# DRIVEWEAVE: FEED-FORWARD DYNAMIC SCENE GEOMETRY FOR AUTONOMOUS DRIVING

**Anonymous authors**
Paper under double-blind review

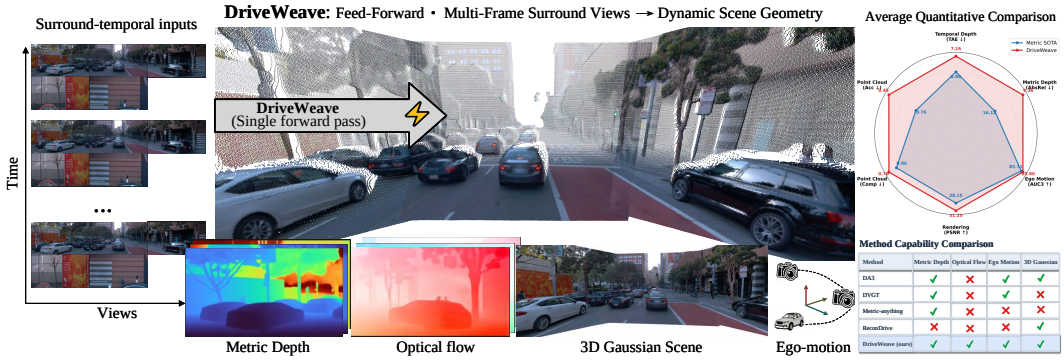


Figure 1: DriveWeave processes multi-frame surround-view images into a unified dynamic scene representation in a single forward pass, jointly predicting spatiotemporally consistent metric depth, optical flow, ego-motion, and 3D Gaussians. It outperforms existing SOTA across six dimensions.

## ABSTRACT

Reconstructing dynamic driving scenes requires jointly modeling 3D geometry and its temporal evolution. Existing paradigms typically estimate geometry and motion independently, lacking explicit mechanisms to couple their evolution and resulting in spatiotemporal scale drift and structural misalignment. We present **DriveWeave**, a unified feed-forward foundation paradigm for autonomous driving that predicts metric depth, optical flow, 6-DoF ego-motion, and 3D Gaussian primitives from surround-view videos in a single pass. To resolve the quadratic complexity of dense 4D attention, we introduce **4D Synchronized Register Attention (4D-SyncAttn)**, which leverages view-frame register synchronization to enable efficient global context aggregation. Furthermore, we formulate a **flow-driven spatiotemporal evolution mechanism** that enforces complementary differentiable photometric and geometric constraints across frames, encouraging a unified and temporally consistent representation of geometry, motion, and 3DGS rendering manifolds. Extensive evaluations across diverse driving benchmarks demonstrate that DriveWeave achieves state-of-the-art performance in metric depth estimation, 3D point cloud reconstruction, ego-motion estimation, and gaussian scene rendering with superior feed-forward efficiency.

## 1 INTRODUCTION

Recent feed-forward geometry models, such as DUSt3R (Wang et al., 2024) and VGGT (Wang et al., 2025a), have driven a paradigm shift by recovering 3D scene structure directly from visual observations via large-scale representation learning. However, these models operate primarily under static scene assumptions. Real-world autonomous driving environments are inherently four-dimensional, governed by continuous temporal evolution, complex 6-DoF ego-motion, and the dynamic trajectories of independent traffic agents (Chen et al., 2026).

Reconstructing dynamic driving scenes requires more than extending static 3D models across time. Feed-forward geometric estimation (Wang et al., 2026b;a; Chen et al., 2025a) and explicit neural



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.

rendering (e.g., 3D Gaussian Splatting) (Charatan et al., 2024; Chen et al., 2025b; Smart et al., 2024; Ye et al., 2025; Jiang et al., 2025) evolve on disjointed, parallel tracks. Reconstructing geometry independently per frame severs temporal continuity, while implicit image-level feature sharing lacks the metric grounding necessary to enforce multi-view and temporal alignment. Consequently, a fundamental gap persists: *geometry, motion, and volumetric appearance are not modeled within a singular, physically grounded framework that consistently explains the 4D dynamic scene.*

While motion cues can guide geometry or attention weights (Cong et al., 2026; Fang et al., 2026), implicit latent feature sharing alone cannot guarantee physical coherence across decoupled task heads. This motivates our core research question: **how can dense temporal correspondences be explicitly leveraged to constrain both geometric evolution and renderable appearance, ensuring that predictions across space and time describe the exact same underlying dynamic scene?** Addressing this gap mandates closed-loop, cross-frame constraints that consistently bind metric depth, ego-motion, and 3D Gaussian rendering across time into a unified 4D scene manifold.

To resolve these fundamental limitations, we present **DriveWeave**, a unified feed-forward foundation framework for dynamic scene geometry in autonomous driving. DriveWeave moves beyond disjointed pipelines by explicitly weaving temporal correspondences into the fabric of scene structure, motion, and appearance, making mutual cross-task agreement a direct source of joint supervision. Without requiring per-scene optimization, DriveWeave embodies this principle through three interlocking technical pillars:

- **Spatiotemporal Feature Coordination via 4D-SyncAttn.** To resolve the quadratic complexity of dense 4D attention, we introduce 4D Synchronized Register Attention (4D-SyncAttn). By routing global interactions through compact view-frame register tokens, 4D-SyncAttn enables scalable spatiotemporal feature exchange at a fraction of the computational cost.
- **Unified Multi-Task Geometric Representation.** DriveWeave formulates dense depth, optical flow, 6-DoF ego-motion, and explicit 3D Gaussian primitives as a consistency-coupled 4D state derived from shared spatiotemporal features. Metric depth and ego-motion anchor Gaussian primitives within a canonical coordinate frame, while optical flow governs their temporal trajectories.
- **Differentiable Cross-Frame Consistency.** We transform predicted spatiotemporal relationships into closed-loop supervisory signals. Flow-guided reprojections and predicted motion enforce complementary photometric and geometric consistency across frames, aligning scene appearance and geometry beyond isolated task-specific objectives.

In summary, our main contributions are fourfold: i) We propose **DriveWeave**, a unified feed-forward foundation model for autonomous driving that jointly predicts metric depth, optical flow, ego-motion, and 3D Gaussians from surround driving sequences in a single forward pass. ii) We design **4D-SyncAttn**, a novel register-synchronized attention mechanism that efficiently aggregates spatiotemporal context across camera views and temporal frames. iii) We formulate a flow-driven evolution mechanism with **differentiable cross-frame consistency constraints**, leveraging dense temporal correspondences to maintain coherent geometry, motion, and 3D Gaussian rendering. iv) Extensive experiments across diverse benchmarks demonstrate that DriveWeave sets a new state-of-the-art in 4D representation quality, metric geometric accuracy, and feed-forward efficiency.

## 2 Related Work

**Feed-Forward Visual Geometry.** Feed-forward 3D geometry foundation models have advanced rapidly, progressing from relative depth estimation (Ranftl et al., 2022; Yang et al., 2024a;b) to camera-aware metric depth recovery (Hu et al., 2024; Piccinelli et al., 2024; 2025). To move beyond single-view predictions, recent architectures reconstruct global scene geometry directly within a shared 3D frame, scaling from pairwise matching (Wang et al., 2024; Leroy et al., 2025) to scalable multi-view representations (Yang et al., 2025b; Wang et al., 2025a; 2026c; Lin et al., 2025). However, unconstrained visual reconstructions often lack absolute metric scale, making them ill-suited for autonomous driving, where depth topology, spatial extent, and vehicle kinematics must strictly align with a physical metric reference. Extending multi-view geometry to 4D dynamic sequences requires continuous temporal feature propagation while accounting for independent object dynamics. Existing approaches employ persistent memory banks (Wang & Agapito, 2025; Wang et al., 2025b), register-based attention (Wang et al., 2026b), or flow/motion masks (Cong et al.,

2026; Fang et al., 2026) to guide cross-frame aggregation. In surround-view autonomous driving, where minimal cross-camera overlap creates severe spatial blind spots, leveraging temporal history (Fei et al., 2024; Zuo et al., 2026), rig calibration (Jia et al., 2025), and metric prompts (Yuan et al., 2026) becomes essential to anchor scale. Unlike decoupled pipelines, **DriveWeave** unifies multi-camera temporal sequences via *4D-SyncAttn* and establishes a metrically grounded dynamic geometric representation.

**Generalizable Gaussian Scene Reconstruction.** While NeRF (Mildenhall et al., 2020) and 3D Gaussian Splatting (3DGS) (Kerbl et al., 2023) achieve high-fidelity novel-view synthesis via per-scene optimization, feed-forward neural rendering conditions predictions on input observations to bypass optimization bottlenecks (Yu et al., 2021; Wang et al., 2021; Chen et al., 2021). Feed-forward Gaussian models directly infer explicit 3D primitives for real-time rendering (Charatan et al., 2024; Chen et al., 2025b; Liu et al., 2025), recent variants even operating on unposed images by jointly estimating structure and rendering parameters (Smart et al., 2024; Ye et al., 2025; Jiang et al., 2025). In these frameworks, predicted geometry provides a critical structural scaffold, ensuring that rendered views adhere to underlying scene topology (Smart et al., 2024; Lin et al., 2025; Jiang et al., 2025). In dynamic driving environments, rendering models must capture continuous temporal evolution. While per-scene dynamic radiance fields and 4D Gaussian representations model scene deformation via deformation fields or actor kinematics (Li et al., 2021; Wu et al., 2024; Yan et al., 2025; Chen et al., 2025c), their heavy computational burden limits practical utility. Feed-forward driving models predict Gaussians directly from surround views (Tian et al., 2025) or forecast future actor states (Yu et al., 2026; Song et al., 2026), yet independent frame-wise prediction frequently induces temporal flickering and geometric drift. Departing from loosely coupled multi-task heads, **DriveWeave** leverages dense optical flow as a temporal correspondence mechanism to guide 3D Gaussian evolution and establishes cross-frame consistency that binds geometry, motion, and appearance into a unified dynamic scene manifold.

# 3 METHOD

**Method Overview.** As illustrated in Fig. 2, **DriveWeave** processes the spatiotemporal sequence $\mathcal{I}$ in a single feed-forward pass. Input images $\mathbf{I}_t^v$ are tokenized alongside camera- and time-indexed ego/scene tokens, then ingested by a Transformer backbone powered by *4D-SyncAttn* (Sec. 3.1). This backbone factorizes intra-image, cross-view, and temporal interactions into a synchronized 4D feature hierarchy. The shared features are decoded into geometrically coupled predictions: metric depth maps $\hat{\mathbf{D}}$, optical flow fields $\hat{\mathbf{F}}$, ego-motion poses $\hat{\mathbf{P}}$, and 3D Gaussian primitives $\hat{\mathcal{G}}$. During training, flow-guided temporal consistency constraints (Sec. 3.2) enforce mutual supervision between dynamic motion, renderings, and 3D geometry.

**Problem Formulation.** Given a multi-camera video sequence $\mathcal{I} = \{\mathbf{I}_t^v\}_{t=1,v=1}^{T,V}$ ($\mathbf{I}_t^v \in \mathbb{R}^{H\times W\times 3}$), DriveWeave infers a unified 4D representation parameterizing metric geometry, 6-DoF ego pose, dense motion, and renderable 3D Gaussians. For each camera $v$, the intrinsic matrix $\mathbf{K}^v \in \mathbb{R}^{3\times 3}$ and rigid camera-to-ego extrinsic transformation $\mathbf{E}^v = \mathbf{T}_{e_t \leftarrow c_t^v} \in \mathrm{SE}(3)$ are provided by the pre-calibrated sensor rig and remain fixed over time. Designating the initial ego frame $e_1$ as the coherent coordinate system across the entire $T \times V$ observation sequence, the predicted ego motion at timestamp $t$ is $\hat{\mathbf{P}}_t = \hat{\mathbf{T}}_{e_1 \leftarrow e_t} \in \mathrm{SE}(3)$, with $\hat{\mathbf{P}}_1 = \mathbf{I}$. The absolute pose of camera $v$ at timestamp $t$ is $\hat{\mathbf{T}}_{e_1 \leftarrow c_t^v} = \hat{\mathbf{P}}_t \mathbf{E}^v$, which decouples static sensor rig geometry ($\mathbf{E}^v$) from dynamic ego motion ($\hat{\mathbf{P}}_t$), anchoring all downstream dynamic geometry and motion representations to a unified metric space.

## 3.1 FEED-FORWARD MULTI-VIEW TEMPORAL GEOMETRY PREDICTION

**I. Multi-View Temporal Transformer** To explicitly model the 4D geometric dynamics of driving scenes, we develop 4D Synchronized Register Attention (4D-SyncAttn), which performs structured cross-view and cross-frame interactions over spatiotemporally organized tokens $\mathbf{Z}^0 \in \mathbb{R}^{B\times T\times V\times P\times C}$ spanning temporal frames, camera views, and intra-image patches.

