METEOVERSE: UNIFIED WEATHER-CONTROLLABLE VIDEO WORLD MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Video world models aim to predict future content from an observed scene while following prescribed camera motion. Real-world scene evolution is determined not only by changes in viewpoint and object dynamics, but also by environmental conditions such as weather, which can substantially alter scene appearance and visibility. Modeling such realistic weather evolution is challenging because the required weather modification depends jointly on the observed and desired weather states. Depending on their relation, the model may need to preserve, introduce, or remove a weather effect. Existing video world models typically leave this weather transition implicit, forcing the generation backbone to infer weather evolution together with scene dynamics and camera motion, which leads to imprecise weather control. To address this limitation, we propose MeteoVerse, a unified weather-controllable video world model that generates future videos from a single sunny or adverse-weather image, conditioned on a weather-free scene description, a target-weather instruction, and a camera trajectory. Rather than conditioning only on the desired weather, MeteoVerse explicitly estimates the observed and target weather states and represents the required weather transition. A transition-aware mixture of weather experts (MeteoMoE) then translates this transition into category-specific residual weather features, unifying weather preservation, introduction, and removal while enabling fine-grained control over introduced weather intensity. We further construct the MeteoVerse dataset with over 50K real-world weather video clips, generated sunny counterparts, disentangled scene and weather descriptions, weather-intensity annotations, and camera trajectories. Extensive quantitative and qualitative experiments demonstrate substantially improved weather controllability while retaining competitive scene consistency and camera-control performance.

1 INTRODUCTION

Video world models aim to predict future content from an observed scene while following prescribed camera motion. While recent advances have enabled increasingly flexible control over scene content and viewpoint, real-world scene evolution is also shaped by environmental conditions. Weather is a particularly important factor, as changes in rain, snow, and fog can substantially alter scene appearance and visibility over time. A world model simulating realistic future observations should therefore model not only what is observed and from where, but also under which weather condition the future unfolds. Accordingly, we study weather-controllable video world modeling, *i.e.*, given a single image captured under sunny or adverse weather, a weather-free scene description, a target-weather instruction, and a camera trajectory, the model generates a future video that preserves scene identity, follows the prescribed camera motion, and realizes the desired weather evolution through weather preservation, introduction, or removal.

Existing approaches address weather manipulation under substantially different assumptions. Reconstruction-based methods can synthesize geometrically consistent weather effects through explicit 3D scene representations or physical simulation, but require multi-view observations and costly per-scene reconstruction (Li et al., 2022; Qian et al., 2025). Video weather editing methods can introduce or remove weather effects while preserving the structure and motion of an existing video (Lin et al., 2025; Qian et al., 2026), but assume that the complete source video is already available. More recently, Holo-World brings weather control into single-image video world modeling by

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.

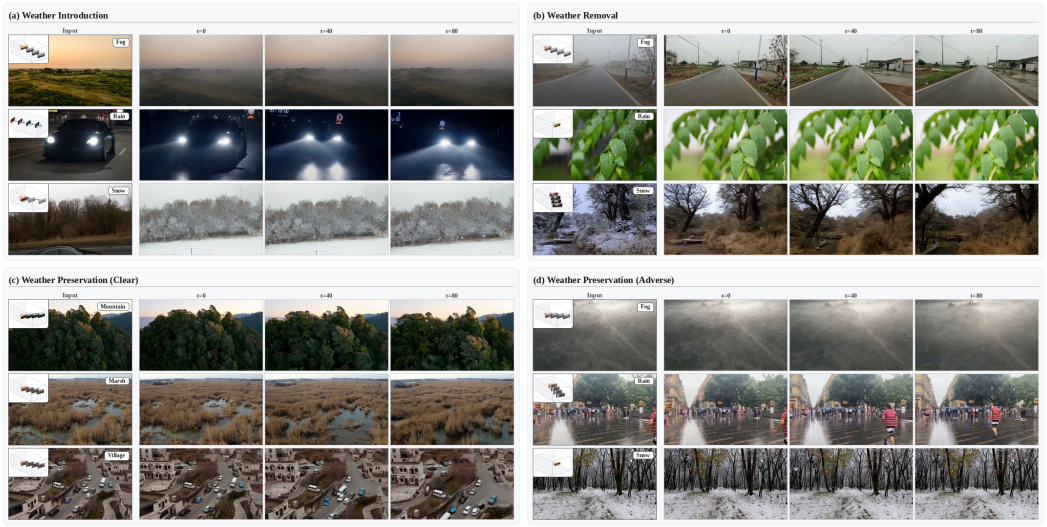

Figure 1: Overview of MeteoVerse, which supports weather preservation, introduction, and removal from sunny or adverse-weather inputs under prescribed camera trajectories.

jointly controlling camera motion and adverse-weather generation (Yin et al., 2026). However, it assumes a sunny input and mainly focuses on weather introduction, leaving weather removal unexplored. These limitations motivate a unified formulation that can operate from either sunny or adverse-weather observations and explicitly control how weather evolves into the future.

The key challenge is that a target-weather instruction specifies the desired weather condition, but not the transformation required to reach it. The same target condition may correspond to fundamentally different operations depending on the weather already present in the input. For example, a rainy target requires introducing rain from a sunny observation or preserving rain from an already rainy observation, while a sunny target requires removing adverse weather from an adverse-weather observation. Therefore, effective weather control requires explicitly reasoning about the transition between the observed and target weather states rather than relying on the target condition alone. When this transition is left implicit, the video backbone must simultaneously infer the observed weather, interpret the target instruction, determine the required modification, and model scene evolution and camera motion. This entanglement makes precise weather control particularly difficult.

To address this challenge, we propose MeteoVerse, a unified weather-controllable video world model that explicitly models the transition between observed and desired weather states. Given the input image and target-weather instruction, a weather-state predictor estimates continuous observed and target states over rain, snow, and fog, from which the required weather transition is derived. Rather than directly conditioning the entire generation backbone on the target weather, we introduce a transition-aware mixture of weather experts (MeteoMoE) to translate the transition into residual weather modifications. Specifically, category-specific rain, snow, and fog experts extract scene-adaptive weather features, whose responses are modulated according to the direction and magnitude of the corresponding transition components and then fused into a transition-aware representation. The resulting representation is injected into the video world-model backbone, while the weather-free scene description and camera trajectory provide semantic and viewpoint conditions independently. This decomposition allows the pretrained backbone to preserve the underlying scene evolution, while MeteoMoE focuses on the weather modification required to reach the target state. As a result, a single model can support weather preservation, introduction, and removal, together with fine-grained control over introduced weather intensity.

Training such a model requires supervision covering diverse weather transitions. We therefore construct the MeteoVerse dataset, containing over 50K real-world rain, snow, and fog video clips together with generated sunny counterparts, disentangled scene and weather descriptions, category-specific weather-intensity annotations, and camera trajectories. Extensive quantitative and qualitative experiments demonstrate that MeteoVerse substantially improves weather controllability over existing controllable video world models, particularly when the desired weather differs from the observed condition, while retaining competitive scene consistency and camera-control performance.

