AW-COMPOSER: COMPOSITIONAL ASSET AND WEATHER EDITING FOR DRIVING VIDEOS

Anonymous authors
Paper under double-blind review

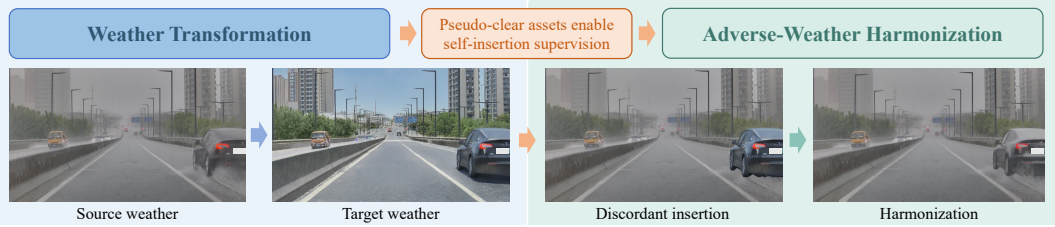

Figure 1: AW-Composer combines asset insertion with weather control. AW-Weather transforms a source video to specified weather while preserving scene structure and also generates paired self-insertion supervision. Trained with these pairs, AW-Harmony locally adapts inserted assets to rain, snow, or fog while preserving their identity, placement, and motion.

ABSTRACT

Compositional editing of driving videos offers a scalable way to create long-tail scenarios by combining asset insertion with weather control. Given a background video, an asset with specified placement, and a retained or newly specified target weather, the goal is to preserve scene structure and asset identity while maintaining spatial and temporal weather consistency. These editing dimensions are coupled: weather transformation changes global appearance and visibility, whereas insertion requires local adaptation to the resulting scene. Achieving this requires coordinated global weather editing and local asset adaptation, yet content-aligned cross-weather supervision is difficult to obtain. We present AW-Composer, a diffusion-based framework that coordinates AW-Weather and AW-Harmony. AW-Weather transforms source videos to requested weather while preserving scene content and also provides paired self-insertion supervision. Trained on these pairs, AW-Harmony uses a single-step, mask-conditioned video DiT with a multi-scale Weather Context Adapter to adapt inserted assets across frames while preserving the scene outside the insertion region. We evaluate weather transformation, paired and open-set harmonization, and specified-weather joint composition. AW-Weather achieves a human score of 93.6/100, compared with 81.2 for VACE. On paired insertion, AW-Harmony reduces FVD by 42.8% relative to the best baseline and achieves the best Weather Consistency on both insertion tracks. On specified-weather joint composition, AW-Composer reduces FVD by 22.5% relative to the strongest harmonization baseline. In the six-way human study, AW-Harmony obtains a Top-1 rate of 49.2%, nearly three times that of the next-best method.

1 INTRODUCTION

Autonomous-driving perception training and validation require long-tail objects and events under diverse weather, yet collecting diverse asset–weather combinations in the real world is difficult to scale. A scalable approach is to place specific assets into different weather environments while preserving scene structure, asset identity, placement, and motion. Existing driving-scene editors can control an inserted object's identity, pose, and trajectory (Bai et al., 2024; Chen et al., 2025b; Ljungbergh et al., 2026; Wang et al., 2025; 2026; Zhan et al., 2026), but these controls alone do not

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.

ensure weather consistency with the surrounding scene. An inserted asset may remain too clean in heavy rain, too sharp in dense fog, or temporally inconsistent with surrounding snowfall. Because the asset is the key perception target in a long-tail sample, such local mismatches make the gap between generated and real data especially salient in the target region and can reduce the utility of these samples for training and evaluation. Applying a full-scene weather transformation after insertion is also not a complete solution: non-weather content drift in the inserted-asset region can compromise the reliability of the generated data.

A viable approach is to coordinate global weather control with local asset adaptation: a weather model transforms the source video when new weather is requested, and a harmonization model then adapts the inserted asset to either the original or transformed scene. General image and video harmonizers (Cong et al., 2020; Ke et al., 2022; Huang et al., 2019; Guo et al., 2024), however, are not designed for local temporal adaptation driven by background weather evidence. Directly capturing content-aligned cross-weather training pairs would require reproducing the same asset and scene under different weather, which is difficult in practice. Instead, combining weather transformation with self-insertion constructs a locally weather-inconsistent input while retaining the original adverse-weather observation as the harmonization target, yielding a paired example from a single real observation. Open-set asset insertion further lacks paired ground truth, motivating an evaluator of local weather consistency between the inserted asset and its surrounding scene.

We therefore present AW-Composer, a diffusion-based framework that decomposes this coupled problem into global weather control and local asset adaptation. AW-Weather transforms the source video to the requested weather using continuous weather-intensity control and staged real-domain adaptation while preserving scene content. It also provides paired harmonization supervision for training AW-Harmony. Building on this supervision, AW-Harmony formulates local weather adaptation as a single-step, mask-guided conditional transport process in a video DiT. Its multi-scale Weather Context Adapter aggregates background weather evidence from near-object, lower-frame, and global regions, while a mask-aware objective strengthens supervision on the edited region and its boundary. We establish complementary Paired Self-Insertion and Open-Set Asset Insertion tracks and introduce a learned Weather Consistency evaluator for cases without paired targets.

We evaluate the framework from components to composition: global weather control and content preservation with AW-Weather, local adaptation with AW-Harmony on paired and open-set insertion, and the complete AW-Composer pipeline under newly specified rain, snow, and fog. Round-trip recovery and downstream perception studies provide complementary tests of content preservation and data utility.

Our contributions are:

- We formulate compositional asset and weather editing for driving videos, combining global weather control and local asset adaptation under retained or newly specified weather.
- We introduce AW-Weather for weather control and paired self-insertion supervision, and AW-Harmony for single-step adaptation using multi-scale background weather cues.
- We establish paired and open-set harmonization tracks with a learned Weather Consistency evaluator, and validate end-to-end composition under newly specified rain, snow, and fog.

2 Related work

Controllable driving-scene editing. Driving-scene editors use geometry, neural rendering, asset banks, and diffusion priors to insert or manipulate traffic participants (Bai et al., 2024; Ljungbergh et al., 2026; Chen et al., 2025a), with recent systems extending to multi-view, multimodal, and temporally coherent control (Lu et al., 2025; Li et al., 2026; Wang et al., 2025; Zhu et al., 2025). HorizonForge, CompoSIA, and related systems further control asset geometry, identity, trajectory, or ego action (Wang et al., 2026; Zhan et al., 2026), while complementary work evaluates synthetic scenes through perception utility (Zeng et al., 2026). Table 1 summarizes representative editing and evaluation protocols. AW-Composer couples asset insertion with global weather control and evaluates local asset–background weather agreement under retained or newly specified weather.

Work	Object insertion	Weather editing	Paired local fidelity	Temporal quality	Asset–weather agreement
SceneCrafter (Zhu et al., 2025)	✓	✓	–	–	–
R3D2 (Ljungbergh et al., 2026)	✓	–	✓	–	–
Dream4Drive (Zeng et al., 2026)	✓	–	–	✓	–
HorizonWeaver (Soroco et al., 2026)	✓	✓	✓	–	–
WeatherWeaver (Lin et al., 2025)	–	✓	–	✓	–
WeatherEdit (Qian et al., 2026)	–	✓	–	✓	–
AutoWeather4D (Liu et al., 2026)	–	✓	–	✓	–
AW-Composer (ours)	✓	✓	✓	✓	✓

Table 1: Comparison of driving-scene and weather editors. ✓: reported; –: not reported. Paired fidelity uses local targets; weather agreement measures asset–background consistency.

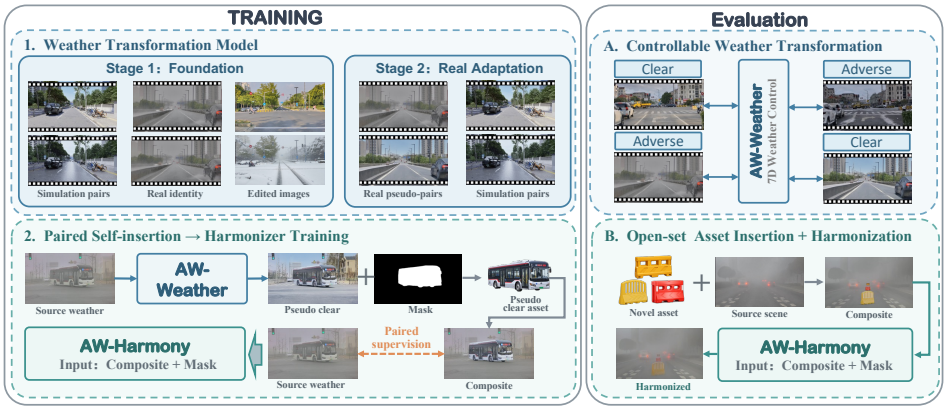

Figure 2: AW-Composer comprises AW-Weather for global weather control and AW-Harmony for local asset adaptation. AW-Weather transforms the source video to a requested condition and generates pseudo-clear foregrounds for paired self-insertion supervision; AW-Harmony adapts pasted assets to either original or transformed scenes.

