how visual effects change over time, improving alignment with scene structure.

The primary limitation of these methods is the inability to achieve **fine-grained semantic and temporal control over weather**. Most general video editors are targeting clean, high-quality outputs; heavy weather degradations are under-represented and hard to elicit reliably, resulting in limited weather appearance, weak particle details, and unstable temporal dynamics during editing. This motivates our structured conditioning interface, which provides stable guidance by injecting appearance semantics, particle dynamics, and 3D geometric cues into a video editor for diverse and consistent weather synthesis.

3 Methodology

Our framework consists of three stages (Fig. 3): **(i) Semantic-aware appearance anchoring** constructs an appearance anchor composed of an edited initial-

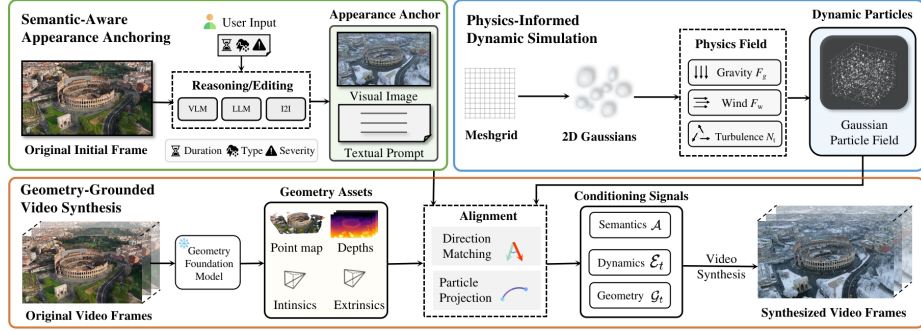


Fig. 3: Overview of our framework. Given an input video, we construct structured conditioning signals from three aspects: *semantics* (semantic-aware appearance anchoring), *dynamics* (physics-informed dynamic simulation), and *geometry* (geometry-grounded video synthesis). These signals are combined to guide a video diffusion model to synthesize realistic weather effects.

frame reference and a scene-grounded text prompt. Specifically, our MLLM-based pipeline employs a Vision-Language Model (VLM) to parse the scene, a Large Language Model (LLM) to refine the instructions, and an image-to-image (I2I) model to synthesize the initial frame. **(ii) Physics-informed particle simulation** simulates a weather particle field under physically motivated forces (*e.g.*, gravity and wind), producing temporally coherent particle cues that encode dynamics. **(iii) Geometry-grounded video synthesis** extracts per-frame geometry from the input video and uses this information to align the outputs from previous stages. This yields the final conditioning signals: semantics $\mathcal{A}$ with edited initial-frame and text prompt, dynamics $\{\mathcal{E}_t\}$ with particle-augmented depth maps, and geometry $\{\mathcal{G}_t\}$ with original scene depths, which jointly condition the video diffusion editor to produce the final video $\tilde{V}$.

3.1 Semantic-Aware Appearance Anchoring

To guide weather video synthesis toward a desired target appearance, we first edit the video’s initial frame using an image editing model. However, a naive prompt such as “*make it snowy*” is inherently underspecified, offering limited control over key factors such as severity and lighting tone. Even with coarse modifiers (*e.g.*, “heavy” or “light”), prompts remain ambiguous and often lead to limited editing effects (Fig. 4). We therefore develop a semantic-aware editing strategy comprising: semantic parsing, effects reasoning, and scene captioning. Our core idea is to craft detailed editing prompts by explicitly binding global appearance changes to scene semantics based on user intent, enabling fine-grained control and achieving desirable editing outcomes, as shown in Fig. 5.

Semantic Parsing. Given the initial frame I_0 , we use a VLM to translate the visual scene into interpretable attributes and produce a structured description P_{scene} via $P_{\text{scene}} \leftarrow \text{VLM}(I_0)$. This structured representation grounds subsequent

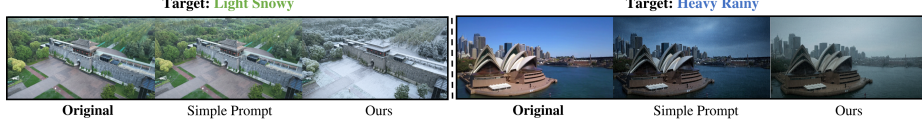


Fig. 4: Comparisons of editing results between the simple prompt and our prompt. With the simple prompt, the editing model struggles to produce light snow with clearly visible effects and heavy rain with accurate, scene-consistent lighting. In contrast, our generated prompt incorporates richer scene semantics and finer-grained details, leading to outputs that more reliably match the intended conditions.

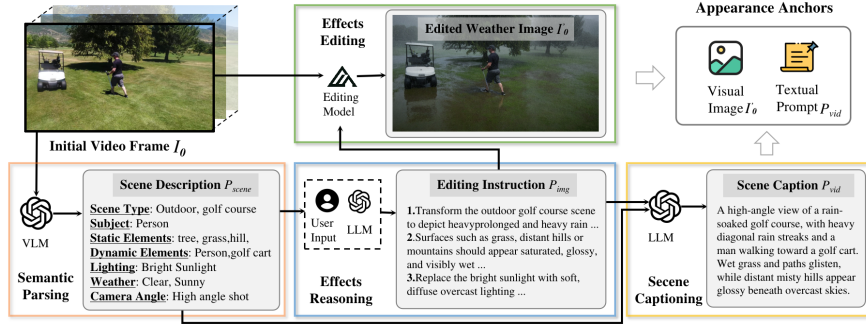


Fig. 5: Semantic-aware appearance anchoring. Given the initial video frame, a VLM parses scene semantics and an LLM reasons about weather-specific effects conditioned on user intent. The resulting instruction is used to edit the initial frame to anchor target appearance, while also being combined with scene semantics to provide a refined text description for subsequent video generation.

edits in scene semantics, enabling precise and localized appearance changes that correctly reflect the intended weather.