Our contributions are summarized as follows:

- We introduce MeteoVerse, a unified weather-controllable video world model that predicts camera-controlled future videos from either sunny or adverse-weather observations and supports weather preservation, introduction, and removal.
- We propose explicit weather transition modeling together with MeteoMoE, which translates weather-state changes into category-specific residual weather features and enables fine-grained control over introduced weather intensity.
- We construct the MeteoVerse dataset with over 50K real-world weather video clips, generated sunny counterparts, disentangled scene and weather descriptions, weather-intensity annotations, and camera trajectories, providing supervision for diverse weather transitions. Extensive experiments demonstrate substantially improved weather controllability while maintaining competitive scene consistency and camera-control performance.

2 Related Work

2.1 Video World Models

Camera-controllable video generation has evolved from motion control to geometry-aware scene exploration. MotionCtrl, CameraCtrl, and CamI2V introduce explicit camera control into video generation (Wang et al., 2024; He et al., 2024; Zheng et al., 2024), while ViewCrafter, ReCamMaster, TrajectoryCrafter, and Voyager further support camera-guided scene exploration or trajectory manipulation (Yu et al., 2024; Bai et al., 2025a; Yu et al., 2025; Huang et al., 2025a). Video world models extend this paradigm toward interactive future prediction, including GAIA-1, Genie, GameNGen, Matrix-Game, and LingBot-World (Hu et al., 2023; Bruce et al., 2024; Valevski et al., 2024; Zhang et al., 2025; Robbyant Team et al., 2026). Most closely related, Holo-World introduces weather control into single-image video world modeling (Yin et al., 2026), but mainly considers adverse-weather generation from sunny inputs. MeteoVerse instead supports both sunny and adverse-weather observations and unifies weather preservation, introduction, and removal.

2.2 Weather Synthesis Methods

Weather manipulation has been studied in restoration, reconstruction-based synthesis, and generative editing. Multi-weather restoration methods remove rain, snow, and fog from degraded observations (Li et al., 2020; Valanarasu et al., 2022; Yang et al., 2023; 2024), while ClimateNeRF, WeatherEdit, and WeatherCity achieve view-consistent weather synthesis through explicit scene representations (Li et al., 2022; Qian et al., 2025; Wu et al., 2026). Generative approaches such as IntrinsicWeather, WeatherWeaver, and AutoAWG reduce the need for explicit reconstruction (Zhu et al., 2026; Lin et al., 2025; Hu et al., 2026), but image editing does not predict unseen views and video editing assumes a complete source video. MeteoVerse instead predicts future observations from a single image while jointly controlling camera motion and weather evolution. Existing weather datasets mainly target adverse-weather perception, restoration, or editing (Sakaridis et al., 2021; Sun et al., 2022; Zhang et al., 2023; Lin et al., 2025; Zhu et al., 2026; Yin et al., 2026). MeteoVerse provides over 50K real-world weather clips with generated sunny counterparts, disentangled scene and weather descriptions, fine-grained weather-intensity annotations, and camera trajectories to support weather-controllable video world modeling.

3 Methods

3.1 Problem Formulation and Overview

Let I denote the current visual observation, p a general text condition, and $\mathcal{C}$ a prescribed camera trajectory. A controllable video world model with parameters θ predicts future observations $V_{1:T}$ as,

$$p_\theta\left(V_{1:T} \mid I, p, \mathcal{C}\right). \tag{1}$$

Although weather information may be implicitly contained in I or p, this formulation does not explicitly specify how the observed weather should evolve toward the requested condition. The

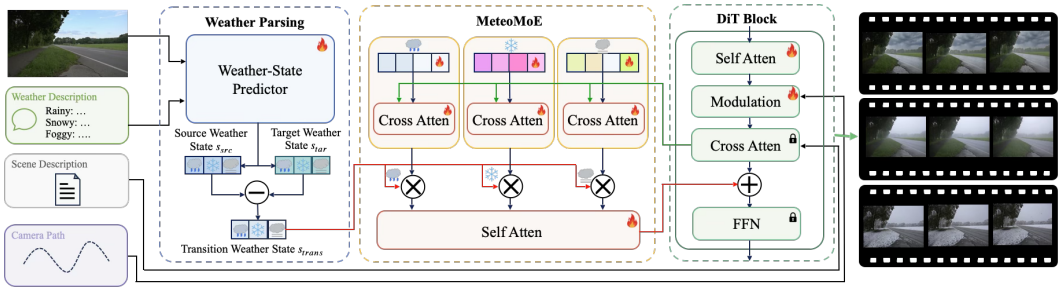

Figure 2: Overview of MeteoVerse. The weather-state predictor estimates the observed and target weather states and derives the required weather transition. MeteoMoE modulates the rain, snow, and fog experts according to the corresponding transition components, fuses their responses through self-attention, and residually injects the resulting transition-aware features into the DiT backbone.

video backbone therefore needs to infer the required weather modification together with scene evolution and camera-induced changes, which can lead to imprecise weather control. As illustrated in Fig. 2, MeteoVerse separates the general text condition into a weather-free scene description p_{s} and a target-weather instruction p_{w}. A weather-state predictor P_ϕ estimates the weather state observed in the input image and the desired target weather state,

$$(\mathbf{s}_{\mathrm{src}}, \mathbf{s}_{\mathrm{tar}}) = P_\phi\left(I, p_{\mathrm{w}}\right), \qquad \mathbf{s}_{\mathrm{trans}} = \mathbf{s}_{\mathrm{tar}} - \mathbf{s}_{\mathrm{src}}. \tag{2}$$

Here, $\mathbf{s}_{\mathrm{src}}$ describes the weather condition observed in the input image, while $\mathbf{s}_{\mathrm{tar}}$ represents the desired future weather condition. Their difference $\mathbf{s}_{\mathrm{trans}}$ characterizes the weather modification required to reach the target state. We further introduce MeteoMoE, a transition-aware mixture of weather experts that converts $\mathbf{s}_{\mathrm{trans}}$ into residual weather conditioning for the pretrained video backbone. The resulting generation process can be written as,

$$p_{\theta,\psi}\left(V_{1:T} \mid I, p_{\mathrm{s}}, \mathcal{C}, \mathbf{s}_{\mathrm{trans}}\right), \tag{3}$$

where θ denotes the frozen pretrained backbone parameters and ψ denotes the trainable adaptation parameters. In this formulation, p_{s} provides scene semantics, $\mathcal{C}$ controls camera motion, and $\mathbf{s}_{\mathrm{trans}}$ specifies the required weather modification.

