ONEFIXER: HIGH-QUALITY AND CONSISTENT ONE-STEP AUTOREGRESSIVE 3DGS REFINEMENT FOR DRIVING SCENES

Anonymous authors
Paper under double-blind review

ABSTRACT

Autoregressive video diffusion is a promising render-time fixer for 3D Gaussian Splatting (3DGS) in autonomous-driving simulation, but deployment demands high visual quality and temporal consistency at low latency. This is particularly challenging for one-step causal generation, where each imperfect prediction immediately becomes context for subsequent frames. Existing approaches improve test-time rollouts through staged training with multiple trained modules and rollout-aware regularization, yet one-step refinement quality can still fall short of what render-time deployment requires. We introduce **OneFixer**, a **one-step** autoregressive video-diffusion fixer trained in **one task-specific adaptation stage**. Our key idea is a *deployment-matched shared rollout*: the model's own one-step predictions serve as the causal context for flow matching, exposing training to deployment-time errors, while the same rollout receives direct pixel-space perceptual supervision to preserve fine detail. Because the predictions optimized for current-frame quality are exactly those reused as future context, fidelity and autoregressive robustness are learned jointly, without bidirectional-to-causal conversion or teacher–student distillation. Targeting driving simulation, OneFixer further leverages cues that setting readily provides—lane geometry and dynamic-agent states—to improve geometric fidelity. Across Waymo and proprietary driving scenes with 900-frame autoregressive rollouts, OneFixer achieves the lowest FVD, LPIPS, and DISTS among all baselines under one-step inference, with strong geometric fidelity and temporal consistency matching or exceeding multi-stage DMD pipelines. Under identical backbone and conditioning, it matches the final quality of a multi-stage DMD-with-Self-Forcing pipeline in under half the GPU-hours and continues to improve beyond its plateau. In closed-loop simulation with a driving policy, OneFixer reduces the collision rate by a third relative to the raw 3DGS rendering. Project page: `https://onefixer-web.vercel.app/`.

1 INTRODUCTION

3D Gaussian Splatting (3DGS) offers autonomous-driving simulation an explicit, controllable scene representation with fast rasterized rendering (Kerbl et al., 2023; Zhou et al., 2024), but novel-view quality degrades severely in driving scenes, where ego-centric capture along a single trajectory leaves regions sparsely observed and dynamic objects introduce occlusions and moving geometry (Zhou et al., 2025; Li et al., 2025; Sun et al., 2025). Diffusion-based fixers restore realistic appearance from such imperfect renderings using learned priors (Wu et al., 2025; 2026), and among them autoregressive (AR) video diffusion (de Lutio et al., 2026; NVIDIA et al., 2026; Zhou et al., 2026) is the natural fit for an online render-time fixer: it is more temporally consistent than image-based fixers (Wu et al., 2025; Zhang et al., 2026c; Wei et al., 2025) and, unlike bidirectional video models (Liu et al., 2024b), streams without a bounded rollout horizon.

Achieving both perceptual quality and temporal consistency in the low-latency, one- or few-step regime of AR generation is, however, challenging. Reducing the denoising budget degrades each prediction, and because that prediction becomes the causal context for the next chunk, errors compound: the history drifts away from the ground-truth histories seen during training, an exposure bias well documented in AR video diffusion (Chen et al., 2024; Yin et al., 2025; Huang et al., 2025; Li

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.

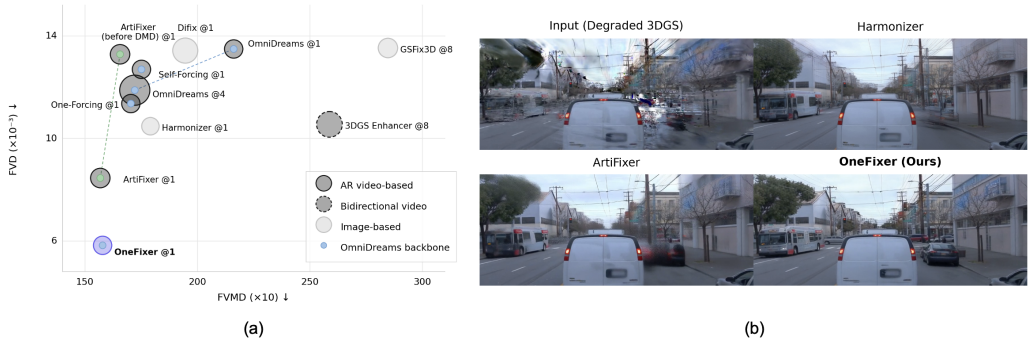

Figure 1: **Quality–consistency–latency trade-off of 3DGS refinement.** (a) Methods are positioned by video fidelity (I3D FVD, ↓) and temporal consistency (FVMD, ↓), with bubble area proportional to inference latency; lower-left and smaller are better. Dashed lines connect a method across its step-reduction or distillation path (e.g., OmniDreams @4 → @1). (b) Qualitative comparison on a degraded Waymo 3DGS rendering; OneFixer recovers finer foliage and a more faithful vehicle under one-step causal refinement.

et al., 2026; Feng et al., 2026). Fig. 1(a) illustrates this on causal OmniDreams: moving from four steps to one substantially lowers latency but worsens both FVD and FVMD, even though every step is anchored by a 3DGS rendering, which limits drift only partially.

Prior AR 3DGS fixers address these two failures with separate machinery. Against drift, they adopt multi-stage training that combines DMD (Yin et al., 2024) with Self-Forcing (Huang et al., 2025), as in ArtiFixer and the Xiaomi EV World Model (de Lutio et al., 2026; Zhou et al., 2026); these recipes are effective but require a bidirectional teacher and carry their own design burden (Zhao et al., 2026; You et al., 2026; Luo et al., 2026). Against the quality loss of one-step prediction, they append critics or reward-based post-training (Feng et al., 2026; Zhang et al., 2026a;b; Wu et al., 2026), each adding a model or a stage. Notably, all of these recipes were designed for generic text- or image-to-video generation from pure noise, and are transferred to 3DGS refinement without exploiting the structure the task provides. This raises the central question of this work: *when the AR task is 3DGS refinement, can we obtain high-quality, temporally consistent one-step causal generation with a simpler, single-stage adaptation procedure?*

We introduce OneFixer, which fine-tunes a pretrained causal video generator (NVIDIA et al., 2026) into the deployment-time one-step 3DGS fixer in a single stage, with no bidirectional-to-causal conversion, teacher–student distillation, critic, or separate post-training: the generator rolls out its own one-step predictions exactly as at deployment, and this rollout serves as the causal history for flow matching and as the source of predictions for direct pixel-space supervision. We observe that 3DGS refinement needs no proxy for this rollout. Because the flow source is the opacity-mixed 3DGS rendering (de Lutio et al., 2026) rather than noise, a single flow step already yields a plausible, though imperfect, prediction, and because deployment is also one step, the training rollout *is* the sampler rather than an approximation of it, unlike multi-step denoising trajectories or re-noised ground truth (Huang et al., 2025; Guo et al., 2026). Flow matching thus conditions on the model's true deployment-time errors, and pixel supervision reaches decoded pixels through a single forward pass, with no truncated backpropagation through a denoising chain (Wang et al., 2026b; Prabhudesai et al., 2023).

Both objectives are optimized *jointly* on this rollout: flow matching enforces consistency in latent space at every position, and the pixel loss recovers detail and flows back through the history that produced it, so preceding predictions are trained as context. A sequential recipe provides neither coupling (Zhang et al., 2026b; He et al., 2026).

OneFixer also exploits cues that driving simulation already exposes: the opacity-derived reliability of the rendering and, where present, lane geometry and dynamic-agent states (Yan et al., 2024; Li et al., 2025; Luo et al., 2023). These matter beyond appearance, since a displaced lane or hallucinated vehicle is an incorrect observation for the policy under test.

We evaluate OneFixer on Waymo (Sun et al., 2020) and proprietary driving scenes with 900-frame autoregressive rollouts ($9.7\times$ the training horizon). OneFixer achieves the lowest FVD, LPIPS, and DISTS at one step with temporal consistency matching multi-stage DMD pipelines; a same-backbone controlled study attributes the gain to shared-rollout training, and closed-loop simulation on 60 scenes shows that the refined observations yield fewer collisions and off-road events for a driving policy.

2 Related Work

2.1 Diffusion Models for Render-Time 3DGS Enhancement

Diffusion priors complement explicit 3DGS reconstruction either as supervision for the scene representation (Wu et al., 2025; Liu et al., 2024b) or as a render-time fixer that refines rendered observations directly (de Lutio et al., 2026; Wu et al., 2026); we focus on the latter. Image-based fixers such as DiFix3D+ (Wu et al., 2025) and GSFix3D (Wei et al., 2025) apply strong image priors to individual renderings and achieve high frame-level quality, and recent work extends them with short temporal conditioning for online use (Zhang et al., 2026c); without a video prior, however, they cannot robustly recognize transient interference such as a Gaussian floater popping into view and tend to refine it as scene content (Fig. 4). Bidirectional video diffusion offers a stronger temporal prior, as in 3DGS-Enhancer (Liu et al., 2024b), but operates on fixed clips well short of minute-scale driving simulation and cannot stream. Driving-oriented fixers therefore adopt causal autoregressive models: ArtiFixer (de Lutio et al., 2026), OmniDreams (NVIDIA et al., 2026), and the Xiaomi EV World Model (Zhou et al., 2026) obtain efficient generators through staged causal adaptation and distillation, and RealityBridge (Wu et al., 2026) follows a supervised curriculum with reward-guided post-training. Nevertheless, none establishes quality-preserving one-step deployment, relying instead on few-step sampling and multi-stage training.

2.2 Autoregressive Video Diffusion and Exposure Bias

Causal video generation suffers an exposure bias: training histories come from clean data while inference conditions on the model's own predictions (Chen et al., 2024; Yin et al., 2025; Huang et al., 2025; Li et al., 2026). The prevailing remedy distills a pretrained bidirectional teacher into a causal few-step student with DMD-style objectives and then exposes the student to its own rollouts (Yin et al., 2025; Huang et al., 2025; Zhu et al., 2026; Zhao et al., 2026; Zheng et al., 2026). This recipe presupposes a bidirectional teacher and is sensitive to student initialization (Zhu et al., 2026; Zhao et al., 2026) and distillation regularization (You et al., 2026); even then one-step quality lags, prompting auxiliary critics or adversarial post-training (Feng et al., 2026; Wang et al., 2026a), whose min–max objective is itself known to be unstable (Wang et al., 2026a), or reward-based follow-up stages (Zhang et al., 2026a;b; He et al., 2026), each adding a model or a stage, and preserving long-rollout capability through these stages is not guaranteed (Zhang et al., 2026b; He et al., 2026). Closest to our motivation, self-resampling (Guo et al., 2026) trains an autoregressive model without a teacher by generating its own causal context, but that context is a proxy: ground truth is noised and re-denoised in one step, so leaked ground-truth content masks the model's true errors and the training rollout differs from the multi-step sampler used at inference. It also provides no signal on output pixel quality. OneFixer instead rolls out the exact one-step deployment generator, so the history carries the model's true errors, and adds paired pixel-space supervision on the same rollout.