Effects Reasoning. Conditioned on the user intent $\mathcal{U}$ and the scene description P_{scene} , we employ an LLM to generate an image editing instruction P_{img} . In particular, $\mathcal{U}$ specifies target weather type, duration, and severity, while P_{scene} provides scene context to ensure the instruction is semantically plausible, yielding $P_{\text{img}} \leftarrow \text{LLM}(P_{\text{scene}}, \mathcal{U})$. We then generate an edited image I'_0 using an I2I model $\mathcal{M}_{\text{img}}$, which establishes scene-level weather appearance (*e.g.*, wet surface) together with a consistent global lighting tone (*e.g.*, overcast illumination):

$$I'_0 \leftarrow \mathcal{M}_{\text{img}}(I_0, P_{\text{img}}). \quad (1)$$

Scene Captioning. To provide a compact and consistent target for subsequent video synthesis, we further produce a video editing prompt P_{vid} with a scene caption that summarizes the expected edited scene under the target weather, capturing global atmosphere and key dynamic cues:

$$P_{\text{vid}} \leftarrow \text{LLM}(P_{\text{scene}}, P_{\text{img}}). \quad (2)$$

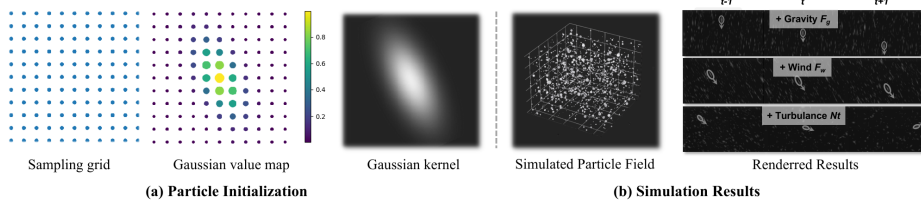


Fig. 6: Physics-informed dynamic simulation. (a) Weather particles are initialized as grid-sampled anisotropic Gaussians. (b) The particle field is evolved under gravity, wind, and turbulence, producing physically plausible motion that serves as explicit motion cues for video synthesis.

At the subsequent video synthesis stage, we integrate explicit weather cues by projecting the particle effects into I'_0 , yielding $\tilde{I}_0$ and feed it together with P_{vid} as the final appearance anchor $\mathcal{A}$ to the video editing model. The first-frame $\tilde{I}_0$ serves as a strong appearance reference that initializes the scene structure and global style, while P_{vid} supplies a high-level description that encourages temporally consistent semantics across frames.

3.2 Physics-Informed Dynamic Simulation

To generate dynamic weather effects, we model weather as a set of temporally persistent particles in a local field. Each particle is represented as an anisotropic Gaussian primitive (Fig. 6a), and its motion is governed by physics priors such as gravity, wind, and turbulence (Fig. 6b). This produces a physically plausible and controllable dynamic weather representation that can be geometrically grounded to video frames in later stages.

Particle initialization. We define a local sampling grid $\{\mathbf{s}\}$ with $\mathbf{s} = (x, y) \in \mathbb{R}^2$ and represent the appearance of particle i by an anisotropic Gaussian kernel:

$$G_i(\mathbf{s}) = \exp\left(-\frac{x^2}{2(\sigma_{x,i}^2)\gamma} - \frac{y^2}{2(\sigma_{y,i}^2)\gamma}\right), \quad (3)$$

where $(\sigma_{x,i}, \sigma_{y,i})$ parameterize per-particle anisotropy and γ controls the depth-dependent scale that is derived from the particle depth during projection. The resulting stamp serves as a compact primitive for fine-grained particle appearance (*e.g.*, rounded snowflakes or elongated raindrops) in the subsequent simulation.

Physics-informed simulation. To simulate particle dynamics, we augment each particle’s state with a depth coordinate, *i.e.*, $\mathbf{x}_i = (X_i, Y_i, Z_i) \in \mathbb{R}^3$, forming a particle field. We evolve each particle with a force-based dynamics that serves as an explicit temporal prior:

$$\dot{\mathbf{x}}_i(t) = \mathbf{v}_i(t), \quad m \dot{\mathbf{v}}_i(t) = \mathbf{F}_g + \mathbf{F}_w + \mathbf{N}_t(\mathbf{x}_i(t), t), \quad (4)$$

where $(\dot{\cdot})$ denotes the time derivative, $\mathbf{v}_i(t)$ corresponds to the velocity of particle i at time t , and m represents its mass. Gravity is modeled as $\mathbf{F}_g = mg\mathbf{d}_g$,

with magnitude g and direction $\mathbf{d}_g$, while wind is parameterized as $\mathbf{F}_w = \alpha \mathbf{d}_w$ with strength α and direction $\mathbf{d}_w$. To capture stochastic local perturbations, we model turbulence as $\mathbf{N}_t(\mathbf{x}, t) = \beta \mathbf{n}(\mathbf{x}, t)$, where β controls its strength and $\mathbf{n}$ is a divergence-free field instantiated via curl noise:

$$\mathbf{n}(\mathbf{x}, t) = \nabla \times \psi(\mathbf{x}, t), \quad \nabla \cdot \mathbf{n}(\mathbf{x}, t) = 0, \quad (5)$$

where ψ is derived from a 3D noise field $o(\cdot)$ (e.g., Perlin noise), such as

$$\psi(x, y, z) = (o(x, y, z), o(y, z, x), o(z, x, y)). \quad (6)$$

Finally, we discretize Eq. (4) using an explicit Euler integrator:

$$\mathbf{v}_i^{t+1} = \mathbf{v}_i^t + \frac{\mathbf{F}_w + \mathbf{F}_g + \mathbf{N}_t(\mathbf{x}_i^t, t)}{m} \Delta t, \quad \mathbf{x}_i^{t+1} = \mathbf{x}_i^t + \mathbf{v}_i^{t+1} \Delta t. \quad (7)$$

This formulation yields an expressive dynamics prior: gravity and wind govern global motion, while turbulence introduces locally varying, volume-preserving perturbations for realistic particle behavior.