3.2 Weather-State Transition Modeling

We represent each weather state using a continuous intensity vector over rain, snow, and fog, *i.e.*,

$$\mathbf{s} = [s_{\mathrm{rain}}, s_{\mathrm{snow}}, s_{\mathrm{fog}}] \in [0,1]^3, \tag{4}$$

where each component denotes the intensity of the corresponding weather effect. Sunny weather is represented by the zero state $\mathbf{s} = [0,0,0]$. Accordingly, each component δ_k of $\mathbf{s}_{\mathrm{trans}}$ describes the direction and magnitude of the required modification for weather category k. A positive value indicates that the corresponding weather effect should be introduced. A negative value indicates that the effect should be removed. When $\delta_k = 0$, no additional modification is required and the observed weather condition is preserved. We LoRA-fine-tune Qwen3-VL-2B Bai et al. (2025b) as the weather-state predictor P_ϕ. Its supervision is constructed from the weather annotations described in Sec. 3.4. The LoRA rank and scaling factor α are set to 64 and 128, respectively. During video-model training, the annotated weather states are directly used to compute $\mathbf{s}_{\mathrm{trans}}$. The predictor is frozen and used only at inference time. For weather introduction from sunny inputs, users may specify $\mathbf{s}_{\mathrm{tar}}$ to control the desired weather intensity.

3.3 Transition-Aware MeteoMoE

MeteoMoE translates the explicit weather transition $\mathbf{s}_{\mathrm{trans}}$ into residual weather features for the video backbone. Rain, snow, and fog exhibit distinct visual characteristics, and representing them with a shared weather feature may entangle category-specific effects. We therefore introduce three weather experts. The expert for weather category k is represented by a learnable token set T_k, where $k \in \{\mathrm{rain}, \mathrm{snow}, \mathrm{fog}\}$. In the ℓ-th DiT block, let H^{ℓ}_{txt} denote the feature after text cross-attention.

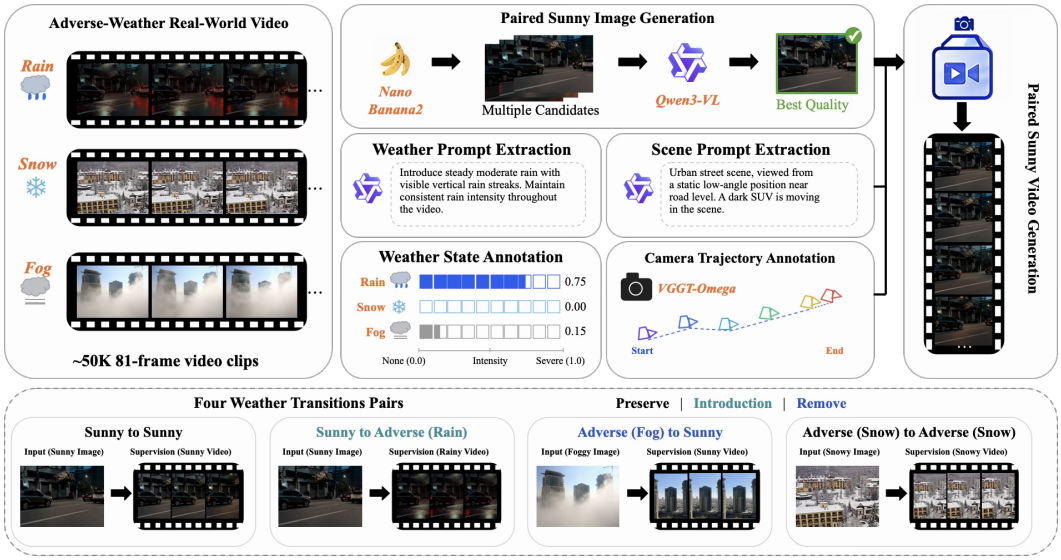

Figure 3: Construction of MeteoVerse dataset. We collect over 50K 81-frame real-world rain, snow, and fog clips with disentangled scene and weather descriptions, weather-intensity annotations, camera trajectories, and pseudo-paired sunny counterparts. The sunny and adverse-weather observations are then recombined to provide supervision for weather preservation, introduction, and removal.

MeteoMoE uses this feature as the query to retrieve category-specific weather information from the expert tokens, *i.e.*,

$$E_k^\ell = \mathrm{CA}_k^\ell \left(\mathrm{Norm}\left(H_{\mathrm{txt}}^\ell\right), T_k, T_k \right), \qquad k \in \{\mathrm{rain}, \mathrm{snow}, \mathrm{fog}\}. \tag{5}$$

Here, H_{txt}^ℓ serves as the query, while T_k provides the keys and values. This produces a scene-adaptive representation for each weather category. Each expert response is then modulated by the corresponding transition component δ_k, which are fused through a lightweight self-attention module. It can be written as,

$$E_{\mathrm{trans}}^\ell = \mathrm{SA}_{\mathrm{w}}^\ell \left(\left\{ \delta_k E_k^\ell \right\}_{k \in \{\mathrm{rain}, \mathrm{snow}, \mathrm{fog}\}} \right). \tag{6}$$

Finally, E_{trans}^ℓ is residually injected into the DiT block as,

$$\widetilde{H}_{\mathrm{txt}}^\ell = H_{\mathrm{txt}}^\ell + E_{\mathrm{trans}}^\ell. \tag{7}$$

In this way, MeteoMoE focuses on the weather modification specified by $\mathbf{s}_{\mathrm{trans}}$, while the pretrained video backbone retains scene evolution and camera-conditioned future prediction.

3.4 METEOVERSE DATASET CONSTRUCTION

The construction pipeline is illustrated in Fig. 3. We collect real-world videos with visible rain, snow, or fog and divide them into 81-frame clips, resulting in over 50K adverse-weather clips. For each clip, Qwen3-VL-8B Bai et al. (2025b) generates a weather-free scene description and a separate weather description. VGGT-Omega Wang et al. (2026) estimates the camera intrinsics and extrinsics, and clips with severe camera jitter or abrupt pose changes are discarded. To obtain fine-grained weather states, we adopt a coarse-to-fine intensity annotation strategy. For each weather category, Qwen3-VL-30B Bai et al. (2025b) first predicts an ordinal intensity level according to visual severity and then estimates a continuous intensity value within the selected level. Each clip is evaluated with a confidence estimate, and samples with low confidence are removed. The remaining annotations are further verified by GPT-5.6 Sol OpenAI (2026).

Pseudo-paired sunny counterpart generation. Since paired sunny and adverse-weather videos of the same real-world scene are difficult to obtain, we construct a pseudo-paired sunny counterpart for each adverse-weather clip. Given the first adverse-weather frame I_{adv}, Nano Banana 2 Google DeepMind (2026) generates multiple sunny candidates. Qwen3-VL-30B Bai et al. (2025b) selects

the candidate that best removes the adverse-weather effect while preserving scene content. Samples without a qualified candidate are discarded. The selected sunny image I_{sun}, together with the weather-free scene description and recovered camera trajectory, is passed to the pretrained LingBot-World Robbyant Team et al. (2026) to generate the pseudo-paired sunny video V_{sun}. We discard generated videos with poor visual quality or obvious temporal artifacts, and re-estimate the camera trajectory of the remaining videos using VGGT-Omega Wang et al. (2026) to obtain more accurate annotations of the camera motion.