**i) Intra-Image Spatial Attention.** For each observation $(t, v)$, local self-attention operates over tokens $\mathbf{Z}_t^v$ to extract instantaneous 2D spatial context prior to 4D aggregation. To preserve strict geometric layout, Rotary Positional Embeddings (RoPE) are explicitly applied to the image patch

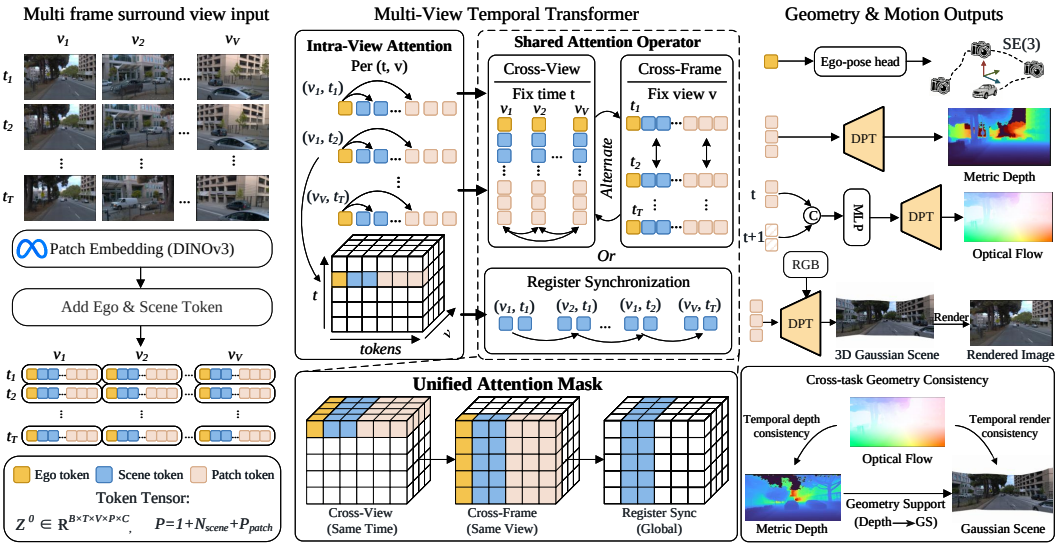


Figure 2: **Overview of DriveWeave.** Multi-frame surround-views are modeled via 4D-SyncAttn to predict ego pose, metric depth, optical flow, and 3DGS under flow-guided temporal consistency.

tokens. Conversely, global state tokens (e.g., ego and scene) remain invariant to image-plane positional encodings, allowing them to effectively abstract position-agnostic scene properties.

**ii) Cross-View and Cross-Frame Attention.** To avoid the $O((T \times V \times P)^2)$ complexity of full 4D attention, we factorize the spatiotemporal interactions while maintaining a unified attention operator: *cross-view attention* aggregates tokens across all $V$ cameras at each synchronized timestamp $t$ to capture cross-view spatial relationships. Subsequently, *cross-frame attention* aggregates tokens across all $T$ timestamps within the same camera stream $v$, tracking motion cues and temporal dynamics. By alternating transformer configurations, this factorized formulation progressively integrates 4D spatial-temporal context, eliminating dense global attention over all patch tokens.

**iii) View-Frame Register Synchronization.** While factorized attention reduces complexity, weakened global coordination may cause scale drift; we introduce synchronized registers as a global anchor. Periodically, register tokens from $T \times V$ observations are gathered into a pool and updated jointly. These synchronized registers encode shared 4D context and are redistributed to their local observations. Since only compact register tokens participate in the exchange, global scene information can be propagated across views and frames with negligible token overhead.

**II. Dense Prediction Heads** Task-specific Dense Prediction Transformers (DPT) aggregate multi-scale features from intermediate backbone stages to progressively decode the shared 4D representation into complementary predictions of metric depth, optical flow, and 3D Gaussian primitives.

**Metric Depth.** For observation $(t, v)$, a dedicated DPT decoder predicts dense metric depth $\hat{\mathbf{D}}_t^v \in \mathbb{R}_+^{H \times W}$ and pixel-wise confidence $\hat{\mathbf{\Sigma}}_t^v$ from multi-level features. To alleviate the sparsity of real-world LiDAR supervision while preserving metric scale, we combine sparse LiDAR measurements with dense structural priors. Dense pseudo-depth maps from DA3 are aligned to projected LiDAR depth using robust optimal estimation (ROE), transferring LiDAR metric scale to dense regions.

The resulting depth objective $\mathcal{L}_{\text{depth}}$ applies spatially differentiated supervision based on measurement availability: sparse pixels with valid LiDAR returns are anchored via an $\ell_1$ loss ($\mathcal{L}_{\text{lidar}}$), while remaining dense regions are supervised via a scale-invariant logarithmic loss ($\mathcal{L}_{\text{silog}}$) on the aligned pseudo-depth. To further refine local geometric boundaries, a confidence-weighted gradient loss ($\mathcal{L}_{\text{grad}}$) mitigates noise while enforcing structural sharpness.

**Optical Flow as a Dynamic Engine.** For each camera stream $v$, we concatenate multi-level Transformer features from consecutive timestamps $t$ and $t+1$ at each stage. The paired feature hierarchy is fed into a specialized flow DPT decoder to infer a forward displacement field $\hat{\mathbf{F}}_{t \to t+1}^v \in \mathbb{R}^{\mathrm{H} \times \mathrm{W} \times 2}$,

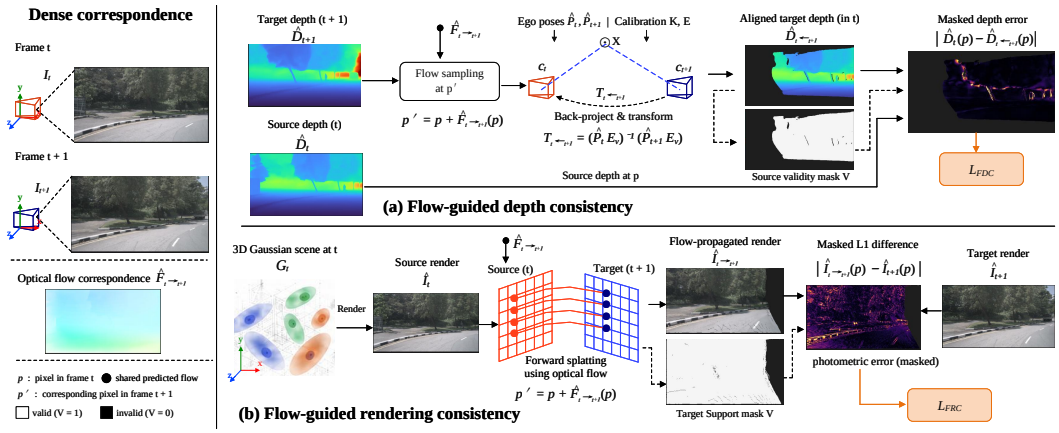


Figure 3: **Flow-guided Cross-task Temporal Consistency.** (a) Depth consistency aligns target depth with source predictions via flow-guided reprojection. (b) Gaussian render consistency forward-warps source Gaussian renderings along optical flow for temporal appearance alignment.

where $\hat{\mathbf{F}}^v_{t\to t+1}(p) = (\Delta u, \Delta v)$ specifies the image-plane displacement vector for pixel $p$. Since input features already encode global temporal context via 4D-SyncAttn, explicit feature pairing directly resolves local inter-frame correspondences while maintaining sequence-level motion awareness. Following RoMa (Edstedt et al., 2024), flow predictions are supervised using soft covisibility weighting and a generalized Charbonnier penalty (Barron, 2019).

**Metric-Grounded Gaussian Representation.** To enable feed-forward novel view synthesis without per-scene optimization, we instantiate a pixel-anchored 3D Gaussian representation. Each pixel $p$ predicts a localized primitive by fusing DPT and raw image features via residual connections. Primitive centroids are explicitly back-projected from metric depth $\hat{D}^v_t$ using camera intrinsics $\mathbf{K}^v$ ($\Pi^{-1}$) and transformed into the canonical ego-frame via pose $\hat{\mathbf{T}}_{e_1\leftarrow c^v_t}$. This depth-guided parameterization grounds centroids in metric geometry, aggregating surround-view predictions into a unified ego-centric scene representation $\hat{\mathcal{G}}_t$. The unified scene enables differentiable rendering to arbitrary target viewpoints, supervised end-to-end via photometric, structural, and perceptual losses ($\mathcal{L}_{\text{GS}}$).

### 3.2 Flow-Guided Cross-Task Temporal Consistency

Optical flow establishes dense pixel-wise correspondences across consecutive frames. We leverage these motion associations as a cross-task bridge, enforcing strict temporal consistency on 3D metric depth and feed-forward 3D Gaussian rendering. For pixel $\mathbf{p}$ in view $v$ at timestamp $t$, its displaced location at $t+1$ is given by $\mathbf{p}' = \mathbf{p} + \hat{\mathbf{F}}^v_{t\to t+1}(\mathbf{p})$.

**Flow-Guided Depth Consistency.** To prevent temporal geometric drift, optical flow identifies co-visible cross-frame pixel correspondences for rigid metric depth enforcement (Figure 3a). Target depth $\hat{\mathbf{D}}^v_{t+1}(\mathbf{p}')$ is bilinearly sampled at valid flow locations, back-projected to 3D metric space using intrinsics $\mathbf{K}^v$, transformed from target frame $c^v_{t+1}$ to source frame $c^v_t$ via estimated ego-motion, and reprojected to obtain depth $\hat{\mathbf{D}}^v_{t+1\to t}$:

$$\begin{bmatrix}\mathbf{P}_{c^v_t}\\ 1\end{bmatrix} = \mathbf{T}_{c^v_t\leftarrow c^v_{t+1}}\begin{bmatrix}\Pi^{-1}\left(\mathbf{p}', \hat{\mathbf{D}}^v_{t+1}(\mathbf{p}'); \mathbf{K}^v\right)\\ 1\end{bmatrix}, \quad \mathbf{T}_{c^v_t\leftarrow c^v_{t+1}} = (\mathbf{T}_{e_1\leftarrow c^v_t})^{-1}\mathbf{T}_{e_1\leftarrow c^v_{t+1}}, \tag{1}$$

The reprojected target depth is defined as $\hat{\mathbf{D}}^v_{t+1\to t}(\mathbf{p}) = [\mathbf{P}_{c^v_t}]_z$, over which metric depth consistency is enforced against the source prediction $\hat{\mathbf{D}}^v_t$ within co-visible regions $\Omega^v_{t\to t+1}$:

$$\mathcal{L}_{\text{FDC}} = \frac{1}{|\Omega^v_{t\to t+1}|}\sum_{\mathbf{p}\in\Omega^v_{t\to t+1}}\left\|\hat{\mathbf{D}}^v_t(\mathbf{p}) - \hat{\mathbf{D}}^v_{t+1\to t}(\mathbf{p})\right\|_1, \tag{2}$$

where $\Omega^v_{t\to t+1}$ denotes the flow-derived validity mask, retaining correspondences within image boundaries, with positive depth, and satisfying forward-backward consistency and visibility checks.

**Flow-Guided Gaussian Rendering Consistency.** Frame-wise photometric supervision treats observations independently, inducing temporal flickering in feed-forward predictions. We propagate rendered Gaussian scenes along optical flow fields to enforce cross-frame appearance continuity (Figure 3b). Specifically, rendered observation $\hat{\mathbf{I}}_t^v$ is forward-splatted via predicted flow $\hat{\mathbf{F}}_{t\to t+1}^v$ to synthesize target frame appearance:

$$\left(\hat{\mathbf{I}}_{t\to t+1}^v, \mathbf{V}_{t\to t+1}^v\right) = \mathcal{W}\left(\hat{\mathbf{I}}_t^v, \hat{\mathbf{F}}_{t\to t+1}^v\right), \tag{3}$$

where $\mathcal{W}$ denotes differentiable forward-splatting and $\mathbf{V}_{t\to t+1}^v$ is the valid visibility mask. Temporal render consistency $\mathcal{L}_{\mathrm{FRC}}$ supervises the warped rendering against target view $\hat{\mathbf{I}}_{t+1}^v$:

$$\mathcal{L}_{\mathrm{FRC}} = \frac{\sum_{\mathbf{p}} \mathbf{V}_{t\to t+1}^v(\mathbf{p}) \left\| \hat{\mathbf{I}}_{t\to t+1}^v(\mathbf{p}) - \hat{\mathbf{I}}_{t+1}^v(\mathbf{p}) \right\|_1}{\sum_{\mathbf{p}} \mathbf{V}_{t\to t+1}^v(\mathbf{p})}. \tag{4}$$

By aligning adjacent-frame renderings along optical flow trajectories, $\mathcal{L}_{\mathrm{FRC}}$ suppresses flickering and enforces appearance consistency across dynamic Gaussian representations.