3.3 Geometry-Grounded Video Synthesis

The simulated particle field is defined in a canonical coordinate system, where gravity is simply a fixed axis. However, the input video is captured in a scene-specific coordinate system: the true “down” direction depends on the 3D orientation of the scene. Additionally, the camera undergoes motion across frames in dynamic scenes. If we directly inject particles into the image plane with a fixed camera in canonical space, raindrops and snowflakes exhibit incorrect falling directions and drift with the camera, leading to implausible weather effects.

To address this, we ground both simulation and synthesis in scene geometry (Fig. 7). In detail, we first employ a geometry foundation model [43] to estimate a 3D point map S_p , per-frame depth maps D , and camera parameters $(K_t, R_t, \mathbf{t}_t)$ from the input video V . Then, we align the particle field with the scene geometry by (i) matching its gravity direction with that of the 3D point map, and (ii) projecting the evolving particle field into each frame using the estimated camera parameters. The resulting particle maps, together with scene geometry and appearance anchors, serve as the final conditioning signals for video synthesis.

Direction matching. Given a point map S_p , we sample 3D points $\mathbf{p}$ and fit the dominant ground plane using point-to-plane distance as the residual. Specifically, we estimate $\boldsymbol{\pi} : \mathbf{c}^\top \mathbf{p} + d = 0$, and count $\mathbf{p}$ as an inlier if $|\mathbf{c}^\top \mathbf{p} + d| < \tau$, where $\mathbf{c}$ is the unit normal and τ is an inlier threshold. Then, we take the scene gravity direction as $\mathbf{d}'_g = -\mathbf{c}$. Let $\mathbf{d}_g$ denote the gravity axis in our canonical simulation. We compute the minimal rotation R_g satisfying $R_g \mathbf{d}_g = \mathbf{d}'_g$, and apply it to particle positions and velocities to obtain the aligned field G' .

Particle projection. For each frame t , we use the estimated camera intrinsics and extrinsics $(K_t, R_t, \mathbf{t}_t)$ to project particle i with its 3D position $\mathbf{x}_i^t$ into the image plane:

$$\mathbf{u}_{i,t} = \pi(K_t, R_t, \mathbf{t}_t, \mathbf{x}_i^t), \quad z_{i,t} = (R_t \mathbf{x}_i^t + \mathbf{t}_t)_z, \quad (8)$$

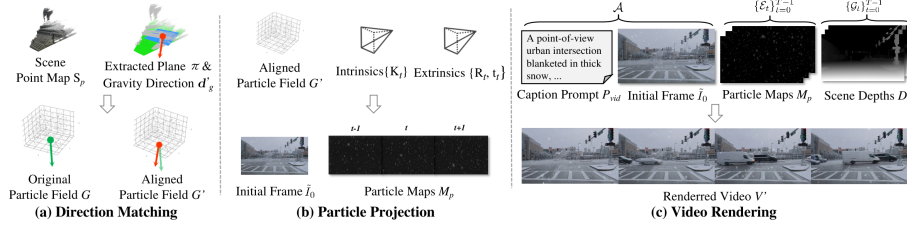


Fig. 7: Geometry-grounded video synthesis. (a) The simulated particle field is aligned with the scene by estimating gravity direction from 3D geometry. (b) The aligned particles are then projected into edited initial frame and each depth frame using camera intrinsics and extrinsics to produce particle maps. (c) Finally, we condition video generation on the particle-augmented inputs together with scene geometry and appearance anchor, yielding plausible and consistent weather effects.

where $\mathbf{u}_{i,t} \in \mathbb{R}^2$ is the pixel coordinate and $z_{i,t}$ is the camera-space depth for scale control. To couple appearance with dynamics (*e.g.*, raindrops elongate along the motion direction), we orient each particle stamp by the projected velocity direction:

$$\theta_{i,t} = \text{atan2}((\mathbf{v}_i^t)^y, (\mathbf{v}_i^t)^x), \quad (9)$$

yielding particle maps M_p that serve as dynamic cues $\mathcal{E}_t$. We also composite the particles onto the edited initial frame as $\tilde{I}_0$ to update appearance anchor $\mathcal{A}$.

Video Synthesis. Finally, we steer a pretrained conditional video editor $\mathcal{M}_{vid}$ with the appearance anchor $\mathcal{A}$, the particle dynamics $\{\mathcal{E}_t\}$, and the geometric cues $\{\mathcal{G}_t\}$ to synthesize final video:

$$\tilde{V} = \mathcal{M}_{vid}(V; \mathcal{A}, \{\mathcal{E}_t\}_{t=0}^{T-1}, \{\mathcal{G}_t\}_{t=0}^{T-1}). \quad (10)$$

The resulting tri-prior conditioning yields disentangled control over global atmosphere, particle dynamics, and geometry-consistent placement. This bridges explicit physical simulation and implicit generative modeling, enabling realistic weather synthesis with flexible control.

4 Experiments

4.1 Implementation Details

Our implementation follows the three-stage design in Sec. 3: appearance anchoring, particle simulation, and geometry grounding. For appearance anchoring, we use GPT-4.1 [1] to parse the initial frame into a structured scene description P_{scene} , and to reason about the target weather effects, yielding an image-editing instruction and a compact video prompt; FLUX.1-Kontext Pro [38] then edits the initial frame to produce the appearance anchor. For geometry grounding, we adopt Depth Anything V3 [43] to obtain geometry information and drive the particle simulation. We use Wan2.1-VACE [27] as the conditional video diffusion



Fig. 8: Qualitative comparison under different weather conditions. Top: snowy scenes. Bottom: rainy scenes. Particle details are highlighted in colored boxes.

model with its default inference settings. The preprocessing steps take ~ 3 minutes per video, while the overall runtime is dominated by video synthesis at ~ 38 minutes. All experiments are conducted on a single NVIDIA RTX A6000 GPU.