Image and video harmonization. Image harmonization adjusts pasted foregrounds through learned domain transformation, explicit appearance operators, or diffusion models (Cong et al., 2020; Ke et al., 2022; Lu et al., 2023; Chadebec et al., 2025), while video methods additionally model temporal appearance variation (Huang et al., 2019; Guo et al., 2024). VACE, DiffusionHarmonizer, and related insertion and removal models provide generic masked video editing, simulation enhancement, or object-associated effect modeling (Jiang et al., 2025; Zhang et al., 2026; Gao et al., 2026; Fu et al., 2026; Chen et al., 2026). These methods do not expose target weather as a coordinated global control or explicitly use the resulting background as evidence for local weather inference. AW-Harmony lets the inserted object query multi-scale background evidence and confines adaptation to the target region, making harmonization the local stage of AW-Composer.

Controllable weather generation. Weather editing controls scene appearance through intrinsic decomposition (Zhu et al., 2026), 3D/4D rendering (Qian et al., 2026; Liu et al., 2026; Wu et al., 2026), and diffusion-based scene translation (Zhu et al., 2025). In our task, AW-Weather has two roles. At inference, it provides the requested global weather before local asset adaptation. During training, it must preserve object identity and geometry because content drift would introduce non-weather discrepancies between composites and real targets. We build on WeatherWeaver's source-conditioned video editing and staged training on multi-source pairs (Lin et al., 2025), adapting its architecture, data construction, and curriculum both for target-weather control and for generating pseudo-clear foregrounds used in paired self-insertion.

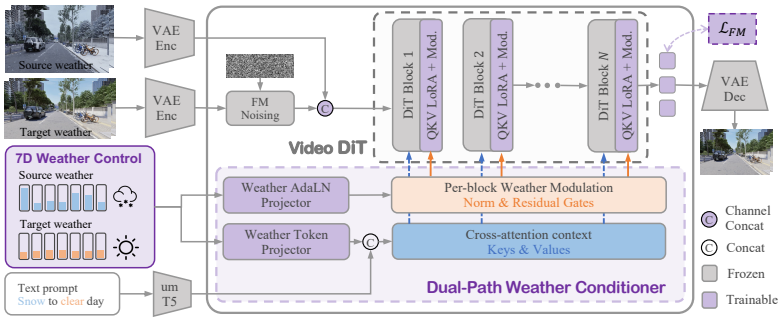

Figure 3: AW-Weather architecture. Source-video latents and raw-7D source/target weather conditions guide a Wan V2V DiT for global weather editing and pseudo-clear training-pair generation.

3 AW-Composer

AW-Composer uses AW-Weather for global weather control and AW-Harmony for single-step local asset adaptation. The former supplies requested-weather backgrounds and pseudo-clear foregrounds for paired training; the latter adapts pasted assets to original or transformed scenes, preserving identity, placement, and motion. Figure 2 shows the pipeline; Appendix A.1 defines the task.

3.1 Preliminaries

AW-Weather and AW-Harmony use separate Wan2.1-T2V-1.3B DiT parameters and share a frozen video VAE (Peebles & Xie, 2023; Wan et al., 2025). The VAE encodes video into 16-channel latents $z = E(V)$. Both modules use velocity outputs $v_\theta(x_\sigma, \sigma)$, where x_σ is the latent state at time σ. AW-Weather follows standard latent Flow Matching (Lipman et al., 2023); Section 3.2 gives its path, training target, and sampling procedure. AW-Harmony predicts the target from the composite latent in one step (mapping and loss in Section 3.4).

3.2 AW-Weather: Global control and paired supervision

Building on WeatherWeaver (Lin et al., 2025), AW-Weather T_ϕ transforms source videos from weather s_s to s_t. For paired self-insertion, it generates a pseudo-clear counterpart V^c of a real adverse-weather video V^w. We reinsert the corresponding object from V^c into V^w under its original mask and use V^w as the target. This construction requires preserving object identity and geometry so that the pair primarily differs in weather appearance.

As shown in Figure 3, we implement AW-Weather on a Wan DiT. The source-video latent is concatenated with x_σ as a content reference, while explicit source and target weather conditions specify the appearance change. These conditions use raw seven-dimensional vectors for cloud, rain, fog, puddle, snowflake, snow, and dust. A learned projector separately encodes source and target weather intensities as weather tokens prepended to the cross-attention context. The zero-initialized low-rank Weather AdaLN branch jointly encodes these intensities and their difference, $\Delta s = s_{\text{target}} - s_{\text{source}}$, to generate per-block normalization scales, shifts, and residual gates.

AW-Weather follows standard latent Flow Matching (Lipman et al., 2023). Let $z_0 = E(V_{\text{target}})$ denote the target-video latent. It uses the straight conditional path

$$\begin{aligned} x_\sigma &= (1-\sigma)z_0 + \sigma\epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \\ u^\star_\sigma &= \frac{\partial x_\sigma}{\partial \sigma} = \epsilon - z_0, \qquad \frac{\mathrm{d}x_\sigma}{\mathrm{d}\sigma} = v_\theta(x_\sigma, \sigma). \end{aligned} \tag{1}$$

$\sigma = 0$ denotes the target latent and $\sigma = 1$ Gaussian noise. Training regresses the DiT velocity to $u^\star_\sigma$ without solving the trajectory, while sampling integrates the learned field from $\sigma = 1$ to $\sigma = 0$.

Stage 1 combines simulated cross-weather pairs, real-video identity pairs, and real-image pairs edited with Qwen-Image-Edit-2511 (Wu et al., 2025). Stage 2 adapts to real videos using bidirec-

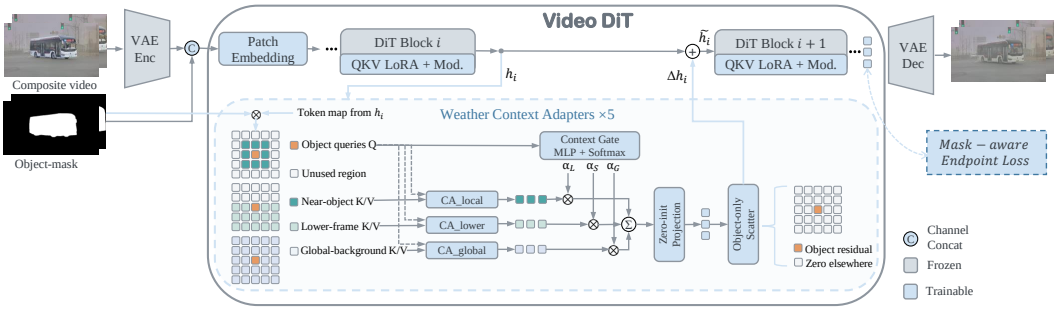

Figure 4: AW-Harmony architecture. Composite and mask latents condition a Wan DiT velocity field. At selected blocks, object queries independently read near-object, lower-frame, and global background tokens; a learned gate fuses these sources, and an adapter residual supported only on object tokens updates the DiT state.

tional observation–Stage-1-output pairs with simulated replay. Filtering removes substantial structural changes and artifacts. Appendix A.6 gives the curriculum and filtering procedure.

3.3 AW-HARMONY: MASK-GUIDED CONDITIONAL TRANSPORT

Given target-weather background B_t, asset A_P, and mask M, the composite is $C = M \odot A_P + (1 - M) \odot B_t$. Paired self-insertion uses $A_P = V^c$ and $B_t = V^w$. With $z_C = E(C)$ and downsampled mask m, AW-Harmony takes a 33-channel input $[z_C, z_C, m]$, where the two copies of z_C serve as transport state and condition. Text enters through cross-attention, and the Weather Context Adapter (Figure 4) supplies background weather evidence to object tokens.

The Weather Context Adapter samples three background pools outside the slightly dilated object mask: a near-object ring B_{local} for local attenuation and precipitation, a lower-frame region B_{lower} for near-field and road cues, and sparse global tokens B_{global} for overall visibility and illumination. Let Q_o denote tokens inside the latent mask. At selected DiT blocks, each $q \in Q_o$ reads

$$R_s(q) = \text{Attn}_s(q, B_s), \qquad s \in \{\text{local}, \text{lower}, \text{global}\}. \tag{2}$$

Each Attn_s uses source-specific projections. A gate MLP G produces logits $g_s(q) = [G(q)]_s$ to fuse the three responses per object token:

$$\alpha_s(q) = \text{softmax}_s\big(g_s(q)\big), \qquad R(q) = \sum_s \alpha_s(q) R_s(q). \tag{3}$$

At block ℓ, the adapter updates hidden states H_ℓ to H_ℓ^+:

$$H_\ell^+ = H_\ell + \mathcal{S}_m[P_{0,\ell}(R)], \tag{4}$$

where $P_{0,\ell}$ is a zero-initialized projection and $\mathcal{S}_m$ writes residuals only to object tokens. After decoding, observed pixels outside the mask are copied deterministically to the output. Adapter and LoRA parameters are trained jointly (Appendix A).

