Extrapolated Urban View Synthesis Benchmark

Supplementary Material

Appendix A: Methods Discussions

Planar-Based vs. Ellipsoid-Based. Planar-based methods (e.g., GSPro [3], PGSR [2], and 2DGS [5]) excel in road representation due to their planar geometry and refinement strategies but struggle with fine-textured urban objects like plants and fences. Conversely, ellipsoid-based methods (e.g., 3DGS [7] and 3DGM [8]) better handle high-textured objects but often overfit, leading to errors in road representation. For instance, in the translation setting (Figure Ia), planar-based methods struggle with plants, while ellipsoid-based methods perform poorly on roads. A hybrid representation could effectively combine the strengths of both approaches to address these challenges in EUVS.

Enhancing View Synthesis with Diffusion Priors. While training cameras may collectively cover the entire scene, the limited number of viewpoints often results in insufficient representation of certain areas. Leveraging diffusion priors proves to be an effective approach in such cases. By supervising augmented views with diffusion priors, unseen or poorly represented views can be generated and corrected. For instance, as shown in Figure Ib, the building rendered by other models is fragmented, but guiding with diffusion priors helps complete the building structure and presents a holistic urban scene. On average, in Table 1 of the main paper, VEGS [6] with diffusion priors significantly outperforms 3DGS [7] in the rotation-only setting, achieving a 19.4% increase in PSNR (23.33 vs. 19.53) and a 5.8% improvement in SSIM (0.7949 vs. 0.7511).

Regularization by Depth Priors. Utilizing depth priors from foundation models, such as Depth Anything [9], has proven to be an effective approach for enhancing training regularization [4]. In our experiments, depth regularization enhances geometric accuracy by utilizing depth information to constrain Gaussians in regions like the sky and road to more geometrically consistent positions. As shown in Figure Ic, the sky is accurately constrained to a distant position, ensuring it does not overlap with the building during lane changes. Similarly, the road is aligned to a consistent plane, effectively mitigating the distortion issues observed in the vanilla baseline. The regularization by depth priors ensures spatial consistency and reduces visual artifacts, leading to more reasonable extrapolated views.

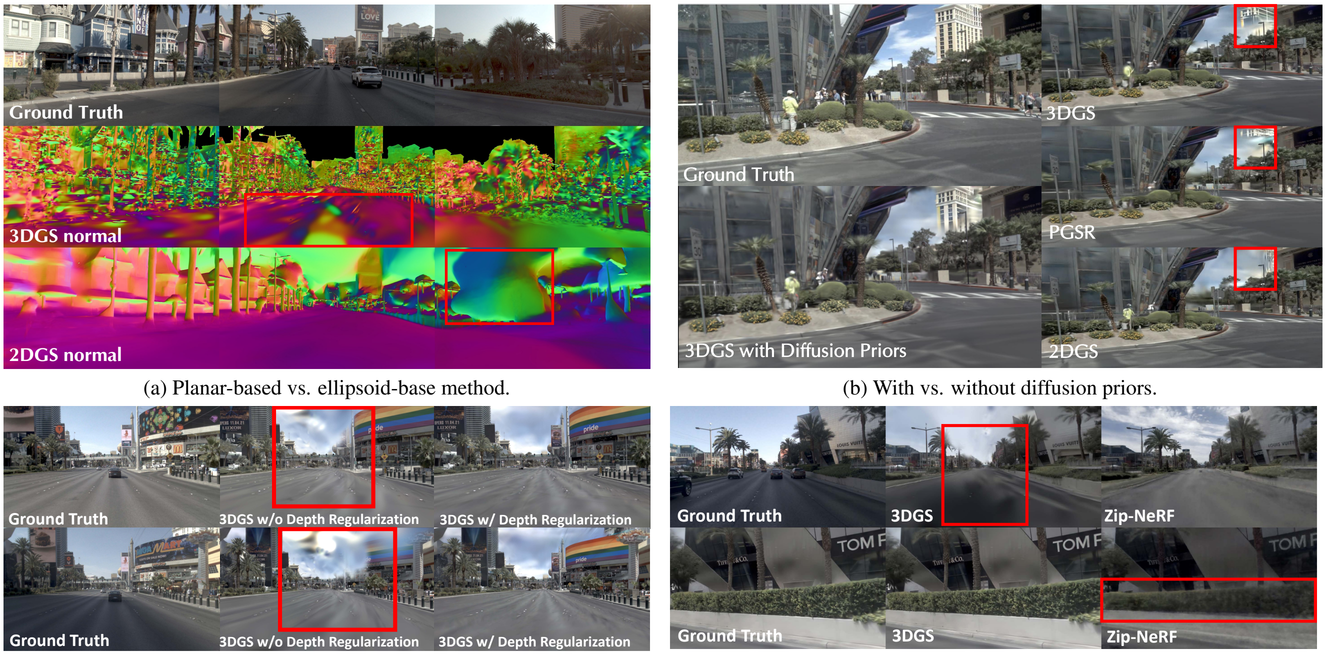
Gaussian-Based vs. NeRF-Based Methods. A fundamental difference between Gaussian-based and NeRF-based approaches lies in their representation: Gaussian-based methods rely on explicit representations, whereas NeRF-based methods use implicit representations. Our experiments reveal that implicit methods, such as Zip-NeRF [1], preserve overall geometry more consistently under large shifts, though it can still lose some sharpness even with small viewpoint extrapolations. In contrast, the explicit representation of Gaussian Splatting-based methods excels in regions with accurate geometry, producing sharper fine details (e.g., foliage), but struggles with incomplete geometry under large shifts. , as illustrated in Figure Id.