4.2 Dataset

To evaluate generalizability and robustness, we curate an evaluation dataset consisting of 58 videos from DAVIS [53], DL3DV-10k [44], PandaSet [65], and nuScenes [7]. It spans a wide range of real-world scenarios captured from aerial and ego-centric perspectives, encompassing both static and dynamic camera setups with diverse motion trajectories. This ensures diversity in scene types, geometric complexity, and illumination, all of which directly affect atmospheric appearance and particle visibility. All sequences are temporally trimmed and processed with a unified preprocessing protocol for fair evaluation. Additionally, we use ACDC [58] and MUSES [6] for downstream task evaluation, covering diverse weather conditions in autonomous driving scenarios.

Table 1: Quantitative evaluation under different weather conditions. We conduct CLIP-based and VLM-based evaluations for snow and rain settings. The overall score is averaged across conditions.

Method	Snow			Rain			Overall		
	CLIP-D $\uparrow$	CLIP-S $\uparrow$	VLM $\uparrow$	CLIP-D $\uparrow$	CLIP-S $\uparrow$	VLM $\uparrow$	CLIP-D $\uparrow$	CLIP-S $\uparrow$	VLM $\uparrow$
CogVideoX [70]	0.08	0.82	0.56	0.06	0.82	0.52	0.07	0.82	0.54
WeatherEdit [55]	0.14	0.90	0.45	0.12	0.92	0.36	0.13	0.91	0.41
LTX-Video [21]	0.15	0.80	0.68	0.12	0.83	0.67	0.13	0.82	0.67
VACE [27]	0.18	0.80	0.76	0.14	0.83	0.74	0.16	0.82	0.75
Ours	0.19	0.79	0.84	0.15	0.82	0.81	0.17	0.81	0.83

4.3 Evaluation Metrics

For each input video, we evaluate synthesized rain and snow effects using metrics that measure semantic preservation, prompt alignment, and perceptual quality. For semantic preservation, we measure CLIP similarity [57] (CLIP-S) between edited and input videos. For prompt alignment, we compute CLIP directional alignment [16] (CLIP-D) between edited frames and the editing prompt. We further employ a VLM model GPT-4V [1] for evaluating high-level semantic alignment and overall perceived quality. In addition, we conduct a human evaluation to assess perceptual realism and collect user preferences for selecting the best result. Finally, we report mIoU for semantic segmentation evaluation.

4.4 Results

We compare our method against two groups of representative video editing methods. The first stream consists of prompt-only methods, such as CogVideoX [70] and WeatherEdit [55], which take an input RGB video along with a text prompt to generate the edited sequence. The second stream includes conditional editors, such as LTX-Video [21] and VACE [27], which incorporate inputs including reference images and depth maps, to guide the editing process. For a fair comparison on fine-grained particle effects, we provide all conditional baselines with the same text prompt as our method and the same reference image but with particle conditioning removed.

Qualitative comparison. For snowy scenes, prompt-only methods such as CogVideoX [70] fail to produce convincing snow effects, while WeatherEdit [55] only yields mild whitening of the input, as illustrated in Fig. 8. Although conditioning-based editors such as LTX-Video [21] and VACE [27] better match the global snowy tone, they rarely synthesize dense particle effects with falling snowflakes. For rainy weather, CogVideoX [70] and WeatherEdit [55] darken the scene and increase surface reflectance (*e.g.*, the dancing sequence), yet their edits are inconsistent across scenes (*e.g.*, limited changes on the statue scene). Although LTX-Video [21] and VACE [27] produce global tone change, they fail to synthesize visible rain streaks in a stable manner.

In general, prompt-only editing tends to underexpress heavy weather degradations, while conditional editors often prioritize background appearance and

Table 2: Realism rating. The rating is based on a 5-point Likert scale. Our method achieves the highest scores on both criteria.

Method	Photo-realism $\uparrow$	Physical-realism $\uparrow$
CogVideoX [70]	2.51	2.66
WeatherEdit [55]	2.50	2.54
LTX-Video [21]	2.87	2.72
VACE [27]	3.58	3.52
Ours	4.16	4.12

Table 3: User preference. Percentage of selections as best under rain and snow.

Method	Rain(%)	Snow(%)
CogVideoX [70]	8.7	2.2
WeatherEdit [55]	15.2	10.9
LTX-Video [21]	6.5	4.3
VACE [27]	28.3	26.1
Ours	41.3	56.5

suppress fine-grained particles. In contrast, our method can effectively edit both the general atmosphere and detail particles, producing realistic weather synthesis results across diverse conditions.

Quantitative comparison. As shown in Tab. 1, our method achieves the highest CLIP-D score across all settings, indicating the closest alignment between the synthesized weather effects and the desired target conditions. Moreover, the top ratings from the VLM-based evaluator suggest superior overall perceptual quality, with more convincing and visually coherent weather phenomena. Notably, WeatherEdit [55] attains the highest CLIP-S score, reflecting strong preservation of the original scene. However, this largely stems from its conservative editing behavior, which introduces only subtle weather effects. VACE [27] and LTX-Video [21] exhibit a similar trend: they maintain structure consistency, yet produce weaker and less realistic weather effects, as evidenced by lower CLIP-D scores compared to ours. In contrast, our method introduces substantial atmospheric changes with dense particle effects while largely preserving scene identity, achieving a balance between weather expressiveness and scene fidelity.

Human Evaluation. To assess both photorealism and physical realism of the generated videos, we follow prior work [46] to conduct human evaluation. Photorealism evaluates perceptual quality and alignment with the target condition, while physical realism measures whether the weather dynamics (*e.g.*, particle motion and scene interactions) are physically plausible. We recruit 46 participants and ask them to rate each video on a 5-point Likert scale with 1 indicating strongly disagree and 5 denoting strongly agree for the statements “the video is physically realistic” and “the video is photorealistic”. To reduce ordering bias, the results are presented in random order. In Tab. 2, our method receives the highest average ratings on both criteria, indicating strong physical plausibility and visual quality. We additionally collect the user preference results in which participants select the most preferred edited result among competing methods. As shown in Tab. 3, our method is selected most frequently under both weather types, receiving 41.3% of votes in rain and 56.5% in snow. This suggests a clear advantage in overall perceived quality, especially for more challenging snow cases where fine-grained particles and realistic dynamics are critical.

