

Extrapolated Urban View Synthesis Benchmark

Xiangyu Han^{1,3*} Zhen Jia^{1*} Boyi Li² Yan Wang² Boris Ivanovic² Yurong You²
 Lingjie Liu³ Yue Wang^{2,4} Marco Pavone^{2,5} Chen Feng¹ Yiming Li^{1,2†}

¹NYU ²NVIDIA ³University of Pennsylvania ⁴USC ⁵Stanford University

<https://ai4ce.github.io/EUVS-Benchmark>

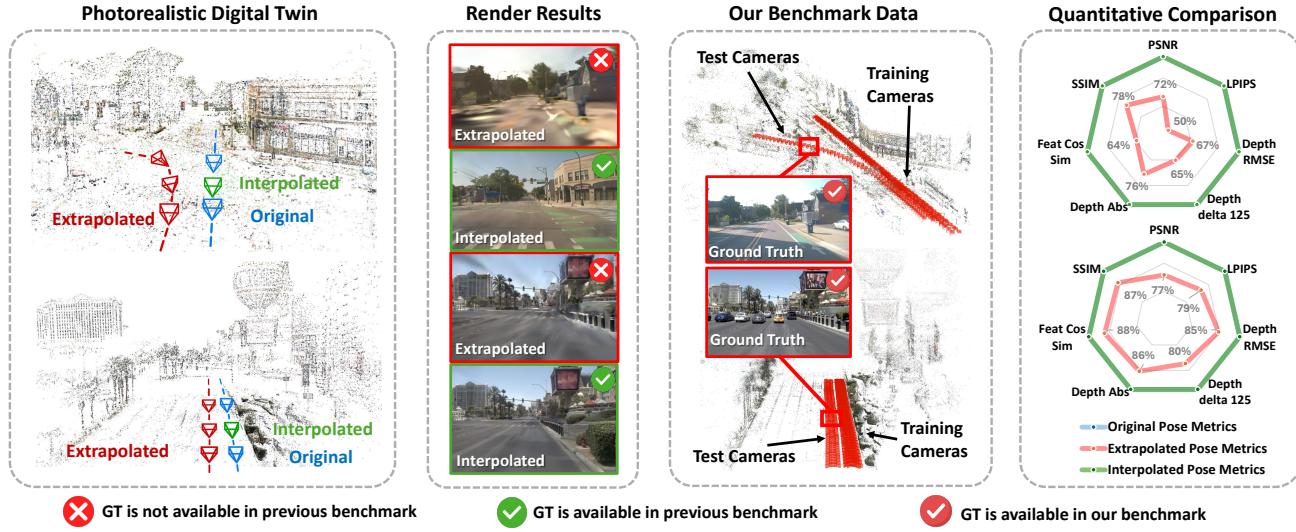


Figure 1. **Our key contributions.** Previous evaluations for urban view synthesis have primarily focused on interpolated poses, as the lack of ground truth data has made it challenging to evaluate extrapolated poses. We address this gap by providing real-world data that enables both quantitative and qualitative evaluations of extrapolated view synthesis in urban scenes. The quantitative results reveal a significant performance drop in 3D Gaussian Splatting [25] when handling extrapolated views, highlighting the need for more robust NVS methods.

Abstract

Photorealistic simulators are essential for the training and evaluation of vision-centric autonomous vehicles (AVs). At their core is Novel View Synthesis (NVS), a crucial capability that generates diverse unseen viewpoints to accommodate the broad and continuous pose distribution of AVs. Recent advances in radiance fields, such as 3D Gaussian Splatting, achieve photorealistic rendering at real-time speeds and have been widely used in modeling large-scale driving scenes. However, their performance is commonly evaluated using an interpolated setup with highly correlated training and test views. In contrast, extrapolation, where test views largely deviate from training views, remains underexplored, limiting progress in generalizable simulation technology. To address this gap, we lever-

age publicly available AV datasets with multiple traversals, multiple vehicles, and multiple cameras to build the first Extrapolated Urban View Synthesis (EUVS) benchmark. Meanwhile, we conduct quantitative and qualitative evaluations of state-of-the-art Gaussian Splatting methods across different difficulty levels. Our results show that Gaussian Splatting is prone to overfitting to training views. Besides, incorporating diffusion priors and improving geometry cannot fundamentally improve NVS under large view changes, highlighting the need for more robust approaches and large-scale training. We have released our data to help advance self-driving and urban robotics simulation technology.

1. Introduction

The development of vision-centric autonomous vehicles (AVs) relies heavily on photorealistic simulators, which provide controlled, reproducible, and scalable environments

*Equal contribution.

†Corresponding author.

for training and evaluation of driving models [14, 43, 57]. These simulators enable AVs to learn and adapt to a variety of real-world scenarios, from crowded urban streets to adverse weather conditions, without the logistical and safety concerns of physical road testing. At the heart of these simulators is the capability for Novel View Synthesis (NVS)—a key technology that generates realistic images of unseen viewpoints, simulating the continuous changes in perspective that occur as AVs navigate through urban environments.

Recent advancements in radiance fields, particularly methods based on 3D Gaussian Splatting [25], have significantly improved the realism and efficiency of NVS. These approaches [8, 53, 65, 67] can produce photorealistic renderings at real-time speeds, making them highly attractive for large-scale driving scene simulation. However, despite their impressive results, the evaluation of NVS methods has predominantly focused on **interpolated** scenarios, where training and test viewpoints are closely related. While interpolation tests are valuable for assessing local consistency, they fall short in addressing the more critical challenge of **extrapolation**—where test viewpoints differ significantly from the training data. As shown in Figure 1, the interpolation test set demonstrates strong performance, with metrics such as PSNR, SSIM, and LPIPS remaining very close to the training set values. In contrast, the extrapolation test set, which includes additional translation and rotation changes relative to the training set, exhibits notable drops in performance. Specifically, the metric decreases relative to the training set are **28%** for PSNR, **22%** for SSIM, and **50%** for LPIPS. These results underscore the urgent need to explore and advance extrapolated view synthesis in complex urban scenes, as real-world driving often involves encountering scenarios with significant viewpoint shifts and diverse spatial transformations that deviate from training distributions. Several recent studies [21, 23] have investigated the generalization capabilities of NVS in 3D Gaussian Splatting. Although they show promising qualitative results, there is no comprehensive quantitative analysis due to the absence of standardized datasets. Moreover, their evaluations are primarily limited to specific scenarios or use cases, without investigating varying levels of difficulty based on the degree of extrapolation. This gap underscores the urgent need for a benchmark that offers diverse and challenging datasets, enabling a rigorous and systematic assessment of NVS methods.

To establish a common platform for assessing the robustness of NVS methods, we introduce a comprehensive benchmark for quantitatively and qualitatively evaluating extrapolated novel view synthesis in large-scale urban scenes. Our benchmark leverages publicly available datasets, including NuPlan [4], MARS [30], and Argoverse2 [45], which feature multi-traversal, multi-agent and multi-camera sensory recordings. Multi-traversal data con-

sists of asynchronous traversals of the same location, while multi-agent data captures multiple vehicles simultaneously present in the same area. These data provide diverse camera poses within a 3D scene, enabling the training and evaluation of extrapolated view synthesis in outdoor environments. For the experimental setup, we define three difficulty levels: (1) translation only, (2) rotation only, and (3) translation + rotation, as shown in Figure 4. In autonomous driving scenarios, Level 1 corresponds to maneuvers such as lane changes, Level 2 involves switching between cameras facing different directions, and Level 3 addresses complex intersections, such as crossroads with diverse traversal paths. These levels represent common challenges in autonomous driving, and addressing them enables the synthesis of complete scenes from sparse image observations.

We conduct pose estimation and sparse reconstruction using COLMAP [38], which facilitates the initialization of Gaussian Splatting. We then evaluate state-of-the-art Gaussian Splatting-based approaches across each difficulty level, identifying performance gaps both qualitatively and quantitatively in extrapolated urban view synthesis.

In summary, our main contributions are as follows:

- We initiate the first comprehensive quantitative and qualitative study on the Extrapolated Urban View Synthesis (EUVS) problem, supported by a robust evaluation framework that categorizes difficulty levels (translation-only, rotation-only, and translation + rotation), while assessing performance using diverse metrics including reconstruction accuracy and visual fidelity.
- We construct a novel dataset by integrating multi-traversal, multi-agent, and multi-camera data from publicly available resources, totaling **90,810** frames across **345** videos. Our dataset effectively addresses the limitations of existing benchmarks, enabling rigorous and robust evaluation for extrapolated urban view synthesis.
- We benchmark state-of-the-art Gaussian Splatting-based and NeRF-based models and analyze key factors that influence the performance of extrapolated NVS, laying a solid foundation for future advancements in this challenging task. Data and code are released on our project page.