### 3.3 Overall Training Objective

**Ego-Motion Estimation.** The sequence-level motion dynamics are jointly captured by the dedicated ego and scene tokens, which are decoded by a lightweight MLP pose head into an estimated ego-motion trajectory $\hat{\mathbf{P}} = \{\hat{\mathbf{P}}_t\}_{t=1}^T$. Each pose parameterizes a rigid 6-DoF transformation $\hat{\mathbf{P}}_t = \hat{\mathbf{T}}_{e_1\leftarrow e_t} \in \mathrm{SE}(3)$, anchored to the ego-first coordinate frame such that $\hat{\mathbf{P}}_1 = \mathbf{I}$. We parameterize each relative pose as a 7D vector composed of a 3D translation vector $\hat{\mathbf{t}}_t \in \mathbb{R}^3$ and a normalized rotation quaternion $\hat{\mathbf{q}}_t \in \mathbb{R}^4$. The predicted motion trajectory is supervised via a smooth $\ell_1$ objective against ground-truth vehicle poses over valid sequence timestamps.

**Overall Joint Objective.** DriveWeave unifies multi-view metric geometry, temporal optical flow correspondence, 6-DoF ego-motion tracking, and metric-grounded 3D Gaussian scene synthesis within an end-to-end differentiable pipeline. By coupling task-specific supervision with flow-guided geometric and photometric consistency regularizers, the multi-task joint objective is formulated as:

$$\mathcal{L} = \mathcal{L}_{\mathrm{depth}} + \mathcal{L}_{\mathrm{flow}} + \lambda_{\mathrm{ego}}\mathcal{L}_{\mathrm{ego}} + \mathcal{L}_{\mathrm{GS}} + \mathcal{L}_{\mathrm{FDC}} + \mathcal{L}_{\mathrm{FRC}}, \tag{5}$$

where each objective is evaluated over its valid support region. We set $\lambda_{\mathrm{ego}} = 5.0$, and all remaining loss weights to $1.0$, ensuring balanced gradient magnitudes across the task branches (**Appendix** A).

## 4 Experiments

We benchmark **DriveWeave** against state-of-the-art feed-forward geometry models across four core autonomous driving tasks: (i) metric depth estimation, (ii) 3D point cloud reconstruction, (iii) 6-DoF ego-pose trajectory estimation, and (iv) 3D Gaussian scene rendering. Additional details regarding our data curation, experimental setup, and extended results are provided in the **Appendix B&C**.

**Data Curation.** We construct a large-scale, multi-source dense geometry corpus for autonomous driving by aggregating synthetic pixel-complete depth and benchmark LiDAR-based geometry supervision from diverse driving datasets (**Appendix** Table 7). Synthetic data provides occlusion-free depth for structural edge cases, while real-world sequences provide sparse LiDAR depth observations from real driving scenarios. To synthesize dense metric targets from relative foundation priors $\tilde{\mathbf{D}}$, we align them to sparse LiDAR observations $\mathbf{D}_l$ via Robust Optimal Estimation (ROE) (Wang et al., 2025c) scale-shift alignment: $(s^*, b^*) = \arg\min_{s,b} \sum_{\mathbf{p}\in\Omega_l} \rho\left(s\tilde{\mathbf{D}}(\mathbf{p}) + b - \mathbf{D}_l(\mathbf{p})\right)$, yielding metric-aligned depth $\mathbf{D}_{\mathrm{align}} = s^*\tilde{\mathbf{D}} + b^*$. Finally, a dual-consistency spatial verification step filters out monocular estimation failures and LiDAR projection artifacts, preserving accurate, dense metric supervision for geometric learning.

### 4.1 Comparison with State-of-the-Art

**Metric Depth Estimation.** DriveWeave sets new state-of-the-art performance across all five benchmarks (Table 1), consistently obtaining the lowest $\mathrm{AbsRel}$ and highest $\delta_1$. For example, on Waymo,

Table 1: Quantitative results for metric depth estimation. * denotes relative-depth methods aligned with LiDAR depth via ROE for metric-scale recovery. : best; : second best.

| Method | Nuscenes | | | DDAD | | | Waymo | | | KITTI | | | Openscenes | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Rel ↓ | $\delta_1$ ↑ | TAE ↓ | Rel ↓ | $\delta_1$ ↑ | TAE ↓ | Rel ↓ | $\delta_1$ ↑ | TAE ↓ | Rel ↓ | $\delta_1$ ↑ | TAE ↓ | Rel ↓ | $\delta_1$ ↑ | TAE ↓ |
| CUT3R* | 13.6 | 80.9 | 21.7 | 18.8 | 73.7 | 11.9 | 13.6 | 82.4 | 15.1 | 9.1 | 88.6 | 4.2 | 15.0 | 79.7 | 21.5 |
| DA3* | 12.6 | 85.3 | 19.2 | 13.8 | 83.6 | 11.7 | 11.2 | 86.1 | 17.4 | 7.9 | 92.5 | 8.0 | 13.6 | 83.4 | 19.2 |
| VGGT-Ω* | 13.6 | 83.4 | 23.1 | 12.6 | 84.7 | 19.0 | 11.0 | 86.4 | 34.1 | 7.1 | 93.2 | 7.3 | 14.6 | 83.5 | 31.6 |
| Unidepth-v2 | 14.9 | 86.7 | 26.7 | 15.1 | 87.0 | 9.9 | 18.7 | 69.0 | 7.3 | 11.3 | 92.3 | 7.2 | 20.6 | 83.2 | 24.6 |
| FoundationGeo | 16.6 | 78.9 | 20.5 | 22.5 | 66.2 | 9.9 | 23.6 | 59.8 | 7.4 | 14.2 | 85.9 | 8.7 | 16.9 | 81.4 | 19.7 |
| DVGT | 12.3 | 85.2 | 17.0 | 21.3 | 73.3 | 8.2 | 17.9 | 80.2 | 4.5 | 15.1 | 82.1 | 6.9 | 16.2 | 82.5 | 17.3 |
| SurroundNEXO | 15.7 | 81.8 | 14.9 | 15.3 | 83.0 | 6.4 | 19.9 | 67.4 | 4.2 | 9.9 | 89.7 | 4.1 | 14.9 | 84.5 | 15.3 |
| Metric-anything | 16.8 | 81.7 | 22.1 | 15.9 | 84.5 | 9.1 | 23.9 | 61.6 | 8.3 | 8.4 | 91.7 | 9.2 | 19.1 | 81.8 | 20.3 |
| DriveWeave | **10.0** | **88.5** | **12.2** | **10.8** | **88.6** | **5.5** | **9.6** | **90.2** | **3.2** | **5.0** | **95.4** | **2.8** | **11.4** | **88.2** | **12.1** |

DriveWeave reduces $\mathrm{AbsRel}$ to $9.6\%$, outperforming DVGT ($17.9\%$) and SurroundNEXO ($19.9\%$). DriveWeave yields exceptional temporal stability, achieving the lowest Temporal Alignment Error ($\mathrm{TAE}$) across all datasets. The significant reduction in post-alignment inter-frame depth discrepancies show that our joint 4D spatiotemporal formulation effectively mitigates scale drift, suppressing temporal flickering and delivering smooth, geometrically consistent depth trajectories. See **Appendix** Fig. 8 for qualitative comparisons of metric depth estimation under diverse driving scenes.

**Ego-motion Estimation.** As shown in Table 2, DriveWeave consistently achieves the best ego-motion estimation performance across nuScenes, DDAD, and Waymo under both AUC3 and AUC30. The larger improvements under the stricter $3°$ threshold indicate that DriveWeave provides more accurate fine-grained motion estimation rather than only recovering coarse trajectory trends. These consistent gains across different datasets support the effectiveness of the unified spatiotemporal representation for ego-motion recovery. **Appendix** Fig. 12 shows the single-pass ego-motion estimation across varying sequence length. DriveWeave maintains tight alignment with ground-truth trajectories.

Table 2: Quantitative results for ego-motion estimation. AUC3/AUC30 denote the area under the pose-error recall curve at thresholds of $3°$ and $30°$.

| Method | Nuscenes | | DDAD | | Waymo | |
|---|---|---|---|---|---|---|
| | AUC3↑ | AUC30↑ | AUC3↑ | AUC30↑ | AUC3↑ | AUC30↑ |
| CUT3R | 41.39 | 80.91 | 20.32 | 83.24 | 36.71 | 74.06 |
| DA3 | 62.81 | 84.89 | 81.98 | 92.97 | 73.57 | 82.18 |
| VGGT-Ω | 70.99 | 85.09 | 77.13 | 92.76 | 75.78 | 82.61 |
| DVGT | 80.17 | 85.84 | 91.22 | 94.45 | 79.75 | 82.55 |
| DriveWeave | **82.19** | **86.67** | **94.27** | **95.63** | **81.24** | **83.81** |

**Point Cloud Reconstruction.** Table 3 evaluates the global 3D reconstruction quality after lifting multi-view predictions into a common metric coordinate system. DriveWeave sets new state-of-the-art performance for global 3D geometric reconstruction across all five benchmarks (Table 3), obtaining the lowest accuracy error ($\mathrm{Acc}$) on nuScenes ($0.53\,\mathrm{m}$), DDAD ($0.53\,\mathrm{m}$), Waymo ($0.52\,\mathrm{m}$), KITTI ($0.22\,\mathrm{m}$), and OpenScenes ($0.47\,\mathrm{m}$). DriveWeave delivers superior completeness ($\mathrm{Comp}$) on Waymo, KITTI, and OpenScenes while remaining highly competitive on nuScenes and DDAD, confirming that multi-view predictions are accurately consolidated across space and time rather than merely producing locally plausible depth. Crucially, DriveWeave processes a full 96-image sequence (16 frames $\times$ 6 views) in just $1.69\,\mathrm{s}$, establishing an optimal Pareto frontier between geometric accuracy and computational throughput. See **Appendix** Fig. 7 and 9 for qualitative visualizations of accumulated temporal point clouds and surround-view metric reconstructions, respectively. DriveWeave does not merely recover visually plausible point clouds, but establishes a metrically grounded and globally consistent dynamic scene representation from surround-view observations.

**Gaussian Scene Reconstruction.** DriveWeave sets new state-of-the-art feed-forward 3D Gaussian reconstruction and rendering on nuScenes and Waymo (Table 4), achieving top $\mathrm{PSNR}$, $\mathrm{SSIM}$, and $\mathrm{LPIPS}$ across both per-view and unified-view protocols. The performance advantage is particularly evident under unified-view rendering (reaching $0.92$ $\mathrm{SSIM}$ and $0.14$ $\mathrm{LPIPS}$ on both benchmarks), confirming globally coherent Gaussian fusion across multi-camera views and temporal frames. Despite jointly predicting metric geometry, 6-DoF ego-motion, and 3DGS appearance, DriveWeave executes the complete rendering pipeline in just $0.58\,\mathrm{s}$, delivering superior feed-forward inference throughput. **Appendix** Fig. 10 and 11 illustrate qualitative Gaussian reconstruction and novel-view rendering, demonstrating not only the effectiveness, but also superior cross-view rendering capability of DriveWeave in learning a geometrically reliable scene representation.