3.4 AW-HARMONY TRAINING OBJECTIVE

Given $z_0 = E(V^w)$, AW-Harmony predicts the target in one step from z_C at fixed $\bar{\sigma} = 0.30$:

$$x_{\bar{\sigma}} = z_C, \qquad \widehat{z}_0 = z_C - \bar{\sigma} v_\theta(z_C, \bar{\sigma}). \tag{5}$$

To prevent background error from overwhelming small objects, we weight the channel-averaged squared error $e = \frac{1}{C_z}\sum_c (\widehat{z}_{0,c} - z_{0,c})^2$, where C_z is the latent channel count. Let $m^+ = \text{Dilate}_1(m)$ denote a spatial dilation by one latent pixel, $b = (m^+ - m)_+$ the outer boundary band, and $\rho = \langle m^+ \rangle$ the dilated-mask occupancy. For a target object weight fraction α, we set

$$w_o = \text{clip}\left(\frac{\alpha(1-\rho)}{\rho(1-\alpha)}, w_{\min}, w_{\max}\right), \qquad W = 1 + (w_o - 1)m^+. \tag{6}$$

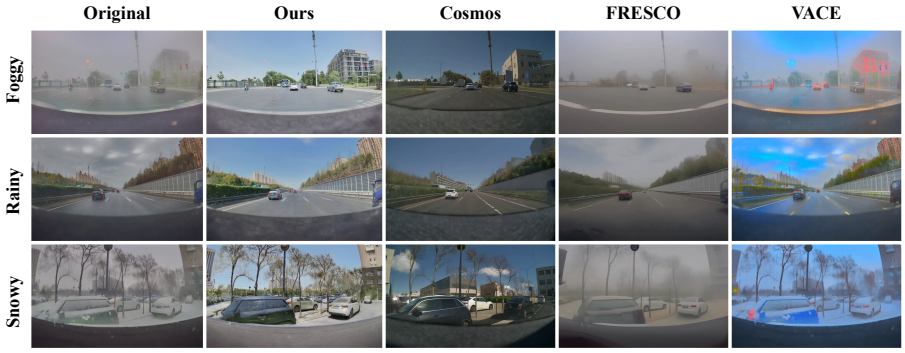

Figure 5: Comparison of clear-target transformation on foggy, rainy, and snowy driving videos.

	Blinded VLM			Human	Source
Method	Weather ↑	Technical ↑	Overall ↑	Score ↑	CLIP-S ↑
Cosmos T2.5	84.6	28.2	80.8	66.6	0.664
FRESCO	85.4	54.1	85.1	38.0	0.738
VACE-1.3B	87.9	70.7	88.3	81.2	0.752
AW-Weather (Stage 2)	**92.7**	**93.1**	**94.9**	**93.6**	**0.784**

Table 2: Clear-target transformation quality on real rain, snow, and fog driving videos.

Here, $\langle\cdot\rangle$ denotes a temporal-spatial mean, and sums below span the same grid. Clipping bounds object weights, and $\varepsilon > 0$ stabilizes normalization; fixed settings are in Appendix A. We minimize

$$\mathcal{L}_{\text{ep}} = \frac{\sum We}{\sum W + \varepsilon}, \qquad \mathcal{L}_{\text{bg}} = \frac{\sum (1-m)e}{\sum (1-m) + \varepsilon},$$
$$\mathcal{L}_{\text{bnd}} = \frac{\sum be}{\sum b + \varepsilon}, \qquad \mathcal{L}_{\text{harm}} = \lambda_{\text{ep}}\mathcal{L}_{\text{ep}} + \lambda_{\text{bg}}\mathcal{L}_{\text{bg}} + \lambda_{\text{bnd}}\mathcal{L}_{\text{bnd}}. \tag{7}$$

$\mathcal{L}_{\text{bg}}$ constrains the region outside the original mask, and $\mathcal{L}_{\text{bnd}}$ its seam-prone outer band.

4 Experiments

4.1 Experimental setup

Data and tracks. AW-Weather uses Unreal Engine simulation (Epic Games, n.d.), real-video identity, and edited real-image pairs in Stage 1; Stage 2 mixes 70% real-domain with 30% simulated pairs. Its 200-video test covers real rain, snow, and fog from scene-disjoint internal holdouts and nuScenes (Caesar et al., 2020). AW-Harmony trains on 26,600 masked 21-frame clips across 23 object categories, eight weather labels, and four scale bins. The Paired Self-Insertion Track has 622 scene-disjoint clips with observed adverse-weather targets; the Open-Set Asset Insertion Track has 475 cross-source composites in heavy rain, dense fog, and heavy snow without paired targets. Data and curricula are detailed in Appendices A.6 and B.

Specified-Weather Joint Composition Track. We transform 50 clear sequences to heavy rain, heavy snow, and fog, then reinsert one source vehicle per sequence at its original location. The 150 cases use AW-Weather outputs as paired targets, not independently captured cross-weather observations. Round-trip recovery maps their harmonized outputs back to clear, using the original videos as targets (Appendix B.2).

Metrics. PSNR, SSIM (Wang et al., 2004), LPIPS (Zhang et al., 2018), FID (Heusel et al., 2017), and FVD (Unterthiner et al., 2019) measure paired fidelity on temporal union-mask crops (full frames for round-trip recovery). CLIP-GT, CLIP-S, and CLIP-D (Radford et al., 2021) measure target agreement, source similarity, and directional adaptation; open-set evaluation adds VBench

	Mask-local fidelity					Weather Consistency			Semantic
Method	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	FVD ↓	WC Score ↑	Emb. Dist. ↓	Same-W. (%) ↑	CLIP-GT ↑
VTT	32.5	0.976	0.044	13.4	44.7	0.499	0.256	69.6%	0.951
Harmonizer	**34.0**	**0.979**	0.029	6.8	26.4	0.510	0.250	70.6%	0.960
LBM	29.9	0.959	0.040	10.4	40.6	0.499	0.258	66.0%	0.952
VACE-1.3B	31.4	0.964	0.046	10.6	31.9	0.516	0.246	72.0%	0.939
DiffusionHarm.	31.0	0.975	0.033	7.7	29.9	0.492	0.262	66.6%	0.957
AW-Harmony	33.5	0.966	**0.028**	**6.5**	**15.1**	**0.546**	**0.228**	**75.3%**	**0.967**

Table 3: Reconstruction quality, weather consistency, and semantic preservation on the Paired Self-Insertion Track. Best values are shown in bold.

	Learned Weather Consistency metrics				Full-frame quality
Method	WC Score ↑	Emb. Dist. ↓	Same-W. (%) ↑	Int. Dist. ↓	VBench ↑
Direct composite*	0.421	0.310	55.8%	0.305	0.762
VTT	0.466	0.273	67.8%	0.283	0.770
Harmonizer	0.437	0.296	60.7%	0.298	0.775
LBM	0.456	0.284	65.4%	0.283	**0.780**
VACE-1.3B	0.473	0.270	66.9%	0.281	0.767
DiffusionHarm.	0.432	0.301	58.9%	0.299	0.778
AW-Harmony	**0.492**	**0.258**	**72.4%**	**0.261**	0.775

Table 4: Weather consistency and full-frame quality on the Open-Set Asset Insertion Track. *Direct composite is the unharmonized input reference and is excluded from best-value ranking.

Quality (Huang et al., 2024). A learned ResNet-50 (He et al., 2016) compares object and shared reference-background crops using WC Score, embedding/intensity distances, and same-weather rate. For round-trip recovery, it also reports full-frame clear-weather similarity (Clear Similarity). Evaluator training and calibration are scene-disjoint from harmonizer training and evaluation; definitions and calibration limits appear in Appendix B.

Baselines and inference. Weather baselines are FRESCO (Yang et al., 2024), Cosmos Transfer2.5-2B (NVIDIA et al., 2026), and VACE-1.3B (Jiang et al., 2025), evaluated on the same 200 videos with a clear target. Harmonization baselines are Harmonizer (Ke et al., 2022), VTT (Guo et al., 2024), frame-wise LBM (Chadebec et al., 2025), VACE-1.3B, and temporal DiffusionHarmonizer (Zhang et al., 2026), using official code with shared composites and masks. AW-Harmony takes 12.5 seconds per 21-frame clip on one RTX 5090, excluding loading (Appendix A).

4.2 AW-Weather clear-target transformation quality

AW-Weather leads all three blinded gpt-5.6-sol axes, Human Score, and Source CLIP-S (Table 2), with the largest VLM advantage in technical quality. Figure 5 illustrates weather removal and scene preservation; Appendix C gives the evaluation protocol.

4.3 Paired Self-Insertion Track: adverse-weather reconstruction

Table 3 shows that AW-Harmony attains the best LPIPS, FID, FVD, and Weather Consistency, reducing the next-best FVD by 42.8%. In Appendix Figure 12, directly inserted objects remain too clean in snow and rain and too sharp in dense fog; AW-Harmony more closely matches the paired ground truth in visibility attenuation and local weather effects across frames. Appendix Figure 13 provides the complete baseline comparison.

4.4 Open-Set Asset Insertion Track: automatic and perceptual evidence

In Table 4, AW-Harmony achieves the best Weather Consistency with comparable full-frame quality. In Figure 6, LBM and Diffusion-Harmonizer often leave assets overly clean or high-contrast, while VTT and VACE introduce intrinsic-color or structural drift; AW-Harmony restores weather features that best match the background while preserving object silhouettes and intrinsic colors. Blinded gpt-



Figure 6: Qualitative comparison on the Open-Set Asset Insertion Track under heavy rain and dense fog. Each example shows four temporal frames, with red boxes enlarging the inserted-asset regions.

5.6-sol comparisons favor AW-Harmony over every baseline (mean pairwise win rate: 82.5%). Its human six-way Top-1 share is 49.2% (95% CI: 45.8–52.6), compared with 16.6% for the next-best method (Appendix C, Table 14).

4.5 Specified-Weather Joint Composition Track

AW-Weather first generates the requested-weather background, and AW-Harmony adapts the pasted clear vehicle. All harmonizers share the transformed videos, composites, and masks. Across 150 cases, AW-Composer leads all aggregate metrics against generated targets (Table 5), reducing FVD from the strongest baseline's 179.05 to 138.73 (22.5%). It also achieves the lowest FVD for each target weather (Appendix B.2). Appendix Figure 11 shows rain and snow examples.