2. Related Works

Extrapolated View Synthesis. Extrapolated view synthesis aims to generate novel views beyond observed perspectives, addressing challenges in visual coherence for unseen regions. RapNeRF [61] proposes a random ray-casting policy that enables training on unseen views based on visible ones. Following work [54] enhances this approach by incorporating holistic priors. Additionally, numerous generalizable models [6, 9, 44] have emerged, capable of generating extrapolated novel views from a limited number of input images. While these methods are designed for indoor scenes, several works address extrapolated view synthesis in out-

door driving scenarios, which typically involve forward-facing cameras and unbounded environments. To tackle the Level 1 challenge in our benchmark and address the scarcity of lane change data, GGS [21] introduces a novel virtual lane generation module. In parallel, AutoSplat [26] tackles lane change in dynamic scenes by applying geometric and reflected consistency constraints. To address the Level 1 or Level 2 challenge, FreeSim [16], VEGS [23], and SGD [59] enhance 3DGS [25] with diffusion model priors to improve model generalization. *Yet existing methods suffer from two major limitations: (1) a lack of real data for quantitative evaluation, which confines them to qualitative analysis, and (2) a narrow focus on a specific level in our benchmark, preventing a comprehensive and systematic exploration.*

3D Gaussian Splatting. Recent advances in radiance fields, particularly NeRF [34] and 3DGS [25], have garnered significant attention due to their impressive advancements in NVS. NeRF employs an implicit representation through a multi-layer perceptron (MLP). Furthermore, 3DGS explicitly represents scenes using anisotropic 3D Gaussian ellipsoids, enabling high-quality real-time rendering. Several works have addressed issues such as difficulties with reflective surfaces [24], aliasing [58], etc. However, urban scenes introduce unique challenges due to their unbounded and dynamic nature. To address the challenge, several works separate dynamic and static elements in the scene by leveraging a composite dynamic Gaussian graph [53, 67], optical flow prediction [55, 65], etc. PVG [8] presents a unified representation model that simultaneously incorporates both dynamic and static components without relying on priors. To achieve realistic geometry and efficient rendering, 2DGS [22] collapses 3D Gaussians onto 2D planes, while hybrid approaches [40, 47] combine different Gaussians to better capture region-specific features. *In summary, current urban NVS methods primarily focus on effectively handling dynamic elements and enhancing geometry representation, while the challenge of extrapolated view synthesis remains largely underexplored.*

Autonomous Driving Simulators. Current simulators focus on three key challenges: parameter initialization [17, 41], traffic simulation [28, 51, 64], and sensor simulation [1, 14, 19, 27, 57, 63]. Sensor simulation is crucial for generating realistic sensory data that AVs depend on for perception and decision-making. Early sensor simulators [14, 39, 49] provide simulated environments that are valuable for research but lack visual realism. Recent studies have focused on data-driven simulators that extract data from real-world driving logs, creating more realistic and adaptable environments. These methods can be classified into two categories: generation-based and reconstruction-based approaches. The former rely on inputs such as text, video, and other data sources for simulation, supported by world models or prior knowledge [19, 27, 63].

Reconstruction-based simulations leverage real-world data to ensure both visual fidelity and geometric consistency [46, 48, 57]. UniSim [57] is a pioneering example of this approach, utilizing NeRF-based scene representation to create dynamic scenes with geometric information that are both editable and controllable. *Extrapolated view synthesis is essential for these simulators, as it enables the generation of realistic and consistent views from diverse angles.*

Autonomous Driving Datasets. High-quality datasets play a vital role in advancing autonomous driving research. The KITTI dataset [20], released in 2012, marked a major milestone, significantly accelerating advances in AVs [18, 33, 42]. Since then, many influential autonomous vehicle datasets have been developed to tackle challenges like adverse weather conditions [37], multimodal fusion [3, 4], repeated driving [4, 13], collaborative driving [29, 30, 52], and motion prediction [5, 15, 45], etc. We leverage publicly available datasets with multi-traversal, multi-agent, and multi-camera recordings, enabling a comprehensive and robust evaluation of extrapolated urban view synthesis.

3. The EUVS Benchmark

3.1. Data Source

We employ three publicly available autonomous driving datasets, nuPlan [4], Argoverse 2 [45], and MARS [30], to comprehensively evaluate extrapolated view synthesis in urban driving scenes while leveraging their unique characteristics to ensure robust and diverse assessments. The nuPlan [4] dataset is the first large-scale planning benchmark, providing 1,200 hours of driving data collected from four cities across the United States and Asia. Argoverse 2 [45] focuses on multimodal perception and forecasting, providing 1,000 annotated 3D scenarios with lidar, stereo imagery, and HD maps. It also includes 20,000 unlabeled lidar sequences for self-supervised learning and 250,000 motion forecasting scenarios highlighting complex interactions in six U.S. cities. The MARS [30] dataset introduces multi-agent and multi-traversal scenarios, supporting collaborative driving research with vehicles interacting within the same area and asynchronous revisits to the same locations. By integrating these datasets, our benchmark enables the evaluation of view synthesis across diverse and realistic urban environments under varying conditions. Figure 3 illustrates the distribution of the integrated datasets.

3.2. Evaluation Framework

To systematically assess model performance in extrapolated urban view synthesis, our evaluation framework incorporates *three difficulty levels* and *three data configurations*. Data configurations include multi-traversal, multi-agent, and multi-camera, while difficulty levels are categorized into (1) translation only, (2) rotation only, and (3)



Figure 2. Dataset visualization. Our dataset features diverse scenes across various locations in different cities, sourced from multiple datasets. Typical driving scenarios include maneuvers such as lane changes, cross intersections, and T-junctions. **Top:** Each column displays images captured at the same location by different agents or traversals. **Bottom:** Each image displays the COLMAP points at a specific location, along with the corresponding camera poses.

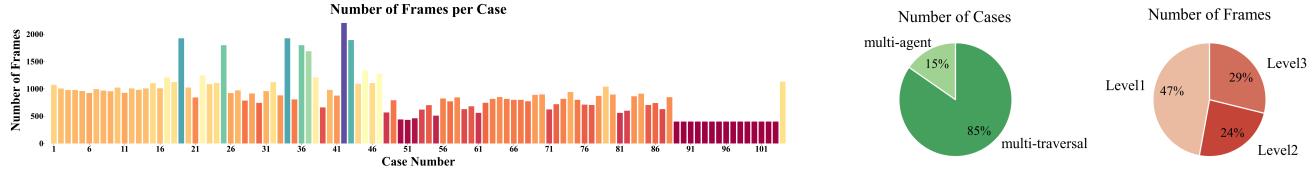


Figure 3. Dataset distribution. Our dataset comprises **90,810** frames distributed over **104** cases, capturing a diverse array of multi-traversal paths, multi-agent interactions, and multi-camera perspectives across varying difficulty levels.

translation plus rotation, as illustrated in [Figure 4](#).

Level 1. The translation-only experimental setup involves scenarios where the vehicle's position shifts without any change in orientation. This scenario is commonly observed in lane changes. We use traversals from different lanes in multi-traversal data, focusing on the three front cameras. The data is sourced from nuPlan [4] and Argoverse 2 [45].

Level 2. The second level, rotation only, evaluates models on views that differ significantly in orientation. In vision-centric autonomous vehicles, this experimental setting involves transitions between views captured from different directions. We leverage multi-camera data from nuPlan [4], training on three forward-facing and three rear-facing cameras to capture diverse perspectives, and evaluating using two side-facing cameras for comprehensive coverage.

Level 3. The third level, translation plus rotation, incorpo-

rates both positional shifts and orientational deviations, presenting the most challenging scenario for NVS. To address this, we utilize multi-traversal driving data collected from the same location but across different traversal routes. For example, the training and test sets may include routes that approach an intersection from different directions. Typical route combinations feature scenarios such as intersections, T-junctions, and Y-junctions, as shown in [Figure 2](#), ensuring diversity and comprehensive evaluation. The data for this level is from MARS [30] and Argoverse 2 [45].

3.3. Algorithm Overview

Vanilla 3D Gaussian Splatting. 3D Gaussian Splatting (3DGS) [25] leverage 3D Gaussians to explicitly represent the scene, which achieves high quality while offering real-time rendering by avoiding unnecessary computation in the

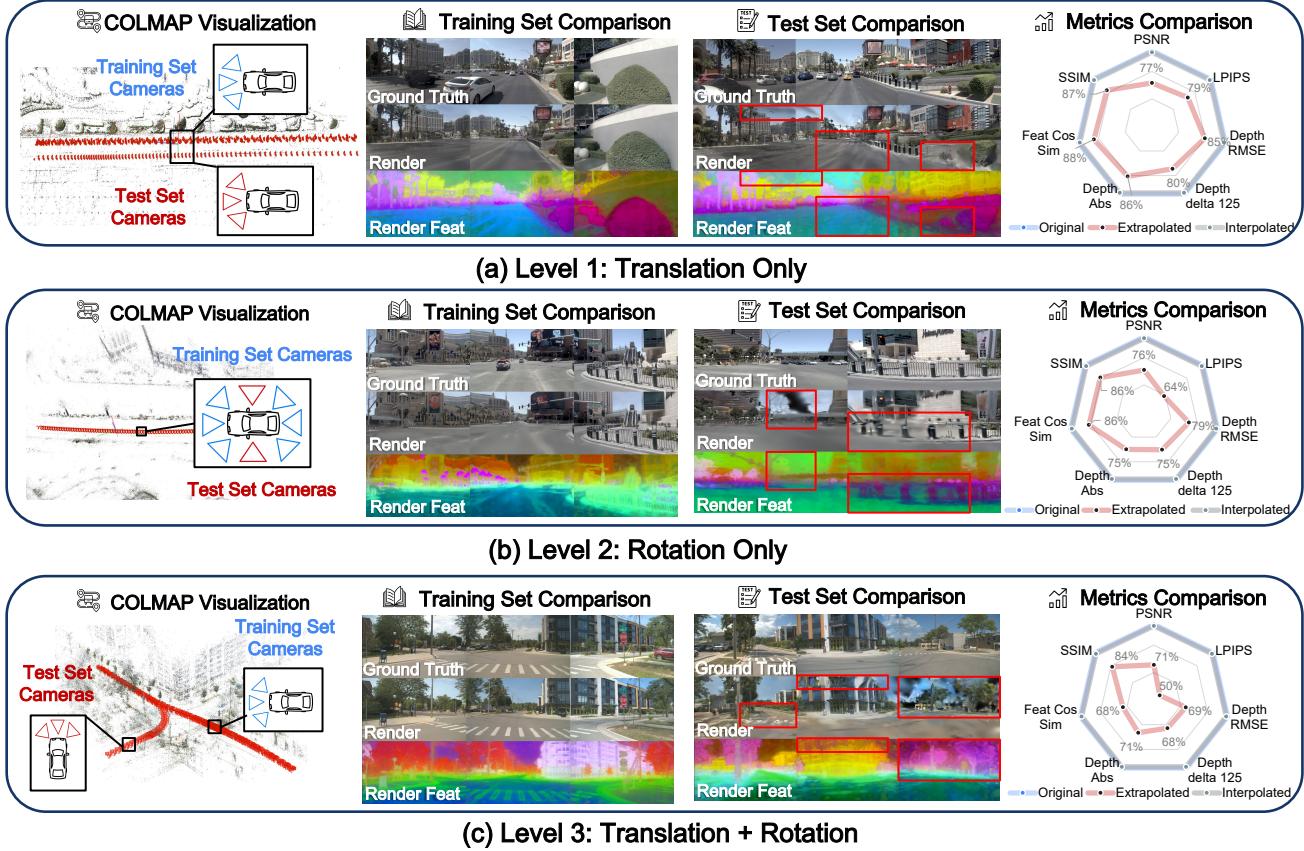


Figure 4. **Qualitative and quantitative results across three difficulty levels.** The results show a clear degradation in performance as the difficulty level increases, highlighting the challenge of maintaining consistency and realism in complex urban scenarios.

empty space. Building on this, 3DGM [31] leverages multi-traversal consensus to differentiate transient and permanent elements, enabling joint 2D segmentation and 3D mapping without using any human supervision.