Table 3: Quantitative results for 3D point cloud reconstruction. Time denotes the end-to-end inference time of each method for point cloud reconstruction from 96 images (16 frames × 6 views).

| Method | Nuscenes | | DDAD | | Waymo | | KITTI | | Openscenes | | Time |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Acc↓ | Comp↓ | Acc↓ | Comp↓ | Acc↓ | Comp↓ | Acc↓ | Comp↓ | Acc↓ | Comp↓ | |
| CUT3R* | 0.68 | 1.81 | 0.76 | 3.78 | 0.58 | 1.24 | 0.42 | 0.93 | 0.70 | 1.85 | 5.77s |
| DA3* | 0.73 | 1.27 | 0.71 | 1.60 | 0.66 | 0.96 | 0.37 | 0.43 | 0.73 | 1.27 | 8.63s |
| VGGT-Ω* | 0.70 | 0.98 | 0.84 | 0.94 | 0.64 | 0.88 | 0.39 | 0.47 | 0.75 | 0.90 | 3.63s |
| Unidepth-v2 | 0.81 | **0.88** | 0.77 | 0.83 | 1.18 | 1.44 | 0.37 | 0.42 | 0.88 | 0.71 | 2.16s |
| FoundationGeo | 0.91 | 1.13 | 0.98 | 1.02 | 1.38 | 2.16 | 0.54 | 0.68 | 0.82 | 0.77 | 6.30s |
| DVGT | 0.73 | 1.09 | 0.98 | 1.17 | 1.13 | 1.24 | 0.46 | 0.63 | 0.96 | 0.94 | 2.61s |
| SurroundNEXO | 0.84 | 1.20 | 0.75 | 1.25 | 1.10 | 1.79 | 0.44 | 0.65 | 0.68 | 0.97 | 7.98s |
| Metric-anything | 0.92 | 1.04 | 0.82 | **0.80** | 1.43 | 1.59 | 0.43 | 0.56 | 0.91 | 0.73 | 4.72s |
| DriveWeave | **0.53** | 0.99 | **0.53** | 1.06 | **0.52** | **0.77** | **0.22** | **0.29** | **0.47** | **0.69** | **1.69s** |

Table 4: Feed-forward Gaussian scene rendering. Benchmarked under per-view and unified-view protocols. Latency denotes end-to-end runtime of the complete unified-view pipeline.

| Method | Nuscenes | | | | | | Waymo | | | | | | Time |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Per-View Reconstruction | | | Unified-view Render | | | Per-View Reconstruction | | | Unified-view Render | | | |
| | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ | |
| DrivingForward | 19.5 | 0.78 | 0.26 | 21.2 | 0.77 | 0.27 | 20.0 | 0.76 | 0.28 | 21.3 | 0.76 | 0.29 | **0.40s** |
| DA3 | 17.5 | 0.48 | 0.56 | 17.2 | 0.48 | 0.58 | 17.9 | 0.50 | 0.55 | 17.6 | 0.50 | 0.57 | 1.09s |
| Envision4D | 20.3 | 0.64 | 0.41 | 19.7 | 0.55 | 0.46 | 25.2 | 0.74 | 0.25 | 26.0 | 0.76 | 0.22 | 2.04s |
| Uni-splat | 24.7 | 0.73 | 0.35 | 23.1 | 0.71 | 0.37 | 28.6 | 0.84 | 0.21 | 27.5 | 0.83 | 0.22 | 0.77s |
| Recon-drive | 29.3 | 0.87 | **0.12** | 27.4 | 0.84 | 0.15 | 29.5 | 0.91 | **0.13** | 28.9 | 0.87 | 0.16 | 0.88s |
| DriveWeave | **30.9** | **0.94** | 0.13 | **31.2** | **0.92** | **0.14** | **30.8** | **0.92** | **0.13** | **31.3** | **0.92** | **0.14** | 0.58s |

## 4.2 Ablation Experiments

**Spatiotemporal Attention Mechanisms.** Table 5 evaluates different spatiotemporal attention variants. While naive factorized attention (sequential cross-camera and cross-frame self-attention) reduces latency from 4.61 s to 1.68 s relative to full 4D attention, it suffers geometric accuracy degradation due to information bottlenecks across non-adjacent views and timestamps. Our 4D-SyncAttn bridges these factorized streams using lightweight synchronized register tokens to restore global spatiotemporal context propagation. Consequently, 4D-SyncAttn achieves superior performance across all trajectory ($\mathrm{AUC}_3$), depth ($\mathrm{AbsRel}$, $\delta_1$), and 3D reconstruction ($\mathrm{Acc}$) metrics, while further slashing runtime to 1.17 s. This confirms that compact register synchronization enables efficient global context exchange without the quadratic cost of dense 4D attention.

Table 5: **Spatiotemporal attention ablation on nuScenes.** Evaluates global interaction (I+G), factorized view/temporal interaction (I+C/V), and synchronized registers (I+C/V+SyncR).

| Method | AUC3↑ | Rel↓ | $\delta_1$ ↑ | Acc↓ | Comp↓ | Time |
|---|---|---|---|---|---|---|
| I+G | 80.68 | 10.8 | 87.6 | 0.55 | **0.95** | 4.61s |
| I+C/V | 78.95 | 10.6 | 88.2 | 0.59 | 1.06 | 1.68s |
| I+C/V+SyncR | **82.19** | **10.0** | **88.5** | **0.53** | 0.99 | **1.17s** |

**Consistency Mechanics Under Dynamic Driving Environments.** Fig. 4 visualizes our consistency behavior under complex driving dynamics. For geometric depth, flow correspondences and predicted ego-motion align cross-frame geometry, while flow-derived validity masks filter invalid correspondences and disoccluded regions, preventing unreliable matches from corrupting the geometric consistency constraint. For photometric rendering, flow propagation maintains appearance alignment across both camera ego-motion and object dynamics while masking newly exposed pixels. Consequently, residual error maps remain narrowly localized around high-gradient motion boundaries and severe occlusions, demonstrating robust spatiotemporal consistency across both static layouts and dynamic traffic scenarios.

**Flow-Guided Cross-Task Temporal Consistency.** Table 6 validates the complementary contributions of Flow-Depth Consistency (FDC) and Flow-Rendering Consistency (FRC). FDC explicitly couples optical flow, 6-DoF ego-motion, and metric depth via cross-frame reprojection, enforcing 3D geometric and trajectory coherence. Conversely, FRC photometrically propagates 3D Gaussian renderings along optical flow to enforce multi-frame appearance continuity on the 2D image manifold. Operating on orthogonal dimensions of the 4D representation, their joint optimization

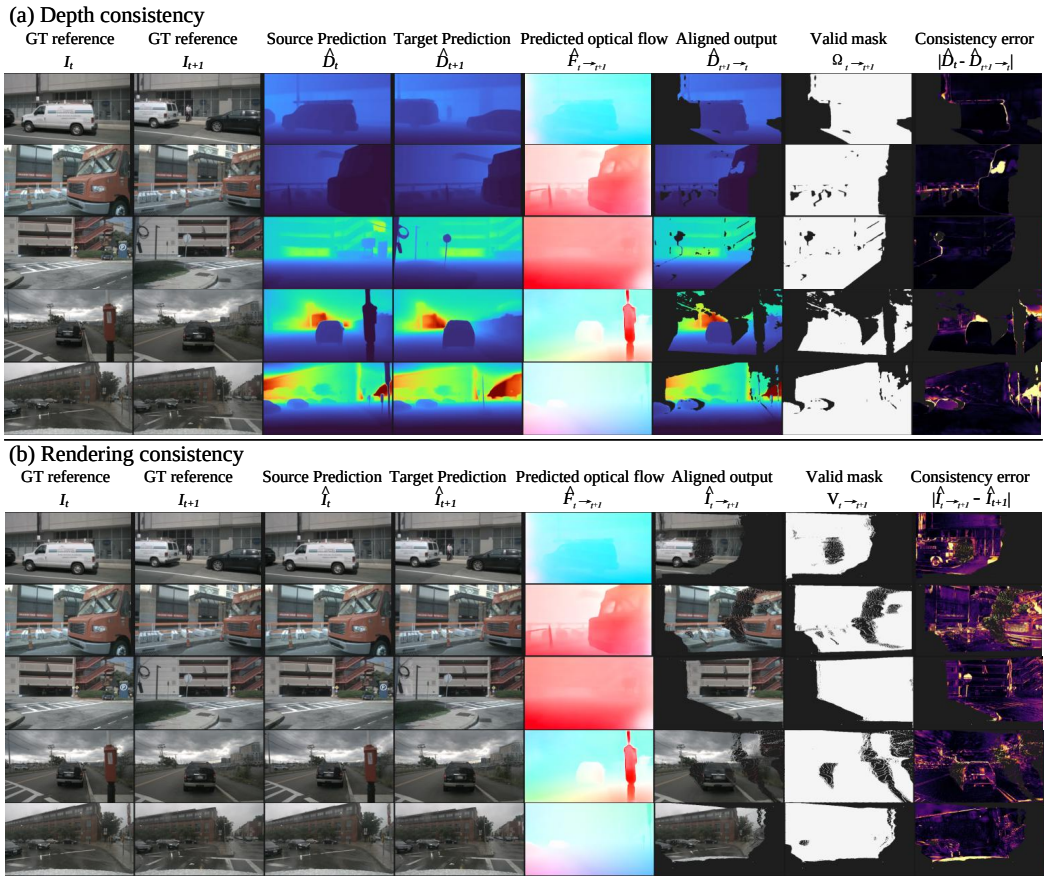


Figure 4: **Flow-guided cross-task temporal consistency mechanics. (a) FDC** aligns target depth via flow correspondences and ego-motion. **(b) FRC** propagates Gaussian renderings along optical flow. Valid masks and residual maps visualize geometric and photometric errors.

Table 6: Ablation of flow-guided temporal consistency. **Bold**: best; Underline: second best.

| Method | FDC | FRC | Nuscenes | | | | | Waymo | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Rel ↓ | TAE↓ | Acc ↓ | AUC3 ↑ | PSNR ↑ | Rel ↓ | TAE↓ | Acc ↓ | AUC3 ↑ | PSNR ↑ |
| 1 | ✗ | ✗ | 11.2 | 17.8 | 0.60 | 72.42 | 29.8 | 11.1 | 6.9 | 0.66 | 74.31 | 30.1 |
| 2 | ✓ | ✗ | 10.7 | **11.7** | 0.58 | 80.53 | 30.4 | 10.1 | 3.5 | 0.61 | 78.59 | 30.7 |
| 3 | ✗ | ✓ | 10.9 | 16.5 | 0.54 | 77.45 | 30.9 | 10.6 | 6.5 | 0.54 | 77.15 | **31.4** |
| 4 | ✓ | ✓ | **10.0** | 12.2 | **0.53** | **82.19** | **31.2** | **9.6** | **3.2** | **0.52** | **81.24** | 31.3 |

establishes a correspondence-driven bridge across geometry, motion, and appearance, achieving optimal performance on metric depth estimation, scene reconstruction, and novel-view rendering.

## 5 Conclusion

We present **DriveWeave**, a feed-forward multi-view temporal geometry foundation model for autonomous driving. Powered by 4D Synchronized Register Attention and flow-driven spatiotemporal evolution mechanism, DriveWeave seamlessly unifies metric depth, optical flow, 6-DoF ego-motion, and Gaussian scene representations under explicit photometric and geometric constraints. By leveraging flow-guided cross-task temporal consistency constraints, DriveWeave explicitly binds geometry, motion, and appearance through spatiotemporal correspondences, enabling coherent 4D scene reconstruction across multi-camera surround-view sequences. Extensive evaluations demonstrate that DriveWeave establishes new state-of-the-art performance across metric geometry, motion estimation, and Gaussian rendering, paving the way for scalable 4D geometry foundation models.

## AI USE STATEMENT

In this work, generative AI tools were used exclusively for [language editing and manuscript polishing]. Generative AI was not used to [generate synthetic datasets, formulate mathematical claims or proofs, design methodologies, or perform experiments]. The authors have thoroughly audited all AI-assisted revisions and take full responsibility for the integrity, validity, and accuracy of the final manuscript and its contents.

## REFERENCES

Jonathan T. Barron. A general and adaptive robust loss function. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4326–4334, 2019. doi: 10.1109/CVPR.2019.00446.

Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2, 2020. URL https://arxiv.org/abs/2001.10773.

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 (CVPR)*, June 2020.

Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, and James Hays. Argoverse: 3d tracking and forecasting with rich maps. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8740–8749, 2019. doi: 10.1109/CVPR.2019.00895.

David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. Pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19457–19467, 2024. doi: 10.1109/CVPR52733.2024.01840.

Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14124–14133, 2021.

Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22831–22840, June 2025a.

Xiaoxue Chen, Ziyi Xiong, Yuantao Chen, Gen Li, Nan Wang, Hongcheng Luo, Long Chen, Haiyang Sun, Bing Wang, Guang Chen, Hongyang Li, Ya-Qin Zhang, Hangjun Ye, and Hao Zhao. Dggt: Feedforward 4d reconstruction of dynamic driving scenes using unposed images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1265–1276, June 2026.

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 *Computer Vision – ECCV 2024*, pp. 370–386, 2025b. ISBN 978-3-031-72664-4.

Ziyu Chen, Jiawei Yang, Jiahui Huang, Riccardo de Lutio, Janick Martinez Esturo, Boris Ivanovic, Or Litany, Zan Gojcic, Sanja Fidler, Marco Pavone, Li Song, and Yue Wang. Omnire: Omni urban scene reconstruction. In *International Conference on Learning Representations*, volume 2025, pp. 85508–85527, 2025c.

Zhongxiao Cong, Qitao Zhao, Minsik Jeon, and Shubham Tulsiani. Flow3r: Factored flow prediction for scalable visual geometry learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 438–447, June 2026.

OpenScene Contributors. Openscene: The largest up-to-date 3d occupancy prediction benchmark in autonomous driving. https://github.com/OpenDriveLab/OpenScene, 2023.

Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. RoMa: Robust Dense Feature Matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2024.

Gasser Elazab, Frank Neuhaus, Tilman Koß, Malte Splietker, Aditya Date, Michael Unterreiner, Maximilian Jansen, and Olaf Hellwich. Card: A multi-modal automotive dataset for dense 3d reconstruction in challenging road topography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17820–17830, June 2026.

Juntong Fang, Zequn Chen, Weiqi Zhang, Donglin Di, Xuancheng Zhang, Chengmin Yang, and Yu-Shen Liu. More: Motion-aware feed-forward 4d reconstruction transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28914–28924, June 2026.

