G-WADE: GEOMETRY-PRESERVING WEATHER-STYLIZED URBAN ASSETS DIFFUSION ENHANCEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Extending 3D Gaussian Splatting (3DGS) urban assets under adverse weather (e.g., rain, snow) directly from real-world captures is prohibitively expensive, motivating the alternative of editing an existing, geometrically accurate 3DGS scene to synthesize new weather conditions. However, existing weather editing methods are limited in asset augmentation and prone to geometry collapse, multi-view texture inconsistency, and long-range drift. We observe that video diffusion models with depth/edge control can preserve scene geometry while re-stylizing appearance, but often produce substantial appearance discrepancies when revisiting previously seen regions, undermining multi-view consistency and degrading the quality of the resulting 3DGS reconstructions. In this paper, we propose G-WADE: Geometry-preserving Weather-stylized urban Assets Diffusion Enhancement, which is a geometry-fixed framework for the augmentation of weather-stylized urban assets. G-WADE fine-tunes two domain-specific variants of Cosmos-Transfer2.5 (Base-Model) by leveraging VAE temporal downsampling property: a *first-frame*-conditioned model (F-Model) and a *first-last-frame*-conditioned model (FL-Model) that extracts a tail latent slice as an additional anchor. Additionally, G-WADE designs trajectory-aware generation strategies with coordinated model scheduling, separately tailored for high-precision single-building capture (orbit-and-descend trajectories) and large-scale urban scenes (center-out grid trajectories). The generated multi-view-consistent, geometry-preserving frames are then used to optimize the attributes of the original 3DGS. Experiments on rain and snow weather stylization over both single-building and large-scale urban 3DGS scenes show that our method achieves effective urban asset augmentation, with better multi-view consistency, perceptual realism, and geometry preservation than existing video and 3D editing baselines.

1 INTRODUCTION

Reconstructing photorealistic, drivable 3D scenes from drone-captured imagery using 3D Gaussian Splatting (3DGS) (Kerbl et al., 2023) has become a cornerstone technique for digital twins, autonomous-driving simulation, and embodied-AI training. A recurring practical bottleneck, however, is weather coverage: collecting real aerial footage of the same city block under clear skies, heavy rain, and snow is costly, time-consuming, and often physically infeasible within a single project timeline. This motivates an appealing alternative: given a geometrically accurate 3DGS reconstruction captured under nominal weather, can we edit its appearance to plausibly and consistently synthesize novel weather conditions, without ever re-flying the drone or re-solving the scene geometry? Such a capability would allow a single high-quality capture to be repurposed for all-weather simulation and rendering, at a fraction of the data-collection cost.

However, existing 3DGS editing methods fall short of this goal. Early approaches built on 2D image diffusion priors for 3D scene editing, such as Instruct-GS2GS style pipelines (Haque et al., 2023; Vachha & Haque, 2024; Chen et al., 2024b), edit each training view independently (or with only weak cross-view regularization), and thus lack the multi-view coherence required for a globally consistent appearance. Later approaches like DGE (Chen et al., 2024a) sample multi-view consistent edits and extend to the remaining views, improving cross-view consistency. Even so, they

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Figure 1: **Renderings of Stylized 3D City Scenes Generated by G-WADE**. We present G-WADE, a geometry-fixed stylization framework for augmenting 3D urban assets with different weather conditions. Given a pre-trained 3D Gaussian Splatting (3DGS) scene, G-WADE enables flexible weather stylization while preserving the underlying scene geometry and producing highly multi-view-consistent and photorealistic renderings.

struggle to handle substantial geometric transformations and suffer from long-range drift. Nevertheless, None of these approaches consider large-range weather editing tasks. Recently, large-scale video diffusion models exhibit greater potential for weather editing. Some works (Fiebelman et al., 2026; Wu et al., 2026) explores weather editing for pre-constructed 3DGS with Video Score Distillation Sampling (Video-SDS) to add dynamic rain and snow, yet they are incapable of altering the attributes of the original Gaussians because of geometric consistency problems. When it comes to the geometry-fixed weather editing, our key observation is that Cosmos-Transfer2.5 (NVIDIA et al., 2025b) can offer strong geometry preservation because the control signal is derived from the original scene's depth/edge maps. However, Cosmos-Transfer2.5 (Base-Model) only conditions generation on the first frame of the video. As the camera trajectory extends over tens to hundreds of frames covering a large urban area, small stylization inconsistencies compound frame-by-frame, producing severe appearance drift. When such drifting video frames are later used as multi-view supervision to optimize a 3DGS scene, overlapping regions observed from different frames disagree, resulting in pronounced ghosting artifacts and spurious floating Gaussians.

In this paper, we tackle geometry-fixed weather augmentation of 3DGS scenes by directly addressing drift in multi-chunk video generation. Our key insight is that the $1{+}4n$ temporal downsampling structure of the Wan 2.1 VAE used in Cosmos-Transfer2.5, which maps $1{+}4n$ RGB frames to $n{+}1$ temporal latent slices, provides a natural way to extract a compact *tail latent slice* from the end of a reference video. We inject this slice as an additional endpoint condition, yielding a first-last-frame-conditioned model (FL-Model) that anchors both ends of a generated chunk, alongside a first-frame-conditioned model (F-Model) for appearance propagation from a known starting view. Building on these conditioning modes, we design trajectory-aware generation strategies that progressively reuse generated views as anchors to suppress drift over interconnected video chunks. Specifically, we adopt an orbit-and-descend strategy for standalone buildings and a center-out grid strategy for large-scale urban scenes, where axis trajectories establish initial anchors and quadrant-wise generation expands the stylized region outward. Finally, we optimize only the appearance attributes of the original 3DGS while freezing its position, opacity, rotation, and scale, thereby preserving the original scene geometry throughout weather stylization.