4.6 Harmonization-first round-trip recovery

The default pipeline applies AW-Weather before AW-Harmony. To test content preservation in the reverse order, each weather editor maps the full harmonized outputs from Section 4.5 back to clear, using the original clear videos as paired full-frame targets. AW-Weather leads all aggregate metrics (Table 6), reducing FID and FVD by 35.8% and 22.6% against their respective strongest baselines.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	FVD ↓	WC Score ↑	CLIP-GT ↑
VTT	20.67	0.909	0.134	79.97	297.69	0.467	0.942
Harmonizer	24.62	0.917	0.114	57.92	192.19	0.486	0.948
LBM	23.29	0.920	0.107	49.38	179.05	0.493	0.945
VACE-1.3B	22.31	0.891	0.137	54.15	195.94	0.481	0.940
DiffusionHarm.	19.75	0.898	0.127	55.02	236.24	0.482	0.932
AW-Harmony	**26.74**	**0.929**	**0.088**	**36.96**	**138.73**	**0.505**	**0.962**

Table 5: Results on the Specified-Weather Joint Composition Track.

4.7 Ablation studies

Table 7 tests context, objective, and supervision. At a fixed context-token budget, adding lower-frame and then local context raises WC Score from 0.469 to 0.481 and 0.492; all three adapter variants outperform ControlNet-style background injection (Zhang et al., 2023). Figure 7(a) shows improved rain and mist attenuation as context pools are added. The mask-aware objective improves

Weather editor	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	FVD ↓	CLIP-GT ↑	Clear Similarity ↑
VACE-1.3B	13.45	0.567	0.515	97.82	1689.50	0.823	0.739
FRESCO	12.75	0.539	0.590	121.96	1476.43	0.778	0.336
Cosmos T2.5	12.16	0.497	0.520	88.26	1167.66	0.835	0.635
AW-Weather	**17.73**	**0.606**	**0.343**	**56.65**	**903.39**	**0.900**	**0.920**

Table 6: Harmonization-First Round-Trip Recovery.

Study	Variant	WC Score ↑	Same-W. (%) ↑	CLIP-S ↑	CLIP-D ↑	VBench ↑
Context	ControlNet	0.465	68.4%	0.9079	0.018	0.7732
	Global	0.469	68.7%	0.9067	0.018	0.7734
	+ Lower-frame	0.481	70.5%	0.9084	0.025	0.7741
	+ Local	**0.492**	**72.4%**	**0.9085**	**0.027**	**0.7752**
Objective	w/o mask-aware loss	0.483	70.8%	**0.9138**	0.021	0.7743
	w/ mask-aware loss	**0.492**	**72.4%**	0.9085	**0.027**	**0.7752**
Supervision	Color-only	0.427	57.5%	0.9009	0.011	0.7516
	AW-Harmony	**0.492**	**72.4%**	**0.9085**	**0.027**	**0.7752**

Table 7: Ablations of spatial context, supervision, and objective on the Open-Set Asset Insertion Track. Results are compared within each study.

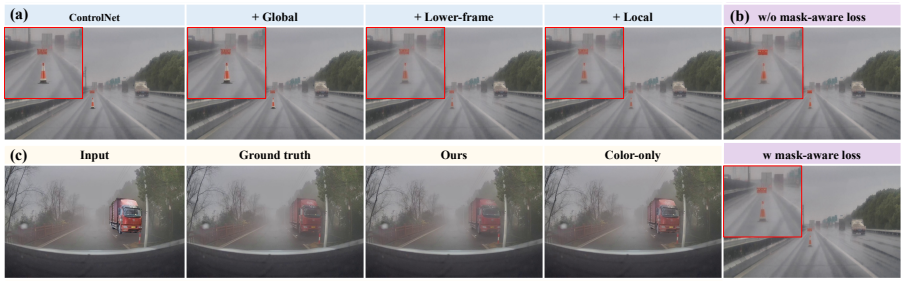

Figure 7: Qualitative comparisons of (a) background feature injection and multi-scale context, (b) removing the mask-aware loss, and (c) replacing our paired supervision with color-only training.

WC Score and CLIP-D over uniform latent reconstruction and restores local weather effects more fully (b). Weather-transformed paired supervision improves both metrics over Color-only training (c), supporting the value of weather-specific supervision beyond color adjustment. Appendix B.6 reports the complete metrics.

4.8 Downstream Data Utility

On nuScenes, we fine-tune Faster R-CNN R50-FPN with five data variants under equal training budgets. On scene-disjoint real validation images, joint asset and weather augmentation improves Clear/Rain AP from 26.85/22.23 to 27.23/24.95 over real-only training (Appendix B.7). For internal deployment, we compare OCC training on approximately 340,000 20-second adverse-weather rare-obstacle insertion clips with and without AW-Harmony (Appendix A.5). Both groups use the same insertions and training conditions. Harmonization raises AEB recall from 88.9% to 94.4% (+5.5 pp), with a relative gain of 33.3% in rainy water-barrier scenarios (Appendix Figure 14).

5 Conclusion

We presented AW-Composer, which combines global weather control with single-step local asset adaptation and constructs paired supervision from real observations. Paired and open-set insertion, specified-weather composition, and round-trip recovery support its generation quality and content preservation, with human and VLM evaluations supporting perceptual quality. Internal deployment further illustrates the utility of harmonized assets for perception training.

ETHICS STATEMENT

The human evaluations used perceptual ratings or preference selection to assess weather-transformation and asset-harmonization results, with method identities hidden. Participants received the evaluation instructions, no personal or sensitive participant attributes were collected, and only aggregate evaluation results are reported. Internal driving data remain access-controlled because they may contain sensitive road-scene and proprietary information. Generated videos are retained as synthetic data and are used for controlled data augmentation and model evaluation; asset placement and trajectories are validated by the upstream simulation pipeline.

REPRODUCIBILITY STATEMENT

Sections 3 and 4 describe the compositional task, model, objectives, data construction, baselines, and evaluation. Appendix A reports model configurations, training curricula, and inference settings; Appendix B specifies the datasets, joint-editing and round-trip protocols, metrics, evaluator training, ablations, and downstream experiments; and Appendix C provides human and VLM preference results for asset harmonization, perceptual-evaluation protocols, and VLM prompts. For the internal OCC/AEB experiments, we report only publicly disclosable aggregate results; the data and production models remain access-controlled. Subject to the necessary authorization, we plan to publicly release the relevant code and test sets before publication.

AI USE STATEMENT

Large language models (LLMs) were used for code assistance, manuscript polishing, and literature-informed drafting of the downstream detection evaluation protocol. The authors take full responsibility for the entire paper, including the experimental design and content generated or modified with LLM assistance.

REFERENCES

Chen Bai, Zeman Shao, Guoxiang Zhang, Di Liang, Jie Yang, Zhuorui Zhang, Yujian Guo, Chengzhang Zhong, Yiqiao Qiu, Zhendong Wang, et al. Anything in any scene: Photorealistic video object insertion. *arXiv preprint arXiv:2401.17509*, 2024.

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11621–11631, 2020.

Clément Chadebec, Onur Tasar, Sanjeev Sreetharan, and Benjamin Aubin. Lbm: Latent bridge matching for fast image-to-image translation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 29086–29098. IEEE, 2025.

Xiaoxue Chen, Bhargav Chandaka, Chih-Hao Lin, Ya-Qin Zhang, David Forsyth, Hao Zhao, and Shenlong Wang. Invrgb+ l: inverse rendering of complex scenes with unified color and lidar reflectance modeling. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 27176–27186. IEEE, 2025a.

Yuqing Chen, Lin Liu, Haisu Wu, Xiaopeng Zhang, Yaowei Wang, Yujiu Yang, and Qi Tian. Generaser: Generalizable video object removal via balanced text-mask guidance and decoupled locator-preserver. *arXiv preprint arXiv:2605.30045*, 2026.

Zhengqing Chen, Ruohong Mei, Xiaoyang Guo, Qingjie Wang, Yubin Hu, Wei Yin, Weiqiang Ren, and Qian Zhang. SynthDrive: Scalable Real2Sim2Real sensor simulation pipeline for high-fidelity asset generation and driving data synthesis, 2025b. URL https://arxiv.org/abs/2509.06798.

Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. Dovenet: Deep image harmonization via domain verification. In *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 8391–8400. IEEE, 2020.

Epic Games. Unreal Engine. https://www.unrealengine.com/, n.d. Accessed September 14, 2026.

Yang Fu, Yike Zheng, Ziyun Dai, and Henghui Ding. Effecterase: Joint video object removal and insertion for high-quality effect erasing. *arXiv preprint arXiv:2603.19224*, 2026.

Xiangbo Gao, Renjie Li, Xinghao Chen, Yuheng Wu, Suofei Feng, Qing Yin, and Zhengzhong Tu. Pisco: Precise video instance insertion with sparse control. *arXiv preprint arXiv:2602.08277*, 2026.

Zonghui Guo, Xinyu Han, Jie Zhang, Shiguang Shan, and Haiyong Zheng. Video harmonization with triplet spatio-temporal variation patterns. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19177–19186. IEEE, 2024.

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

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Hao-Zhi Huang, Sen-Zhe Xu, Jun-Xiong Cai, Wei Liu, and Shi-Min Hu. Temporally coherent video harmonization using adversarial networks. *IEEE Transactions on Image Processing*, 29:214–224, 2019.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17191–17202. IEEE, 2025.