Xin Fei, Wenzhao Zheng, Yueqi Duan, Wei Zhan, Masayoshi Tomizuka, Kurt Keutzer, and Jiwen Lu. Driv3r: Learning dense 4d reconstruction for autonomous driving. *arXiv preprint arXiv:2412.06777*, 2024.

Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013.

Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10579–10596, 2024. doi: 10.1109/TPAMI.2024.3444912.

Arpit Jadon, Haoran Wang, Phillip Thomas, Michael Stanley, S. Nathaniel Cibik, Rachel Laurat, Omar Maher, Lukas Hoyer, Ozan Unal, and Dengxin Dai. Realdrivesim: A realistic multi-modal multi-task synthetic dataset for autonomous driving. In *2025 IEEE Intelligent Vehicles Symposium (IV)*, pp. 316–323, 2025. doi: 10.1109/IV64158.2025.11097728.

Xiaosong Jia, Yanhao Liu, Yu Hong, Renqiu Xia, Junqi You, Bin Sun, Zhihui Hao, and Junchi Yan. Drivevggt: Calibration-constrained visual geometry transformers for multi-camera autonomous driving. *arXiv preprint arXiv:2511.22264*, 2025.

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.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4), July 2023. ISSN 0730-0301. doi: 10.1145/3592433.

Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3d with mast3r. In *Computer Vision – ECCV 2024*, pp. 71–91, 2025. ISBN 978-3-031-73220-1.

Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6494–6504, 2021. doi: 10.1109/CVPR46437.2021.00643.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Muxin Liu, Xiaoyang Lyu, Tianhe Ren, Peng Dai, Xiaoshan Wu, Zhiyue Zhang, Jiaqi Zhang, Jiehong Lin, Shaoshuai Shi, and Xiaojuan Qi. Foundationgeo: Learning spatial pixel-wise fields for monocular metric geometry. In *European Conference on Computer Vision*, pp. 353–371. Springer, 2026.

Tianqi Liu, Guangcong Wang, Shoukang Hu, Liao Shen, Xinyi Ye, Yuhang Zang, Zhiguo Cao, Wei Li, and Ziwei Liu. Mvsgaussian: Fast generalizable gaussian splatting reconstruction from multi-view stereo. In *European Conference on Computer Vision*, pp. 37–53. Springer, 2025.

Baorui Ma, Jiahui Yang, Donglin Di, Xuancheng Zhang, Jianxun Cui, Hao Li, Xie Yan, and Wei Chen. Metricanything: Scaling metric depth pretraining with noisy heterogeneous sources. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2026.

Goodarz Mehr and Azim Eskandarian. Simbev: A synthetic multi-task multi-sensor driving data generation tool and dataset. *arXiv preprint arXiv:2502.01894*, 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. In *Computer Vision – ECCV 2020*, pp. 405–421, 2020. ISBN 978-3-030-58452-8.

Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. UniDepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. UniDepthV2: Universal monocular metric depth estimation made simpler, 2025. URL https://arxiv.org/abs/2502.20110.

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3), 2022.

Ahmed Rida Sekkat, Rohit Mohan, Oliver Sawade, Elmar Matthes, and Abhinav Valada. Amodal-synthdrive: A synthetic amodal perception dataset for autonomous driving, 2024. URL https://arxiv.org/abs/2309.06547.

Chen Shi, Shaoshuai Shi, Xiaoyang Lyu, Chunyang Liu, Kehua Sheng, Bo Zhang, and Li Jiang. Unisplat: Unified spatio-temporal fusion via 3d latent scaffolds for dynamic driving scene reconstruction. In *International Conference on Learning Representations*, volume 2026, pp. 125218–125235, 2026.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. URL https://arxiv.org/abs/2508.10104.

Brandon Smart, Chuanxia Zheng, Iro Laina, and Victor Adrian Prisacariu. Splatt3r: Zero-shot gaussian splatting from uncalibrated image pairs. 2024. URL https://arxiv.org/abs/2408.13912.

Qi Song, Yifei He, Chi Zhang, Zheng Fu, Xuhe Zhao, Mengmeng Yang, Kun Jiang, Rui Huang, and Diange Yang. Envision4d: Envisioning visual futures via feed-forward 4d gaussian splatting for autonomous driving. *arXiv preprint arXiv:2606.10656*, 2026.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurélien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2443–2451, 2020. doi: 10.1109/CVPR42600.2020.00252.

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.

Haonan Wang, Hanyu Zhou, Haoyue Liu, and Luxin Yan. 4d-vggt: A spatiotemporal foundation model for dynamic scene geometry estimation. In *Computer Vision – ECCV 2026*, pp. 112–134, 2026a. ISBN 978-3-032-37035-8.

Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory. In *2025 International Conference on 3D Vision (3DV)*, pp. 78–89, 2025. doi: 10.1109/3DV66043.2025.00013.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306, 2025a. doi: 10.1109/CVPR52734.2025.00499.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026b.

Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4688–4697, 2021. doi: 10.1109/CVPR46437.2021.00466.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10510–10522, 2025b. doi: 10.1109/CVPR52734.2025.00983.

Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5261–5271, 2025c.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20697–20709, 2024. doi: 10.1109/CVPR52733.2024.01956.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. \pi^3: Permutation-equivariant visual geometry learning. In *International Conference on Learning Representations*, volume 2026, pp. 10481–10497, 2026c.

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

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 (CVPR)*, pp. 20310–20320, June 2024.

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 *Computer Vision – ECCV 2024*, pp. 156–173, 2025. ISBN 978-3-031-73464-9.

Honghui Yang, Di Huang, Wei Yin, Chunhua Shen, Haifeng Liu, Xiaofei He, Binbin Lin, Wanli Ouyang, and Tong He. Depth any video with scalable synthetic data. In *International Conference on Learning Representations*, volume 2025, pp. 97335–97349, 2025a.

Jianing Yang, Alexander Sax, Kevin J. Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025b.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10371–10381, 2024a. doi: 10.1109/CVPR52733.2024.00987.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *Advances in Neural Information Processing Systems*, volume 37, pp. 21875–21911, 2024b. doi: 10.52202/079017-0688.

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. In *International Conference on Learning Representations*, volume 2025, pp. 54009–54033, 2025.

Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4576–4585, 2021. doi: 10.1109/CVPR46437.2021.00455.

Haibao Yu, Kuntao Xiao, Jiahang Wang, Ruiyang Hao, Guoran Hu, Yuxin Huang, Haifang Qin, Bowen Jing, Yuntian Bo, and Ping Luo. Recondrive: Fast feed-forward 4d gaussian splatting for autonomous driving scene reconstruction. In *https://arxiv.org/abs/2603.07552*, 2026.

Shuai Yuan, Runxi Tang, Yuzhou Ji, Fudong Ge, Hanshi Wang, Yifei Wang, Xianming Zeng, Jianyun Xu, Xingliang Liu, Yanfeng Wang, and Zhipeng Zhang. Surroundnexo: Ego-centric metric bridging for spatially consistent geometry in autonomous driving. *arXiv preprint arXiv:2606.16960*, 2026.

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

Yuchen Zhang, Nikhil Keetha, Chenwei Lyu, Bhuvan Jhamb, Yutian Chen, Yuheng Qiu, Jay Karhade, Shreyas Jha, Yaoyu Hu, Deva Ramanan, Sebastian Scherer, and Wenshan Wang. Ufm: A simple path towards unified dense correspondence with flow. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 109372–109399, 2025. doi: 10.52202/085713-3648.

Sicheng Zuo, Zixun Xie, Wenzhao Zheng, Shaoqing Xu, Fang Li, Shengyin Jiang, Long Chen, Zhi-Xin Yang, and Jiwen Lu. Dvgt: Driving visual geometry transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14658–14668, June 2026.

# A Details of Geometry Training

This section provides the detailed task-specific objectives omitted from the main paper for compactness. DriveWeave jointly supervises metric depth, optical flow, ego-motion, and Gaussian reconstruction, while the flow-guided cross-task consistency terms $\mathcal{L}_{\mathrm{FDC}}$ and $\mathcal{L}_{\mathrm{FRC}}$ are described in the main content.

**Metric Depth Supervision.** For each observation $(t, v)$, the depth head predicts a dense metric depth map $\hat{\mathbf{D}}_t^v$ alongside an associated pixel-wise confidence map $\hat{\mathbf{\Sigma}}_t^v$. To unify the precise metric scale of sparse LiDAR with dense structural supervision, we leverage valid LiDAR projections $\mathbf{D}^{\mathrm{lidar}}$ as metric anchors and scale-aligned pseudo-depth $\mathbf{D}^{\mathrm{pseudo}}$ across remaining unobserved regions. The depth supervision objective is formulated as:

$$\mathcal{L}_{\mathrm{depth}} = \mathcal{L}_{\mathrm{lidar}}(\hat{\mathbf{D}}, \mathbf{D}^{\mathrm{lidar}}) + \lambda_{\mathrm{silog}}\mathcal{L}_{\mathrm{silog}}(\hat{\mathbf{D}}, \mathbf{D}^{\mathrm{pseudo}}) + \mathcal{L}_{\mathrm{grad}}(\hat{\mathbf{D}}, \mathbf{D}^{\mathrm{pseudo}}; \hat{\mathbf{\Sigma}}), \tag{6}$$

where $\mathcal{L}_{\mathrm{lidar}}$ is an $\ell_1$ loss evaluated over valid LiDAR pixels, and $\mathcal{L}_{\mathrm{silog}}$ provides dense scale-invariant structural supervision. To preserve local geometric discontinuities while reducing the influence of unreliable pseudo-labels, the confidence-weighted gradient term is:

$$\mathcal{L}_{\mathrm{grad}} = \sum_{p \in \Omega_{\mathrm{dense}}} \left( \hat{\Sigma}_p \left\| \nabla \hat{D}_p - \nabla D_p^{\mathrm{pseudo}} \right\|_1 - \lambda_c \log \hat{\Sigma}_p \right), \tag{7}$$

where $\Omega_{\mathrm{dense}}$ denotes the valid dense-supervision region. The first term encourages structural agreement between predicted and pseudo depth gradients, while the logarithmic regularizer prevents trivial confidence suppression. To balance LiDAR and pseudo-depth supervision, we discard the top 10% highest-error pixels during loss aggregation.

**Optical Flow Supervision.** For consecutive timestamps $t$ and $t+1$, the flow decoder predicts a dense displacement field $\hat{\mathbf{F}}_{t\to t+1}^v \in \mathbb{R}^{H\times W\times 2}$. Following robust dense correspondence regression, flow supervision is weighted by a soft covisibility mask $\mathbf{C}_{t\to t+1}^v$:

$$\mathcal{L}_{\mathrm{flow}} = \frac{\sum_{t,v,p} C_{t\to t+1}^v(p)\ell_{\mathrm{robust}}\left( \left\| \phi\left(\hat{\mathbf{F}}_{t\to t+1}^v(p)\right) - \phi\left(\mathbf{F}_{t\to t+1}^v(p)\right) \right\|_2 \right)}{\sum_{t,v,p} C_{t\to t+1}^v(p)}, \tag{8}$$

where $C_{t\to t+1}^v \in [0,1]$ denotes the soft covisibility mask. To compress the vast dynamic range of driving speeds, regression is performed in a sign-preserving logarithmic space, $\phi(x) = \mathrm{sign}(x)\log(1+|x|)$, ensuring robust gradients across both subtle background motion and rapid ego-vehicle maneuvers.

**Metric-Grounded 3D Gaussian Supervision.** For each pixel $p$ in camera $v$ at timestamp $t$, the predicted Gaussian primitive is parameterized as:

$$\hat{\mathbf{g}}_{t,v,p} = \left(\hat{\boldsymbol{\mu}}_{t,v,p}^{e_1}, \hat{\mathbf{s}}_{t,v,p}, \hat{\mathbf{q}}_{t,v,p}^{e_1}, \hat{\alpha}_{t,v,p}, \hat{\mathbf{c}}_{t,v,p}^{e_1}\right), \quad \hat{\boldsymbol{\mu}}_{t,v,p}^{e_1} = \hat{\mathbf{T}}_{e_1\leftarrow c_t^v}\,\Pi^{-1}\left(p, \hat{D}_t^v; \mathbf{K}^v\right), \tag{9}$$

where $\hat{\boldsymbol{\mu}}$, $\hat{\mathbf{s}}$, $\hat{\mathbf{q}}$, $\hat{\alpha}$, and $\hat{\mathbf{c}}$ represent the 3D Gaussian centroid, anisotropic scale, rotation quaternion, opacity, and spherical harmonic (SH) appearance coefficients, respectively. The depth-guided centroid grounds all Gaussian primitives in the canonical ego-first metric frame.