Planar-based and Geometry Refined Gaussian Splatting. GaussianPro [11] builds on 3DGS [25] by introducing multi-frame geometric optimization, which guides the densification of 3D Gaussians, enhancing scene consistency in complex geometries. It further refines geometry by encouraging Gaussian primitives to adopt flat structures. Similarly, 2DGS [22] projects the 3D volume into a set of 2D oriented planar Gaussian disks, enabling high-fidelity surface reconstruction. PGSR [7] introduces an unbiased depth rendering method and integrates single-view geometric, multi-view photometric, and geometric regularization techniques to improve global geometry accuracy.

Gaussian Splatting with Diffusion Priors. VEGS [23] introduces a novel view generalization approach that harnesses pre-extracted surface normals to align 3D Gaussians while generating augmented camera views guided by diffusion priors. These diffusion priors serve a dual purpose: providing denoising loss guidance and supervising the training of augmented cameras. This process effectively mitigates floating artifacts and fragmented geometries, resulting in more accurate and coherent 3D representations.

gates floating artifacts and fragmented geometries, resulting in more accurate and coherent 3D representations.

Feature-Enhanced Gaussian Splatting. Feature 3DGS [66] extends 3D Gaussian Splatting with a Parallel N-dimensional Gaussian Rasterizer, allowing simultaneous rendering of radiance fields and high-dimensional semantic features. By embedding semantic features directly into 3D Gaussians, the approach enhances optimization, enabling better correspondence with scene semantics and achieving more detailed and accurate spatial representations.

NeRF-based Method. Instant-NGP [35] uses a multiresolution hash encoding to map spatial coordinates into compact latent representations via hash tables across multiple resolutions. This approach efficiently encodes high-frequency details by combining trainable feature vectors with interpolation, enabling adaptive and scalable input encodings without the need for structural updates or explicit collision handling. Zip-NeRF [2] leverages multisampling with isotropic Gaussians for scale-aware features and introduces a smooth anti-aliasing loss to address z-aliasing. In addition, it incorporates a novel distance normalization technique to better manage close and distant objects, achieving high-quality rendering and fast training.

Table 1. **Quantitative rendering results across three difficulty levels.** The results reveal significant performance drops between the training and test sets, highlighting the challenges of extrapolated view synthesis. PSNR drops from 24.6% to 30.6%, SSIM from 12.8% to 19.5%, Feature Cosine Similarity from 11.2% to 35.9%, and LPIPS shows the largest decline, from 25.3% to 70.0%. The average drop becomes more pronounced as the difficulty of the task increases, indicating a clear trend of performance degradation.

Method	PSNR \uparrow			SSIM \uparrow			LPIPS \downarrow			Feat Cos Sim \uparrow			
	Train	Test	Drop	Train	Test	Drop	Train	Test	Drop	Train	Test	Drop	
Level 1	3DGS [25]	21.36	16.37	23.4%	0.8275	0.7203	13.0%	0.2041	0.2599	27.3%	0.6828	0.6039	11.6%
	3DGM [31]	20.96	16.35	22.0%	0.8293	0.7248	12.6%	0.2003	0.2542	26.9%	0.6802	0.6087	10.5%
	GSPro [11]	21.51	16.39	23.8%	0.8310	0.7189	13.5%	0.1804	0.2450	35.8%	0.7081	0.6130	13.4%
	VEGS [23]	21.26	15.88	25.3%	0.8107	0.7047	13.1%	0.2498	0.3062	22.6%	0.6323	0.5521	12.7%
	PGSR [7]	20.57	16.32	20.7%	0.8104	0.7102	12.4%	0.2262	0.2733	20.8%	0.6515	0.5848	10.2%
	2DGS [22]	20.87	16.30	21.9%	0.8076	0.7103	12.0%	0.2438	0.2890	18.5%	0.6256	0.5644	9.8%
	Feature 3DGS [66]	21.02	16.01	23.8%	0.8096	0.7243	10.5%	0.1876	0.2575	37.3%	0.6958	0.6122	12.0%
	Zip-NeRF [2]	19.68	14.06	28.6%	0.7856	0.6917	12.0%	0.2711	0.3418	26.1%	0.6318	0.5542	12.3%
Level 2	Instant-NGP [35]	18.77	12.65	32.6%	0.7631	0.6252	18.1%	0.4874	0.5938	21.8%	0.5465	0.4837	11.5%
	AVERAGE	20.67	15.59	24.6%	0.8083	0.7046	12.8%	0.2501	0.3134	25.3%	0.6505	0.5774	11.2%
	3DGS [25]	25.75	19.53	24.2%	0.8766	0.7511	14.3%	0.1536	0.2668	73.7%	0.7327	0.6319	13.8%
	3DGM [31]	25.75	18.78	27.1%	0.8786	0.7464	15.0%	0.1556	0.2813	80.8%	0.7278	0.6344	12.8%
	GSPro [11]	26.42	19.39	26.6%	0.8821	0.7470	15.3%	0.1329	0.2246	69.0%	0.7523	0.6487	13.8%
	VEGS [23]	24.54	23.33	4.9%	0.8366	0.7949	5.0%	0.2301	0.2811	22.2%	0.6595	0.6133	7.0%
	PGSR [7]	24.53	18.38	25.1%	0.8612	0.7119	17.3%	0.1555	0.2532	62.8%	0.7200	0.5817	19.2%
	2DGS [22]	25.15	18.83	25.1%	0.8578	0.7204	16.0%	0.1756	0.2917	66.1%	0.6898	0.5785	16.1%
Level 3	Feature 3DGS [66]	24.91	19.59	21.4%	0.8800	0.7864	10.6%	0.1377	0.2278	65.4%	0.7427	0.6464	13.0%
	Zip-NeRF [2]	29.06	17.36	40.3%	0.8660	0.6715	22.5%	0.2078	0.3582	72.4%	0.7479	0.5843	21.9%
	Instant-NGP [35]	25.61	17.15	33.0%	0.8596	0.7212	16.1%	0.3340	0.5171	54.8%	0.7254	0.6182	14.8%
	AVERAGE	25.75	19.15	25.6%	0.8665	0.7390	14.7%	0.1870	0.3002	62.5%	0.7220	0.6153	14.8%
	3DGS [25]	21.22	14.99	29.4%	0.8550	0.7169	16.1%	0.2252	0.4050	79.8%	0.7002	0.4774	31.8%
	3DGM [31]	20.62	14.60	29.2%	0.8543	0.7233	15.3%	0.2254	0.4049	79.6%	0.6874	0.4672	32.0%
	GSPro [11]	21.58	14.82	31.3%	0.8634	0.6996	19.0%	0.2010	0.3877	92.9%	0.7093	0.4541	36.0%
	VEGS [23]	21.13	14.25	32.6%	0.8266	0.6475	21.7%	0.2359	0.4422	87.5%	0.6785	0.4442	34.5%
Level 3	PGSR [7]	19.60	14.17	27.7%	0.8238	0.6984	15.2%	0.2934	0.4363	48.7%	0.5867	0.3787	35.5%
	2DGS [22]	17.35	11.36	34.5%	0.7568	0.5447	28.0%	0.4296	0.5459	27.1%	0.3552	0.2327	34.5%
	Feature 3DGS [66]	21.88	14.33	34.5%	0.8643	0.6386	26.1%	0.1400	0.3816	172.6%	0.7411	0.4669	37.0%
	Zip-NeRF [2]	20.61	14.42	30.0%	0.8383	0.6565	21.7%	0.2197	0.4546	106.9%	0.7108	0.3645	48.7%
	Instant-NGP [35]	19.63	14.39	26.7%	0.8179	0.7104	13.1%	0.4956	0.6592	33.0%	0.6083	0.4157	31.7%
	AVERAGE	20.40	14.15	30.6%	0.8334	0.6707	19.5%	0.2740	0.4575	70.0%	0.6419	0.4113	35.9%

Table 2. **Quantitative comparison of extrapolated depth estimation at Level 1.** 3DGM [31] demonstrates superior performance on most evaluation metrics, while VEGS [23] and GSPro [11] excel in SqRel and Delta1, respectively.

Baseline	AbsRel \downarrow	RMSE \downarrow	SqRel \downarrow	Delta1 \uparrow	Delta2 \uparrow	Delta3 \uparrow
3DGS [25]	0.361	14.44	10.41	0.649	0.824	0.895
3DGM [31]	0.301	13.93	8.906	0.651	0.846	0.915
PGSR [7]	0.366	17.57	21.50	0.759	0.834	0.883
GSPro [11]	0.355	19.66	32.01	0.643	0.839	0.909
VEGS [23]	0.368	15.00	8.398	0.441	0.691	0.827

4. Experiment

4.1. Experiment Setup

Implementation Details. We initialize 3DGS-based methods with COLMAP [38]. We use SegFormer [50] to mask out movable objects during training and evaluation. We also exclude movable objects during 3DGS initialization.