Zhanghan Ke, Chunyi Sun, Lei Zhu, Ke Xu, and Rynson WH Lau. Harmonizer: Learning to perform white-box image and video harmonization. In *European conference on computer vision*, pp. 690–706. Springer, 2022.

Jiusi Li, Jackson Jiang, Jinyu Miao, Miao Long, Tuopu Wen, Peijin Jia, Shengxiang Liu, Chunlei Yu, Maolin Liu, Yuzhan Cai, et al. Realistic and controllable 3d gaussian-guided object editing for driving video generation. In *2026 IEEE Intelligent Vehicles Symposium (IV)*, pp. 60–67. IEEE, 2026.

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. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13580–13591. IEEE, 2025.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Tianyu Liu, Weitao Xiong, Kunming Luo, Manyuan Zhang, Peng Li, Yuan Liu, and Ping Tan. Autoweather4d: Autonomous driving video weather conversion via g-buffer dual-pass editing. *arXiv preprint arXiv:2603.26546*, 2026.

William Ljungbergh, Bernardo Taveira, Wenzhao Zheng, Adam Tonderski, Chensheng Peng, Fredrik Kahl, Christoffer Petersson, Michael Felsberg, Kurt Keutzer, Masayoshi Tomizuka, et al. R3d2: Realistic 3d asset insertion via diffusion for autonomous driving simulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 925–934, 2026.

Lingxiao Lu, Jiangtong Li, Junyan Cao, Li Niu, and Liqing Zhang. Painterly image harmonization using diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 233–241, 2023.

Shouyi Lu, Zihan Lin, Chao Lu, Huanran Wang, Guirong Zhuo, and Lianqing Zheng. Multieditor: Controllable multimodal object editing for driving scenarios using 3d gaussian splatting priors. *arXiv preprint arXiv:2507.21872*, 2025.

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, 2026. URL `https://arxiv.org/abs/2511.00062v2`.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Chenghao Qian, Wenjing Li, Yuhu Guo, and Gustav Markkula. Weatheredit: Controllable weather editing with 4d gaussian field. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 8511–8519, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, 2015. URL https://arxiv.org/abs/1506.01497.

Mauricio Soroco, Francesco Pittaluga, Zaid Tasneem, Abhishek Aich, Bingbing Zhuang, Wuyang Chen, Manmohan Chandraker, and Ziyu Jiang. HorizonWeaver: Generalizable multi-level semantic editing for driving scenes, 2026. URL https://arxiv.org/abs/2604.04887v1.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. FVD: A new metric for video generation. In *ICLR Workshop*, 2019. URL https://openreview.net/pdf?id=rylgEULtdN.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Shuyun Wang, Haiyang Sun, Bing Wang, Hangjun Ye, and Xin Yu. Mirage: One-step video diffusion for photorealistic and coherent asset editing in driving scenes. *arXiv preprint arXiv:2512.24227*, 2025.

Yifan Wang, Francesco Pittaluga, Zaid Tasneem, Chenyu You, Manmohan Chandraker, and Ziyu Jiang. Horizonforge: Driving scene editing with any trajectories and any vehicles. *arXiv preprint arXiv:2602.21333*, 2026.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. doi: 10.1109/TIP.2003.819861.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. URL https://arxiv.org/abs/2508.02324.

Wenhua Wu, Huai Guan, Zhe Liu, and Hesheng Wang. Weathercity: Urban scene reconstruction with controllable multi-weather transformation. *arXiv preprint arXiv:2602.22096*, 2026.

Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Fresco: Spatial-temporal correspondence for zero-shot video translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8703–8712, 2024.

Kai Zeng, Zhanqian Wu, Kaixin Xiong, Xiaobao Wei, Xiangyu Guo, Zhenxin Zhu, Kalok Ho, Lijun Zhou, Bohan Zeng, Ming Lu, et al. Rethinking driving world model as synthetic data generator for perception tasks. In *International Conference on Learning Representations*, volume 2026, pp. 135507–135534, 2026.

Yifan Zhan, Zhengqing Chen, Qingjie Wang, Zhuo He, Muyao Niu, Xiaoyang Guo, Wei Yin, Weiqiang Ren, Qian Zhang, and Yinqiang Zheng. Composing driving worlds through disentangled control for adversarial scenario generation, 2026.

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*, pp. 3836–3847, 2023.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Yuxuan Zhang, Katarína Tóthová, Zian Wang, Kangxue Yin, Haithem Turki, Riccardo de Lutio, Yen-Yu Chang, Or Litany, Sanja Fidler, and Zan Gojcic. Diffusionharmonizer: Bridging neural reconstruction and photorealistic simulation with online diffusion enhancer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 43494–43504, 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://openaccess.thecvf.com/content/CVPR2026/html/Zhu_IntrinsicWeather_Controllable_Weather_Editing_in_Intrinsic_Space_CVPR_2026_paper.html.

Zehao Zhu, Yuliang Zou, Chiyu Max Jiang, Bo Sun, Vincent Casser, Xiukun Huang, Jiahao Wang, Zhenpei Yang, Ruiqi Gao, Leonidas Guibas, et al. Scenecrafter: Controllable multi-view driving scene editing. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6812–6822. IEEE, 2025.

A Implementation details

A.1 Compositional task formulation

Let B be a background video, A_P an asset rendered under placement condition P with spatiotemporal mask M, and s_t the target weather. We seek an edited video that preserves the structure of B and the identity, placement, and motion of A_P, while making the asset and background consistent with s_t across frames. AW-Composer implements this mapping with AW-Weather as the weather transformation model T_ϕ and AW-Harmony as the video harmonizer H_θ:

$$\begin{aligned} B_t &= \begin{cases} B, & \text{if the input weather is retained,} \\ T_\phi(B, s_t), & \text{if a new weather is specified,} \end{cases} \\ C &= M \odot A_P + (1 - M) \odot B_t, \\ \widehat{V} &= H_\theta(C, M). \end{aligned} \tag{8}$$

The target-weather background remains visible outside M in C, where AW-Harmony reads it as context. Thus the same local adaptation stage handles both task settings; only the source of its background changes. Figure 2 summarizes the inference path and the paired supervision that connects the two modules.

A.2 Driving-video VAE

We use a Wan-VAE adapted for driving video and initialize it from the pretrained VAE of Wan2.1-T2V-1.3B (Wan et al., 2025), whose original spatiotemporal compression ratio is $4 \times 8 \times 8$. Driving clips often contain substantial inter-frame displacement from ego motion, steering, braking, and nearby-object interactions. We therefore remove the intermediate temporal-compression stage and set the temporal stride to one while retaining 8×8 spatial compression. The resulting $1 \times 8 \times 8$ tokenizer preserves one latent timestep per input frame.

Removing temporal compression changes the feature distribution throughout the encoder–decoder. Starting from the pretrained weights, we therefore fine-tune all VAE modules on 100 hours of self-collected driving video for seven days using eight NVIDIA A100 80GB GPUs. We retain the losses, learning rate, and optimization schedule of the original Wan-VAE training recipe (Wan et al., 2025). The resulting VAE is frozen when training AW-Harmony.

A.3 AW-Harmony configuration

AW-Harmony uses Wan2.1-T2V-1.3B with a 33-channel latent input: 16 transport-state channels initialized with the composite latent, 16 composite-conditioning channels, and one object-mask channel. Weather Context Adapters are inserted at blocks $[10, 14, 18, 22, 26]$ with width 1,536 and 12 heads. Each adapter samples at most 192 near-object background tokens, 64 lower-frame background tokens, and 32 global background tokens. The near-object ring uses inner and outer radii of one and six latent pixels. The trainable low-rank path is a rank-96 LoRA.

AW-Harmony uses conditional-latent initialization at the fixed transport coordinate $\bar{\sigma} = 0.30$. Its loss weights are $(\lambda_{\text{ep}}, \lambda_{\text{bg}}, \lambda_{\text{bnd}}) = (1.0, 1.0, 0.3)$, its object-weight bounds are $(w_{\min}, w_{\max}) = (3, 80)$, and the target object share of the spatial weight is $\alpha = 0.5$.

A.4 Optimization details

AW-Harmony is trained for 30,000 steps on 32 NVIDIA RTX 5090 GPUs with bfloat16 precision and DeepSpeed ZeRO-2. We use a per-GPU batch size of one without gradient accumulation. AdamW uses $\beta = (0.9, 0.99)$, weight decay 0.01, and gradient clipping at 0.3. The Weather Context Adapters, rank-96 LoRA, time projection, and modulation layers use a learning rate of 10^{-4}; the input patch embedding uses 2×10^{-5}. The learning rate has 1,000 warm-up steps followed by cosine decay. The remaining pretrained DiT parameters and the VAE are frozen.

