CONFIXGS: LEARNING TO FIX FEEDFORWARD 3D GAUSSIAN SPLATTING WITH CONFIDENCE-AWARE DIFFUSION PRIORS IN DRIVING SCENES

Anonymous authors
Paper under double-blind review

ABSTRACT

Feedforward 3D Gaussian Splatting (3DGS) often struggles in trajectory-based sparse-view driving scenes. Existing Gaussian repair methods mainly target optimization-based 3DGS, while diffusion-based repair is typically restricted to iterative refinement near observed viewpoints, leaving feedforward 3DGS repair underexplored. We propose **ConFixGS**, a plug-and-play method that learns to fix feedforward 3DGS with confidence-aware diffusion priors. Starting from a pretrained feedforward model, ConFixGS generates diffusion-enhanced local pseudo-targets and validates them through reprojection-based cross-checking against support views. The resulting dense confidence maps guide refinement, enhancing reliable details while suppressing hallucinated or inconsistent evidence. On Waymo, nuScenes, and KITTI, ConFixGS improves challenging novel view synthesis, with PSNR gains of up to 3.68 dB and FID reduced by nearly half. Our results highlight confidence-aware fusion of generative priors and support-view consistency as a key principle for robust feedforward 3D scene reconstruction.

1 INTRODUCTION

3D reconstruction of driving scenes is essential for autonomous driving development and simulation, such as novel view synthesis and scenario replay (Zhou et al., 2025a; Cao et al., 2025; Zhou et al., 2026a; Zhao et al., 2026; Li et al., 2025; Zhou et al., 2026b; Wang et al., 2025a). While vehicle-mounted cameras provide scalable real-world observations, they are typically collected along forward-moving trajectories, leading to sparse, anisotropic views with limited overlap, wide baselines, and severe occlusions. 3D Gaussian Splatting (3DGS) enables real-time high-fidelity rendering (Kerbl et al., 2023), but conventional 3DGS requires costly per-scene optimization and is difficult to scale. Feedforward 3DGS addresses this limitation by predicting Gaussian primitives from sparse input images in a single forward pass (Charatan et al., 2024; Chen et al., 2024b; Wang et al., 2024b; Hong et al., 2025; Ye et al., 2024; Wewer et al., 2024; Smart et al., 2024; Ye et al., 2025), making it appealing for driving-scene modeling (Hou et al., 2026; Song et al., 2026; Lin et al., 2025; Han et al., 2025). However, in trajectory-based sparse-view settings, feedforward models often produce blurred objects, low contrast, and texture smearing.

Our key observation is that feedforward 3DGS exhibits different behavior at varying trajectory scales. When applied to a small set of nearby, highly overlapping views, it often produces sharper and more reliable local novel views than when it is forced to reconstruct a large scene from many sparsely connected trajectory images, and applying a 2D diffusion prior on top of these local renderings produces the repairable pseudo-targets (Fig. 4 on Waymo, nuScenes, and KITTI). Although such local reconstructions are incomplete, they provide useful intermediate renderings that can serve as repair candidates for a global representation. This suggests a natural strategy: decompose a sparse driving trajectory into local overlapping subsets, generate local pseudo views, enhance them with strong 2D priors, and use them to refine the initial global feedforward reconstruction.

Recent diffusion-assisted 3DGS repair methods show that 2D generative priors can enhance degraded renderings and distill high-frequency details back into 3D representations (Wu et al., 2025; Liu et al., 2024; Wang et al., 2025b; Wei et al., 2025; Yin et al., 2025; Paliwal et al., 2025; Yu et al., 2024a). However, these methods are mostly designed for optimization-based 3DGS or rely on iter-

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: **ConFixGS: a plug-and-play repair of feedforward 3DGS in sparse driving scenes.** *Left:* Our method enhances state-of-the-art feedforward backbones, yielding better novel view rendering and consistent gains across metrics. *Right:* The repaired 3D Gaussians generalize well beyond the original camera trajectory, supporting novel view synthesis under large lateral offsets and elevated drone-like viewpoint settings where unrepaired feedforward 3DGS typically fails. More results are shown in Sec. E and the supplementary videos.

ative repair near the original viewpoints, limiting their deployment in a feedforward setting. More critically, directly optimizing a global 3DGS from independently diffusion-enhanced local views is ill-posed: diffusion outputs are stochastic and not inherently multi-view consistent. Treating all pseudo views as equally reliable can introduce hallucinated textures, ambiguous geometry, ghosting, and spurious Gaussians. Thus, the central challenge is not whether diffusion priors are useful, but how to decide which generated details are confident enough to supervise feedforward 3DGS repair.

In this paper, we take a different view: *diffusion-enhanced pseudo views should not be blindly distilled into 3D, and their confidence should be explicitly estimated through the geometry of the feedforward reconstruction itself.* Inspired by reprojection-based consistency reasoning in controllable novel view and video synthesis (Yan et al., 2025; Seo et al., 2024; Müller et al., 2024; Yang et al., 2024c; Fan et al., 2025; Wang et al., 2024a), we propose **ConFixGS**, a confidence-guided approach for repairing feedforward 3DGS in sparse driving scenes. Starting from an initial feedforward reconstruction, ConFixGS builds local feedforward reconstructions from overlapping view subsets and enhances their rendered novel views with a 2D diffusion prior. Each enhanced view is treated as a candidate pseudo-target rather than direct ground truth. To estimate its reliability, we unproject pseudo-target pixels using depth rendered from the initial reconstruction, reproject them into nearby real support views, and compare the reprojected appearance with observed images. This step produces a dense confidence map that identifies which generated details are supported by multi-view evidence and which regions are uncertain or likely hallucinated.

The resulting confidence maps guide a global 3DGS repair initialized from the original feedforward prediction: reliable diffusion details are injected into the Gaussian representation, while uncertain regions are down-weighted to avoid corrupting geometry or spawning spurious Gaussians. Since ConFixGS repairs rather than reconstructs from scratch, it requires fewer than 1K optimization iterations, compared with the 30K–50K iterations commonly used by optimization-based driving-scene 3DGS pipelines. Experiments on Waymo, nuScenes, and KITTI show consistent improvements over strong feedforward baselines, and extend to challenging novel view settings with large lateral camera offsets and elevated drone-like viewpoints, as shown in Fig. 1. Our contributions are fourfold:

- We identify a scale-induced failure mode of feedforward 3DGS in driving scenes: with long observation sequences, global processing can oversmooth local details, whereas processing nearby subsets with the same backbone preserves sharper and more repairable evidence.
- We propose **ConFixGS**, the first plug-and-play repair approach for feedforward 3DGS without backbone retraining, reducing artifacts, and yielding high-quality novel views.

- We estimate continuous $[0, 1]$ confidence maps for diffusion-enhanced local pseudo-targets via support-view reprojection, separating reliable generative details from hallucinations.
- We use these confidence maps to guide global Gaussian repair, modulating both photometric supervision and densification to improve large-offset and elevated-view rendering.

2 Related Work

3D Reconstruction for Driving Scenes. Driving-scene reconstruction is central to autonomous driving, supporting realistic simulation, controllable data generation, and evaluation under rare trajectories and viewpoints. Although neural radiance fields (NeRFs) and their unbounded-scene variants established high-quality view synthesis as a reconstruction objective (Mildenhall et al., 2021; Barron et al., 2021; Zhang et al., 2020; Barron et al., 2022), 3DGS represents scenes with explicit anisotropic Gaussian primitives and enables real-time rendering via efficient rasterization (Kerbl et al., 2023). Follow-up 3DGS methods improve anti-aliasing, geometry, surfaces, propagation, resource usage, and sparse-view robustness (Yu et al., 2024b; Lu et al., 2024; Huang et al., 2024; Chen et al., 2024a; Guédon & Lepetit, 2024; Cheng et al., 2024; Mallick et al., 2024; Li et al., 2024c; Turkulainen et al., 2025; Chung et al., 2024; Li et al., 2024a; Huang et al., 2025a; Zhang et al., 2025b; Huang et al., 2025b; Fang et al., 2025). For outdoor and autonomous-driving scenes, prior work further leverages street-view assumptions, dynamic decomposition, laser scanning or depth priors, object structure, and large-scale spatio-temporal modeling (Guo et al., 2023; Zhou et al., 2024; Yan et al., 2024; Jiang et al., 2024; Shen et al., 2025; Lee et al., 2025; Xia et al., 2025; Cui et al., 2025; Miao et al., 2025; Hess et al., 2025; Wu et al., 2024; Xu et al., 2025b; Chen et al., 2026; Yang et al., 2024a). However, driving data is inherently trajectory-based, featuring sparse forward-moving viewpoints, wide baselines, limited overlap, and large unobserved regions. These properties make per-scene optimization slow and fragile, especially for novel views far from the original vehicle trajectory.

Feedforward 3DGS. To avoid costly per-scene optimization, recent work has shifted toward feedforward reconstruction, where large networks regress Gaussian parameters or related scene representations from a few input images. Existing feedforward 3DGS or Gaussian reconstruction models target image pairs, sparse multi-view inputs, single-view or object-centric generation, free-view synthesis, pose-free inputs, and large reconstruction models (Charatan et al., 2024; Chen et al., 2024b;c; Wang et al., 2024b; Zhang et al., 2025a; Zou et al., 2024; Tang et al., 2024; Zhang et al., 2024; Hong et al., 2025; 2024). Recent systems further improve depth coupling, unconstrained-view generalization, universal world reconstruction, and driving-specific surround-view reconstruction (Xu et al., 2025a; Jiang et al., 2025; Liu et al., 2025; Tian et al., 2025). While feedforward methods are appealing for their single-pass global scene estimate, they lack a native per-scene optimization loop. Consequently, once a pretrained predictor outputs an imperfect scene, it cannot repair missing structures, gray uncertain regions, or over-smoothed textures. Adding more trajectory views does not necessarily solve this issue, as weak overlap and fixed network capacity complicate global fusion. We therefore utilize feedforward prediction as an initial reconstruction, focusing on how to repair it with local generative evidence without compromising global consistency.

Diffusion Priors for Novel View Synthesis. Diffusion models provide strong 2D generative priors for hallucinating plausible details in unobserved regions (Rombach et al., 2022; Brooks et al., 2023). This has inspired sparse-view novel view synthesis and 3D generation methods that synthesize additional viewpoints or view-consistent image sets (Liu et al., 2023a;b; Shi et al., 2024), or distill 2D diffusion priors into 3D representations via score distillation, 3D-aware diffusion, iterative editing, and optimization (Alldieck et al., 2024; Haque et al., 2023; Gu et al., 2023; Karnewar et al., 2023). Closest to our setting, recent methods enhance low-quality 3DGS renderings with view-consistent, video, single-step, or score-distillation diffusion priors, followed by 3D optimization or fine-tuning (Liu et al., 2024; Wang et al., 2025b; Wu et al., 2025; Wei et al., 2025). While these approaches successfully recover high-frequency details and fill missing content, they are primarily designed for optimization-based representations. To maintain stability, existing methods often restrict diffusion-based repair to views very close to the observations, severely limiting their repair range in sparse driving trajectories. Additionally, generic 2D diffusion does not guarantee 3D consistency across independently generated frames. Disagreeing pseudo targets can make 3D optimization average incompatible evidence or create spurious structures, causing blur, ghosting, and repeated artifacts. A very recent feedforward method, DGGT (Chen et al., 2025), also incorporates

diffusion into its architecture, but uses it mainly for 2D post-processing of rendered images instead of repairing the 3D Gaussians.

Controllable Novel View Video Synthesis and Reprojection Cues. Recent controllable novel view video synthesis methods show that camera trajectories can be redirected by combining generative video priors with geometric reprojection cues (Yu et al., 2025). Related view-consistent generation work also studies how to identify low-certainty or inconsistent generated content under multi-view constraints (Tao et al., 2024). These results suggest that unprojection and reprojection can indicate which generated content is supported by source observations. Unlike methods that directly produce a 2D video, such as (Chen et al., 2025), we use reprojection to evaluate diffusion-enhanced local pseudo targets before they supervise an explicit 3D Gaussian scene. This converts local generative outputs into confidence-weighted 3D training signals, retaining diffusion flexibility while avoiding unconditional reliance on hallucinated regions.

Robust Optimization and Confidence Estimation. Multi-view consistency has long been used in classical multi-view geometry and MVS to detect unreliable observations, where reprojection and photometric errors help handle occlusions, outliers, and non-Lambertian effects (Schönberger et al., 2016). Similar ideas appear in robust losses, robust NeRFs, and reconstruction pipelines that downweight inconsistent pixels or views during optimization (Barron, 2019; Li et al., 2024b). Confidence-weighted learning is also widely used in self-training and pseudo-labeling, including heteroscedastic confidence formulations (Kendall & Gal, 2017). Recent 3DGS methods further introduce confidence-aware or depth, normal, or feature consistency regularization to stabilize geometry under sparse observations (Chung et al., 2024; Turkulainen et al., 2025; Li et al., 2024a; Huang et al., 2025a; Zhang et al., 2025b; Huang et al., 2025b). Existing methods usually reweight photometric objectives or mask training signals, but rarely intervene in the structural update rules of explicit 3D representations under generative pseudo supervision. We adapt these principles to diffusion-assisted feedforward repair. Instead of treating generated views as ground truth, we evaluate diffusion-enhanced pseudo-targets via deterministic reprojection-based cross-validation against support images, which yields a dense confidence map. By learning propagating this confidence to reweight the photometric loss and modulate both gradients and densification during 3DGS refinement, ConFixGS suppresses unreliable generative residuals from spawning spurious Gaussians while empowering local details to improve the feedforward results.