2.3 Direct Supervision for One-Step Fidelity

Recovering perceptual quality at very low step counts typically requires supervision beyond the diffusion objective. Image models pioneered this by backpropagating pixel- or reward-space losses through a truncated denoising chain (Clark et al., 2024; Prabhudesai et al., 2023) or a single step (Parmar et al., 2024; Wu et al., 2024; Zhang et al., 2024; Kim et al., 2026). Video models follow the same principle in the bidirectional setting: DOVE (Chen et al., 2025) and OASIS (Guo et al., 2025) pair pixel-space supervision with flow-based temporal losses, since frame-wise supervision alone does not enforce consistency, SeedVR2 (Wang et al., 2026a) uses feature matching and adversarial post-training, and ChopGrad (Rivkin et al., 2026) makes latent-space pixel losses tractable through truncated decoder backpropagation, including for 3DGS refinement. OASIS additionally starts its



Figure 2: **Shared-rollout training for one-step 3DGS refinement.** From opacity-mixed 3DGS inputs and scene conditioning, the deployment-time one-step causal generator rolls out its own predictions $\{\hat{\mathbf{z}}_0^1, \ldots, \hat{\mathbf{z}}_0^N\}$ exactly as at inference. The rollout serves two objectives: self-rollout flow matching supervises every position in latent space conditioned on the self-generated history (detached, sg), and a frame-wise pixel-space perceptual loss on one sampled prediction back-propagates through the last k latent positions of that history to recover fine detail.

single step from the degraded latent rather than noise, a structured source akin to ours, but in a clip-level model. Streaming one-step restoration has also been reached, but through multi-stage distillation from a bidirectional teacher (Zhuang et al., 2025). None of these couples temporal consistency and paired pixel fidelity in a single stage for autoregressive generation: the bidirectional methods have no causal history to reuse, while causal pipelines train the reused prediction through distribution matching or distillation rather than against its clean reference (Huang et al., 2025; Zhuang et al., 2025).

3 Method

Fig. 2 summarizes OneFixer: starting from the pre-DMD causal OmniDreams checkpoint (NVIDIA et al., 2026), we train two objectives on one deployment-matched rollout, self-rollout flow matching (Sec. 3.2) and frame-wise perceptual supervision (Sec. 3.3); Fig. 3 previews their complementary effects.

3.1 One-Step Causal Refinement from a 3DGS Source

At generation position i, the model is conditioned on the causal history $\mathbf{h}^{<i}$ of preceding latents and a map-derived scene condition $\mathbf{c}_m^i$ encoding lane geometry and dynamic actors, following OmniDreams (NVIDIA et al., 2026). Let $\mathbf{z}_r^i$ and $\mathbf{z}_0^i$ denote the latents of the 3DGS render and the ground-truth frame, and $\mathbf{m}^i \in [0,1]$ the opacity-derived reliability mask (de Lutio et al., 2026). Rather than pure noise $\boldsymbol{\epsilon}^i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, the flow source is

$$\mathbf{z}_1^i = \mathbf{m}^i \odot \mathbf{z}_r^i + (1 - \mathbf{m}^i) \odot \boldsymbol{\epsilon}^i, \tag{1}$$

which retains rendered content where the reconstruction is reliable and leaves generative freedom where it is not. The backbone is a causal flow-matching model: along the interpolant $\mathbf{z}_t^i = (1-t)\,\mathbf{z}_0^i + t\,\mathbf{z}_1^i$, a velocity network $v_\theta(\mathbf{z}_t^i, t \mid \mathbf{h}^{<i}, \mathbf{c}_m^i)$ is trained to predict $\mathbf{z}_1^i - \mathbf{z}_0^i$, and sampling integrates from $t{=}1$ to $t{=}0$. OneFixer performs a single flow step from this source,

$$\hat{\mathbf{z}}_0^i = G_\theta(\mathbf{z}_1^i \mid \mathbf{h}^{<i}, \mathbf{c}_m^i) = \mathbf{z}_1^i - v_\theta(\mathbf{z}_1^i, 1 \mid \mathbf{h}^{<i}, \mathbf{c}_m^i). \tag{2}$$

Because the source already carries scene structure, this single step is a correction rather than a synthesis from noise, which is what makes the self-rollout of Sec. 3.2 viable; replacing the mixed source with pure noise and channel-concatenated 3DGS latents, the alternative explored in ArtiFixer (de Lutio et al., 2026), degrades every metric (V6, Table 2a).

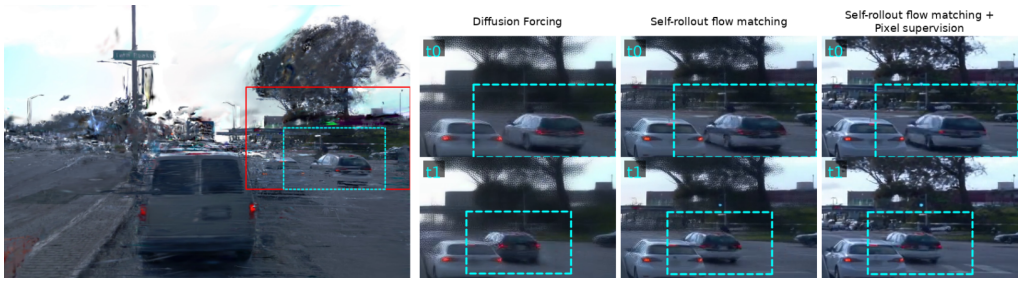

Figure 3: **Complementary effects of self-rollout and perceptual supervision.** Left: a Waymo example and its degraded 3DGS input. Right: one-step autoregressive rollouts and magnified regions. Diffusion Forcing exhibits appearance variation across generated frames; self-rollout flow matching improves temporal consistency, while direct pixel-space perceptual supervision further recovers fine visual details, particularly in the tree branches and foliage.

3.2 Self-Rollout Flow Matching

At deployment, $\mathbf{h}^{<i}$ consists of the model's own preceding one-step predictions $\hat{\mathbf{z}}_0$. We construct the training history the same way, applying Eq. (2) recursively,

$$\hat{\mathbf{z}}_0^i = G_\theta\big(\mathbf{z}_1^i \mid \mathrm{sg}(\hat{\mathbf{z}}_0^{<i}), \mathbf{c}_m^i\big), \qquad i = 1, \ldots, N, \tag{3}$$

with $\mathrm{sg}(\cdot)$ the stop-gradient. For each position we sample $t \sim \mathcal{U}(0,1)$ and minimize

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{i,t}\Big[\big\| v_\theta(\mathbf{z}_t^i, t \mid \mathrm{sg}(\hat{\mathbf{z}}_0^{<i}), \mathbf{c}_m^i) - (\mathbf{z}_1^i - \mathbf{z}_0^i)\big\|_2^2\Big]. \tag{4}$$

The flow objective is unchanged; only its causal context is replaced by the self-generated history, and it supervises every position of the sequence in latent space, which is where its effect on temporal consistency comes from (Fig. 3, middle).

3.3 Frame-Wise Pixel Supervision on the Shared Rollout

Latent flow matching supervises the one-step output only in latent space and never sees the decoded pixels, so it does not directly optimize perceptual quality. We therefore sample a single latent position $w \sim \mathcal{U}\{1, \ldots, N\}$ from the training horizon, causally decode $\hat{\mathbf{z}}_0^w$ to RGB, and minimize

$$\mathcal{L}_{\mathrm{perc}} = d_{\mathrm{LPIPS}}(\hat{\mathbf{x}}^w, \mathbf{x}^w), \tag{5}$$

where, with a slight abuse of notation, $\hat{\mathbf{x}}^w$ and $\mathbf{x}^w$ denote the several RGB frames that the decoder's temporal compression maps to latent position w and their paired clean references; the loss is averaged over them. Because each chunk is a single forward pass, this loss reaches the generator with no backpropagation through a denoising chain; the only truncation is across autoregressive history and the causal Wan decoder, whose temporal cache is detached in the manner of ChopGrad (Rivkin et al., 2026). On the autoregressive side, gradients are retained through the last k latent positions ending at w and earlier history is detached. The truncation bounds memory (App. B.2), but the retained path is also a second channel for consistency: the loss on $\hat{\mathbf{x}}^w$ reaches the k preceding predictions through their role as w's history, so they are trained to be good *context*, not only good frames. The ablation confirms that each channel improves FVMD on its own and that the full model needs both (Sec. 4.3). The full objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{FM}} + \lambda_{\mathrm{perc}}\mathcal{L}_{\mathrm{perc}}, \tag{6}$$

with both terms evaluated on the same shared rollout.

4 Results

4.1 Experimental Setup

Training data and setup. We train OneFixer separately on Waymo (Sun et al., 2020) and our proprietary Seoul driving dataset, using 350 and 400 training scenes, respectively. Following prior

3DGS enhancement pipelines (Wu et al., 2025; de Lutio et al., 2026; Wu et al., 2026), we construct paired degraded-render/ground-truth samples using cycle-rendered views and under-trained 3DGS reconstructions. For both domains, we initialize from the official OmniDreams causal checkpoint released before DMD distillation (NVIDIA et al., 2026). Training uses 544×960 resolution, an effective batch size of 64, and 93-frame causal sequences. On Seoul, these sequences are sampled from 300-frame training source clips. We use AdamW with a learning rate of 3×10^{-5} and $\lambda_{\text{perc}} = 10$, with 400 optimization steps on Waymo and 600 on Seoul. Appendix A provides data-generation and training details.

Evaluation protocol. We evaluate on 80 held-out Waymo scenes using 198-frame autoregressive rollouts and 32 held-out Seoul scenes using 900-frame rollouts. These correspond to approximately $2.1\times$ and $9.7\times$ the 93-frame causal training horizon. We report complementary perceptual, geometric, and temporal metrics. Definitions, aggregation rules, and evaluation windows are provided in Appendix C.1.

Baselines and implementation. We compare with image-based fixers DiFix3D+ (Wu et al., 2025), DiffusionHarmonizer (Zhang et al., 2026c), and GSFix3D (Wei et al., 2025); the bidirectional video model 3DGS-Enhancer (Liu et al., 2024b); and causal video fixers ArtiFixer (de Lutio et al., 2026) and OmniDreams (NVIDIA et al., 2026). Each method uses its officially released checkpoint or recommended pretrained initialization for task-specific adaptation. For DiffusionHarmonizer we report both its frame-wise mode and a temporal mode; because the official teacher-forced temporal recipe drifted on our long rollouts, the temporal variant is trained with self-generated context (App. C.2). We distinguish three baselines using the OmniDreams backbone. *OmniDreams* starts from the released DMD-distilled causal checkpoint and undergoes task-specific diffusion-forcing fine-tuning, following the fixer adaptation process in the original paper. *Self-Forcing* follows the staged distillation recipe of ArtiFixer (de Lutio et al., 2026)[1]: the bidirectional teacher is fine-tuned from the official pre-DMD bidirectional checkpoint to 3DGS refinement (Stage 1); the causal student is fine-tuned from the official pre-DMD causal checkpoint—OneFixer's initialization—to the fixer task (Stage 2); the Stage-2 student is then distilled with DMD and Self-Forcing (Yin et al., 2024; Huang et al., 2025) against the Stage-1 teacher (Stage 3). ArtiFixer converts its Stage-1 model to causal in Stage 2 because Wan2.1 has no causal release; we instead start Stage 2 from the released causal checkpoint, as in OmniDreams' own distillation pipeline, so that the student shares OneFixer's initialization. *One-Forcing* shares Stages 1–2 and adds the One-Forcing critic (Feng et al., 2026) in Stage 3. Details are in Appendix C.2.

4.2 Comparison with 3DGS Enhancement Methods

Overall comparison. Table 1 shows that OneFixer achieves the strongest one-step balance across perceptual, geometric, and temporal metrics: it improves temporal consistency over image-based fixers while retaining their frame-level fidelity, and improves perceptual quality over bidirectional and causal video fixers while matching or exceeding them on temporal metrics. Harmonizer's temporal variant, trained with a warping loss, wins Warp L1 on both datasets but is otherwise no better than its frame-wise counterpart, so short-range smoothness alone does not yield video-level consistency. On Waymo, the same-backbone Self-Forcing and One-Forcing baselines attain the best PSNR and SSIM but trail OneFixer on FVD, LPIPS, DISTS, and FVMD, a perception–distortion trade-off in which distillation favors smoother, more conservative outputs. On Seoul, over 900-frame rollouts, OneFixer leads every metric except lane jitter and Warp L1, and the two distilled baselines land close to OmniDreams @1, the diffusion-forcing fine-tune of the released DMD checkpoint. Fig. 6(b) shows that OneFixer keeps the lowest position-wise FVD throughout the 900-frame rollout rather than gaining its aggregate advantage from a subset of the sequence, and Fig. 4 shows representative cases of improved local fidelity and preserved scene structure under 3DGS artifacts.