Evaluation Metrics. We use three widely-used metrics to evaluate visual quality: peak signal-to-noise ratio (PSNR),

Table 3. **Comparison between interpolated and extrapolated view synthesis results across three difficulty levels of 3DGS [25].** There is a notable performance drop from interpolated to extrapolated results across all levels.

Level	PSNR \uparrow		SSIM \uparrow		LPIPS \downarrow		Feat Cos Sim \uparrow	
	In.	Ex.	In.	Ex.	In.	Ex.	In.	Ex.
Level 1	19.61	15.57	0.80	0.65	0.21	0.29	0.66	0.55
Level 2	19.16	17.23	0.79	0.68	0.21	0.31	0.67	0.58
Level 3	22.89	16.34	0.88	0.78	0.19	0.32	0.73	0.57

structural similarity index measure (SSIM), and learned perceptual image patch similarity (LPIPS) [62]. We also employ DINOv2 [36] feature cosine similarity to evaluate image quality in latent space. For geometry evaluation, we use depth metrics, including RMSE and $\delta_{1.25}$.

4.2. Experimental Results

Table 1 and 3 present the quantitative results across Levels 1-3, while Figure 5 illustrates the qualitative outcomes on

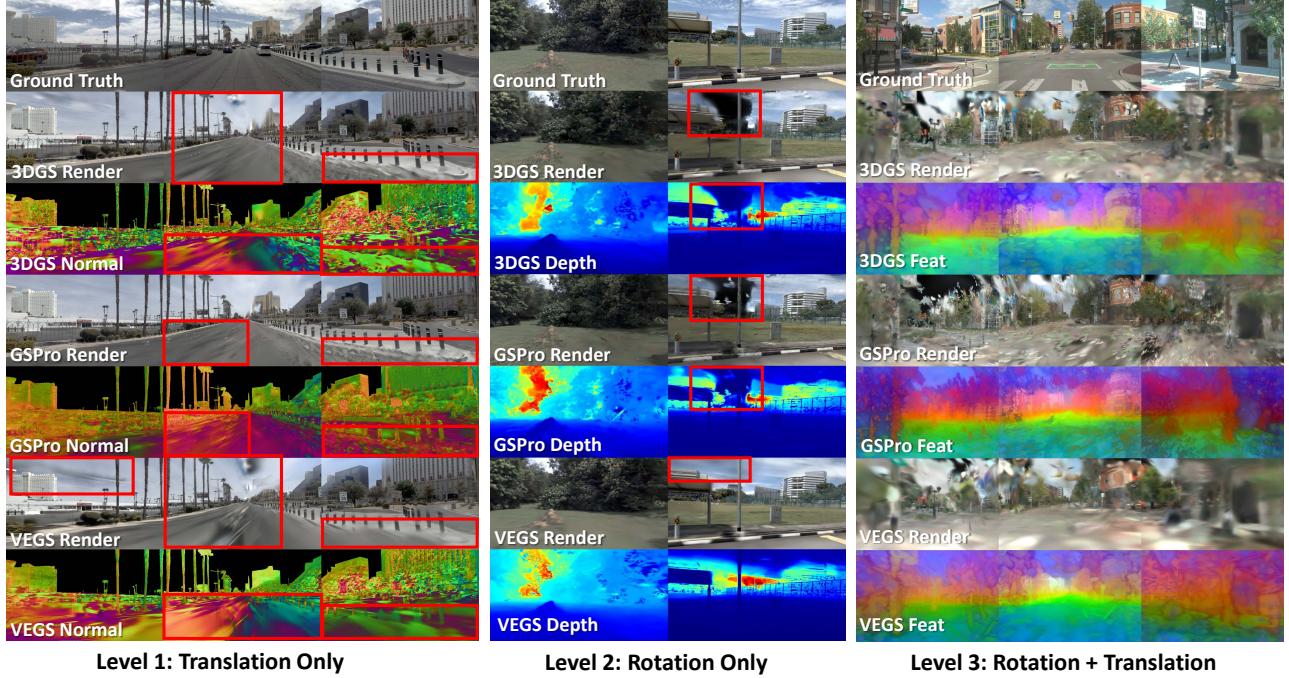


Figure 5. **Qualitative comparison of extrapolated view synthesis across different levels of camera movement (translation, rotation, and combined).** For each level, results from different methods (3DGS [25], GSPro [11], and VEGS [23]) are compared against the ground truth. Red boxes highlight areas where methods are limited in capturing fine details, such as road surfaces, sky regions, or object boundaries, demonstrating the challenges faced by each approach under varying movement complexities.

the extrapolated test set. The results indicate that, while the metrics perform relatively well in the training set and interpolated test set, there is a significant drop in performance in the extrapolated test set across all baselines. We detail the rendering results of each level in the following.

Level 1: Translation-only. At Level 1, the views from the training cameras fully cover the test views, with moderate translational changes. (1) The results reveal a consistent drop in train-to-test performance across all metrics, underscoring the challenge of generalizing to unseen views. Notably, the relative performance drop varies by metric: for instance, PSNR decreases by approximately 23–25% across most methods (e.g., GSPro [11]: $21.51 \rightarrow 16.39$, a 23.8% drop), with SSIM and LPIPS showing similar proportional declines. (2) On the test set, the methods perform comparably: GSPro achieves the highest PSNR (16.39) and lowest LPIPS (0.2450), while 3DGM [31] leads in SSIM (0.7248). Both GSPro and 3DGM deliver similar feature cosine similarity scores (GSPro: 0.6130, 3DGM: 0.6087). Other methods, such as VEGS [23] and PGSR [7], lag slightly, indicating opportunities for improvement. The results suggest that while GSPro marginally outperforms the others, the differences are small, emphasizing the need for better solutions.

Level 2: Rotation-only. At Level 2, the training views provide substantial coverage of the surrounding scene. However, most methods exhibit poor generalization ability, with

a PSNR drop of approximately 22.75%. Rotation variations are particularly challenging for areas rich in texture, such as trees and intricate elements, where blurring artifacts often appear. Regions perpendicular to the vehicle are also difficult to capture. Furthermore, distant regions pose significant reconstruction challenges, frequently resulting in missing buildings and black holes in the sky. VEGS [23] and GSPro [11] stand out as the best-performing baselines at this level: VEGS benefits from its diffusion prior, while GSPro’s robust geometry handling enhances generalization.

Level 3: Translation + Rotation. At Level 3, the view changes are the largest. (1) All methods exhibit notable train-to-test performance drops across all metrics, reflecting the difficulty of generalizing to Level 3. For instance, 3DGS [25] experiences a PSNR drop of 29.4% ($21.22 \rightarrow 14.99$), while GSPro [11] undergoes a similar drop of 31.3% ($21.58 \rightarrow 14.82$). These declines underscore the increasing challenge of handling extrapolated views as complexity rises. (2) Feature 3DGS [66] and 3DGM [31] stand out as leading methods, excelling in LPIPS (0.3816) and SSIM (0.7233), respectively, and demonstrating notable improvements compared to other methods. However, overall performance remains limited, with PSNR consistently falling below 15, highlighting significant room for improvement in generating high-fidelity outputs for extrapolated views.

4.3. Comparison of Baselines

Quantitative Comparison. We report the quantitative performance comparison across all levels and baselines in [Figure I](#). (1) At Level 1, as shown in [Figure Ia](#), GSPro [11] slightly outperforms other baselines on the training set. However, the performance gaps on the test set are small, with most baselines performing comparably poorly. Among them, 3DGS [25], 3DGM [31], and GSPro achieve relatively better results. (2) At Level 2 ([Figure Ib](#)), Zip-NeRF [2] demonstrates over a 10% improvement in PSNR compared to all other baselines on the training set. In extrapolated poses, VEGS [23] significantly outperforms all other methods, achieving at least 20% higher PSNR. These results highlight the effectiveness of diffusion priors in rotation-only settings. (3) At Level 3, as shown in [Figure Ic](#), none of the baselines exhibit a clear advantage, as all methods fail equally in this challenging setting. Among them, GSPro maintains a slight lead on the training set. On the 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 levels and baselines in [Figure II](#), [Figure III](#) and [Figure IV](#). (1) At Level 1, as shown in [Figure IIb](#), all methods exhibit imperfections in ground rendering, while planar-based methods such as 2DGs [22] and PGSR [7] show comparatively fewer flaws on the ground surface. GSPro [11] produces more accurate geometry reconstruction, achieving realistic surfaces and high-fidelity representations of street objects like trees and buildings. (2) At Level 2, as shown in [Figure III](#), most baselines suffer from sky artifacts such as holes and floating objects. In contrast, VEGS [23] produces the more accurate renderings, exhibiting minimal floating artifacts and broken geometry, attributed to the guidance provided by diffusion priors. (3) At Level 3, as shown in [Figure IVb](#), 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 urban scenes.

4.4. Discussion and Analysis

Planar-Based vs. Ellipsoid-Based. Planar-based methods (e.g., GSPro [11], PGSR [7], and 2DGs [22]) 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 [25] and 3DGM [31]) better handle high-textured objects but often overfit, leading to errors in road

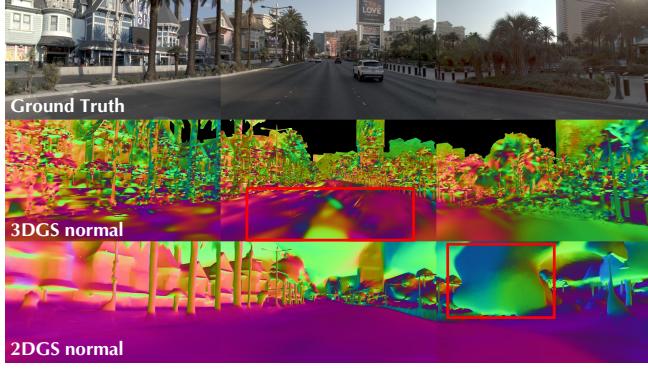
representation. For instance, in the translation setting ([Figure 6a](#)), planar-based methods struggle with plants, while ellipsoid-based methods perform poorly on roads. A hybrid representation could effectively combine the strengths of both approaches to address these challenges in EUVS.