3 METHODOLOGY

3.1 FEEDFORWARD REPAIR PROBLEM

Background. 3DGS (Kerbl et al., 2023) represents a scene with a set of anisotropic Gaussians (mean μ, covariance Σ, color c, opacity α), renders pixel color via depth-sorted α-blending, and updates topology through densification and pruning driven by accumulated 2D position gradients (we follow standard formulations and detail them in Appendix B). Pre-trained image-to-image diffusion models $\mathcal{F}_{\text{diff}}$ further provide generative priors that map a low-quality rendering $\hat{I}$ to a high-fidelity target $I^* = \mathcal{F}_{\text{diff}}(\hat{I}, c_{\text{cond}})$, where c_{cond} denotes optional conditioning.

Problem Setting. Given a sparse set of support observations $\mathcal{S}$ with RGB images $\{I_s^{\text{obs}}\}_{s\in\mathcal{S}}$ and cameras $\{(K_v, T_v)\}_{v=1}^{V}$, our goal is to optimize a 3DGS scene Θ that faithfully reconstructs the support views while rendering high-fidelity, geometrically consistent novel views.

Why Feedforward Repair Is Different. A feedforward 3DGS backbone produces an initial scene in a single forward pass. Since this prediction is a deterministic output of a pre-trained network, simply re-running the model cannot fix its errors. Repair must instead operate downstream in the Gaussian parameter space, making novel view supervision the main available lever. A natural strategy is to generate diffusion-enhanced pseudo-targets I_n^* for novel views $n \notin \mathcal{S}$ using $\mathcal{F}_{\text{diff}}$, and optimize Θ by matching renderings $\hat{C}_n$ to them. However, independently generated diffusion targets can contain multi-view inconsistent hallucinations. Directly minimizing $\mathcal{L}(\hat{C}_n, I_n^*)$ may therefore corrupt the 3D representation, causing warped geometry and floating artifacts under conflicting 2D supervision. ConFixGS thus treats diffusion outputs as candidate evidence rather than ground truth.

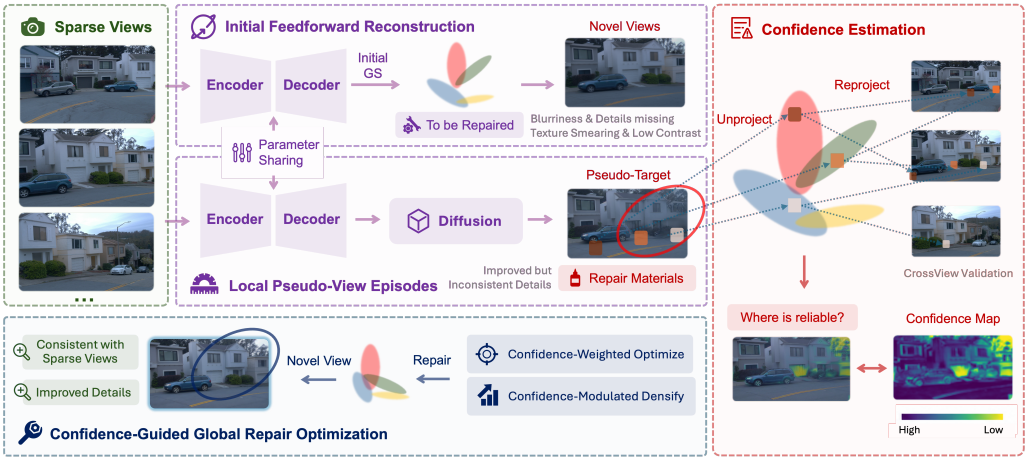

Figure 2: **ConFixGS Framework Overview.** ConFixGS consists of three stages: (*i*) initial feedforward reconstruction with local pseudo-view episodes, (*ii*) input-observation-guided confidence estimation through reprojection, and (*iii*) confidence-modulated global repair optimization.

3.2 ConFixGS Overview

ConFixGS repairs a feedforward reconstruction by asking which locally generated details can be supported by the input observations (Fig. 2). The pipeline consists of three stages: (*i*) take the original feedforward 3DGS prediction $\mathcal{G}_0$ as the initial reconstruction to be repaired and construct detail-rich local pseudo-view episodes using local feedforward rendering and diffusion; (*ii*) compute a dense confidence map W_n for each pseudo view by reprojecting generated pixels into the captured support observations, using $\mathcal{G}_0$ purely as a geometric scaffold for establishing correspondences; (*iii*) refine the global 3D Gaussians with a confidence-weighted objective, where low-confidence pixels are suppressed in both photometric gradients and standard 3DGS densification signals.

3.3 Initial Feedforward Reconstruction and Local Pseudo-View Episodes

Initial Feedforward Reconstruction. We employ a pre-trained feedforward 3DGS backbone F_{global} to produce the initial reconstruction from sparse support views:

$$\mathcal{G}_0 = F_{\text{global}}(\{I_s^{\text{obs}}\}_{s\in\mathcal{S}}). \tag{1}$$

The predicted scene is a set of N_0 Gaussians $\mathcal{G}_0 = \{(\mu_i, s_i, q_i, o_i, c_i)\}_{i=1}^{N_0}$, where $\mu_i \in \mathbb{R}^3$ is the 3D mean, $s_i \in \mathbb{R}^3_+$ is the per-axis scale, $q_i \in \mathbb{R}^4$ is the rotation quaternion, $o_i \in [0,1]$ is the opacity, and $c_i \in \mathbb{R}^3$ is the RGB color. During the stage in Sec. 3.5 we adopt the standard logit/log reparameterization $\hat{o}_i = \sigma^{-1}(o_i)$ and $\hat{s}_i = \log s_i$ so that activated values stay in their valid ranges throughout training. This is the feedforward result we aim to repair: it can be over-smoothed or incomplete under sparse driving views, but it provides a useful initial state for later global refinement. Rendering $\mathcal{G}_0$ at any viewpoint v yields a base RGB image $\hat{I}_v^{(0)}$, an accumulated opacity map $\hat{\alpha}_v^{(0)}$, and a depth proxy $d_v^{(0)}$. The base RGB is consumed by the local pseudo-view branch below, while the opacity and depth are reused as a geometric scaffold for confidence estimation in Sec. 3.4.

Local Pseudo-View Episodes. For each novel view $n \notin \mathcal{S}$, we build a local episode around nearby trajectory observations and generate a pseudo-target I_n^* by enhancing a local feedforward rendering conditioned on temporal context I_n^{ctx} and the base render $\hat{I}_n^{(0)}$. This local formulation preserves the empirical advantage of small, highly overlapping inputs: local predictions may lack full-scene coverage, but their nearby novel views are often cleaner and easier to repair than a single global feedforward output. Following Wu et al. (2025) and Chen et al. (2025), we use SD-Turbo as the generative prior:

$$I_n^* = \mathcal{F}_{\text{SD-Turbo}}\Big(F_{\text{local}}(I_n^{\text{ctx}}, \hat{I}_n^{(0)})\Big), \quad \forall n \notin \mathcal{S}. \tag{2}$$

For support views, we preserve absolute fidelity by setting $I_s^* = I_s^{\text{obs}}$, yielding a complete target set $\{I_v^*\}_{v=1}^{V}$.

3.4 Confidence Estimation via Input-Observation Reprojection

Novel-view pseudo-targets may include diffusion hallucinations. We compute a reprojection-based confidence score from the intuition that a reliable pseudo-target pixel should be photo-consistent with the captured support observations supplied to the global feedforward model. The decision of whether a pseudo-target pixel is credible is grounded in this observed RGB evidence, and the initial reconstruction $\mathcal{G}_0$ enters only as a geometric scaffold, where the rendered depth $d_n^{(0)}$ provides a cheap proxy for establishing correspondences, and its accumulated opacity $\hat{\alpha}_n^{(0)}$ flags regions where this scaffold is too sparse to support reliable correspondence.

Cross-View Unprojection. For a pixel $p = (u_p, v_p)$ in a novel view n, we unproject it into 3D using the depth proxy $d_n^{(0)}(p)$:

$$X(p) = R_n \left(d_n^{(0)}(p)\, K_n^{-1} \tilde{p} \right) + t_n, \tag{3}$$

where $\tilde{p} = [u_p, v_p, 1]^\top$ and (R_n, t_n) is the camera-to-world transformation. We reproject $X(p)$ to a support view $s \in \mathcal{S}$:

$$p_s = \mathrm{dehom}\left(K_s R_s^\top \left(X(p) - t_s \right) \right). \tag{4}$$

Consensus Color and Discrepancy. Let $\mathcal{V}(p) \subset \mathcal{S}$ be the set of support views where the reprojection is valid (positive depth, within image bounds). We sample the support images and compute a reprojection consensus color:

$$\bar{I}_{\text{reproj}}(p) = \frac{1}{|\mathcal{V}(p)| + \epsilon} \sum_{s \in \mathcal{V}(p)} \mathrm{Sample}(I_s^{\text{obs}}, p_s). \tag{5}$$

We define the discrepancy between the pseudo-target and the consensus as

$$e_n(p) = \frac{1}{3} \sum_{c \in \{R,G,B\}} \left| I_n^*(p,c) - \bar{I}_{\text{reproj}}(p,c) \right|. \tag{6}$$

Support-Validated Confidence. We define a support-validated pixel-wise confidence score $\tilde{w}_n(p)$:

$$\tilde{w}_n(p) = \begin{cases} \exp\left(-\frac{e_n(p)^2}{2\sigma_e^2} \right), & \text{if } \mathcal{V}(p) \neq \emptyset \text{ and } \hat{\alpha}_n^{(0)}(p) \geq \alpha_{\min}, \\ \gamma, & \text{if } \mathcal{V}(p) = \emptyset \text{ and } \hat{\alpha}_n^{(0)}(p) \geq \alpha_{\min}, \\ 0, & \text{if } \hat{\alpha}_n^{(0)}(p) < \alpha_{\min}, \end{cases} \tag{7}$$

where σ_e controls the confidence-decay bandwidth, γ is a conservative baseline confidence score assigned to regions whose correspondences are valid in geometry but cannot be photometrically validated by any support image, and the third branch zeros out pixels whose underlying scaffold $\mathcal{G}_0$ has no opacity coverage: there the depth proxy $d_n^{(0)}$ is unreliable, so the reprojection itself, and any photometric comparison built on it, is meaningless. We then apply spatial smoothing to obtain the final *per-pixel confidence score*:

$$w_n(p) = (G_k * \tilde{w}_n)(p) \ \in [0,1], \qquad W_n \in [0,1]^{H \times W}, \quad (W_n)_p = w_n(p), \tag{8}$$

where G_k is a $k \times k$ mean filter. We refer to the resulting continuous-valued field W_n as the *confidence map* (in contrast to a binary segmentation mask), and to the scalar $w_n(p)$ as the per-pixel confidence score. For support views, we set the confidence map to all-ones:

$$W_s = \mathbf{1}, \qquad \forall s \in \mathcal{S}. \tag{9}$$

3.5 Confidence-Guided Global Repair Optimization

With support-validated confidence maps $\{W_v\}_{v=1}^V$ and fixed targets $\{I_v^*\}_{v=1}^V$, we initialize $\Theta \leftarrow \mathcal{G}_0$ and refine 3DGS by optimizing a confidence-weighted objective. This stage is a lightweight repair step rather than a full reconstruction from scratch: the original feedforward prediction supplies the initial state, and the optimization selectively assimilates only reliable local pseudo-view evidence.

Confidence-Weighted Objective. For a rendered image $\hat{C}_v$ and its target I_v^*, we define a confidence-weighted L_1 loss:

$$\mathcal{L}_{\text{tgt}}^{L_1}(v) = \frac{\sum_p w_v(p) \, \|\hat{C}_v(p) - I_v^*(p)\|_1}{\sum_p w_v(p) + \epsilon}. \tag{10}$$

To apply pixel-wise weighting within SSIM while preserving local normalization, we use a blending approximation:

$$\tilde{C}_v = w_v \odot \hat{C}_v + (1 - w_v) \odot I_v^*, \qquad \mathcal{L}_{\text{tgt}}^{\text{SSIM}}(v) = 1 - \text{SSIM}(\tilde{C}_v, I_v^*). \tag{11}$$

The total target loss is

$$\mathcal{L}_{\text{tgt}}(v) = (1 - \lambda_s)\mathcal{L}_{\text{tgt}}^{L_1}(v) + \lambda_s \mathcal{L}_{\text{tgt}}^{\text{SSIM}}(v). \tag{12}$$

For support views, we apply an anchoring loss $\mathcal{L}_{\text{sup}}(s)$ with weight λ_{sup} to prevent drift.

Confidence-Modulated Densification. We extend the standard 3DGS topology update by using the per-pixel confidence score to gate densification. Specifically, the normalized confidence weight $\lambda_v(p)$ modulates both the photometric gradient and the accumulated 2D position-gradient signal used for clone/split decisions. Low-confidence pixels are therefore prevented from driving either parameter updates or densification, while standard clone/split/prune thresholds remain unchanged. Details are provided in Appendix B, with full pseudocode in Algorithm 1.

4 Experiments

4.1 Setup

Dataset. We evaluate our framework on the Waymo Open Dataset (Mei et al., 2022), nuScenes (Caesar et al., 2020), and KITTI (Voigtlaender et al., 2019). To evaluate under sparse-view conditions, we use every 5th view as input, and the task is to synthesize the 4 intermediate views. Each scene contains 50 frames, unless the maximum sequence length is shorter.

Metrics. To assess photometric reconstruction quality, we report Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) (Wang et al., 2004), and Learned Perceptual Image Patch Similarity (LPIPS) (Dosovitskiy & Brox, 2016). For perceptual and temporal consistency, we report Fréchet Inception Distance (FID) (Heusel et al., 2017) and Fréchet Video Distance (FVD) (Unterthiner et al., 2018).

