
The Road Ahead in Autonomous Driving: The KITScenes Multimodal Dataset

Richard Schwarzkopf^{1,2*} Fabian Immel^{1,2*}
Alexander Blumberg² Jonas Merkert² Nils Rack² Kaiwen Wang²
Fabian Konstantinidis² Julian Truetsch^{1,2} Carlos Fernandez² Annika Bätz²
Kevin Rösch^{1,2} Marlon Steiner¹ Willi Poh² Yinzhe Shen² Royden Wagner²
Felix Hauser² Dominik Strutz² Jaime Villa³ Gleb Stepanov²
Holger Caesar⁴ Ömer Şahin Taş^{1,2} Frank Bieder^{1,2}
Jan-Hendrik Pauls^{2†}✉ Christoph Stiller^{1,2}

¹FZI Research Center for Information Technology ²Karlsruhe Institute of Technology
{immel, schwarzkopf}@fzi.de jan-hendrik.pauls@kit.edu

³ University Charles III of Madrid ⁴Delft University of Technology

Abstract

Existing autonomous driving datasets have enabled major progress, but fall short in sensor fidelity, map completeness, or geographic diversity. We present KITScenes Multimodal, a European dataset built around high-fidelity sensors and maps. Our fully synchronized sensor suite combines high-resolution global-shutter cameras, long-range lidar beyond 400 m, 4D imaging radar, and redundant GNSS/INS localization. Our HD maps are, to our knowledge, the most complete of any sensor dataset, validated through autonomous driving trials on open-source software. For the first time in a public dataset, all driving-relevant traffic elements, such as traffic lights, are mapped in 3D to a reprojection-accurate level with full topological connectivity. Recorded in cities with irregular street layouts and mixed traffic modes, our dataset complements existing datasets by broadening the available geographic diversity. We also introduce four benchmarks, each advancing spatial learning for embodied AI: online HD map construction, long-range depth estimation, novel view synthesis, and end-to-end driving. Project page: <https://kitscenes.com/>.

1 Introduction

Autonomous driving datasets [1–4] have enabled significant progress in both computer vision and autonomous driving research. However, existing datasets still fall short of capturing the complexity required for spatially aware driving in dense urban environments. Some lack public annotations or topology-aware map references [2, 3], while others focus on comparatively simple driving scenarios such as motorways [5, 6]. As autonomous driving systems move toward deeper spatial understanding, datasets must support reasoning not only about objects, but also about geometry, road structure, and their geospatial relationships. High-fidelity datasets enriched with geospatial annotations, HD maps, and 3D labels are essential for evaluating such capabilities.

High-fidelity datasets with geospatial annotations have a limited geographic footprint, with coverage heavily skewed toward North America and Asia. KITTI [1], though seminal, is small-scale; ZOD [7] annotates only single keyframes with image-space labels; and large-scale recording efforts such as those from Nvidia [8] still lack public annotations. Consequently, complex European urban

*Joint first authors

†Project lead and ✉ Corresponding author

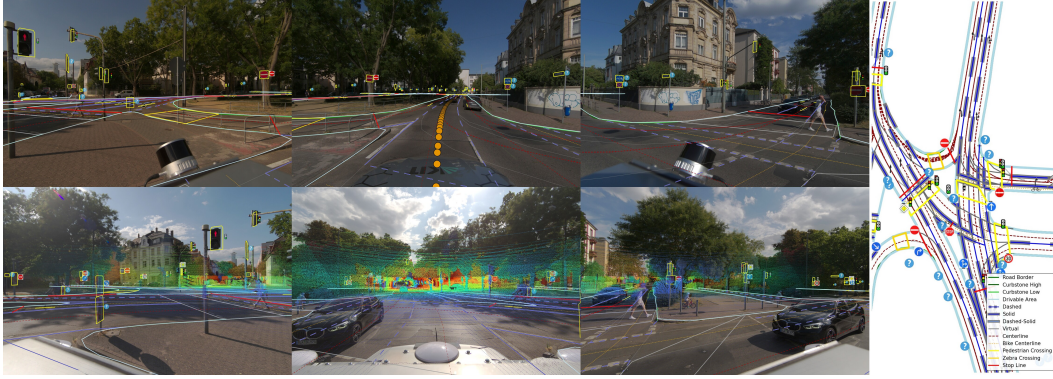


Figure 1: A showcase of 3D HD map elements and the ground truth reprojected into 6 out of 9 cameras, the dense long-range lidar pointcloud reprojected into the rear cameras, and a top-view of the HD map benchmark labels. Best viewed zoomed in.

environments remain underrepresented in current autonomous driving benchmarks, arguably being the most difficult to spatially reason about. This leaves a clear need for datasets that combine high-fidelity sensing, complete geospatial context, and dense 3D annotations.

We present **KITScenes Multimodal**, a dataset recorded across diverse European urban environments using a state-of-the-art robotaxi sensor platform. Our dataset addresses the geographic gap in existing benchmarks while simultaneously raising the bar on both sensor fidelity and geospatial understanding. Our sensor platform combines high-resolution cameras (up to 16.2 Mpx), long-range lidar with effective range beyond 400 m, 4D imaging radar, and redundant GNSS, all hardware-synchronized and processed with high-fidelity pipelines that make the data suitable for applications such as neural rendering and novel view synthesis. Besides high fidelity sensor data, we provide the most complete HD maps of any public autonomous driving dataset. Annotated in Lanelet2 [9], our maps visualized in Figure 1 cover all regulatory road feature and traffic sign classes, and host our annotated 3D traffic lights, signs, and poles with reprojection-accurate localization.