Training and inference efficiency. Fig. 6-(a) shows that OneFixer is substantially faster than existing video-based fixers and approaches the latency of image-based enhancement, since a single flow step per causal block yields several RGB frames at once through the chunked causal de-

[1] ArtiFixer's project page describes two phases, but its official implementation trains three stages: supervised fine-tuning, diffusion-forcing adaptation from the stage-1 checkpoint, and DMD initialized from the stage-2 student and the stage-1 critic; we follow the implementation's count.



Figure 4: **Qualitative comparison of 3DGS refinement methods.** OneFixer better preserves realistic appearance and scene structure under challenging reconstruction errors. Its geometry-aware conditioning improves structural fidelity, while the video prior increases robustness to transient artifacts that are difficult to handle independently frame by frame, such as the popping Gaussian floater in the bottom row.

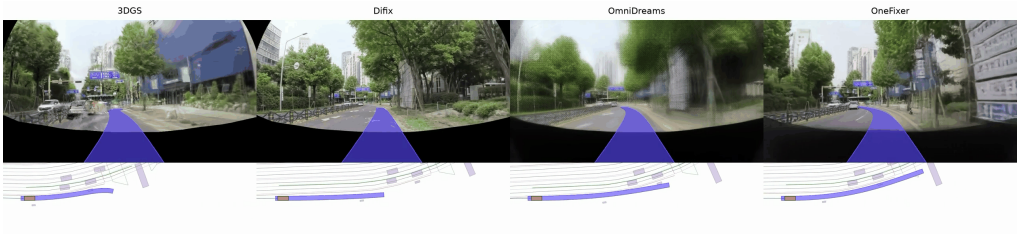

Figure 5: **Closed-loop replay with different renderers.** Top: policy observation and planned path (blue) where the 3DGS reconstruction degrades. Bottom: ego trajectory over the lane map. Raw 3DGS shows ghosted structure that bends the trajectory; DiFix smears the lane markings across the road and the policy leaves it; OmniDreams blurs the scene at one step. OneFixer's faithful observation keeps the policy on route.

coder, amortizing the backbone cost over the block. Fig. 6-(c) compares training trajectories against the same-backbone Self-Forcing and One-Forcing baselines, whose three stages appear as regime changes in the curves. OneFixer falls below both baselines' final FVD after roughly 345 GPU-hours, before either baseline has completed teacher and causal adaptation, and reaches the Self-Forcing FVMD plateau at the same point; it ends below both baselines on both metrics. The DMD stage delivers the expected one-step gain but saturates within its first checkpoint, and a further 300–400 GPU-hours change either metric only marginally; adding a One-Forcing critic (Feng et al., 2026) raises the cost per iteration without closing the gap. The plateau is therefore a ceiling of the distillation recipe rather than of its budget, and since backbone, conditioning, and causal initialization are held fixed, the difference is attributable to shared-rollout training.

Closed-loop replay fidelity. We drive a vision-language-action (VLA) policy in closed loop on 60 proprietary scenes with 5 seeds each, varying only the renderer (Table 2b). OneFixer yields the lowest collision rate, off-road rate, and deviation from the logged route, reducing collisions by a third relative to the raw 3DGS rendering (6.9% vs. 10.6%) and halving the off-road rate with non-overlapping 95% intervals. The comparison is informative about what the policy needs: the image-based DiFix improves the off-road rate but increases collisions relative to raw 3DGS, suggesting that improved frame-level appearance alone does not necessarily translate to policy-relevant consistency. The same-backbone OmniDreams and Self-Forcing stay close to the unrefined baseline. Because the scenes overlap with the policy's training data, these numbers measure replay fidelity rather than policy generalization. Fig. 5 shows a representative case: where the 3DGS rendering degrades,

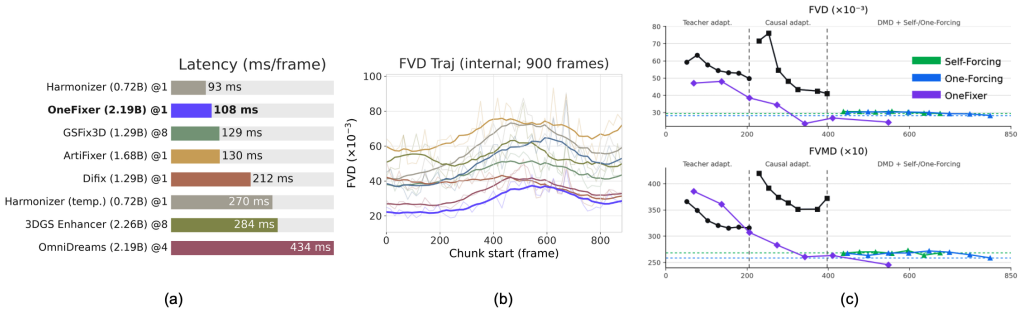

Figure 6: **Inference efficiency, long-horizon stability, and training efficiency.** (a) Per-frame enhancement latency measured on a single H200 GPU and bf16 precision for all models. @N denotes the number of denoising steps, and model parameter counts (in billions) are shown in parentheses. (b) Position-wise FVD over 900-frame autoregressive rollouts on the proprietary dataset. Self-Forcing is shown in steel blue; other colors match (a). (c) Validation FVD and FVMD vs cumulative GPU-hours, 32 Waymo scenes, 93-frame one-step rollouts, same backbone and conditioning. Baselines follow the standard fixer-distillation recipe (de Lutio et al., 2026; Zhou et al., 2026): teacher adaptation, causal fixer adaptation, then DMD with Self-Forcing (green) or One-Forcing (blue).

OneFixer's sharper, structurally faithful observation is the only one under which the policy keeps following the road.

4.3 Ablation and Analysis

Table 2a compares variants V1–V9, all initialized from the same pre-DMD OmniDreams causal checkpoint and evaluated at one step; V9 is the full model, SF denotes self-rollout flow matching, and k is the perceptual backpropagation horizon in latent positions.

Structured source. Replacing the opacity-mixed source with noise and channel-concatenated 3DGS latents (V6) degrades every metric substantially, confirming that the structured source is what makes one-step fine-tuning practical.

Shared rollout: objectives, coupling, and history. Self-rollout flow matching alone (V2) mainly improves temporal consistency over V1 while leaving a perceptual gap; LPIPS alone (V3) improves fidelity but is less consistent; combining both (V9) gives the best balance (Fig. 3). How they are combined matters: sequential training under the same budget (V7) matches V9 on frame-level fidelity but loses on FVD and FVMD, and shortening the perceptual horizon to k=1 (V4) slightly improves LPIPS while worsening both. What the rollout contains matters as well: replacing the model's own predictions with re-denoised ground truth as in self-resampling (Guo et al., 2026) (V8) degrades every metric, most in LPIPS (0.158 vs. 0.117), since leaked clean content lets the model rely on context it will not have at deployment. Together these support optimizing both objectives jointly on the model's own deployment-time rollout, so that preceding predictions are trained for their role as future context.

Geometry conditioning. Removing lane and actor conditioning (V5) lowers Cuboid F1 and modestly worsens perceptual and temporal metrics, yet V5 still beats every one-step baseline in Table 1 on FVD, LPIPS, and Cuboid F1 with competitive FVMD, so the gains do not hinge on conditioning. Robustness to lane and actor conditioning errors is evaluated in Appendix B.1.

5 Conclusion and Limitations

We introduced OneFixer, a one-step autoregressive video-diffusion fixer for 3DGS refinement in driving scenes. Training the deployment-time generator on a single shared rollout, as both history for flow matching and source for perceptual supervision, yields a single-model, single-stage pipeline that delivers strong perceptual quality and temporal consistency far beyond the training horizon at

Table 1: Comparison on Waymo and proprietary Seoul dataset. Denoising steps are denoted by @N. Smaller raised values denote 95% t-intervals over clips; FVD and FVMD are single set-level Fréchet distances and therefore carry no per-clip interval. For ArtiFixer, parentheses give the stage of its official pipeline at which the model is evaluated (pre-DMD: after causal adaptation; DMD: after DMD with Self-Forcing). Harmonizer (temporal) conditions on its four previous outputs and is trained with self-generated context (App. C.2). OmniDreams, Self-Forcing, One-Forcing, and OneFixer share the OmniDreams backbone and differ only in training recipe (Sec. 4.1). Blue, green, and red cells indicate the best, second-best, and worst results within each dataset, respectively.

	Perception					Geometry			Temporal		
Model	FVD ↓	LPIPS ↓	DISTS ↓	PSNR ↑	SSIM ↑	Cuboid F1 ↑	Lane F1 ↑	Lane x-err ↓	Lane jit. ↓	FVMD ↓	Warp L1 ↓
Waymo (80 scenes × 198 frames)											
GSFix3D @8	13.5	0.182±0.015	0.185±0.010	22.9±0.6	0.781±0.017	0.702±0.058	0.574±0.074	0.178±0.029	0.167±0.024	285	270±23
DiFix @1	13.4	0.142±0.015	0.126±0.010	23.6±0.7	0.810±0.015	0.776±0.049	0.708±0.078	0.136±0.021	0.094±0.017	195	123±20
Harmonizer @1	10.5	0.136±0.015	0.124±0.011	24.5±0.7	0.819±0.015	0.776±0.050	0.742±0.076	0.127±0.022	0.093±0.017	179	115±18
Harmonizer (temporal) @1	12.5	0.147±0.015	0.131±0.010	24.1±0.7	0.820±0.014	0.768±0.050	0.712±0.077	0.137±0.025	0.101±0.019	184	101±17
3DGS Enhancer @8	10.5	0.216±0.019	0.211±0.015	21.7±0.6	0.784±0.016	0.609±0.068	0.575±0.075	0.187±0.027	0.099±0.021	259	112±20
ArtiFixer (pre-DMD) @1	13.3	0.170±0.019	0.187±0.012	24.0±0.7	0.805±0.016	0.705±0.056	0.615±0.081	0.157±0.028	0.107±0.024	166	115±21
ArtiFixer (pre-DMD) @4	6.88	0.147±0.017	0.145±0.009	23.4±0.8	0.797±0.017	0.728±0.052	0.636±0.080	0.155±0.031	0.108±0.026	165	116±22
ArtiFixer (DMD) @1	8.46	0.146±0.016	0.158±0.011	24.1±0.7	0.811±0.015	0.744±0.054	0.675±0.079	0.142±0.022	0.104±0.022	157	119±23
OmniDreams @1	13.5	0.184±0.018	0.177±0.012	23.7±0.7	0.829±0.015	0.758±0.055	0.645±0.078	0.134±0.021	0.094±0.019	216	125±21
OmniDreams @4	11.9	0.132±0.015	0.119±0.008	23.6±0.7	0.829±0.016	0.791±0.053	0.691±0.078	0.123±0.020	0.093±0.023	172	119±22
Self-Forcing @1	12.7	0.159±0.017	0.158±0.011	24.6±0.7	0.838±0.015	0.751±0.058	0.652±0.078	0.146±0.034	0.087±0.015	175	102±20
One-Forcing @1	11.4	0.153±0.017	0.152±0.011	24.6±0.7	0.839±0.015	0.753±0.058	0.667±0.078	0.128±0.027	0.095±0.017	170	103±20
OneFixer @1	5.83	0.117±0.013	0.101±0.007	24.0±0.7	0.830±0.015	0.808±0.052	0.726±0.073	0.120±0.019	0.088±0.016	158	102±21
Proprietary Seoul (32 scenes × 900 frames)											
GSFix3D @8	15.7	0.250±0.011	0.211±0.011	21.2±0.6	0.731±0.024	0.610±0.072	0.406±0.047	0.368±0.037	0.306±0.028	654	301±28
DiFix @1	7.78	0.195±0.011	0.131±0.004	19.3±0.8	0.695±0.024	0.742±0.064	0.683±0.029	0.262±0.024	0.175±0.011	446	187±26
Harmonizer @1	30.7	0.209±0.014	0.179±0.009	22.5±0.6	0.761±0.023	0.735±0.063	0.709±0.034	0.251±0.024	0.167±0.011	370	121±18
Harmonizer (temporal) @1	28.0	0.244±0.015	0.211±0.010	21.4±0.7	0.756±0.021	0.720±0.071	0.652±0.043	0.281±0.028	0.153±0.010	423	99±17
3DGS Enhancer @8	18.9	0.282±0.013	0.242±0.009	21.2±0.6	0.735±0.024	0.499±0.071	0.537±0.050	0.337±0.034	0.137±0.010	406	105±17
ArtiFixer (DMD) @1	20.2	0.239±0.014	0.251±0.011	20.6±0.5	0.724±0.025	0.671±0.066	0.614±0.053	0.366±0.032	0.145±0.013	344	111±20
OmniDreams @1	27.4	0.251±0.016	0.241±0.010	22.1±0.5	0.770±0.021	0.698±0.067	0.532±0.071	0.278±0.032	0.140±0.015	432	134±18
OmniDreams @4	13.1	0.193±0.011	0.148±0.005	21.1±0.6	0.747±0.022	0.733±0.067	0.723±0.030	0.242±0.022	0.142±0.012	368	206±29
Self-Forcing @1	27.2	0.238±0.015	0.239±0.010	22.7±0.6	0.768±0.022	0.690±0.065	0.563±0.058	0.272±0.032	0.147±0.017	375	106±19
One-Forcing @1	26.1	0.237±0.015	0.239±0.010	22.7±0.6	0.769±0.021	0.690±0.063	0.564±0.055	0.272±0.031	0.149±0.014	369	106±19
OneFixer @1	7.13	0.148±0.010	0.123±0.006	22.9±0.6	0.785±0.021	0.766±0.062	0.757±0.029	0.220±0.020	0.141±0.012	312	101±18

