

UFO: Unifying Feed-Forward and Optimization-based Methods for Large Driving Scene Modeling

Supplementary Material

1. Video Results

We provide video demonstrations of our method’s reconstruction quality and dynamic modeling capabilities in the `video` folder. Each video showcases novel view synthesis results on sequences of varying lengths from the Waymo Open Dataset validation set. The videos include renderings of RGB images, depth maps, as well as visualizations of predicted lifespan maps and object motion assignments to illustrate how our method handles transient phenomena and dynamic objects.

The provided video files follow the naming convention `[scene_idx]_[ctx_freq]_[length].mp4`, where: `scene_idx` is the index of the scene in the validation set. `ctx_freq` is the number of context frames used per second (maximum: 10 fps). `length` is the total sequence duration in seconds.

2. Additional Dataset & Implementation Details

We train and evaluate our model using the Waymo Open Dataset [2], following their official train-validation split. We downsample the images to 160×240 and report results on the front, front-left, and front-right cameras, consistent with previous work [7]. The model is trained on 16 NVIDIA H200 GPUs with a total batch size of 64 for approximately one day using the AdamW optimizer.

We warm up the model for 5,000 iterations with a linearly increasing learning rate from 0 to 1×10^{-4} , after which a constant learning rate of 1×10^{-4} is applied. Our training objective consists of the following loss components:

- RGB MSE Loss ($\lambda_{\text{RGB}} = 1.0$): Base reconstruction loss.
- Perceptual Loss ($\lambda_{\text{LPIPS}} = 0.05$): LPIPS loss, activated after iteration 5000.
- Depth Loss ($\lambda_{\text{depth}} = 1.0$): L1 loss on LiDAR depth.
- Flow Regularization ($\lambda_{\text{flow}} = 0.005$): MSE on forward flow.
- Lifespan Regularization ($\lambda_{\text{lifespan}} = 0.0001$): Encourages persistent Gaussians.
- Sky Depth Loss ($\lambda_{\text{sky-depth}} = 0.01$): MSE for sky regions.
- Sky Opacity Loss ($\lambda_{\text{sky-opacity}} = 0.1$): L1 loss on sky opacity.
- Object Assignment Loss ($\lambda_{\text{obj}} = 1.0$): Cross-entropy for object classification.

3. Recurrent Scene Update

Algorithm 1 provides a detailed specification of our recurrent scene update mechanism described in Section ?? of the main paper. At each timestep t , the algorithm processes a new observation by: (1) selecting visible scene tokens from the previous state \mathcal{S}_{t-1} based on camera frustum and proximity, (2) transforming these tokens to the local camera-centric coordinate system to stabilize learning, (3) applying the transformer update $\mathcal{T}_{\text{update}}$ to refine existing tokens and generate new ones, and (4) merging the refined and new tokens back into the global scene representation. The visibility filtering mechanism ensures computational efficiency by limiting attention to the most relevant K tokens at each step, enabling near-linear scaling with sequence length. This selective updating strategy allows the model to iteratively refine previously observed geometry as new viewpoints become available, distinguishing our approach from standard feed-forward methods that lack refinement capabilities. Auxiliary tokens for sky modeling and affine color correction are maintained and updated throughout the sequence. After processing all timesteps, the final scene representation \mathcal{S}_T is decoded into 3D Gaussians for rendering.

Algorithm 1 Recurrent Scene Update for UFO

Require: Images $\{I_t\}_{t=1}^T$, poses $\{P_t\}_{t=1}^T$, intrinsics $\{\mathbf{K}_t\}_{t=1}^T$, tracked boxes $\{B_t\}_{t=1}^T$, visible-token budget K