Our contributions are summarized as follows:

- We propose a tail-frame latent conditioning strategy that reuses a temporal latent slice from the end of the preceding video chunk as an additional condition for the next chunk, yielding first-frame (F-Model) and first-last-frame (FL-Model) conditioned video diffusion models that effectively suppress drift accumulation during multi-chunk video generation.

- We develop a data preparation pipeline for multi-condition urban weather stylization using paired RGB, depth, and edge videos together with a fixed set of LLM-generated rain and snow prompts to fine-tune a Cosmos-Transfer2.5 backbone for geometry-preserving weather stylization.
- We design trajectory-aware generation strategies tailored to both single-building and large-scale urban aerial scenes and demonstrate state-of-the-art multi-view consistency and geometric fidelity for rain and snow weather stylization.

2 Related Work

2.1 Video Diffusion Backbones for Stylization

Large-scale video diffusion models such as Sora (Brooks et al., 2024), Stable Video Diffusion (Blattmann et al., 2023), and the Cosmos world-foundation model family (NVIDIA et al., 2025a) have demonstrated strong capabilities for controllable video generation and stylization. Control mechanisms including AnimateDiff-style motion modules (Guo et al., 2024), ControlVideo (Zhang et al., 2024), and ControlNet-style spatial conditioning extended to video (Zhang et al., 2023) allow structural signals (e.g., depth, edge, or pose maps) to guide generation while altering appearance, which is directly applicable to geometry-preserving weather stylization. Recent works also explore specialized conditioning for stylization, such as video-to-video translation with explicit temporal correspondence (Geyer et al., 2024; Qi et al., 2023) and attention-based feature injection that preserves structural layout while swapping texture and atmosphere (Wang et al., 2023). In particular, Cosmos-Transfer2.5 (NVIDIA et al., 2025b) supports depth- and edge-conditioned video-to-video transfer, making it a natural backbone for our task. However, the native Cosmos-Transfer2.5 model conditions only on the first frame. This design is known to suffer from compounding drift over long or closed-loop sequences, which is directly addressed by our first-and-last-frame conditioning.

2.2 Diffusion-Guided 3D Scene Editing

Diffusion models (Nichol et al., 2022; Brooks et al., 2023) have driven significant progress in image editing. Building on these advances, early text-driven 3D editing methods adopt two predominant paradigms. In the Iterative Dataset Update (IDU) framework, 2D edits are applied to rendered views and back-propagated to the 3D representation iteratively. Instruct-NeRF2NeRF (Haque et al., 2023) pioneered this approach on Neural Radiance Fields (Mildenhall et al., 2020), while Instruct-GS2GS (Vachha & Haque, 2024) and GaussianEditor (Chen et al., 2024b) later adapted it to 3D Gaussian Splatting (Kerbl et al., 2023) for efficiency. In parallel, Score Distillation Sampling (SDS)-based methods, such as DreamEditor (Zhuang et al., 2023), FocalDreamer (Li et al., 2024), and Progressive3D (Cheng et al., 2024), distill 2D diffusion priors as optimization signals to directly update the 3D representation. However, both families process views independently, often producing artifacts under novel viewpoints. To address this, recent works explicitly enforce multi-view consistency: DGE (Chen et al., 2024a) shares cross-view attention during iterative editing, EditSplat (Lee et al., 2025) leverages 3D geometry priors for coherent inpainting, and GaussCtrl (Wu et al., 2024) couples 2D diffusion control with explicit Gaussian constraints. Other environmental manipulation tackles holistic scene-level appearance changes such as relighting and season translation, exemplified by WeatherCity (Wu et al., 2026), which disentangles scene geometry and weather appearance, yet its naive 2D image editing initialization can still introduce multi-view inconsistency and blurry Gaussians. Despite these advances, existing methods either rely on per-view 2D diffusion edits without multi-view consistency, or on hand-crafted weather models that cannot capture the complex, spatially varying appearance of real rain and snow at urban-block scale.

2.3 Weather-Specific Editing

More recently, weather-specific editing of 3D Gaussian scenes has attracted increasing attention. WeatherEdit (Qian et al., 2026) augments a 3DGS scene with a dynamic 4D Gaussian field to synthesize controllable snow, rain, and fog particles, and StyledStreets (Chen et al., 2025) disentangles persistent scene geometry from transient style attributes to translate street scenes across seasons and weather with multi-camera consistency. Weather-Magician (Sang et al., 2025) reconstructs real scenes and renders synthesized 4D weather in real time by modeling rain and snow as explicit

Gaussian particles, while the inverse task—removing adverse-weather artifacts to recover a clean scene—is addressed by WeatherGS (Qian et al., 2025). Let it Snow (Fiebelman et al., 2026) instead applies video score distillation sampling (Video-SDS) to inject dynamic rain and snow into pre-built 3DGS scenes. Despite these advances, existing methods either rely on per-view 2D diffusion edits without multi-view consistency, or treat weather as an additive particle layer decoupled from the underlying scene appearance; none re-stylizes the surface appearance of a large-scale 3DGS scene (wet specularity, snow accumulation, overcast global illumination) under a strict geometry-fixed constraint, which is the problem we target.

3 PRELIMINARIES

3.1 3D GAUSSIAN SPLATTING

A 3DGS scene $\mathcal{G} = \{g_i\}_{i=1}^{N}$ represents geometry and appearance using N anisotropic 3D Gaussians. Each Gaussian g_i is parameterized by a center $\boldsymbol{\mu}_i \in \mathbb{R}^3$, a quaternion $\boldsymbol{q}_i \in \mathbb{R}^4$, a scale $\boldsymbol{s}_i \in \mathbb{R}^3$, an opacity $\alpha_i \in [0, 1]$, and a set of spherical harmonics (SH) coefficients $\boldsymbol{c}_i \in \mathbb{R}^k$ encoding view-dependent color. The scene is rendered via alpha-blending of the projected 2D Gaussians: $\boldsymbol{C}(p) = \sum_{i=1}^{N} \alpha_i T_i \boldsymbol{C}_i(p)$, where $T_i = \prod_{j=1}^{i-1}(1 - \alpha_j)$. We explicitly partition the parameters of g_i into a geometry group $\{\boldsymbol{\mu}_i, \boldsymbol{q}_i, \boldsymbol{s}_i, \alpha_i\}$ and an appearance group $\{\boldsymbol{c}_i\}$. This separation is central to our method (Section 4.4): we freeze the geometry group and optimize only the appearance group when editing weather, so that the edited scene $\mathcal{G}_{\text{new}}$ is guaranteed to share the exact same geometry as the original scene $\mathcal{G}_{\text{orig}}$.