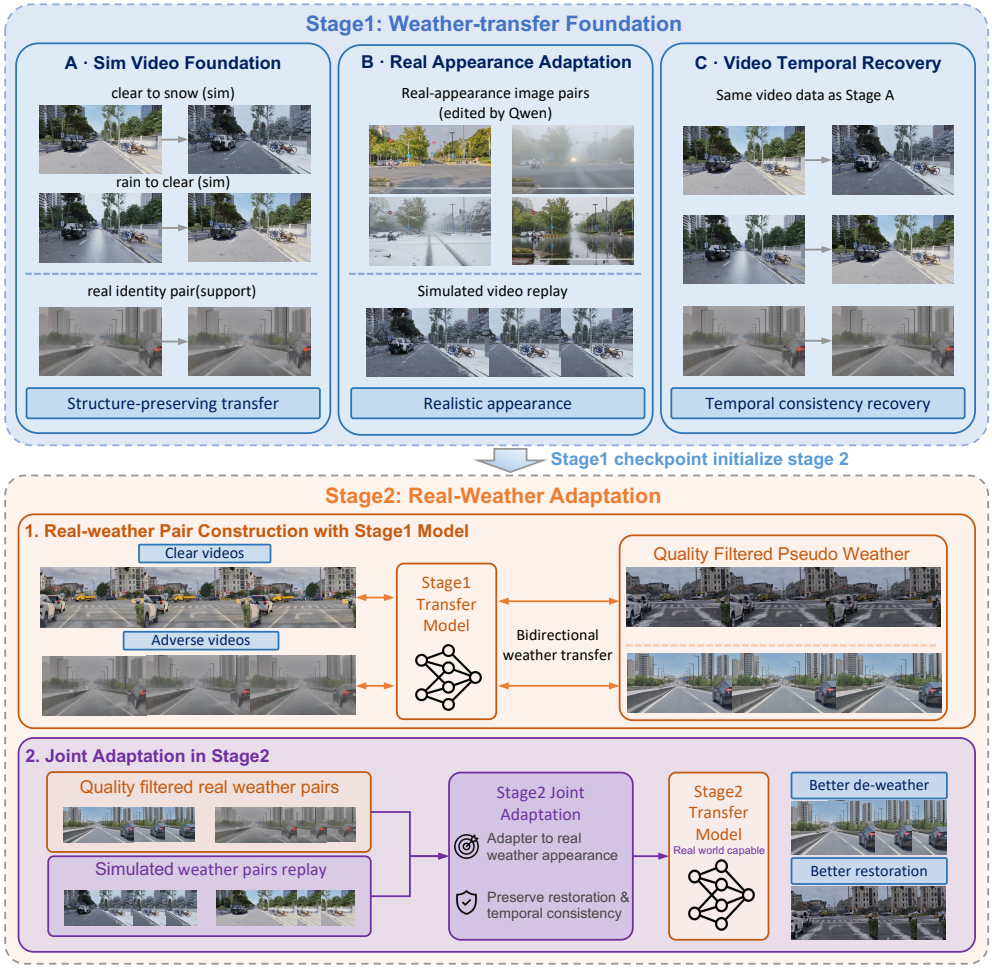

Figure 8: AW-Weather curriculum. Stage 1 comprises simulated-pair training, real-image adaptation with selected temporal parameters frozen, and a final video-only training phase. Stage 2 combines bidirectional real-domain adaptation with simulated-pair replay.

A.5 Inference Details

On the two evaluation tracks, AW-Harmony takes 21 frames at 832×480 and 10 fps and starts inference from the composite latent at the fixed transport coordinate $\bar{\sigma} = 0.30$. We export all 21 frames. Every comparison method receives the same composite and mask, uses no paired ground truth at inference, and follows the same evaluation and export protocol. Public methods follow their official open-source implementations; VACE uses 50-step UniPC inference.

Long-sequence deployment. Deployment videos are processed with 21-frame sliding windows and 10-frame overlap. Each window after the first uses one test-time update of the Weather Context Adapter, with sequence-specific adaptation state carried across windows to promote consistent appearance. States are initialized independently for each sequence, and overlapping output frames are combined by linear blending.

A.6 AW-Weather Curriculum

Stage 1. We train 21-frame clips at 832×480 for 30,000 steps. The 15,000-step foundation phase uses a video stream containing 90% simulated transfer pairs and 10% training pairs in which the input and target are the same real video. Its learning rate starts at 10^{-4} with 1,000 warm-up steps and decays to 10^{-5} over the final 3,000 steps. The following 10,000-step domain-adaptation phase

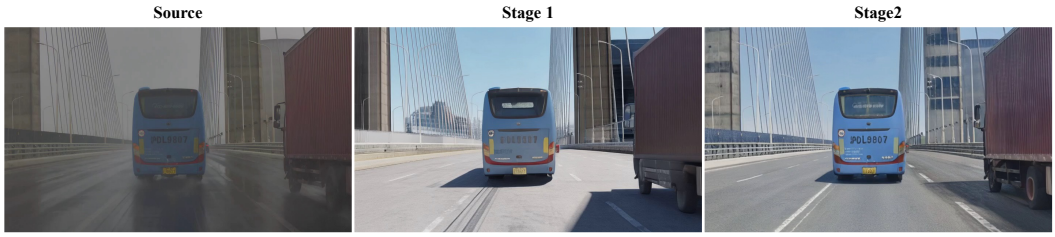

Figure 9: Stage-wise AW-Weather transformation on a real adverse-weather source. Compared with Stage 1, Stage 2 retains the pseudo-clear target with better source-content preservation.

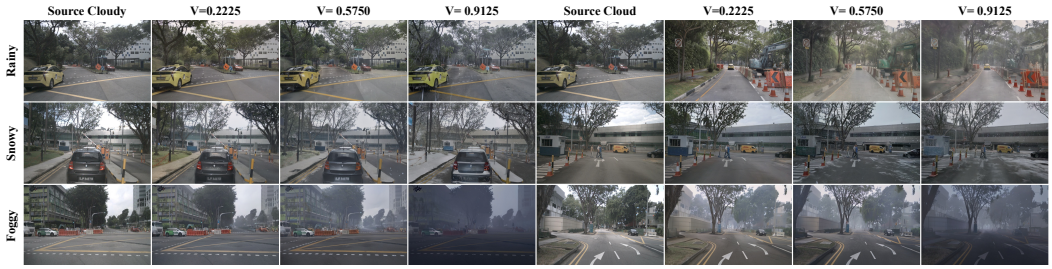

Figure 10: Weather-intensity control. For each source video, increasing the target control value V from 0.2225 to 0.9125 progressively strengthens rain, snow, or fog while preserving the scene layout.

uses the same video stream and introduces real image pairs on 10% of optimizer steps at loss weight 0.1. Temporal LoRA and Weather AdaLN parameters are frozen on these image steps. The final 5,000-step video-only phase restores temporal training while cosine-decaying the learning rate from 10^{-5} to 3×10^{-6}.

Stage 2. Starting from the Stage-1 checkpoint, we continue training for 10,000 steps with 70% real-domain pairs and 30% simulated pairs. Each real observation and its Stage-1 transformation are used in both directions; within the real-domain portion, generated-to-real and real-to-generated pairs are sampled equally. We use AdamW with a learning rate of 10^{-5}, 500 warm-up steps, cosine decay, $\beta = (0.9, 0.99)$, weight decay 0.01, and gradient clipping at 0.3. Both stages use 32 GPUs, bfloat16, DeepSpeed ZeRO-2, a per-GPU batch size of one, and no gradient accumulation. We optimize rank-96 LoRA together with the patch embedding, time projection, modulation, weather projector, and rank-16 Weather AdaLN modules; the remaining DiT parameters and the VAE are frozen.

Real-pair construction. Each pair uses the same 21-frame real clip and frame window: one side is the real observation and the other is its Stage-1 transformation with matching frame names. We reject missing, unreadable, misordered, resolution-mismatched, severely blurred, saturated, or near-duplicate sequences. Canny-edge overlap and Sobel-gradient correlation reject transformations with large structural changes. For rain data, a required road segmentation mask limits an additional filter to road surfaces and rejects red reflection artifacts. The remaining clips are split 95/5 by real scene and frame window before bidirectional expansion, preventing a source window from crossing the train/validation split. These checks reduce non-weather differences but do not establish exact pixel correspondence.

Conditioning and inference. The model uses raw-7D source and target conditions throughout training and inference. The seven channels are cloud, rain, fog, puddle, snowflake, snow, and dust. Projected weather tokens enter the cross-attention context, while Weather AdaLN receives source, target, and target-minus-source values at every DiT block. Inference takes 21 frames at 832×480, uses a FlowUniPC solver for 20 flow-matching steps with guidance scale 1.5, and sets both training and sampling shifts to 1.0. Figure 10 visualizes graded control of rain, snow, and fog.

Method	Heavy rain	Heavy snow	Fog
VTT	418.67	428.08	223.06
Harmonizer	261.23	286.40	245.00
LBM	248.91	265.99	272.21
VACE-1.3B	334.51	319.79	258.23
DiffusionHarm.	290.72	312.75	354.40
AW-Harmony	**220.78**	**243.18**	**130.91**

Table 8: Per-weather FVD (↓) on the Specified-Weather Joint Composition Track. Each column contains 50 cases.

Human evaluation. Three independent raters view each complete 21-frame source–result pair with method identities hidden and output order randomized and balanced. They assign 0–4 scores to four criteria in each of three axes: target-weather realization, source-content preservation, and technical quality. For each axis, ratings are averaged within rain, snow, and fog and then macro-averaged with equal weight across the three source-weather categories. Let the resulting axis scores be $(q_{\mathrm{W}}, q_{\mathrm{P}}, q_{\mathrm{T}})$. We report $25(0.55q_{\mathrm{W}} + 0.30q_{\mathrm{P}} + 0.15q_{\mathrm{T}})$ as Human Score. The macro-averaged axis triples are $(2.97, 1.81, 3.25)$ for Cosmos Transfer2.5-2B, $(1.26, 1.53, 2.44)$ for FRESCO, and $(3.78, 3.89, 3.31)$ for AW-Weather, yielding Human Scores of 66.6, 38.0, and 93.6, respectively. VACE-1.3B yields a Human Score of 81.2 under the same protocol.

