STRUCTURED OBSERVATION MODELING FOR FEED-FORWARD 4D DRIVING RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Feed-forward 4D driving reconstruction aims to recover coherent dynamic scenes from sparse multi-camera temporal observations, while existing methods still struggle with coherent reconstruction from such sparse observations. Through a data-centric analysis, we find that farther target views receive sparser and more distributed geometric support across cameras and time, while driving datasets exhibit heterogeneous reconstruction conditions, motivating structured observation modeling and heterogeneous training, respectively. In this paper, we propose SOM4D, a feed-forward framework designed to better model sparse multi-view temporal observations for 4D driving scene reconstruction. Specifically, we introduce Pose-Time Relational Adaptation (PTRA) to explicitly model camera–time relations within a frozen visual-geometry backbone. We further introduce Hierarchical Observation Consolidation (HOC) to integrate multi-level features and shared scene context while preserving dense spatial details. To improve explicit scene reconstruction, we develop a Corrective Gaussian Head (CGH) that decodes dynamic Gaussians and progressively corrects them through geometry-aware scene correction and recurrent source-render feedback. Additionally, we train SOM4D on mixed-data to broaden the training conditions and assess the benefit of heterogeneous training on Waymo. Extensive experiments under both Waymo-only and mixed-data settings validate the proposed model design and the benefit of heterogeneous training.

1 INTRODUCTION

Reconstructing dynamic 4D scenes from sparse visual observations is a fundamental challenge in computer vision, with broad implications for realistic simulation, perception, planning, and generative data synthesis (Yang et al., 2025; Lu et al., 2025). Feed-forward 4D reconstruction takes sparse multi-camera images, camera calibration, and timestamps as input, and predicts a 4D scene that can be rendered from arbitrary viewpoints and times. This problem becomes particularly challenging in driving scenes, where ego motion continuously changes viewpoints and visibility, while independently moving objects alter the scene geometry over time. Meanwhile, large-scale spatial coverage, sparse temporal sampling, and limited camera overlap make the available observations inherently incomplete. As a result, different views and timestamps capture only partial and changing portions of the scene, leaving many target regions weakly constrained by the sources. Therefore, the central challenge is to recover a coherent 4D scene from sparse and incomplete visual observations.

Recently, several feed-forward reconstruction methods have emerged for autonomous driving. In 3D reconstruction, VGD and VGGD recover static driving scenes from sparse surround-view images (Lin et al., 2025b; 2026). Some recent 4D methods extend reconstruction to dynamic scenes, but adopt different input settings, such as unposed images or additional object-level conditions (Chen et al., 2025; Tan et al., 2026; Chi et al., 2026). STORM and SLARM are most closely aligned with this task, reconstructing dynamic driving scenes from calibrated multi-view temporal observations (Yang et al., 2025; Qiu et al., 2026). Both largely follow a direct observation-to-Gaussian pipeline, where camera and time cues are mixed in token space before Gaussian decoding. This leaves pairwise camera–time relations, cross-level evidence consolidation, and source-based scene correction largely implicit. These aspects become more important when useful evidence is distributed across cameras and timestamps. This motivates a more structured observation-modeling framework for 4D reconstruction.

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.

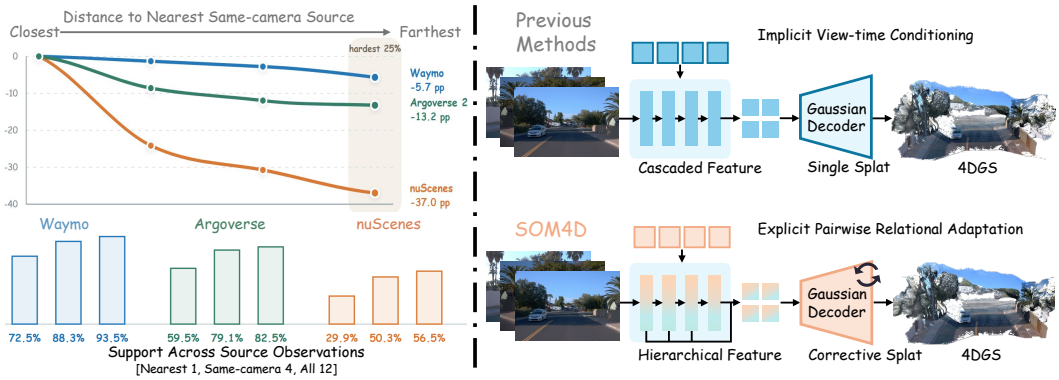


Figure 1: **Observation support and structured reconstruction.** Left: reference-verifiable geometric support decreases with nearest-source displacement, while other source times and cameras provide complementary support. Right: schematic contrast between a cascaded decoder and SOM4D's relational encoding, hierarchical consolidation, and source-based correction.

To better understand reconstruction difficulty, we analyze how sparse observations support target views. Waymo serves as our principal benchmark because its calibrated multi-camera temporal sequences closely match the task setting (Sun et al., 2020). As shown in Figure 1, we group target novel views by their distance to the nearest same-camera source and measure reference-verifiable geometric support using calibrated depth-consistent reprojection. We divide targets into four quartiles, Q_1–Q_4, from nearest to farthest source distance. Support decreases from Q_1 to Q_4, with drops of 5.67, 36.95, and 13.23 percentage points on Waymo, nuScenes, and Argoverse 2, respectively (Caesar et al., 2020; Wilson et al., 2023). A nested source-set analysis further shows that the available support is distributed across different times and cameras. Figure 1 and the further analysis in Appendix A show that the three datasets also exhibit heterogeneous reconstruction conditions. Together, these observations reveal reference-verifiable geometric support within scenes and heterogeneous reconstruction conditions across datasets.

These observations therefore motivate a data–model perspective. On the data side, we hypothesize that heterogeneous auxiliary datasets may expose the model to additional reconstruction conditions beyond Waymo alone. On the model side, these observations motivate a more structured 4D representation that can organize scene information throughout reconstruction while preserving local detail and source consistency. Data diversity broadens the reconstruction regimes seen during training, while model design determines how observations are organized and exploited. This perspective motivates both heterogeneous training and the structured reconstruction framework shown in Figure 1.

To this end, we propose SOM4D, a feed-forward 4D reconstruction framework for better modeling sparse multi-view temporal observations. We first introduce **P**ose-**T**ime **R**elational **A**daptation (**PTRA**) to explicitly model camera–time relations within a frozen visual-geometry backbone. To better aggregate complementary information while preserving dense spatial details, we introduce **H**ierarchical **O**bservation **C**onsolidation **(HOC)**, which maintains a dense feature hierarchy and compact shared context across cameras and time. We then develop a **C**orrective **G**aussian **H**ead **(CGH)** that decodes dynamic Gaussians and progressively corrects geometry, appearance, and motion using source observations. Moreover, we train SOM4D jointly on Waymo, nuScenes, and Argoverse 2 to expose the model to a broader range of reconstruction conditions. Our main contributions are:

- We conduct a data-centric analysis of feed-forward 4D driving reconstruction, revealing reference-verifiable geometric support within scenes and heterogeneous reconstruction conditions across datasets. Based on these observations, we propose SOM4D for more effective sparse 4D driving scene reconstruction.
- We propose Pose-Time Relational Adaptation (PTRA) to strengthen the backbone representation of sparse multi-view temporal observations. It explicitly models camera–time relations while preserving the pretrained visual-geometry capability of the frozen backbone.

- We develop a hierarchical scene modeling and decoding mechanism to better reconstruct 4D driving scenes. Specifically, we introduce HOC for multi-level dense and shared-context aggregation, together with CGH for geometry-aware scene correction and recurrent source-render feedback.
- Extensive experiments under both Waymo-only and mixed-data training validate the effectiveness of SOM4D, while controlled comparisons further quantify the impact of heterogeneous multi-dataset training on feed-forward 4D driving reconstruction.

2 Related Work

2.1 Feed-forward reconstruction in general scenes

Feed-forward reconstruction has rapidly advanced with large-scale visual-geometry models. DUSt3R and MASt3R predict point maps and correspondences, while VGGT and VGGT-Omega extend this paradigm toward generic multi-view geometry priors (Wang et al., 2024; Leroy et al., 2024; Wang et al., 2025a; 2026). Recent methods further improve reconstruction by aggregating information across larger image sets and maintaining scene-level geometric context. In parallel, renderable predictors such as pixelSplat, MVSplat, and GS-LRM directly infer Gaussian scene representations for novel-view synthesis (Charatan et al., 2024; Chen et al., 2024; Zhang et al., 2024). Beyond static scenes, MonST3R and CUT3R extend feed-forward geometry estimation to temporal observations, while dynamic Gaussian representations and generalizable 4D Gaussian models introduce explicit time-varying scene primitives (Zhang et al., 2025; Wang et al., 2025b; Luiten et al., 2023; Lin et al., 2025a). These advances establish strong geometric, multi-view, and temporal priors. Our work builds on these developments but focuses on how sparse calibrated observations should be represented and integrated throughout feed-forward 4D reconstruction in large-scale driving scenes.

2.2 Feed-forward reconstruction in driving scenes

Driving scenes introduce additional challenges from ego motion, dynamic objects, large spatial extent, and sparse camera overlap. Existing 3D methods mainly address static geometry or novel-view synthesis, including surround-depth estimation and feed-forward Gaussian reconstruction with VGD and VGGD (Wei et al., 2022; Lin et al., 2025b; 2026). Recent 4D methods consider dynamic driving scenes under substantially different task formulations. Some focus on scene flow or occupancy (Hur & Roth, 2021), while others use unposed images, object-level conditions, recurrent reconstruction, or future-scene prediction (Chen et al., 2025; Chi et al., 2026; Tan et al., 2026; Yu et al., 2026; Huang et al., 2026; Song et al., 2026). DrivingRecon, STORM, and SLARM instead reconstruct explicit dynamic driving scenes from temporal observations (Lu et al., 2025; Yang et al., 2025; Qiu et al., 2026). Among them, STORM and SLARM are most closely aligned with our setting and demonstrate effective spatio-temporal interaction and explicit motion modeling. Our work focuses on jointly preserving observation-aligned detail and shared cross-observation context from visual encoding to explicit 4D scene construction. This motivates our view of feed-forward 4D driving reconstruction as a structured observation modeling problem.

3 Method

3.1 Preliminaries

Problem formulation. Given sparse calibrated source observations $\mathcal{O} = \{(I_i, K_i, T_i, \tau_i)\}_{i=1}^{N_s}$, the feed-forward 4D reconstruction predicts a single dynamic scene $\mathcal{X}$ and renders it at a target query $q = (K_q, T_q, \tau_q)$:

$$\mathcal{X} = f_\theta(\mathcal{O}), \qquad \hat{I}_q = \mathcal{R}(\mathcal{X}, q), \tag{1}$$

where I_i, K_i, T_i, and τ_i denote the source image, camera intrinsics, camera-to-world pose, and timestamp. The reconstructed scene $\mathcal{X}$ is shared by all target queries, which never condition f_θ.