(a) Component ablation on Waymo.

ID	SF		LPIPS	Map	Mix	FVD ↓	LPIPS ↓	Cuboid F1 ↑	FVMD ↓
V1	×		×	○	○	13.5	0.187±0.018	0.758±0.055	232
V2	○		×	○	○	13.3	0.155±0.017	0.759±0.058	180
V3	×		$\circ_4$	○	○	6.9	0.123±0.014	0.795±0.053	210
V4	○	+	$\circ_1$	○	○	7.9	**0.114±0.022**	0.793±0.084	188
V5	○	+	$\circ_4$	×	○	7.3	0.121±0.013	0.787±0.055	162
V6	○	+	$\circ_4$	○	×	104.8	0.405±0.028	0.715±0.084	711
V7	○	→	$\circ_4$	○	○	7.7	0.119±0.014	0.803±0.052	193
V8	△	+	$\circ_4$	○	○	9.0	0.158±0.018	0.785±0.054	189
V9	○	+	$\circ_4$	○	○	**5.8**	0.117±0.013	**0.808±0.053**	**158**

(b) Closed-loop evaluation.

Renderer	Closed-loop Replay Collision (%) ↓	Route dist. (m) ↓	Off-road (%) ↓
3DGS	10.6±2.2	4.18±0.22	5.1±1.3
DiFix	13.0±2.4	4.00±0.19	3.3±1.0
OmniDreams @4	10.2±2.0	4.31±0.23	4.3±1.2
Self-Forcing @1	10.7±1.8	4.32±0.22	3.6±1.1
OneFixer @1	**6.9±1.7**	**3.96±0.22**	**2.4±0.9**

Table 2: **Ablation and closed-loop evaluation.** (a) SF: causal history for flow matching — × ground truth, △ re-denoised ground truth (Sec. 4.3), ○ own rollout. $\circ_k$: LPIPS with backpropagation horizon k (k=4 default). Map: lane and actor conditioning. Mix: opacity-mixed source; × uses noise with channel-concatenated 3DGS latents. +: joint; →: 250 SF then 150 LPIPS-only steps. (b) Closed-loop replay on 60 proprietary scenes × 5 policy seeds, varying only the renderer; mean and 95% interval over scene–seed runs. Collision is the at-fault rate (front or lateral); route distance is the mean deviation from the logged trajectory.

one-step latency. Two limitations remain. Fine structures such as text and road markings are not reliably recovered, and strong perceptual scores do not guarantee their exactness. Perceptual supervision through decoded predictions also makes peak training memory grow with both the parallel temporal context and the perceptual gradient window (Appendix B.2); truncation keeps the default configuration on a single device, and scaling to higher resolutions or more cameras trades context for memory. Cheaper feature-space supervision (Wang et al., 2026a), lightweight decoders (Zhuang et al., 2025), and reduced backbone redundancy (Guo et al., 2025) are natural next steps, provided they preserve causal refinement quality and the simplicity of shared-rollout training.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, clarity, and formatting of the manuscript. All technical content, experimental results, and claims were reviewed and verified by the authors, who take full responsibility for the final submission.

REPRODUCIBILITY STATEMENT

Source code for training, inference, and the frozen evaluation protocol is provided as supplementary material, including the exact Waymo recipe used for Table 1 (`reproduce_sv_16gpu.sh`) and the metric implementations described in App. C.1. The pretrained checkpoint is the publicly released pre-DMD causal OmniDreams model (App. A.3). Waymo data generation is described in App. A.1; the Seoul dataset is proprietary and cannot be released, but its preprocessing and training configuration are documented in App. A.2 and A.3 and included in the code. Baseline configurations follow App. C.2.

REFERENCES

Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. TransFusion: Robust LiDAR-camera fusion for 3D object detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1090–1099, 2022.

Yunpeng Bai, Yossi Gandelsman, Michaël Gharbi, and Qixing Huang. Qwen-video-edit: Instruction-based video editing by repurposing an image editing model. *arXiv preprint arXiv:2608.14790*, 2026.

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*, 2020.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Li Chen, Chonghao Sima, Yang Li, Zehan Zheng, Jiajie Xu, Xiangwei Geng, Hongyang Li, Conghui He, Jianping Shi, Yu Qiao, and Junchi Yan. PersFormer: 3D lane detection via perspective transformer and the OpenLane benchmark. In *European Conference on Computer Vision (ECCV)*, pp. 550–567, 2022.

Zheng Chen, Zichen Zou, Kewei Zhang, Xiongfei Su, Xin Yuan, Yong Guo, and Yulun Zhang. DOVE: Efficient one-step diffusion model for real-world video super-resolution. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Kevin Clark, Paul Vicol, Kevin Swersky, and David J. Fleet. Directly fine-tuning diffusion models on differentiable rewards. In *International Conference on Learning Representations*, 2024.

Riccardo de Lutio, Tobias Fischer, Yen-Yu Chang, Yuxuan Zhang, Jay Zhangjie Wu, Xuanchi Ren, Tianchang Shen, Katarina Tothova, Zan Gojcic, and Haithem Turki. Artifixer: Enhancing and extending 3D reconstruction with auto-regressive diffusion models. In *ACM SIGGRAPH*, 2026.

Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Image Analysis: 13th Scandinavian Conference (SCIA)*, volume 2749 of *Lecture Notes in Computer Science*, pp. 363–370. Springer, 2003.

Jiaqi Feng, Justin Cui, Yuanhao Ban, and Cho-Jui Hsieh. One-forcing: Towards stable one-step autoregressive video generation. *arXiv preprint arXiv:2605.23458*, 2026.

Jinpei Guo, Yifei Ji, Zheng Chen, Yufei Wang, Sizhuo Ma, Yong Guo, Yulun Zhang, and Jian Wang. Towards redundancy reduction in diffusion models for efficient video super-resolution. *arXiv preprint arXiv:2509.23980*, 2025.

Yuwei Guo, Ceyuan Yang, Hao He, Yang Zhao, Meng Wei, Zhenheng Yang, Weilin Huang, and Dahua Lin. End-to-end training for autoregressive video diffusion via self-resampling. In *European Conference on Computer Vision (ECCV)*, 2026.

Dailan He, Guanlin Feng, Xingtong Ge, Yi Zhang, Bingqi Ma, Guanglu Song, Yu Liu, and Hongsheng Li. Ar-copo: Align autoregressive video generation with contrastive policy optimization. In *European Conference on Computer Vision*, pp. 57–73. Springer, 2026.

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.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023.

Aro Kim, Myeongjin Jang, Chaewon Moon, Youngjin Shin, Jinwoo Jeong, and Sang-hyo Park. FiDeSR: High-fidelity and detail-preserving one-step diffusion super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 38270–38280, 2026.

Haodong Li, Shaoteng Liu, Zhe Lin, and Manmohan Chandraker. Rolling sink: Bridging limited-horizon training and open-ended testing in autoregressive video diffusion. *arXiv preprint arXiv:2602.07775*, 2026.

Tianyu Li, Yihang Qiu, Zhenhua Wu, Carl Lindström, Peng Su, Matthias Nießner, and Hongyang Li. MTGS: Multi-traversal gaussian splatting. *arXiv preprint arXiv:2503.12552*, 2025.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, Xuanmao Li, Xingpeng Sun, Rohan Ashok, Aniruddha Mukherjee, Hao Kang, Xiangrui Kong, Gang Hua, Tianyi Zhang, Bedrich Benes, and Aniket Bera. DL3DV-10K: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22160–22169, 2024.

Jiahe Liu, Youran Qu, Qi Yan, Xiaohui Zeng, Lele Wang, and Renjie Liao. Fréchet video motion distance: A metric for evaluating motion consistency in videos. *arXiv preprint arXiv:2407.16124*, 2024a.

Xi Liu, Chaoyi Zhou, and Siyu Huang. 3dgs-enhancer: Enhancing unbounded 3d gaussian splatting with view-consistent 2d diffusion priors. In *Advances in Neural Information Processing Systems*, 2024b.

Yang Luo, Shengju Qian, Xiaohang Tang, Zirui Zhu, Yong Liu, Xin Wang, and Yang You. On-policy adversarial flow distillation for autoregressive video generation. *arXiv preprint arXiv:2605.26105*, 2026.

Yueru Luo, Chaoda Zheng, Xu Yan, Tang Kun, Chao Zheng, Shuguang Cui, and Zhen Li. LATR: 3D lane detection from monocular images with transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7941–7952, 2023.

NVIDIA. Cosmos 3: Omnimodal world models for physical AI. *arXiv preprint arXiv:2606.02800*, 2026.

NVIDIA, Yulong Cao, Riccardo de Lutio, Sanja Fidler, Guillermo Garcia Cobo, Zan Gojcic, Maximilian Igl, Boris Ivanovic, Peter Karkus, Janick Martinez Esturo, Marco Pavone, Aaron Smith, Ellie Tanimura, Michal Tyszkiewicz, Michael Watson, Qi Wu, and Le Zhang. Alpasim: A modular, lightweight, and data-driven research simulator for autonomous driving, October 2025. URL `https://github.com/NVlabs/alpasim`. Software.