3.2 VIDEO DIFFUSION BACKBONE

We use Cosmos-Transfer2.5 (NVIDIA et al., 2025b) as our backbone, which is a controllable video diffusion model that accepts structural control signals. In our case, depth video and edge maps rendered from the original 3DGS scene along a target camera trajectory. Cosmos-Transfer2.5 generates a stylized RGB video that respects the given geometry while altering appearance according to a text prompt P_{weather} (e.g., "heavy snow accumulation on rooftops, overcast sky").

The VAE encoder WAN 2.1(Team Wan et al., 2025) of Cosmos-Transfer2.5 compresses a chunk of $T = 1 + 4n$ RGB frames (for integer $n \geq 1$) into $n + 1$ latent frames along the temporal axis, i.e., it treats the first frame specially (as an anchor) and groups the remaining $4n$ frames into n blocks of 4, each block being compressed into one latent frame. Formally, denoting the VAE encoder as $\mathcal{E}$ and an input clip as $\mathbf{x}_{1:T} \in \mathbb{R}^{T \times H \times W \times 3}$,

$$\mathbf{z}_{0:n} = \mathcal{E}(\mathbf{x}_{1:T}), \qquad \mathbf{z}_{0:n} \in \mathbb{R}^{(n+1) \times h \times w \times c}, \tag{1}$$

where $\mathbf{z}_0$ corresponds to the first input frame $\mathbf{x}_1$ and $\mathbf{z}_j$ $(j = 1, \ldots, n)$ corresponds to the j-th subsequent block of 4 frames $\mathbf{x}_{4j-2:4j+1}$.

4 METHOD

4.1 OVERVIEW

Given a geometry-accurate 3DGS scene $\mathcal{G}_o$ reconstructed under nominal weather and a target weather prompt P_w, our goal is to produce a weather-stylized scene $\mathcal{G}_e$ that is photorealistic, multi-view consistent, and geometry-preserving. Our method consists of data curation, model fine-tuning, and trajectory-aware inference. We first construct a paired weather-editing dataset from rendered RGB, depth, and edge videos with weather-stylized RGB targets generated by the base Cosmos-Transfer2.5 model (Section 4.2). We then fine-tune Cosmos-Transfer2.5 with our tail-latent conditioning mechanism to support both first-frame and first-last-frame conditioning modes (Section 4.3). At inference, scene-specific trajectory strategies sequentially generate interconnected video chunks by reusing previously generated views as anchors, and the resulting stylized views are used to optimize only the appearance coefficients of $\mathcal{G}_o$ while keeping its geometry fixed, yielding $\mathcal{G}_e$ (Section 4.4).

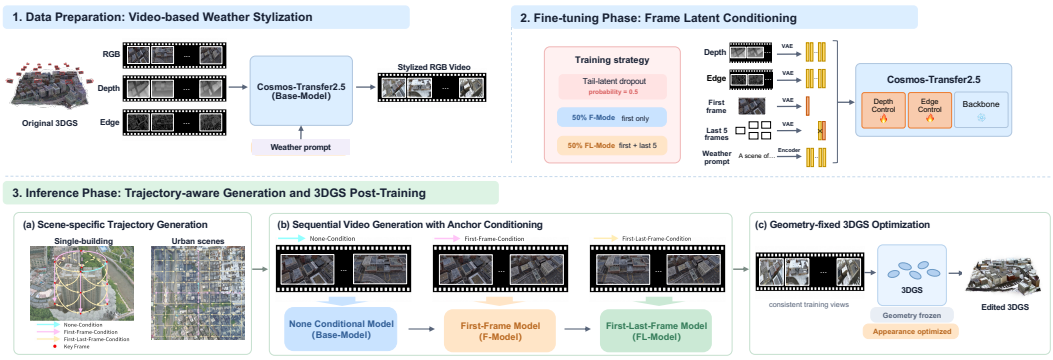

Figure 2: **Architecture Overview.** Given a pre-trained 3DGS, 1) we first construct a weather-stylization dataset via Cosmos-Transfer2.5, manually filtering out low-quality results and dynamic objects. 2) We retain the corresponding prompt, depth, edge, and stylized RGB video as training data, and introduce an additional tail-latent conditioning mechanism. 3) We design separate trajectory schemes for standalone buildings and large-scale urban scenes. Leveraging our fine-tuned model, inference is performed through scene-specific trajectory. Stylized 3DGS is optimized with these consistent views.

4.2 Data Preparation

We construct a weather-stylization dataset for fine-tuning Cosmos-Transfer2.5 from large-scale urban 3DGS scenes reconstructed from drone captures. For each scene, we render RGB and depth videos along diverse camera trajectories, including orbital, straight, and polyline paths with varying angular spans and motion ranges, and extract edge videos using Canny detection. We prepare 10 rain and 10 snow prompts covering diverse weather and lighting conditions and expand them with a large language model. The depth and edge videos together with the weather prompt are fed into the base Cosmos-Transfer2.5 model to generate weather-stylized RGB videos. We manually filter out low-quality results and those containing dynamic objects, and retain the corresponding prompt, depth, edge, and stylized RGB video as training data.