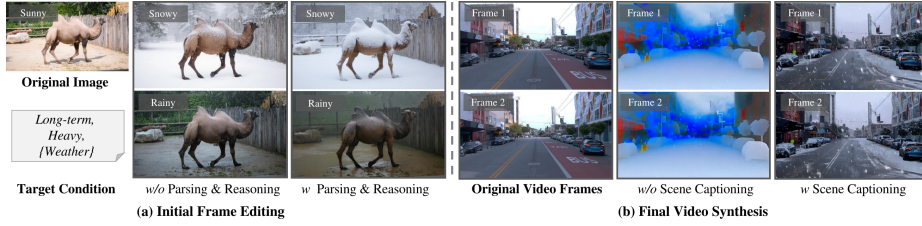


Fig. 9: Ablation on appearance anchoring. (a) Without semantic parsing and effect reasoning to generate a detailed prompt, results either change background content (snowy) or produce limited effects on the subject (rainy). With our prompt, the desired effects are achieved while retaining scene structure. (b) Without detailed prompts derived from scene captioning, results often exhibit hallucinations. With captioning, the model preserves scene identity and faithfully reflects the intended weather edits.

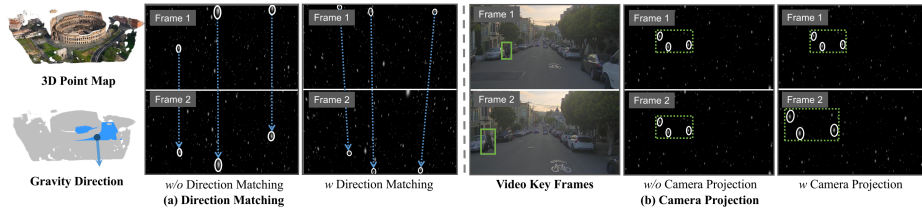


Fig. 10: Ablation on geometry grounding. Representative particles are highlighted in white. (Left) The blue solid arrow denotes the true gravity direction in 3D, and the dotted arrows show falling trends in 2D. (Right) The green solid boxes indicate the reference object scale, while the dashed boxes represent the particle-group scale.

4.5 Ablation Study

We validate our key designs in (i) appearance anchoring (semantic parsing, effect reasoning, and scene captioning) and (ii) geometry grounding (gravity direction matching and particle camera projection), indicating that each component is critical to achieving diverse, dynamic, and physically plausible weather synthesis.

Appearance anchoring. In Fig. 9, for initial frame editing, we compare a naive prompt (*e.g.*, “make it long-term, heavy {weather}”) with a detailed prompt generated by semantic parsing and effects reasoning. Under snow, the naive prompt exhibits hallucinations and alters the background. Under rain, it fails to induce heavy precipitation, leaving the camel visually dry. In contrast, our prompt yields more faithful atmospheric cues including low-lighting, wetness and haze while preserving scene identity. For final video synthesis, we compare a naive prompt (“a heavy snowy driving scene with snowflakes.”) with our captioned prompt. Without detailed guidance, the model is prone to hallucinations because injected particles alter the depth map and can contradict the model’s structural prior. The detailed prompt provides stronger semantic constraints, helping the model interpret these perturbations as weather effects, yielding the desired, visually consistent results across frames.

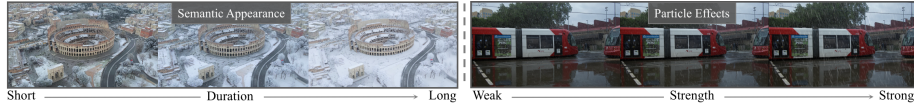


Fig. 11: Examples of controllable weather effects. (Left) User inputs specify the weather duration category (short/medium/long) to control overall semantic appearance. (Right) Simulation parameters adjust particle shape and motion to reflect wind strength (weak to strong), enabling fine-grained particle control.

Method	ACDC		MUSES	
	Snow	Rain	Snow	Rain
DAFormer	50.9	50.8	41.4	33.8
HRDA	50.0	51.6	40.4	35.4
DAFormer+ours	56.3 +5.4	56.6 +5.8	47.2 +5.8	41.5 +7.7
HRDA+ours	60.4 +10.4	66.1 +14.5	50.4 +10.0	44.4 +9.0

Table 4: Quantitative segmentation results for baselines and models augmented with our data. Improvements over the baseline are shown in green.

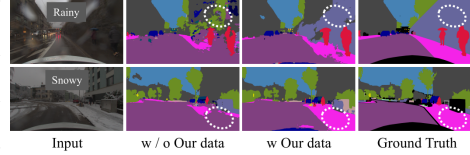


Fig. 12: Qualitative segmentation results comparison under adverse weather: w/ vs. w/o our augmentation. Improved regions are highlighted in white circles.

Geometry grounding. As shown in Fig. 10, we ablate the influence of gravity direction matching and camera-consistent projection on particle motion in geometry grounding. For clarity, we visualize particle trajectories using the rendered particle maps. Without gravity direction alignment, snowflake motion is dominated by the canonical axis, yielding vertical falling; with alignment, particles follow the real scene gravity direction estimated from the fused point map, producing consistent trajectories under tilted cameras. To demonstrate the effectiveness of camera projection, we disable physics-driven motion and keep only motion induced by the camera. In a forward-driving sequence, particles should exhibit strong parallax and appear to approach the viewer; without camera projection, particles remain stationary relative to the image, whereas with projection, they move coherently with ego-motion, yielding geometry-consistent particle motion.