To demonstrate the unique strengths of the dataset, we introduce four benchmarks: (1) **Complete online HD map perception**, evaluating relational Lanelet2 map prediction from sensor data; (2) **long-range monocular depth estimation**, targeting depth beyond 200 m where current methods degrade severely; (3) **novel view synthesis**, exploiting our high-fidelity imagery and dense lidar for 3D scene reconstruction; and (4) **multimodal end-to-end models for autonomous driving**, predicting future trajectories and scene evolution from camera, lidar, and radar inputs. Our contributions include:

- A multimodal European driving dataset, recorded in three cities with a high-fidelity robotaxi sensor suite: 72.5 Mpx of synchronized global-shutter cameras, seven lidars with over $3\times$ the point density and twice the effective range of the next closest dataset, three 4D imaging radars, and redundant GNSS/INS.
- Production-grade Lanelet2 HD maps covering 62 km^2 with 29 road-feature classes, 120 traffic-sign classes, and 3D traffic lights, signs, and poles localized to reprojection accuracy. The maps include all regulatory elements required for autonomous navigation and are validated for use in the open-source Autoware [10] stack, both online and in simulation.
- Four benchmarks designed to expose the limits of current methods on the path to Level 4 autonomy, targeting capabilities existing datasets cannot benchmark at this fidelity: holistic HD map prediction, depth estimation beyond 200 m, high-fidelity novel view synthesis, and multi-modal end-to-end driving.

2 Related Work

Autonomous Driving Datasets for Perception The past decade has seen a rapid growth of autonomous driving datasets. Foundational datasets such as nuScenes [2], Waymo Open [3], and Argoverse 2 [11] established the multimodal paradigm with complementary sensor configurations

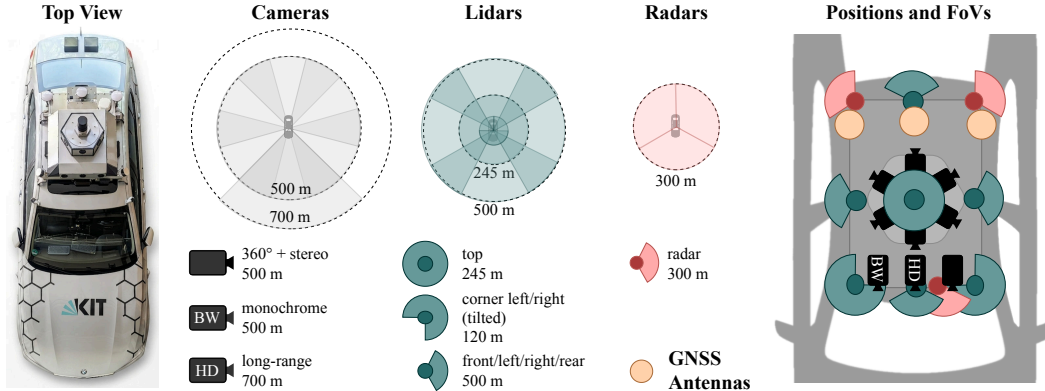


Figure 2: KITScenes Multimodal Sensor Setup. Our sensor rack (left) is depicted along with nominal sensing range (center), as well as sensor positions and their field of view (right).

and annotation schemes. Further datasets [12, 13] broaden the range of traffic layouts and driving conditions, although detailed map annotations and deployment-oriented perception support remain limited. KITTI [1] and KITTI-360 [14] remain influential but limited in scale and sensor diversity by current standards. ZOD [7] provides large-scale recordings, yet annotates only a single keyframe per scenario and mainly provides image-space labels. MAN TruckScenes [5] focuses on motorway trucking rather than complex urban perception. While TruckDrive [6] features long-range sensors, it likewise targets trucking scenarios, relies on automotive RCCB cameras, and has not released any public data to date. Large-scale fleet recordings such as Nvidia Physical AI AV [8] provide broad real-world coverage but lack public annotations. A quantified comparison of the sensor setups is shown in Table 1.

HD Maps and Map Perception Benchmarks Map representations accompanying public datasets vary substantially in completeness. nuScenes [2] and Argoverse 2 [11] expose lane geometry via dataset-specific APIs but omit regulatory structure from traffic lights and signs. OpenLaneV2 [15] adds lane-topology links, but as image-space annotations rather than metric 3D maps. To our knowledge, no prior dataset provides HD maps that are simultaneously reprojection-accurate, complete in regulatory structure (traffic signs, lights, lane assignments), and validated in a planning stack (Table 2). As a consequence, so far online HD map construction methods [16? –28] are evaluated on simple geometric primitives only (lane dividers without type, pedestrian crossings, road borders). Lanelet2 [9] has emerged as the open academic standard for HD maps, encoding geometry, topology, and 3D regulatory elements in a single graph; it is the native input of Autoware [10] and translatable to learning-friendly representations using [29].