4.3 Fine-Tuning with First- and First-Last-Frame Conditioning

Starting from the base Cosmos-Transfer2.5, we introduce a tail-latent conditioning mechanism that enables first-last-frame conditioning in addition to first-frame conditioning. Following the $1{+}4n$ temporal structure of the VAE, the first-frame latent $\mathbf{z}_{\text{first}}$ is obtained by encoding the first reference frame, while the tail latent $\mathbf{z}_{\text{end}}$ is extracted from the end of a reference clip. Specifically, we encode its final $5 = 1{+}4{\cdot}1$ frames, discard the first temporal slice, and retain the second slice corresponding to the trailing four frames as $\mathbf{z}_{\text{end}}$. The remaining latent positions are initialized as noise.

During fine-tuning, the pretrained Cosmos-Transfer2.5 backbone is kept fixed and only the depth and edge control branches are updated. We randomly drop $\mathbf{z}_{\text{end}}$ with probability 0.5, allowing the same model to operate with either first-frame-only or first-last-frame conditioning at inference. The latter provides an additional endpoint anchor that helps suppress drift when chaining multiple video chunks.

4.4 Trajectory-Based 3DGS Post-Training

Given our fine-tuned model, inference is performed through scene-specific trajectory generation with progressive anchor reuse. Depending on the available anchors, the Base-Model initializes generation, the F-Model propagates appearance from a known starting view, and the FL-Model generates trajectories constrained by both endpoints. Based on this principle, we design separate trajectory schemes for standalone buildings and large-scale urban scenes, as illustrated in 3.(a) of Fig. 2.

Trajectory Design. For a standalone building, we use an intersecting mesh trajectory that first generates multi-view-consistent surrounding views along a closed loop and then progressively ex-

Table 1: Quantitative comparison against 3D editing and video-to-3D baselines on rain and snow weather editing. Best results in **bold**.

	Multi-building				Single-building			
Method	PSNR ↑	SSIM ↑	CLIP-S ↑	CLIP-DS ↑	PSNR ↑	SSIM ↑	CLIP-S ↑	CLIP-DS ↑
Instruct-GS2GS	17.82	0.453	0.704	0.031	16.48	0.660	0.655	0.090
WeatherCity	18.26	0.573	0.801	0.171	18.44	0.578	**0.735**	**0.138**
transfer2.5 + 3DGS	21.07	0.663	0.760	0.129	22.64	0.723	0.700	0.098
cosmos3 + 3DGS	19.81	0.631	0.766	0.091	18.15	0.622	0.677	0.058
Ours	**22.47**	**0.748**	**0.814**	**0.175**	**23.57**	**0.749**	0.715	0.116

tends downward to cover the facades. The Base-Model, F-Model, and FL-Model are combined along the loop to maintain consistency. We sample anchor views at approximately quarter positions, use the F-Model to extend them to lower elevations, and finally apply the FL-Model to connect anchors across different azimuths.

For large-scale urban scenes, we adopt a center-out grid strategy. We first generate four axis-aligned trajectory sequences from the scene center, using the Base-Model for initialization and the F-Model to continue from available starting anchors. We then generate L-shaped trajectories sequentially within each quadrant using the FL-Model, with endpoints anchored by the axis trajectories or previously generated L-shaped trajectories. This progressive anchor reuse expands the stylized region outward while limiting drift accumulation.

3DGS Post-Training Given the generated stylized views $\{V_k\}$, we optimize the original 3DGS scene $\mathcal{G}_{\text{o}}$ to obtain the weather-edited scene $\mathcal{G}_{\text{e}}$. We freeze all geometric parameters $\{\boldsymbol{\mu}_i, \boldsymbol{q}_i, \boldsymbol{s}_i, \alpha_i\}$ and optimize only the color parameters $\{\boldsymbol{c}_i\}$ using a photometric loss between rendered and generated views. Since the geometry is fixed throughout optimization, $\mathcal{G}_{\text{e}}$ preserves the geometry and depth of $\mathcal{G}_{\text{o}}$ while modifying only its appearance.

5 Experiments

5.1 Experimental Setup

Implementation details. Our models are built on Cosmos-Transfer2.5 (NVIDIA et al., 2025b) and trained on the aerial weather dataset described in Section 4.3. During fine-tuning, we keep the pretrained Cosmos-Transfer2.5 generative backbone frozen and optimize only the depth and edge control branches using Adam with a learning rate of 10^{-5} for 5 epochs and a batch size of 8. For spatial augmentation, each training video is randomly cropped to either a 16:9 or 1:1 aspect ratio with equal probability. We further randomly reverse the temporal order of each video and apply one of no flip, vertical flip, or horizontal flip with equal probability. Independently of these augmentations, each training iteration samples the conditioning mode with equal probability between first-frame-only and first-last-frame conditioning. For the geometry-fixed 3DGS optimization stage, we optimize only the appearance coefficients $\boldsymbol{c}_i$ using Adam with a learning rate of 0.0025 for 10,000 iterations while keeping all geometric attributes fixed. All experiments are conducted on 8 NVIDIA A100 80GB GPUs.

Baselines. We compare against two categories of baselines. *3D editing baselines*: Instruct-GS2GS (Vachha & Haque, 2024) and WeatherCity. *Video-to-3D baselines*: Cosmos-Transfer2.5 + 3DGS and Cosmos3 + 3DGS, directly stylizing the aerial control video with the official, non-fine-tuned model and using the resulting frames to optimize the 3DGS scene.

Evaluation metrics. We evaluate semantic alignment using CLIP-S (Radford et al., 2021) and CLIP-DS (Gal et al., 2022), and multi-view consistency using PSNR and SSIM. For the latter, we hold out a subset of views from 3DGS optimization and compare their independently generated stylized images with renderings of the optimized stylized 3DGS at the same viewpoints. Higher PSNR and SSIM indicate stronger multi-view consistency.