Performance Gains from Multi-Traversal Data. Multi-traversal data plays a critical role in Extrapolated View Synthesis. Using the GaussianPro model [3] in Setting 1, we progressively increase the number of training traversals to observe its impact. The results, shown in Figure III and Figure II, indicate that as the number of traversals increases, the NVS metrics for the test view gradually improve, then plateau. This consistent improvement stems from increased unique observations, enabling diverse perspectives and more accurate background reconstruction while reducing dynamic object influence. This suggests that incorporating more visual data can help improve the performance of extrapolated view synthesis.

Appendix B: Comparison of Baselines

Quantitative Comparison. We report the quantitative performance comparison across all settings and baselines in Figure IV. **(1)** In Setting 1, the performance gaps on the extrapolative test set are small, with most baselines performing comparably poorly. Among them, 3DGS [7], 3DGM [8], and GSPro achieve relatively better results. **(2)** In Setting 2 (Figure IVb), in extrapolated views, VEGS [6] significantly outperforms all other methods, achieving at least 20% higher PSNR. These results highlight the effectiveness of diffusion priors in rotation-only settings. **(3)** In Setting 3, as shown in Figure IVc, none of the baselines exhibit a clear advantage, as all methods fail equally in this challenging setting. On the extrapolative test set, different baselines exhibit strengths in specific metrics, but no method demonstrates superiority across all metrics, indicating that all baselines struggle with extrapolated view synthesis and fail to address it fundamentally.

Qualitative Comparison. We present the qualitative baseline comparison across all settings and baselines in Figure V, Figure VI and Figure VII. **(1)** In Setting 1, as shown in Figure Vb, all methods exhibit imperfections in ground rendering, while planar-based methods such as 2DGS [5] and PGSR [2] show comparatively fewer flaws on the ground surface. GSPro [3] produces more accurate geometry reconstruction, achieving realistic surfaces and high-fidelity representations of street objects like trees and


Ground Truth
3DGS normal
2DGS normal
Ground Truth
3DGS
PGSR
3DGS with Diffusion Priors
2DGS
Ground Truth
3DGS w/o Depth Regularization
3DGS w/ Depth Regularization
Ground Truth
3DGS w/o Depth Regularization
3DGS w/ Depth Regularization
Ground Truth
3DGS
Zip-NeRF
Ground Truth
3DGS
Zip-NeRF

(a) Planar-based vs. ellipsoid-base method.

(b) With vs. without diffusion priors.

(c) With vs. without depth priors.

(d) GS-based vs. NeRF-based.

Figure I. **Qualitative comparison of different techniques.** The various techniques excel in different aspects, showing some trade-offs in extrapolated view synthesis. Although they can partially address the challenges, they fail to resolve the underlying issues fundamentally.