NVIDIA, Aarti Basant, Amlan Kar, Despoina Paschalidou, Fangyin Wei, Francesco Ferroni, Guillermo Garcia Cobo, Haithem Turki, Huan Ling, Jaewoo Seo, James Lucas, Jay Zhangjie Wu, Jialiang Wang, Jonathan Lorraine, Jun Gao, Kai He, Katarina Tothova, Kevin Xie, Michał Tyszkiewicz, Qi Wu, Riccardo de Lutio, Ruilong Li, Sanja Fidler, Seung Wook Kim, Tianchang Shen, Tianshi Cao, Tobias Pfaff, William Lew, Xindi Wu, Xuanchi Ren, Yifan Lu, Yuxuan Zhang, Zan Gojcic, and Zian Wang. NVIDIA omnidreams: Real-time generative world model for closed-loop autonomous vehicle simulation. *arXiv preprint arXiv:2606.03159*, 2026.

Gaurav Parmar, Taesung Park, Srinivasa Narasimhan, and Jun-Yan Zhu. One-step image translation with text-to-image models. *arXiv preprint arXiv:2403.12036*, 2024.

Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation. *arXiv preprint arXiv:2310.03739*, 2023.

Dmitriy Rivkin, Parker Ewen, Lili Gao, Julian Ost, Stefanie Walz, Rasika Kangutkar, Mario Bijelic, and Felix Heide. Chopgrad: Pixel-wise losses for latent video diffusion via truncated backpropagation. *arXiv preprint arXiv:2603.17812*, 2026. doi: 10.48550/arXiv.2603.17812.

Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3626–3636, June 2022.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurélien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Su Sun, Cheng Zhao, Zhuoyang Sun, Yingjie Victor Chen, and Mei Chen. SplatFlow: Self-supervised dynamic gaussian splatting in neural motion flow field for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27487–27496, 2025.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

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

Jianyi Wang, Shanchuan Lin, Zhijie Lin, Yuxi Ren, Meng Wei, Zongsheng Yue, Shangchen Zhou, Hao Chen, Yang Zhao, Ceyuan Yang, Xuefeng Xiao, Chen Change Loy, and Lu Jiang. SeedVR2: One-step video restoration via diffusion adversarial post-training. In *International Conference on Learning Representations*, 2026a.

Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. Probabilistic and geometric depth: Detecting objects in perspective. In *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pp. 1475–1485. PMLR, 2022.

Yifan Wang, Yanyu Li, Gordon Guocheng Qian, Sergey Tulyakov, Yun Fu, and Anil Kag. Diffusion-DRF: Free, rich, and differentiable reward for video diffusion fine-tuning. *arXiv preprint arXiv:2601.04153*, 2026b.

Jiaxin Wei, Stefan Leutenegger, and Simon Schaefer. GSFix3D: Diffusion-guided repair of novel views in Gaussian splatting. *arXiv preprint arXiv:2508.14717*, 2025.

Jay Zhangjie Wu, Yuxuan Zhang, Haithem Turki, Xuanchi Ren, Jun Gao, Mike Zheng Shou, Sanja Fidler, Zan Gojcic, and Huan Ling. DIFIX3D+: Improving 3D reconstructions with single-step diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26024–26035, 2025.

Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Zhenhua Wu, Yun Pang, Mingkun Chang, Yuwei Ning, Liangzhi Wang, Yi Xiao, and Guanbin Li. Realitybridge: Bridging editable 3D gaussian splatting driving simulations and real-world videos. *arXiv preprint arXiv:2606.16278*, 2026.

Yunzhi Yan, Haotong Lin, Chenxu Zhou, Weijie Wang, Haiyang Sun, Kun Zhan, Xianpeng Lang, Xiaowei Zhou, and Sida Peng. Street gaussians: Modeling dynamic urban scenes with gaussian splatting. In *European Conference on Computer Vision (ECCV)*, 2024. doi: 10.1007/978-3-031-73464-9_10.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6613–6623, 2024.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22963–22974, 2025.

Yuyang You, Yongzhi Li, Jiahui Li, Yadong Mu, Quan Chen, and Peng Jiang. Adaptive video distillation: Mitigating oversaturation and temporal collapse in few-step generation. *arXiv preprint arXiv:2603.21864*, 2026.

Aiping Zhang, Zongsheng Yue, Renjing Pei, Wenqi Ren, and Xiaochun Cao. Degradation-guided one-step image super-resolution with diffusion priors. *arXiv preprint arXiv:2409.17058*, 2024.

Jingran Zhang, Ning Li, Yuanhao Ban, Andrew Bai, and Justin Cui. Reward-forcing: Autoregressive video generation with reward feedback. *arXiv preprint arXiv:2601.16933*, 2026a. doi: 10.48550/arXiv.2601.16933.

Songchun Zhang, Zeyue Xue, Siming Fu, Jie Huang, Xianghao Kong, Y Ma, Haoyang Huang, Nan Duan, and Anyi Rao. Astrolabe: Steering forward-process reinforcement learning for distilled autoregressive video models. *arXiv preprint arXiv:2603.17051*, 2026b. doi: 10.48550/arXiv.2603.17051.

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 (CVPR)*, 2026c.

Min Zhao, Hongzhou Zhu, Kaiwen Zheng, Zihan Zhou, Bokai Yan, Xinyuan Li, Xiao Yang, Chongxuan Li, and Jun Zhu. Causal forcing++: Scalable few-step autoregressive diffusion distillation for real-time interactive video generation. *arXiv preprint arXiv:2605.15141*, 2026.

Kaiwen Zheng, Guande He, Min Zhao, Jintao Zhang, Huayu Chen, Jianfei Chen, Chen-Hsuan Lin, Ming-Yu Liu, Jun Zhu, and Qianli Ma. Causal-rcm: A unified teacher-forcing and self-forcing open recipe for autoregressive diffusion distillation in streaming video generation and interactive world models. *arXiv preprint arXiv:2606.25473*, 2026.

Yang Zheng, Adam W. Harley, Bokui Shen, Gordon Wetzstein, and Leonidas J. Guibas. PointOdyssey: A large-scale synthetic dataset for long-term point tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19855–19865, 2023.

Jingqiu Zhou, Lue Fan, Linjiang Huang, Xiaoyu Shi, Si Liu, Zhaoxiang Zhang, and Hongsheng Li. FlexDrive: Toward trajectory flexibility in driving scene gaussian splatting reconstruction and rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1549–1558, 2025.

Lijun Zhou, Hongcheng Luo, Zhenxin Zhu, Cheng Chi, Mingfei Tu, Kaixin Xiong, et al. Xiaomi EV World Model: A joint world model integrating reconstruction and generation for autonomous driving. *arXiv preprint arXiv:2605.18137*, 2026.

Xiaoyu Zhou, Zhiwei Lin, Xiaojun Shan, Yongtao Wang, Deqing Sun, and Ming-Hsuan Yang. DrivingGaussian: Composite gaussian splatting for surrounding dynamic autonomous driving scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21634–21643, 2024.

Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. *arXiv preprint arXiv:2602.02214*, 2026. doi: 10.48550/arXiv.2602.02214.

Junhao Zhuang, Shi Guo, Xin Cai, Xiaohui Li, Yihao Liu, Chun Yuan, and Tianfan Xue. FlashVSR: Towards real-time diffusion-based streaming video super-resolution. *arXiv preprint arXiv:2510.12747*, 2025.

A Implementation Details

A.1 3DGS Data Generation - Waymo

We construct a multi-modal driving corpus in which each sequence provides a *degraded* RGB render, the *logged* front-camera RGB, an opacity map, and a rasterized semantic map. The objective is to train models that recover photometric fidelity and geometric consistency when novel views are synthesized from an imperfect 3D scene representation—a setting related to recent one-step diffusion refinements of 3D reconstructions (Wu et al., 2025). Our pairs are produced deterministically from explicit 3D Gaussian reconstructions rather than from a diffusion model, yielding controllable and repeatable degradation without generative stochasticity.

Reference reconstruction. Each sequence begins with a *Street Gaussians* (Yan et al., 2024) reconstruction of a dynamic urban clip: static background, sky, and moving actors are modeled as separate Gaussian fields with time-varying actor poses, trained on the logged camera trajectory with LiDAR depth and sky-mask supervision. This reconstruction defines geometry, appearance, and actor motion for all subsequent rendering.

Cycle-render degradation. To synthesize a structurally grounded but visually degraded input, we apply a four-stage *cycle render*:

1. Render training views from the reference model with the camera shifted laterally by $1\,\mathrm{m}$ in the vehicle frame.
2. Train a *second* Street-Gaussian scene representation on these shifted views only. Actors are largely pruned during this stage because motion is inconsistent with static geometry in the shifted views.
3. *Splice* actor Gaussians and their pose tracks from the reference checkpoint into the cycle model so dynamic objects remain present at render time.
4. Re-render at the *logged* camera poses. Because the cycle model was trained $1\,\mathrm{m}$ to the left, this nominal-trajectory view corresponds to a $1\,\mathrm{m}$ right-shifted novel view relative to the cycle training cameras—introducing parallax, blur, and exposure mismatch while preserving coarse scene layout.

The degraded RGB sequence forms `rgb_input`; the original logged frames form `rgb_gt`. Accumulated renderer alpha, with sky regions filled from segmentation masks, forms `opacity`. This

Table 3: Cycle-render paired corpus statistics (front camera).

Quantity	Train	Validation
Scenes (four modalities present)	350	80
with dynamic actors	241	52
Median frames per scene	198	198
Total aligned frame pairs	69,238	15,812

lateral-extrapolation artifact is representative of novel-view errors from feed-forward 3D representations and motivates learning-based restoration in the spirit of reconstruction-oriented diffusion methods (Wu et al., 2025).

Semantic map (`sdmap`). Semantic-map videos are rasterized bird's-eye-view-style conditioning frames (1920×1280 at 10 fps) aligned to each sequence's front camera. They follow the OmniDreams *Ludus* HD-map renderer, which drives NVIDIA's FlashDreams inference stack (`https://github.com/NVIDIA/flashdreams`) to turn vector scene geometry into temporally consistent map video. Concretely, per-frame *clipGT* assets are built from public annotations: OpenLane `lane3d` lane centerlines (white/yellow polylines and road boundaries), together with 3D bounding boxes from the original dataset's native object labels (laser-based detection boxes on dynamic actors and traffic signs where available). These vector pools and the logged camera trajectory are fed to `LudusRenderer`, which produces stylized HD-map frames matched to each timestamp. In the cycle-render corpus, `sdmap` clips are taken from a precomputed library generated with this pipeline rather than re-rendered at cycle time. When Ludus is unavailable, an equivalent CPU rasterizer can project the same OpenLane polylines and 3D boxes into the camera using standard calibration, preserving semantic content at the cost of the learned Ludus appearance. Map length may be shorter than the RGB stream on some sequences; in that case, all modalities are truncated to the shortest available length before training.

Corpus statistics. Table 3 summarizes the constructed train/validation splits. Each scene contributes approximately 198 temporally aligned samples; after truncation to the common length, the training split contains **350** scenes and **69,238** frame-level pairs, and the validation split contains **80** scenes and **15,812** pairs. A smaller quality-filtered subset (verified actor splice and uniform frame counts) contains 276 train and 59 validation scenes, of which 241 and 52 include non-static traffic. RGB and opacity are 1600×1072 at 24 fps; semantic maps are 1920×1280 at 10 fps.

A.2 3DGS Data Generation - Internal

We use an internal multi-camera urban driving dataset collected from production vehicles operating on public roads in South Korea, covering diverse urban and suburban environments such as arterial roads, intersections, highways, and dense city streets. Each clip contains synchronized surround-view RGB streams with logged camera calibration, ego-vehicle poses, and temporally aligned semantic-map metadata. The original front-camera videos are recorded at 1920×1080 and are processed at 10 fps. Training source clips use 30-second windows (300 frames), from which 93-frame sequences are sampled; the 32 held-out evaluation scenes use merged 90-second windows (900 frames) to test long-horizon rollouts.