Observation diagnosis. For a target q, we denote its distance to the nearest same-camera source by $b(q; \mathcal{O})$ and its reference-verifiable geometric support by $s(q; \mathcal{O})$. Empirically, a larger $b(q; \mathcal{O})$ is associated with a lower $s(q; \mathcal{O})$, while for high-baseline targets the available support is distributed

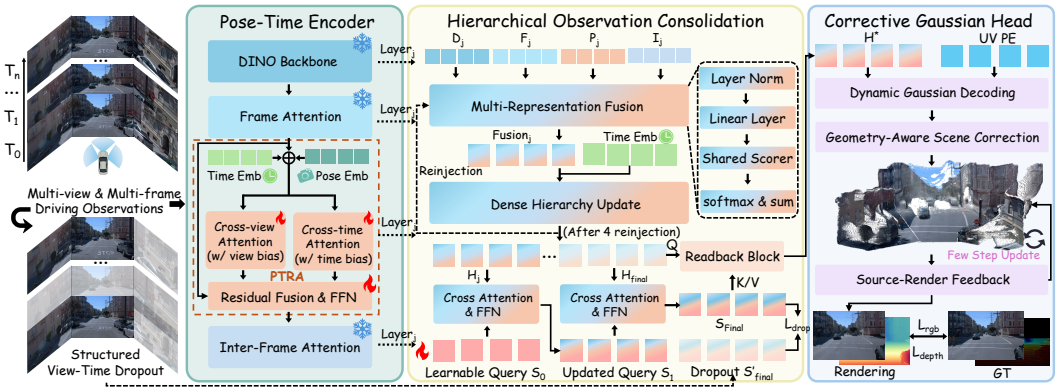


Figure 2: **Structured observation modeling in SOM4D.** PTRA introduces explicit camera–time relational modeling into the frozen visual-geometry backbone and produces relation-aware multi-level representations. HOC hierarchically fuses DINO, frame, PTRA, and inter-frame features while jointly refining dense features and shared context, which is read back before scene decoding. CGH then decodes dynamic Gaussians, performs geometry-aware scene correction through local metric correction and cross-view source consolidation, and applies recurrent source-render feedback correction.

across cameras and time. Across datasets, we denote the training distributions of Waymo, nuScenes, and Argoverse 2 by p_W, p_N, and p_A, which cover different measured reconstruction conditions. The corresponding definitions and data analysis are detailed in Appendix A.2.

Design guideline. On the model side, sparse multi-view temporal observations should be structured by explicitly modeling camera–time relations, consolidating complementary evidence without losing dense detail, and maintaining consistency with the source observations. On the data side, let p_W, p_N, and p_A denote the training distributions of Waymo, nuScenes, and Argoverse 2; we use the balanced mixture $p_{\text{Mix3}} = \frac{1}{3}(p_W + p_N + p_A)$ with equal iteration over the three datasets, while keeping the inference formulation and optimization budget unchanged. Detailed model and training configurations are provided in Appendices B and C, respectively.

3.2 Overall

Given the source observations $\mathcal{O}$, SOM4D reconstructs a dynamic 4D scene through three stages, as shown in Figure 2. First, the frozen visual-geometry backbone extracts multi-level features, where **P**ose-**T**ime **R**elational **A**daptation (**PTRA**) explicitly models camera–time relations. At each selected layer, we obtain DINO, frame, PTRA, and inter-frame representations, denoted as D_j, F_j, P_j, and I_j. **H**ierarchical **O**bservation **C**onsolidation **(HOC)** then hierarchically integrates these representations across multiple scales. It progressively refines a dense feature hierarchy H_j while updating a compact shared context S_j through cross-attention. The final context is read back into the dense representation to obtain H_{out}, allowing scene-level information from different cameras and times to complement local spatial details. Finally, **C**orrective **G**aussian **H**ead **(CGH)** decodes H_{out} into a dynamic Gaussian scene. Geometry-aware scene correction first combines local metric correction with cross-view source consolidation, after which source-render feedback correction progressively updates geometry, appearance, and motion. The resulting scene is constructed entirely from the source observations and can be rendered at arbitrary target cameras and times.

3.3 Pose-Time Relational Encoder

VGGT-Omega provides strong pretrained visual-geometry features, while the sparse driving observations additionally carry explicit camera and temporal relations. We introduce **P**ose-**T**ime **R**elational **A**daptation (**PTRA**) at four selected backbone stages to incorporate these relations without modifying the frozen backbone. Let $X_\ell \in \mathbb{R}^{B\times T\times V\times P\times C}$, denote the image patch tokens at stage ℓ, where T, V, and P are the numbers of source times, cameras, and spatial tokens per image, respectively,

and $C = 1024$ is the channel dimension. We augment the tokens with learnable camera and time embeddings as $\widetilde{X}_\ell = X_\ell + e_{\text{cam}} + e_{\text{time}}$.

PTRA factorizes relational modeling along the camera and temporal dimensions. The view branch connects cameras at the same time, while the temporal branch connects different times from the same camera. Relative camera geometry and temporal offsets are encoded by attention biases B_v and B_t. The PTRA update is:

$$\Delta_v = \mathrm{Attn}_v\Big(\mathrm{LN}(\widetilde{X}_\ell); B_v\Big), \qquad \Delta_t = \mathrm{Attn}_t\Big(\mathrm{LN}(\widetilde{X}_\ell); B_t\Big),$$
$$Y_\ell = X_\ell + \gamma_v\Delta_v + \gamma_t\Delta_t, \qquad X'_\ell = Y_\ell + \gamma_f\,\mathrm{FFN}(\mathrm{LN}(Y_\ell)). \tag{2}$$

Each branch attends over all spatial tokens within its observation group, allowing relational information to propagate beyond corresponding patch locations. The residual scales are initialized to 10^{-3} so that PTRA starts as a small adaptation of the pretrained representation. At each selected stage, we collect four representations for HOC, $\mathcal{F}_\ell = \{F^D_\ell, F^F_\ell, F^P_\ell, F^I_\ell\}$, corresponding to DINO, frame, PTRA, and inter-frame features, respectively.

3.4 Hierarchical Observation Consolidation

PTRA structures camera–time relations, but useful scene evidence remains distributed across sparse observations. To address this, we introduce Hierarchical Observation Consolidation (HOC), which preserves observation-aligned details in a dense hierarchy H_ℓ while progressively aggregating cross-observation information into a compact shared context S_ℓ.

Multi-representation fusion. At each stage $\ell \in \{4, 11, 17, 23\}$, HOC receives four backbone representations $\mathcal{F}_\ell = \{F^D_\ell, F^F_\ell, F^P_\ell, F^I_\ell\}$. After normalization and projection to a common feature space, they are adaptively fused as:

$$Z_\ell = \sum_{r\in\{D,F,P,I\}} \alpha^r_\ell \bar{F}^r_\ell, \qquad \alpha_\ell = \mathrm{softmax}_r\big(g(\bar{F}^r_\ell)\big), \tag{3}$$

where $\bar{F}^r_\ell = W^r_\ell\,\mathrm{LN}(F^r_\ell)$. The fused features are decoded from coarse to fine to form $H_{23} \to H_{17} \to H_{11} \to H_4$, preserving dense spatial information throughout the hierarchy.

Context readback. Dense features remain observation-aligned, while complementary information is distributed across the source set. We therefore initialize $M = 512$ learnable context tokens and progressively update them at each level. Let E_ℓ denote the flattened dense features from all source cameras and times. The context update is:

$$\widehat{S}_\ell = S_{\text{prev}} + \mathrm{CA}(\mathrm{LN}(S_{\text{prev}}), \mathrm{LN}(E_\ell), \mathrm{LN}(E_\ell)),$$
$$S_\ell = \widehat{S}_\ell + \mathrm{FFN}\Big(\mathrm{LN}(\widehat{S}_\ell)\Big). \tag{4}$$

We use the complete dense features without spatial pooling, allowing S_ℓ to aggregate shared scene information while H_ℓ retains local structure.

Context readback. Since scene prediction remains dense, the accumulated context is returned to the final dense representation:

$$A = \mathrm{softmax}\left(\frac{\mathrm{LN}(H_4)\,\mathrm{LN}(S_4)^\top}{\sqrt{256}}\right), \qquad H_{\text{out}} = H_4 + \lambda_{\text{read}} W_{\text{read}}(AS_4). \tag{5}$$

Thus, H_{out} preserves local spatial details while incorporating information aggregated across cameras and time, and is passed to the reconstruction head.

3.5 Corrective Gaussian Head

Dynamic Gaussian decoding. Two convolutional tails upsample H_{out} to the source-image resolution. For each source pixel p in image i, the decoder predicts depth, scale, rotation, opacity, appearance, and cubic motion parameters (v, a, j). The predicted depth z_{ip} is lifted to the world-space center $\mu^0_{ip} = t_i + R_i K_i^{-1}\widetilde{p}\, z_{ip}$. Its position at time τ is then modeled by:

$$\mu_{ip}(\tau) = \mu^0_{ip} + v_{ip}\delta + \frac{1}{2}a_{ip}\delta^2 + \frac{1}{6}j_{ip}\delta^3, \tag{6}$$

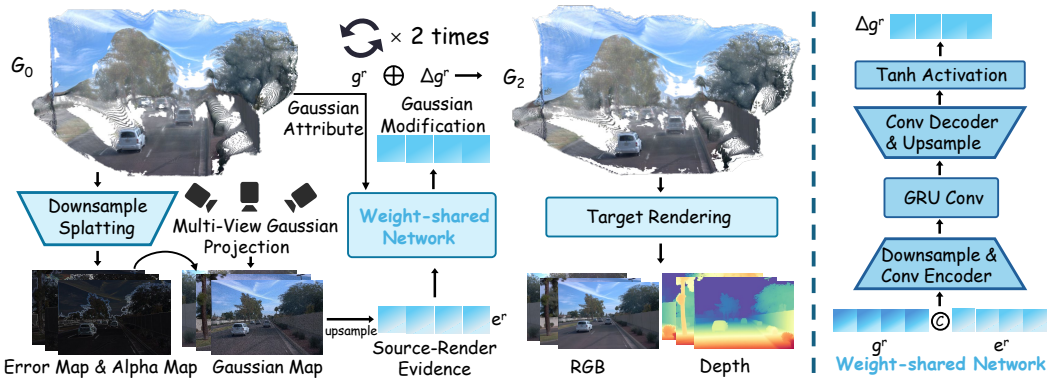


Figure 3: **Corrective Gaussian Head (CGH).** The dense representation H_{out} is first decoded into source-anchored dynamic Gaussians. Geometry-aware scene correction then combines local metric correction with cross-view source consolidation, followed by two shared-weight rounds of source-render feedback correction before target rendering.

where $\delta = 2(\tau - \tau_i)$ denotes the relative time offset. Appearance combines the source RGB with a learned residual and degree-1 spherical harmonics. The decoder therefore produces a dense source-anchored dynamic Gaussian scene $\mathcal{G}_{\text{dec}}$.

Geometry-aware scene correction. The decoded scene is further corrected using explicit scene geometry. *Local metric correction* combines calibrated observations, backbone features, and local neighborhoods constructed from $\mathcal{G}_{\text{dec}}$ to update depth, motion, dynamic score, and confidence. *Cross-view source consolidation* then uses peer-view reprojection consistency to reweight Gaussian opacity and scale.

Source-render feedback. Geometry-aware correction does not directly account for the rendered appearance of the reconstructed scene. We therefore render the current Gaussians back to the source cameras and use the resulting feedback to update the scene. At refinement round r:

$$(h^{r+1}, \Delta g^r) = U_\phi(h^r, e^r, g^r), \qquad g^{r+1} = \text{Compose}(g^r, \Delta g^r), \tag{7}$$

where g^r denotes the current Gaussian parameters and e^r denotes source-render feedback. The update jointly corrects Gaussian position, scale, rotation, opacity, appearance, and motion. We apply two recurrent rounds with shared weights $\mathcal{G}^0 \rightarrow \mathcal{G}^1 \rightarrow \mathcal{G}^2$, and use final scene $\mathcal{G}^2$ for target rendering.