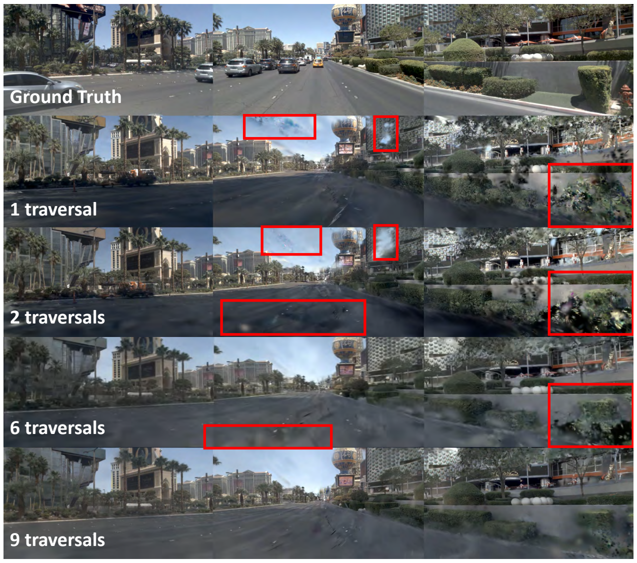


Figure II. **As the number of traversals increases, the performance of NVS improves.** This is highlighted in the red box, where the texture progressively enriches and errors in areas like the sky and ground are reduced.

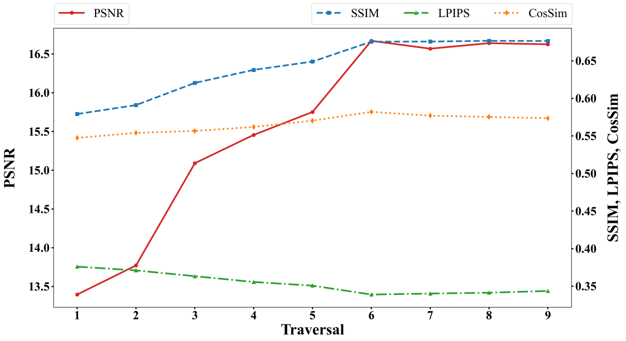
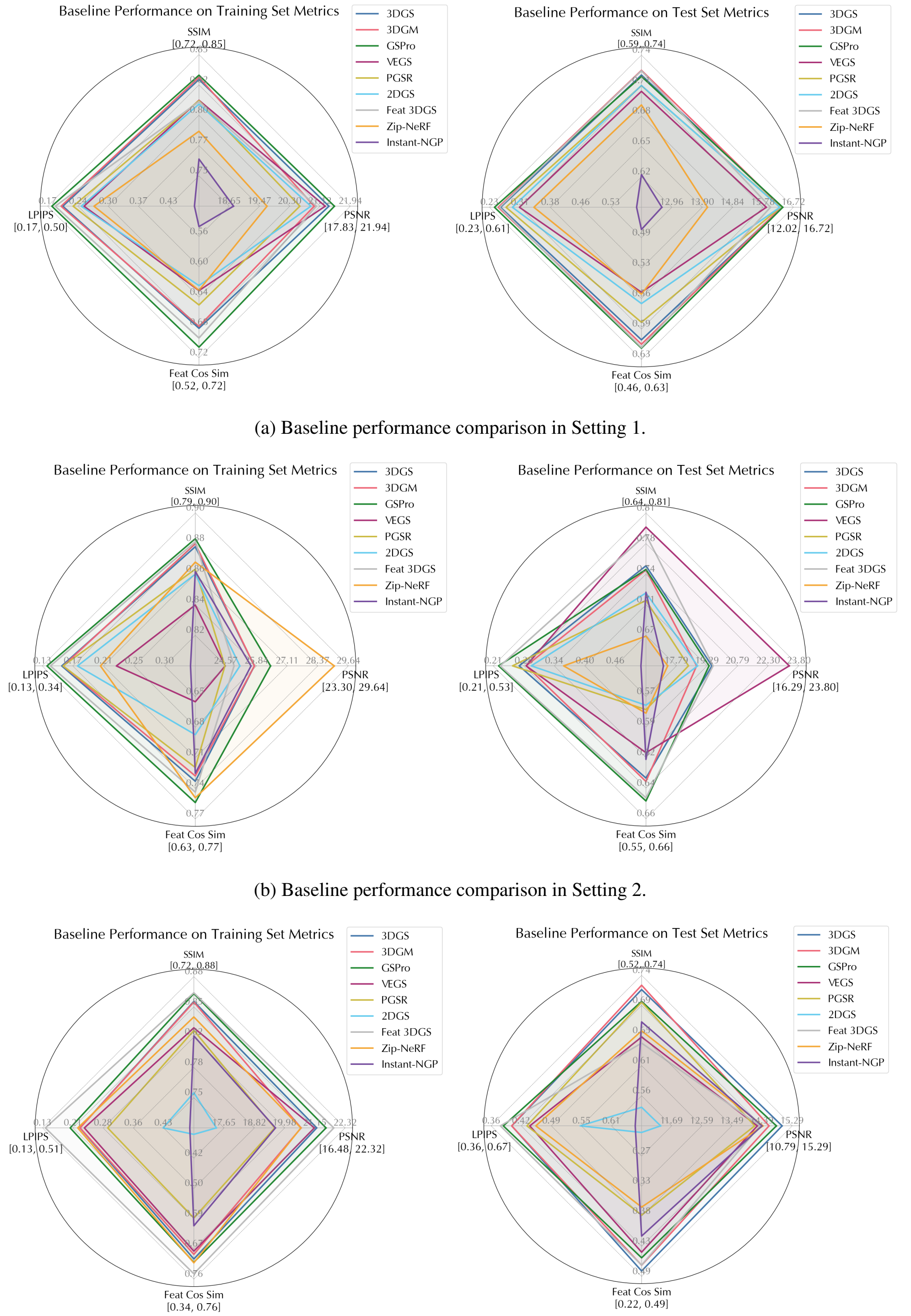