We construct a paired driving-video corpus using multi-camera logged urban sequences and 3D Gaussian scene reconstruction. For each scene, we first build an MTGS scene representation following Li et al. (2025), then render the trained representation to obtain a degraded RGB sequence. The rendered sequence is paired with the original logged RGB sequence at the same camera trajectory, together with renderer opacity and a rasterized semantic map.

Scene reconstruction. Each source clip is represented as a dynamic 3D Gaussian scene. The reconstruction is trained from synchronized multi-camera video using the logged camera calibration and trajectory. We use the front, front-left, front-right, side-left, and side-right cameras when available. The training objective follows the MTGS/3DGS renderer stack: the Gaussian scene is optimized to reproduce the logged views, while geometry-aware visibility is captured through the

rendered opacity channel. Ego-vehicle regions are masked during preprocessing so that the reconstructed scene models the external environment rather than the vehicle body.

A.3 OneFixer Training Details

Backbone and clip representation. We initialize OneFixer from the official OmniDreams causal checkpoint released prior to DMD distillation.[2] This differs from the separately released DMD-distilled OmniDreams checkpoint used for the OmniDreams baseline in Sec. 4.

Training uses 544×960 crops and 24 latent positions corresponding to 93 RGB frames. Each causal block contains two latent positions, corresponding to 8 RGB frames per generation. The Waymo model uses a local-attention window of 12 latent positions; the proprietary model uses a window of 6. We do not use CFG at inference.

Self-rollout and truncated gradients. To supervise a self-generated rollout, we add a perceptual term on RGB reconstructed from the model's own autoregressive history. A target latent index is sampled uniformly from 0–23, covering the full $N{=}24$ clip. The corresponding decoded frames (a few RGB frames per latent, following Wan's temporal compression) are compared with identically decoded ground truth using a frozen VGG LPIPS model, weighted approximately an order of magnitude above the flow-matching term.

Backpropagation is truncated in two places. On the autoregressive side, gradients are retained only through the last 4 latent frames ending at the sampled index (two autoregressive blocks of size 2); earlier history is stop-gradient. On the Wan decoder side, decoding is causal and cache-based, but the decoder feature cache is detached at every latent frame. Consequently, perceptual gradients do not propagate through decoder recurrence, corresponding to a truncation distance of 1 following Rivkin et al. (2026).

B Further Analysis

B.1 Robustness to Geometry-Conditioning Errors

Evaluation protocol. To keep the computational cost tractable, we randomly sample 32 scenes from the 80-scene Waymo evaluation set and use the same subset for all perturbation settings and reference methods. We perturb the lane and dynamic-agent geometry used to construct the sdmap conditioning, while keeping the input 3DGS RGB renderings, opacity, and evaluation targets unchanged. The perturbations are parameterized by a lane displacement $\ell \in \{-0.50, -0.25, 0, 0.25, 0.50\}$ m and planar object-box translations $\delta_x, \delta_y \in \{-0.25, 0, 0.25\}$ m. The Cartesian sweep comprises 45 configurations. For visualization, we group results by the absolute lane displacement $|\ell|$ or object translation magnitude $\sqrt{\delta_x^2 + \delta_y^2}$, whose maximum is approximately 0.354 m. A zero value on one axis therefore does not necessarily indicate fully unperturbed conditioning, since the other parameters can remain nonzero.

Perturbation scale. The tested displacements are comparable in magnitude to reported errors in driving-scene perception. On OpenLane, LATR with a ResNet-50 backbone reports near- and far-range lateral errors of approximately 0.22 and 0.26 m (Luo et al., 2023); our maximum lane displacement is approximately twice these values. For object geometry, TransFusion reports a mean translation error of approximately 0.26 m on nuScenes (Bai et al., 2022), comparable to the 0.25–0.354 m nonzero translations tested here. These benchmarks ground the sweep in representative decimeter-scale errors. The experiment is a controlled sensitivity test: fixed geometric offsets do not reproduce the full distribution of detection, localization, or map-alignment errors encountered in deployment.

Results. Fig. 7 shows that OneFixer retains the benefits of geometry conditioning under the tested perturbations. Lane F1 remains above the unconditioned reference throughout the sweep, while

[2] https://huggingface.co/nvidia/omni-dreams-models/tree/main/single_view/student-init

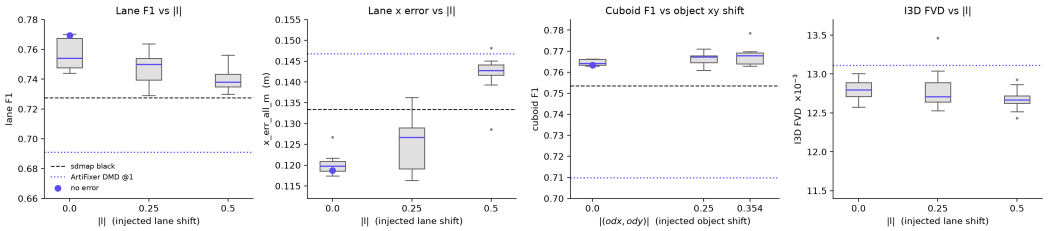

Figure 7: **Robustness to geometry-conditioning errors.** Evaluation uses a fixed random subset of 32 Waymo scenes. From left to right: Lane F1, lane lateral error, Cuboid F1, and I3D FVD, grouped by absolute lane displacement or planar object translation magnitude. Injected displacements are measured in meters. Boxplots summarize variation across perturbation configurations within each group. Black dashed lines indicate unconditioned OneFixer (black sdmap input), purple dotted lines indicate ArtiFixer DMD @1 on the same subset, and purple dots mark the fully unperturbed reference where shown. Lane F1 and Cuboid F1 remain above the unconditioned reference, although the largest lane displacement produces higher lateral error.

Cuboid F1 stays stable across object translations and consistently exceeds the unconditioned reference. I3D FVD likewise shows no systematic degradation with increasing lane displacement, suggesting that overall video fidelity is robust to these errors. Although lane lateral error exceeds the unconditioned reference at the largest displacement, the retained Lane F1 advantage indicates that conditioning remains useful for recovering lane structure. Overall, these results support the practical value of imperfect geometric guidance, with a trade-off in positional accuracy under larger lane misalignment.

B.2 Training Memory

Direct perceptual supervision on the shared rollout retains gradients through decoded predictions and the last k latent positions of the autoregressive history, so peak training memory depends on both the number of latent positions processed in parallel and the perceptual gradient window. Fig. 8 measures single-GPU peak allocated memory at 544×960 with the perceptual loss placed on the final latent, the configuration that maximizes the retained history. The default setting of 24 latent positions with $k = 4$ requires about 102 GiB, and memory increases monotonically along both axes: at 24 positions it reaches 124 GiB for $k = 8$ and exceeds device capacity for $k \geq 12$, whereas shortening the parallel context to 16 positions accommodates k up to 16. Because each causal block spans two latent positions, $k = 1$ and $k = 2$ coincide. Truncation therefore keeps the default configuration comfortably within a single device, and the two knobs offer a direct trade-off: a longer perceptual gradient window is available whenever a shorter parallel temporal context is acceptable, which is the trade made for the multi-camera and high-resolution settings in Appendix E. This memory–horizon trade-off is not specific to OneFixer and also arises in rollout-based autoregressive and distillation methods that optimize self-generated trajectories.

B.3 Generalization to Non-Driving Scenes.

To test transfer beyond driving, we fine-tune OneFixer on 32 DL3DV scenes (Ling et al., 2024) with degraded renders produced by under-trained 3dgs checkpoints, without lane or actor conditioning (black sdmap input). DiFix and ArtiFixer use their official DL3DV-finetuned checkpoints. Table 4 and Fig. 10 report results on the held-out scenes.

C Comparison

C.1 Evaluation Protocol and Metrics

Unless otherwise specified, metrics are computed against the corresponding clean reference sequence. For metrics evaluated independently per clip, we report the mean together with a 95% t-interval across clips. I3D FVD and FVMD are distributional metrics computed over an evaluation

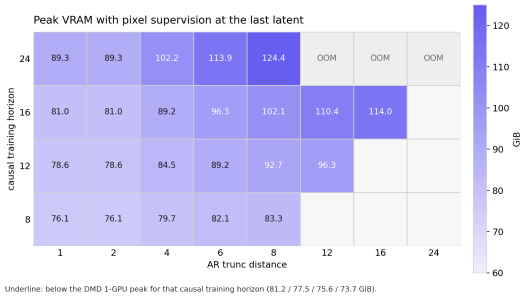

Figure 8: **Peak training memory versus parallel context and perceptual gradient window.** Single-GPU peak allocated memory (GiB) at 544×960 with one-step generation, a full rollout, and the perceptual loss on the final latent position, as a function of the number of latent positions processed in parallel (rows) and the perceptual backpropagation window k in latent steps (columns). Grey cells exceed device capacity; blank cells have k larger than the context. The default configuration (24 positions, $k = 4$) is 102 GiB. Multi-GPU training with sharded parameters shows comparable per-device peaks.

set and therefore do not receive per-clip confidence intervals. For compact presentation in Table 1, I3D FVD, FVMD, and Warp L1 are reported scaled by 10^{-3}, 10^{-1}, and 10^{4}, respectively.

FVD. We report I3D-based Fréchet Video Distance (FVD) (Unterthiner et al., 2018) using the full-clip evaluation protocol adopted by StyleGAN-V (Skorokhodov et al., 2022). Unlike frame-wise perceptual metrics, FVD compares the distributions of spatiotemporal video features extracted from generated and reference clips, providing a complementary measure of overall video fidelity. A single Fréchet distance is computed over the evaluation set. Because the Fréchet estimate is biased upward for smaller sets and shorter clips, FVD values on the 32-scene subsets used in Fig. 6(c) (93-frame rollouts) and Fig. 7 (198-frame rollouts) are higher than, and not comparable to, the 80-scene, 198-frame values in Table 1, nor to each other; only within-figure comparisons are meaningful there.

Perceptual metrics. We report LPIPS, PSNR, and SSIM for frame-level perceptual and reconstruction quality. Each metric is computed between generated and clean-reference frames and aggregated over the evaluated sequence. For evaluation, LPIPS uses the AlexNet backbone, following the common protocol used in prior 3DGS enhancement benchmarks; this evaluation backbone is distinct from the frozen VGG-based LPIPS used for training. The reported value is the mean across clips, with the corresponding 95% t-interval.

DISTS. We additionally report DISTS, which compares deep-feature structure and texture statistics and is less sensitive than PSNR to plausible high-frequency detail that does not align pixel-wise with the reference. We use the `pyiqa` implementation, compute the score per frame against the clean reference, aggregate over the evaluated sequence, and report the mean across clips with the corresponding 95% t-interval.

Cuboid F1. We run a PGD monocular 3D detector (Wang et al., 2022) with a ResNet-101 FPN and the nuScenes (Caesar et al., 2020) mono3D-finetuned checkpoint. Images are rescaled so that the focal length matches 1266.42 pixels. We retain nuScenes vehicle labels $\{0, 1, 2, 3, 4\}$ with score at least 0.10. The same detector is run on generated and ground-truth videos; the ground-truth reference is not derived from the map raster.

For each frame, a detected 7-DoF box is projected to its eight image-space corners; boxes with any corner behind $z = 0.5\,\mathrm{m}$ are removed. We fill the projected convex hulls to obtain binary masks and aggregate intersection, predicted area, and reference area over the clip. With $P = I/A_{\mathrm{pred}}$ and $R = I/A_{\mathrm{ref}}$, we report

$$\text{Cuboid F1} = \frac{2PR}{P + R}. \tag{7}$$

Thus, Cuboid F1 measures projected-mask overlap rather than 3D-IoU box matching.