3.6 Learning Strategy

Training supervision. We train SOM4D with RGB reconstruction loss $\mathcal{L}_{\text{RGB}}$, perceptual loss $\mathcal{L}_{\text{LPIPS}}$ (Zhang et al., 2018), and camera-depth supervision $\mathcal{L}_{\text{depth}}$. We additionally use a subset-consistency loss $\mathcal{L}_{\text{subset}}$ to regularize the shared HOC context under reduced source observations, together with auxiliary supervision $\mathcal{L}_{\text{aux}}$ for source details, dynamic regions, appearance, and intermediate predictions. Target images and reference annotations are used only for supervision and never condition scene construction. Detailed definitions and weights are provided in Appendix C.

Optimization. We optimize the trainable components with AdamW while keeping the visual-geometry backbone frozen with 32 batch size in 4 NVIDIA 5090. We use the same architecture and supervision for Waymo-only and Mix3 training, varying only the training-data distribution.

4 Experiments

4.1 Experimental Setup

Tasks. We study feed-forward 4D driving reconstruction from sparse calibrated multi-view temporal observations. Following the sparse-context setting, each scene window contains 20 temporal frames from three frontal cameras. The model receives four source times at t, t+0.5s, t+1.0s, and t+1.5s,

Table 1: **Comparison to state-of-the-art methods on the Waymo Open Dataset.** Published results are shown together with our controlled Waymo-only and Mix3 reruns on the 2019-clip validation protocol. Dynamic-only and Full image report held-out target performance; Full Rec reports source-view reconstruction. *: reproduced or adapted results under the corresponding protocol. †: Non-sky RGB metrics in our reruns; D-RMSE follows the valid-depth-GT protocol.

Methods	Dynamic NVS			Full NVS†			Full Rec†		
	PSNR↑	SSIM↑	D-RMSE↓	PSNR↑	SSIM↑	D-RMSE↓	PSNR↑	SSIM↑	D-RMSE↓
Per-scene optimization & Generalizable feed-forward methods									
EmerNeRF	17.79	0.255	40.88	24.51	0.738	33.99	-	-	-
3DGS	17.13	0.267	13.88	25.13	0.741	19.68	-	-	-
PVG	15.51	0.128	15.91	22.38	0.661	13.01	-	-	-
DeformableGS	17.10	0.266	12.14	25.29	0.761	14.79	-	-	-
LGM	17.36	0.216	11.09	18.53	0.447	9.07	-	-	-
LGM*	19.58	0.443	9.43	23.59	0.691	8.02	-	-	-
GS-LRM*	20.02	0.520	9.95	25.18	0.753	7.94	-	-	-
Feed-forward 4D driving reconstruction methods									
STORM	22.10	0.624	7.50	26.38	0.794	5.48	-	-	-
STORM*	21.59	0.596	7.44	25.91	0.781	5.30	26.99	0.831	5.21
SLARM*	22.27	0.639	6.72	26.67	0.796	4.95	28.14	0.860	4.83
SOM4D	22.97	0.681	6.01	27.16	0.825	4.11	29.24	0.895	4.01
STORM* (Mix3)	21.77	0.603	7.00	26.12	0.786	4.76	27.26	0.835	4.68
SLARM* (Mix3)	22.55	0.653	6.30	26.97	0.806	4.49	28.46	0.866	4.39
SOM4D (Mix3)	**23.03**	**0.684**	**5.94**	**27.32**	**0.830**	**3.99**	**29.47**	**0.899**	**3.89**

and reconstructs a single dynamic scene that is rendered at the remaining held-out camera–time queries, covering both temporal interpolation and post-context extrapolation. Waymo is our principal benchmark, while nuScenes and Argoverse 2 are additionally used to study heterogeneous multi-dataset training and cross-domain reconstruction (Sun et al., 2020; Caesar et al., 2020; Wilson et al., 2023). We consider *Waymo-only* (W) training and *Mix3* training, which jointly samples all three datasets under a fixed total training budget. Training and sampling details are provided in Appendix C.

Baselines. We compare against representative optimization-based and feed-forward reconstruction methods. For broader reference, we include published results from per-scene approaches such as EmerNeRF, 3DGS, PVG, and Deformable GS (Yang et al., 2023; Kerbl et al., 2023; Chen et al., 2023; Yang et al., 2024), as well as feed-forward reconstruction models including LGM and GS-LRM (Tang et al., 2024; Zhang et al., 2024). Our primary comparisons are with STORM and SLARM, which are most closely matched to our calibrated multi-view temporal 4D reconstruction setting (Yang et al., 2025; Qiu et al., 2026).

Metrics. We report PSNR and SSIM for appearance reconstruction and camera-Z depth RMSE (D-RMSE) for geometry. The primary image metrics are evaluated over non-sky regions; on Waymo, we additionally report the same metrics on strict dynamic regions using the independent evaluation mask. Interpolation and post-context extrapolation are reported separately when applicable.

4.2 Comparison Results

Main results. Table 1 compares SOM4D with representative optimization-based and feed-forward methods on Waymo. Under Waymo-only training, SOM4D outperforms the reproduced STORM and SLARM baselines across all NVS metrics, reaching 22.97/0.681/6.01 on dynamic regions and 27.16/0.825/4.11 on full images in PSNR/SSIM/D-RMSE. With Mix3, SOM4D further improves to 23.03/0.684/5.94 and 27.32/0.830/3.99, respectively, achieving the best NVS performance in the table. The improvement is particularly pronounced for reconstruction, where SOM4D reaches 29.24 PSNR under Waymo-only training and 29.47 with Mix3, exceeding the corresponding SLARM baselines by 1.10 and 1.01 dB while also substantially improving SSIM and depth accuracy.

Controlled training. The paired Waymo-only and Mix3 results separate architectural gains from the effect of heterogeneous training. With the same Waymo-only data, SOM4D improves over

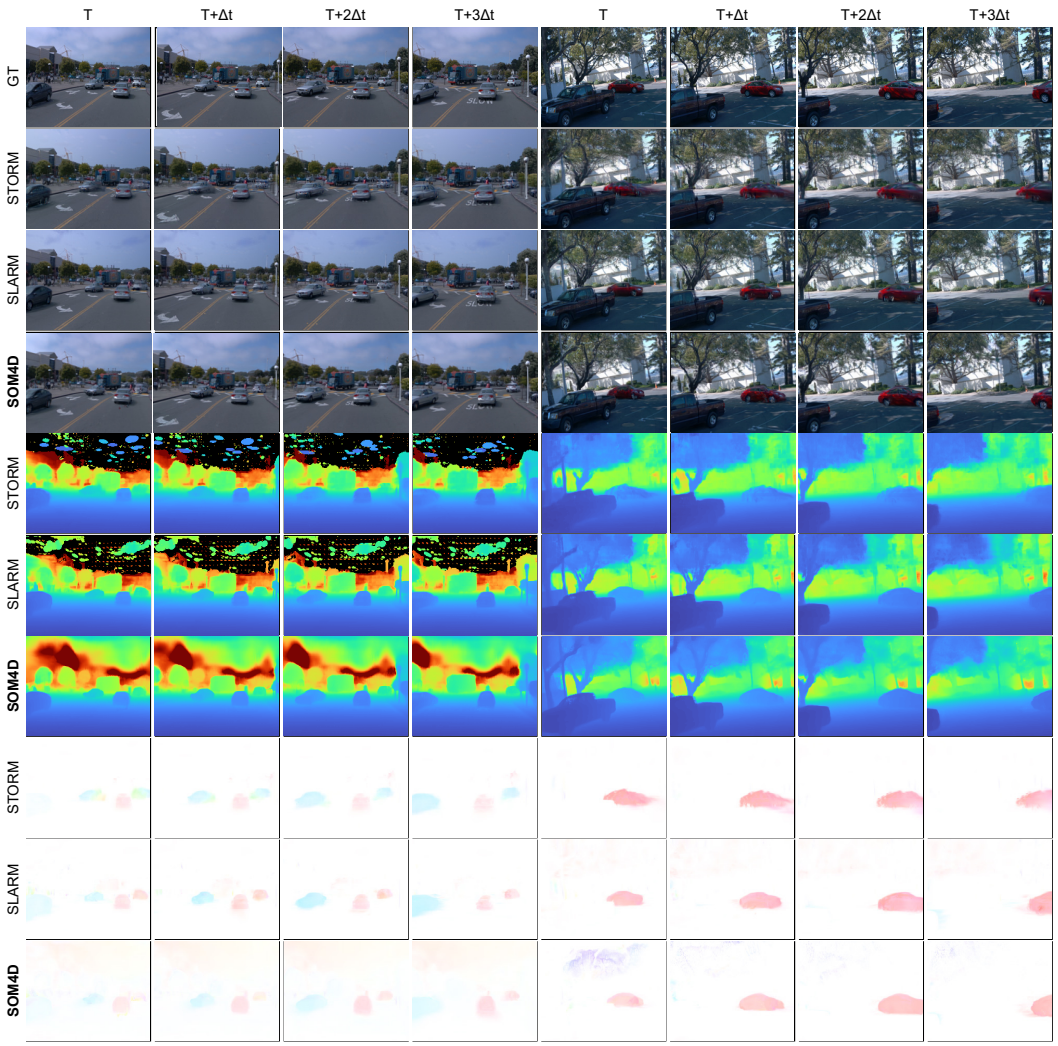


Figure 4: **Qualitative comparison on Waymo.** We compare STORM, SLARM, and SOM4D on the same held-out target view across multiple timestamps. The top block shows RGB reconstruction against the ground truth, while the middle and bottom blocks show the corresponding predicted depth and optical flow. SOM4D produces more consistent appearance and geometry over time, particularly around dynamic objects and distant scene structures.

reproduced SLARM from 26.67 to 27.16 PSNR on full-image NVS and from 28.14 to 29.24 on reconstruction, while reducing D-RMSE from 4.95 to 4.11 and from 4.83 to 4.01, respectively. Similar gains are observed over reproduced STORM. Mix3 further improves all three architectures, and SOM4D remains the strongest model across NVS and reconstruction metrics. This consistent W-to-Mix3 improvement supports heterogeneous training as a complementary source of gain beyond the architectural improvements of SOM4D.

Visualizations. Figure 4 compares STORM, SLARM, and SOM4D on identical held-out target views across multiple timestamps. SOM4D produces cleaner dynamic-object boundaries, more stable appearance, and more consistent scene geometry, in agreement with the quantitative improvements in both RGB quality and depth accuracy. The strong reconstruction results further indicate that these gains are not limited to novel-view synthesis, but also translate to more accurate recovery of the observed scene. These qualitative improvements remain consistent across challenging viewpoints and timestamps. Additional qualitative examples are provided in the Apppendix D.3.

Table 2: Reconstruction across target baseline quartiles on Waymo. Each entry is non-sky PSNR↑/D-RMSE↓; Q_1–Q_4 denote increasing nearest-source displacement.

Method	Q_1	Q_2	Q_3	Q_4
Waymo-only				
STORM	27.72/6.28	25.89/4.92	25.26/4.73	24.76/5.19
SLARM	**29.57**/5.61	26.36/4.63	25.60/4.53	25.13/4.97
SOM4D	28.78/**4.50**	**27.44/3.78**	**26.59/3.82**	**25.81/4.21**
Mix3				
STORM	27.76/5.80	26.25/4.32	25.53/4.20	24.95/4.66
SLARM	**29.74**/5.12	26.83/4.17	25.94/4.11	25.35/4.51
SOM4D	29.13/**4.35**	**27.60/3.68**	**26.70/3.72**	**25.85/4.10**

Table 3: Core architecture ablations on Waymo. All variants use the same 50k Waymo-only training schedule and supervision. Entries are PSNR↑/D-RMSE↓.