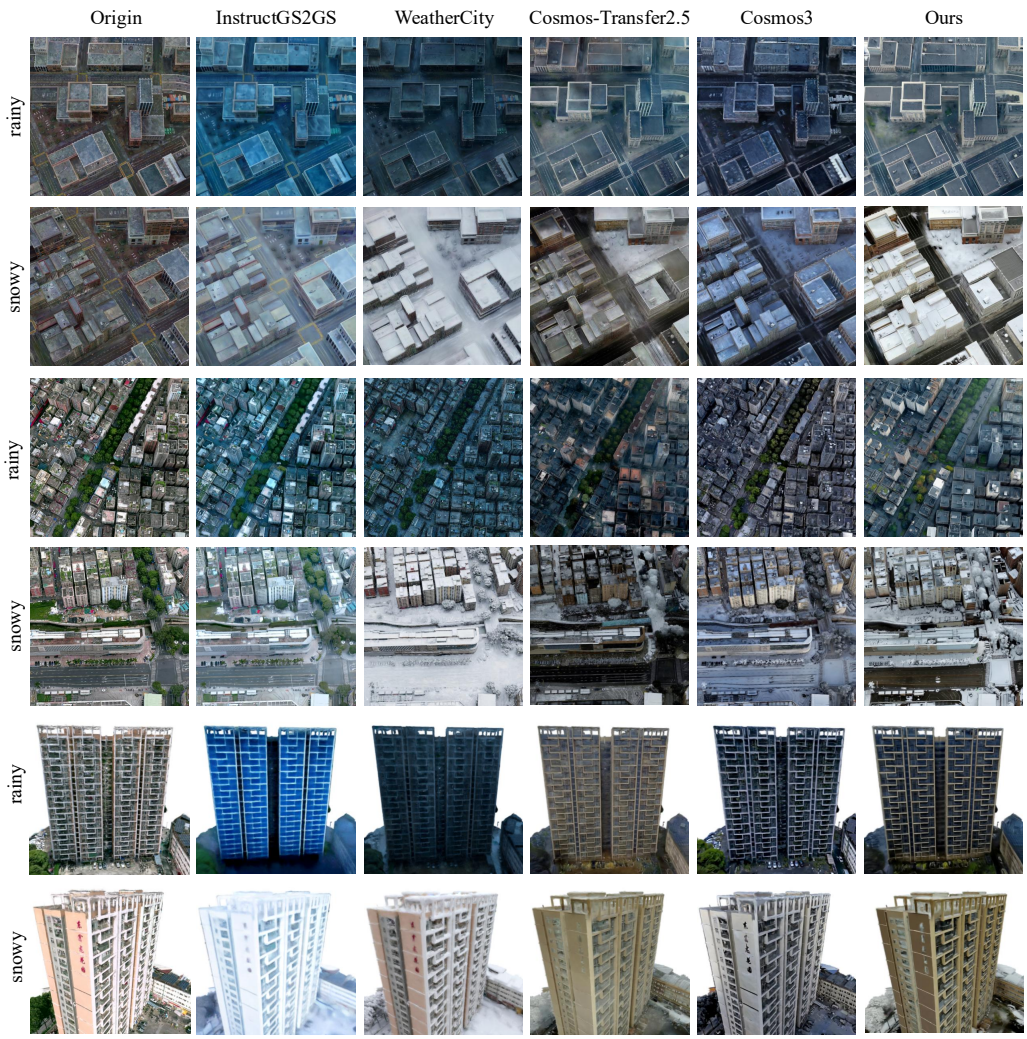

Figure 3: **Qualitative 3DGS rendering results of weather stylization on our dataset.** Our method can not only produce realistic weather editing effects, but also maintain the multi-view consistency. By contrast, other baselines exhibit either scene distortion or blurry Gaussians.

5.2 Experimental Results

Long-term controllable video generation. To evaluate the performance of long-range inference, we generate videos along straight-line camera trajectories using an ABB scheme, where A denotes the Cosmos-Transfer2.5 (Base-Model) and B our fine-tuned first-frame model (F-Model): the sequence is bootstrapped once with the Base-Model, after which the F-Model is applied iteratively, each step feeding the previously generated frame back as the new first-frame condition so that the entire trajectory is produced in a single contiguous pass. We compare this against a segment-wise baseline that splits the same trajectory into short clips generated independently and then concatenated. The visualization results are shown as Fig.4. The long-range ABB inference exhibits substantially smaller exposure bias and does not produce the filter effect, which is the frame-level drift that degrades the video into an artifact-laden, filtered appearance. It is confirmed that our fine-tuned first-frame conditioning model preserves stronger temporal consistency than the segment-wise alternative.

To further support loop-closure inference on circular trajectories, we adopt an ABC scheme, where C denotes our fine-tuned first-last-frame model (FL-Model). The loop is bootstrapped with the Base

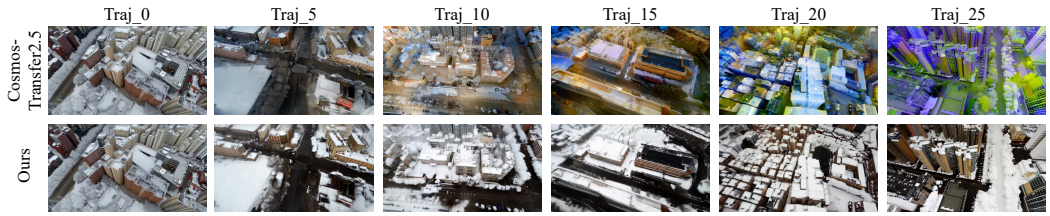

Figure 4: **Comparison on long-horizon inference.** Both our model and the Base Model generate long video sequences in a chunk-wise autoregressive manner, where the last frame of each preceding chunk is used as the conditioning frame for the subsequent chunk. On sequences of up to 2.4K frames, our model substantially mitigates long-horizon drift while maintaining a more consistent stylization across the generated sequence.