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 6b](#), 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, as shown in [Table 1](#), VEGS [23] with diffusion priors significantly outperforms 3DGS [25] 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 [56], has proven to be an effective approach for enhancing training regularization [12]. 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 6c](#), 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 [2], excel at handling low-texture regions like roads and sky in extrapolated views, due to their ability to ensure coherent depth and color transitions across large, gradual surfaces. However, NeRF-based methods struggle with capturing high-frequency or fine-grained details, such as lane markings and fences, due to the inherent limitations of their implicit representation. In contrast, the explicit representation of Gaussian-based methods demonstrates a distinct advantage in preserving detail and sharpness in high-frequency regions, as illustrated in [Figure 6d](#).

Dynamic Scenes. Current methods that focus on dynamic scenes also struggle with extrapolated view synthesis. To evaluate extrapolated view synthesis in dynamic scenes, we



(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.



(e) Dynamic scenes qualitative comparison.



(f) With vs. without lighting handling.

Figure 6. Qualitative comparison of different techniques. The various techniques show some ability to partially address the challenges but fail to fundamentally resolve the underlying issues. Additionally, a lack of generalization persists in dynamic scene baselines. Incorporating separate intrinsic and dynamic appearance features emerges as a promising solution to effectively mitigate lighting inconsistencies.

implement OmniRe [10] as a baseline, training on six front and back cameras and testing on two side cameras. OmniRe organizes rigid-deformable nodes and background nodes to capture dynamic scene structures and employs SMPL [32] for non-rigid object modeling. As illustrated in Figure 6e, the rendering results reveal a significant performance gap between the training and test sets. In the test views, objects such as trees and stakes lose texture and geometric details, resulting in noticeably blurry outputs, whereas the training views maintain high fidelity. As shown in Table 4, the performance metrics indicate an average drop of 25%. The results highlight the challenges of extrapolated view synthesis in dynamic scenes and the need for further research.

Lighting Inconsistency Handling. Lighting inconsistencies between training and test traversals, driven by varia-

tions in lighting and weather conditions, pose significant challenges for multi-traversed datasets. To address the impact of lighting variations, we introduce the Gaussian in the Wild (GS-W) [60] baseline. GS-W abandons the conventional color modeling approach that relies on spherical harmonic coefficients and instead introduces separated intrinsic properties for each Gaussian point and dynamic appearance features for each image. An object’s intrinsic appearance is determined by its material and surface properties, while its dynamic appearance is influenced by environmental factors such as highlights and shadows. This novel approach captures the unchanged scene appearance while accommodating dynamic variations like illumination and weather changes. We extract dynamic appearance features from both training and test images to ensure fairness and

Table 4. Quantitative performance of OmniRe [10] at Level 2. The experiment uses a single traversal with dynamic objects, showing a noticeable drop from train to test.

Level	PSNR \uparrow		SSIM \uparrow		LPIPS \downarrow		Feat Cos Sim \uparrow	
	Train	Test	Train	Test	Train	Test	Train	Test
Level 2	19.78	15.32	0.65	0.45	0.38	0.53	0.73	0.58

Table 5. Quantitative performance of GS-W [60] at each level. After learning the lighting features, the train and test metrics show significant improvement compared to other baselines. However, there is still a considerable drop from train to test.

Level	PSNR \uparrow		SSIM \uparrow		LPIPS \downarrow		Feat Cos Sim \uparrow	
	Train	Test	Train	Test	Train	Test	Train	Test
Level 1	28.15	20.22	0.89	0.78	0.15	0.23	0.71	0.64
Level 2	30.10	21.21	0.91	0.82	0.13	0.20	0.76	0.67
Level 3	28.62	19.36	0.87	0.73	0.15	0.31	0.74	0.50

consistency. The results of GS-W are presented in Table 5 and Figure 6f. It can be observed that even after eliminating the influence of dynamic appearance inconsistencies, there is still a significant performance drop, particularly at Level 3. The performance drop is primarily due to the Gaussian-based model’s limited generalization capability, underscoring its inherent challenges in adapting to extrapolated and unseen scenarios.

Performance Gains from Multi-Traversal Data. Multi-traversal data plays a critical role in Extrapolated View Synthesis. Using the GaussianPro model [11] at Level 1, we progressively increase the number of training traversals to observe its impact. The results, shown in Figure 8 and Figure 7, 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.

5. Conclusion

We introduce the first Extrapolated Urban View Synthesis (EUVS) benchmark to advance photorealistic simulation technologies for self-driving and robotics. Our benchmark integrates real-world multi-traversal, multi-agent, and multi-camera data, categorizes scenes by difficulty, and evaluates state-of-the-art NVS models. Experimental results reveal that while some methods address specific challenges, current models demonstrate limited generalization, with significant overfitting to training views and suboptimal performance in extrapolated view synthesis. Additionally, we identify key challenges across difficulty levels and high-

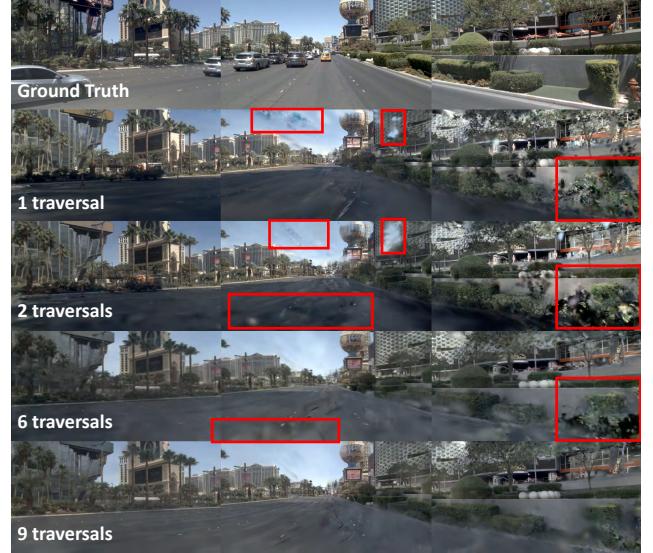


Figure 7. As the number of traversals increases, the performance of NVS improves. This is highlighted in the red box, where the texture progressively enriches and errors in areas like the sky and ground are reduced.

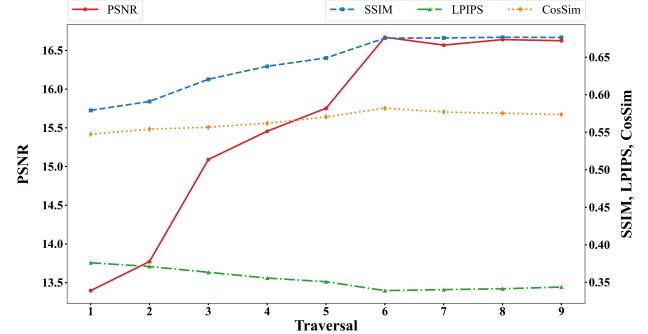


Figure 8. NVS performance vs. number of traversals. With more traversals, PSNR and SSIM exhibit notable improvements, indicating enhanced image quality and structural similarity. LPIPS values decrease, reflecting better perceptual consistency, while CosSim stabilizes after an initial rise. These results highlight the importance of more visual data for improving NVS performance.

light the pros and cons of different models, offering valuable insights for developing more generalizable NVS solutions. To support further research, we have released the dataset and benchmark, addressing the long-standing data scarcity and providing robust baseline evaluations. We believe the EUVS benchmark will catalyze meaningful advancements in self-driving and urban robotics innovation.

Acknowledgment. This work was supported in part through NSF grants 2238968 and 2121391, and the NYU IT High Performance Computing resources, services, and staff expertise. Yiming Li is supported by NVIDIA Graduate Fellowship (2024-2025).