Long-range Perception, Neural Rendering, and End-to-End Driving Monocular depth estimation is predominantly benchmarked on KITTI [1] and DDAD [30]; recent foundation models [31, 32] achieve strong near-range performance, but existing benchmarks rarely assess depth beyond 80–100 m. Neural scene representations for driving like NeRF-based [33, 34] and 3D Gaussian Splatting methods [35–37], are similarly constrained by input image fidelity and lidar density. End-to-end driving models [38, 39] and world models are evaluated almost exclusively on nuScenes, limiting the sensor configurations and geographies under which they are assessed.



3 The KITScenes Multimodal Dataset

3.1 High-Resolution Long-Range Multi-Modal Sensor Setup

KITScenes Multimodal uses a fully synchronized sensor suite. Figure 2 depicts the sensor positions and their nominal fields of view. To enable sensor fusion up to maximum effective sensing range, we perform intrinsic and extrinsic calibration across all modalities, achieving subpixel intrinsic and 1 cm and 0.1° extrinsic accuracy. Further details are listed in Section A and Section B.

Table 1: KITScenes Multimodal sets a new state of the art for temporally consistent high-resolution high-fidelity RGB surround vision, highly dense long-range lidar, and ranging modality coverage. We triple the average lidar point density and almost double the typical maximum range; see also Figure 5.

Dataset	Cameras						Radar		Lidar			
			MPix	FOV	Shutter	Cam. Sync	Comp.	Config	#	Avg pts	Max pts	Max range
nuScenes [2]	6	–	8.4	360°	Rolling	to lidar	JPEG	5×3D	1	34.7 k	34.8 k	102.1 m
ONCE [13]	7	–	14.5	360°	Rolling	to lidar	JPEG	–	1	64.7 k	69.7 k	196.8 m
nuPlan Sensors [41]	8	–	19.2	360°	Rolling	to lidar	JPEG	–	5	93.0 k	100.3 k	215.5 m
Argoverse 2 Sensor [11]	7	1	28.6	360°	Rolling	to lidar	JPEG	–	2	96.9 k	106.3 k	217.4 m
WOD Perception [3]	5	–	10.4	230°	Rolling	to lidar	JPEG	–	5	175.5 k	215.9 k	75.0 m
MAN TruckScenes [5]	4	–	9.3	360°	Rolling	to lidar	JPEG	6×4D	6	231.7 k	296.7 k	221.6 m
Zenseact Open [7]	1	–	8.3	120°	Rolling	–	PNG	–	3	253.7 k	311.1 k	244.0 m
Nvidia PhysicalAI AV [8]	7	–	14.5	360°	Rolling	no	H.264	9×4D	1	297.2 k	344.1 k	206.0 m
KITScenes Multimodal	7	1	72.5	360°	Global	all cameras	JPEGLI	3×4D	7	906.4 k	1235.2 k	409.2 m

 Monocular cameras,  Stereo camera pair, MPix = Total resolution per frame, Comp. = Image compression

Cameras. The camera suite comprises six 7.1 Mpx surround cameras providing full 360° coverage, one 16.2 Mpx high-resolution long-range camera, and a tilted forward-facing stereo setup, yielding a combined resolution of 72.5 Mpx per frame, which is more than twice that of the next closest dataset (Table 1). Existing setups put their focus on dynamic object perception [2, 3, 5, 11], triggering the cameras when the lidar swept across the image center to ensure a minimal delay between both modalities. All cameras use global shutter sensors and are hardware-synchronized, ensuring pixel-accurate temporal alignment. The images are anonymized and compressed with JPEGLI [40], a state-of-the-art visually lossless codec described in Section A.1. This is the foundation for our high fidelity ground truth for neural rendering and novel synthesis. At the same time, we ensure lidar coverage by redundantly combining multiple lidars with varying sweeping directions.

Lidar. Seven lidar sensors provide 360° coverage with substantial overlap between adjacent units. As shown in Table 1, the fused point cloud contains on average more than 900 k points per frame with peaks above 1.2 M points, tripling the effective point density over existing datasets. The use of 1550 nm lidars enables an average maximum range of more than 400 m, nearly doubling that of the next-best dataset. This long-range capability is essential for both online long-range perception and for providing ground truth for benchmarks, such as monocular depth estimation. Figure 5 compares the per-distance-bin return density for KITScenes and existing autonomous driving datasets, showing that KITScenes provides higher effective point density in every bin and extends usable range beyond 250 m, where prior datasets fall to zero.

3.2 HD Map Annotation

We provide pixel-accurate 3D maps that can be directly used in the open-source Autoware [10] stack, both for simulation and real-world autonomous driving. All maps are annotated in Lanelet2 [9], an established open-source format for semantic HD maps. Beyond geometry, each map encodes the full regulatory structure required for autonomous driving: Road level polylines are annotated with one of 29 classes, (*e.g.*, road border, dashed, zebra-crossing *etc.*) traffic signs are classified based on 220 German road traffic code classes [43] (with 120 observed), traffic lights types are grouped into four categories (car, bike, pedestrian, misc). All traffic signs and lights are explicitly assigned to the lanes they govern via topological links in the Lanelet2 format. Traffic lights, road signs, and poles are annotated based on lidar and camera data as 3D shapes including orientation that are reprojection-accurate to the calibrated camera images [44]. This reprojection accuracy directly connects map labels to image pixels, enabling HD map annotations to be used as pixel-level training signal for perception models without any additional alignment step, as shown in Section 4.1.