B Data and metric details

B.1 Coverage and masks

The Paired Self-Insertion Track evaluation set contains 622 21-frame sequences. Its weather distribution is fog 91, heavy rain 105, heavy snow 89, mist 102, rain 94, rainstorm 74, and snow 67. The harmonization training set contains 26,600 clips with no source-scene overlap with paired evaluation. Its weather distribution is fog 3,115, heavy rain 3,142, heavy snow 8,685, mist 3,295, rain 3,704, rainstorm 3,314, snow 1,338, and dust 7; its object-area bins are huge 8,778, large 8,543, medium 7,192, and small 2,087.

The Open-Set Asset Insertion Track evaluation set contains 475 clips: 159 heavy-rain, 169 fog, and 147 heavy-snow.

B.2 Joint-composition and round-trip protocols

Specified-Weather Joint Composition Track. We draw 50 clear-weather sequences from two internal clear-driving collections. Each sequence contributes one suitably sized vehicle and one spatiotemporal mask; no sequence is repeated for multiple objects. AW-Weather converts the complete clear sequence to each of three targets: heavy rain, heavy snow, and fog. We then copy the selected vehicle from the original clear sequence into the same location in each transformed sequence. This creates 150 composites with a weather-mismatched foreground and a paired target: the corresponding complete AW-Weather transformation before the clear vehicle was pasted. The six harmonizers receive identical composites and masks. Metrics use the same temporal union-mask crop and definitions as the Paired Self-Insertion Track. Weather Consistency compares the harmonized vehicle with its surrounding transformed background. All six methods complete all 150 cases without evaluation failures. Figure 11 provides qualitative comparisons for rain and snow.

Harmonization-first round trip. This evaluation reuses the track's 150 AW-Harmony outputs as common adverse-weather inputs. Each weather editor receives the full harmonized video and a clear target condition. The original clear source sequence is the paired full-frame reference. We compute PSNR, SSIM, LPIPS, FID, FVD, and CLIP-GT over all frames and report Clear Similarity from the frozen weather evaluator over full frames. All four weather editors have complete outputs for all 150 cases, and Table 6 reports the final four-method comparison.



Figure 11: Qualitative comparison on the Specified-Weather Joint Composition Track under rain and snow, with three frames per example. After AW-Weather transforms a clear source video, its original vehicle is reinserted into the transformed scene. Rows show the mask, composite input, clear source, generated weather target, AW-Harmony, and selected baselines.

B.3 METRIC DEFINITIONS AND PAIRED CROP PROTOCOL

All CLIP metrics use the pretrained CLIP ViT-L/14 image and text encoders (Radford et al., 2021). Strict-mask CLIP-GT applies the same object mask to each output and paired ground-truth frame, fills pixels outside the mask with the CLIP mean color, and averages image-embedding cosine similarity over valid frames. CLIP-S compares the output with its source image. CLIP-D compares the source-to-output visual direction with the clear-to-target-weather text direction. VBench Quality is the aggregate of Temporal Flickering, Motion Smoothness, Imaging Quality, and Aesthetic Quality (Huang et al., 2024).

Paired pixel and distribution metrics use a temporal union-mask box, mask threshold 0.10, 10% padding, zero fill outside the object mask, and a fixed 832×480 evaluation canvas. Weather Consistency uses per-frame mask-aware object and background crops at 224×224, filling mask-excluded pixels with the mean of the retained region. For the reference background, we center a square neighborhood on the object-mask bounding box and try side lengths of 1.8, 2.2, 2.8, and 3.4 times the box's longer side, selecting the first candidate with at least 35% background pixels. If none qualifies, we use the candidate or full frame with the highest background fraction. Every method uses the same crop and reference semantics.

B.4 PAIRED SELF-INSERTION TRACK QUALITATIVE COMPARISON

Figure 12 shows paired examples in snow, fog, and rain. Figure 13 provides the complete baseline comparison omitted from the compact teaser. The heavy-rain example tests vehicle appearance under wet roads and reduced visibility, while the dense-fog example tests foreground attenuation and structural preservation across frames.



Figure 12: Paired self-insertion examples in snow, fog, and rain. Each example shows three temporal frames of the discordant input, AW-Harmony output, and paired ground truth. Red boxes enlarge selected asset regions.

B.5 WEATHER CONSISTENCY TRAINING AND VALIDATION

Given an object crop o and reference-background crop b, the evaluator returns unit-normalized embeddings z_o, z_b and predicted raw-7D intensity vectors $\widehat{w}_o, \widehat{w}_b$. We define

$$d_z = \mathrm{dist}(z_o, z_b)/\kappa_z, \qquad d_w = \mathrm{dist}(\widehat{w}_o, \widehat{w}_b)/\kappa_w, \tag{9}$$

$$S_{\mathrm{weather}} = \lambda_z \exp(-d_z/\tau_z) + \lambda_w \exp(-d_w/\tau_w), \tag{10}$$

where dist is Euclidean distance. The predicted same-weather rate is the fraction of pairs satisfying $d_z \leq \tau_{\mathrm{same}}$ and is not classification accuracy. The reference background crop is taken from the corresponding original adverse-weather background and shared by all methods for each sample.

The evaluator comprises a ResNet-50 backbone (He et al., 2016), a 256-dimensional unit-normalized embedding head, and a sigmoid-bounded raw-7D weather-intensity regression head. For a training pair (a, b), its objective is

$$\mathcal{L}_{\mathrm{WC}} = \mathcal{L}^{\mathrm{cat}}_{\mathrm{SupCon}}(\{z_i, c_i\}) + \frac{1}{2}[\mathrm{SmoothL1}(\widehat{w}_a, w_a) + \mathrm{SmoothL1}(\widehat{w}_b, w_b)]\,, \tag{11}$$

where c_i is the weather category. The category-supervised contrastive term treats samples of the same weather category in a batch as positives and the remaining categories as negatives, making z

Validation/calibration statistic	Value
Pair predictions	8,192
Same-weather embedding distance	0.180
Different-weather embedding distance	0.350
Distance gap	0.170
Raw-7D regression MAE	0.116
Best-threshold accuracy	79.4%
Best-threshold balanced accuracy	75.8%
Distance threshold fit on these pairs	0.341

Table 9: Validation and threshold calibration of the Weather Consistency evaluator. Threshold performance is measured on the same calibration set used to select the threshold and is not an independent test result.

	Weather Consistency				Region CLIP		Full-frame quality
Context	WC Score ↑	Emb. Dist. ↓	Same-W. (%) ↑	Int. Dist. ↓	CLIP-S ↑	CLIP-D ↑	VBench ↑
Global	0.469	0.270	68.7%	0.276	0.9067	0.018	0.7734
+ Lower-frame	0.481	0.262	70.5%	0.269	0.9084	0.025	0.7741
+ Local	**0.492**	**0.258**	**72.4%**	**0.261**	**0.9085**	**0.027**	**0.7752**

Table 10: Multi-scale context ablation of the Weather Context Adapter on the complete test set of the Open-Set Asset Insertion Track.

a category-oriented weather-similarity representation; the regression term retains raw-7D intensity differences. Both terms have unit weight, the contrastive temperature is 0.10, and the final checkpoint uses no additional alignment loss. We train for 12,000 steps with batch size 64 and learning rate 10^{-4}, including 1,000 warm-up steps.

The evaluator training and calibration indexes are mutually scene-disjoint and are also fully scene-disjoint from the harmonization training set and the evaluation sets of both tracks; they contain no generated weather, harmonization inputs, or outputs from any compared method. The normalized distances use $(\kappa_z, \kappa_w) = (2, \sqrt{7})$, and the Weather Score uses $(\lambda_z, \lambda_w) = (0.65, 0.35)$ and $(\tau_z, \tau_w) = (0.35, 0.30)$. The calibration index contains 1,430 records grouped by real scene before 8,192 pairs are constructed deterministically. A record may contribute to multiple pairs, so these pairs are not 8,192 independent real samples. The same-weather threshold $\tau_{\text{same}} = 0.3412$ maximizes pair accuracy; all weights and thresholds are then frozen before harmonizer comparison. Because threshold selection and the reported accuracy use the same pairs, the values in Table 9 quantify calibration fit rather than independent generalization. We have not calibrated the evaluator against human local-weather judgments or tested ranking sensitivity to its weights, thresholds, or simple photometric perturbations; consequently, we treat its outputs as descriptive metrics.

B.6 Detailed ablation results

All three ablations below are evaluated on the same complete test set of the Open-Set Asset Insertion Track used in Table 7. Within the multiscale adapter variants, the total number of sampled context tokens is held constant across Global, + Lower-frame, and + Local; only their allocation among the active pools changes.

In the supervision ablation, Color-only is trained only on color-augmented data, whereas AW-Harmony uses paired supervision constructed by AW-Weather (Table 11).

B.7 Downstream 2D object detection

We assess whether the compositional editor produces useful training data using Faster R-CNN (Ren et al., 2015) with a ResNet-50-FPN backbone on nuScenes front-camera keyframes (Caesar et al., 2020). The evaluation protocol uses scene-disjoint training and validation data, ten detection classes, and 2D boxes projected from the official 3D annotations. Starting from a common real-data checkpoint, five variants receive the same number of fine-tuning updates: real data only (R), weather augmentation (W), direct asset pasting (P), harmonized asset insertion (A), and joint asset inser-

	Weather Consistency				Region CLIP		Full-frame quality
Training supervision	WC Score ↑	Emb. Dist. ↓	Same-W. (%) ↑	Int. Dist. ↓	CLIP-S ↑	CLIP-D ↑	VBench ↑
Color-only	0.427	0.3055	57.5%	0.3019	0.9009	0.011	0.7516
AW-Harmony	**0.492**	**0.2578**	**72.4%**	**0.2611**	**0.9085**	**0.027**	**0.7752**

Table 11: Full comparison of our weather-transformed paired supervision and color-only training on the complete test set of the Open-Set Asset Insertion Track.