The Gaussian rendering objective is formulated as:

$$\mathcal{L}_{\mathrm{GS}} = \lambda_2\mathcal{L}_2 + \lambda_{\mathrm{SSIM}}\mathcal{L}_{\mathrm{SSIM}} + \lambda_{\mathrm{LPIPS}}\mathcal{L}_{\mathrm{LPIPS}}, \tag{10}$$

where $\mathcal{L}_2$, $\mathcal{L}_{\mathrm{SSIM}}$, and $\mathcal{L}_{\mathrm{LPIPS}}$ supervise pixel-level photometric fidelity, structural coherence, and perceptual quality, respectively. We empirically set $\lambda_2 = 1.0$, $\lambda_{\mathrm{SSIM}} = 0.2$, and $\lambda_{\mathrm{LPIPS}} = 0.05$ across all experiments.

Table 7: **Summary of datasets for dense geometry supervision.** Overview of public driving benchmarks utilized for metric depth supervision and geometric evaluation.

| Name | Syn | Scenes | Views | Name | Syn | Scenes | Views |
|---|---|---|---|---|---|---|---|
| VKITTI2 (Cabon et al., 2020) | ✓ | 50 | 2 | nuScenes (Caesar et al., 2020) | ✗ | 700 | 6 |
| AmodalSynthDrive (Sekkat et al., 2024) | ✓ | 105 | 4 | DDAD (Guizilini et al., 2020) | ✗ | 150 | 6 |
| RealDriveSim (Jadon et al., 2025) | ✓ | 6,343 | 1 | KITTI (Geiger et al., 2013) | ✗ | 156 | 2 |
| SimBEV (Mehr & Eskandarian, 2025) | ✓ | 220 | 6 | Waymo (Sun et al., 2020) | ✗ | 798 | 5 |
| CARD (Elazab et al., 2026) | ✗ | 23 | 2 | Openscenes (Contributors, 2023) | ✗ | 1,310 | 8 |
| Argoverse (Chang et al., 2019) | ✗ | 65 | 7 | | | | |

**Ego-Motion Supervision.** The ego-motion head predicts the trajectory $\hat{\mathbf{P}} = \{\hat{\mathbf{P}}_t\}_{t=1}^{T}$, where $\hat{\mathbf{P}}_t = \hat{\mathbf{T}}_{e_1 \leftarrow e_t} \in \mathrm{SE}(3)$ and $\hat{\mathbf{P}}_1 = \mathbf{I}$. Each pose is represented by a translation vector $\hat{\mathbf{t}}_t \in \mathbb{R}^3$ and a normalized rotation quaternion $\hat{\mathbf{q}}_t \in \mathbb{R}^4$. We supervise the trajectory using:

$$\mathcal{L}_{\text{ego}} = \frac{1}{T-1} \sum_{t=2}^{T} \left( \left\| \hat{\mathbf{t}}_t - \mathbf{t}_t \right\|_1 + \| \hat{\mathbf{q}}_t - \mathbf{q}_t \|_1 \right). \tag{11}$$

# B Data Collection and Experimental Details

## B.1 Dense Geometry Supervision

**Multi-Source Geometry Corpus.** Learning a metric-grounded 4D scene representation requires pixel-dense, scale-consistent, and spatially continuous depth supervision. However, acquiring ground-truth metric geometry across diverse real-world driving environments presents a fundamental trade-off: raw LiDAR point clouds deliver reliable absolute metric scale but suffer from severe spatial sparsity and non-uniform angular resolution, whereas off-the-shelf monocular depth models yield dense surface topology but inherit scale ambiguity and temporal instability.

To resolve this limitation, we construct a large-scale, multi-source dense geometry corpus that combines three complementary supervision avenues: (i) synthetic pixel-complete dense depth, (ii) metric depth completed from sparse LiDAR measurements in real-world driving benchmarks, and (iii) our high-density self-collected LiDAR dataset. The detailed statistics and distribution breakdown across all three sources are summarized in Table 7. For all benchmark datasets, training uses only the official training split, while evaluation is performed on strictly scene-disjoint test splits.

**Synthetic Dense Depth.** Synthetic simulation environments generate pixel-complete metric depth without sensor occlusion masks, beam sparsity, or atmospheric measurement noise. Unlike physical sensors, synthetic data offers dense and accurate geometric annotations across diverse driving conditions, including different illumination periods, weather variations, and challenging viewpoints. This provides essential supervision for structural edge cases where physical sensors routinely underperform, including distant highway horizons, ultra-thin structures (e.g., traffic light poles, cables, and fences), and specular or textureless road surfaces.

**Real-World Metric Depth Completion.** Real-world LiDAR provides metric depth measurements, but its sparse sampling leaves substantial gaps between scan lines and at distant or weakly observed regions. We use these measurements as geometric anchors and apply a depth completion pipeline to obtain denser supervision for real driving sequences. Together with the spatially complete depth available from synthetic rendering, the completed real-world depth targets provide complementary supervision for learning scene geometry across synthetic and real domains.

**LiDAR-Aligned Dense Depth.** In real-world driving environments, dense metric depth annotations are inherently unavailable due to the spatial sparsity of physical sensors. To synthesize high-quality, pixel-dense metric targets, we leverage dense relative depth predictions $\tilde{\mathbf{D}}$ from a foundation geometry model as structural topological priors and ground them against metric LiDAR observations $\mathbf{D}_l$. Specifically, raw multi-frame LiDAR point clouds are transformed into the camera coordinate system and subjected to occlusion-aware visibility filtering and ego-motion compensation prior to image plane projection. To bridge the scale-and-shift ambiguity inherent in monocular predictions,

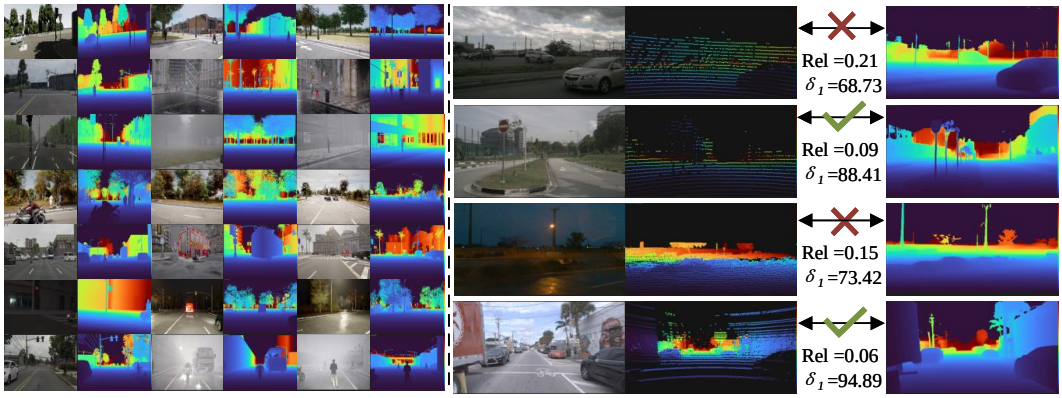


Figure 5: **Overview of data curation and quality filtering.** Qualitative examples illustrating real-world dataset filtering, highlighting the automated removal of noisy, corrupted, or low-fidelity samples to ensure high-quality training supervision.

we formulate a robust global alignment over the valid LiDAR support mask $\Omega_l$ using the ROE objective (Wang et al., 2025c):

$$(s^*, b^*) = \arg\min_{s,b} \sum_{\mathbf{p}\in\Omega_l} \rho\left(s\tilde{\mathbf{D}}(\mathbf{p}) + b - \mathbf{D}_l(\mathbf{p})\right), \tag{12}$$

where $\rho(\cdot)$ denotes a robust loss function resilient to gross sensor outliers. The metric-aligned dense depth map is subsequently obtained via $\mathbf{D}_{\text{align}}(\mathbf{p}) = s^*\tilde{\mathbf{D}}(\mathbf{p}) + b^*$.

**Dual-Consistency Quality Filtering.** While global alignment transfers absolute metric scale to dense monocular predictions, pseudo-ground-truth targets can still suffer from two distinct failure modes: (i) Model Prediction Failures. Monocular estimators struggle under extreme conditions, such as specular reflection, low-light nighttime scenes, or textureless surfaces. (ii) Sensor Projection Artifacts. LiDAR point clouds are subject to edge bleed, motion blur, and parallax occlusions, where laser returns incorrectly project onto background objects near depth discontinuities. To eliminate these corruptions, we institute a spatial consistency verification step. We evaluate local error residuals $|\mathbf{D}_{\text{align}}(\mathbf{p}) - \mathbf{D}_l(\mathbf{p})|$ across $\Omega_l$ alongside structural gradient alignment. Samples exhibiting severe residual disagreement or depth edge misalignment are discarded. As illustrated in Fig. 5, this dual-filtering strategy removes unreliable pseudo-labels originating from both monocular estimation failures and projection artifacts, preserving only high-confidence metric supervision for end-to-end model training.

**Real-World Dense LiDAR Collection.** To complement public benchmarks, we collect a 6-million-frame real-world dataset captured with a Pandar LiDAR and six surround-view cameras. Rather than directly projecting sparse, single-frame LiDAR scans, we aggregate multi-frame LiDAR observations via SLAM-based ego-motion alignment for static scene elements and motion compensation for dynamic objects. Following geometric consistency filtering and camera projection, the fused point clouds yield dense metric depth maps. This aggregation pipeline dramatically improves depth coverage over conventional sparse LiDAR projections (Fig. 6), delivering high-density metric geometry supervision for model training.

### B.2 Implementation Details

**Experimental Settings and Optimization.** Given $V$ surround-view cameras, we construct input sequences by sampling $T = \min(32, \lfloor 64/V \rfloor)$ consecutive temporal frames, capping the total batch sequence footprint at 64 images. Inputs are processed while preserving the native aspect ratio of each camera image through principal-point-centered cropping and resizing, with the corresponding camera intrinsic matrices $\mathbf{K}^v$ adjusted accordingly. All dataset-specific coordinate conventions are standardized to the Right-Down-Forward (RDF) camera frame, and time-varying ego poses are expressed relative to the canonical reference frame $e_1$ at timestamp $t = 1$. For real-world training

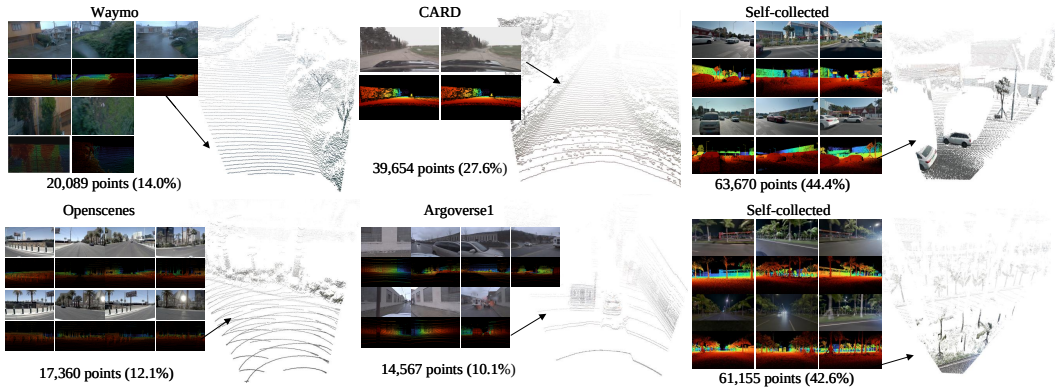


Figure 6: **Comparison of dense geometry supervision.** Qualitative visualization of colored point clouds across public driving datasets versus our self-collected dense LiDAR corpus. Valid point counts are annotated beneath each sample.

sequences lacking ground-truth optical flow annotations, we extract dense optical flow pseudo-labels using UFM (Zhang et al., 2025).

We initialize the visual feature encoder of DriveWeave with pretrained DINOv3 (Siméoni et al., 2025) weights and fully activate it throughout training. We optimize DriveWeave using AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a weight decay of $0.05$. The learning rate is linearly warmed up to a peak value of $1 \times 10^{-5}$ over the first $100{,}000$ iterations, followed by a cosine annealing decay schedule. Distributed training is accelerated using bfloat16 (BF16) mixed precision, TensorFloat-32 (TF32) tensor math, PyTorch Distributed Data Parallelism (DDP) via NCCL, and activation gradient checkpointing. Gradients are clipped independently for each module to a maximum $\ell_2$-norm threshold of $1.0$. The model is trained on $64\times$ NVIDIA H20 GPUs for approximately five days ($100$ epochs at $10{,}000$ iterations/epoch), while all inference runtime and rendering throughput metrics are benchmarked on a single NVIDIA H20 GPU. All models are evaluated under a unified pixel resolution of $512 \times 288$ to ensure fair comparison.