Model and extended frame-by-frame with F-Model, and at the final step the FL-Model is applied while conditioning jointly on the start and end frames, so that the generated ring closes back onto its beginning. Because FL-Model explicitly models the correspondence between the first and last frames, this loop-closure inference achieves a smooth, seamless end-to-start transition for circular data, eliminating the visible seam or jump that would otherwise break the continuity of the ring.

Generation consistency. As shown in Fig. 3 and Tab. 1, our method consistently produces cleaner weather-edited 3DGS representations with fewer blur, ghosting, and view-dependent artifacts than the compared methods. This observation is also supported quantitatively by PSNR and SSIM on held-out viewpoints, where our method achieves substantially higher reconstruction scores than all baselines. Since the held-out stylized views are not used during 3DGS optimization, these metrics reflect how well the independently generated views can be explained by a single optimized 3D representation, providing a direct measure of multi-view consistency.

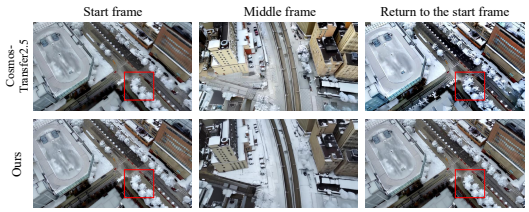

Figure 5: **The comparison results of our ABC scheme and raw Cosmos-Transfer2.5 model.** After loop closure and returning to the starting viewpoint, our method achieves better consistency.

The comparison with the original Cosmos-Transfer2.5 further highlights the effect of long-range generation. Around the initial trajectories, the two methods produce similar weather appearance because generation in this region mainly relies on the Base-Model and first-frame conditioning. As generation progressively expands outward, however, the original model exhibits increasing appearance drift and gradual degradation of the target weather style. In contrast, our first-last-frame conditioning provides additional endpoint constraints for later trajectories and maintains more stable stylization across long chains of interconnected video chunks.

WeatherCity achieves higher CLIP-S or CLIP-DS in some settings, indicating strong per-view semantic alignment with the target weather prompt. However, these metrics do not directly reflect cross-view consistency. Its lower PSNR and SSIM, together with the qualitative results in Fig. 3, show that the independently edited views are less consistent with a single 3D representation, resulting in noisier and less coherent 3DGS appearance under close-range inspection. Instruct-GS2GS shows weaker weather stylization and multi-view consistency in our experiments, while Cosmos3 produces less convincing outdoor weather appearance, which is mainly attributable to differences in the underlying generation backbone.

5.3 Ablation Studies

Ablation 1: selection of checkpoints. We select the optimal checkpoint by evaluating the FVD of models trained with different iterations, as illustrated in Fig.6. The top-left subplot shows the orig-

inal input image. At low iterations, the model fails to fully learn the weather stylization capability, resulting in severe color distortion and inconsistent weather appearance across frames. As training proceeds, the model gradually captures the mapping from geometric conditions to weathered 3DGS outputs. In later training stages, the model becomes stable and FVD no longer decreases significantly, which indicates that additional training iterations bring marginal gains to the generation quality. We adopt the checkpoint at 96000 iterations as our final model weights for all subsequent evaluations.

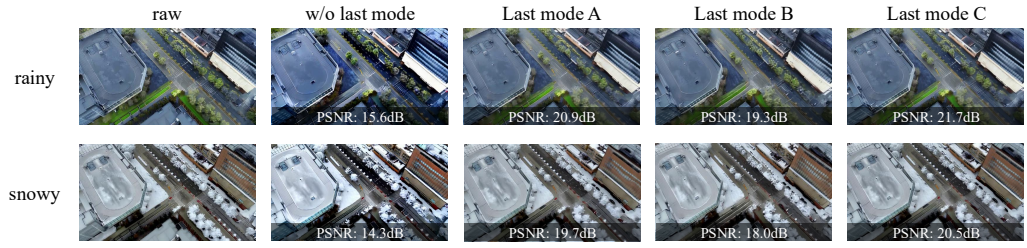

Figure 7: **Ablation study on the last-frame encoding strategy.** We investigate different strategies for encoding the last-frame condition and use the FL-Model to generate a complete chunk under each strategy. We compare these strategies by evaluating the quality of the generated frame immediately preceding the conditioning frame.

Ablation 2: last-frame encoding strategy. We ablate different encoding strategies for the end-frame condition. Last mode A encodes a single end frame, Last mode B averages the latents of the last four frames, and Last mode C, our proposed strategy, jointly encodes the last five frames and retains the second temporal slice corresponding to the trailing four frames. We measure cross-chunk consistency by PSNR on the overlapping view shared by the conditioning window and the current chunk. Since only the trailing-four-frame latent slice is injected, the remaining leading frame provides a same-pose reference from the preceding chunk for direct PSNR evaluation. As shown in Fig.7, our strategy achieves the highest consistency.

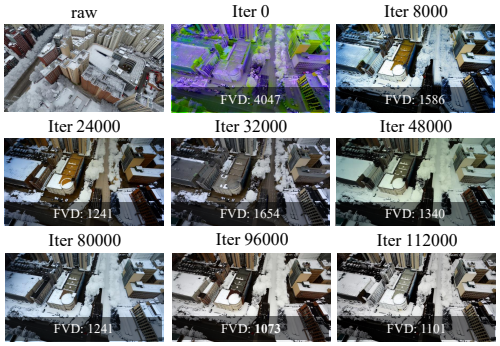

Figure 6: **Ablation study on the number of training iterations.** We evaluate the model quality at different training stages using the FVD and select the model with the best performance.