3.3 Dataset Statistics

Our current release contains 1007 10 s to 60 s scenarios totaling 5.7 h and 162 km of synchronized multimodal recording at 10 Hz. Details on the split and label statistics can be found in Section G. The dataset currently spans Karlsruhe, Frankfurt, and Sindelfingen, chosen for their unique environments of a planned 18th century radial layout, a dense metropolitan financial district core, and a suburban-

Table 2: Comparison of related datasets comprised of HD maps and sensor data, datasets from Table 1 without HD maps are not listed. Legend: \checkmark yes, \checkmark partial/limited, \times no; $()$: unreleased data; \uparrow : large coverage based on dataset description which is not reported or reproduced area coverage.

	Dataset	Area (km ²)	Region	All sensors	360°	3D lanes	Lane border type	Bike Lanes	3D Traffic elements	Full topology	Human HD map	OSS AD stk.
Limited spatial learning	WOD Perception [3]	76 km ²	US	\checkmark	\times	\checkmark	\checkmark	\checkmark	\times	\times	\checkmark	\times
	nuPlan Sensors [†] [41]	\uparrow	US, Asia	\checkmark [†]	\checkmark	\times	\checkmark	\times	\times	\checkmark [†]	\checkmark	\times
	AV2 TbV [42]	42 km ²	US	\times	\checkmark	\checkmark	\checkmark	\checkmark	\times	\times	\checkmark	\times
	Nvidia PhysicalAI AV [8]	$\uparrow\uparrow$	US, EU	\checkmark	\checkmark	(\checkmark)	(\times)	(\times)	(\checkmark)	(\times)	(\times)	\times
Full spatial learning	nuScenes [2]	5 km ²	US, Asia	\checkmark	\checkmark	\times	\checkmark	\times	\times	\times	\checkmark	\times
	Argoverse 2 Sensor [11]	17 km ²	US	\checkmark	\checkmark	\checkmark	\checkmark	\checkmark	\times	\times	\checkmark	\times
	OpenLane-V2 [15]	22 km ²	US, Asia	\times	\checkmark	\checkmark	\checkmark	\times	\times	\checkmark [†]	\checkmark	\times
	KITScenes Multimodal	62 km ²	EU	\checkmark	\checkmark	\checkmark	\checkmark	\checkmark	\checkmark	\checkmark	\checkmark	\checkmark

[†]Remarks: nuPlan Sensor [41]: shorthand for the 10% of scenes in nuPlan with available sensor data. traffic light states available through offline state estimation, no linkage to sensor data. NVIDIA PhysicalAI AV: entries transparent filled based on current publically available release plans, not verified. OpenLaneV2: built on top of sensor data of AV2 and nuScenes, with limited set of labeled traffic element 2D bounding boxes in a visible range of 25x50m at 2Hz. **All sensors**: full suite and quality of original sensor dataset available. **OSS AD stack**: Native support of HD map for simulation and closed-loop driving with open-source software autonomous driving stack. **Full spatial learning**: support for full resolution multimodal 360° surround view learning with a at least a base set of BEV annotations.

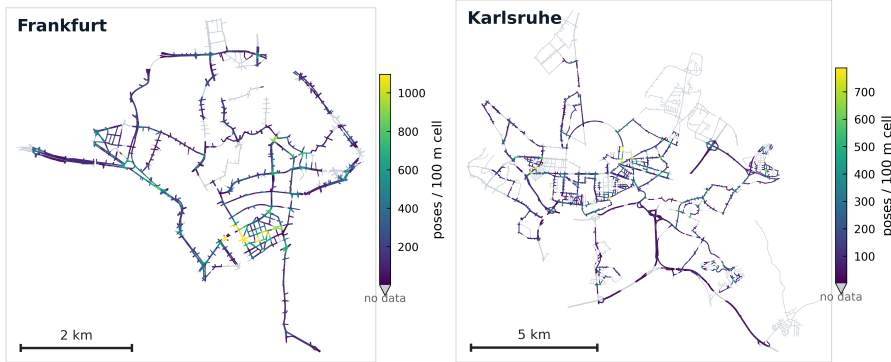


Figure 3: Spatial coverage for two KITScenes cities. The color indicates the number of poses within a 100m grid cell on top of our HD map outlined in the background.

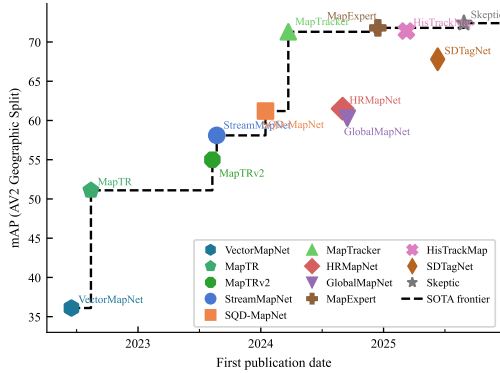
industrial mix. Recordings took place across summer 2025 and winter 2025/26 to expose models to seasonal appearance changes and a wide coverage as visualized in Figure 3.