Baselines. We benchmark against state-of-the-art feedforward 3DGS methods, including pixelSplat (Charatan et al., 2024), MVSplat (Chen et al., 2024b), TranSplat (Zhang et al., 2025a)), and driving-specific models DrivingForward (Tian et al., 2025). To demonstrate our method's versatility as a plug-and-play enhancement module, we integrate it atop two recent feedforward architectures: DepthSplat (Xu et al., 2025a) and WorldMirror (Liu et al., 2025). We also apply a diffusion model to further refine the novel view renderings, denoting this enhanced variant as ConFixGS+. Each scene is repaired with only K=1,000 Adam iterations on top of the initial feedforward prediction. The full set of hyperparameters, learning rates, and densification schedule is reported in Appendix A. All datasets share one hyperparameter setting. The complexity and efficiency analysis is provided in Appendix C.

4.2 Quantitative Results

Photometric Reconstruction Quality. As shown in Tab. 1, ConFixGS consistently improves existing feedforward architectures. ConFixGS-DS and ConFixGS-WM substantially outperform their baselines, DepthSplat and WorldMirror, with PSNR gains of up to **3.30 dB** and **3.68 dB** on Waymo, respectively. ConFixGS also improves cross-dataset generalization: on KITTI, ConFixGS-DS raises DepthSplat from below AnySplat to above it, achieving 14.48 dB better than 14.08 dB PSNR.

Perceptual and Temporal Consistency. Tab. 2 evaluates perceptual quality with FID and FVD. Although diffusion-based post-processing slightly degrades standard photometric metrics in Tab. 1, it substantially improves perceptual realism and temporal consistency. This reflects the perception-distortion trade-off, where pixel-level metrics penalize plausible high-frequency details introduced

Table 1: **Quantitative evaluation on the Waymo, nuScenes, and KITTI datasets.** The best, second-best, and third-best results are highlighted in **bold**, underline, and *italics*, respectively. This will be used for the following tables.

Method	Waymo			nuScenes			KITTI		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
pixelSplat (Charatan et al., 2024)	20.20	0.6043	0.2818	16.23	0.3018	0.5267	12.70	0.2456	0.5004
MVSplat (Chen et al., 2024b)	13.64	0.3849	0.6478	12.66	0.2072	0.6643	7.14	0.0467	0.6980
TranSplat (Zhang et al., 2025a)	15.34	0.4605	0.5125	15.17	0.3144	0.6126	8.08	0.1138	0.6922
DepthSplat (Xu et al., 2025a)	*24.56*	*0.7990*	*0.2550*	*16.96*	*0.5057*	*0.5224*	12.52	*0.3190*	0.5863
DrivingForward (Tian et al., 2025)	20.63	0.6765	0.3174	14.24	0.3422	0.5354	8.71	0.2644	0.5914
WorldMirror (Liu et al., 2025)	20.38	0.6000	0.4949	15.07	0.3621	0.6136	9.13	0.1855	0.6652
AnySplat (Jiang et al., 2025)	20.49	0.6683	0.3159	15.85	0.4486	0.5384	*14.08*	0.3117	**0.4719**
ConFixGS-WM	24.06	0.7193	0.4176	16.60	0.3448	0.5679	12.27	0.2056	0.6323
ConFixGS-WM+	22.88	0.6808	0.3631	16.10	0.3125	0.5400	12.31	0.1997	0.5891
ConFixGS-DS	**27.86**	**0.8670**	**0.2060**	**19.37**	**0.5469**	0.4585	**14.48**	**0.3456**	0.5475
ConFixGS-DS+	25.39	0.8138	0.2540	19.05	0.5209	**0.4451**	14.20	0.3259	*0.5332*

Table 2: **Quantitative evaluation of perceptual and temporal consistency** across the Waymo, nuScenes, and KITTI datasets.

Method	Waymo		nuScenes		KITTI	
	FID ↓	FVD ↓	FID ↓	FVD ↓	FID ↓	FVD ↓
WorldMirror	185.3	90.3	324.0	160.2	327.1	249.5
DepthSplat	68.60	*24.38*	297.99	178.02	308.74	231.64
ConFixGS-WM	92.3	36.4	191.6	*67.5*	298.6	232.9
ConFixGS-WM+	*57.2*	25.9	**116.8**	**50.9**	168.3	128.4
ConFixGS-DS	50.02	17.26	*168.37*	70.04	*285.02*	*197.17*
ConFixGS-DS+	**48.46**	**14.96**	119.15	53.00	**151.18**	**110.26**

Table 3: **Comparison against Difix3D+** (Wu et al., 2025) on the Waymo dataset.

Method	PSNR↑	SSIM↑	LPIPS↓	FID↓	FVD↓
Difix3D	*21.12*	**0.7817**	0.5065	269.64	169.23
Difix3D+	19.04	*0.6856*	0.4914	167.85	139.08
Difix3D*	20.25	0.6094	0.4790	*85.89*	47.86
Difix3D*+	19.46	0.5713	*0.4375*	79.80	*42.67*
ConFixGS	24.06	0.7193	0.4176	92.26	36.39
ConFixGS+	**22.88**	0.6808	**0.3631**	**57.20**	**25.88**

* denotes the Difix3D baseline is a modified feedforward version.

by diffusion, *e.g.*, ConFixGS-DS+ nearly halves the FID on KITTI compared with ConFixGS-DS (151.18 vs. 285.02) and achieves the best or second-best FVD scores across all benchmarks.

Comparison with Difix3D+. Tab. 3 compares ConFixGS with Difix3D+ (Wu et al., 2025) on Waymo. ConFixGS consistently outperforms the original Difix3D+ baseline across key metrics. To isolate the source of improvement, we further compare against a modified Difix3D+ using the same WorldMirror feedforward initialization as our pipeline. Even under this controlled setting, ConFixGS maintains a clear quantitative advantage, showing that the gains come from the proposed support-validated confidence module rather than a stronger geometric initialization.

4.3 Ablation Study

We evaluate three ablated variants on nuScenes in Tab. 4. *w/o Diffusion* extracts I^* directly from local or $\mathcal{G}_0$ renderings, retaining competitive PSNR (18.91 for ConFixGS+) but substantially increasing FID to 235.0, highlighting the perception-distortion trade-off. *w/o Confidence Guidance* assigns uniform weights to diffusion-enhanced pseudo-targets, reducing PSNR from 19.03 to 17.46 due to conflicting pseudo-supervision. *w/o Global Opt* removes the final 3DGS optimization and relies on independent local outputs, yielding the worst FVD of 159.3 and underscoring the importance of global optimization for temporal coherence. Dynamic-region, pose-noise, hyperparameter, and prior-corruption studies are reported in Appendix D.

4.4 Qualitative Analysis

Fig. 1 left shows ConFixGS with WorldMirror as the feedforward backbone. Compared with the original results, our method reduces blur, ghosting, and unstable textures, yielding clearer structures and more coherent appearance. The diffusion prior enhances high-frequency details in under-observed regions, while the local feedforward branch supplies complementary geometry and object cues through fine-grained local Gaussians. This helps recover missing or inaccurate content, such as an incomplete bicyclist or a waiting vehicle at a traffic light. More results are provided in Sec. E.

Fig. 3 compares ConFixGS with state-of-the-art baselines across multiple datasets and driving scenes. ConFixGS produces higher-quality novel view renderings with fewer artifacts, sharper de-

Table 4: **Ablation study on nuScenes.** We compare metrics additionally in ConFixGS+ modes.

Method	ConFixGS					ConFixGS+				
	PSNR↑	SSIM↑	LPIPS↓	FID↓	FVD↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	FVD↓
Full Method (Ours)	**19.03**	0.5252	**0.4602**	166.9	**65.4**	18.84	0.4934	0.4468	**112.3**	**45.4**
w/o Diffusion	18.86	**0.5407**	0.4846	*235.0*	87.6	**18.91**	**0.5087**	**0.4399**	121.7	47.6
w/o Confidence Guidance	*17.46*	0.4587	*0.4921*	**161.6**	*95.5*	*17.22*	0.4195	0.5013	*134.3*	*76.4*
w/o Global Opt	17.01	*0.4792*	0.5281	316.3	159.3	17.15	*0.4394*	*0.4947*	147.5	90.6

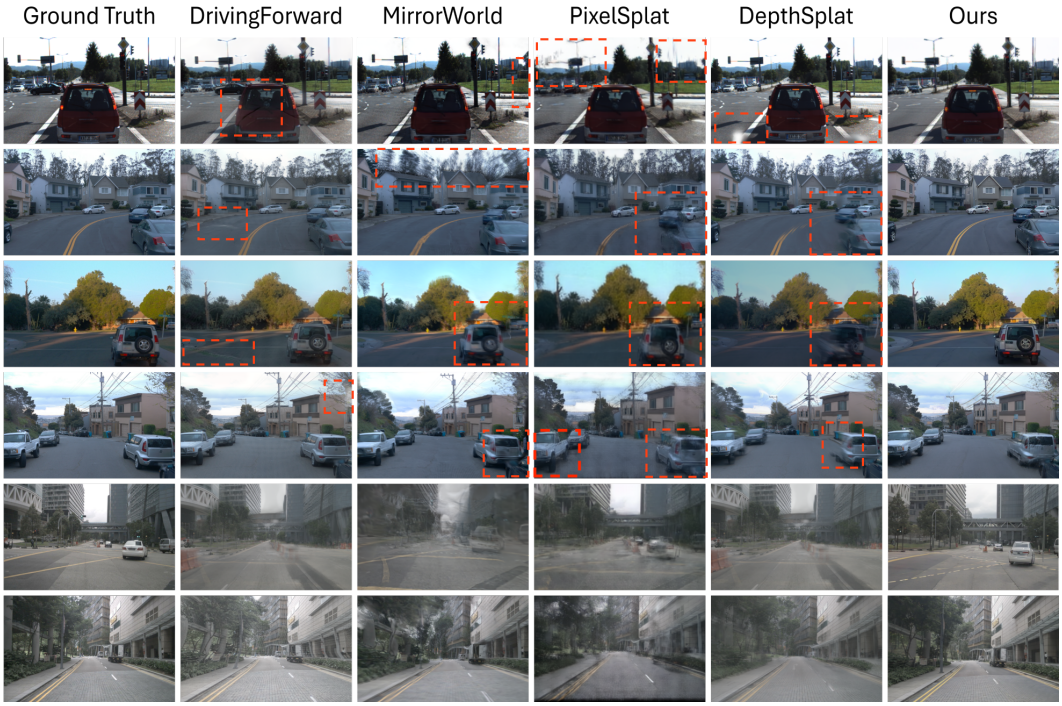

Figure 3: **Visual comparison.** We use DepthSplat as the feedforward backbone for our approach. More results are shown in Figs. 6 and 7, with additional backbone comparisons in Figs. 8, 9, and 10.

tails, better completeness, and stronger fidelity to the ground truth. It remains stable across diverse scene layouts and observation sparsity, avoiding common baseline failure modes.

5 Conclusion

In summary, we present ConFixGS, a confidence-guided repair approach specifically designed to fix feedforward 3DGS in sparse-view driving scenes. ConFixGS uses a training-free reprojection check to cross-validate diffusion-enhanced pseudo-targets against the captured sparse support images, producing dense per-pixel confidence maps without auxiliary predictors. These maps guide lightweight per-scene optimization by modulating photometric supervision and Gaussian densification, enabling the model to absorb reliable generative details while suppressing hallucinated or multi-view inconsistent evidence. More broadly, our results suggest that geometry-grounded confidence estimation offers a simple yet effective principle for routing generative priors into 3D Gaussian representations, without relying on iterative diffusion-optimization loops.

Limitations. ConFixGS neither estimates poses nor models motion, so such inconsistencies are gated conservatively and reduce repair gains (Appendix D). Extreme sparsity or corrupted priors can still degrade results. Temporal confidence and joint pose refinement remain future work.

ETHICS STATEMENT

This work uses only publicly released driving datasets (Waymo Open Dataset, nuScenes, and KITTI) under their respective licenses. No new data were collected and no human subjects were involved. The released code and videos contain only imagery from these datasets. We are not aware of ethical concerns specific to this work and declare no conflicts of interest.

REPRODUCIBILITY STATEMENT

Appendix A specifies the full evaluation protocol and hyperparameters of the three stages, which are shared across datasets. Appendix B gives the exact confidence-weighted objective and its gradients. Appendix C reports hardware and runtime. The supplementary material contains anonymized source code implementing all three stages, the configuration files with the exact hyperparameters used, and scripts that reproduce the main tables and figures, together with the qualitative videos referenced in Appendix E and supplementary material. All base models (DepthSplat, AnySplat, DrivingForward, STORM, and the diffusion prior) are used with their public checkpoints.

AI USE STATEMENT

In this work, we used generative AI tools only to polish the manuscript for readability and assist in editing the released code. The method, its core implementation, the experiments, and the analysis were developed and carried out by the authors without generative AI tools. All AI-assisted text and code were reviewed by the authors, and we take responsibility for the final content of this work.

REFERENCES

Thiemo Alldieck, Nikos Kolotouros, and Cristian Sminchisescu. Score distillation sampling with learned manifold corrective. In *European Conference on Computer Vision*, pp. 1–18. Springer, 2024.

Jonathan T Barron. A general and adaptive robust loss function. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4331–4339, 2019.

Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5855–5864, 2021.

Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5470–5479, 2022.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18392–18402, 2023.

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.

Wei Cao, Marcel Hallgarten, Tianyu Li, Daniel Dauner, Xunjiang Gu, Caojun Wang, Yakov Miron, Marco Aiello, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, Andreas Geiger, and Kashyap Chitta. Pseudo-simulation for autonomous driving. In *Conference on Robot Learning (CoRL)*, 2025.

David Charatan, Sizhe Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *CVPR*, 2024.

Danpeng Chen, Hai Li, Weicai Ye, Yifan Wang, Weijian Xie, Shangjin Zhai, Nan Wang, Haomin Liu, Hujun Bao, and Guofeng Zhang. Pgsr: Planar-based gaussian splatting for efficient and high-fidelity surface reconstruction. *IEEE Transactions on Visualization and Computer Graphics*, 2024a.