6 Conclusion and Limitations

We present G-WADE, a geometry-fixed diffusion framework for weather stylization of urban 3DGS assets. We fine-tune two Cosmos-Transfer2.5 variants, F-Model and FL-Model, to retain geometric fidelity during rainy and snowy appearance synthesis. Trajectory-aware generation pipelines are tailored for single-building and large urban scenes. Multi-view consistent outputs optimize the original 3DGS. Experiments verify that G-WADE reduces geometry collapse, texture inconsistency and long-range drift, outperforming baselines for multi-view consistency, realism and geometry preservation for adverse-weather urban asset augmentation. Our framework has limitations for future exploration. We only validate rain and snow editing; extending to fog, haze or sandstorm with volumetric scattering needs extra volumetric rendering beyond the SH-based appearance model. Besides, our fine-tuning dataset covers only urban aerial scenes, and the generalization of models to natural environments like forests and mountains remains untested.

REFERENCES

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets. *arXiv preprint arXiv:2311.15127*, 2023. URL https://arxiv.org/abs/2311.15127.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. InstructPix2Pix: Learning to Follow Image Editing Instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18392–18402, 2023. doi: 10.1109/CVPR52729.2023.01764. URL https://openaccess.thecvf.com/content/CVPR2023/html/Brooks_InstructPix2Pix_Learning_To_Follow_Image_Editing_Instructions_CVPR_2023_paper.html.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video Generation Models as World Simulators. OpenAI Technical Report, 2024. URL https://openai.com/index/video-generation-models-as-world-simulators/.

Minghao Chen, Iro Laina, and Andrea Vedaldi. DGE: Direct Gaussian 3D Editing by Consistent Multi-View Editing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 74–92, 2024a. doi: 10.1007/978-3-031-72904-1_5. URL https://arxiv.org/abs/2404.18929.

Yiwen Chen, Zilong Chen, Chi Zhang, Feng Wang, Xiaofeng Yang, Yikai Wang, Zhongang Cai, Lei Yang, Huaping Liu, and Guosheng Lin. GaussianEditor: Swift and Controllable 3D Editing with Gaussian Splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21476–21485, 2024b. doi: 10.1109/CVPR52733.2024.02029. URL https://openaccess.thecvf.com/content/CVPR2024/html/Chen_GaussianEditor_Swift_and_Controllable_3D_Editing_with_Gaussian_Splatting_CVPR_2024_paper.html.

Yuyin Chen, Yida Wang, Xueyang Zhang, Kun Zhan, Peng Jia, Yifei Zhan, and Xianpeng Lang. StyledStreets: Multi-Style Street Simulator with Spatial and Temporal Consistency. *arXiv preprint arXiv:2503.21104*, 2025. URL https://arxiv.org/abs/2503.21104.

Xinhua Cheng, Tianyu Yang, Jianan Wang, Yu Li, Lei Zhang, Jian Zhang, and Li Yuan. Progressive3D: Progressively Local Editing for Text-to-3D Content Creation with Complex Semantic Prompts. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=O072Rc8uUy.

Gal Fiebelman, Hadar Averbuch-Elor, and Sagie Benaim. Let it Snow! Animating 3D Gaussian Scenes with Dynamic Weather Effects via Physics-Guided Score Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 37322–37331, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Fiebelman_Let_it_Snow_Animating_3D_Gaussian_Scenes_with_Dynamic_Weather_CVPR_2026_paper.html.

Rinon Gal, Or Patashnik, Haggai Maron, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. StyleGAN-NADA: CLIP-Guided Domain Adaptation of Image Generators. *ACM Transactions on Graphics*, 41(4), 2022. doi: 10.1145/3528223.3530164. URL https://doi.org/10.1145/3528223.3530164.

Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. TokenFlow: Consistent Diffusion Features for Consistent Video Editing. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/069274129275ec1c6511abb5f6d96a31-Abstract-Conference.html.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/94894cf9bab20c00f6eaaff3b7f2be70-Abstract-Conference.html.

Ayaan Haque, Matthew Tancik, Alexei A. Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-NeRF2NeRF: Editing 3D Scenes with Instructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19740–19750, 2023. doi: 10.1109/ICCV51070.2023.01808. URL https://openaccess.thecvf.com/content/ICCV2023/html/Haque_Instruct-NeRF2NeRF_Editing_3D_Scenes_with_Instructions_ICCV_2023_paper.html.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics*, 42(4):1–14, 2023. doi: 10.1145/3592433. URL https://doi.org/10.1145/3592433.

Dong In Lee, Hyeongcheol Park, Jiyoung Seo, Eunbyung Park, Hyunje Park, Ha Dam Baek, Sangheon Shin, Sangmin Kim, and Sangpil Kim. EditSplat: Multi-View Fusion and Attention-Guided Optimization for View-Consistent 3D Scene Editing with 3D Gaussian Splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11135–11145, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Lee_EditSplat_Multi-View_Fusion_and_Attention-Guided_Optimization_for_View-Consistent_3D_Scene_CVPR_2025_paper.html.

Yuhan Li, Yishun Dou, Yue Shi, Yu Lei, Xuanhong Chen, Yi Zhang, Peng Zhou, and Bingbing Ni. FocalDreamer: Text-Driven 3D Editing via Focal-Fusion Assembly. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(4):3279–3287, 2024. doi: 10.1609/aaai.v38i4.28113. URL https://doi.org/10.1609/aaai.v38i4.28113.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 405–421, 2020. doi: 10.1007/978-3-030-58452-8_24. URL https://www.ecva.net/papers/eccv_2020/papers_ECCV/html/1473_ECCV_2020_paper.php.

Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16784–16804. PMLR, 2022. URL https://proceedings.mlr.press/v162/nichol22a.html.