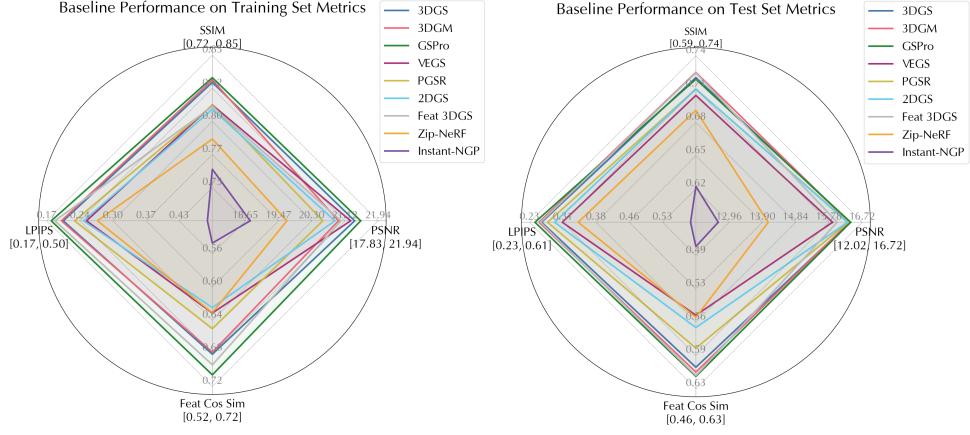
References

- [1] Alexander Amini, Tsun-Hsuan Wang, Igor Gilitschenski, Wilko Schwarting, Zhijian Liu, Song Han, Sertac Karaman, and Daniela Rus. Vista 2.0: An open, data-driven simulator for multimodal sensing and policy learning for autonomous vehicles. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2419–2426. IEEE, 2022. 3
- [2] 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. 5, 6, 8
- [3] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liang, Qiang Xu, Anush Krishnan, Yu Pan, Gi-ancarlo Baldan, and Oscar Beijbom. nuscenes: A multi-modal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 3
- [4] Holger Caesar, Juraj Kabzan, Kok Seang Tan, Whye Kit Fong, Eric Wolff, Alex Lang, Luke Fletcher, Oscar Beijbom, and Sammy Omari. nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles. *arXiv preprint arXiv:2106.11810*, 2021. 2, 3, 4
- [5] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, et al. Argoverse: 3d tracking and forecasting with rich maps. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8748–8757, 2019. 3
- [6] David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19457–19467, 2024. 2
- [7] 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. 5, 6, 7, 8
- [8] 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. 2, 3
- [9] 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. *arXiv preprint arXiv:2403.14627*, 2024. 2
- [10] Ziyu Chen, Jiawei Yang, Jiahui Huang, Riccardo de Lutio, Janick Martinez Esturo, Boris Ivanovic, Or Litany, Zan Gajcic, Sanja Fidler, Marco Pavone, et al. Omnisre: Omni urban scene reconstruction. *arXiv preprint arXiv:2408.16760*, 2024. 9, 10
- [11] 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. 5, 6, 7, 8, 10
- [12] 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. 8
- [13] Carlos A Diaz-Ruiz, Youya Xia, Yurong You, Jose Nino, Junnan Chen, Josephine Monica, Xiangyu Chen, Katie Luo, Yan Wang, Marc Emond, et al. Ithaca365: Dataset and driving perception under repeated and challenging weather conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21383–21392, 2022. 3
- [14] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017. 2, 3
- [15] Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R Qi, Yin Zhou, et al. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9710–9719, 2021. 3
- [16] Lue Fan, Hao Zhang, Qitai Wang, Hongsheng Li, and Zhaoxiang Zhang. Freesim: Toward free-viewpoint camera simulation in driving scenes, 2024. 3
- [17] Lan Feng, Quanyi Li, Zhenghao Peng, Shuhan Tan, and Bolei Zhou. Trafficgen: Learning to generate diverse and realistic traffic scenarios. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3567–3575. IEEE, 2023. 3
- [18] Jannik Fritsch, Tobias Kuehnl, and Andreas Geiger. A new performance measure and evaluation benchmark for road detection algorithms. In *16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013)*, pages 1693–1700. IEEE, 2013. 3
- [19] Shenyuan Gao, Jiazhui Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. *arXiv preprint arXiv:2405.17398*, 2024. 3
- [20] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 3
- [21] Huasong Han, Kaixuan Zhou, Xiaoxiao Long, Yusen Wang, and Chunxia Xiao. Ggs: Generalizable gaussian splatting for lane switching in autonomous driving. *arXiv preprint arXiv:2409.02382*, 2024. 2, 3
- [22] 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. 3, 5, 6, 8
- [23] 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. 2, 3, 5, 6, 7, 8

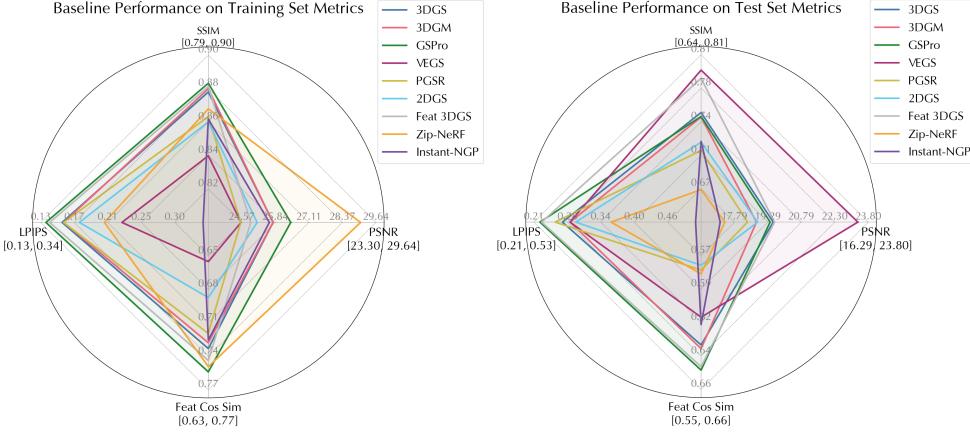
- [24] Yingwenqi Jiang, Jiadong Tu, Yuan Liu, Xifeng Gao, Xiaoxiao Long, Wenping Wang, and Yuexin Ma. Gaussian-shader: 3d gaussian splatting with shading functions for reflective surfaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5322–5332, 2024. 3
- [25] 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, 2, 3, 4, 5, 6, 7, 8
- [26] Mustafa Khan, Hamidreza Fazlali, Dhruv Sharma, Tongtong Cao, Dongfeng Bai, Yuan Ren, and Bingbing Liu. Autosplat: Constrained gaussian splatting for autonomous driving scene reconstruction. *arXiv preprint arXiv:2407.02598*, 2024. 3
- [27] Seung Wook Kim, Jonah Philion, Antonio Torralba, and Sanja Fidler. Drivegan: Towards a controllable high-quality neural simulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5820–5829, 2021. 3
- [28] Quanyi Li, Zhenghao Mark Peng, Lan Feng, Zhizheng Liu, Chenda Duan, Wenjie Mo, and Bolei Zhou. Scenarionet: Open-source platform for large-scale traffic scenario simulation and modeling. *Advances in neural information processing systems*, 36, 2024. 3
- [29] Yiming Li, Dekun Ma, Ziyan An, Zixun Wang, Yiqi Zhong, Siheng Chen, and Chen Feng. V2x-sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving. *IEEE Robotics and Automation Letters*, 7(4):10914–10921, 2022. 3
- [30] Yiming Li, Zhiheng Li, Nuo Chen, Moonjun Gong, Zonglin Lyu, Zehong Wang, Peili Jiang, and Chen Feng. Multiagent multitraversal multimodal self-driving: Open mars dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22041–22051, 2024. 2, 3, 4
- [31] 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. 5, 6, 7, 8
- [32] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 9
- [33] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3061–3070, 2015. 3
- [34] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 3
- [35] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, 2022. 5, 6
- [36] Maxime Oquab, Timothée Darcret, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 6
- [37] Matthew Pitropov, Danson Evan Garcia, Jason Rebello, Michael Smart, Carlos Wang, Krzysztof Czarnecki, and Steven Waslander. Canadian adverse driving conditions dataset. *The International Journal of Robotics Research*, 40(4-5):681–690, 2021. 3
- [38] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 6
- [39] Shital Shah, Debadatta Dey, Chris Lovett, and Ashish Kapoor. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and Service Robotics: Results of the 11th International Conference*, pages 621–635. Springer, 2018. 3
- [40] Xi Shi, Lingli Chen, Peng Wei, Xi Wu, Tian Jiang, Yonggang Luo, and Lecheng Xie. Dhgs: Decoupled hybrid gaussian splatting for driving scene. *arXiv preprint arXiv:2407.16600*, 2024. 3
- [41] Shuhan Tan, Kelvin Wong, Shenlong Wang, Sivabalan Manivasagam, Mengye Ren, and Raquel Urtasun. Scenegen: Learning to generate realistic traffic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 892–901, 2021. 3
- [42] Izzeddin Teeti, Valentina Musat, Salman Khan, Alexander Rast, Fabio Cuzzolin, and Andrew Bradley. Vision in adverse weather: Augmentation using cyclegans with various object detectors for robust perception in autonomous racing. *arXiv preprint arXiv:2201.03246*, v3, 2023. 3
- [43] Adam Tonderski, Carl Lindström, Georg Hess, William Ljungbergh, Lennart Svensson, and Christoffer Petersson. Neurad: Neural rendering for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14895–14904, 2024. 2
- [44] Christopher Wewer, Kevin Raj, Eddy Ilg, Bernt Schiele, and Jan Eric Lenssen. latentsplat: Autoencoding variational gaussians for fast generalizable 3d reconstruction. *arXiv preprint arXiv:2403.16292*, 2024. 2
- [45] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemel Pontes, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv preprint arXiv:2301.00493*, 2023. 2, 3, 4
- [46] Chenming Wu, Jiadai Sun, Zhelun Shen, and Liangjun Zhang. Mapnerf: Incorporating map priors into neural radiance fields for driving view simulation. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7082–7088. IEEE, 2023. 3
- [47] Ke Wu, Kaizhao Zhang, Zhiwei Zhang, Shanshuai Yuan, Muer Tie, Julong Wei, Zijun Xu, Jieru Zhao, Zhongxue Gan, and Wenchao Ding. Hgs-mapping: Online dense mapping using hybrid gaussian representation in urban scenes. *arXiv preprint arXiv:2403.20159*, 2024. 3