Figure III. **NVS performance vs. number of traversals.** With more traversals, PSNR and SSIM exhibit notable improvements, indicating enhanced image quality and structural similarity. LPIPS values decrease, reflecting better perceptual consistency, while CosSim stabilizes after an initial rise. These results highlight the importance of more visual data for improving NVS performance.

buildings. **(2)** In Setting 2, as shown in Figure VI, most baselines suffer from sky artifacts such as holes and floating objects. In contrast, VEGS [6] produces the more accurate renderings, exhibiting minimal floating artifacts and broken geometry, attributed to the guidance provided by diffusion priors. **(3)** In Setting 3, as shown in Figure VIIb, all baselines face significant challenges on the test set. The geometry across all methods appears highly fragmented, and the color consistency is compromised, reflecting a tendency to overfit to the training views. Among the baselines, 2DGS and PGSR show relatively weaker performance, underscoring the limitations of planar representations in effectively capturing the complexity of whole urban scenes.

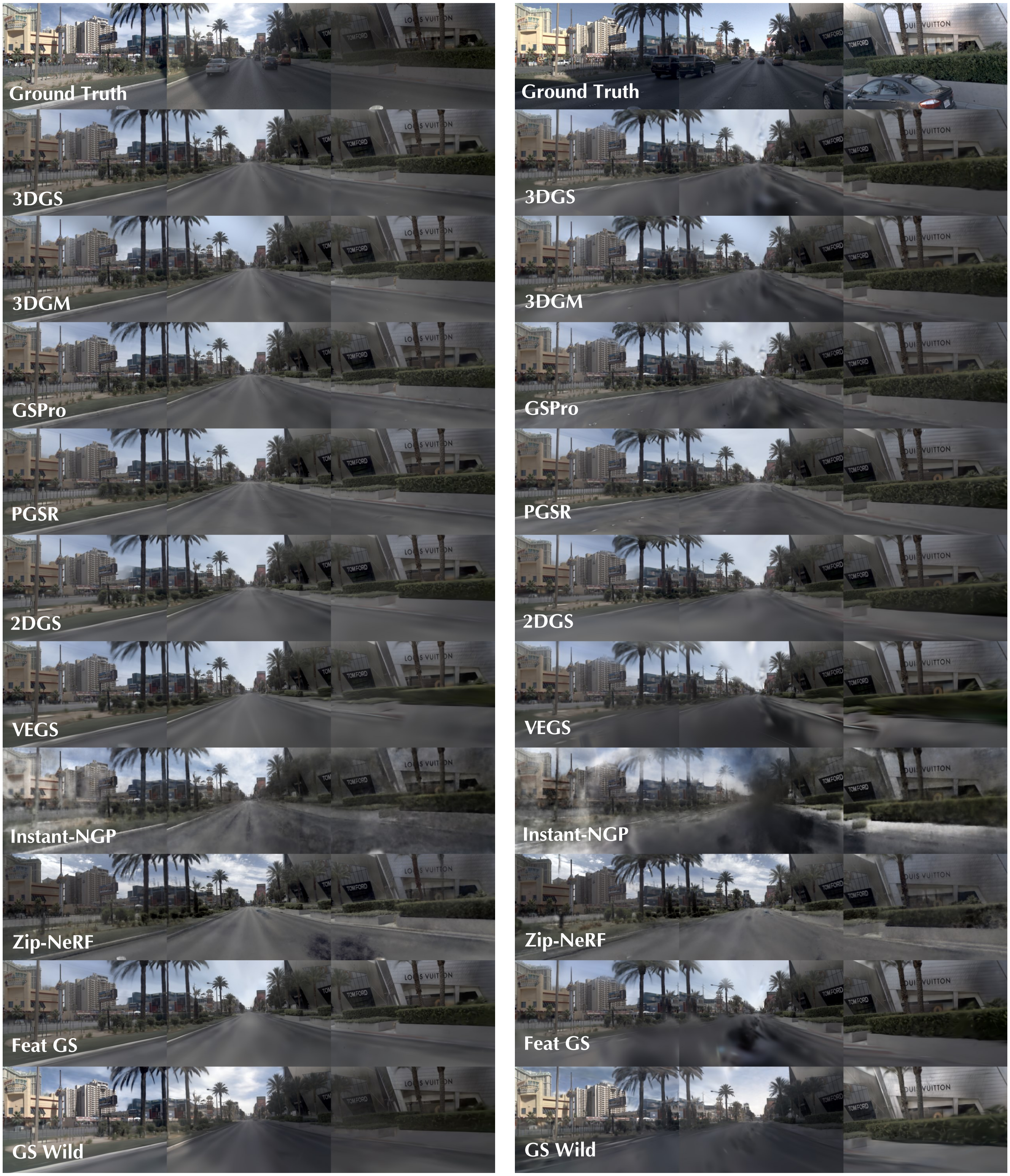


(a) Baseline performance comparison in Setting 1.

(b) Baseline performance comparison in Setting 2.

(c) Baseline performance comparison in Setting 3.

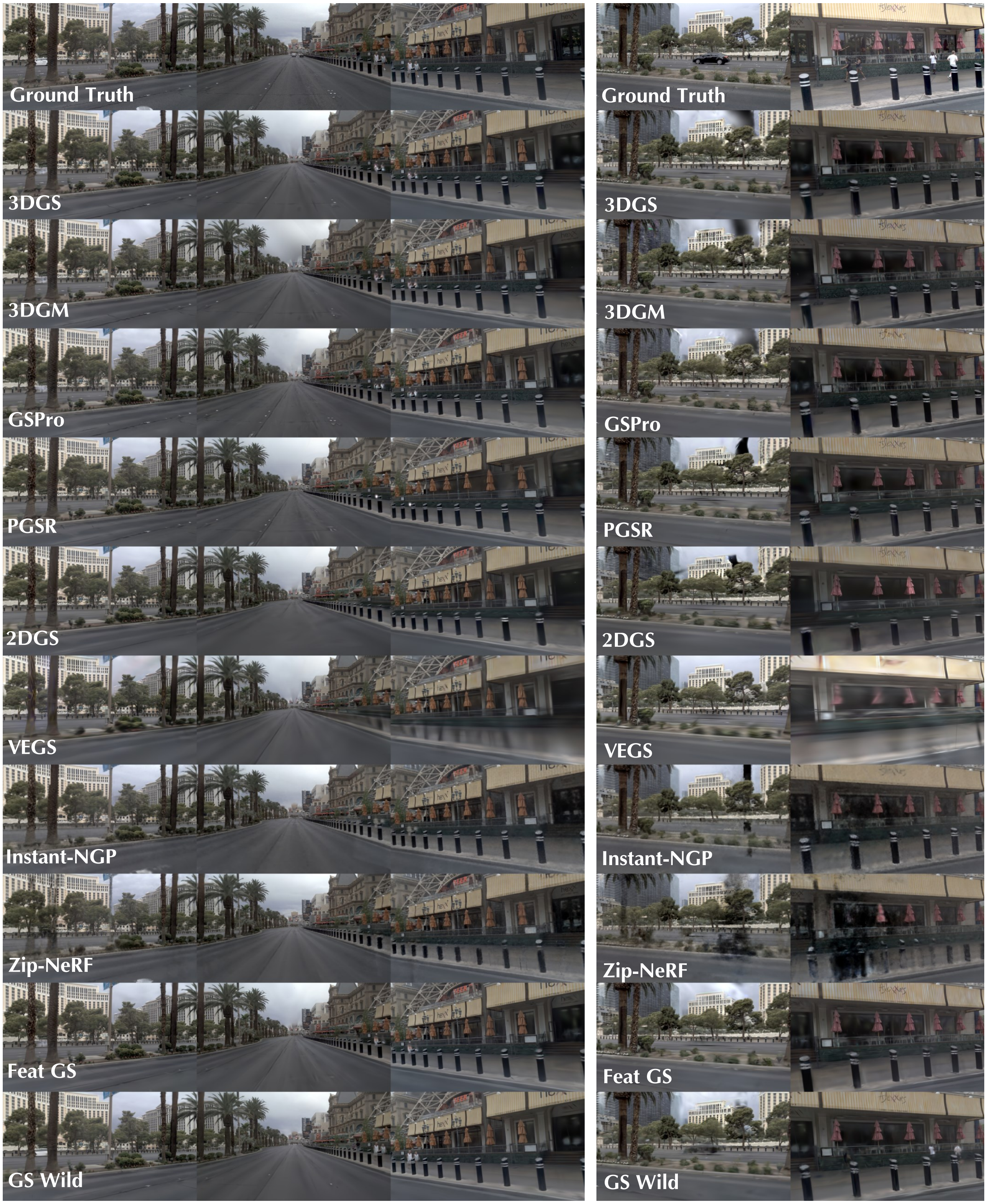
Figure IV. **Baseline performance comparison across different settings.** Since scenes in different settings evaluate varying capabilities, different baselines demonstrate strengths in different evaluation settings.