Variant	Dynamic	Full†
w/o PTRA	21.54/6.47	26.19/4.30
w/o cam&time emb.	22.37/6.03	26.81/4.01
w/o HOC	22.29/6.02	26.69/3.99
w/o shared context S	22.41/6.00	26.85/3.98
w/o CGH	22.04/6.06	26.39/4.02
w/o render feedback	22.24/6.00	26.46/3.99
Full SOM4D	**22.51/6.00**	**26.86/3.97**

4.3 Reconstruction under Reduced Support

Our data analysis shows that larger target-to-source displacement is associated with lower reference-verifiable geometric support. We therefore evaluate all controlled Waymo checkpoints using the fixed Q_1–Q_4 baseline quartiles defined in Appendix A.2. Table 2 reports non-sky PSNR and D-RMSE. PSNR consistently decreases from Q_1 to Q_4 for all methods, showing that the data-defined baseline axis is reflected in appearance reconstruction quality. While SLARM achieves the highest PSNR in Q_1, SOM4D performs best from Q_2 through Q_4 and yields the lowest D-RMSE in every quartile under both training settings. In Q_4, SOM4D improves over SLARM by 0.68 dB / 0.76 D-RMSE with Waymo-only training and by 0.50 dB / 0.41 D-RMSE with Mix3.

These results show that SOM4D maintains stronger geometric reconstruction across all baseline regimes, while its appearance advantage becomes evident beyond the closest-source regime and persists toward Q_4. Mix3 further improves both PSNR and D-RMSE for all three architectures in every Waymo quartile, indicating broader benefits across target regimes rather than a gain specific to reduced-support targets. Appendix D.1 extends this analysis to nuScenes and Argoverse 2, where SOM4D likewise achieves the strongest performance from Q_2 through Q_4 under Mix3 training.

4.4 Ablation Studies

We ablate the main components of SOM4D using the same 50k Waymo-only training protocol. Removing PTRA also removes its feature stream F^P; the w/o shared context S variant retains the dense HOC hierarchy but removes shared-context aggregation and readback. For the reconstruction head, we separately remove the full corrective head and only the source-render correction.

Table 3 reports PSNR and D-RMSE on dynamic-only and full-image regions. Removing PTRA causes the largest drop, reducing PSNR by 0.97 and 0.67 dB, respectively, while also increasing depth error. Removing only the camera/time embeddings has a much smaller effect, indicating that the gain mainly comes from relational adaptation. HOC also improves reconstruction, with the dense multi-representation hierarchy contributing most of the gain and shared context S providing a smaller benefit. CGH gives a consistent 0.47 dB improvement, while source-render feedback correction mainly improves appearance with little effect on D-RMSE.

5 Conclusion

We study feed-forward 4D driving reconstruction through the structure of sparse multi-view temporal observations. Our analysis reveals reference-verifiable geometric support and complementary evidence across cameras and time, motivating SOM4D with relation-aware encoding, hierarchical consolidation, and source-based correction. Controlled Waymo-only and Mix3 comparisons further separate the benefits of model design and heterogeneous training. The current motion model is limited to short temporal horizons, and structured correction stages introduce additional computation. Future work could improve long-term motion modeling and efficiency.

AI USE STATEMENT

We used generative AI tools to assist with manuscript writing and language polishing, including improving grammar, clarity, and readability, and to support literature searches by suggesting potentially relevant papers. We did not use generative AI for generating new experimental results or changing the model implementation. The authors reviewed all AI-assisted text for accuracy and consistency with the intended scientific meaning and verified AI-suggested references through a manual literature survey. We take full responsibility for the final content of this paper, including all AI-assisted text.

ETHICS STATEMENT

All authors have read the ICLR Code of Ethics, and this work adheres to its principles. This study concerns feed-forward 4D driving-scene reconstruction using the existing Waymo, nuScenes, and Argoverse 2 datasets; no human-subject experiments were conducted. Reconstructed driving scenes may retain sensitive visual information present in the source data, and dataset coverage may limit generalization across environments and road users. Downstream use should therefore account for privacy, potential biases, and reconstruction errors. The reported reconstruction benchmarks do not establish the safety or reliability of the method for deployment in safety-critical driving systems.

REPRODUCIBILITY STATEMENT

Section 3 and Appendix B describe the model, architectural configurations, and source-render update algorithm. Section 4.1 specifies the evaluation setting, and Appendix C summarizes supervision, optimization, sampling, and the pathway-ablation protocol. Appendix A defines the independent data diagnosis and baseline groups, while Appendix D provides additional quantitative, qualitative, and ablation results.

REFERENCES

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. URL https://arxiv.org/abs/1903.11027.

David Charatan, Sizhe Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelSplat: 3D gaussian splats from image pairs for scalable generalizable 3D reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2312.12337.

Xiaoxue Chen, Ziyi Xiong, Yuantao Chen, Gen Li, Nan Wang, Hongcheng Luo, Long Chen, Haiyang Sun, Bing Wang, Guang Chen, Hangjun Ye, Hongyang Li, Ya-Qin Zhang, and Hao Zhao. DGGT: Feedforward 4D reconstruction of dynamic driving scenes using unposed images. *arXiv preprint arXiv:2512.03004*, 2025. URL https://arxiv.org/abs/2512.03004.

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*, 2024. URL https://arxiv.org/abs/2403.14627.

Yurui Chen, Chun Gu, Junzhe Jiang, Xiatian Zhu, and Li Zhang. Periodic vibration gaussian: Dynamic urban scene reconstruction and real-time rendering. *arXiv preprint arXiv:2311.18561*, 2023. URL https://arxiv.org/abs/2311.18561.

Cheng Chi, Xianqi Wang, Hongcheng Luo, Mingfei Tu, Gangwei Xu, Zehan Zhang, Bing Wang, Guang Chen, Hangjun Ye, Sida Peng, Xin Yang, and Haiyang Sun. PointForward: Feedforward driving reconstruction through point-aligned representations. *arXiv preprint arXiv:2605.11594*, 2026. URL https://arxiv.org/abs/2605.11594.

Jiahui Huang, Jiawei Ren, Michal Tyszkiewicz, Bjoern Haefner, Michael Shelley, Xin Kang, Seung Wook Kim, Ning Xu, Qi Wu, Janick Martinez Esturo, Shengyu Huang, Nick Schneider,

Laura Leal-Taixé, Zan Gojcic, and Sanja Fidler. Instant NuRec: Feed-forward 3D gaussian reconstruction for driving scene simulation. *arXiv preprint arXiv:2607.14203*, 2026. URL https://arxiv.org/abs/2607.14203.

Junhwa Hur and Stefan Roth. Self-supervised multi-frame monocular scene flow. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. URL https://arxiv.org/abs/2105.02216.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):139:1–139:14, 2023. doi: 10.1145/3592433.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3D with MASt3R. In *European Conference on Computer Vision*, 2024. URL https://arxiv.org/abs/2406.09756.

Chieh Hubert Lin, Zhaoyang Lv, Songyin Wu, Zhen Xu, Thu Nguyen-Phuoc, Hung-Yu Tseng, Julian Straub, Numair Khan, Lei Xiao, Ming-Hsuan Yang, Yuheng Ren, Richard Newcombe, Zhao Dong, and Zhengqin Li. DGS-LRM: Real-time deformable 3D gaussian reconstruction from monocular videos. *arXiv preprint arXiv:2506.09997*, 2025a. URL https://arxiv.org/abs/2506.09997.

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*, 2025b. URL https://arxiv.org/abs/2510.19578.

Junhong Lin, Jinlong Wang, Xianda Guo, Yanlun Peng, Wei Zheng, Guoqing Liu, Hanli Wang, Tiesong Zhao, and Wei Gao. Visual geometry foundation-aware gaussians for single-frame surround-view driving reconstruction. *arXiv preprint arXiv:2608.10682*, 2026. URL https://arxiv.org/abs/2608.10682.

Hao Lu, Tianshuo Xu, Wenzhao Zheng, Yunpeng Zhang, Wei Zhan, Dalong Du, Masayoshi Tomizuka, Kurt Keutzer, and Yingcong Chen. DrivingRecon: Large 4D gaussian reconstruction model for autonomous driving. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/f5717c76feff4f751604c0678c46627b-Abstract-Conference.html.

Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3D gaussians: Tracking by persistent dynamic view synthesis. *arXiv preprint arXiv:2308.09713*, 2023. URL https://arxiv.org/abs/2308.09713.

Zhicheng Qiu, Jiarui Meng, Tong an Luo, Yican Huang, Xuan Feng, Xuanfu Li, and Zhan Xu. SLARM: Streaming and language-aligned reconstruction model for dynamic scenes. *arXiv preprint arXiv:2603.22893*, 2026. URL https://arxiv.org/abs/2603.22893.

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. URL https://arxiv.org/abs/2606.10656.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. URL https://arxiv.org/abs/1912.04838.

Kaiyuan Tan, Yingying Shen, Mingfei Tu, Haohui Zhu, Bing Wang, Guang Chen, Hangjun Ye, and Haiyang Sun. UFO: Unifying feed-forward and optimization-based methods for large driving scene modeling. *arXiv preprint arXiv:2602.20943*, 2026. URL https://arxiv.org/abs/2602.20943.

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. *arXiv preprint arXiv:2402.05054*, 2024. URL https://arxiv.org/abs/2402.05054.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025a. URL https://arxiv.org/abs/2503.11651.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Omega. *arXiv preprint arXiv:2605.15195*, 2026. URL https://arxiv.org/abs/2605.15195.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3D perception model with persistent state. *arXiv preprint arXiv:2501.12387*, 2025b. URL https://arxiv.org/abs/2501.12387.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D vision made easy. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2312.14132.

Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Yongming Rao, Guan Huang, Jiwen Lu, and Jie Zhou. SurroundDepth: Entangling surrounding views for self-supervised multi-camera depth estimation. In *Conference on Robot Learning*, 2022. URL https://arxiv.org/abs/2204.03636.

Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv preprint arXiv:2301.00493*, 2023. URL https://arxiv.org/abs/2301.00493.

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. *arXiv preprint arXiv:2311.02077*, 2023. URL https://arxiv.org/abs/2311.02077.

Jiawei Yang, Jiahui Huang, Yuxiao Chen, Yan Wang, Boyi Li, Yurong You, Apoorva Sharma, Maximilian Igl, Peter Karkus, Danfei Xu, Boris Ivanovic, Yue Wang, and Marco Pavone. STORM: Spatio-temporal reconstruction model for large-scale outdoor scenes. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2501.00602.

Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20331–20341, 2024. doi: 10.1109/CVPR52733.2024.01922.

Zhongrui Yu, Zhao Wang, Yijia Xie, Yida Wang, Xueyang Zhang, Yifei Zhan, and Kun Zhan. StreetForward: Perceiving dynamic street with feedforward causal attention. *arXiv preprint arXiv:2603.19552*, 2026. URL https://arxiv.org/abs/2603.19552.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. MonST3R: A simple approach for estimating geometry in the presence of motion. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.03825.

Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. GS-LRM: Large reconstruction model for 3D gaussian splatting. *arXiv preprint arXiv:2404.19702*, 2024. URL https://arxiv.org/abs/2404.19702.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. URL https://arxiv.org/abs/1801.03924.

APPENDIX CONTENTS

Appendix A Data Analysis . . . 13
A.1 Reconstruction Conditions across Datasets . . . 13
A.2 Target-Dependent Geometric Support . . . 13
A.3 Support across Cameras and Time . . . 15
A.4 Robustness and Scope . . . 16
Appendix B Model Design . . . 16
B.1 PTRA and Hierarchical Observation Consolidation . . . 17
B.2 Corrective Gaussian Head . . . 17
Appendix C Training Details . . . 18
C.1 Training Objective and Supervision . . . 18
C.2 Optimization and Training Protocol . . . 19
Appendix D Experimental Results . . . 19
D.1 Cross-Dataset Reconstruction Results . . . 19
D.2 Quantitative Results . . . 23
D.3 Qualitative Results . . . 23
D.4 Ablation Studies . . . 23

A DATA ANALYSIS

This appendix defines the observation diagnosis and the baseline groups used in Section 4.3. We first characterize dataset-level reconstruction conditions, then measure target support and its distribution across source observations. We use images, calibration, reference depth, and masks for this analysis.

A.1 RECONSTRUCTION CONDITIONS ACROSS DATASETS

Analysis protocol. We analyze 4,246 validation clips from Waymo, nuScenes, and Argoverse 2, comprising 1,817, 1,387, and 1,042 clips, respectively. Each 20-frame window uses three cameras and source offsets $\{0, 5, 10, 15\}$, yielding 12 source images and 48 held-out targets. The resulting 203,808 targets include 87,216 Waymo targets. Clip descriptors summarize acquisition motion and content. target images define baseline quartiles, and eligible target pixels determine support percentages. Bootstrap resampling retains whole scenes.

Validation-domain conditions. We express trajectories relative to their initial ego pose and measure endpoint translation and unwrapped yaw change. Depth statistics use finite, positive camera-Z references. Fine-mask area divided by non-sky image area measures dynamic-content prevalence rather than verified object motion.

Figure 5 summarizes the validation-domain reconstruction conditions. Waymo has a heavier translation tail, nuScenes and Argoverse 2 have larger validation yaw tails, and Argoverse 2 extends to longer reference depths. Dynamic-content prevalence also differs across domains. These measurements describe different reconstruction conditions rather than an ordering of dataset difficulty.

Training-domain conditions. To check whether this diversity is also present in the data used for learning, we separately scan 18,612 deterministic training windows: 7,181 from Waymo, 6,510 from nuScenes, and 4,921 from Argoverse 2. Motion is measured over each 20-frame window, while clip depth D90 pools three cameras at offsets $\{0, 5, 10, 15\}$, giving 12 fixed depth images per window.

Figure 6 shows distinct motion and reference-depth distributions in the three training domains. These deterministic scans characterize available conditions, not the realized Mix3 sampling frequencies or the performance effect of mixing datasets. We therefore use them to motivate the training-distribution comparison, while target-level support is examined separately below.

A.2 TARGET-DEPENDENT GEOMETRIC SUPPORT

Source–target displacement. For target q, we define the nearest same-camera positional baseline as:

$$b(q; \mathcal{O}) = \min_{i:\, v_i = v_q} \|c_q - c_i\|_2, \tag{8}$$

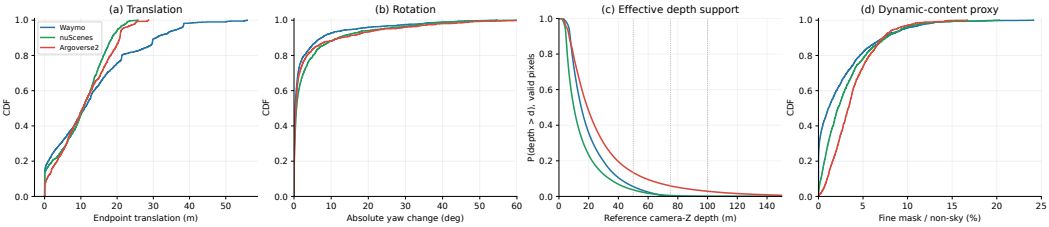


Figure 5: **Reconstruction conditions across validation domains.** Translation, yaw, and dynamic-content ratios use clip-level CDFs, while the depth panel pools valid reference pixels. The three datasets cover overlapping but distinct motion, reference-depth, and dynamic-content conditions. The depth range is a dataset descriptor and is distinct from the geometric support $s(q; \mathcal{O})$ defined below.

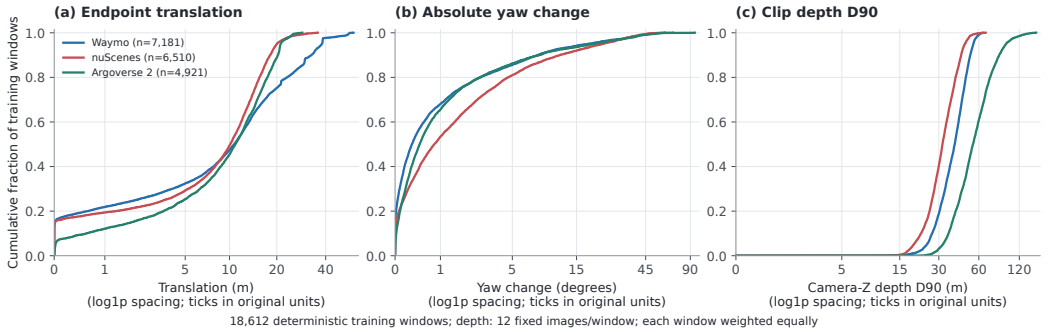


Figure 6: **Reconstruction conditions in deterministic training windows.** Empirical CDFs of endpoint translation, absolute yaw change, and clip depth D90 across 18,612 windows. Each window is weighted equally. The scan characterizes the conditions available in the three training domains; it does not represent the realized sampling frequencies of Mix3.

Table 4: **Nearest-source baseline quantiles in meters.** P25, P50, and P75 delimit target-image quartiles within each validation manifest. Q4 therefore spans different physical distances across datasets.

Dataset	P25	P50	P75	P90	Maximum
Waymo	0.192	0.834	1.708	3.101	11.825
nuScenes	0.267	0.770	1.401	2.160	5.911
Argoverse 2	0.348	0.822	1.383	2.178	6.122

where c_i is the translation of camera-to-world pose T_i in a common metric frame and v_i identifies the camera. The minimum is taken over the four same-camera source times, not over all cameras or the closest timestamp. We partition target images by the within-dataset P25, P50, and P75 boundaries in Table 4, from Q1 (smallest baseline) to Q4 (largest). These geometry-defined groups are also used in Section 4.3.

Reference-verifiable support. Let $\mathcal{P}_q$ be the nonempty set of static, non-sky target samples with valid reference depth. Each sample is unprojected to 3D and reprojected into the source images. A source verifies the target sample when the projected point is valid and its local reference depth satisfies:

$$|z_{\text{projected}} - z_{\text{reference}}| \leq \max(0.25\,\text{m}, \eta z_{\text{projected}}), \qquad \eta = 0.02. \tag{9}$$

Let $\chi_i(p)$ indicate this agreement in source i. We define:

$$s(q; \mathcal{O}) = \frac{1}{|\mathcal{P}_q|} \sum_{p \in \mathcal{P}_q} \mathbf{1}[\exists\, o_i \in \mathcal{O} : \chi_i(p) = 1]. \tag{10}$$

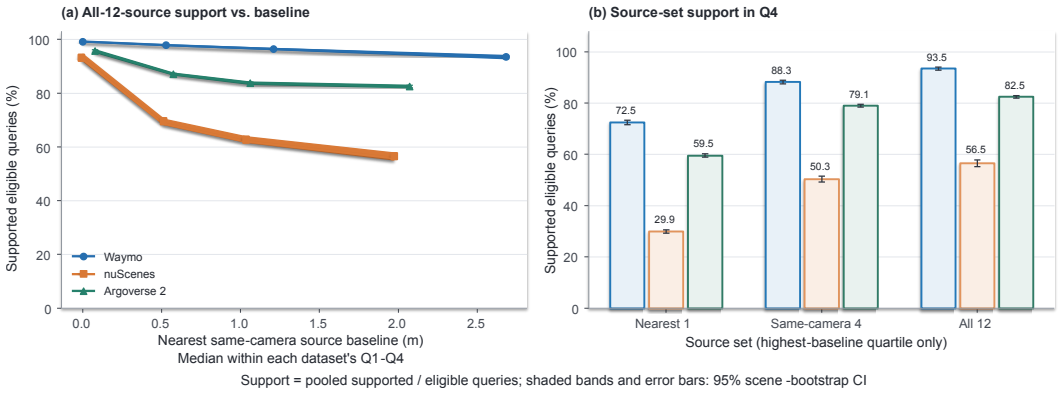


Figure 7: **Target-dependent support and its distribution across source observations.** (a) All-source geometric support decreases from Q1 to Q4 as the median nearest same-camera baseline increases. Shaded bands show 95% scene-bootstrap intervals. (b) For the same Q4 targets, support is measured with three nested source subsets: the nearest same-camera source, all four same-camera times, and all 12 camera–time sources. The source-set comparison is a diagnostic decomposition of the available observations, not a model-input ablation.

Table 5: **All-source support across target-baseline quartiles.** Support percentages use pooled eligible-query counts. Q4–Q1 is in percentage points with 95% scene-bootstrap intervals from 1,000 resamples.

Dataset	Q1	Q2	Q3	Q4	Q4–Q1 [95% CI]
Waymo	99.161	97.846	96.410	93.487	-5.674 $[-6.353, -5.139]$
nuScenes	93.461	69.305	62.728	56.509	-36.952 $[-38.722, -35.105]$
Argoverse 2	95.701	87.114	83.743	82.475	-13.226 $[-13.873, -12.559]$

Table 6: **Support within nested source subsets at Q4.** The same target queries are evaluated against each subset. Support columns are percentages; time/view increments are percentage points.

Dataset	Nearest 1	Same-camera 4	All 12	Time gain (pp)	View gain (pp)
Waymo	72.465	88.291	93.487	15.826	5.196
nuScenes	29.940	50.345	56.509	20.405	6.164
Argoverse 2	59.542	79.050	82.475	19.508	3.425

The main support measurement searches all 12 source images. Within each quartile, we pool supported and eligible target-sample counts rather than averaging per-target percentages. The analysis uses a stride-4 target grid and excludes dilated dynamic masks and sky. In total, 100,164,241 eligible target samples are evaluated. Because reference-depth density differs across datasets, raw support values should not be interpreted as a cross-dataset difficulty ranking. Lower s means less reference-verifiable support, not necessarily less total RGB information or larger model error.

Table 5 and Figure 7(a) show lower pooled support in the larger-baseline groups of all three datasets. The Q_4–Q_1 changes are -5.674, -36.952, and -13.226 percentage points on Waymo, nuScenes, and Argoverse 2, with all scene-bootstrap intervals below zero. This is a group-level association between source–target displacement and reference-verifiable support. The next analysis keeps the Q_4 targets fixed to locate that support within the source set.

A.3 Support across Cameras and Time

For the same Q_4 targets, we compare three nested source subsets: the nearest same-camera source, all four same-camera times, and all 12 camera–time sources. This decomposition keeps the target set

Table 7: **Sensitivity of the baseline–support association.** Q4–Q1 support changes in percentage points, with 95% scene-bootstrap intervals. All listed contrasts remain negative in all three domains.