- [48] Zirui Wu, Tianyu Liu, Liyi Luo, Zhide Zhong, Jianteng Chen, Hongmin Xiao, Chao Hou, Haozhe Lou, Yuantao Chen, Runyi Yang, et al. Mars: An instance-aware, modular and realistic simulator for autonomous driving. In *CAAI International Conference on Artificial Intelligence*, pages 3–15. Springer, 2023. 3
- [49] Bernhard Wyman, Eric Espié, Christophe Guionneau, Christos Dimitrakakis, Rémi Coulom, and Andrew Sumner. Torcs, the open racing car simulator. *Software available at http://torcs.sourceforge.net*, 4(6):2, 2000. 3
- [50] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34: 12077–12090, 2021. 6
- [51] Danfei Xu, Yuxiao Chen, Boris Ivanovic, and Marco Pavone. Bits: Bi-level imitation for traffic simulation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2929–2936. IEEE, 2023. 3
- [52] Runsheng Xu, Xin Xia, Jinlong Li, Hanzhao Li, Shuo Zhang, Zhengzhong Tu, Zonglin Meng, Hao Xiang, Xiaoyu Dong, Rui Song, et al. V2v4real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13712–13722, 2023. 3
- [53] Yunzhi Yan, Haotong Lin, Chenxu Zhou, Weijie Wang, Haiyang Sun, Kun Zhan, Xianpeng Lang, Xiaowei Zhou, and Sida Peng. Street gaussians for modeling dynamic urban scenes. *arXiv preprint arXiv:2401.01339*, 2024. 2, 3
- [54] Chen Yang, Peihao Li, Zanwei Zhou, Shanxin Yuan, Bingbing Liu, Xiaokang Yang, Weichao Qiu, and Wei Shen. Nervfs: Neural radiance fields for free view synthesis via geometry scaffolds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16549–16558, 2023. 2
- [55] Jiawei Yang, Boris Ivanovic, Or Litany, Xinshuo Weng, Sung Wook Kim, Boyi Li, Tong Che, Danfei Xu, Sanja Fidler, Marco Pavone, et al. Emernerf: Emergent spatial-temporal scene decomposition via self-supervision. *arXiv preprint arXiv:2311.02077*, 2023. 3
- [56] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024. 8
- [57] Ze Yang, Yun Chen, Jingkang Wang, Sivabalan Manivasagam, Wei-Chiu Ma, Anqi Joyce Yang, and Raquel Urtasun. Unisim: A neural closed-loop sensor simulator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1389–1399, 2023. 2, 3
- [58] Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-splatting: Alias-free 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19447–19456, 2024. 3
- [59] Zhongrui Yu, Haoran Wang, Jinze Yang, Hanzhang Wang, Zeke Xie, Yunfeng Cai, Jiale Cao, Zhong Ji, and Mingming Sun. Sgd: Street view synthesis with gaussian splatting and diffusion prior. *arXiv preprint arXiv:2403.20079*, 2024. 3
- [60] Dongbin Zhang, Chuming Wang, Weitao Wang, Peihao Li, Minghan Qin, and Haoqian Wang. Gaussian in the wild: 3d gaussian splatting for unconstrained image collections. *arXiv preprint arXiv:2403.15704*, 2024. 9, 10
- [61] Jian Zhang, Yuanqing Zhang, Huan Fu, Xiaowei Zhou, Bowen Cai, Jinchi Huang, Rongfei Jia, Binqiang Zhao, and Xing Tang. Ray priors through reprojection: Improving neural radiance fields for novel view extrapolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18376–18386, 2022. 2
- [62] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6
- [63] Guosheng Zhao, Chaojun Ni, Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Boyuan Wang, Youyi Zhang, Wenjun Mei, and Xingang Wang. Drivedreamer4d: World models are effective data machines for 4d driving scene representation. *arXiv preprint arXiv:2410.13571*, 2024. 3
- [64] Ziyuan Zhong, Davis Rempe, Danfei Xu, Yuxiao Chen, Sushant Veer, Tong Che, Baishakhi Ray, and Marco Pavone. Guided conditional diffusion for controllable traffic simulation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3560–3566. IEEE, 2023. 3
- [65] Hongyu Zhou, Jiahao Shao, Lu Xu, Dongfeng Bai, Weichao Qiu, Bingbing Liu, Yue Wang, Andreas Geiger, and Yiyi Liao. Hugs: Holistic urban 3d scene understanding via gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21336–21345, 2024. 2, 3
- [66] Shijie Zhou, Haoran Chang, Sicheng Jiang, Zhiwen Fan, Zehao Zhu, Dejia Xu, Pradyumna Chari, Suya You, Zhangyang Wang, and Achuta Kadambi. Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21676–21685, 2024. 5, 6, 7
- [67] Xiaoyu Zhou, Zhiwei Lin, Xiaojun Shan, Yongtao Wang, Deqing Sun, and Ming-Hsuan Yang. Drivinggaussian: Composite gaussian splatting for surrounding dynamic autonomous driving scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21634–21643, 2024. 2, 3

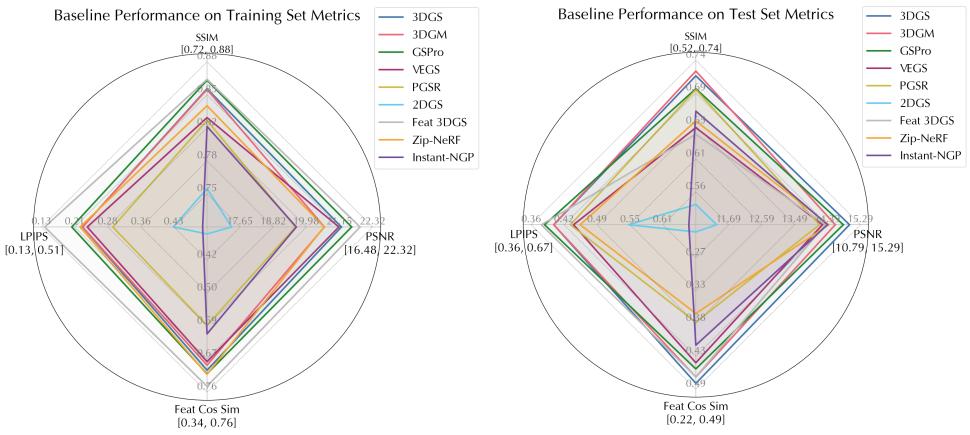
Appendix



(a) Baseline performance comparison at Level 1.

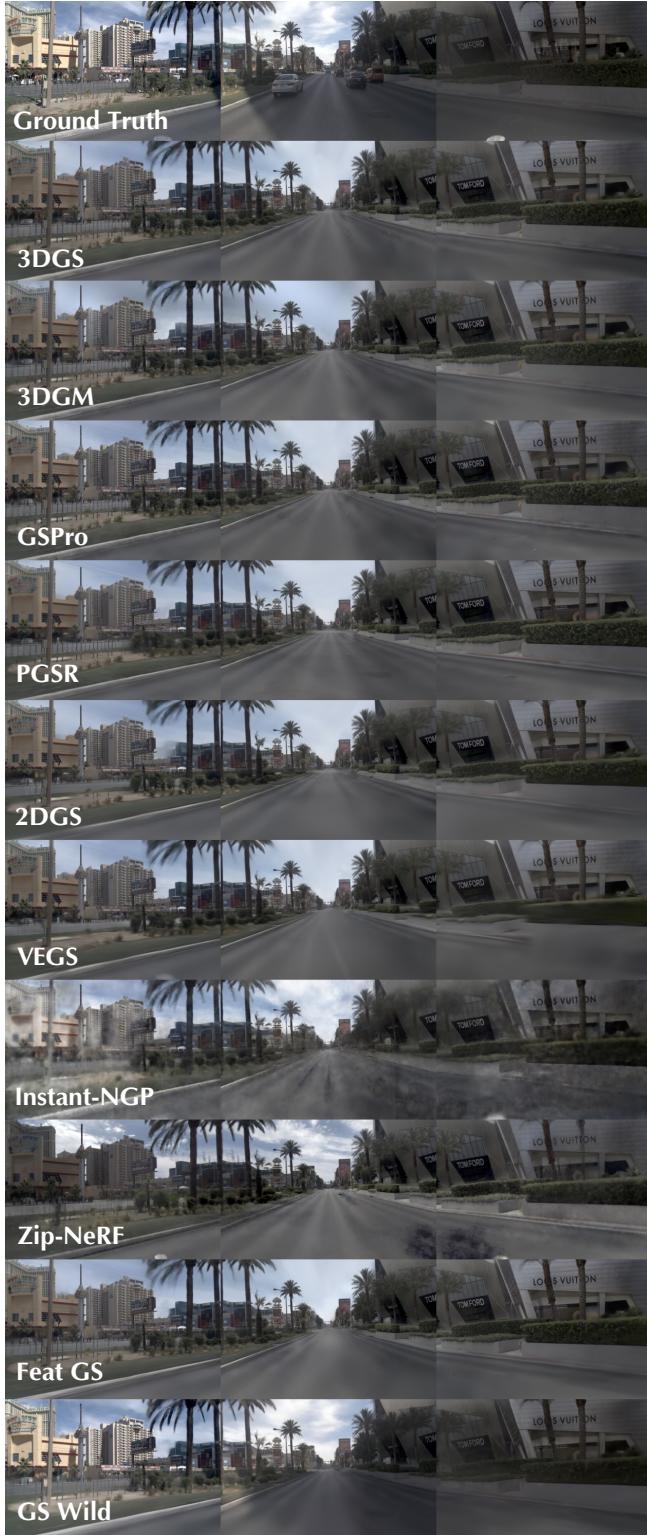


(b) Baseline performance comparison at Level 2.



(c) Baseline performance comparison at Level 3.

Figure I. Baseline performance comparison across different levels. Since scenes at different difficulty levels evaluate varying capabilities, different baselines demonstrate strengths at different difficulty levels. However, as the difficulty increases, the baselines tend to fail more uniformly, leading to a situation where no single baseline demonstrates a significant advantage.