4 Benchmarks

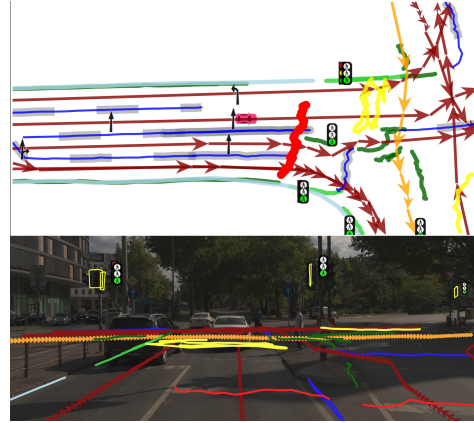
Our benchmarks span spatial learning from map-level scene understanding to multimodal end-to-end driving. They expose limitations of existing methods that prior datasets cannot reveal.

4.1 Online HD Map Construction

Online HD map construction aims to predict a structured, drivable map directly from onboard sensor data, without relying on pre-built prior maps. Existing benchmarks evaluate the prediction of simple geometric primitives such as lane dividers and pedestrian crossings [19], leading to a saturation of existing benchmarks, as shown in Figure 4a. We enable a substantially more complete formulation: our Lanelet2 maps encode lane topology, regulatory elements, traffic signs, and traffic lights with their lane assignments, allowing models to be evaluated on predicting the full Lanelet2 map structure. As a baseline for topology prediction, we extend MapQR [45] with a graph neural network (GNN) head that consumes the map element tokens from the decoder and predicts pairwise relations between all predicted map elements (hereafter called MapQR-Topo). Architecture and implementation details are described in Section H.1.



(a) Historical SOTA progression of online HD map construction models [17–19, 47, 22–27, 46, 28] on AV2 [11]. A saturation on the current datasets, perception range and task complexity can be seen after the introduction of Maptracker [23].



(b) Example online HD map construction prediction of SDTagNet [46] on a validation sample. While showing new capabilities such as 3D detection of non-ground elements thanks to the extensive map labels, a large gap in prediction of complete 3D HD maps remains.

Figure 4: Historical SOTA progression of online HD map construction models and example online HD map construction prediction on KITScenes Multimodal.

Results. In Table 3, we evaluate MapTRv2 [19] as a widely adopted camera-only baseline and SDTagNet [46] as a representative of methods that leverage SD map priors. Both exhibit a large performance drop on our complete formulation compared to existing benchmarks, revealing a gap hidden by the currently limited task scope, with SDTagNet benefiting more from the richer formulation. This suggests that structured prior knowledge becomes increasingly valuable as the task approaches real-world complexity. An example of prediction outputs is provided in Figure 4b. A qualitative example the predicted topology by MapQR-Topo is shown in Figure 17 in the Appendix.

Table 3: Evaluation of online HD map perception models. For readability, the classes are grouped into 6 categories for the average precision: Lane Markings (LM), Lane Centerlines (LC), Road Infrastructure (RI), Traffic Lights (TL), Traffic Signs (TS) and Road Markings (RM). For the topology prediction baseline MapQR-Topo we additionally report the topology score.

Model	AP_{LM}	AP_{LC}	AP_{RI}	AP_{TL}	AP_{TS}	AP_{RM}	AP_{Topo}
MapTRv2 [19]	5.1	18.0	6.7	5.8	3.0	8.1	-
SDTagNet [46]	4.5	19.4	7.1	6.3	2.4	9.0	-
MapQR-Topo	4.1	16.0	5.9	3.6	1.9	5.6	16.4

4.2 Long-range Monocular Depth Estimation

Monocular depth estimation has made rapid progress on near-range benchmarks, yet autonomous driving at highway speeds and in complex intersections requires reliable depth estimates well beyond 100 m. We show that current depth estimation models trained and evaluated on existing datasets fail to generalize to long-range distances, as their training signal is dominated by close-range lidar returns. We provide a dedicated benchmark for long-range monocular depth estimation, enabling the first systematic evaluation of depth estimation at ranges that extend beyond 400 m.

We report established metrics for monocular depth evaluation: absolute relative error (AbsRel) and threshold accuracy δ_1 . Scores are reported stratified into close range (0 m–100 m), medium range (100 m–200 m) and far range (>200 m), and overall. A detailed description of the setup and ground truth generation can be found in Section H.2.

Results. We evaluate UniDAC [32], Depth Anything 3 [31], and MapAnything [48], all reported to achieve dataset-agnostic SOTA monocular depth estimation. They provide strong performance at close

Figure 5: Lidar distribution across major autonomous driving datasets. Lines show the per-bin mean over 500 train samples. KITScenes Multimodal sets a new benchmark in both density and range.