**Two-Stage Curriculum Schedule.** DriveWeave is trained over a 100-epoch schedule with a progressive data-and-task curriculum:

- **Data Mixture Curriculum:** Training initializes exclusively on high-fidelity synthetic data to establish clean, uncorrupted geometric representations. The synthetic mixture ratio is gradually annealed as multi-source real-world driving sequences are introduced.
- **Stage I: Dynamic Scene Geometric Warmup (Epochs 1–25):** Training first focuses on establishing stable geometric and motion representations through the shared Transformer backbone, metric depth, optical flow, and 6-DoF ego-motion heads, supervised by $\mathcal{L}_{\text{depth}}$, $\mathcal{L}_{\text{flow}}$, and $\mathcal{L}_{\text{ego}}$. Starting from Epoch 10, the 3D Gaussian branch is additionally activated and optimized with $\mathcal{L}_{\text{GS}}$, allowing the renderable scene representation to be progressively grounded on the learned metric geometry.
- **Stage II: Full 4D Joint Optimization (Epochs 26–100):** From Epoch 25 onward, the flow-guided depth-consistency loss $\mathcal{L}_{\text{FDC}}$ and rendering-consistency loss $\mathcal{L}_{\text{FRC}}$ are introduced on top of the existing task-specific objectives. The depth, optical flow, ego-motion, and 3D Gaussian branches continue to be supervised by $\mathcal{L}_{\text{depth}}$, $\mathcal{L}_{\text{flow}}$, $\mathcal{L}_{\text{ego}}$, and $\mathcal{L}_{\text{GS}}$, respectively, while the newly introduced consistency terms explicitly couple these predictions across frames. All representation heads and backbone weights are jointly end-to-end optimized to enforce cross-task alignment across geometry, motion, and rendering.

### B.3 Evaluation Metrics

**Baseline Comparisons.** For 3D geometry and motion evaluation, we compare against leading general-purpose and driving-specific foundation models, including CUT3R (Wang et al., 2025b),

DA3 (Lin et al., 2025), VGGT-Ω (Wang et al., 2026b), UniDepth-v2 (Piccinelli et al., 2025), FoundationGeo (Liu et al., 2026), Metric Anything (Ma et al., 2026), DVGT (Zuo et al., 2026), and SurroundNEXO (Yuan et al., 2026). For feed-forward 3D Gaussian reconstruction and novel-view rendering, we benchmark against DrivingForward (Tian et al., 2025), DA3 (Lin et al., 2025), Envision4D (Song et al., 2026), Uni-splat (Shi et al., 2026), and Recon-drive (Yu et al., 2026).

**Depth Estimation.** We benchmark metric depth accuracy against sparse ground-truth LiDAR points using absolute relative error ($\mathrm{AbsRel}$) and threshold accuracy ($\delta_1$). For unaligned scale-ambiguous baseline models, predictions are first aligned to absolute metric scale via ROE solver using reference LiDAR measurements. Let $\hat{d}_{\mathbf{p}}$ and $d_{\mathbf{p}}$ denote the predicted and ground-truth depth values at pixel location $\mathbf{p} \in \Omega$, respectively. The single-frame depth metrics are formulated as:

$$\mathrm{AbsRel}(\hat{\mathbf{D}}, \mathbf{D}) = \frac{1}{|\Omega|} \sum_{\mathbf{p}\in\Omega} \frac{|\hat{d}_{\mathbf{p}} - d_{\mathbf{p}}|}{d_{\mathbf{p}}}, \quad \delta_k = \frac{1}{|\Omega|} \sum_{\mathbf{p}\in\Omega} \mathbb{I}\left(\max\left(\frac{\hat{d}_{\mathbf{p}}}{d_{\mathbf{p}}}, \frac{d_{\mathbf{p}}}{\hat{d}_{\mathbf{p}}}\right) < 1.25^k\right), \tag{13}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function, and $\Omega$ represents the set of valid LiDAR projection mask coordinates.

To evaluate temporal geometric stability across video frames, we follow DAV (Yang et al., 2025a) and compute the Temporal Alignment Error ($\mathrm{TAE}$). For adjacent frame pairs, depth maps are back-projected to 3D points, warped into the target frame coordinate system using camera extrinsics and GT ego-motion, and re-projected back to the target image plane. We measure $\mathrm{AbsRel}$ over overlapping valid pixel regions in both forward ($t \to t+1$) and backward ($t+1 \to t$) directions, averaging over all $T-1$ sequence transitions:

$$\mathrm{TAE} = \frac{1}{2(T-1)} \sum_{t=1}^{T-1} \left[\mathrm{AbsRel}\left(f_{t\to t+1}(\hat{\mathbf{D}}_t), \hat{\mathbf{D}}_{t+1}\right) + \mathrm{AbsRel}\left(f_{t+1\to t}(\hat{\mathbf{D}}_{t+1}), \hat{\mathbf{D}}_t\right)\right], \tag{14}$$

where $f_{t\to s}$ represents the depth reprojection operator from frame $t$ to frame $s$. Lower $\mathrm{AbsRel}$ and $\mathrm{TAE}$, along with higher $\delta_1$, signify superior geometric accuracy and temporal stability.

**Point Cloud Reconstruction.** We evaluate global 3D metric geometric fidelity using point cloud Accuracy ($\mathrm{Acc}$) and Completeness ($\mathrm{Comp}$) (Wang et al., 2025a). Let $\hat{\mathcal{P}}$ and $\mathcal{P}$ denote the predicted and ground-truth reference point clouds, respectively, expressed within a canonical metric reference frame. These Chamfer-distance-based components are defined as:

$$\mathrm{Acc} = \frac{1}{|\hat{\mathcal{P}}|} \sum_{\hat{\mathbf{p}}\in\hat{\mathcal{P}}} \min_{\mathbf{q}\in\mathcal{P}} \|\hat{\mathbf{p}} - \mathbf{q}\|_2, \quad \mathrm{Comp} = \frac{1}{|\mathcal{P}|} \sum_{\mathbf{q}\in\mathcal{P}} \min_{\hat{\mathbf{p}}\in\hat{\mathcal{P}}} \|\mathbf{q} - \hat{\mathbf{p}}\|_2. \tag{15}$$

Here, $\mathrm{Acc}$ measures the average Euclidean distance from predicted points to their nearest ground-truth neighbors (evaluating noise and artifacts), whereas $\mathrm{Comp}$ quantifies how thoroughly the prediction covers the ground-truth surface (evaluating structural missingness). Lower values indicate superior reconstruction quality for both metrics.

**Ego-Motion Estimation.** We quantify 6-DoF vehicle trajectory accuracy using normalized Area Under the Curve ($\mathrm{AUC}$) metrics at 3° and 30° error thresholds ($\mathrm{AUC}_3$ and $\mathrm{AUC}_{30}$)(Zuo et al., 2026). For baseline methods (e.g., CUT3R (Wang et al., 2025b), DA3 (Lin et al., 2025), VGGT-Ω (Wang et al., 2026b)), camera trajectories estimated from front-view image sequences are converted to vehicle ego-frame poses via calibrated extrinsic transformations prior to comparison. The normalized $\mathrm{AUC}$ metric over a maximum error threshold $\tau_{\max}$ is defined as:

$$\mathrm{AUC}_{\tau_{\max}} = \frac{1}{\tau_{\max}} \int_0^{\tau_{\max}} P(e \le x)\, \mathrm{d}x, \tag{16}$$

where $e$ represents the maximum of the relative rotation error and translation-direction angular error across frame pairs, and $P(e \le x)$ denotes the cumulative distribution function of frame pairs with joint orientation/translation error below $x$. Higher $\mathrm{AUC}_3$ and $\mathrm{AUC}_{30}$ percentages reflect more precise motion trajectory recovery.

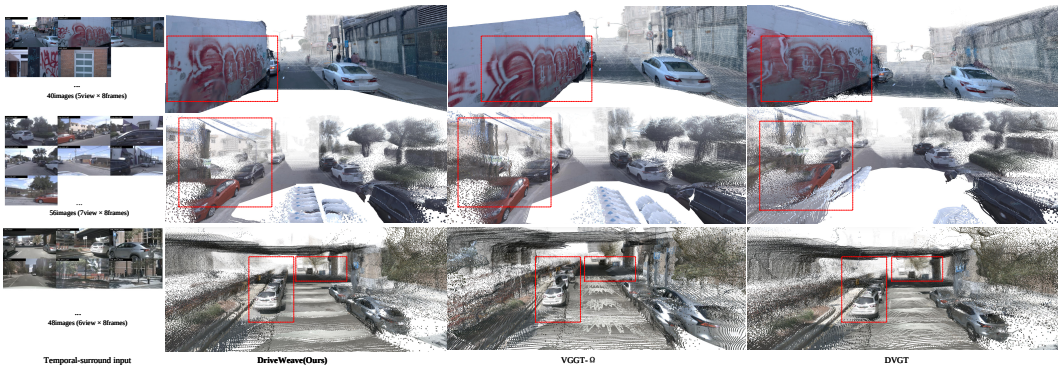


Figure 7: **Temporal point cloud reconstruction from in-the-wild sequences.** Qualitative comparison of accumulated temporal point clouds across static, dynamic, and ego-motion scenarios. DriveWeave is evaluated against VGGT-Ω and DVGT, exhibiting superior geometric fidelity and spatiotemporal alignment.

Table 8: **Ablation study of geometric task supervision combinations.** Impact of combining metric depth, point cloud, 6-DoF ego-motion, and optical flow losses on 3D geometric reconstruction and 3D Gaussian scene rendering quality.

| Method | Nuscenes | | | | | Waymo | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Rel ↓ | TAE↓ | Acc ↓ | AUC3 ↑ | PSNR ↑ | Rel ↓ | TAE↓ | Acc ↓ | AUC3 ↑ | PSNR ↑ |
| Depth + Ego + GS | 12.0 | 19.0 | 0.68 | 69.86 | 28.9 | 12.3 | 8.1 | 0.73 | 71.53 | 29.1 |
| Point + Ego + GS | 13.8 | 22.4 | 0.79 | 61.25 | 27.4 | 14.3 | 9.8 | 0.81 | 62.89 | 27.9 |
| Depth + Point + Ego + GS | 12.5 | 19.6 | 0.69 | 69.14 | 28.7 | 12.4 | 8.3 | 0.74 | 71.22 | 29.4 |
| Depth + Ego + Flow + GS | **11.2** | **17.8** | **0.60** | **72.42** | **29.8** | 11.1 | **6.9** | **0.66** | **74.31** | **30.1** |
| Depth + Point + Ego + Flow + GS | 11.5 | 18.1 | 0.64 | 72.16 | 29.5 | **10.9** | 7.3 | **0.66** | 73.42 | 29.6 |

**Rendering Quality.** We evaluate feed-forward 3D Gaussian Splatting rendering performance by generating novel-view rendering images from the learned 4D scene representation. Rendered camera views are evaluated against corresponding ground-truth RGB images using Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) (Wang et al., 2004), and Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018) using VGG features. These metrics comprehensively capture pixel-level accuracy (PSNR), structural fidelity (SSIM), and deep perceptual realism (LPIPS).

# C Additional Experiments

## C.1 Metric Geometry

**Temporal Point Cloud.** Beyond frame-level reconstruction accuracy, Fig. 7 further examines temporal geometric consistency by accumulating multi-frame surround-view predictions into a unified 3D scene. Across different temporal-surround configurations, DriveWeave preserves substantially cleaner and more coherent scene structures than VGGT-Ω (Wang et al., 2026b) and DVGT (Zuo et al., 2026). Competing methods exhibit visible cross-frame misalignment, duplicated structures, and geometric drift as observations are accumulated, particularly around scene boundaries, vehicles, and distant structures. In contrast, DriveWeave maintains stable spatial alignment over time and produces more compact surfaces with fewer temporal artifacts. This qualitative comparison demonstrates that the improvement is not limited to individual-frame geometry, but extends to temporally consistent reconstruction of the complete driving scene.

**Geometric Supervision.** Table 8 studies how different geometric supervision signals affect the unified representation. Direct point supervision alone performs notably worse than depth supervision, suggesting that dense pixel-aligned metric depth provides a more suitable geometric anchor for subsequent pose estimation and Gaussian construction. Moreover, jointly introducing point supervision does not consistently improve the depth-based model, indicating potential optimization

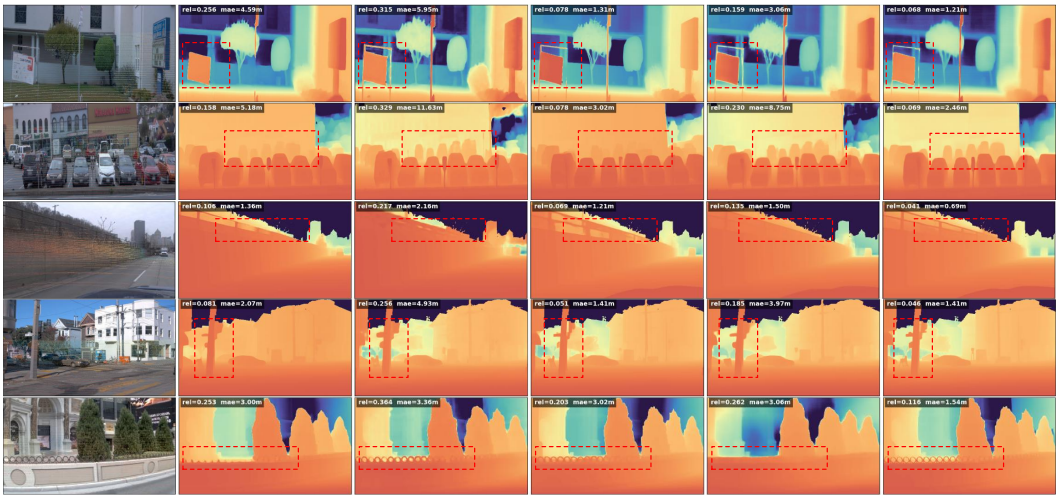


Figure 8: **Qualitative comparison of metric depth estimation in driving scenes.** Column 1 presents RGB images with ground-truth LiDAR overlays. Red dashed boxes highlight fine structural details and thin depth boundaries across predictions. Quantitative metrics (AbsRel and MAE in meters), evaluated over valid LiDAR pixels, are annotated in the top-left corner of each depth map.