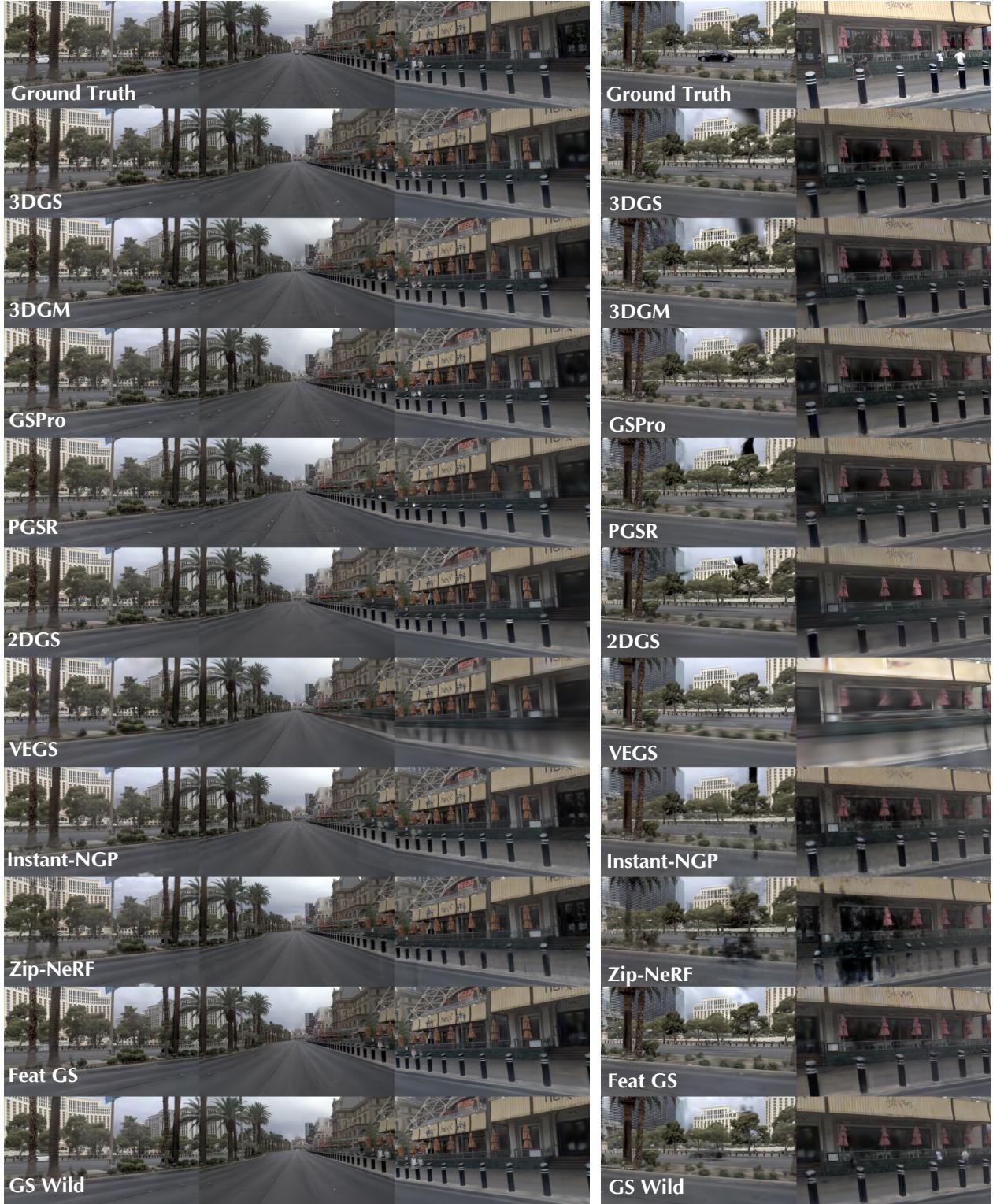


(a) Rendering results comparison in original view.



(b) Rendering results comparison in extrapolated view.

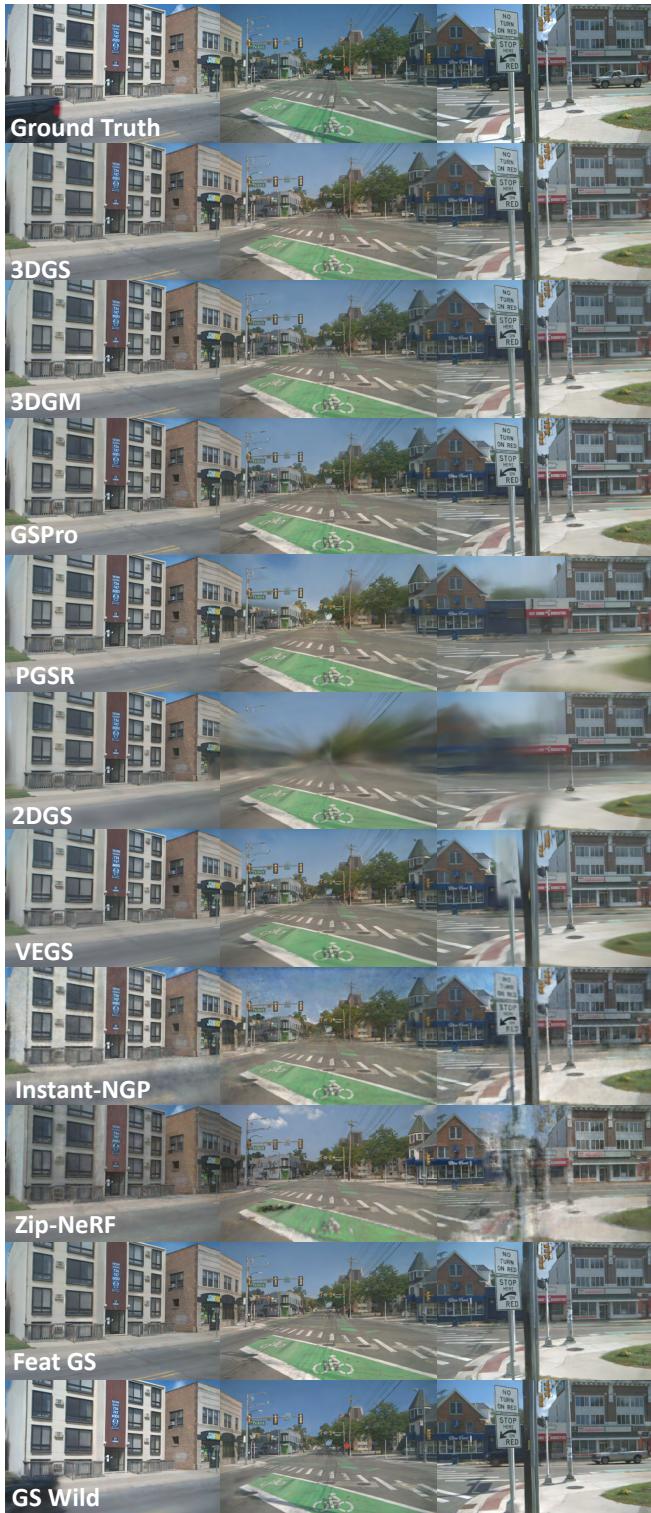
Figure II. **Qualitative comparison of baseline methods at Level 1.** Ground reconstruction failures and floating artifacts in the sky are particularly noticeable, highlighting the challenges in the lane change.



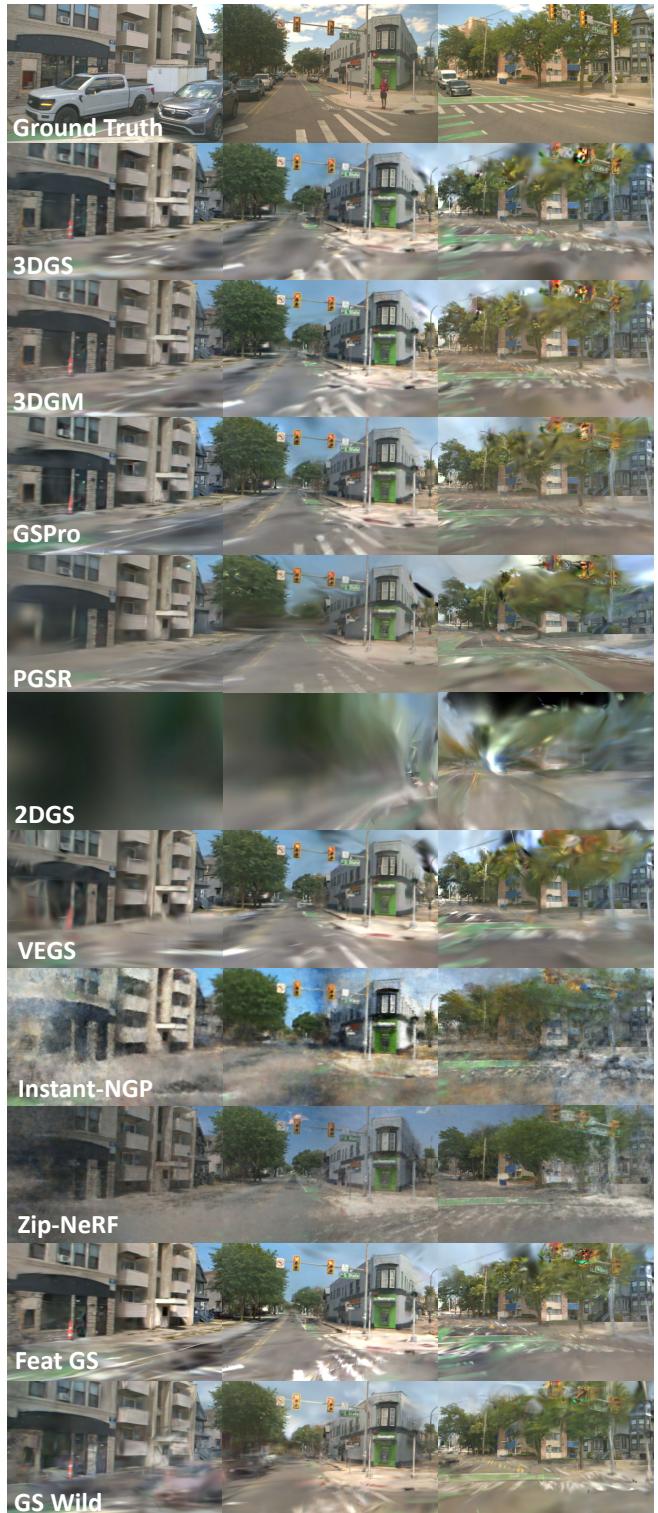
(a) Rendering results comparison in original view.

(b) Rendering results comparison in extrapolated view.

Figure III. Qualitative comparison of baseline methods at Level 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 IV. Qualitative comparison of baseline methods at Level 3. The rendering quality deteriorates significantly in extrapolated viewpoints. The geometry becomes fragmented, especially in trees, traffic lights, and lane marks.