4.6 Applications

By leveraging the structured conditioning, our design enables flexible and interpretable control over both semantic appearance and particle effects, as shown in Fig. 11. This disentangled control allows users to flexibly synthesize diverse weather conditions for the same scene, ranging from subtle environment changes to extreme conditions with dense particles. We further demonstrate the utility of our synthesized data for augmenting semantic segmentation in the domain generalization setting. When training DAFormer [25] and HRDA [26] with our generated data, we observe consistent improvements across evaluation settings, boosting mIoU by at least 5.4% and up to 14.5% (Tab. 4). Qualitatively, our augmentation mitigates prediction errors of the HRDA [26] baseline on MUSES [6], yielding cleaner, more complete segmentation masks (Fig. 12). These results in-

indicate that our synthesized weather data provides a valuable supplement for training perception models robust to adverse weather.

5 Limitations

Our method builds upon several upstream modules, including first-frame anchoring, geometry estimation, gravity alignment, and particle projection. Errors at any of these stages may propagate directly into the final video. For instance, the geometry grounding relies on recovering a dominant ground plane to estimate gravity, which can be fragile in scenes that lack clear ground structure or involve challenging viewpoints. Moreover, the achievable weather particle effects are inherently bounded by the pretrained video diffusion model, occasionally leading to failed particle generation. Future work may address these limitations by jointly optimizing the upstream modules to mitigate cascading errors and incorporating stronger or domain-adapted diffusion priors to improve the fidelity and controllability of particle effects.

6 Conclusion

In this work, we propose a semantic-aware, physics-informed, and geometry-grounded framework for controllable weather synthesis in real-world videos. We factorize the synthesis into a structured tri-prior interface comprising semantics, dynamics, and geometry, enabling precise control over both global atmospheric appearance and fine-grained particle motion. This design produces diverse and physically plausible weather effects across varied scenes, effectively eliciting weather priors that are often underexpressed with text-only prompting. Furthermore, the resulting high-fidelity synthetic data improves downstream robustness, yielding up to 14.5% mIoU improvement in semantic segmentation under extreme weather conditions. It highlights the effectiveness of our method as a scalable engine for generating diverse, safety-critical corner cases, enabling controllable synthesis of rare adverse weather conditions that are difficult to capture in the real world, thereby improving downstream perception robustness for resilient autonomous systems.

Acknowledgment

This work was funded by the National Natural Science Foundation of China (Nos. 62572166 and 62402157) and the Fundamental Research Funds for the Central Universities (No. JZ2025HG TB0219). It was also partially supported by the Ministry of Education and Science of Bulgaria (support for INSAIT, part of the Bulgarian National Roadmap for Research Infrastructure).

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: GPT-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Alhaija, H.A., Alvarez, J., Bala, M., Cai, T., Cao, T., Cha, L., Chen, J., Chen, M., Ferroni, F., Fidler, S., et al.: Cosmos-Transfer1: Conditional world generation with adaptive multimodal control. arXiv preprint arXiv:2503.14492 (2025)
3. 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)
4. Bahmani, S., Skorokhodov, I., Siarohin, A., Menapace, W., Qian, G., Vasilkovsky, M., Lee, H.Y., Wang, C., Zou, J., Tagliasacchi, A., et al.: VD3D: Taming large video diffusion transformers for 3D camera control. arXiv preprint arXiv:2407.12781 (2024)
5. Bai, J., Xia, M., Fu, X., Wang, X., Mu, L., Cao, J., Liu, Z., Hu, H., Bai, X., Wan, P., et al.: ReCamMaster: Camera-controlled generative rendering from a single video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14834–14844 (2025)
6. Brödermann, T., Bruggemann, D., Sakaridis, C., Ta, K., Liagouris, O., Corkill, J., Van Gool, L.: MUSES: The multi-sensor semantic perception dataset for driving under uncertainty. In: European Conference on Computer Vision. pp. 21–38. Springer (2024)
7. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuScenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11621–11631 (2020)
8. Chen, W., Ji, Y., Wu, J., Wu, H., Xie, P., Li, J., Xia, X., Xiao, X., Lin, L.: Control-a-video: Controllable text-to-video diffusion models with motion prior and reward feedback learning. arXiv preprint arXiv:2305.13840 (2023)
9. Chen, Y., He, X., Ma, X., Ma, Y.: ContextFlow: Training-free video object editing via adaptive context enrichment. arXiv preprint arXiv:2509.17818 (2025)
10. Chow, W., Li, L., Kong, L., Li, Z., Xu, Q., Song, H., Ye, T., Wang, X., Bai, J., Xu, S., Li, X., Pan, J., Liu, S., Zhou, R., Yang, T., Liu, S.: EditMGT: Unleashing potentials of masked generative transformers in image editing. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 38038–38048 (2026)
11. Cong, Y., Xu, M., Simon, C., Chen, S., Ren, J., Xie, Y., Perez-Rua, J.M., Rosenhahn, B., Xiang, T., He, S.: FLATTEN: optical flow-guided attention for consistent text-to-video editing. arXiv preprint arXiv:2310.05922 (2023)
12. 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 Conference on Computer Vision and Pattern Recognition. pp. 16153–16162 (2025)
13. Engine, U.: Unreal Engine. Retrieved from Unreal Engine: <https://www.unrealengine.com/en-US/what-is-unreal-engine-4> (2018)
14. Feng, R., Weng, W., Wang, Y., Yuan, Y., Bao, J., Luo, C., Chen, Z., Guo, B.: CCEdit: Creative and controllable video editing via diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6712–6722 (2024)
15. Fiebelman, G., Averbuch-Elor, H., Benaim, S.: Let it snow! animating 3D gaussian scenes with dynamic weather effects via physics-guided score distillation. arXiv preprint arXiv:2504.05296 (2025)