Training pair construction. Let $(I_{\mathrm{adv}}, V_{\mathrm{adv}})$ denote the original adverse-weather pair and $(I_{\mathrm{sun}}, V_{\mathrm{sun}})$ denote its pseudo-paired sunny counterpart. We construct four input-target pairs,

$$\mathcal{P} = \{(I_{\mathrm{sun}}, V_{\mathrm{sun}}), (I_{\mathrm{adv}}, V_{\mathrm{adv}}), (I_{\mathrm{sun}}, V_{\mathrm{adv}}), (I_{\mathrm{adv}}, V_{\mathrm{sun}})\}. \tag{8}$$

The first two pairs provide supervision for sunny and adverse-weather preservation. The third pair provides supervision for weather introduction, while the fourth provides supervision for weather removal. Sunny observations are assigned the zero weather state, while adverse-weather observations use their annotated states. The diverse weather intensities in the collected clips provide supervision for introducing rain, snow, and fog at different intensity levels.

3.5 Network Architecture and Optimization

We build MeteoVerse upon the pretrained LingBot-World Robbyant Team et al. (2026), which consists of a high-noise model for coarse generation and a low-noise model for refinement. The input image, weather-free scene description, and camera trajectory $\mathcal{C}$ follow the original visual, textual, and camera-conditioning pathways, respectively. MeteoMoE is inserted after text cross-attention in every DiT block of both stages, with independent parameters for the high-noise and low-noise models. We freeze the pretrained backbone and optimize MeteoMoE together with LoRA adapters inserted into the self-attention and feed-forward layers. The LoRA rank and scaling factor α are set to 64 and 128, respectively.

Following LingBot-World, the video model is optimized using a flow-matching objective. Let z_t denote the latent state at flow timestep t and v_t denote the corresponding target velocity. The flow-matching loss can be written as,

$$\mathcal{L}_{\mathrm{fm}} = \mathbb{E}_{t,z_t}\left[\left\|G_{\theta,\psi}\left(z_t, t, I, p_{\mathrm{s}}, \mathcal{C}, \mathbf{s}_{\mathrm{trans}}\right) - v_t\right\|_2^2\right], \tag{9}$$

where θ denotes the frozen pretrained parameters and ψ denotes the trainable MeteoMoE and LoRA parameters. To provide additional reconstruction-level supervision, we recover the clean latent estimate $\hat{z}_0$ from z_t and the predicted velocity and decode it into $\hat{V}_{1:T}$. We apply frame-wise reconstruction and perceptual losses Johnson et al. (2016), *i.e.*,

$$\mathcal{L}_{\ell_1} = \frac{1}{T}\sum_{i=1}^{T}\left\|\hat{V}_i - V_i\right\|_1, \qquad \mathcal{L}_{\mathrm{vgg}} = \frac{1}{T}\sum_{i=1}^{T}\mathrm{VGG}\left(\hat{V}_i, V_i\right). \tag{10}$$

The overall training objective can be written as,

$$\mathcal{L} = \mathcal{L}_{\mathrm{fm}} + \lambda_{\ell_1}\mathcal{L}_{\ell_1} + \lambda_{\mathrm{vgg}}\mathcal{L}_{\mathrm{vgg}}, \tag{11}$$

where λ_{ℓ_1} and λ_{vgg} are set to 0.1 and 0.05, respectively.

4 Experiments

4.1 Implementation Details

We adopt a progressive training strategy that increases the clip length from 21 to 41 and finally 81 frames, while increasing the spatial resolution from 240×416 to 480×832. Training is performed on 8 NVIDIA A800-SXM4 GPUs with a per-GPU batch size of 1 using BF16 precision. We use AdamW with cosine learning-rate decay and gradient clipping at 1.0. The learning rate is set to 1×10^{-5} for the LoRA parameters and 5×10^{-5} for the remaining trainable parameters. The weather-state predictor is LoRA-fine-tuned using AdamW with a learning rate of 1×10^{-4} and a per-device batch size of 4. Greedy decoding is used for weather-state prediction at inference.

Table 1: Quantitative comparison across *Weather Preservation*, *Weather Introduction*, and *Weather Removal*. Overall Score is reported only for Weather Preservation, as intentional weather changes in Weather Introduction and Weather Removal make the aggregate input-consistency score less appropriate. The best and second-best results are highlighted in **bold** and underlined, respectively.

Method	VBench-I2V Evaluation							Camera Evaluation		Weather Evaluation		
	Overall Score↑	Subject Consistency↑	Background Consistency↑	Motion Smoothness↑	Dynamic Degree↑	Aesthetic Quality↑	Imaging Quality↑	RotErr↓	TransErr↓	Weather Alignment↑	VLM Evaluation↑	User Study↑
Weather Preservation												
Uni3C	86.53	95.80	94.15	99.21	30.50	50.87	63.86	0.755	0.102	88.00	74.99	76.80
GEN3C	86.22	**96.89**	**94.34**	**99.50**	23.50	50.11	60.46	**0.331**	0.073	85.50	72.30	74.90
VerseCrafter	85.01	95.20	92.68	99.17	23.00	50.86	**66.21**	5.926	0.128	90.75	**77.02**	78.60
NeoVerse	85.73	92.83	92.15	99.15	38.00	48.76	61.93	0.678	**0.054**	85.75	63.38	68.70
LingBot-World	86.56	95.57	93.84	99.06	38.00	**51.44**	63.45	4.225	0.173	88.76	74.43	75.80
Ours	**86.70**	94.78	94.03	98.86	**43.50**	50.93	63.66	3.416	0.132	**93.00**	76.07	**80.90**
Weather Introduction												
Uni3C	–	95.37	94.26	99.17	28.00	50.09	64.46	0.796	0.161	12.00	22.50	28.60
GEN3C	–	**96.84**	**94.55**	**99.49**	23.00	49.15	60.90	**0.364**	**0.056**	5.00	14.76	21.40
VerseCrafter	–	94.85	92.48	99.11	19.00	50.10	**67.46**	5.991	0.134	29.00	33.70	38.20
NeoVerse	–	92.85	91.84	99.21	39.00	47.98	61.06	0.679	0.068	9.00	17.98	24.10
LingBot-World	–	95.61	93.55	99.10	35.00	50.49	64.05	4.316	0.155	13.00	20.13	26.70
Ours	–	95.28	94.15	98.90	**47.00**	**51.09**	61.51	3.207	0.122	**61.00**	**54.15**	**66.70**
Weather Removal												
Uni3C	–	95.31	93.78	99.25	35.00	51.82	63.70	0.828	0.130	31.00	45.03	49.20
GEN3C	–	**96.94**	94.12	**99.51**	24.00	51.02	60.08	**0.327**	0.082	34.00	46.17	50.70
VerseCrafter	–	94.25	92.77	99.22	30.00	51.31	65.68	6.784	0.122	31.00	41.67	46.80
NeoVerse	–	92.66	92.43	99.07	37.00	49.47	62.55	0.671	**0.047**	38.00	40.98	45.60
LingBot-World	–	95.08	93.72	99.05	44.00	**52.15**	63.54	4.384	0.303	34.00	46.45	51.40
Ours	–	95.66	**94.14**	99.00	**48.00**	50.95	**67.03**	3.476	0.110	**89.00**	**77.92**	**82.40**