Control	Waymo	nuScenes	Argoverse 2
Query-weighted bins	-5.623	-36.823	-13.238
95% CI	$[-6.287, -5.084]$	$[-38.600, -34.952]$	$[-13.884, -12.577]$
Input-reference present	-5.273	-29.368	-10.696
95% CI	$[-5.795, -4.830]$	$[-30.973, -27.692]$	$[-11.237, -10.150]$
Internal targets	-5.033	-34.939	-14.190
95% CI	$[-5.787, -4.451]$	$[-36.792, -33.058]$	$[-14.893, -13.508]$
Tail targets	-6.508	-40.596	-11.576
95% CI	$[-7.158, -5.925]$	$[-42.413, -38.646]$	$[-12.303, -10.871]$
1% tolerance	-9.276	-38.787	-15.459
95% CI	$[-10.082, -8.551]$	$[-40.611, -36.895]$	$[-16.235, -14.701]$
5% tolerance	-2.571	-28.765	-8.473
95% CI	$[-3.042, -2.154]$	$[-30.428, -27.085]$	$[-8.985, -7.966]$

fixed and locates the available support within the source set. It does not change the reconstruction task or measure learned fusion gains.

Figure 7(b) and Table 6 show additional support when the nearest same-camera source is expanded to four same-camera times and then to all 12 sources. These increments describe the specified nesting order, not the isolated contribution of a temporal or cross-view network branch. They establish that reference-verifiable evidence for the same targets extends beyond the nearest observation, motivating access to complementary sources.

A.4 Robustness and Scope

Support sensitivity. Table 7 varies bin weighting, reference availability, target time range, and depth tolerance. Unless stated otherwise, the main baseline boundaries remain fixed. Internal targets exclude the four post-context offsets, while tail targets contain only those offsets.

All listed Q4–Q1 contrasts remain negative. Spatial-stride checks on 51 clips change support by at most 0.050, 0.311, and 0.786 percentage points on Waymo, nuScenes, and Argoverse 2. Motion differences also persist when elapsed time is restricted to $1.85, 1.95$ seconds. On Waymo, the fine-mask proxy has a clip-level Spearman correlation of approximately 0.994 with independent dynamic GT under a full-image denominator, supporting ranking agreement but not pixelwise equivalence.

Alternative descriptors. Nearest-source yaw shows the same negative Q4–Q1 association, but translation and rotation are correlated and do not establish independent effects. In contrast, clip depth D90 shows no universal monotonic support relationship across domains and tolerances. We therefore use reference-depth range only as a dataset descriptor, not as a target-level difficulty variable.

Scope and model validation. The diagnosis establishes distinct dataset conditions, a baseline–support association, and complementary support across source times and cameras. It does not measure dynamic-surface support, provide an information bound, or determine a unique architecture. Section 4.3 evaluates model predictions within the same baseline groups, while W/Mix3 comparisons change the training distribution. These are separate empirical tests of reconstruction behavior and training exposure, rather than a causal decomposition of model gains.

B Model Design

This appendix complements Sec. 3 by specifying the implementation choices omitted from the main text. Table 8 summarizes the principal architectural configuration; the following text focuses on the camera–time relation biases, the exact HOC context update, and the source-corrective Gaussian head.

Table 8: **Architecture configuration of SOM4D.** Spatial resolutions correspond to the Waymo input size 160×240; module widths and configurations are shared across datasets.

Module	Component	Configuration
PTRA	Backbone	Frozen Omega, width 1024
	Adaptation	Stages $\{4, 11, 17, 23\}$; 16 heads; FFN 4096; residual scale 10^{-3}
	Conditioning	Camera embedding + Fourier time embedding
HOC	Dense hierarchy	Width 256; grids $5\times8 \to 10\times15 \to 20\times30 \to 40\times60$
	Shared context	512 tokens, width 256; 8 heads; FFN 1024; LayerScale 0.1
	Output	$H_{\text{out}} \in \mathbb{R}^{B\times12\times256\times40\times60}$
CGH	Gaussian decoding	Depth $[0.2, 400]$ m; max scale 0.5; source-RGB residual + degree-1 SH
	Motion	Cubic trajectory with velocity, acceleration, and jerk
	PointWorld	Width 128; 4 pre-, 4 local-, and 9 post-attention blocks; 24 neighbors
	Render correction	2 shared-weight rounds; 96-D recurrent state; 40×60 update grid
Scene	Gaussians	460,800 pixel Gaussians + 12,288 sky Gaussians

B.1 PTRA and Hierarchical Observation Consolidation

PTRA. The frozen Omega patch encoder has width 1024, and four independent PTRA modules are inserted at stages $\{4, 11, 17, 23\}$. PTRA operates on patch tokens, while the camera token and 16 register tokens bypass the adapter. Camera conditioning uses a 17-dimensional descriptor formed by the first two rotation columns, scaled translation, log-distance, translation direction, and four normalized intrinsics; it is encoded by a $17 \to 1024 \to 1024$ GELU MLP. Normalized time is mapped through 256-dimensional Fourier features and a $256 \to 1024 \to 1024$ SiLU MLP. The view-relation bias is predicted from relative pose and paired intrinsics by a $21 \to 64 \to 16$ network, while the temporal-relation bias uses relative pose and Fourier time difference with a $269 \to 64 \to 16$ network, producing one bias value per attention head. Omega remains frozen while gradients pass through the frozen operations to optimize the inserted adapters.

HOC. At each retained level $\ell \in \{4, 11, 17, 23\}$, the four representations F_ℓ^D, F_ℓ^F, F_ℓ^P, and F_ℓ^I are independently normalized and projected from 1024 to 256 dimensions, then fused by the shared softmax scorer in Eq. 3. The dense hierarchy is decoded coarse-to-fine as $H_{23} \to H_{17} \to H_{11} \to H_4$. The shared context starts from 512 learned tokens and is progressively updated from all dense source locations without spatial pooling. The implementation uses the update in Eq. 4 with learned LayerScale coefficients and an additional attention-statistic gate,

$$\widehat{S}_\ell = S_{\text{prev}} + \rho_\ell \gamma_\ell^a \,\mathrm{CA}(\mathrm{LN}(S_{\text{prev}}), \mathrm{LN}(E_\ell), \mathrm{LN}(E_\ell)), \qquad S_\ell = \widehat{S}_\ell + \gamma_\ell^f \,\mathrm{FFN}\Big(\mathrm{LN}(\widehat{S}_\ell)\Big), \tag{11}$$

where γ_ℓ^a and γ_ℓ^f are initialized to 0.1 and $\rho_\ell = 0.1 + 0.9(1 - u_\ell)$ uses the normalized attention-entropy statistic u_ℓ. Equation 4 shows the core update, the implementation additionally uses the residual scales and entropy gate specified here. Final readback follows Eq. 5, with its learned channel scale initialized to 0.1, and produces $H_{\text{out}} \in \mathbb{R}^{B\times12\times256\times40\times60}$.

B.2 Corrective Gaussian Head

Gaussian decoding and motion. The Gaussian and motion tails apply a 3×3 convolution at width 256, bilinear upsampling to the source-image resolution, and UV position encoding. The Gaussian tail concatenates normalized source RGB and predicts depth, scale, quaternion, opacity, and base RGB. Depth follows the camera-Z convention used in Sec. 3.5, with $z = 0.2 + 399.8\sigma(\hat{z})$; scales use a capped exponential with maximum 0.5, quaternions are normalized, and opacity uses a sigmoid. The nine motion channels represent three 3D vectors for velocity, acceleration, and jerk in Eq. 6, while a separate branch predicts the dynamic score. Appearance combines a bounded residual around the source RGB with degree-1 spherical harmonics.

Geometry-aware scene correction. The local metric correction block combines cached frame/inter-frame features with the current Gaussian geometry at width 128. It uses four pre-attention blocks, four local metric blocks with eight heads and 24 neighbors, and nine per-image post-attention blocks. Its

Algorithm 1 Two-round source-render feedback.

Input: corrected scene $\mathcal{G}^0$, source observations $\mathcal{O}$, shared updater U_ϕ

$h^0 \leftarrow 0$
for $r = 0, 1$ **do**
  $\widehat{\mathcal{I}}^{\,r} \leftarrow \mathrm{RenderSrc}(\mathcal{G}^r, \mathcal{O})$
  $e^r \leftarrow \mathrm{sg}\Big(\mathrm{Evidence}(\widehat{\mathcal{I}}^{\,r}, \mathcal{O})\Big)$
  $g^r \leftarrow \mathrm{Attr}(\mathcal{G}^r)$
  $(h^{r+1}, \Delta g^r) \leftarrow U_\phi(h^r, e^r, g^r)$
  $g^{r+1} \leftarrow \mathrm{Compose}(g^r, \Delta g^r)$
  $\mathcal{G}^{r+1} \leftarrow \mathrm{UpdateScene}(\mathcal{G}^r, g^{r+1})$
end for
return $\mathcal{G}^2$

controller predicts residuals for depth, the nine cubic-motion channels, dynamic score, and confidence. The depth update moves Gaussian centers along the corresponding camera rays, while scale, rotation, opacity, and spherical-harmonic appearance remain unchanged at this stage. Cross-view source consolidation then uses peer-view reprojection consistency to reweight opacity and scale while retaining every pixel Gaussian; the configuration uses stride 2, depth tolerances of 0.75 m and 2%, color tolerance 0.2, and minimum peer opacity 0.02. These two steps produce $\mathcal{G}^0$.

Two-round source-render correction. The final correction feeds source-view rendering evidence back to the explicit Gaussian scene. Algorithm 1 expands the recurrent update in Eq. 7; the same updater U_ϕ is reused in both rounds.

Each round renders the current scene to the source cameras at 80×120 and performs the recurrent update on a 40×60 grid. The 35-channel input contains 13 detached source-render evidence channels and a 22-channel Gaussian state consisting of RGB, depth, scale, quaternion, opacity, nine motion channels, and dynamic score. A $35 \to 96 \to 96$ convolutional encoder and a GRU-like 96-channel hidden state feed a $96 \to 96 \to 22$ decoder whose final layer is zero-initialized. The decoder produces 22 residual channels, which are composed with the current Gaussian state to obtain the next scene. Two shared-weight rounds therefore produce $\mathcal{G}^0 \to \mathcal{G}^1 \to \mathcal{G}^2$.

Sky and inference. The sky branch adds $12{,}288$ hemispherical Gaussians at radius 300 m with source-conditioned color and zero motion, giving $473{,}088$ Gaussians in the Waymo scene. The final $\mathcal{G}^2$ is constructed once from the source observations and reused for all target queries; target images and annotations never enter scene construction. For a source frame at normalized time $\tau_i = f_i/15$, motion at query time τ_q uses $\delta = 2(\tau_q - \tau_i)$, so $\delta = 0$ recovers source-anchored Gaussian position.

C Training Details

This appendix summarizes the objective, supervision sources, and optimization protocol used for SOM4D. Appendix B describes the inference architecture. The W/Mix3 comparison retains the objective and optimizer settings below and changes the sampled training distribution.

C.1 Training Objective and Supervision

Following Sec. 3.6, we optimize

$$\mathcal{L} = \mathcal{L}_{\mathrm{RGB}} + 0.05\mathcal{L}_{\mathrm{LPIPS}} + \mathcal{L}_{\mathrm{depth}} + 0.05\mathcal{L}_{\mathrm{subset}} + \mathcal{L}_{\mathrm{aux}}. \tag{12}$$

Target reconstruction. $\mathcal{L}_{\mathrm{RGB}}$ compares the rendered target image with the reference RGB, while $\mathcal{L}_{\mathrm{LPIPS}}$ provides perceptual supervision and is activated after a 1k-step warmup. Target images supervise the rendered output but never enter source-scene construction.