Xiaoxue Chen, Ziyi Xiong, Yuantao Chen, Gen Li, Nan Wang, Hongcheng Luo, Long Chen, Haiyang Sun, Bing Wang, Guang Chen, et al. Dggt: Feedforward 4d reconstruction of dynamic driving scenes using unposed images. *arXiv preprint arXiv:2512.03004*, 2025.

Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In *European conference on computer vision*, pp. 370–386. Springer, 2024b.

Yuedong Chen, Chuanxia Zheng, Haofei Xu, Bohan Zhuang, Andrea Vedaldi, Tat-Jen Cham, and Jianfei Cai. Mvsplat360: Feed-forward 360 scene synthesis from sparse views. *Advances in Neural Information Processing Systems*, 37:107064–107086, 2024c.

Yurui Chen, Chun Gu, Junzhe Jiang, Xiatian Zhu, and Li Zhang. Periodic vibration gaussian: Dynamic urban scene reconstruction and real-time rendering. *International Journal of Computer Vision*, 134(3):83, 2026.

Kai Cheng, Xiaoxiao Long, Kaizhi Yang, Yao Yao, Wei Yin, Yuexin Ma, Wenping Wang, and Xuejin Chen. Gaussianpro: 3d gaussian splatting with progressive propagation. In *Forty-first International Conference on Machine Learning*, 2024.

Jaeyoung Chung, Jeongtaek Oh, and Kyoung Mu Lee. Depth-regularized optimization for 3d gaussian splatting in few-shot images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 811–820, 2024.

Xiao Cui, Weicai Ye, Yifan Wang, Guofeng Zhang, Wengang Zhou, Tong He, and Houqiang Li. Streetsurfgs: Scalable urban street surface reconstruction with planar-based gaussian splatting. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. *Advances in neural information processing systems*, 29, 2016.

Lue Fan, Hao Zhang, Qitai Wang, Hongsheng Li, and Zhaoxiang Zhang. Freesim: Toward free-viewpoint camera simulation in driving scenes. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12004–12014, 2025.

Shuangkang Fang, I Shen, Takeo Igarashi, Yufeng Wang, ZeSheng Wang, Yi Yang, Wenrui Ding, Shuchang Zhou, et al. Nerf is a valuable assistant for 3d gaussian splatting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 26230–26240, 2025.

Jiatao Gu, Alex Trevithick, Kai-En Lin, Joshua M. Susskind, Christian Theobalt, Lingjie Liu, and Ravi Ramamoorthi. NerfDiff: Single-image view synthesis with NeRF-guided distillation from 3D-aware diffusion. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 11808–11826. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/gu23a.html.

Antoine Guédon and Vincent Lepetit. Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5354–5363, 2024.

Jianfei Guo, Nianchen Deng, Xinyang Li, Yeqi Bai, Botian Shi, Chiyu Wang, Chenjing Ding, Dongliang Wang, and Yikang Li. Streetsurf: Extending multi-view implicit surface reconstruction to street views. *arXiv preprint arXiv:2306.04988*, 2023.

Huasong Han, Kaixuan Zhou, Xiaoxiao Long, Yusen Wang, and Chunxia Xiao. Ggs: Generalizable gaussian splatting for lane switching in autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3329–3337, 2025.

Ayaan Haque, Matthew Tancik, Alexei Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-nerf2nerf: Editing 3d scenes with instructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Georg Hess, Carl Lindström, Maryam Fatemi, Christoffer Petersson, and Lennart Svensson. Splatad: Real-time lidar and camera rendering with 3d gaussian splatting for autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 11982–11992, 2025.

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. *Advances in neural information processing systems*, 30, 2017.

Sunghwan Hong, Jaewoo Jung, Heeseong Shin, Jisang Han, Jiaolong Yang, Chong Luo, and Seungryong Kim. Pf3plat: Pose-free feed-forward 3d gaussian splatting. *arXiv preprint arXiv:2410.22128*, 2024.

Sunghwan Hong, Jaewoo Jung, Heeseong Shin, Jisang Han, Jiaolong Yang, Chong Luo, and Seungryong Kim. PF3plat: Pose-free feed-forward 3d gaussian splatting for novel view synthesis. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=VjI1NnsW4t.

Qirui Hou, Wenzhang Sun, Chang Zeng, Chunfeng Wang, Hao Li, and Jianxun Cui. Drivingscene: A multi-task online feed-forward 3d gaussian splatting method for dynamic driving scenes. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 10287–10291. IEEE, 2026.

Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *SIGGRAPH 2024 Conference Papers*. Association for Computing Machinery, 2024. doi: 10.1145/3641519.3657428.

Han Huang, Yulun Wu, Chao Deng, Ge Gao, Ming Gu, and Yu-Shen Liu. Fatesgs: Fast and accurate sparse-view surface reconstruction using gaussian splatting with depth-feature consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3644–3652, 2025a.

Zexu Huang, Min Xu, and Stuart Perry. Det-gs: Depth-and edge-aware regularization for high-fidelity 3d gaussian splatting. *arXiv preprint arXiv:2508.04099*, 2025b.

Changjian Jiang, Ruilan Gao, Kele Shao, Yue Wang, Rong Xiong, and Yu Zhang. Li-gs: Gaussian splatting with lidar incorporated for accurate large-scale reconstruction. *IEEE Robotics and Automation Letters*, 2024.

Lihan Jiang, Yucheng Mao, Linning Xu, Tao Lu, Kerui Ren, Yichen Jin, Xudong Xu, Mulin Yu, Jiangmiao Pang, Feng Zhao, et al. Anysplat: Feed-forward 3d gaussian splatting from unconstrained views. *ACM Transactions on Graphics (TOG)*, 44(6):1–16, 2025.

Animesh Karnewar, Niloy J Mitra, Andrea Vedaldi, and David Novotny. Holofusion: Towards photo-realistic 3d generative modeling. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 22976–22985, 2023.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), July 2023. URL `https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/`.

Jaewon Lee, Mangyu Kong, Minseong Park, and Euntai Kim. Geomgs: Lidar-guided geometry-aware gaussian splatting for robot localization. *arXiv preprint arXiv:2501.13417*, 2025.

Jiahe Li, Jiawei Zhang, Xiao Bai, Jin Zheng, Xin Ning, Jun Zhou, and Lin Gu. Dngaussian: Optimizing sparse-view 3d gaussian radiance fields with global-local depth normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20775–20785, 2024a.

Mengfei Li, Ming Lu, Xiaofang Li, and Shanghang Zhang. Rustnerf: Robust neural radiance field with low-quality images. *arXiv preprint arXiv:2401.03257*, 2024b.

Peizheng Li, Zhenghao Zhang, David Holtz, Hang Yu, Yutong Yang, Yuzhi Lai, Rui Song, Andreas Geiger, and Andreas Zell. Spacedrive: Infusing spatial awareness into vlm-based autonomous driving. *arXiv preprint arXiv:2512.10719*, 2, 2025.

Yanyan Li, Chenyu Lyu, Yan Di, Guangyao Zhai, Gim Hee Lee, and Federico Tombari. Geogaussian: Geometry-aware gaussian splatting for scene rendering. In *European conference on computer vision*, pp. 441–457. Springer, 2024c.

Junhong Lin, Kangli Wang, Shunzhou Wang, Songlin Fan, Ge Li, and Wei Gao. Vgd: Visual geometry gaussian splatting for feed-forward surround-view driving reconstruction. *arXiv preprint arXiv:2510.19578*, 2025.

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9298–9309, 2023a.

Xi Liu, Chaoyi Zhou, and Siyu Huang. 3dgs-enhancer: Enhancing unbounded 3d gaussian splatting with view-consistent 2d diffusion priors. *Advances in Neural Information Processing Systems*, 37:133305–133327, 2024.

Yifan Liu, Zhiyuan Min, Zhenwei Wang, Junta Wu, Tengfei Wang, Yixuan Yuan, Yawei Luo, and Chunchao Guo. Worldmirror: Universal 3d world reconstruction with any-prior prompting. *arXiv preprint arXiv:2510.10726*, 2025.

Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023b.

Tao Lu, Mulin Yu, Linning Xu, Yuanbo Xiangli, Limin Wang, Dahua Lin, and Bo Dai. Scaffold-gs: Structured 3d gaussians for view-adaptive rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20654–20664, 2024.

Saswat Subhajyoti Mallick, Rahul Goel, Bernhard Kerbl, Markus Steinberger, Francisco Vicente Carrasco, and Fernando De La Torre. Taming 3dgs: High-quality radiance fields with limited resources. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024.

Jieru Mei, Alex Zihao Zhu, Xinchen Yan, Hang Yan, Siyuan Qiao, Liang-Chieh Chen, and Henrik Kretzschmar. Waymo open dataset: Panoramic video panoptic segmentation. In *European Conference on Computer Vision*, pp. 53–72. Springer, 2022.

Sheng Miao, Jiaxin Huang, Dongfeng Bai, Xu Yan, Hongyu Zhou, Yue Wang, Bingbing Liu, Andreas Geiger, and Yiyi Liao. Evolsplat: Efficient volume-based gaussian splatting for urban view synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 11286–11296, 2025.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

Norman Müller, Katja Schwarz, Barbara Rössle, Lorenzo Porzi, Samuel Rota Bulo, Matthias Nießner, and Peter Kontschieder. Multidiff: Consistent novel view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10258–10268, 2024.

Avinash Paliwal, Xilong Zhou, Wei Ye, Jinhui Xiong, Rakesh Ranjan, and Nima Khademi Kalantari. Ri3d: Few-shot gaussian splatting with repair and inpainting diffusion priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 25094–25103, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pp. 87–103. Springer, 2024.

Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *European conference on computer vision*, pp. 501–518. Springer, 2016.

Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016.

Junyoung Seo, Kazumi Fukuda, Takashi Shibuya, Takuya Narihira, Naoki Murata, Shoukang Hu, Chieh-Hsin Lai, Seungryong Kim, and Yuki Mitsufuji. Genwarp: Single image to novel views with semantic-preserving generative warping. *Advances in Neural Information Processing Systems*, 37:80220–80243, 2024.

Jian Shen, Huai Yu, Ji Wu, Wen Yang, and Gui-Song Xia. Lidar-enhanced 3d gaussian splatting mapping. *arXiv preprint arXiv:2503.05425*, 2025.

Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. MVDream: Multi-view diffusion for 3d generation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=FUgrjq2pbB.

Brandon Smart, Chuanxia Zheng, Iro Laina, and Victor Adrian Prisacariu. Splatt3r: Zero-shot gaussian splatting from uncalibrated image pairs. *arXiv preprint arXiv:2408.13912*, 2024.

Rui Song, Chenwei Liang, Hu Cao, Zhiran Yan, Walter Zimmer, Markus Gross, Andreas Festag, and Alois Knoll. Collaborative semantic occupancy prediction with hybrid feature fusion in connected automated vehicles. In *2024 IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE/CVF, 2024.

Rui Song, Chenwei Liang, Yan Xia, Walter Zimmer, Hu Cao, Holger Caesar, Andreas Festag, and Alois Knoll. Coda-4dgs: Dynamic gaussian splatting with context and deformation awareness for autonomous driving. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE/CVF, 2025.

Rui Song, Tianhui Cai, Markus Gross, Yun Zhang, Walter Zimmer, Zhiyu Huang, Olaf Wysocki, and Jiaqi Ma. Energs: Energy-based gaussian splatting with partial geometric priors. *arXiv preprint arXiv:2604.26238*, 2026.

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 Computer Vision and Pattern Recognition Conference*, pp. 27487–27496, 2025.

Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision*, pp. 1–18. Springer, 2024.

Zeng Tao, Tong Yang, Junxiong Lin, Xinji Mai, Haoran Wang, Beining Wang, Enyu Zhou, Yan Wang, and Wenqiang Zhang. Lcgen: Mining in low-certainty generation for view-consistent text-to-3d. *Advances in Neural Information Processing Systems*, 37:20276–20303, 2024.

Qijian Tian, Xin Tan, Yuan Xie, and Lizhuang Ma. Drivingforward: Feed-forward 3d gaussian splatting for driving scene reconstruction from flexible surround-view input. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

Matias Turkulainen, Xuqian Ren, Iaroslav Melekhov, Otto Seiskari, Esa Rahtu, and Juho Kannala. Dn-splatter: Depth and normal priors for gaussian splatting and meshing. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2421–2431. IEEE, 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.

Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. Mots: Multi-object tracking and segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Qitai Wang, Lue Fan, Yuqi Wang, Yuntao Chen, and Zhaoxiang Zhang. Freevs: Generative view synthesis on free driving trajectory. *arXiv preprint arXiv:2410.18079*, 2024a.

Yunsong Wang, Tianxin Huang, Hanlin Chen, and Gim Hee Lee. Freesplat: Generalizable 3d gaussian splatting towards free view synthesis of indoor scenes. *Advances in Neural Information Processing Systems*, 37:107326–107349, 2024b.

Yuping Wang, Shuo Xing, Cui Can, Renjie Li, Hongyuan Hua, Kexin Tian, Zhaobin Mo, Xiangbo Gao, Keshu Wu, Sulong Zhou, et al. Generative ai for autonomous driving: Frontiers and opportunities. *arXiv preprint arXiv:2505.08854*, 2025a.

Zhaorui Wang, Yi Gu, Deming Zhou, and Renjing Xu. Fixinggs: Enhancing 3d gaussian splatting via training-free score distillation. *arXiv preprint arXiv:2509.18759*, 2025b.

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.

Jiaxin Wei, Stefan Leutenegger, and Simon Schaefer. Gsfix3d: Diffusion-guided repair of novel views in gaussian splatting. *arXiv preprint arXiv:2508.14717*, 2025.

Christopher Wewer, Kevin Raj, Eddy Ilg, Bernt Schiele, and Jan Eric Lenssen. latentsplat: Autoencoding variational gaussians for fast generalizable 3d reconstruction. In *European conference on computer vision*, pp. 456–473. Springer, 2024.

Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20310–20320, 2024.

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*, pp. 26024–26035, 2025.

Kejing Xia, Jidong Jia, Ke Jin, Yucai Bai, Li Sun, Dacheng Tao, and Youjian Zhang. D^2GS: Dense depth regularization for LiDAR-free urban scene reconstruction. *Advances in Neural Information Processing Systems*, 38:56899–56927, 2025.

Haofei Xu, Songyou Peng, Fangjinhua Wang, Hermann Blum, Daniel Barath, Andreas Geiger, and Marc Pollefeys. Depthsplat: Connecting gaussian splatting and depth. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 16453–16463, 2025a.

Jiawei Xu, Kai Deng, Zexin Fan, Shenlong Wang, Jin Xie, and Jian Yang. Ad-gs: Object-aware b-spline gaussian splatting for self-supervised autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24770–24779, 2025b.

Runsheng Xu, Xin Xia, Jinlong Li, Hanzhao Li, Shuo Zhang, Zhengzhong Tu, Zonglin Meng, Hao Xiang, Xiaoyu Dong, Rui Song, et al. V2v4real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13712–13722, 2023.

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*, pp. 156–173. Springer, 2024.

Yunzhi Yan, Zhen Xu, Haotong Lin, Haian Jin, Haoyu Guo, Yida Wang, Kun Zhan, Xianpeng Lang, Hujun Bao, Xiaowei Zhou, et al. Streetcrafter: Street view synthesis with controllable video diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 822–832, 2025.

Jiawei Yang, Jiahui Huang, Yuxiao Chen, Yan Wang, Boyi Li, Yurong You, Apoorva Sharma, Maximilian Igl, Peter Karkus, Danfei Xu, et al. Storm: Spatio-temporal reconstruction model for large-scale outdoor scenes. *arXiv preprint arXiv:2501.00602*, 2024a.

Jiawei Yang, Boris Ivanovic, Or Litany, Xinshuo Weng, Seung Wook Kim, Boyi Li, Tong Che, Danfei Xu, Sanja Fidler, Marco Pavone, and Yue Wang. EmerneRF: Emergent spatial-temporal scene decomposition via self-supervision. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=ycv2z8TYur.

Zeyu Yang, Zijie Pan, Yuankun Yang, Xiatian Zhu, and Li Zhang. Driving view synthesis on free-form trajectories with generative prior. *arXiv preprint arXiv:2412.01717*, 2024c.

Botao Ye, Sifei Liu, Haofei Xu, Xueting Li, Marc Pollefeys, Ming-Hsuan Yang, and Songyou Peng. No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images. *arXiv preprint arXiv:2410.24207*, 2024.

Botao Ye, Boqi Chen, Haofei Xu, Daniel Barath, and Marc Pollefeys. Yonosplat: You only need one model for feedforward 3d gaussian splatting. *arXiv preprint arXiv:2511.07321*, 2025.

Xingyilang Yin, Qi Zhang, Jiahao Chang, Ying Feng, Qingnan Fan, Xi Yang, Chi-Man Pun, Huaqi Zhang, and Xiaodong Cun. Gsfixer: Improving 3d gaussian splatting with reference-guided video diffusion priors. *arXiv preprint arXiv:2508.09667*, 2025.

Haibao Yu, Yizhen Luo, Mao Shu, Yiyi Huo, Zebang Yang, Yifeng Shi, Zhenglong Guo, Hanyu Li, Xing Hu, Jirui Yuan, et al. Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21361–21370, 2022.

Hanyang Yu, Xiaoxiao Long, and Ping Tan. Lm-gaussian: Boost sparse-view 3d gaussian splatting with large model priors. *arXiv preprint arXiv:2409.03456*, 2024a.

Mark Yu, Wenbo Hu, Jinbo Xing, and Ying Shan. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 100–111, October 2025.

Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-splatting: Alias-free 3d gaussian splatting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19447–19456, 2024b.

Chuanrui Zhang, Yingshuang Zou, Zhuoling Li, Minmin Yi, and Haoqian Wang. Transplat: Generalizable 3d gaussian splatting from sparse multi-view images with transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 9869–9877, 2025a.

Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020.

Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. Gs-lrm: Large reconstruction model for 3d gaussian splatting. In *European Conference on Computer Vision*, pp. 1–19. Springer, 2024.

Qilin Zhang, Olaf Wysocki, Steffen Urban, and Boris Jutzi. Cdgs: Confidence-aware depth regularization for 3d gaussian splatting. *arXiv preprint arXiv:2502.14684*, 2025b.

Seth Z Zhao, Luobin Wang, Hongwei Ruan, Yuxin Bao, Yilan Chen, Ziyang Leng, Abhijit Ravichandran, Honglin He, Zewei Zhou, Xu Han, et al. Bridgesim: Unveiling the ol-cl gap in end-to-end autonomous driving. *arXiv preprint arXiv:2604.10856*, 2026.

Hongyu Zhou, Longzhong Lin, Jiabao Wang, Yichong Lu, Dongfeng Bai, Bingbing Liu, Yue Wang, Andreas Geiger, and Yiyi Liao. Hugsim: A real-time, photo-realistic and closed-loop simulator for autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025a.

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*, pp. 21634–21643, 2024.

Zewei Zhou, Hao Xiang, Zhaoliang Zheng, Seth Z Zhao, Mingyue Lei, Yun Zhang, Tianhui Cai, Xinyi Liu, Johnson Liu, Maheswari Bajji, et al. V2XPnP: Vehicle-to-everything spatio-temporal fusion for multi-agent perception and prediction. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025b.

Zewei Zhou, Seth Z Zhao, Tianhui Cai, Zhiyu Huang, Bolei Zhou, and Jiaqi Ma. Turbotrain: Towards efficient and balanced multi-task learning for multi-agent perception and prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4391–4402, 2025c.

Zewei Zhou, Tianhui Cai, Seth Zhao, Yun Zhang, Zhiyu Huang, Bolei Zhou, and Jiaqi Ma. Autovla: A vision-language-action model for end-to-end autonomous driving with adaptive reasoning and reinforcement fine-tuning. *Advances in Neural Information Processing Systems*, 38: 27920–27956, 2026a.

Zewei Zhou, Ruining Yang, Yiluan Guo, Sherry X Chen, Tao Feng, Kateryna Pistunova, Yishan Shen, Lili Su, Jiaqi Ma, et al. Spanvla: Efficient action bridging and learning from negative-recovery samples for vision-language-action model. *arXiv preprint arXiv:2604.19710*, 2026b.

Walter Zimmer, Gerhard Arya Wardana, Suren Sritharan, Xingcheng Zhou, Rui Song, and Alois C Knoll. Tumtraf v2x cooperative perception dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22668–22677, 2024.

Zi-Xin Zou, Zhipeng Yu, Yuan-Chen Guo, Yangguang Li, Ding Liang, Yan-Pei Cao, and Song-Hai Zhang. Triplane meets gaussian splatting: Fast and generalizable single-view 3d reconstruction with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10324–10335, 2024.

We provide additional implementation details, methodological analysis, efficiency evaluation, robustness studies, and qualitative results to support the main paper. To facilitate reproducibility, Sec. A details the complete benchmark configuration, including dataset protocols and all key hyperparameters used in our experiments. Sec. B provides additional methodological details on our confidence-weighted objective and gradients, clarifying how the support-validated confidence is incorporated into optimization and densification. Sec. C reports a complexity and efficiency analysis. Sec. D provides analyses of dynamic regions, camera-pose errors, hyperparameter sensitivity, diffusion-prior corruption, input sparsity, and compatibility with feedforward 4D reconstruction. Finally, Sec. E presents additional qualitative results in three parts: Sec. E.1 novel view synthesis along out-of-distribution trajectories, including lateral offsets and a virtual drone view; Sec. E.2 plug-and-play enhancement results on additional feedforward backbones; and Sec. E.3 further visual comparisons with Difix3D+. The videos complement these results by comparing our method against multiple baselines, highlighting the stability of ConFixGS under challenging viewpoint changes and its effectiveness in enhancing state-of-the-art feedforward Gaussian Splatting methods. More **videos** and our **code** are provided in the supplementary material.

A Implementation Details

Datasets and protocol. We evaluate on driving scenes from Waymo Open Dataset (Mei et al., 2022), nuScenes (Caesar et al., 2020), and KITTI (Voigtlaender et al., 2019). For each scene, we extract 50 consecutive frames from the front-facing camera, using every 5th frame as a support (context) view (10 views) and the remaining 40 frames as novel views, yielding an extremely sparse-view setting (80% novel views). In comparison, prior driving novel view synthesis benchmarks often use protocols with much denser context views, e.g., holding out every 4th frame (Zhou et al., 2024; Sun et al., 2025; Yang et al., 2024b; Song et al., 2025) (25% novel views) or every 2nd frame (Tian et al., 2025) (50% novel views) as test views. All images are resized with aspect ratio preserved (longer side 784 px), with the height rounded to multiples of 14 for compatibility with the 14×14 patch size of the DINOv2 ViT-L/14 in WorldMirror (Liu et al., 2025). For the DepthSplat (Xu et al., 2025a) backbone, images are also resized to match the pretrained model and the encoder's 14×14 patch size. We use dataset-provided camera intrinsics and poses, and additionally apply COLMAP-based scale calibration (Schönberger & Frahm, 2016; Schönberger et al., 2016).

Global-local feedforward backbones. Our framework is plug-and-play w.r.t. the feedforward initializer $\mathcal{G}_0$. We report results using two representative global architectures: DepthSplat (Xu et al., 2025a) and WorldMirror (Liu et al., 2025). For practical compatibility, DepthSplat is executed via a lightweight subprocess bridge, while WorldMirror runs in-process at the same target resolution. The local feedforward operator F_{local} in Eq. 2 is *not* a separate network: it shares the same backbone weights as the global model F_{global}, and is obtained by re-running the same feedforward predictor over a small local window of nearby trajectory views centered on each novel view n. In Sec. E.2, we further provide results using DrivingForward (Tian et al., 2025) and AnySplat (Jiang et al., 2025) as alternative backbones, along with both qualitative and quantitative comparisons before and after enhancement, to demonstrate the generalizability of our approach across diverse feedforward initializers.

Diffusion refinement module. For image-space generative enhancement, we adopt SD-Turbo (Sauer et al., 2024) as our diffusion prior, applied to refine novel view renderings from local feedforward models, and produce pseudo-targets I^*. We rescale the diffusion outputs back to the rendering resolution and retain the camera parameters in corresponding Gaussians.

Support-validated reprojection confidence. A core component is our support-validated confidence estimation for novel views. Using the depth proxy $d_n^{(0)}$ rendered from the initial feedforward reconstruction $\mathcal{G}_0$ as a geometric scaffold, we unproject diffusion-refined pixels into 3D and reproject them into multiple support views to measure photometric disagreement with the input observations. This disagreement is converted into the per-pixel confidence score with $\sigma_e = 0.10$. Pixels that cannot be validated by any support image receive the conservative baseline confidence score $\gamma = 0.3$, and pixels falling in regions whose accumulated opacity $\hat{\alpha}_n^{(0)}$ is below the scaffold-coverage threshold $\alpha_{\min} = 0.3$ are zeroed out, since their depth proxy and the corresponding reprojection are

unreliable. We then apply spatial smoothing with a 15×15 mean-filter kernel ($k = 15$) to obtain the final confidence map W_n.

Confidence-weighted global optimization. We optimize the global 3DGS for 1k iterations using Adam, with learning rates 1.6×10^{-4} for Gaussian positions and 2.5×10^{-3} for spherical harmonics. We set $\lambda_s = 0.2$ and $\lambda_{sup} = 0.5$. During training, the confidence map gates distillation gradients on novel views and also modulates densification to prevent low-confidence regions from spawning spurious Gaussians. Densification is performed every 100 steps under confidence control.

Hardware and software environment. All experiments are conducted on a single NVIDIA L40S GPU (46 GiB VRAM) hosted on a workstation with two AMD EPYC 9354 32-core CPUs and 750 GB system RAM, running Ubuntu 22.04 (Linux kernel 5.15). ConFixGS is implemented in PyTorch 2.4 with CUDA 12.4. To preserve dependency parity with each released baseline, every comparison method is executed inside its own original conda environment, cross-environment communication (e.g., the DepthSplat feedforward call) is handled through a lightweight subprocess bridge.