4.2 Evaluation Configurations

We evaluate on 100 held-out real-world scenes under four weather settings, *i.e.*, sunny-to-sunny, sunny-to-adverse, adverse-to-sunny, and adverse-to-adverse, resulting in 400 test cases. Sunny-to-sunny and adverse-to-adverse constitute *Weather Preservation*, while sunny-to-adverse and adverse-to-sunny correspond to *Weather Introduction* and *Weather Removal*, respectively. Weather Preservation is equally averaged over its sunny and adverse-weather subsets. All methods are evaluated using the same input images, scene descriptions, target-weather instructions, and camera trajectories. We evaluate video quality using VBench-I2V Huang et al. (2025b). Camera-control performance is measured using rotation and translation errors estimated by VGGT-Omega Wang et al. (2026). Weather controllability is evaluated using Weather Alignment, VLM Evaluation, and User Study. Overall Score is reported only for Weather Preservation, as intentional weather changes in Weather Introduction and Weather Removal make the aggregate input-consistency score less appropriate. Detailed metric definitions, generation settings, and the user-study protocol are in Appendix A.

4.3 Comparison with State-of-the-Art Methods

We compare MeteoVerse with five state-of-the-art controllable video world models with publicly available implementations or checkpoints, including Uni3C Cao et al. (2025), GEN3C Ren et al. (2025), VerseCrafter Zheng et al. (2026), NeoVerse Yang et al. (2026), and LingBot-World Robbyant Team et al. (2026). All methods are evaluated under the same protocol.

Quantitative comparisons. The quantitative results are reported in Tab. 1. Under Weather Preservation, MeteoVerse achieves the highest Overall Score, Weather Alignment, and User Study score, while remaining competitive on the individual VBench-I2V and camera-control metrics. The advantage becomes much clearer when the weather condition needs to change. For Weather Introduction, MeteoVerse improves Weather Alignment from 29.00 to 61.00, VLM Evaluation from 33.70 to 54.15, and User Study from 38.20 to 66.70. For Weather Removal, the corresponding scores improve from 38.00 to 89.00, 46.45 to 77.92, and 51.40 to 82.40. Although several baselines perform better on individual video-quality or camera metrics, these metrics mainly reflect scene consistency and trajectory following and can remain high even when the requested weather change is not realized. By explicitly modeling the weather transition, MeteoVerse achieves substantially stronger weather controllability while retaining competitive overall generation quality.

Qualitative comparisons. Fig. 4 presents qualitative comparisons on Weather Introduction and Weather Removal. For Weather Introduction, existing world models often retain the original weather

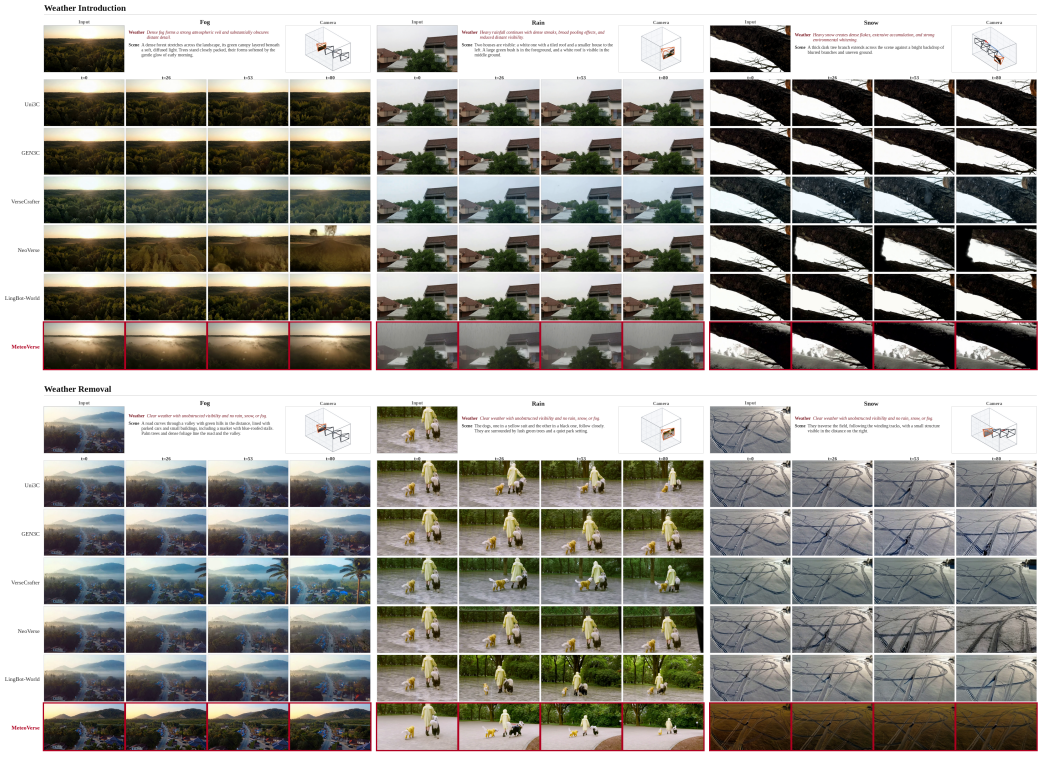

Figure 4: Visual comparisons on weather introduction and removal. More results are in the *Suppl*.

Table 2: Ablation on weather-control representations. All metrics are equally averaged over Weather Preservation, Weather Introduction, and Weather Removal.

Methods	VBench-I2V Evaluation						Camera Evaluation		Weather Evaluation		
	Subject Consistency↑	Background Consistency↑	Motion Smoothness↑	Dynamic Degree↑	Aesthetic Quality↑	Imaging Quality↑	RotErr↓	TransErr↓	Weather Alignment↑	VLM Evaluation↑	User Study↑
Target-weather Prompt p_{w}	92.60	93.14	98.87	48.50	51.88	61.30	4.084	**0.115**	80.85	63.98	73.63
Target State $\mathbf{s}_{\mathrm{tar}}$	93.96	93.29	98.78	**58.33**	**51.92**	61.76	4.372	0.189	66.03	58.73	64.07
Transition State $\mathbf{s}_{\mathrm{trans}}$ (Ours)	**95.24**	**94.11**	**98.92**	46.17	50.99	**64.07**	**3.366**	0.121	**81.00**	**69.38**	**76.67**

Table 3: Ablation on MeteoMoE and backbone adaptation. All metrics are equally averaged over Weather Preservation, Weather Introduction, and Weather Removal.