(a) Rendering results comparison in original view.

(b) Rendering results comparison in extrapolated view.

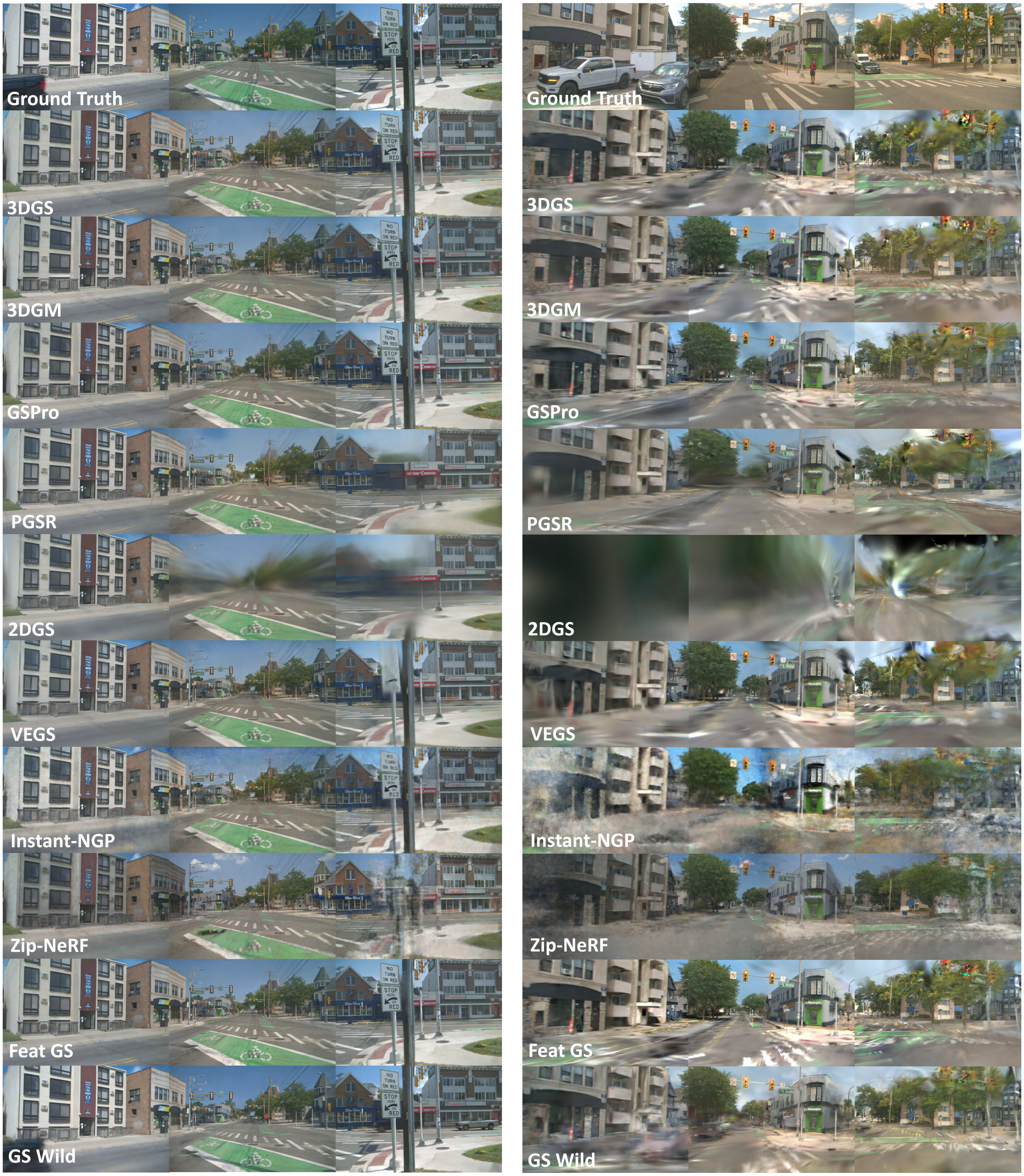
Figure V. **Qualitative comparison of baseline methods in Setting 1.** Ground reconstruction failures and floating artifacts in the sky are particularly noticeable, highlighting the challenges in the lane change.