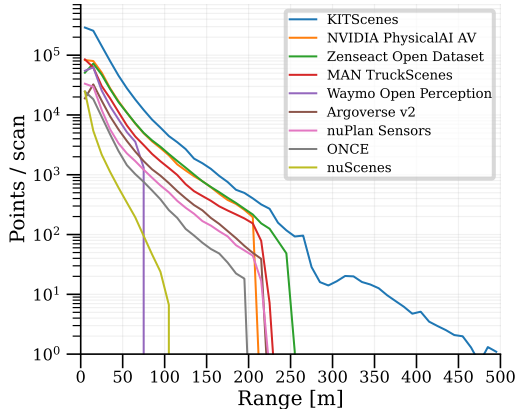


Figure 6: Depth distributions of pixels with valid lidar depth. Depending on the method and compared to the ground, a systematic shortfall compared to the ground truth is observable starting at 75 m–125 m

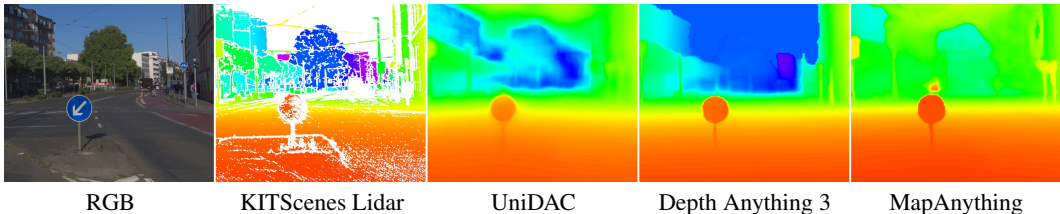
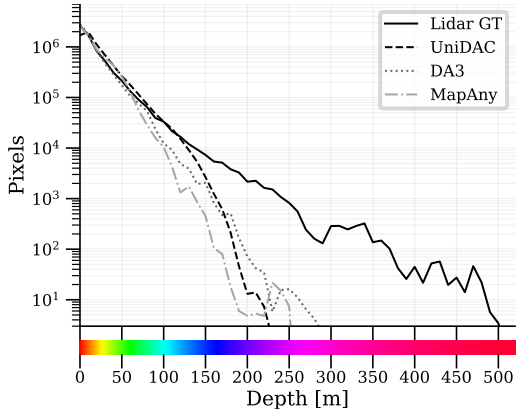


Figure 7: Qualitative comparison of monocular depth estimation methods. The corresponding non-linear depth scale is introduced in Figure 6. All methods systematically underestimate depth at long range relative to the lidar reference, with MapAnything exhibiting the largest deviation.

Table 4: Range-stratified metric depth estimation exposes a ranking inversion: MapAnything dominates overall and at 0 m–100 m but degrades severely beyond it, while UniDAC, ranked last overall, is the strongest long-range estimator. Regardless, all methods perform poorly beyond 200 m.

Method	0 m–100 m		100 m–200 m		>200 m		Overall	
	AbsRel ↓	δ_1 ↑	AbsRel ↓	δ_1 ↑	AbsRel ↓	δ_1 ↑	AbsRel ↓	δ_1 ↑
UniDAC [32]	0.386	24.12	0.302	40.17	0.540	1.78	0.384	24.36
Depth Anything 3 [31]	0.278	48.64	0.472	12.32	0.689	0.86	0.282	47.91
MapAnything [48]	0.149	83.04	0.485	16.34	0.772	0.03	0.156	81.70

range, but fall short as early as 75 m (see Figure 6). Table 4 reveals a critical limitation of aggregate evaluation: overall metrics mask severe performance inversions across depth ranges. MapAnything dominates the 0 m–100 m range and ranks first overall, yet degrades significantly beyond it. UniDAC, ranked last overall, is in fact the strongest long-range estimator by a significant margin. Regardless, no method achieves reliable performance beyond 200 m (further evaluations in Section H.2). With its comprehensive LiDAR setup, KITScenes is uniquely positioned (see Figure 5) to expose such limitations, providing the long-range ground truth density necessary to benchmark methods where current autonomous driving datasets fall short.

4.3 Novel View Synthesis

Neural scene representations and novel view synthesis (NVS) methods have emerged as powerful tools for autonomous driving simulation and data augmentation. Common NVS methods [35–37] are evaluated using pixel-based metrics, but this strongly relies on the availability of ground truth images at target viewpoints, which are typically restricted to the original driven trajectory. While

Table 5: Traffic sign recall on the front camera at seven lateral offsets. The “low” and “high” rows denote evaluations at 280×518 (model scale) and 1600×2844 (cropped sensor scale), respectively. “Photo” is the detector’s recall on the real photograph (upper bound). \uparrow denotes higher is better.

	Photo (%) \uparrow	-3 m	-2 m	-1 m	0 m	+1 m	+2 m	+3 m
low	19.7	4.1 (-79.2%)	6.7 (-66.0%)	11.4 (-42.1%)	18.2 (-7.6%)	11.0 (-44.2%)	5.5 (-72.1%)	3.7 (-81.2%)
high	21.6	3.4 (-84.3%)	5.5 (-74.5%)	9.5 (-56.0%)	15.6 (-27.8%)	9.4 (-56.5%)	4.6 (-78.7%)	3.0 (-86.1%)

lateral novel view synthesis is critical for autonomous driving simulation, its quality is often judged only through qualitative inspection [37] and image-based metrics [49, 50]. However, those often fail to reveal subtle structural distortions that can significantly impact downstream perception tasks. To probe geometric fidelity at novel lateral poses, we introduce a map-based NVS evaluation benchmark using traffic sign recall.