Methods	VBench-I2V Evaluation						Camera Evaluation		Weather Evaluation		
	Subject Consistency↑	Background Consistency↑	Motion Smoothness↑	Dynamic Degree↑	Aesthetic Quality↑	Imaging Quality↑	RotErr↓	TransErr↓	Weather Alignment↑	VLM Evaluation↑	User Study↑
Global Transition Conditioning	93.53	92.98	98.76	**64.17**	**51.90**	**67.51**	4.291	**0.095**	58.33	54.75	58.42
Shared Weather Expert	93.60	93.09	98.80	62.33	51.69	66.30	4.370	0.114	65.00	58.96	63.74
w/o Backbone LoRA	94.31	93.56	98.90	45.50	51.60	66.83	4.396	0.141	52.02	51.55	53.82
Full MeteoMoE (Ours)	**95.24**	**94.11**	**98.92**	46.17	50.99	64.07	**3.366**	0.121	**81.00**	**69.38**	**76.67**

appearance or produce only weak target-weather effects, whereas MeteoVerse more clearly realizes the requested rain, snow, and fog conditions. For Weather Removal, competing methods frequently leave visible adverse-weather effects, while MeteoVerse more effectively suppresses them and produces a cleaner scene appearance. The underlying scene content and camera-driven evolution remain visually consistent during weather manipulation. These qualitative observations are consistent with the quantitative weather-control results. Video comparisons are provided in the *Suppl*.

5 Ablation Studies

5.1 Effect of Weather Transition Modeling

We compare the target-weather prompt p_{w}, the target weather state $\mathbf{s}_{\mathrm{tar}}$, and the transition state $\mathbf{s}_{\mathrm{trans}}$. The target state describes the desired weather, whereas the transition state additionally ac-

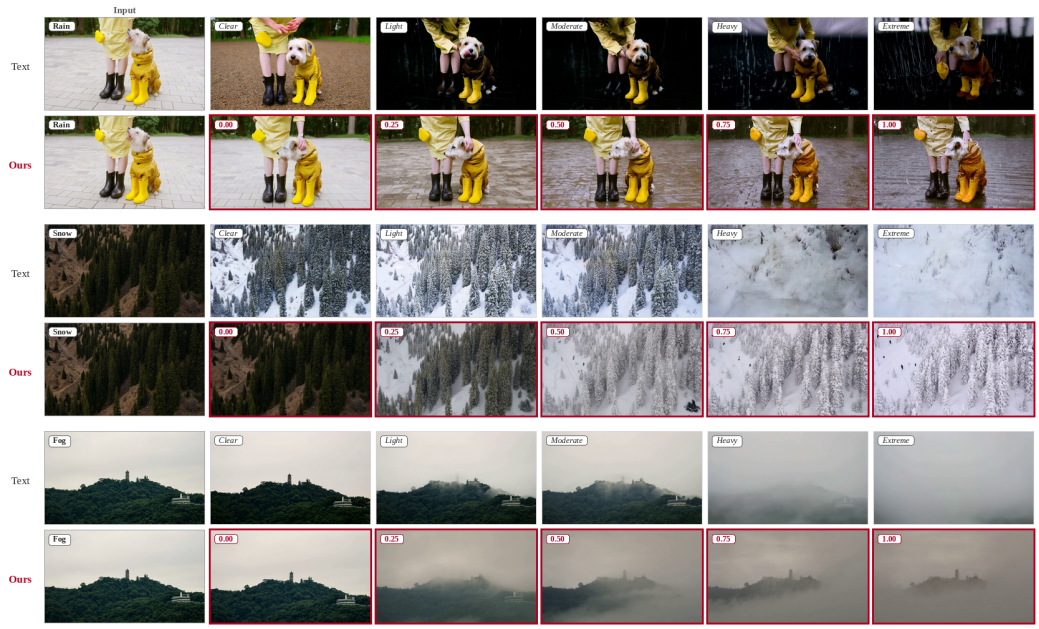

Figure 5: Visual results of fine-grained weather control. More results are in the *Suppl*.

counts for the observed condition and directly specifies the required weather modification. As shown in Tab. 2, $\mathbf{s}_{\text{trans}}$ improves Weather Alignment, VLM Evaluation, and User Study from 66.03, 58.73, and 64.07 with $\mathbf{s}_{\text{tar}}$ to 81.00, 69.38, and 76.67. Compared with p_{w}, it achieves similar Weather Alignment but higher VLM Evaluation and User Study, showing the benefit of explicitly representing the required weather change. We further evaluate fine-grained intensity control for weather introduction by fixing a sunny input, scene description, camera trajectory, and random seed while varying only the target weather intensity. As shown in Fig. 5, explicit transition control produces a clearer and more gradual progression of introduced rain, snow, and fog than text-based control. Video demonstrations are in the *Suppl*.

5.2 Effect of MeteoMoE Architecture

We evaluate how the weather transition is incorporated into the video backbone. *Global Transition Conditioning* projects $\mathbf{s}_{\text{trans}}$ into a global embedding and injects it directly into each DiT block, while *Shared Weather Expert* replaces the category-specific rain, snow, and fog experts with a single shared expert. As shown in Tab. 3, Weather Alignment improves from 58.33 with global conditioning and 65.00 with the shared expert to 81.00 with the full MeteoMoE, with consistent gains in VLM Evaluation and User Study. These results demonstrate the benefit of category-specific weather experts. Removing the backbone LoRA adapters further reduces Weather Alignment to 52.02, showing that lightweight backbone adaptation is important for effectively incorporating the transition-aware features. The full MeteoMoE also achieves the highest subject consistency, background consistency, and motion smoothness. More ablation results about the weather-state predictor are provided in Appendix B.

6 Conclusion

We presented MeteoVerse, a unified weather-controllable video world model for camera-controlled future prediction from a single sunny or adverse-weather image. MeteoVerse explicitly models the required weather transition and uses MeteoMoE to realize weather preservation, introduction, and removal, with fine-grained intensity control for weather introduction. We further constructed the MeteoVerse dataset with over 50K real-world weather clips, generated sunny counterparts, disentangled scene and weather descriptions, weather-intensity annotations, and camera trajectories. Extensive experiments demonstrate substantially improved weather controllability while retaining competitive scene consistency, temporal coherence, and camera-control performance.