Algorithm 1 Support-Validated Confidence Optimization

Input: Captured support images $\{I_s^{\text{obs}}\}_{s\in\mathcal{S}}$, cameras $\{(K_v, T_v)\}_{v=1}^{V}$, hyperparameters $\sigma_e, \gamma, \alpha_{\min}, \lambda_s, \lambda_{sup}, T, T_d$
Output: Optimized 3D scene $\mathcal{G}_{\text{final}}$
$\mathcal{G}_0 \leftarrow F_{\text{global}}(\{I_s^{\text{obs}}\}_{s\in\mathcal{S}})$ ▷ Initial feedforward reconstruction, Eq. 1
$(\hat{I}_v^{(0)}, \hat{\alpha}_v^{(0)}, d_v^{(0)}) \leftarrow \text{Render}(\mathcal{G}_0; K_v, T_v),\ \forall v$ ▷ Base render, opacity, depth (Sec. 3.3)
for $n \notin \mathcal{S}$ **do**
    $I_n^* \leftarrow \text{LocalEpisode}(I_n^{\text{ctx}}, \hat{I}_n^{(0)}, \mathcal{F}_{\text{SD-Turbo}})$ ▷ Pseudo-target generation, Eq. 2
end for
$I_s^* \leftarrow I_s^{\text{obs}},\ \forall s \in \mathcal{S}$ ▷ Support targets are captured inputs
for $n \notin \mathcal{S}$ **do**
    $W_n \leftarrow \text{ReprojConfidence}\Big(I_n^*, d_n^{(0)}, \hat{\alpha}_n^{(0)}, \{I_s^{\text{obs}}\}_{s\in\mathcal{S}}\Big)$ ▷ Support-validated confidence map via Sec. 3.4, Eq. 7
end for
$W_s \leftarrow \mathbf{1},\ \forall s \in \mathcal{S}$ ▷ All-ones confidence map for support views
Initialize parameters $\Theta \leftarrow \mathcal{G}_0$
for iteration $t = 1, \ldots, T$ **do**
    Sample mini-batch $\mathcal{B} \subset \{1, \ldots, V\}$, let $\mathcal{B}_{sup} = \mathcal{B} \cap \mathcal{S}$
    $\hat{C}_v \leftarrow \text{Render}(\Theta; K_v, T_v),\ \forall v \in \mathcal{B}$ ▷ Current renderings
    $\mathcal{L} \leftarrow \frac{1}{|\mathcal{B}|}\sum_{v\in\mathcal{B}} \mathcal{L}_{\text{tgt}}(v)$ ▷ Weighted objective, Eqs. 10–12
    if $|\mathcal{B}_{sup}| > 0$ **then**
        $\mathcal{L} \leftarrow \mathcal{L} + \lambda_{sup}\frac{1}{|\mathcal{B}_{sup}|}\sum_{s\in\mathcal{B}_{sup}} \mathcal{L}_{\text{sup}}(s)$ ▷ Captured-support anchoring (Sec. 3.5)
    end if
    Backpropagate and update Θ via Adam ▷ Gradients are modulated by W_v through Appendix Eq. 13
    if $t \bmod T_d == 0$ **then**
        Densify/Prune Gaussians using the standard 3DGS strategy ▷ Topology updates follow standard 3DGS thresholds. Densification signals are weighted by Appendix Eq. 15
    end if
end for
return $\mathcal{G}_{\text{final}} \leftarrow \Theta$

B Confidence-Weighted Gradients and Objective

To mitigate both geometric and photometric degradation caused by diffusion hallucinations, we propose a confidence-weighted optimization scheme guided by support-validated, discrepancy-based confidence maps. For each novel view pixel, we reproject the pseudo-target into the captured support images used by the global feedforward model and compute its photometric discrepancy from the reprojected multi-view consensus. Pixels with high disagreement receive a low confidence score.

This section details the formulations governing our modulated backward pass. We demonstrate how this per-pixel confidence score acts as an explicit gate that embeds input-observation consistency into both the continuous parameter optimization and the discrete densification of the repaired 3DGS scene.

B.1 Explicit Coupling in the Backward Pass

Rather than unconditionally backpropagating the 2D photometric residual, our framework modulates the gradient using the support-validated confidence map W_v, whose per-pixel confidence score $w_v(p) = (G_k * \tilde{w}_v)(p)$ is obtained by spatially smoothing the raw confidence score $\tilde{w}_v$ with a mean filter G_k. By expanding the normalized weight $\lambda_v(p) = \frac{w_v(p)}{\sum_{p'} w_v(p')}$, the confidence-weighted gradient with respect to any 3D Gaussian parameter $\theta_i \in \{\mu_i, \hat{s}_i, q_i, \hat{o}_i, c_i\}$ is explicitly formulated as:

$$\frac{\partial \mathcal{L}}{\partial \theta_i} = \frac{1}{|\mathcal{B}|} \sum_{v \in \mathcal{B}} \sum_{p} \underbrace{\left[\frac{(G_k * \tilde{w}_v)(p)}{\sum_{p'} w_v(p')} \right]}_{\text{Confidence Gating } \lambda_v(p)} \cdot \underbrace{\nabla_{\hat{C}} \ell_v(p)}_{\text{2D Loss Gradient}} \cdot \underbrace{\frac{\partial \hat{C}_v(p)}{\partial \theta_i}}_{\text{Rendering Derivative}} . \tag{13}$$

where the raw per-pixel confidence score $\tilde{w}_v(p)$ encodes support-image validation through the valid support view set $\mathcal{V}(p)$, the photometric discrepancy $e_v(p) = \frac{1}{3} \sum_{c \in \{R,G,B\}} |I_v^*(p,c) - \bar{I}_{\text{reproj}}(p,c)|$ between the pseudo-target and the reprojected multi-view consensus, and the geometric coverage indicator $\hat{\alpha}_v^{(0)}(p)$ rendered from the initial feedforward reconstruction $\mathcal{G}_0$:

$$\tilde{w}_v(p) = \begin{cases} \exp\left(-\frac{e_v(p)^2}{2\sigma_e^2}\right), & \text{if } \mathcal{V}(p) \neq \emptyset \text{ and } \hat{\alpha}_v^{(0)}(p) \geq \alpha_{\min}, \\ \gamma, & \text{if } \mathcal{V}(p) = \emptyset \text{ and } \hat{\alpha}_v^{(0)}(p) \geq \alpha_{\min}, \\ 0, & \text{if } \hat{\alpha}_v^{(0)}(p) < \alpha_{\min}. \end{cases} \tag{14}$$

This expanded formulation reveals the essence of our method: the parameter update is modulated by the target's consistency with the photometric consensus established by the support views, while the initial reconstruction $\mathcal{G}_0$ enters only as a geometric scaffold for establishing correspondences. Note that in the main paper, we use n to index novel views ($n \notin \mathcal{S}$), and here we use v as a generic view index over all V views.

B.2 Mechanism of Densification Gating

Within our framework, the confidence-modulated 2D position gradient accumulation $\bar{g}_i$ for Gaussian i across its associated views S_i becomes deeply coupled with the support-validated per-pixel confidence score $w_v(p)$:

$$\bar{g}_i = \frac{1}{|S_i|} \sum_{t \in S_i} \left\| \sum_{p} \lambda_{v_t}(p) \cdot \nabla_{\hat{C}} \ell_{v_t}(p) \cdot \frac{\partial \hat{C}_{v_t}(p)}{\partial \mu_i^{2D}} \right\| . \tag{15}$$

This explicit expansion illustrates three distinct gating behaviors dictated by the underlying raw confidence score $\tilde{w}_v(p)$:

1. **Geometric Hallucinations ($\hat{\alpha}_v^{(0)}(p) < \alpha_{\min}$):** If the pixel falls in a region where the initial reconstruction $\mathcal{G}_0$ has no opacity coverage, the rendered depth $d_v^{(0)}(p)$ is unreliable, so any reprojection based on it—and the photometric comparison built on top—is meaningless. We set $\tilde{w}_v(p) \equiv 0$, mathematically zeroing out the positional gradient and preventing the creation of floaters in empty space.
2. **Photometric Inconsistencies ($e_v(p) \gg 0$):** For regions with valid scaffold coverage but severe multi-view conflicts (where the pseudo-target diverges from the support-view consensus $\bar{I}_{\text{reproj}}$), the exponential decay forces $\tilde{w}_v(p) \to 0$, naturally suppressing conflicting gradient updates.
3. **Unobserved Regions ($\mathcal{V}(p) = \emptyset$):** When the geometry is covered by $\mathcal{G}_0$ but no support view can validate pixel p photometrically (e.g., occluded or out-of-view), we assign a baseline $\tilde{w}_v(p) = \gamma$ instead of zero. This preserves a moderate gradient flow, allowing the

diffusion prior to guide expansion where multi-view consensus is unavailable, while remaining less influential than validated regions.

Therefore, our support-validated confidence map suppresses gradient accumulation from unreliable regions with high reprojection discrepancy or with no scaffold coverage, reducing the likelihood that spurious artifacts trigger densification, while preserving limited gradient flow in unobserved-but-covered regions via γ. Note that S_i denotes the set of views where Gaussian i is visible, and v_t is the view index of the t-th element in S_i, *i.e.*, $v_t \in S_i$.

C Complexity and Efficiency Analysis

We analyze the computational cost of ConFixGS in a form aligned with the three-stage pipeline described in the Methodology section. Given K total views, including K_s support views and $K_n = K - K_s$ novel views, the overall computation is dominated by three components: (1) initial feedforward reconstruction and local diffusion, (2) reprojection-based confidence estimation, and (3) confidence-modulated global repair optimization.

Stage 1: Initial feedforward reconstruction and local diffusion. In the first stage, the feedforward backbone F_{global} is executed once on the support views to predict the initial Gaussian scene $\mathcal{G}_0$ to be repaired, after which $\mathcal{G}_0$ is rendered at all K camera poses to produce base renderings $\hat{I}_v^{(0)}$. Then, for each novel view $n \notin \mathcal{S}$, the local branch F_{local} and the diffusion prior $\mathcal{F}_{\text{SD-Turbo}}$ are used to generate the pseudo-target I_n^*. Let T_{ff}^g and T_{ff}^l denote the costs of one global and one local feedforward call, respectively, T_{render} the cost of one 3DGS rendering, and T_{diff} the cost of one diffusion refinement. The complexity of Stage 1 is therefore

$$T_{\text{stage1}} = T_{\text{ff}}^g + K\,T_{\text{render}} + K_n\left(T_{\text{ff}}^l + T_{\text{diff}}\right). \tag{16}$$

This stage scales linearly with the number of novel views, and in practice is typically dominated by the per-view rendering of $\mathcal{G}_0$, the repeated local feedforward, and the diffusion refinement. Empirically, on a representative Waymo scene (50 frames, single L40S GPU), Stage 1 takes $\sim$12.5 min with the DepthSplat backbone (where the per-view DepthSplat call is dispatched as a subprocess bridge) and $\sim$6.4 min with the WorldMirror backbone (where the model stays resident in process across calls).

Stage 2: Reprojection-based confidence estimation. For each novel view pixel, Stage 2 unprojects the pixel using the depth proxy $d_n^{(0)}$ rendered from the initial feedforward reconstruction $\mathcal{G}_0$ (a cheap geometric scaffold for correspondence), reprojects it into all support views, samples valid support observations, and computes the discrepancy-based confidence map W_n. Assuming image resolution $H \times W$, the cost of reprojection-based confidence estimation is linear in the number of novel views, support views, and pixels:

$$T_{\text{stage2}} = \mathcal{O}(K_n K_s H W). \tag{17}$$

Since this stage involves geometric reprojection, image sampling, and photometric discrepancy computation, its cost is substantially smaller than feedforward inference or iterative 3DGS optimization in practice. In our setting (K_n=40, K_s=10, 448×768), Stage 2 finishes in $\sim$15 s, accounting for less than 2% of the total pipeline cost.

Stage 3: Confidence-modulated global repair optimization. In the final stage, we initialize $\Theta \leftarrow \mathcal{G}_0$ and optimize the Gaussian scene for S iterations using mini-batches of size B. Each iteration requires rendering the current scene for the sampled views, computing the confidence-weighted photometric objective, and back-propagating gradients to Gaussian parameters. Let N be the number of Gaussians. The resulting complexity can be expressed as

$$T_{\text{stage3}} = \mathcal{O}(S \cdot B \cdot T_{\text{render}} + S \cdot N), \tag{18}$$

where the rendering term captures forward rasterization and the N-dependent term reflects parameter updates and accumulated densification statistics. Because confidence maps suppress gradients in unreliable regions, the effective optimization is concentrated on high-confidence pixels, which reduces unnecessary gradient competition and stabilizes densification. With S=1,000 and B=4 on $\sim$3.4 M Gaussians, Stage 3 runs at $\sim$0.11 s/step and completes in under 2 min, making it the second-largest contributor to total runtime after the local feedforward and base rendering of Stage 1.

Table 5: **Per-scene wall-clock and peak GPU memory** on a representative Waymo scene (50 frames, K_s=10 support / K_n=40 novel, 448×768, single NVIDIA L40S 46 GiB GPU). *Train time* reports the full pipeline wall-clock, and *Test time* reports the wall-clock per synthesized novel view. Confidence-modulated supervision lets ConFixGS converge in S=1,000 optimization steps with either the WorldMirror or DepthSplat backbone, more than **40**× faster than the S=30,000 schedule of per-scene 3DGS or Difix3D+, while ConFixGS-DS additionally uses noticeably less peak GPU memory than either baseline.

Method	**Backbone**	**Opt. steps S**	**Train time↓**	**Test time↓**	**Peak GPU Mem.↓**
Plain 3DGS	–	30 000	9 h 39 min	1.28 s/view	27 GiB
Difix3D+ (Wu et al., 2025)	–	30 000 + 10 fix	10 h 39 min	1.28 s/view	29 GiB
ConFixGS-WM (*ours*)	WorldMirror	1 000	**8 min 53 s**	1.28 s/view	35 GiB
ConFixGS-DS (*ours*)	DepthSplat	1 000	**14 min 39 s**	**1.28 s/view**	**21 GiB**

Overall complexity. Combining the three stages, the total cost of ConFixGS is

$$T_{\text{total}} = T_{\text{ff}}^{g} + K\,T_{\text{render}} + K_n\left(T_{\text{ff}}^{l} + T_{\text{diff}}\right) + \mathcal{O}(K_n K_s HW) + \mathcal{O}(S \cdot B \cdot T_{\text{render}} + S \cdot N)\,. \quad (19)$$

This expression highlights three important properties of our pipeline. First, the method scales linearly with the number of novel views K_n, since pseudo-target construction and confidence estimation are both performed per novel view. Second, the reprojection-based confidence estimation is lightweight compared to feedforward inference and global optimization. Third, unlike iterative diffusion-refinement pipelines that repeatedly alternate between rendering and generative correction, our method performs pseudo-target generation only once before optimization, after which refinement proceeds entirely in 3D under confidence-aware supervision. As a result, the overall efficiency is primarily governed by the chosen feedforward backbone and the number of global optimization steps.