We re-render the scene at seven lateral offsets $\Delta y \in \{-3, \dots, +3\}$ m and project ground-truth traffic signs from our HD map into each shifted viewpoint, applying lidar-based occlusion filtering to retain only unoccluded signs. We report traffic sign recall at both a low resolution (280×518 , matching the model’s output) and a high resolution (1600×2844 , the cropped sensor resolution), with the real photograph serving as the per-scale upper bound. A full description is given in Section H.3.

Results. As shown in Table 5, evaluating ReconDrive [37] reveals a sharp collapse in structural fidelity: even at the driven trajectory ($\Delta y=0$), upsampling to the sensor’s cropped resolution yields a 27.8% relative recall drop, nearly four times the 7.6% drop at low resolution. This indicates that the reconstruction lacks fine-grained structural detail. With lateral translation, degradation exceeds 80% relative recall loss at $\Delta y = \pm 3$ m, showing that current NVS methods struggle to maintain geometric integrity in novel views, a limitation hidden by standard photometric metrics. A qualitative example of lacking 3D consistency is shown in Figure 8, where the traffic sign fails to maintain its true 3D position after a viewpoint shift. More details, further qualitative comparison in Figure 19 and standard photometric metrics are provided in Section H.3.

4.4 End-to-End Driving

End-to-end driving and neural world models are evaluated almost exclusively on nuScenes, narrowing the sensor configurations, geographies, and map-grounded behaviours under which they are assessed. KITScenes Multimodal supports three input tiers on identical scenes, i.e., a single front-view camera, the full 360° surround-view, and the complete multi-modal suite with lidar and radar, enabling controlled modality ablations with a novel combination of benchmark metrics. Headline baselines reported here are camera-only; sensor and timing data for all tiers are released, leaving multi-modal e2e training as an open challenge. Evaluation setup, split details, and the held-out test-e2e leaderboard split are described in Section H.4.

Beyond standard ADE and FDE [51], we leverage our centimetre-accurate Lanelet2 maps and a lidar-derived occupancy layer to evaluate three map-grounded safety metrics: *drivable-surface survival*, *collision-free rate*, and *centerline distance*, serving as an offline proxy for safety properties usually assessed only in closed-loop simulation. To decouple correctness from a single expert trajectory, we additionally adopt the *Multi-Maneuver Score* (MMS) [52], scoring each prediction against the best of at least three human-annotated admissible maneuvers per scene. Metric definitions and per-horizon profiles are detailed in Section H.4.

Results We zero-shot evaluate four open-source baselines: UniAD [38] and DMAD [53], multi-task perception, prediction, and planning models trained with navigation commands on nuScenes; SSR [54], which plans directly with a self-supervised BEV regulariser; and Epona [55], an autoregressive front-view diffusion world model trained on nuPlan without navigation commands. Table 6 reveals a substantial domain gap, least pronounced for Epona, which is consistent with its larger pretraining corpus. The same ordering holds under the multi-maneuver criterion in Table 18. Figure 9 illustrates a qualitative example of end-to-end predictions.



Figure 8: Example of lacking 3D geometric integrity in current NVS methods. The traffic sign in the shifted view on the right is inconsistent with its true 3D position shown by the reprojected bounding box.

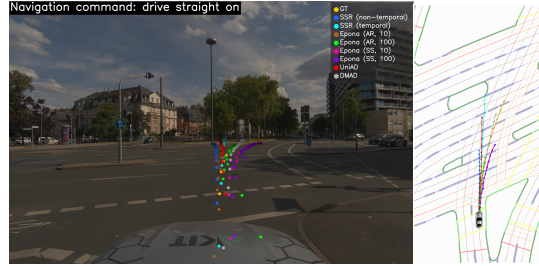


Figure 9: Qualitative end-to-end predictions, showing the front-view camera image with a top-view of all model trajectories overlaid on the HD map and ground truth. Epona tracks the road curvature better than the over-committed navigation-conditioned models, see Figure 21 for more examples.

Table 6: End-to-end results on 200 nine-second e2e samples with all metrics evaluated at the 3 s horizon. ADE and FDE follow [51]; the map-grounded metrics drivable-surface survival, collision-free rate, and centerline distance leverage our HD maps together with a lidar-based occupancy layer. ADE is additionally broken out by scene category. Best values are bold, second-best underlined.

Model	FDE@3 s ↓		ADE@3 s ↓						Survival / Tracking @3 s			
	avg.		avg.	selec.	constr.	overt.	inters.	night	nom.	Drv. surv.↑	Coll.-free↑	CL dist.↓
UniAD [38]	4.85		2.43	3.37	<u>1.96</u>	2.27	2.29	4.87	2.26	55.5	80.9	0.84
DMAD [53]	4.49		2.30	3.59	1.78	2.27	2.06	5.23	2.09	58.4	85.0	0.59
SSR [54] (non-temp.)	7.57		3.97	6.36	2.07	4.50	3.06	8.54	3.96	65.9	78.0	0.68
SSR [54] (temporal)	5.05		2.49	4.25	2.49	2.30	2.59	5.16	1.99	67.6	79.8	0.78
Epona [55] (AR, 10)	7.70		3.62	4.31	5.47	3.93	3.25	6.51	3.44	63.0	81.5	0.62
Epona [55] (AR, 100)	6.04		2.86	3.57	4.27	3.24	2.56	5.48	2.63	57.2	82.1	0.66
Epona [55] (SS, 10)	3.98		<u>1.99</u>	<u>2.71</u>	2.57	2.14	<u>1.85</u>	<u>4.43</u>	<u>1.73</u>	81.5	<u>97.7</u>	0.46
Epona [55] (SS, 100)	<u>3.99</u>		1.97	2.63	2.67	<u>2.17</u>	1.83	4.41	1.71	<u>78.6</u>	98.3	<u>0.47</u>

For SSR, *non-temp.* uses only the current keyframe whereas *temporal* aggregates BEV features across multiple frames. Epona is evaluated with single-step (SS) or autoregressive (AR) rollouts; 10 and 100 denote the number of diffusion denoising steps.