16. Gal, R., Patashnik, O., Maron, H., Bermano, A.H., Chechik, G., Cohen-Or, D.: StyleGAN-NADA: CLIP-guided domain adaptation of image generators. *ACM Transactions on Graphics* **41**(4), 1–13 (2022)
17. Gan, Q., Ren, Y., Zhang, C., Ye, Z., Xie, P., Yin, X., Yuan, Z., Peng, B., Zhu, J.: HumanDiT: Pose-guided diffusion transformer for long-form human motion video generation. *arXiv preprint arXiv:2502.04847* (2025)
18. Gao, X., Li, R., Chen, X., Wu, Y., Feng, S., Yin, Q., Tu, Z.: PISCO: Precise video instance insertion with sparse control. *arXiv preprint arXiv:2602.08277* (2026)
19. Gissler, C., Henne, A., Band, S., Peer, A., Teschner, M.: An implicit compressible SPH solver for snow simulation. *ACM Transactions on Graphics* **39**(4), 36–1 (2020)
20. Guo, Y., Yang, C., Rao, A., Agrawala, M., Lin, D., Dai, B.: SparseCtrl: Adding sparse controls to text-to-video diffusion models. In: *European Conference on Computer Vision*. pp. 330–348. Springer (2024)
21. HaCohen, Y., Chiprut, N., Brazowski, B., Shalem, D., Moshe, D., Richardson, E., Levin, E., Shiran, G., Zabari, N., Gordon, O., Panet, P., Weissbuch, S., Kulikov, V., Bitterman, Y., Melumian, Z., Bibi, O.: LTX-Video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103* (2024)
22. Hahner, M., Dai, D., Sakaridis, C., Zaeck, J.N., Van Gool, L.: Semantic understanding of foggy scenes with purely synthetic data. In: *IEEE Intelligent Transportation Systems Conference*. pp. 3675–3681 (2019)
23. Hao, X., et al.: Is your HD map constructor reliable under sensor corruptions? In: *Advances in Neural Information Processing Systems*. vol. 37, pp. 22441–22482 (2024)
24. Hao, X., et al.: SafeMap: Robust HD map construction from incomplete observations. In: *International Conference on Machine Learning*. pp. 22091–22102. PMLR (2025)
25. Hoyer, L., Dai, D., Van Gool, L.: DAFormer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 9924–9935 (2022)
26. 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)
27. Jiang, Z., Han, Z., Mao, C., Zhang, J., Pan, Y., Liu, Y.: VACE: All-in-one video creation and editing. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 17191–17202 (2025)
28. Jones, M., Abdal, R., Patashnik, O., Salakhutdinov, R., Tulyakov, S., Zhu, J.Y., Wang, K.C.J.: Tuning-free visual effect transfer across videos. *arXiv preprint arXiv:2601.07833* (2026)
29. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G., et al.: 3D Gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics* **42**(4), 139–1 (2023)
30. Kong, L., Liu, Y., Li, X., Chen, R., Zhang, W., Ren, J., Pan, L., Chen, K., Liu, Z.: Robo3D: Towards robust and reliable 3D perception against corruptions. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 19994–20006 (2023)
31. Kong, L., Xie, S., Hu, H., Ng, L.X., Cottureau, B.R., Ooi, W.T.: RoboDepth: Robust out-of-distribution depth estimation under corruptions. In: *Advances in Neural Information Processing Systems*. vol. 36, pp. 21298–21342 (2023)

32. Kong, L., Xie, S., Hu, H., Niu, Y., Ooi, W.T., Cottureau, B.R., Ng, L.X., Ma, Y., Zhang, W., Pan, L., Chen, K., Liu, Z., Qiu, W., Zhang, W., Cao, X., Lu, H., Chen, Y.C., Kang, C., Zhou, X., Ying, C., Shang, W., Wei, X., Dong, Y., Yang, B., Jiang, S., Ma, Z., Ji, D., Li, H., Huang, X., Tian, Y., Kou, G., Jia, F., Liu, Y., Wang, T., Li, Y., Hao, X., Yang, Y., Zhang, H., Wei, M., Zhou, Y., Zhao, H., Zhang, J., Li, J., He, X., Cheng, X., Zhang, B., Zhao, L., Ding, D., Liu, F., Yan, Y., Wang, H., Ye, N., Luo, L., Tian, Y., Zuo, Y., Cao, Z., Ren, Y., Li, Y., Liu, W., Wu, X., Mao, Y., Li, M., Liu, J., Liu, J., Qin, Z., Chu, C., Xu, J., Zhao, W., Jiang, J., Liu, X., Wang, Z., Li, C., Li, S., Yuan, C., Yang, S., Liu, W., Chen, P., Zhou, B., Wang, Y., Zhang, C., Sun, J., Chen, H., Yang, X., Wang, L., Fu, D., Lin, Y., Yang, H., Li, H., Luo, Y., Cheng, X., Xu, Y.: The RoboDrive challenge: Drive anytime anywhere in any condition. *arXiv preprint arXiv:2405.08816* (2024)
33. Kong, L., Xu, X., Ren, J., Zhang, W., Pan, L., Chen, K., Ooi, W.T., Liu, Z.: Multi-modal data-efficient 3D scene understanding for autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **47**(5), 3748–3765 (2025)
34. Kong, L., Yang, W., Mei, J., Liu, Y., Liang, A., Zhu, D., Lu, D., Yin, W., Hu, X., Jia, M., Deng, J., Zhang, K., Wu, Y., Yan, T., Gao, S., Wang, S., Li, L., Pan, L., Liu, Y., Zhu, J., Ooi, W.T., Hoi, S.C.H., Liu, Z.: 3D and 4D world modeling: A survey. *arXiv preprint arXiv:2509.07996* (2025)
35. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: HunyuanVideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603* (2024)
36. Ku, M., Wei, C., Ren, W., Yang, H., Chen, W.: AnyV2V: A tuning-free framework for any video-to-video editing tasks. *arXiv preprint arXiv:2403.14468* (2024)
37. Kuang, Z., Cai, S., He, H., Xu, Y., Li, H., Guibas, L.J., Wetzstein, G.: Collaborative video diffusion: Consistent multi-video generation with camera control. *Advances in Neural Information Processing Systems* **37**, 16240–16271 (2024)
38. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., et al.: FLUX. 1 Kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742* (2025)
39. Li, X., Kou, K., Zhao, B.: Weather GAN: Multi-domain weather translation using generative adversarial networks. *arXiv preprint arXiv:2103.05422* (2021)
40. 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)
41. Liang, A., Kong, L., Yan, T., Liu, H., Yang, W., Huang, Z., Yin, W., Zuo, J., Hu, Y., Zhu, D., Lu, D., Liu, Y., Jiang, G., Li, L., Li, X., Zhuo, L., Ng, L.X., Cottureau, B.R., Gao, C., Pan, L., Ooi, W.T., Liu, Z.: WorldLens: Full-spectrum evaluations of driving world models in real world. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 36385–36399 (2026)
42. 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)
43. Lin, H., Chen, S., Liew, J., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647* (2025)
44. Ling, L., Sheng, Y., Tu, Z., Zhao, W., Xin, C., Wan, K., Yu, L., Guo, Q., Yu, Z., Lu, Y., et al.: DL3DV-10K: A large-scale scene dataset for deep learning-based