NVIDIA, Hassan Abu Alhaija, Jose Alvarez, Maciej Bala, Tiffany Cai, Tianshi Cao, Liz Cha, Joshua Chen, Mike Chen, Francesco Ferroni, Sanja Fidler, Dieter Fox, Yunhao Ge, Jinwei Gu, Ali Hassani, Michael Isaev, Pooya Jannaty, Shiyi Lan, Tobias Lasser, Huan Ling, Ming-Yu Liu, Xian Liu, Yifan Lu, Alice Luo, Qianli Ma, Hanzi Mao, Fabio Ramos, Xuanchi Ren, Tianchang Shen, Shitao Tang, Ting-Chun Wang, Jay Wu, Jiashu Xu, Stella Xu, Kevin Xie, Yuchong Ye, Xiaodong Yang, Xiaohui Zeng, and Yu Zeng. Cosmos-Transfer1: Conditional World Generation with Adaptive Multimodal Control. *arXiv preprint arXiv:2503.14492*, 2025a. URL https://arxiv.org/abs/2503.14492.

NVIDIA, Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, Yu-Wei Chao, Prithvijit Chattopadhyay, Mike Chen, Yongxin Chen, Yu Chen, Shuai Cheng, Yin Cui, Jenna Diamond, Yifan Ding, Jiaojiao Fan, Linxi Fan, Liang Feng, Francesco Ferroni, Sanja Fidler, Xiao Fu, Ruiyuan Gao, Yunhao Ge, Jinwei Gu, Aryaman Gupta, Siddharth Gururani, Imad El Hanafi, Ali Hassani, Zekun Hao, Jacob Huffman, Joel Jang, Pooya Jannaty, Jan Kautz, Grace Lam, Xuan Li, Zhaoshuo Li, Maosheng Liao, Chen-Hsuan Lin, Tsung-Yi Lin, Yen-Chen Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Yifan Lu, Alice Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Seungjun Nah, Yashraj Narang, Abhijeet Panaskar, Lindsey Pavao, Trung Pham, Morteza Ramezanali, Fitsum Reda, Scott Reed, Xuanchi Ren, Haonan Shao, Yue Shen, Stella Shi, Shuran Song, Bartosz Stefaniak, Shangkun Sun, Shitao Tang, Sameena Tasmeen, Lyne Tchapmi, Wei-Cheng Tseng, Jibin Varghese, Andrew Z. Wang, Hao Wang, Haoxiang Wang, Heng Wang, Ting-Chun Wang, Fangyin Wei, Jiashu Xu, Dinghao Yang, Xiaodong Yang, Haotian Ye, Seonghyeon Ye, Xiaohui Zeng, Jing Zhang, Qinsheng Zhang, Kaiwen Zheng, Andrew Zhu, and Yuke Zhu. World Simulation with Video Foundation Models for Physical AI. *arXiv preprint arXiv:2511.00062*, 2025b. URL https://arxiv.org/abs/2511.00062.

Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. FateZero: Fusing Attentions for Zero-Shot Text-Based Video Editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15932–15942, 2023. doi: 10.1109/ICCV51070.2023.01460. URL https://openaccess.thecvf.com/content/ICCV2023/html/QI_FateZero_Fusing_Attentions_for_Zero-shot_Text-based_Video_Editing_ICCV_2023_paper.html.

Chenghao Qian, Yuhu Guo, Wenjing Li, and Gustav Markkula. WeatherGS: 3D Scene Reconstruction in Adverse Weather Conditions via Gaussian Splatting. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 185–191, 2025. doi: 10.1109/ICRA55743.2025.11128699. URL https://doi.org/10.1109/ICRA55743.2025.11128699.

Chenghao Qian, Wenjing Li, Yuhu Guo, and Gustav Markkula. WeatherEdit: Controllable Weather Editing with 4D Gaussian Field. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(10):8511–8519, 2026. doi: 10.1609/aaai.v40i10.37802. URL https://doi.org/10.1609/aaai.v40i10.37802.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Chen Sang, Yeqiang Qian, Jiale Zhang, Chunxiang Wang, and Ming Yang. Weather-Magician: Reconstruction and Rendering Framework for 4D Weather Synthesis in Real Time. *arXiv preprint arXiv:2505.19919*, 2025. URL https://arxiv.org/abs/2505.19919.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and Advanced Large-Scale Video Generative Models. *arXiv preprint arXiv:2503.20314*, 2025. URL https://arxiv.org/abs/2503.20314.

Cyrus Vachha and Ayaan Haque. Instruct-GS2GS: Editing 3D Gaussian Splats with Instructions. Project page, 2024. URL https://instruct-gs2gs.github.io/.

Wen Wang, Yan Jiang, Kangyang Xie, Zide Liu, Hao Chen, Yue Cao, Xinlong Wang, and Chunhua Shen. Zero-Shot Video Editing Using Off-The-Shelf Image Diffusion Models. *arXiv preprint arXiv:2303.17599*, 2023. URL https://arxiv.org/abs/2303.17599.

Jing Wu, Jia-Wang Bian, Xinghui Li, Guangrun Wang, Ian Reid, Philip H. S. Torr, and Victor Adrian Prisacariu. GaussCtrl: Multi-View Consistent Text-Driven 3D Gaussian Splatting Editing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 55–71, 2024. doi: 10.1007/978-3-031-72630-9_4. URL https://gaussctrl.active.vision/.

Wenhua Wu, Huai Guan, Zhe Liu, and Hesheng Wang. WeatherCity: Urban Scene Reconstruction with Controllable Multi-Weather Transformation, 2026. URL https://arxiv.org/abs/2602.22096.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3836–3847, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Zhang_Adding_Conditional_Control_to_Text-to-Image_Diffusion_Models_ICCV_2023_paper.html.

Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. ControlVideo: Training-Free Controllable Text-to-Video Generation. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/ef74413c7bf1d915c3e45c72e19a5d32-Abstract-Conference.html.

Jingyu Zhuang, Chen Wang, Liang Lin, Lingjie Liu, and Guanbin Li. DreamEditor: Text-Driven 3D Scene Editing with Neural Fields. In *SIGGRAPH Asia 2023 Conference Papers*, 2023. doi: 10.1145/3610548.3618190. URL https://doi.org/10.1145/3610548.3618190.