REFERENCES

Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, and Di Zhang. ReCamMaster: Camera-controlled generative rendering from a single video. *arXiv preprint arXiv:2503.11647*, 2025a. doi: 10.48550/arXiv.2503.11647. URL https://arxiv.org/abs/2503.11647.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025b.

Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal Behbahani, Stephanie Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments. *arXiv preprint arXiv:2402.15391*, 2024. doi: 10.48550/arXiv.2402.15391. URL https://arxiv.org/abs/2402.15391.

Chenjie Cao, Jingkai Zhou, Shikai Li, Jingyun Liang, Chaohui Yu, Fan Wang, Xiangyang Xue, and Yanwei Fu. Uni3C: Unifying precisely 3d-enhanced camera and human motion controls for video generation. *arXiv preprint arXiv:2504.14899*, 2025.

Google DeepMind. Nano banana 2. *Model Documentation*, 2026.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. CameraCtrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024. doi: 10.48550/arXiv.2404.02101. URL https://arxiv.org/abs/2404.02101.

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. GAIA-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023. doi: 10.48550/arXiv.2309.17080. URL https://arxiv.org/abs/2309.17080.

Jiagao Hu, Daiguo Zhou, Danzhen Fu, Fuhao Li, Zepeng Wang, Fei Wang, Wenhua Liao, Jiayi Xie, and Haiyang Sun. AutoAWG: Adverse weather generation with adaptive multi-controls for automotive videos. *arXiv preprint arXiv:2604.18993*, 2026. doi: 10.48550/arXiv.2604.18993. URL https://arxiv.org/abs/2604.18993.

Tianyu Huang, Wangguandong Zheng, Tengfei Wang, Yuhao Liu, Zhenwei Wang, Junta Wu, Jie Jiang, Hui Li, Rynson W. H. Lau, Wangmeng Zuo, and Chunchao Guo. Voyager: Long-range and world-consistent video diffusion for explorable 3D scene generation. *ACM Transactions on Graphics*, 44(6):1–15, 2025a. doi: 10.1145/3763330. URL https://doi.org/10.1145/3763330.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, et al. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025b.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pp. 694–711, 2016.

Ruoteng Li, Robby T. Tan, and Loong-Fah Cheong. All in one bad weather removal using architectural search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3175–3185, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Li_All_in_One_Bad_Weather_Removal_Using_Architectural_Search_CVPR_2020_paper.html.

Yuan Li, Zhi-Hao Lin, David Forsyth, Jia-Bin Huang, and Shenlong Wang. ClimateNeRF: Extreme weather synthesis in neural radiance field. *arXiv preprint arXiv:2211.13226*, 2022. doi: 10.48550/arXiv.2211.13226. URL https://arxiv.org/abs/2211.13226.

Chih-Hao Lin, Zian Wang, Ruofan Liang, Yuxuan Zhang, Sanja Fidler, Shenlong Wang, and Zan Gojcic. Controllable weather synthesis and removal with video diffusion models. *arXiv preprint arXiv:2505.00704*, 2025. doi: 10.48550/arXiv.2505.00704. URL https://arxiv.org/abs/2505.00704.

OpenAI. Gpt-5.6 sol. OpenAI Model Documentation, 2026. GPT-5.6 Sol with max reasoning effort.

Chenghao Qian, Wenjing Li, Yuhu Guo, and Gustav Markkula. WeatherEdit: Controllable weather editing with 4D gaussian field. *arXiv preprint arXiv:2505.20471*, 2025. doi: 10.48550/arXiv.2505.20471. URL https://arxiv.org/abs/2505.20471.

Chenghao Qian, Nedko Savov, Lingdong Kong, Yeying Jin, Rui Song, Wenjing Li, Zhun Zhong, Jiaqi Ma, Gustav Markkula, and Luc Van Gool. Semantic-aware, physics-informed, geometry-grounded weather video synthesis. In *European Conference on Computer Vision (ECCV)*, 2026.

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. GEN3C: 3d-informed world-consistent video generation with precise camera control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6121–6132, 2025.

Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, Hanlin Wang, Yinghao Xu, Shuailei Ma, Yihang Chen, Jie Liu, Yansong Cheng, Yao Yao, Jiayi Zhu, Yihao Meng, Kecheng Zheng, Qingyan Bai, Jingye Chen, Zehong Shen, Yue Yu, Xing Zhu, Yujun Shen, and Hao Ouyang. Advancing open-source world models. *arXiv preprint arXiv:2601.20540*, 2026. doi: 10.48550/arXiv.2601.20540. URL https://arxiv.org/abs/2601.20540.

Christos Sakaridis, Haoran Wang, Ke Li, René Zurbrügg, Arpit Jadon, Wim Abbeloos, Daniel Olmeda Reino, Luc Van Gool, and Dengxin Dai. ACDC: The adverse conditions dataset with correspondences for robust semantic driving scene perception. *arXiv preprint arXiv:2104.13395*, 2021. doi: 10.48550/arXiv.2104.13395. URL https://arxiv.org/abs/2104.13395.

Tao Sun, Mattia Segu, Janis Postels, Yuxuan Wang, Luc Van Gool, Bernt Schiele, Federico Tombari, and Fisher Yu. SHIFT: A synthetic driving dataset for continuous multi-task domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21371–21382, 2022. URL https://arxiv.org/abs/2206.08367.

Qwen Team. Qwen3. 8-max: A new bar for coding and cowork, august 2026. *URL https://qwen.ai/blog*.

Jeya Maria Jose Valanarasu, Rajeev Yasarla, and Vishal M. Patel. TransWeather: Transformer-based restoration of images degraded by adverse weather conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2353–2363, 2022. URL https://arxiv.org/abs/2111.14813.

Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. *arXiv preprint arXiv:2408.14837*, 2024. doi: 10.48550/arXiv.2408.14837. URL https://arxiv.org/abs/2408.14837.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. Vggt-omega. *arXiv preprint arXiv:2605.15195*, 2026.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. MotionCtrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024. doi: 10.1145/3641519.3657518. URL https://arxiv.org/abs/2312.03641.

Wenhua Wu, Huai Guan, Zhe Liu, and Hesheng Wang. WeatherCity: Urban scene reconstruction with controllable multi-weather transformation. *arXiv preprint arXiv:2602.22096*, 2026. doi: 10.48550/arXiv.2602.22096. URL https://arxiv.org/abs/2602.22096.

Yijun Yang, Angelica I. Aviles-Rivero, Huazhu Fu, Ye Liu, Weiming Wang, and Lei Zhu. Video adverse-weather-component suppression network via weather messenger and adversarial backpropagation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13200–13210, 2023. URL https://arxiv.org/abs/2309.13700.

Yijun Yang, Hongtao Wu, Angelica I. Aviles-Rivero, Yulun Zhang, Jing Qin, and Lei Zhu. Genuine knowledge from practice: Diffusion test-time adaptation for video adverse weather removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25606–25616, 2024. URL https://arxiv.org/abs/2403.07684.