(a) Rendering results comparison in original view.

(b) Rendering results comparison in extrapolated view.

Figure VI. **Qualitative comparison of baseline methods in Setting 2.** The three front and three back cameras (six in total) are used for training, while the two side cameras are reserved for testing. To ensure clarity and conciseness, only a subset of the training cameras is visualized here due to space limitations.



(a) Rendering results comparison in original view.

(b) Rendering results comparison in extrapolated view.

Figure VII. **Qualitative comparison of baseline methods in Setting 3.** The rendering quality deteriorates significantly in extrapolated viewpoints. The geometry becomes fragmented, especially in trees, traffic lights, and lane marks.

References

[1] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 19697–19705, 2023. 1

[2] Danpeng Chen, Hai Li, Weicai Ye, Yifan Wang, Weijian Xie, Shangjin Zhai, Nan Wang, Haomin Liu, Hujun Bao, and Guofeng Zhang. Pgsr: Planar-based gaussian splatting for efficient and high-fidelity surface reconstruction. arXiv preprint arXiv:2406.06521, 2024. 1

[3] Kai Cheng, Xiaoxiao Long, Kaizhi Yang, Yao Yao, Wei Yin, Yuexin Ma, Wenping Wang, and Xuejin Chen. Gaussianpro: 3d gaussian splatting with progressive propagation. In Forty-first International Conference on Machine Learning, 2024. 1

[4] Jaeyoung Chung, Jeongtaek Oh, and Kyoung Mu Lee. Depth-regularized optimization for 3d gaussian splatting in few-shot images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 811–820, 2024. 1

[5] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In ACM SIGGRAPH 2024 Conference Papers, pages 1–11, 2024. 1

[6] Sungwon Hwang, Min-Jung Kim, Taewoong Kang, Jayeon Kang, and Jaegul Choo. Vegs: View extrapolation of urban scenes in 3d gaussian splatting using learned priors. arXiv preprint arXiv:2407.02945, 2024. 1, 2

[7] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics, 42(4), 2023. 1

[8] Yiming Li, Zehong Wang, Yue Wang, Zhiding Yu, Zan Gojcic, Marco Pavone, Chen Feng, and Jose M. Alvarez. Memorize what matters: Emergent scene decomposition from multitraverse. In Advances in Neural Information Processing Systems (NeurIPS), 2024. 1

[9] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. arXiv preprint arXiv:2406.09414, 2024. 1