- 3D vision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22160–22169 (2024)
45. Liu, C., Li, R., Zhang, K., Lan, Y., Liu, D.: StableV2V: Stabilizing shape consistency in video-to-video editing. *IEEE Transactions on Circuits and Systems for Video Technology* (2025)
 46. Liu, S., Ren, Z., Gupta, S., Wang, S.: PhysGen: Rigid-body physics-grounded image-to-video generation. In: European Conference on Computer Vision. pp. 360–378. Springer (2024)
 47. Lu, Y., Ren, X., Yang, J., Shen, T., Wu, Z., Gao, J., Wang, Y., Chen, S., Chen, M., Fidler, S., et al.: InfiniCube: Unbounded and controllable dynamic 3D driving scene generation with world-guided video models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 27272–27283 (2025)
 48. Ma, Y., He, Y., Cun, X., Wang, X., Chen, S., Li, X., Chen, Q.: Follow your pose: Pose-guided text-to-video generation using pose-free videos. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 4117–4125 (2024)
 49. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: NeRF: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM* **65**(1), 99–106 (2021)
 50. Nayar, S.K., Narasimhan, S.G.: Vision in bad weather. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. vol. 2, pp. 820–827 (1999)
 51. Pang, Y., Zhu, B., Lin, B., Zheng, M., Tay, F.E., Lim, S.N., Yang, H., Yuan, L.: DreamDance: Animating human images by enriching 3D geometry cues from 2D poses. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14039–14050 (2025)
 52. Parmar, G., Park, T., Narasimhan, S., Zhu, J.Y.: One-step image translation with text-to-image models. *arXiv preprint arXiv:2403.12036* (2024)
 53. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 724–732 (2016)
 54. Qian, C., Guo, Y., Mo, Y., Li, W.: WeatherDG: LLM-assisted procedural weather generation for domain-generalized semantic segmentation. *IEEE Robotics and Automation Letters* **10**, 5919–5926 (2025)
 55. Qian, C., Li, W., Guo, Y., Markkula, G.: WeatherEdit: Controllable weather editing with 4D gaussian field. *arXiv preprint arXiv:2505.20471* (2025)
 56. Qian, C., Rezaei, M., Anwar, S., Li, W., Hussain, T., Azarmi, M., Wang, W.: AllWeather-Net: Unified image enhancement for autonomous driving under adverse weather and low-light conditions. In: International Conference on Pattern Recognition. pp. 151–166 (2024)
 57. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
 58. 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)
 59. 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)

60. Schmidt, V., Luccioni, A.S., Teng, M., Zhang, T., Reynaud, A., Raghupathi, S., Cosne, G., Juraver, A., Vardanyan, V., Hernandez-Garcia, A., et al.: ClimateGAN: Raising climate change awareness by generating images of floods. arXiv preprint arXiv:2110.02871 (2021)
61. Stomakhin, A., Schroeder, C., Chai, L., Teran, J., Selle, A.: A material point method for snow simulation. *ACM Transactions on Graphics* **32**(4), 1–10 (2013)
62. Tremblay, M., Halder, S.S., De Charette, R., Lalonde, J.F.: Rain rendering for evaluating and improving robustness to bad weather. *International Journal of Computer Vision* **129**(2), 341–360 (2021)
63. Wang, Y., Wang, L., Ma, Z., Hu, Q., Xu, K., Guo, Y.: VideoDirector: Precise video editing via text-to-video models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 2589–2598 (2025)
64. Wu, W., Guan, H., Liu, Z., Wang, H.: Weathercity: Urban scene reconstruction with controllable multi-weather transformation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 40949–40958 (2026)
65. Xiao, P., Shao, Z., Hao, S., Zhang, Z., Chai, X., Jiao, J., Li, Z., Wu, J., Sun, K., Jiang, K., et al.: PandaSet: Advanced sensor suite dataset for autonomous driving. In: *IEEE International Intelligent Transportation Systems Conference*. pp. 3095–3101 (2021)
66. Xie, S., Kong, L., Dong, Y., Sima, C., Zhang, W., Chen, Q.A., Liu, Z., Pan, L.: Are VLMs ready for autonomous driving? an empirical study from the reliability, data, and metric perspectives. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 6585–6597 (2025)
67. Xie, S., Kong, L., Zhang, W., Ren, J., Pan, L., Chen, K., Liu, Z.: Benchmarking and improving bird's eye view perception robustness in autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **47**(5), 3878–3894 (2025)
68. Xie, Y., Zhang, M., Hao, Q.: ClimateGS: Real-time climate simulation with 3D gaussian style transfer. arXiv preprint arXiv:2503.14845 (2025)
69. Xing, J., Xia, M., Liu, Y., Zhang, Y., Zhang, Y., He, Y., Liu, H., Chen, H., Cun, X., Wang, X., et al.: Make-your-video: Customized video generation using textual and structural guidance. *IEEE Transactions on Visualization and Computer Graphics* **31**(2), 1526–1541 (2024)
70. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: CogVideoX: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)