Lane F1 and lateral error. We detect lanes with LATR (Luo et al., 2023) using the OpenLane (Chen et al., 2022) checkpoint, a class-score threshold of 0.50, at least three visible points, and at least $10\,\mathrm{m}$ longitudinal span. Each lane is resampled at $y = 3, 5, \ldots, 60\,\mathrm{m}$. Lanes with less than $20\,\mathrm{m}$ support or $|\operatorname{mean} x| > 10\,\mathrm{m}$ are excluded. Candidate pairs must overlap by at least $10\,\mathrm{m}$ and satisfy $|\Delta x| \leq 1.5\,\mathrm{m}$ for at least 75% of their overlapping samples. We then perform Hungarian matching using mean absolute lateral error. Lane F1 is computed from the resulting instance-level true positives, false positives, and false negatives. Lane lateral error is the mean $|\Delta x|$ over matched samples; the main table reports the combined near-and-far value, with the boundary at $30\,\mathrm{m}$.

Lane jitter. For a tracked lateral lane position v_t, we define the high-pass residual

$$h_t = v_t - \tfrac{1}{2}(v_{t-1} + v_{t+1}), \tag{8}$$

and report the root mean square of valid residuals. Lanes are associated across frames using a $1.5\,\mathrm{m}$ lateral threshold and a minimum track length of five frames. The reported lane jitter evaluates lateral position at $y = 40\,\mathrm{m}$; the evaluator additionally records a near-range value at $y = 20\,\mathrm{m}$.

FVMD. FVMD (Liu et al., 2024a) measures motion consistency using trajectory statistics and complements appearance-based I3D FVD. We compute FVMD on 16-frame windows sampled every 25 frames throughout each evaluation sequence; thus, the reported score summarizes local motion consistency across the full rollout rather than direct long-range consistency between distant frames. Each window is JPEG-encoded at quality 95, resized to 256×256, and processed by PIPs++ (Zheng et al., 2023) to track 400 points with stride 8. The resulting trajectories are represented by velocity and acceleration histograms using a 5×5 spatial grid, four-frame temporal subcubes, and eight angle bins. FVMD is the Fréchet distance between the generated and reference trajectory feature distributions.

Flow-warped L1. Warp L1 measures short-term temporal consistency over the late evaluation window. We estimate dense optical flow from the current grayscale frame to the previous frame using Farnebäck flow (Farnebäck, 2003), with pyramid scale 0.5, three levels, window size 21, three iterations, polynomial neighborhood 5, and polynomial sigma 1.2. The previous RGB frame is backward-warped with bilinear interpolation, and samples whose warped coordinates fall outside the image are excluded. When a valid road-region mask is available, the error is restricted to that region; otherwise, all valid in-bounds pixels are used. For valid pixels Ω, we compute

$$\frac{1}{|\Omega|} \sum_{p \in \Omega} \frac{\|I_t(p) - \mathcal{W}(I_{t-1})(p)\|_1}{3 \cdot 255}. \tag{9}$$

The metric is averaged over the designated late temporal window for each clip, and Table 1 reports the mean and 95% t-interval across clips.

Closed-loop replay evaluation. We evaluate closed-loop replay fidelity on 60 proprietary scenes, each run with 5 policy seeds, using the AlpaSim evaluators (NVIDIA et al., 2025). Each scene–seed run is repeated several times and averaged; a small number of runs lost to system failures are excluded. We use the same per-timestep scorers and aggregation as the public AlpaSim evaluation: warm-up frames are discarded until the evaluator marks them as relevant (`eval_relevant`), and frames after the first off-road or collision event are excluded. Binary metrics are aggregated by taking the per-rollout MAX, averaging within each scene–seed run, and then averaging across runs; results are reported as percentages with a 95% interval ($1.96\,\sigma/\sqrt{N}$) over the N runs. The ego footprint is represented by a centered axis-aligned bounding box of size $4.855 \times 1.88\,\mathrm{m}$, without shrinkage or corner rounding.

Collision. At each step, we construct bird's-eye-view polygons for the ego vehicle and all other actors and query intersections with an STRtree. A colliding actor is classified as *front* if it intersects the ego front-bumper segment, *rear* if it intersects the rear-bumper segment, and *lateral* otherwise. We report AlpaSim's at-fault collision rate, the union of front and lateral collisions; rear impacts are scored but not reported.

Route distance. At each relevant step we measure the distance (m) from the ego position to the logged trajectory and average over the rollout; the table reports the mean over runs.

Off-road. Off-road evaluation uses the vector lane map. Nearby lanes are queried within $6\,\mathrm{m}$ of the ego vehicle without a heading constraint. A step is considered on-road if any individual lane polygon fully contains the ego footprint or, when the vehicle straddles adjacent lanes, if the union of nearby lane polygons contains it. Lane polygons are constructed from left and right lane edges when both are available; otherwise, a $1.85\,\mathrm{m}$ half-width buffer around the centerline is used. If neither containment test passes, the step is classified according to its distance to the nearest road-edge polyline using the evaluator's $0.001\,\mathrm{m}$ threshold. The reported value is the mean over runs of the per-rollout MAX of this binary indicator.

C.2 Baseline Configuration

All baseline videos are evaluated using the same frozen protocol described in App. C.1. We report each method at the operating point used in the main comparison rather than forcing all methods to use the same number of denoising steps.

GSFix3D. We initialize from the official GSFix3D checkpoint (Wei et al., 2025) (`gsfixer-base`) and fine-tune the UNet on paired degraded/clean videos from the internal dataset and Waymo at 544×960. Training uses 8 GPUs, per-GPU batch 2, and gradient accumulation 4 (effective batch 64), Adam at 3×10^{-5}, bfloat16, and 4 denoising steps. The reported checkpoint is iteration 10,000. The table uses 8-step inference.

DiFix. We initialize from the official NVIDIA DiFix checkpoint (Wu et al., 2025) (`nvidia/difix`) and fine-tune on the same internal dataset and Waymo. Training uses learning rate 2×10^{-5}, per-GPU batch 4, accumulation 2, bfloat16, and L2 + LPIPS + Gram losses (Gram warmup 2,000 steps). The reported run is 800 optimization steps. Inference is independent per frame at diffusion timestep 199 with prompt `remove degradation`: generation is performed at 576×1024 and then resized to 544×960. Evaluation uses one-step inference.

Harmonizer. We initialize from the official DiffusionHarmonizer checkpoint (Zhang et al., 2026c) (`nvidia/Harmonizer`) and fine-tune the full network at 544×960 on the internal dataset and Waymo for 4k steps each, following the released training command: 8 GPUs, per-GPU batch 4, accumulation 2 (effective batch 64), learning rate 2×10^{-5}, bfloat16, timestep 250, $\lambda_{\mathrm{L2}} = 1$, $\lambda_{\mathrm{LPIPS}} = 0.3$. *Harmonizer* (frame-wise) trains and infers on single frames. *Harmonizer (temporal)* continues from the frame-wise checkpoint for 1,500 further steps and conditions on the four previously enhanced frames at inference ($K{=}4$, offsets -1 to -4), as in the official inference script. The released temporal training path conditions on ground-truth prior frames (with 50% context dropout) and applies a RAFT warping loss; on our sequences this teacher-forced recipe drifted within a few dozen frames when rolled out on its own outputs. We therefore trained the temporal variant with self-generated context: the four references are the model's own detached, re-encoded outputs from a short autoregressive window, with the warping loss retained. This is the strongest temporal Harmonizer we could obtain on our data; it is a modification of the official recipe, not a reproduction of the paper's checkpoint. Both variants are evaluated at one step.

3DGS Enhancer. We initialize from Stable Video Diffusion (`img2vid-xt`) and fine-tune 3DGS-Enhancer (Liu et al., 2024b), a *bidirectional* (non-causal) video diffusion UNet. Fine-tuning uses 25-frame clips at 640×960 on the internal dataset and Waymo, with Adam at 1×10^{-5}, weight decay 0.01, fp16, batch 1 per GPU, and 500 steps. Outputs are resized to 544×960 for evaluation. Because the backbone is bidirectional and clip-length limited, full evaluation sequences are produced using overlapping 25-frame windows (overlap 8) and stitching. Guidance is 1.0. The main comparison uses 8-step inference.

ArtiFixer. We initialize from the official ArtiFixer 1.3B checkpoint (`artifixer-1.3b`) (de Lutio et al., 2026), built on Wan2.1-T2V-1.3B (Wan et al., 2025), and follow the three-stage pipeline on the internal dataset and Waymo at 544×960 using 93-frame clips. Training uses 16 GPUs with accumulation 4 (effective batch 64): 400 steps of bidirectional teacher training, 600 steps of causal adaptation (3×10^{-5}), and 200 steps of DMD/Self-Forcing (generator 2×10^{-6}, critic 4×10^{-7}, EMA 0.99). The reported result is the one-step DMD model with text guidance 1.0.

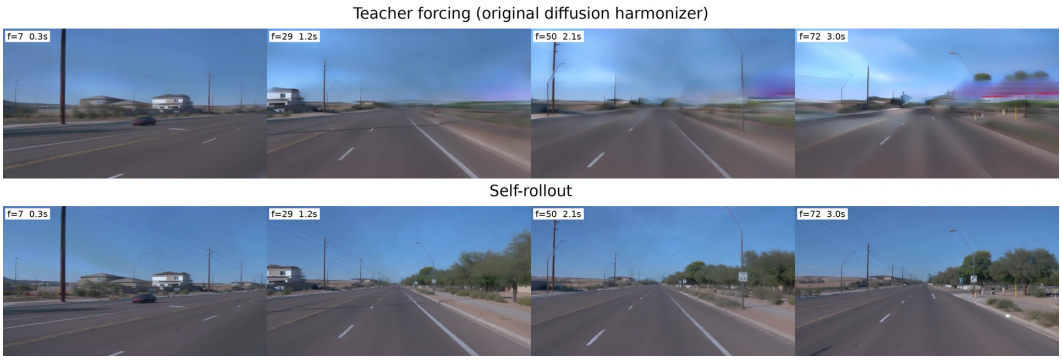

Figure 9: **Exposure bias in a frame-recurrent image fixer.** Harmonizer in temporal mode on a Waymo sequence, conditioning on its own four previous outputs at inference. Top: trained with the official teacher-forced recipe (ground-truth prior frames as context, 50% context dropout); appearance drifts within a few dozen frames and the model begins copying its own artifacts. Bottom: the same model trained with self-generated context; the rollout remains stable. This is the variant reported as Harmonizer (temporal) in Table 1. Frame index and elapsed time are shown in each panel.

Table 4: Generalization to non-driving scenes on DL3DV. Denoising steps are denoted by @N. Best and second-best results are shown in **bold** and underlined, respectively. FVMD is reported unscaled here.

	Temporal		Perception		
Model	Warp L1 ↓	FVMD ↓	LPIPS-Mid ↓	LPIPS-Last ↓	PSNR ↑
DiFix @1	0.0306	6744	0.140	0.164	21.5
ArtiFixer DMD @4	**0.0179**	3010	0.148	0.164	22.2
ArtiFixer DMD @1	0.0181	3411	0.176	0.193	22.3
OneFixer (Ours) @1	0.0201	**2326**	**0.126**	**0.136**	**22.6**

OmniDreams. We initialize from the official DMD-distilled causal OmniDreams checkpoint (NVIDIA et al., 2026) (`https://huggingface.co/nvidia/omni-dreams-models/tree/main/single_view/distilled`) and fine-tune it with diffusion forcing and the opacity-mixing rule on the internal dataset and Waymo. In contrast, OneFixer starts from the separately released causal checkpoint prior to DMD distillation. Training uses 8 GPUs, per-GPU batch 1, and accumulation 8 (effective batch 64) for 400 iterations at 544×960 using 93-frame windows. The main comparison uses the 4-step operating point with guidance 1.0.