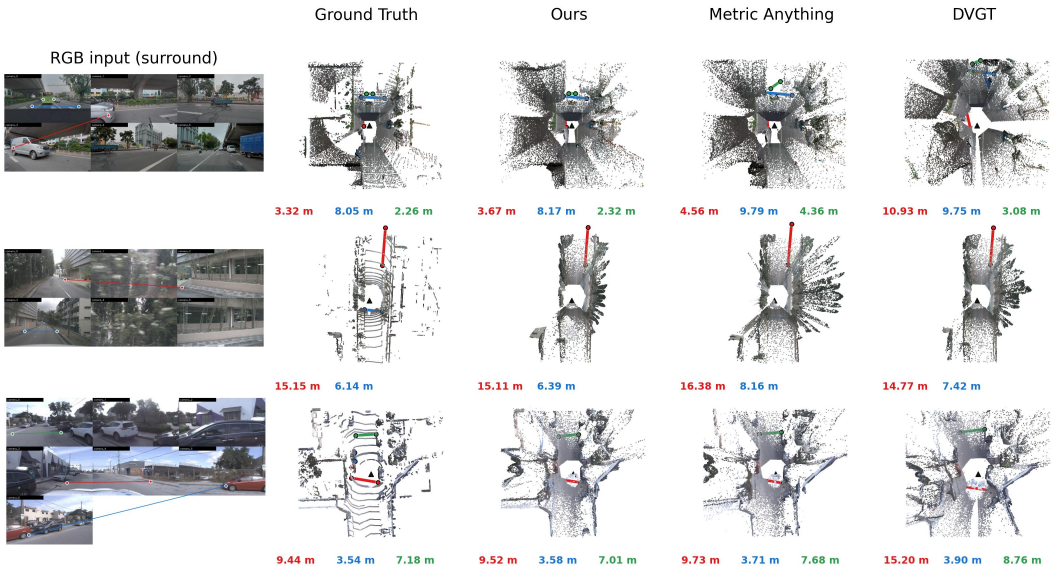


Figure 9: **Qualitative comparison of metric point cloud reconstruction.** Surround-view inputs and reconstructed 3D point clouds are evaluated against ground truth. Colored segments highlight corresponding keypoint pairs, with 3D Euclidean distances (in meters) annotated to measure metric scale fidelity.

interference between direct 3D point regression and dense depth learning, meanwhile, the additional point prediction branch introduces extra model parameters and computational overhead. In contrast, incorporating optical flow consistently strengthens temporal geometry, leading to better depth stability, ego-motion estimation, point-cloud reconstruction, and Gaussian rendering. Overall, these results favor a depth-centered geometric representation, with flow serving as a complementary temporal cue that strengthens geometry and subsequently benefits Gaussian reconstruction.




Figure 10: **Qualitative comparison of feed-forward Gaussian scene reconstruction.** Novel-view rendering outputs from DA3 (Lin et al., 2025), UniSplat (Shi et al., 2026), ReconDrive (Yu et al., 2026), and DriveWeave are evaluated against ground-truth observations on driving scenes. PSNR values are annotated in the top-left corner of each rendered view.

### C.2 Qualitative Analysis

**Depth Quality and Structural Details.** Figure 8 presents qualitative comparisons of metric depth estimation under diverse driving scenes. Compared with existing methods, DriveWeave produces cleaner object boundaries and more structurally coherent depth transitions, particularly around vehicles, roadside poles, building contours, and distant scene regions highlighted by the dashed boxes. Several competing methods exhibit over-smoothed boundaries, local depth discontinuities, or inconsistent scale in challenging regions, whereas DriveWeave better preserves both fine-grained local geometry and the global depth layout. These results indicate that the learned representation captures not only accurate metric scale, but also sharper scene structure across heterogeneous urban environments.

**Metric Scale and Cross-View Geometry.** Figure 9 further examines whether the reconstructed geometry preserves physically meaningful metric relationships after aggregating surround-view observations into a unified 3D space. The colored segments indicate corresponding spatial intervals in the input images and reconstructed point clouds. DriveWeave consistently recovers inter-object

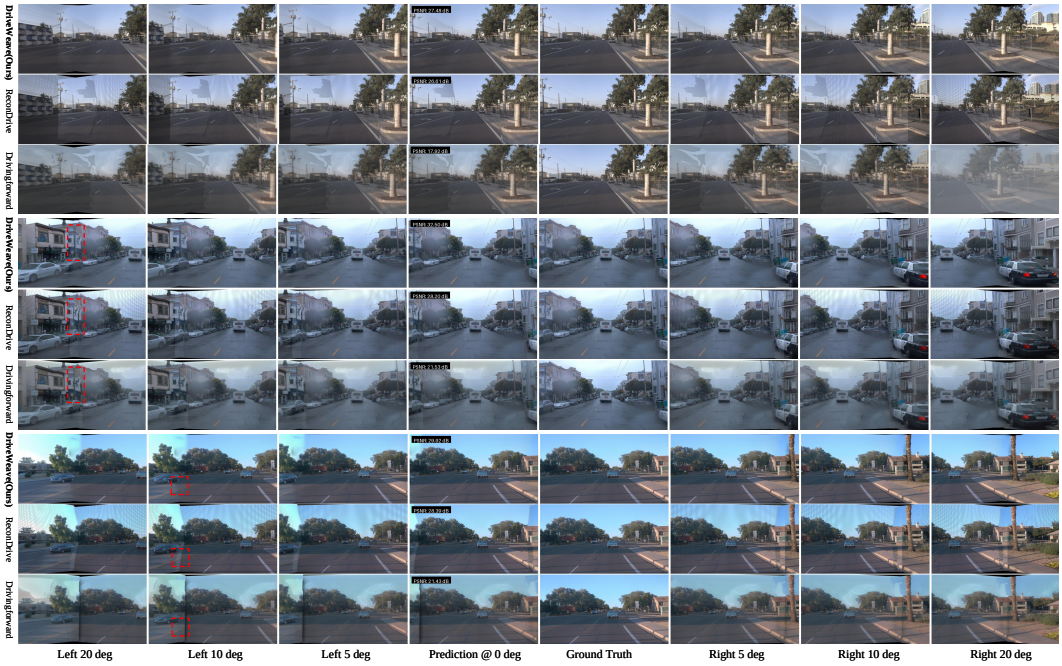


Figure 11: **Qualitative comparison of novel-view rendering under camera yaw perturbations.** Novel-view rendering outputs from DriveWeave, ReconDrive (Yu et al., 2026), and DrivingForward (Tian et al., 2025) are evaluated across viewpoint rotations ranging from $-20°$ to $+20°$. Reference-view predictions and ground-truth observations are displayed at $0°$.

distances that closely match the ground-truth measurements, while maintaining a coherent global scene layout across cameras. In contrast, Metric Anything and DVGT exhibit noticeable scale distortion and cross-view misalignment, particularly for long-range structures and laterally separated objects. These results demonstrate that DriveWeave does not merely recover visually plausible point clouds, but establishes a metrically grounded and globally consistent dynamic scene representation from surround-view observations.

### C.3 Gaussian Rendering

**Cross-View Appearance and Detail Preservation.** Figure 10 qualitatively compares feed-forward Gaussian scene reconstruction across diverse driving scenes. DriveWeave maintains more consistent appearance across different viewpoints, preserving stable colors and structural details even for small or partially observed objects such as distant vehicles, pedestrians, traffic signs, and roadside structures highlighted by the dashed boxes. In contrast, competing methods often exhibit noticeable blurring, color shifts, or ghosting around these fine-grained regions, indicating insufficient cross-view alignment of the reconstructed Gaussians. DriveWeave also better preserves global illumination and scene appearance while retaining sharp local structures, demonstrating that the underlying geometry provides a reliable anchor for cross-view Gaussian aggregation. The consistently high PSNR across the evaluated views further corroborates the visual improvements and confirms the fidelity of the reconstructed scene appearance.

**Novel-View Robustness under Viewpoint Perturbation.** Figure 11 further evaluates the robustness of Gaussian rendering under camera yaw perturbations from $-20°$ to $+20°$. Across all scenes, DriveWeave maintains more stable appearance and structure as the viewpoint changes, producing smoother cross-view transitions and more faithful rendering of buildings, vehicles, trees, and thin roadside structures. In contrast, competing methods are more prone to view-dependent distortion, local blurring, ghosting, or geometric stretching, especially in regions highlighted by the dashed boxes. The strong rendering consistency of DriveWeave under continuous viewpoint shifts also indicates that its underlying 3D geometry is more accurate and globally coherent, since high-quality

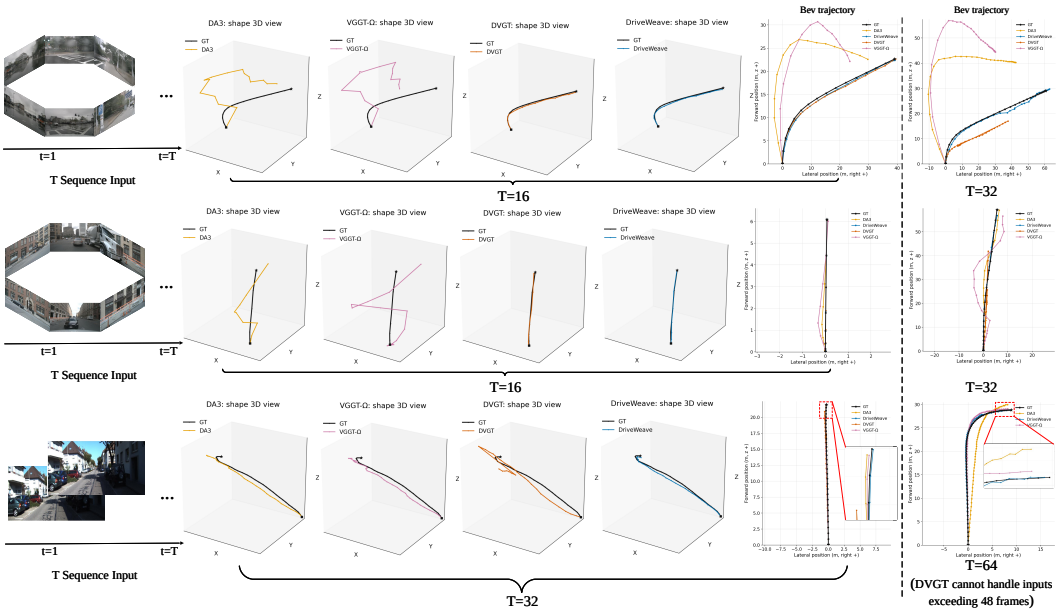


Figure 12: **Qualitative comparison of ego-motion estimation across sequence lengths.** Reconstructed 6-DoF ego-trajectories from evaluated methods are compared against ground truth across varying temporal horizons, highlighting spatiotemporal alignment and drift accumulation over extended sequences.

novel-view synthesis fundamentally relies on correct geometric anchoring of Gaussian primitives. These results therefore demonstrate not only superior cross-view rendering capability, but also the effectiveness of DriveWeave in learning a geometrically reliable scene representation.

### C.4 Ego-Motion Scalability & Drift Analysis

To evaluate temporal scalability and drift accumulation, we analyze the model's single-pass ego-motion estimation across varying sequence lengths in three scenarios: turning, straight driving, and long-horizon trajectories (Fig. 12). While DVGT yields reasonable pose estimates on short sequences, notably, increasing the input sequence length substantially degrades its geometric quality, indicating that additional temporal context does not consistently improve its predictions and can instead interfere with geometric representations established from earlier frames. Moreover, its temporal encoding limits processing to at most 48 frames. In contrast, DriveWeave maintains tight alignment with ground-truth trajectories across all sequence lengths without progressive deviation. This confirms that 4D geometric synchronization effectively preserves long-range temporal dependencies and suppresses error accumulation in ego-motion estimation.