	Weather Consistency				Region CLIP		Full-frame quality
Training objective	WC Score ↑	Emb. Dist. ↓	Same-W. (%) ↑	Int. Dist. ↓	CLIP-S ↑	CLIP-D ↑	VBench ↑
w/o mask-aware loss	0.483	0.2637	70.8%	0.2674	**0.9138**	0.021	0.7743
w/ mask-aware loss	**0.492**	**0.2578**	**72.4%**	**0.2611**	0.9085	**0.027**	**0.7752**

Table 12: Mask-aware loss ablation on the Open-Set Asset Insertion Track.

	Training data	Clear AP ↑	Rain AP ↑
R_0	Real only (initial checkpoint)	26.79	22.27
R	Real only	26.85	22.23
W	Real + weather augmentation	26.76	24.08
P	Real + direct asset pasting	26.81	22.14
A	Real + harmonized assets	27.22	23.07
WA	Real + assets and weather augmentation	**27.23**	**24.95**

Table 13: Downstream detection on real nuScenes images. R_0: initial 12-epoch, single-seed baseline. R–WA: equal-update fine-tuning results.

tion and weather augmentation (WA). Augmented variants mix real and generated images at a 1:1 ratio; P and A share assets and placements. Evaluation uses unmodified real images and reports COCO AP@[0.50:0.95] on the Clear and Rain subsets. Weather subsets follow scene descriptions; Clear denotes non-rainy scenes and includes night scenes. The A–P comparison isolates local harmonization, W–R isolates global weather editing, and WA–R measures the complete asset–weather composition pipeline. In Table 13, W and WA improve Rain AP over R by 1.85 and 2.72 points, respectively. Harmonized insertion (A) improves over direct pasting (P) by 0.41 Clear AP and 0.93 Rain AP. Appendix B.7 specifies the training protocol.

Optimization and data isolation. The training protocol initializes the detector backbone from ImageNet and trains the real-data baseline for 12 epochs. Each variant, including the real-only control R, then receives six real-data-equivalent epochs of fine-tuning with SGD (batch size 16, learning rate 0.002, momentum 0.9, weight decay 10^{-4}, and 500 warm-up iterations); the baseline learning rate is 0.02. All variants share preprocessing, resolution, and optimization budgets. Development scenes are drawn only from the training split. Validation scenes are excluded from detector training, generator fine-tuning, and asset sourcing. Only frames aligned with annotated keyframes are used; image transformations are applied to their boxes. Weather edits retain annotations after geometry checks, and inserted assets receive additional boxes under the same annotation convention. P and A use identical assets, placements, and sampling. AP averages IoU thresholds from 0.50 to 0.95. W tests the global weather module, A tests local asset adaptation, and WA tests their composition. The initial baseline (R_0) precedes this additional fine-tuning and is shown separately.

B.8 Internal deployment evaluation

Section 4.8 presents an internal deployment case study in which generated rare-obstacle videos are added to OCC training and performance is assessed on a non-public AEB evaluation set. Both groups use the same asset-insertion videos for OCC training: the control uses the unharmonized versions, while the experimental group uses their AW-Harmony counterparts. The reported recall values compare these two training-data treatments. Public disclosure is limited to the aggregate outcomes and the representative case in Figure 14; detailed OCC training and AEB evaluation protocols, production data, model checkpoints, and sample-level records remain confidential.



Figure 13: Complete qualitative comparison on the Paired Self-Insertion Track under heavy rain and dense fog. Each example shows the mask and four temporal frames, comparing the input, paired ground truth, AW-Harmony, and all baselines.

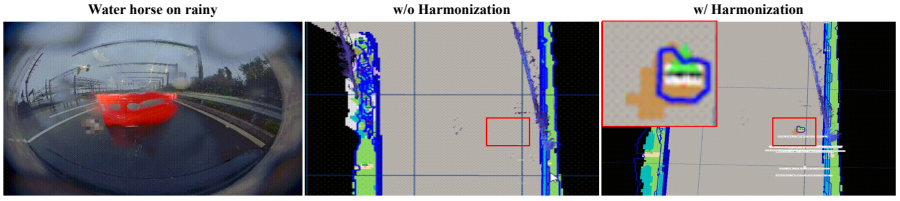

Figure 14: OCC outputs for a rainy water-barrier case from models trained on unharmonized and harmonized versions of the same asset-insertion data: a miss becomes a detection.

Weather-transformation VLM prompt (concise presentation)
You are a blinded visual evaluator. Given five temporally ordered SOURCE–RESULT frame pairs, you do not know the method identity or transformation target. Judge only the visible evidence as follows:
Weather evidence: independently record cloud cover, precipitation, road wetness or accumulated snow, haze, and visibility in SOURCE and RESULT; do not judge whether the transformation succeeded.
Content preservation: assess the consistency of vehicles, people, traffic controls, road geometry, and background structure; do not count weather or illumination changes as content changes.
Technical quality: assess single-frame visual quality, generation artifacts, cross-frame consistency, and motion plausibility; do not count weather or visibility changes themselves as technical defects.
Rate each visible criterion from 0 to 4, briefly describe the observed evidence, and return the scores.

Figure 15: Concise VLM prompt for the weather-transformation component. The evaluation covers visible weather evidence, source-content preservation, and technical quality.

C VLM PERCEPTUAL-EVALUATION PROTOCOL AND PROMPT

C.1 WEATHER-TRANSFORMATION COMPONENT EVALUATION

For weather-transformation evaluation, gpt-5.6-sol receives five temporally ordered source–result frame pairs. Without access to the method identity or transformation target, it records visible weather evidence and separately assesses source-content preservation and technical quality. The scores in Table 2 are aggregated from these subratings according to the clear-weather target and fixed scoring rules. Figure 15 gives the core instruction.

C.2 HARMONIZATION EVALUATION

Table 14 reports the VLM and human preference results on the Open-Set Asset Insertion Track.

(a) VLM pairwise comparison		**(b) Human six-way selection**		
Comparator	AW-Harmony win rate (%) ↑	Method	Top-1 share (%) ↑	95% CI
VTT	76.1%	VTT	12.9%	[10.4–15.4]
Harmonizer	72.3%	Harmonizer	9.6%	[7.6–11.7]
LBM	78.0%	LBM	16.6%	[14.1–19.2]
VACE-1.3B	94.4%	VACE-1.3B	2.8%	[1.5–4.3]
DiffusionHarm.	91.5%	DiffusionHarm.	8.9%	[6.6–11.5]
Mean	**82.5%**	**AW-Harmony**	**49.2%**	**[45.8–52.6]**

Table 14: VLM and human perceptual preference on the Open-Set Asset Insertion Track.

VLM protocol. We use gpt-5.6-sol for blinded pairwise comparisons between AW-Harmony and each baseline generated from the same composite and mask. Method identity is hidden, and A/B assignment is randomized and balanced. Each anonymous method pair is evaluated at five temporally ordered time points. At each time point, the evaluator receives separate full-frame candidates A and B with the inserted asset marked by the same pure-red box; the method-to-label assignment remains fixed within a request. The evaluator makes a forced choice for every frame, and the sequence-level result is the five-frame majority. Figure 16 shows an English rendering of the fixed Chinese prompt. Sample-specific image paths, frame indices, and tool-routing text are replaced by placeholders; the decision criteria and output schema are unchanged.

Human protocol. The human study uses six-way Top-1 selection with 25 participants, each evaluating 30 or 31 clips sampled from the Open-Set Asset Insertion Track. Each sampled clip receives judgments from five distinct participants, and candidate positions are balanced across methods. We obtain 95% confidence intervals by bootstrapping sampled clips.

VLM evaluation prompt (English translation)

You are an independent, blinded visual evaluator. Inspect only the images explicitly provided in this request. Do not access method identities or infer a method from file names.

Five time points from the same sequence are provided in temporal order. Each time point contains image A and image B: two full-frame model outputs for the same time, scene, and inserted asset, without local cropping. The mapping between models and the A/B labels remains fixed across all five time points.

Time t: image A $[A_t]$; image B $[B_t]$, for $t = 1, \ldots, 5$.

The pure-red box in each image locates the same inserted asset to be compared. At each time point, judge only whether the adverse-weather appearance of the asset inside the box is harmonious with the immediately surrounding background outside the box.

Do not evaluate regions outside the target neighborhood, the plausibility of the asset category, or general image quality. Do not evaluate geometric distortion, deformation, or texture detail; none of these factors may determine a frame-level or final choice.

Choose A or B at every time point; ties are not allowed. `FINAL_WINNER` must be the majority of the five frame-level choices. `EVIDENCE` may describe only directly observable, cross-frame weather-consistency evidence and must not invoke any excluded factor.

Return exactly the following seven lines and no other text:

```
FRAME_1:  A or B
FRAME_2:  A or B
FRAME_3:  A or B
FRAME_4:  A or B
FRAME_5:  A or B
FINAL_WINNER: A or B
EVIDENCE: one compliant sentence describing cross-frame weather
consistency
```

Figure 16: Pairwise VLM prompt for the Open-Set Asset Insertion Track. Candidate order is randomized and method identities are hidden.