- 1: $\mathcal{S}_0 \leftarrow \emptyset$ ▷ Scene-token memory
- 2: $\theta_{\text{sky}}, \theta_{\text{affine}} \leftarrow \text{InitAuxTokens}()$ ▷ Shared auxiliary tokens
- 3: **for** $t = 1$ to T **do**
- 4: $R_t \leftarrow \text{PlückerRays}(I_t, P_t, \mathbf{K}_t)$ ▷ Per-pixel ray dirs from pose & intrinsics
- 5: $X_t \leftarrow \text{ImageTokens}(I_t, R_t, t)$ ▷ RGB + Plücker rays + time/pos. enc.
- 6: $\mathcal{V}_{t-1} \leftarrow \text{Visible}(\mathcal{S}_{t-1}, P_t, K)$ ▷ Frustum filter + nearest K tokens
- 7: $\mathcal{V}_{t-1} \leftarrow \text{ToLocal}(\mathcal{V}_{t-1}, P_t)$ ▷ Camera-centric coordinates
- 8: $\mathcal{V}'_{t-1}, \Delta\mathcal{S}_t, \theta_{\text{sky}}, \theta_{\text{affine}} \leftarrow \mathcal{T}_{\text{update}}(X_t, \mathcal{V}_{t-1}, B_t, \theta_{\text{sky}}, \theta_{\text{affine}})$
- 9: $\mathcal{S}_t \leftarrow (\mathcal{S}_{t-1} \setminus \mathcal{V}_{t-1}) \cup \mathcal{V}'_{t-1} \cup \Delta\mathcal{S}_t$
- 10: **end for**
- 11: $\mathcal{G}_T \leftarrow \text{GaussianDecoder}(\mathcal{S}_T, \theta_{\text{sky}}, \theta_{\text{affine}})$

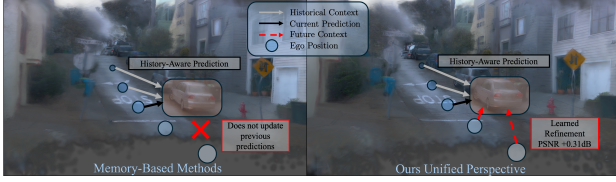


Figure 1. **Comparison with memory-based streaming methods.** Existing streaming pipelines use memory mainly to condition current predictions under temporal causality. UFO explicitly revisits and refines the persistent scene representation as new frames arrive, enabling future-to-past error correction.

4. Comparison with Memory-Based Streaming Methods

Several recent methods employ memory mechanisms for streaming 3D reconstruction, including CUT3R [4], Point3R [5], Spann3R [3], and StreamVGGT [8]. While these methods share a recurrent processing structure with UFO, they differ in a fundamental aspect: the treatment of historical predictions.

As illustrated in Figure 1, existing streaming pipelines use memory as historical context under temporal causality—once a prediction is made, it remains fixed and cannot be revised. Memory serves primarily to condition the current prediction on past observations. In contrast, UFO explicitly revisits and refines previously predicted scene content as new frames arrive. This design draws inspiration from optimization-based methods (3DGS, NeRF) that iteratively refine scenes through render-compare-update loops, but abstracts these loops into a learned module that achieves similar error-correction without explicit rendering or gradient computation.

This distinction is not merely incremental: it enables UFO to correct uncertain predictions—such as distant objects first observed at low resolution—once higher-quality observations become available from closer viewpoints. This future-to-past refinement capability is architecturally impossible in causal streaming pipelines without fundamental redesign. We validate the importance of this refinement mechanism through ablations in Table ?? of the main paper, where our refinement paradigm yields +0.31 dB PSNR over 3D-aware memory alone.

5. Comparison with LiDAR-Supervised Methods

In the main paper, we disable LiDAR-based initialization and depth supervision for optimization-based baselines to ensure a fair comparison with feed-forward methods, which do not require LiDAR at test time. Here, we additionally compare against per-scene optimization methods with LiDAR supervision enabled, to demonstrate the effectiveness of UFO under the strongest baseline configurations.

Method	PSNR \uparrow	SSIM \uparrow	D-RMSE \downarrow
PVG	21.79	0.599	17.21
PVG (+LiDAR)	23.11	0.717	6.43
StreetGS	22.67	0.675	14.88
StreetGS (+LiDAR)	21.40	0.627	5.38
<i>Ours-UFO</i>	27.04	0.816	5.08

Table 1. **Comparison with LiDAR-supervised optimization methods on 16s sequences.** UFO outperforms per-scene optimization baselines across all metrics, even when LiDAR initialization and depth supervision are enabled for the baselines.

As shown in Table 1, LiDAR supervision significantly improves geometric accuracy (Depth RMSE) for optimization-based methods, particularly for PVG. However, UFO still achieves superior performance across all metrics without requiring LiDAR at test time. This result highlights the strong geometric reasoning learned by our model during training, and demonstrates its potential for closed-loop simulation scenarios where LiDAR data may not be available.

6. Adversarial Lighting Conditions

Beyond modeling deformable objects, the lifespan parameter β naturally handles transient lighting artifacts such as lens flare, windshield reflections, and sunlight glare. These phenomena are temporally unstable and view-dependent, causing traditional methods to incorrectly bake them into static geometry, resulting in ghosting effects that degrade visual and geometric quality.