C.1 Empirical Wall-Clock Comparison

To complement the theoretical analysis above, we report measured per-scene wall-clock and peak GPU memory in Table 5 on a representative Waymo scene (50 frames, K_s=10 support and K_n=40 novel views, 448×768 resolution) using a single NVIDIA L40S 46 GiB GPU. Because confidence-modulated supervision concentrates gradients on confident regions and suppresses diffusion hallucinations early in optimization, ConFixGS converges in S=1,000 steps. This is more than an order of magnitude fewer than the S=30,000 schedule used by per-scene 3DGS and Difix3D+ (Wu et al., 2025) and is the dominant source of the observed speed-up. As baselines we include a per-scene 3DGS reconstruction, and a Difix3D+ (Wu et al., 2025) reproduction that interleaves 10 rounds of diffusion fixing into the 30,000-step 3DGS schedule. *Test time* reports the wall-clock for synthesizing a single novel view using the final Gaussian scene. This cost is dominated by gsplat rasterization and PNG encoding to disk, and is therefore identical across methods.

Per-step supervision cost. ConFixGS computes the pseudo-targets I_n^* and the confidence maps W_n once before optimization, in Stages 1d and 2, and applies them as cached tensors at every iteration of Stage 3. Each ConFixGS optimization step therefore performs only a rasterization, a per-pixel multiplication with the cached W_n, the photometric loss, and an Adam update. Difix3D+ instead refreshes its diffusion supervision through periodic *fix events*, each of which re-renders all K_n=40 novel views and re-invokes the diffusion fixer subprocess end-to-end. Table 7 contrasts the wall-clock of one such step under each scheme on the same Waymo scene: a Difix3D+ fix-injection step is more than three orders of magnitude slower than a single ConFixGS optimization step, which is why dense per-step diffusion supervision is feasible only when the supervision is precomputed and cached as in ConFixGS.

Why ConFixGS is efficient. Three factors make ConFixGS roughly **40**× faster than per-scene 3DGS and **43**× faster than Difix3D+ on the same scene, while consuming comparable or lower peak memory: (*i*) pseudo-target generation runs only *once* before optimization, avoiding the expensive iterative render–diffuse–retrain loop of Difix3D+, where each fixing round adds another full pass through the 40 novel views; (*ii*) Stage 3 starts from a near-converged $\mathcal{G}_0$ and only needs to assimilate

Table 6: **Stage-wise wall-clock breakdown of ConFixGS-DS** (DepthSplat backbone, S=1,000 optimization steps, single L40S). Stage 1 dominates the total cost; Stage 2 (confidence estimation) is negligible; and Stage 3 (confidence-modulated optimization) finishes in under two minutes because the optimizer starts from a near-converged $\mathcal{G}_0$.

Stage	Component	Time	% of pipeline
Stage 1a	$\mathcal{G}_0$ global feedforward	53 s	6.0%
Stage 1b	Render $\mathcal{G}_0$ at all K=50 views	4 m 34 s	31.2%
Stage 1c	50 local feedforward calls	6 m 55 s	47.2%
Stage 1d	SD-Turbo diffusion (K_n=40 novel views)	9 s	1.0%
Stage 1 total (initial reconstruction + pseudo-targets)		**12 m 30 s**	**85.4%**
Stage 2	Reprojection-based confidence estimation	15 s	1.7%
Stage 3	Confidence-modulated 3DGS optimization (S=1,000, B=4)	1 m 53 s	12.9%
Pipeline total		**14 m 39 s**	100%

Table 7: **Per-step wall-clock** on the same Waymo scene (single L40S).

Step type	Per-step time↓
ConFixGS optimization step	$\sim$**115** ms
Difix3D+ (Wu et al., 2025) step with diffusion-fix injection	$\sim$180 s

the trusted pseudo-view evidence, reducing the required optimization horizon by an order of magnitude (from S=30,000 to S=1,000); (*iii*) confidence-modulated densification suppresses gradient accumulation in unreliable regions, so the per-step rendering cost stays low throughout optimization rather than growing as densification produces spurious Gaussians under conflicting supervision. The two backbones we evaluate exhibit complementary efficiency profiles: ConFixGS-WM has shorter total wall-clock because the WorldMirror model stays resident in process across the K_n local feedforward calls, but at higher peak GPU memory, ConFixGS-DS dispatches DepthSplat as a per-call subprocess, which adds $\sim$7 s of bridge overhead per call but keeps memory low and avoids carrying the model state across the whole pipeline. In both cases, Stage 2 (confidence estimation) and Stage 3 (confidence-modulated optimization) are negligible compared to Stage 1, confirming that the additional overhead of ConFixGS beyond the chosen backbone is small and bounded.

D Robustness, Scope, and Additional Analyses

This section reports selected analyses that directly test the reliability and scope of ConFixGS. Unless noted otherwise, all settings follow the main experiments without dataset-, scene-, or backbone-specific tuning.

D.1 Dynamic Regions and Compatibility with 4D Reconstruction

Dynamic/static evaluation. ConFixGS does not explicitly reconstruct object motion. We use official 3D boxes and speed annotations only to construct evaluation masks. Tab. 8 reports novel-view performance. Motion-induced cross-view inconsistency makes dynamic pixels less likely to receive high confidence: the fraction with $w > 0.5$ is 14% for dynamic versus 33% for static pixels on nuScenes, and 30% versus 62% on Waymo. Nevertheless, dynamic regions improve on both datasets, showing that conservative confidence assignment does not systematically degrade moving foreground.

Table 8: **Dynamic/static novel-view PSNR.** Dynamic masks are constructed from official 3D boxes and speed annotations for evaluation only.

Dataset	Region	Feedforward	ConFixGS	Gain
nuScenes	Dynamic	15.18	16.10	+0.92
nuScenes	Static	19.48	22.83	+3.35
Waymo	Dynamic	22.26	22.94	+0.68
Waymo	Static	26.77	28.67	+1.90

Proof-of-concept 4D integration. We additionally integrate ConFixGS with STORM (Yang et al., 2024a), a feedforward 4D Gaussian backbone. ConFixGS repairs only STORM's static component. Official 3D boxes exclude moving vehicles, pedestrians, and cyclists from confidence estimation and repair losses, and STORM's original time-dependent rendering is restored in the masked regions during final composition. We evaluate the three most dynamic labeled Waymo segments, where moving objects cover 5.2–9.9% of image pixels.

Table 9: **Proof-of-concept integration with STORM.** The final composition improves static content while preserving STORM's dynamic rendering by construction.

Method	PSNR all↑	PSNR dynamic↑	PSNR static↑	FID↓
STORM 4D rendering	22.82	20.40	23.09	224.3
Repaired static component only	23.87	17.06	25.45	116.4
STORM + ConFixGS (4D output)	24.80 (+1.98)	20.40 (= 0.00)	25.45 (+2.36)	107.1 (−117.2)

The repaired-static row is evaluated against the full image and therefore does not reconstruct moving objects. The final 4D output demonstrates compatibility with static–dynamic composition, but full dynamic-object repair would require temporal correspondence and motion uncertainty in the confidence formulation.

D.2 Camera-Pose Dependence

Native pose-free operation. ConFixGS can also consume poses predicted by a pose-free backbone: it improves AnySplat by 3.37 dB on Waymo and 0.93 dB on nuScenes without accessing dataset poses.

Controlled pose perturbation. To isolate sensitivity to camera geometry, we inject SE(3) noise into the input poses of five Waymo scenes while keeping images and evaluation ground truth fixed. Tab. 10 shows that both the baseline and repair degrade as camera geometry is corrupted. ConFixGS remains beneficial under small perturbations, but its gain drops sharply and approaches zero under larger errors. Since ConFixGS does not optimize poses, it cannot recover clean-pose quality once upstream geometry is corrupted. Its gain degrades gracefully rather than remaining invariant.

Table 10: **Robustness to controlled SE(3) pose noise** (PSNR in dB).

Translation / rotation noise	DepthSplat	ConFixGS-DS	Gain	ConFixGS-DS+ FID
Clean	25.75	26.77	+1.02	55.2
1 cm / 0.1°	24.26	24.49	+0.23	60.7
3 cm / 0.3°	22.03	22.14	+0.11	66.2
5 cm / 0.5°	20.92	20.90	−0.02	74.6
10 cm / 1.0°	19.32	19.25	−0.07	89.8

D.3 Hyperparameter Sensitivity

We perform controlled single-parameter sweeps on nine scenes spanning Waymo, nuScenes, and KITTI. We cache the initial reconstruction and pseudo-targets and use seeded optimization so that only one parameter changes. The same defaults are then fixed for every dataset, scene, and backbone.

Tab. 11 reports the range of mean PSNR observed over each sweep. Three confidence parameters vary performance by at most 0.05 dB. The photometric bandwidth σ_e has the largest, but smooth, effect because it directly controls the tolerated support-view discrepancy.

Table 11: **Single-parameter sensitivity** over nine scenes from three datasets. PSNR variation is the range of mean PSNR over the tested values.

Parameter	Default	Interpretation	PSNR variation
σ_e	0.10	Photometric confidence bandwidth	0.51 dB
γ	0.3	Confidence for unverifiable valid regions	0.02 dB
$\alpha_{\min}$	0.3	Minimum geometric scaffold coverage	0.05 dB
k	15	Confidence smoothing kernel size	0.02 dB

D.4 Diffusion-Prior Robustness

Corrupted prior. We inject seeded Gaussian noise into an increasing fraction of SD-Turbo VAE-decoder weights. The noise standard deviation equals that of each perturbed weight tensor. The average feedforward baseline over these nine scenes is 17.82 dB. Tab. 12 shows that confidence tracks pseudo-target degradation and consistently improves over ungated supervision. Under severe corruption, however, even gated repair may fall below the feedforward baseline.

Table 12: **Robustness to a corrupted diffusion prior** (PSNR in dB).

Perturbed weights	Confidence on	Confidence off	Pseudo-target PSNR	Mean confidence	On − off
0%	19.88	19.03	17.81	62.7%	+0.85
5%	17.41	16.33	15.05	51.0%	+1.08
10%	17.17	15.89	14.57	47.5%	+1.28
25%	16.70	15.03	13.41	38.8%	+1.67
50%	17.33	15.15	10.22	17.1%	+2.18

D.5 Input Sparsity

Initialization and support sparsity. We progressively subsample the trajectory and evaluate nine runs at each interval, as shown in Tab. 13. As both initialization quality and observable evidence weaken, the average gain decreases from 4.37 dB to approximately 1 dB. ConFixGS remains beneficial on average at every level, but one of the 36 runs falls 0.40 dB below its feedforward baseline.

Table 13: **Input-sparsity stress test** (PSNR in dB). Support counts are averaged over evaluated scenes.

Sampling interval	Approx. supports	Before repair	After repair	Gain
3	15.9	17.05	21.42	+4.37
8	6.3	16.63	18.28	+1.65
10	4.7	16.32	17.52	+1.20
15	3.7	15.32	16.41	+1.09

D.6 Comparison with Feedforward 4D Methods

Tab. 14 compares against dedicated feedforward 4D reconstruction methods under unified metric code, image resolution, and support/novel-view splits. These methods solve a broader temporal reconstruction problem, so this comparison is complementary to, rather than a replacement for, the feedforward 3DGS comparisons in the main paper. DGGT's ungated diffusion head improves the more difficult nuScenes setting but worsens Waymo FID, further motivating selective validation of generative enhancement.

Table 14: **Comparison with feedforward 4D methods** under a unified evaluation protocol.

Method	Waymo PSNR↑	nuScenes PSNR↑	Waymo FID↓	nuScenes FID↓
ConFixGS-DS	27.86	19.37	50.02	168.37
ConFixGS-DS+	25.39	19.05	48.46	119.15
STORM (Yang et al., 2024a)	25.52	17.25	133.12	265.88
DGGT (Chen et al., 2025)	24.38	16.37	58.26	186.18
DGGT+diff (Chen et al., 2025)	24.63	17.73	64.76	120.66

D.7 Broader Impact

Hallucinated content in reconstructed driving scenes can create safety risks if synthesized views are used for autonomous-driving simulation, training, or evaluation. ConFixGS reduces this risk by suppressing generated content that cannot be validated against captured observations, but confidence cannot guarantee that every reconstruction error is detected. Synthesized views should therefore not be treated as safety-certified sensor measurements. We recommend retaining confidence maps as reliability annotations for downstream use.

E Additional Comparisons and Results

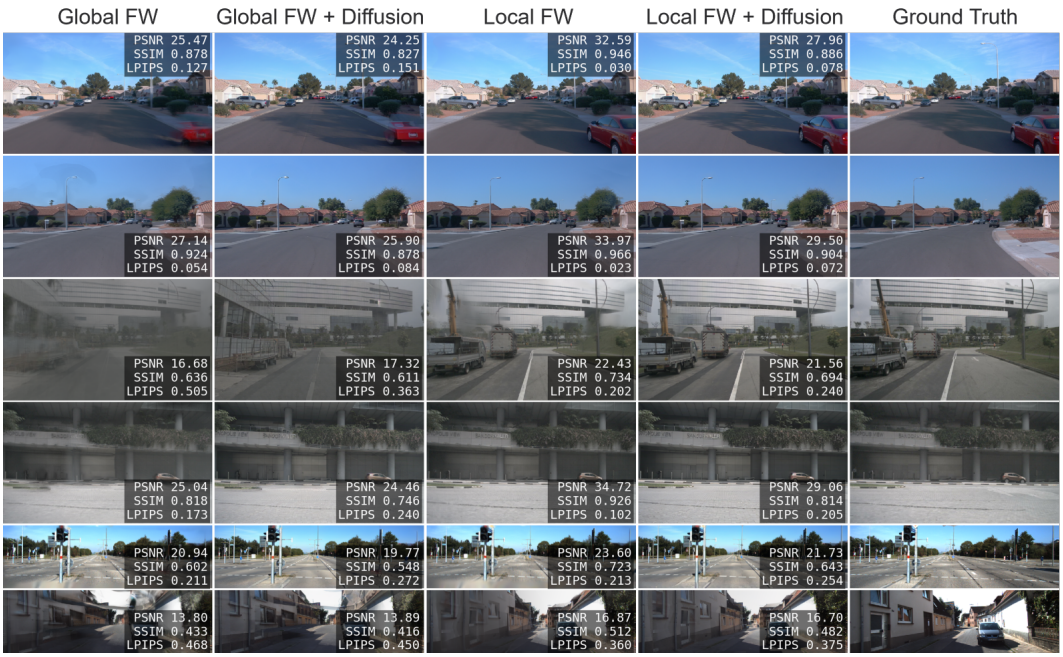

Figure 4: **Comparison of global and local feedforward 3DGS rendering on Waymo, nuScenes, and KITTI.** For each scene we compare the global feedforward (FW) rendering of $\mathcal{G}_0$, which has to reconstruct the entire trajectory from sparse, weakly overlapping support views, against the local FW rendering produced from a small subset around each novel view, and report the diffusion-enhanced version. Across all three datasets, local FW yields visibly sharper and more reliable novel views than global FW, and the diffusion-enhanced local renderings are the most repairable, supporting the key observation in Sec. 1 that motivates our local pseudo-view episode design.