Depth supervision. $\mathcal{L}_{\mathrm{depth}}$ is the implemented masked normalized absolute error between rendered target depth and reference camera-Z depth. We retain finite positive reference depths above 0.01 and use the recorded per-rank batch reference-depth normalization. This supervision is training-only and does not introduce scale alignment at evaluation.

Subset consistency. To regularize the shared HOC context under reduced observations, an independent 3-time/2-camera subset is encoded by the same backbone and HOC. The subset preserves the full branch's reference coordinate system and normalized time labels. $\mathcal{L}_{\text{subset}}$ uses a support/entropy-weighted cosine objective with a stop-gradient full-context target. The subset branch is used only during training and does not change the source observations used at inference.

Auxiliary supervision. $\mathcal{L}_{\text{aux}}$ groups training-only regularizers for motion, source-view appearance and boundaries, intermediate geometry/depth predictions, geometry-correction outputs, and dynamic regions. The full SOM4D W and Mix3 runs retain these terms. Frozen-Omega targets, intermediate probes, and subset forwards supply training supervision and are discarded at inference. No ground-truth scene flow or object tracks are used.

For component ablations, supervision shared by all variants is fixed; a loss defined only on a removed pathway is removed with that pathway. The ablations therefore measure pathway contributions together with their associated training signals, rather than isolating architectural mechanisms at matched capacity.

C.2 Optimization and Training Protocol

We optimize all trainable components with AdamW while keeping the visual-geometry backbone frozen. Main runs use BF16, $\beta = (0.9, 0.999)$, $\epsilon = 10^{-8}$, weight decay 0.05, and global-norm gradient clipping at 1.0. Training lasts 100k optimizer updates with global batch size 32 and no gradient accumulation. The learning rate linearly warms up for 5k updates to 4×10^{-4} and then follows the implemented cosine decay toward the configured minimum 4×10^{-5}. The reference runs use seed 0.

Each training sample uses four fixed source times and four sampled target times. Within SOM4D, Waymo-only and Mix3 runs use the same architecture, supervision, optimizer, and update budget, only the sampled training distribution changes. Mix3 is constructed following the definition in the main text, with equal mixture probabilities over Waymo, nuScenes, and Argoverse 2. Target annotations, frozen-teacher signals, and subset forwards are training-only; inference constructs the scene solely from the source observations.

All component ablations use a separate controlled 50k Waymo-only schedule with 2.5k linear warmup followed by 47.5k cosine decay. Ablation results are therefore compared within this common shortened schedule rather than against a 100k main-model checkpoint.

D Experimental Results

This appendix complements the main evaluation with auxiliary-domain reconstruction, baseline-stratified results, qualitative examples, and source-view ablations. Each subsection addresses a different comparison, and metrics follow Section 4.1.

D.1 Cross-Dataset Reconstruction Results

Tables 9 and 10 evaluate W and Mix3 checkpoints on nuScenes and Argoverse 2. W measures transfer of the Waymo-trained reconstruction model, and Mix3 includes the evaluated domain in training. SOM4D leads all reported NVS and REC metrics under W on both datasets. Under Mix3, it leads all six nuScenes metrics. On Argoverse 2, SLARM has 0.05 dB higher NVS PSNR, while SOM4D leads the remaining metrics. Thus, model rankings remain favorable across the two evaluation settings, but the W-to-Mix3 changes also include target-domain exposure.

Table 9: **Reconstruction on nuScenes.** PSNR↑, SSIM↑, and D-RMSE↓ for held-out NVS and source-view REC on 1,387 validation clips.

Method	Train	NVS			REC		
		PSNR↑	SSIM↑	D-RMSE↓	PSNR↑	SSIM↑	D-RMSE↓
STORM	W	22.44	0.640	8.86	23.60	0.734	8.77
STORM	Mix3	25.01	0.745	5.27	26.34	0.813	5.15
SLARM	W	22.41	0.626	8.87	23.83	0.743	8.76
SLARM	Mix3	26.04	0.767	4.89	27.75	0.850	4.71
SOM4D	W	**22.65**	**0.652**	**7.55**	**24.41**	**0.777**	**7.40**
SOM4D	Mix3	**26.32**	**0.794**	**4.84**	**28.77**	**0.888**	**4.70**

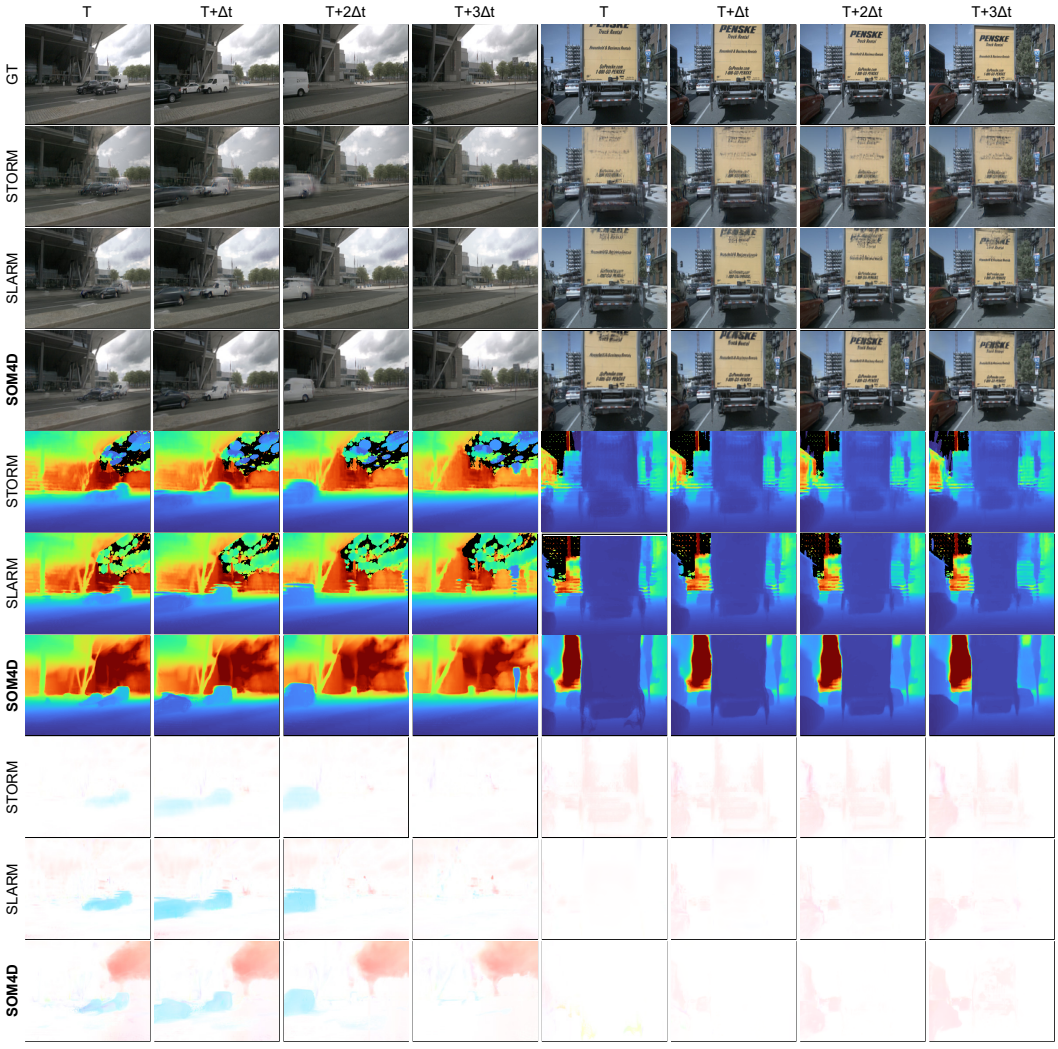


Figure 8: **Qualitative comparison on nuScenes.** Two held-out sequences over four timestamps. Rows show RGB reconstruction, predicted depth, and flow visualizations. In these examples, SOM4D retains clearer vehicle contours and depth boundaries.

Table 10: **Reconstruction on Argoverse 2.** Results on 1,042 validation clips. W is zero-shot from Waymo-only training; Mix3 includes all three training domains.

Method	Train	NVS			REC		
		PSNR↑	SSIM↑	D-RMSE↓	PSNR↑	SSIM↑	D-RMSE↓
STORM	W	19.89	0.603	18.60	21.24	0.708	18.39
STORM	Mix3	24.56	0.782	12.19	25.75	0.836	12.08
SLARM	W	20.06	0.615	17.35	21.57	0.727	17.16
SLARM	Mix3	**25.83**	0.806	11.71	27.43	0.869	11.59
SOM4D	W	**21.07**	**0.637**	**15.99**	**23.22**	**0.767**	**15.77**
SOM4D	Mix3	25.78	**0.824**	**10.90**	**27.83**	**0.892**	**10.77**

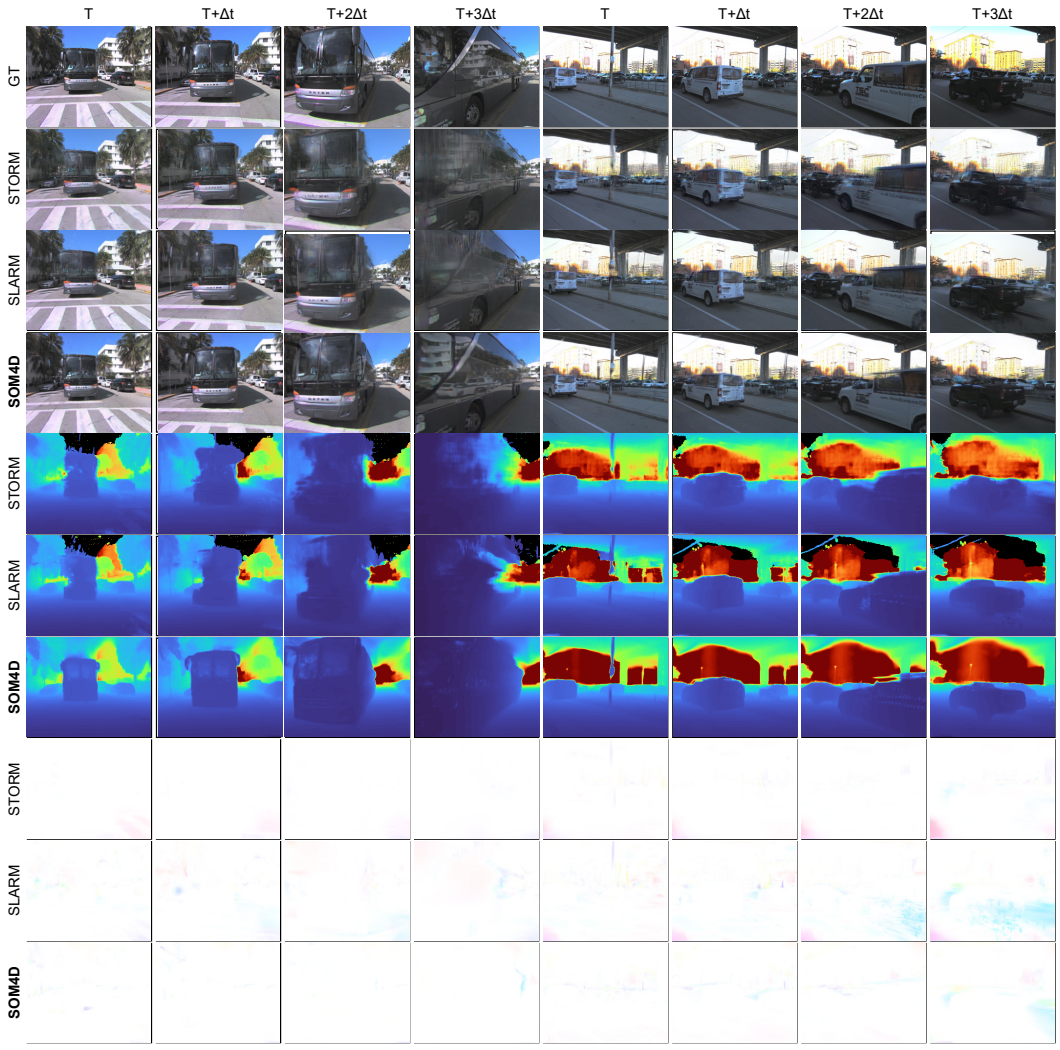


Figure 9: **Qualitative comparison on Argoverse 2.** Two held-out sequences are shown over four timestamps. From top to bottom, the panels show RGB reconstruction (with ground truth), predicted depth, and predicted optical flow. SOM4D shows cleaner large-object boundaries and object–background depth transitions in these examples.