Yuxue Yang, Lue Fan, Ziqi Shi, Junran Peng, Feng Wang, and Zhaoxiang Zhang. NeoVerse: Enhancing 4d world model with in-the-wild monocular videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 40340–40351, 2026.

Xiangchen Yin, Wenzhang Sun, Jiahui Yuan, Zijie Liu, Yinda Chen, Wei Li, Dachun Kai, Chunfeng Wang, and Xiaoyan Sun. Holo-World: Unified camera, object and weather control for video world model. *arXiv preprint arXiv:2606.20083*, 2026. doi: 10.48550/arXiv.2606.20083. URL https://arxiv.org/abs/2606.20083.

Mark Yu, Wenbo Hu, Jinbo Xing, and Ying Shan. TrajectoryCrafter: Redirecting camera trajectory for monocular videos via diffusion models. *arXiv preprint arXiv:2503.05638*, 2025. doi: 10.48550/arXiv.2503.05638. URL https://arxiv.org/abs/2503.05638.

Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. ViewCrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024. doi: 10.48550/arXiv.2409.02048. URL https://arxiv.org/abs/2409.02048.

Howard Zhang, Yunhao Ba, Ethan Yang, Varan Mehra, Blake Gella, Akira Suzuki, Arnold Pfahnl, Chethan Chinder Chandrappa, Alex Wong, and Achuta Kadambi. WeatherStream: Light transport automation of single image deweathering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13499–13509, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Zhang_WeatherStream_Light_Transport_Automation_of_Single_Image_Deweathering_CVPR_2023_paper.html.

Yifan Zhang, Chunli Peng, Boyang Wang, Puyi Wang, Qingcheng Zhu, Fei Kang, Biao Jiang, Zedong Gao, Eric Li, Yang Liu, and Yahui Zhou. Matrix-Game: Interactive world foundation model. *arXiv preprint arXiv:2506.18701*, 2025. doi: 10.48550/arXiv.2506.18701. URL https://arxiv.org/abs/2506.18701.

Guangcong Zheng, Teng Li, Rui Jiang, Yehao Lu, Tao Wu, and Xi Li. CamI2V: Camera-controlled image-to-video diffusion model. *arXiv preprint arXiv:2410.15957*, 2024. doi: 10.48550/arXiv.2410.15957. URL https://arxiv.org/abs/2410.15957.

Sixiao Zheng, Minghao Yin, Wenbo Hu, Xiaoyu Li, Ying Shan, and Yanwei Fu. VerseCrafter: Dynamic realistic video world model with 4d geometric control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 40277–40290, 2026.

Yixin Zhu, Zuo-Liang Zhu, Jian Yang, Miloš Hašan, Jin Xie, and Beibei Wang. IntrinsicWeather: Controllable weather editing in intrinsic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30772–30781, 2026. URL https://arxiv.org/abs/2508.06982.

A Additional Evaluation Details

Benchmark. The held-out benchmark contains 100 real-world scenes with no clip- or source-level overlap with the training set. Each scene is evaluated under sunny-to-sunny, sunny-to-adverse, adverse-to-sunny, and adverse-to-adverse settings, yielding 400 test cases. Weather Preservation is averaged equally over its sunny and adverse-weather subsets. MeteoVerse generates 81-frame videos at 480×832 using 40 denoising steps, with classifier-free guidance set to 5.0.

Table 4: Effect of weather-state prediction on downstream generation. All metrics are equally averaged over Weather Preservation, Weather Introduction, and Weather Removal.

Methods	VBench-I2V Evaluation						Camera Evaluation		Weather Evaluation		
	Subject Consistency↑	Background Consistency↑	Motion Smoothness↑	Dynamic Degree↑	Aesthetic Quality↑	Imaging Quality↑	RotErr↓	TransErr↓	Weather Alignment↑	VLM Evaluation↑	User Study↑
GT States	94.35	93.65	98.87	**52.33**	50.68	63.26	3.695	**0.120**	**82.00**	**70.18**	**77.56**
Predicted States (Ours)	**95.24**	**94.11**	**98.92**	46.17	**50.99**	**64.07**	**3.366**	0.121	81.00	69.38	76.67

Evaluation metrics. We evaluate three complementary aspects. Video quality is measured using the VBench-I2V Huang et al. (2025b) metrics, including subject consistency, background consistency, motion smoothness, dynamic degree, aesthetic quality, and imaging quality. Overall Score is reported only for Weather Preservation, as intentional weather changes in Weather Introduction and Weather Removal make the aggregate input-consistency score less appropriate. Camera-control performance is evaluated using camera trajectories estimated from all generated frames by VGGT-Omega Wang et al. (2026). We report geodesic rotation error (RotErr) and scale-aligned translation error (TransErr). Weather controllability is evaluated using Weather Alignment and VLM Evaluation. Qwen 3.8 Max Team receives only the target-weather instruction and generated video. Weather Alignment measures the percentage of samples that exhibit the requested weather condition, while VLM Evaluation measures the overall quality of weather realization using a 0–100 score over weather-scene compatibility, weather dynamics, and overall video quality. For the ablation studies, all metrics are averaged over Weather Preservation, Weather Introduction, and Weather Removal.

User study. We conduct a blind user study on 36 benchmark cases, with 12 cases for each of Weather Preservation, Weather Introduction, and Weather Removal. The preservation subset contains equal numbers of sunny- and adverse-weather cases, while the introduction and removal subsets are balanced across rain, snow, and fog. We recruit 24 participants. All videos are anonymized and presented in randomized order. Each participant evaluates a randomly assigned subset, and each generated result receives at least five independent ratings. Given the input image, target-weather instruction, and generated video, participants rate weather faithfulness, preservation of non-weather scene content, and temporal naturalness on a five-point scale. For each rating, the three scores are equally averaged and linearly mapped to $[0, 100]$. We then average the scores across participants and test cases. For the main comparison, we report a separate User Study score for each weather-control capability. For the ablation and predictor analyses, we report the equal-weight macro-average over Weather Preservation, Weather Introduction, and Weather Removal.

B Effect of Weather-State Predictor

We evaluate the weather-state predictor on 400 test cases from 100 held-out scenes. It achieves 99.00% transition accuracy, measuring whether weather preservation, introduction, or removal is correctly identified. For continuous state estimation, it obtains an intensity MAE of 0.0342 and a state MAE of 0.0119. Under a maximum component-wise error tolerance of 0.10, the state accuracy reaches 89.75%. We further examine the impact of prediction errors on downstream generation by replacing ground-truth weather states with predicted states in the same frozen MeteoVerse model. As shown in Tab. 4, Weather Alignment changes from 82.00 to 81.00, while VLM Evaluation and User Study change from 70.18 and 77.56 to 69.38 and 76.67, respectively. The small differences indicate that the predicted weather states provide conditioning close to the ground-truth states for downstream weather control.