E.1 Novel View Synthesis on New Trajectories

Beyond the original camera trajectory, we additionally render novel views along *out-of-distribution* viewpoints in Fig. 5, including lateral offsets (*e.g.*, 1 m and 3 m) and a virtual drone-like camera



Figure 5: **Additional novel view synthesis results** under more challenging viewpoint shifts, including lateral offsets of 1 m and 3 m, as well as a drone-style viewpoint at approximately 2.5 m height with a 20° downward pitch.

(about 2.5 m height with a 20° downward pitch). Qualitatively, ConFixGS remains stable under these more challenging trajectory shifts, producing consistent geometry and appearance without noticeable artifacts. This trajectory-level controllability suggests a promising direction for data aug-

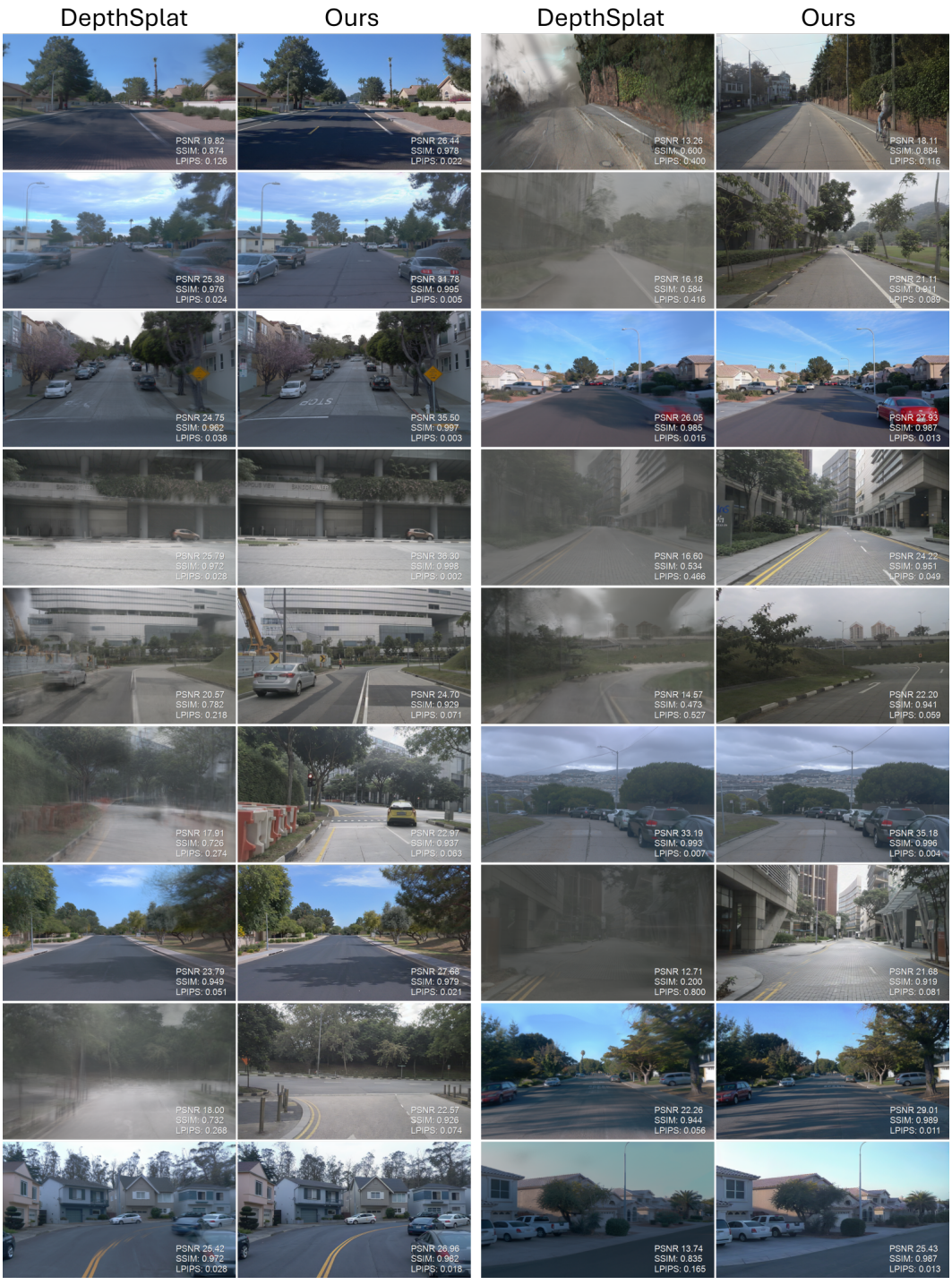

Figure 6: **Qualitative comparison of DepthSplat** (Xu et al., 2025a) before and after ConFixGS enhancement – I.

mentation in driving-centric novel view synthesis benchmarks, enabling new viewpoints that are absent from the original dataset. In particular, it opens up opportunities for (*i*) introducing drone viewpoints for driving scenes and (*ii*) synthesizing multi-agent perspectives for collaborative per-



Figure 7: **Qualitative comparison of DepthSplat** (Xu et al., 2025a) before and after ConFixGS enhancement – II.

ception (Yu et al., 2022; Zhou et al., 2025c; Zimmer et al., 2024; Xu et al., 2023; Zhou et al., 2025b; Song et al., 2024). More visual results can be found in the videos.



Figure 8: **Qualitative comparison of WorldMirror** (Liu et al., 2025) before and after ConFixGS enhancement.



Figure 9: **Qualitative comparison of AnySplat** (Jiang et al., 2025) before and after ConFixGS enhancement.

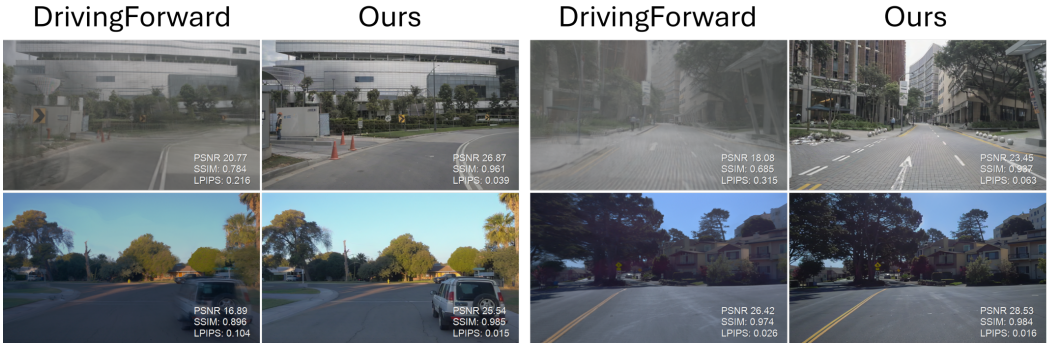

Figure 10: **Qualitative comparison of DrivingForward** (Tian et al., 2025) before and after ConFixGS enhancement.

E.2 Plug-and-Play Enhancement

We further illustrate our enhancement on DepthSplat in Figs. 6 and 7 and on WorldMirror in Fig. 8. Moreover, to demonstrate that our approach is plug-and-play, we apply ConFixGS as an enhancement module on top of additional feedforward backbones, including AnySplat (Jiang et al., 2025) and DrivingForward (Tian et al., 2025). Figs. 9 and 10 report the corresponding before and after

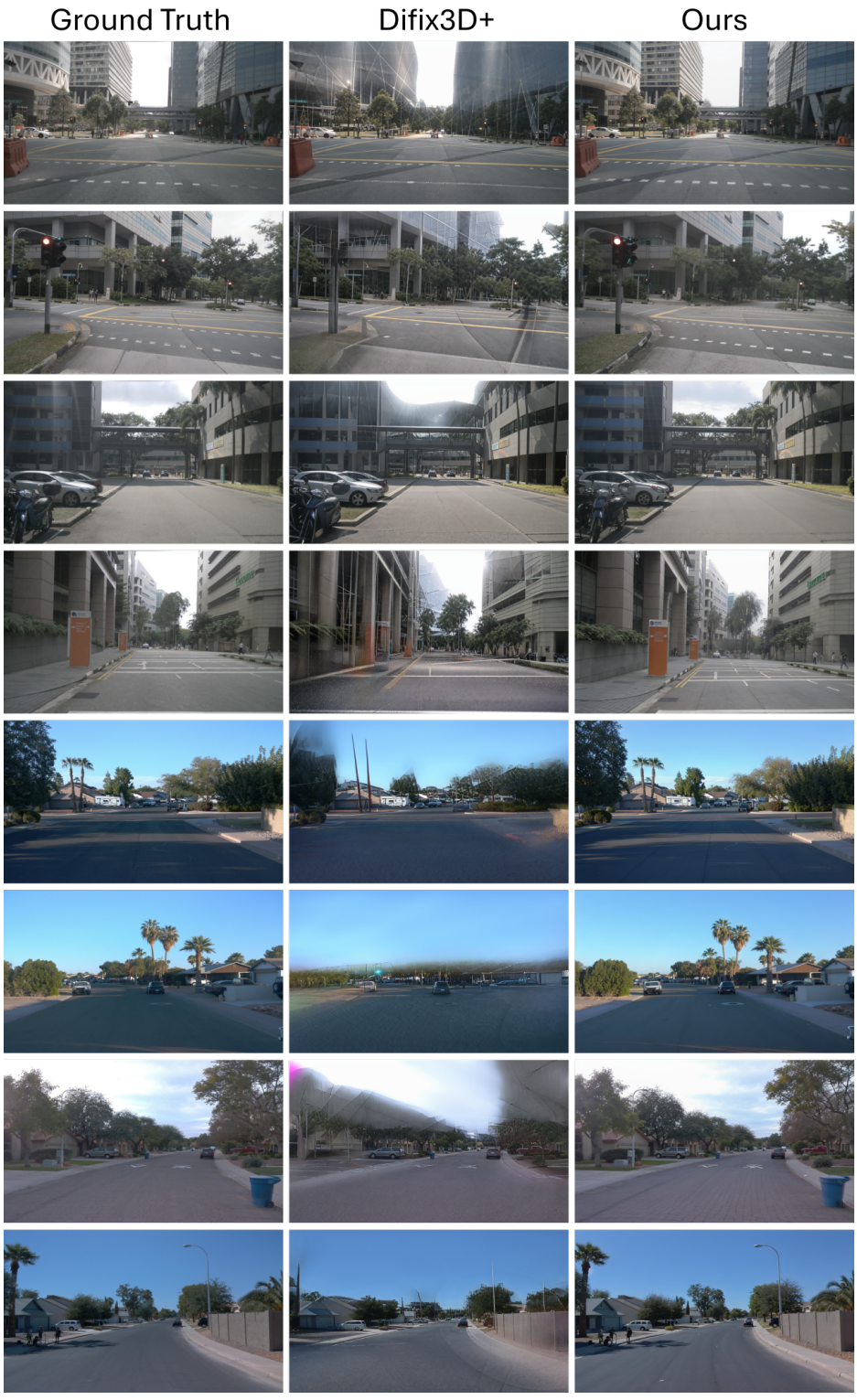

Figure 11: **Qualitative comparison with Difix3D+** (Wu et al., 2025). We use DepthSplat (Xu et al., 2025a) as the feedforward backbone for our approach.

comparisons for AnySplat and DrivingForward, respectively. Using the same experimental settings,

we conduct qualitative comparisons, showing consistent improvements across diverse backbones and highlighting the generality of our method. More visual results can be found in the videos.

E.3 Visual Comparison with Difix3D+

We further compare against Difix3D+ (Wu et al., 2025) to examine the iterative 3D optimization and diffusion fixing paradigm under sparse-view driving scenarios. Our reproduced Difix3D+ is built on gsplat 3DGS trained from scratch for 30k steps. Every 3k steps, we apply Difix to all novel view renderings and replay the refined outputs as additional supervision (10 rounds total). We set the novel view loss weight to 0.3 and the SSIM weight to 0.2. For diffusion inference, we use the official implementation with all default settings.

As shown in Fig. 11, Difix3D+ typically produces brighter, sharper textures and visually crisper renderings, but it also tends to introduce prominent ghosting patterns, while distant regions are often over-smoothed or blurred. We conjecture this stems from the tight coupling between diffusion fixing and the 3D optimization loop: once diffusion-refined signals are injected during training, intermediate Gaussians can be prematurely driven by diffusion-induced residuals before the geometry is sufficiently constrained by multi-view evidence. Applying diffusion iteratively further compounds this effect, accumulating style shifts across rounds and effectively amplifying errors rather than correcting them, which manifests as repeated textures, novel view inconsistency, and degraded far-range fidelity. In contrast, our method derives a per-pixel confidence map via reprojection-based cross-validation and uses it to gate both modulated densification and gradient updates. This effectively suppresses inconsistent diffusion residuals from being absorbed into 3D, yielding more stable structures and improved cross-view consistency while retaining visual enhancement.