Table 11: **Additional metrics across fixed Waymo target-baseline quartiles of Table 2.** Each entry is SSIM↑/D-RMSE↓; Q_1–Q_4 denote increasing nearest-source displacement.

Method	Train	Q_1	Q_2	Q_3	Q_4
STORM	W	0.864 / 6.28	0.786 / 4.92	0.753 / 4.73	0.724 / 5.19
SLARM	W	**0.900** / 5.61	0.795 / 4.63	0.758 / 4.53	0.731 / 4.97
SOM4D	W	0.891 / **4.50**	**0.840** / **3.78**	**0.804** / **3.82**	**0.764** / **4.21**
STORM	Mix3	0.860 / 5.80	0.798 / 4.32	0.761 / 4.20	0.728 / 4.66
SLARM	Mix3	**0.903** / 5.12	0.811 / 4.17	0.770 / 4.11	0.740 / 4.51
SOM4D	Mix3	0.899 / **4.35**	**0.844** / **3.68**	**0.808** / **3.72**	**0.768** / **4.10**

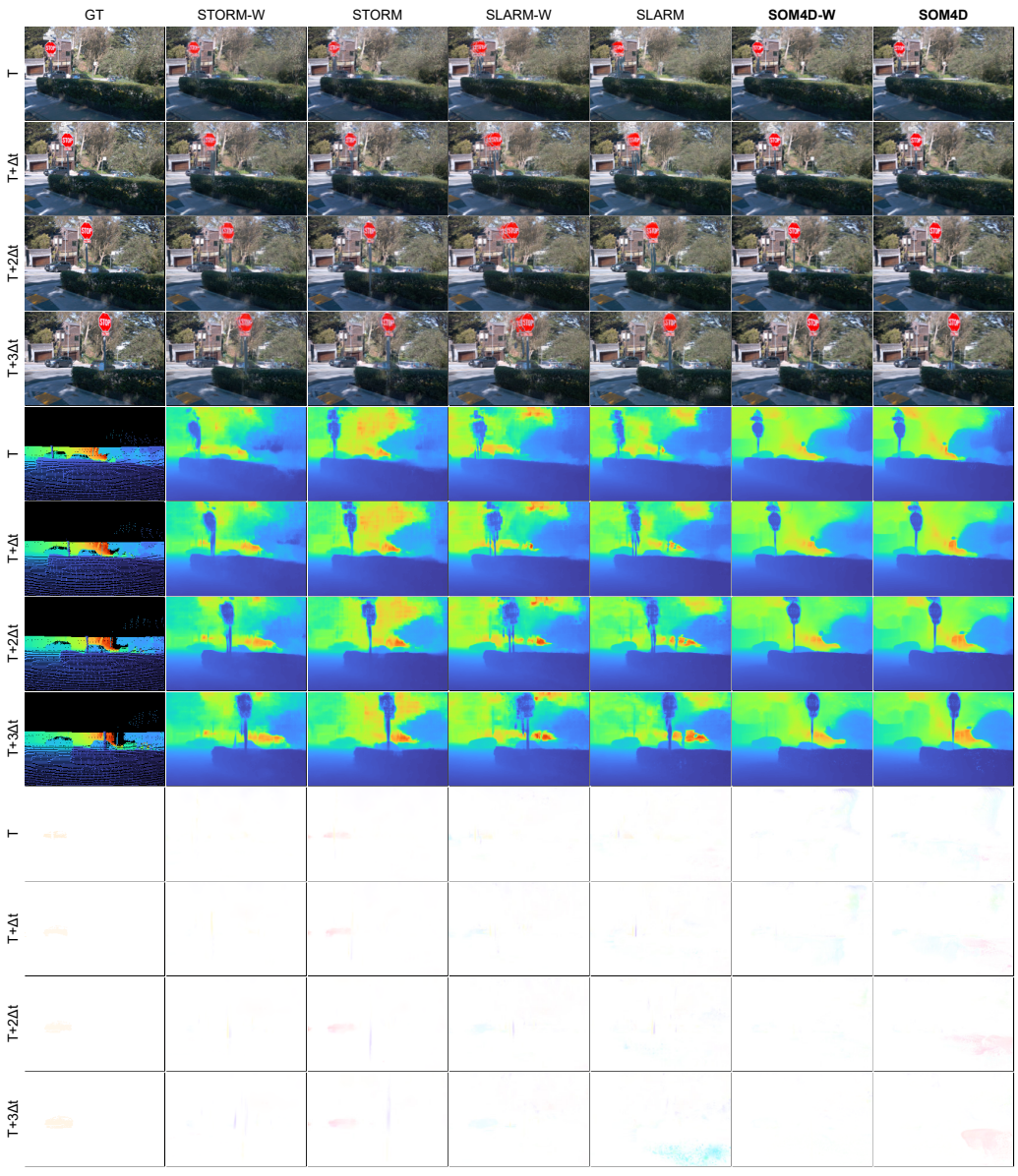


Figure 10: **Waymo comparison under Waymo-only and Mix3 training.** "-W" denotes Waymo-only checkpoints; paired unmarked columns use Mix3 checkpoints on the same target sequence. The three panel groups show RGB, depth, and optical-flow visualizations.

Table 12: **Reconstruction across target-baseline quartiles on nuScenes and Argoverse 2.** All models use Mix3 training. Each entry is PSNR↑/SSIM↑/D-RMSE↓; Q_1–Q_4 denote increasing nearest same-camera source displacement.

Method	Q_1	Q_2	Q_3	Q_4
nuScenes				
STORM	27.30/0.841/5.44	24.97/0.752/5.19	24.40/0.720/5.10	23.38/0.669/5.29
SLARM	**29.16**/0.877/4.86	25.80/0.769/4.80	25.16/0.736/4.77	24.04/0.685/5.02
SOM4D	28.64/**0.879/4.80**	**26.52/0.805/4.79**	**25.70/0.772/4.75**	**24.44/0.720/4.97**
Argoverse 2				
STORM	26.14/0.847/12.81	24.68/0.792/12.07	24.15/0.766/11.54	23.25/0.724/11.64
SLARM	**28.16/0.885**/11.95	25.93/0.815/11.65	25.18/0.785/11.21	24.07/0.741/11.36
SOM4D	26.99/0.870/**11.02**	**26.16/0.839/10.88**	**25.57/0.816/10.49**	**24.38/0.772/10.68**

Table 13: **Model resource comparison under a controlled Waymo profiling protocol.** All methods use 12 source images at 160×240, batch size 1, BF16 autocast, and a single NVIDIA RTX 5090. Latency is measured over 200 clips. Params reports trainable/total parameters.

Method	**Params (M) Train./Total**	**Latency (ms) Mean / P95**	**Peak Mem. (GiB)**
STORM	100.6 / 100.6	127.0 / 139.4	3.23
SLARM	179.5 / 179.5	110.8 / 112.3	4.65
SOM4D	97.0 / 1070.9	226.6 / 233.3	8.08

D.2 Quantitative Results

Table 11 adds SSIM to the fixed Waymo baseline groups, SLARM leads in Q_1, whereas SOM4D leads in Q_2–Q_4 under both W and Mix3. The pattern matches the main PSNR analysis: SLARM is strongest in the closest-source regime Q_1, whereas SOM4D is strongest from Q_2 through Q_4 and has the lowest D-RMSE in every quartile under both W and Mix3. Table 12 shows the corresponding cross-dataset pattern under Mix3. SLARM retains the Q_1 PSNR advantage on both auxiliary domains, while SOM4D leads all three metrics in Q_2–Q_4. The Q_1 SSIM ranking is not identical across domains, SOM4D is slightly higher on nuScenes, whereas SLARM is higher on Argoverse 2. SOM4D has the lowest D-RMSE in every reported quartile. These results locate its advantage across baseline regimes rather than demonstrate a uniform appearance improvement or identify its mechanism.

Table 13 reports higher latency and peak memory for SOM4D despite its smaller trainable parameter count. Its frozen backbone contributes to the much larger total parameter count. The comparison describes a reconstruction-quality versus inference-cost trade-off under the stated profiling protocol.

D.3 Qualitative Results

Figures 8 and 9 compare methods on matched held-out queries over four timestamps within each dataset. The displayed SOM4D results show clearer contours of foreground vehicles and more distinct object–background depth transitions. Figure 10 pairs W and Mix3 checkpoints on the same Waymo sequence, where Mix3 generally yields cleaner object boundaries and more coherent depth around foreground vehicles, with corresponding qualitative differences also visible in the predicted optical flow.

D.4 Ablation Studies

Table 14 extends the NVS ablation with source-view REC. PTRA removal gives the largest NVS PSNR drop, whereas removing the source-corrective head gives the largest REC PSNR drop. Removing HOC reduces both NVS and REC PSNR. Source-render feedback mainly benefits appearance

Table 14: **Complete pathway ablation on Waymo.** Each entry is PSNR↑/D-RMSE↓. NVS evaluates held-out targets and REC evaluates source-view reconstruction.

Variant	NVS Dynamic	NVS Full†	REC Dynamic	REC Full†
w/o PTRA	21.54 / 6.47	26.19 / 4.30	24.69 / 5.99	28.20 / 4.16
w/o camera/time embeddings	22.37 / 6.03	26.81 / 4.01	25.24 / 5.70	28.83 / 3.90
w/o HOC	22.29 / 6.02	26.69 / 3.99	25.11 / 5.69	28.64 / 3.88
w/o shared context S	22.41 / 6.00	26.85 / 3.98	25.32 / 5.68	28.88 / 3.87
w/o source-corrective head	22.04 / 6.06	26.39 / 4.02	24.57 / 5.73	28.13 / 3.92
w/o source-render correction	22.24 / 6.00	26.46 / 3.99	24.76 / 5.69	28.22 / 3.88
Full SOM4D	**22.51 / 6.00**	**26.86 / 3.97**	**25.33 / 5.67**	**28.89 / 3.86**

rather than depth. These comparisons retain shared supervision and remove pathway-specific losses with their pathway, as defined in Section 4.4.