Our method addresses this through learned temporal consistency. As shown in Figure 2, when the model observes localized bright regions appearing inconsistently across viewpoints and timesteps, it automatically assigns very short lifespans to the corresponding Gaussians, causing them to decay rapidly outside their narrow temporal window. Conversely, genuine scene content with high spatio-temporal consistency receives long lifespans, ensuring stable geometry. This behavior emerges naturally from our training objective without explicit artifact detection or supervision.

7. Limitations

Bounding-Box Dependency. Our dynamic object modeling relies on 3D bounding boxes from external tracking systems. Noisy or missing boxes can degrade reconstruction quality for the affected objects. While this assumption is common in driving-scene Gaussian splatting methods (e.g., Street Gaussians [6], OmniRe [1]) to enable scene editability, it introduces a dependency on the quality of the upstream perception pipeline. Our contribution in this regard is learning soft Gaussian-to-object assignment with lifespan prediction in a feed-forward model, which enables

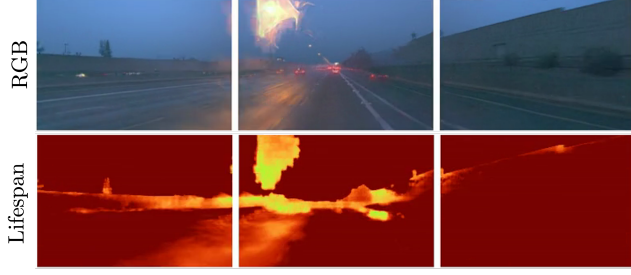


Figure 2. **Lifespan under adverse lighting conditions.** Short-lived Gaussians (blue) automatically capture transient artifacts such as lens flare and windshield reflections, while long-lived Gaussians (warm colors) represent persistent scene geometry.

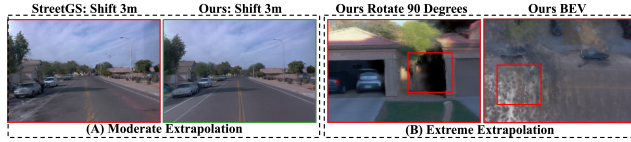


Figure 3. **Failure case: extreme view extrapolation.** UFO handles moderate viewpoint shifts well (A) but produces artifacts under extreme extrapolation (B) where no observations are available.

complex long-range motion modeling without kinematic assumptions such as constant velocity.

Extreme View Extrapolation. As shown in Figure 3, UFO performs significantly better than optimization-based methods under moderate viewpoint shifts (A), producing coherent geometry where per-scene methods fail. However, under extreme viewpoint extrapolation (B), our method produces holes and artifacts in regions that were never observed during reconstruction. Since UFO reconstructs scenes from observed visual evidence, it cannot hallucinate plausible content for entirely unobserved regions. Incorporating generative priors to fill in missing regions is a promising direction for future work.

References

- [1] 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 *ICLR*, 2025. 2
- [2] Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R. Qi, Yin Zhou, Zoey Yang, Aur’elien Chouard, Pei Sun, Jiquan Ngiam, Vijay Vasudevan, Alexander McCauley, Jonathon Shlens, and Dragomir Anguelov. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9710–9719, 2021. 1
- [3] Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory. *arXiv preprint arXiv:2408.16061*, 2024. 2
- [4] Qianqian Wang*, Yifei Zhang*, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *CVPR*, 2025. 2
- [5] Yuqi Wu, Wenzhao Zheng, Jie Zhou, and Jiwen Lu. Point3r: Streaming 3d reconstruction with explicit spatial pointer memory. *arXiv preprint arXiv:2507.02863*, 2025. 2
- [6] 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 *ECCV*, 2024. 2
- [7] Jiawei Yang, Jiahui Huang, Yuxiao Chen, Yan Wang, Boyi Li, Yurong You, Maximilian Igl, Apoorva Sharma, Peter Karkus, Danfei Xu, Boris Ivanovic, Yue Wang, and Marco Pavone. Storm: Spatio-temporal reconstruction model for large-scale outdoor scenes. In *ICLR*, 2025. 1
- [8] Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou, and Jiwen Lu. Streaming 4d visual geometry transformer. *arXiv preprint arXiv:2507.11539*, 2025. 2