Self-Forcing and One-Forcing. Following the staged recipe in Sec. 4.1, the bidirectional teacher is fine-tuned from the official pre-DMD bidirectional OmniDreams checkpoint (NVIDIA et al., 2026) to 3DGS refinement (Stage 1); the causal student is fine-tuned from the official pre-DMD causal checkpoint, OneFixer's initialization, to the fixer task (Stage 2); the Stage-2 student is distilled with DMD and Self-Forcing (Yin et al., 2024; Huang et al., 2025) against the Stage-1 teacher (Stage 3). Stages use 400/400/400 steps on Waymo and 500/500/400 on Seoul at 544×960, effective batch 64, with the same lane and dynamic-agent conditioning as OneFixer. One-Forcing adds the critic module and loss of Feng et al. (2026) to the third stage, which raises the per-iteration cost and accounts for its higher cumulative GPU-hours. Both are evaluated at one denoising step.

D Appearance-Conditioned Refinement

We further investigate whether OneFixer can extend beyond clear-weather restoration to *appearance-conditioned 3DGS refinement*. To construct paired supervision without requiring matched captures under rare conditions, we use AutoAWG (Hu et al., 2026) to transform the clean reference videos from our Waymo cycle-render training set into three target conditions: *night*, *fog*,

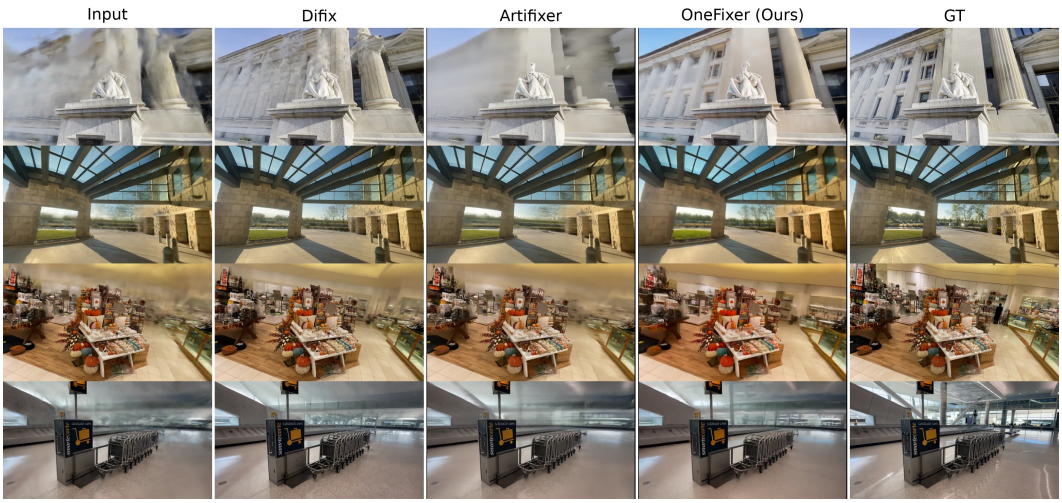

Figure 10: **Qualitative generalization on DL3DV.** OneFixer readily transfers beyond driving scenes after fine-tuning on DL3DV, while DiFix and ArtiFixer are evaluated using their official DL3DV-finetuned checkpoints. Under one-step inference, OneFixer matches or exceeds the frame-level visual quality of the image-based DiFix while providing stronger temporal consistency through video diffusion and produces higher-fidelity results than one-step ArtiFixer.



Figure 11: **Appearance-conditioned 3DGS refinement.** Given the same degraded 3DGS rendering, we compare prompt-conditioned generation for night, fog, and snow. General-purpose video editing models such as Cosmos3 and Qwen-Video-Edit can produce visually convincing target styles, but often substantially alter the underlying scene content and geometry inherited from the 3DGS input. In contrast, after paired prompt-conditioned training, OneFixer jointly performs 3DGS refinement and appearance transformation in a single pass, removing reconstruction artifacts while better preserving the underlying scene structure.

and *snow*. The generated targets preserve the original camera trajectory and scene content, and outputs exhibiting noticeable structural drift are filtered using alignment with the corresponding clean sequence. The original clear-weather sequence is retained as an identity condition. Importantly, the input to OneFixer is unchanged across styles: the model always receives the degraded, viewpoint-shifted 3DGS rendering together with its opacity and HD-map conditioning.

We initialize from the same official OmniDreams causal checkpoint released prior to DMD distillation (NVIDIA et al., 2026) and post-train OneFixer on an equal mixture of identity, night, fog, and snow targets. The requested appearance is provided through text conditioning, while training otherwise follows the same shared-rollout objective used by the clear-weather model, including self-rollout flow matching and direct perceptual supervision. Consequently, the model is not trained as

Table 5: Style conditioning.

Model	Style		Geometry	
	CLIP acc ↑	VLM acc ↑	Lane F1 ↑	Lane x-err ↓
Cosmos3	<u>0.962</u>	<u>0.992</u>	0.098	0.632
Qwen-Video-Edit	**1.000**	**1.000**	<u>0.333</u>	<u>0.627</u>
OneFixer	0.918	0.956	**0.590**	**0.219**

a conventional video stylizer that operates on an already clean input. Instead, it must jointly interpret the structure implied by an imperfect 3DGS rendering, correct reconstruction artifacts, and synthesize the requested appearance in a single end-to-end refinement process.

We compare this capability with general-purpose video editing models, including Cosmos 3 (NVIDIA, 2026) and Qwen-Video-Edit (Bai et al., 2026). While these models can produce strong appearance changes, they do not consistently resolve reconstruction errors inherited from the underlying 3DGS rendering. OneFixer instead combines structural restoration and appearance conditioning within the same generative process, enabling the refined output to remain consistent with the inferred scene structure while adapting its illumination and weather appearance. We evaluate target-style alignment and geometric fidelity to the corresponding clean condition using complementary vision-language and recognition-based metrics, aggregated across frames and scenes (Table 5). These results suggest that a 3DGS fixer can serve not only as a restoration module but also as an end-to-end interface for controllable rendering enhancement. Further improving fine-grained appearance control through broader and more diverse training data remains an interesting direction for future work.

E Scaling to Multi-Camera and High-Resolution Refinement

Multi-camera refinement. We extend OneFixer to synchronized multi-camera generation by adopting the multi-view formulation of OmniDreams (NVIDIA et al., 2026). Joint processing across cameras substantially increases the latent sequence length and memory footprint. To make training feasible, we reduce the latent sequence length used per training iteration by half and compensate with a longer optimization schedule. Despite this reduced training horizon, the resulting model can jointly refine multiple views while preserving cross-view scene structure. As shown in Fig. 12, OneFixer removes prominent reconstruction artifacts across views, while LiDAR points projected from neighboring cameras remain well aligned with the refined images. The resulting views also produce a coherent bird's-eye-view composition, suggesting that refinement does not simply improve each camera independently but preserves the shared scene geometry across viewpoints.

High-resolution refinement. We further study whether the same one-step refinement formulation can scale beyond the resolution used in our main experiments. High-resolution training places substantially greater memory pressure on both the autoregressive rollout and the perceptual branch. We therefore use the same memory-efficient strategy of shortening the self-rollout segment during training, allowing OneFixer to operate at up to 1920-pixel width. Fig. 13 compares refinement at 960- and 1920-pixel widths. The higher-resolution model retains the overall restoration behavior of OneFixer while recovering finer scene details, such as distant road signs and small structural elements. These experiments indicate that the proposed refinement framework can be extended to more demanding spatial and multi-view settings, although doing so requires a shorter rollout during training and therefore reduces the amount of temporal context that can be processed in parallel.

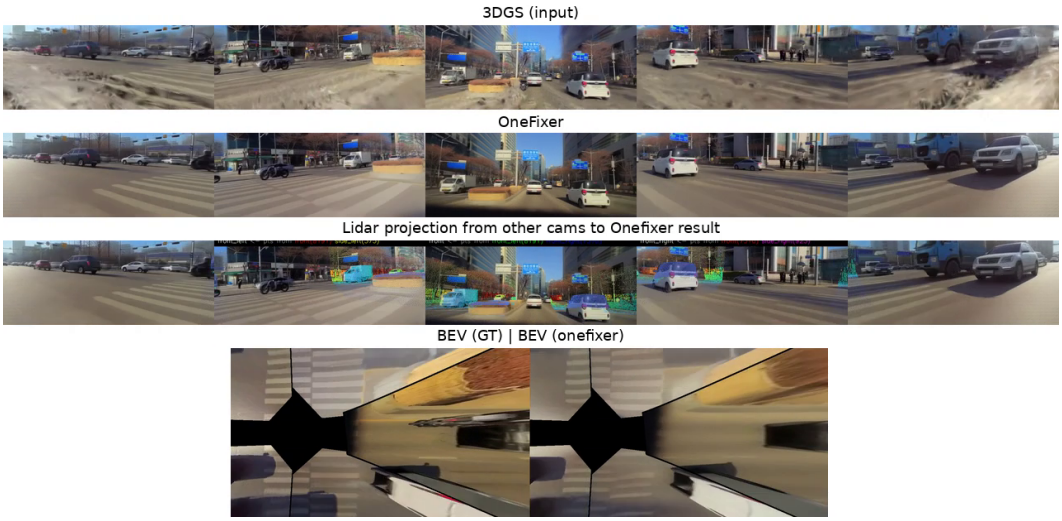

Figure 12: **Multi-camera 3DGS refinement.** From top to bottom: degraded multi-camera 3DGS renderings, OneFixer outputs, LiDAR points from neighboring cameras projected onto the refined views, and bird's-eye-view compositions from the ground truth and OneFixer outputs. OneFixer jointly removes reconstruction artifacts across views while preserving cross-view scene structure. The agreement of projected LiDAR points with the refined images and the resulting BEV composition provides qualitative evidence that refinement remains consistent with the shared underlying scene geometry.

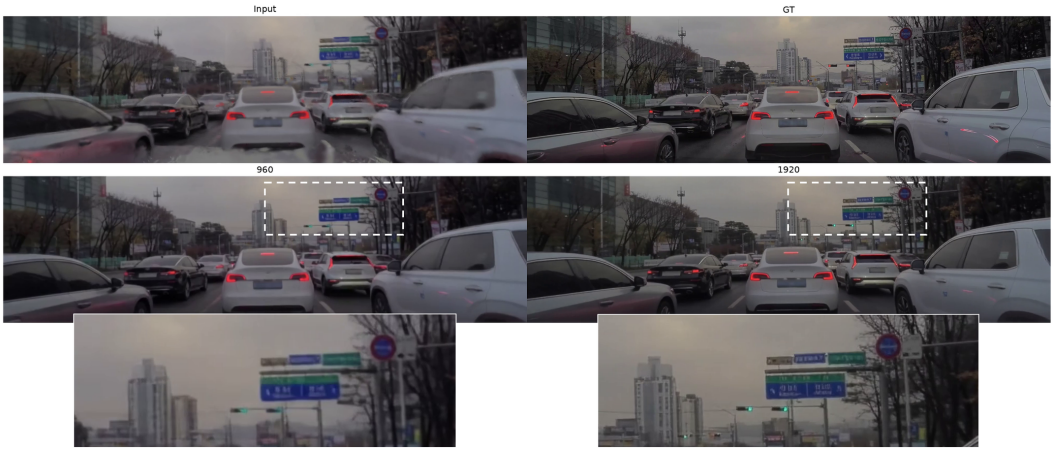

Figure 13: **High-resolution 3DGS refinement.** Top: degraded 3DGS input and clean reference. Bottom: OneFixer outputs at 960- and 1920-pixel widths, with enlarged regions highlighting distant scene details. Scaling to higher resolution preserves the overall refinement behavior while improving the recovery of fine structures such as road signs and other small visual details. High-resolution training is enabled by shortening the self-rollout segment during training to reduce memory consumption.