5 Limitations

Dynamic-object annotations. The current release does not include 3D bounding boxes, tracks, or instance segmentation for dynamic agents. These annotations will be added in a future release.

Dataset scale. At 5.7 h of current recorded data, KITScenes Multimodal is smaller in raw volume than recent large-scale sensor corpora such as nuPlan Sensor (≈ 120 h) or Nvidia Physical AI AV (≈ 1700 h). However, these datasets target fundamentally different tasks and provide neither the same annotation types nor comparable sensor fidelity. Progress in spatial machine learning is increasingly driven by two complementary regimes: large-scale pre-training, where data volume is central, and curated evaluation or fine-tuning data with benchmark protocols that reflect target deployment behavior. Our dataset primarily supports the latter, offering sensor fidelity, annotation completeness, and benchmark breadth that are difficult to replicate at corpus scale.

Open-loop end-to-end evaluation. While the maps are validated end-to-end through closed-loop driving trials in Autoware [10] as shown in Section F, our end-to-end benchmark evaluates open-loop trajectory prediction only. While the released artifacts enable closed-loop evaluation in the Autoware simulator, we leave such experiments to future work.

6 Conclusion

We presented KITScenes Multimodal, a European multi-modal driving dataset that pairs a state-of-the-art sensor suite with high-resolution synchronized global-shutter cameras, lidar reaching beyond

400 m, and 4D imaging radar with the most complete public HD maps of any dataset, covering 62 km² of area and validated by closed-loop autonomous-driving trials. Across our four benchmarks, online HD map construction, long-range depth estimation, novel view synthesis, and end-to-end driving, current state-of-the-art methods leave systematic capability gaps that prior datasets cannot surface, from complete map prediction at full Lanelet2 fidelity, through long-range depth and geometrically consistent novel views, to map-grounded trajectory evaluation in cluttered European urban scenes. By coupling deployment-grade maps with long-range, high-fidelity sensing, KITScenes Multimodal offers a controlled testbed for the spatial-reasoning capabilities required on the path to L4 autonomy.

References

- [1] Andreas Geiger, Philip Lenz, Raquel Urtasun, and Christoph Stiller. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. doi: 10.1177/0278364913491297.
- [2] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [3] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [4] Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R. Qi, Yin Zhou, Zoey Yang, Aurélien 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, October 2021.
- [5] Felix Fent, Fabian Kutenreich, Florian Ruch, Farija Rizwin, Stefan Juergens, Lorenz Lechermann, Christian Nissler, Andrea Perl, Ulrich Voll, Min Yan, and Markus Lienkamp. Man truckscenes: A multimodal dataset for autonomous trucking in diverse conditions. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 62062–62082. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/71ac06f0f8450e7d49063c7bfb3257c2-Paper-Datasets_and_Benchmarks_Track.pdf.
- [6] Filippo Ghilotti, Edoardo Palladin, Samuel Brucker, Adam Sigal, Mario Bijelic, and Felix Heide. Truck-drive: Long-range autonomous highway driving dataset, 2026. URL <https://arxiv.org/abs/2603.02413>.
- [7] Mina Alibeigi, William Ljungbergh, Adam Tonderski, Georg Hess, Adam Lilja, Carl Lindstrom, Daria Motorniuk, Junsheng Fu, Jenny Widahl, and Christoffer Petersson. Zenseact open dataset: A large-scale and diverse multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.
- [8] NVIDIA Corporation. PhysicalAI-Autonomous-Vehicles. <https://huggingface.co/datasets/nvidia/PhysicalAI-Autonomous-Vehicles>, oct 2025. Accessed 2026-05-06, released 2025-10-28.
- [9] Fabian Poggenhans, Jan-Hendrik Pauls, Johannes Janosovits, Stefan Orf, Maximilian Naumann, Florian Kuhnt, and Matthias Mayr. Lanelet2: A high-definition map framework for the future of automated driving. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 1672–1679, 2018. doi: 10.1109/ITSC.2018.8569929.
- [10] Autoware Foundation. Autoware. <https://github.com/autowarefoundation/autoware>. Accessed: 2026-05-02.
- [11] 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. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS Datasets and Benchmarks 2021)*, 2021.

- [12] Xinyu Huang, Peng Wang, Xinjing Cheng, Dingfu Zhou, Qichuan Geng, and Ruigang Yang. The ApolloScape Open Dataset for Autonomous Driving and Its Application. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 42(10):2702–2719, October 2020. ISSN 1939-3539. doi: 10.1109/TPAMI.2019.2926463. URL <https://doi.ieeecomputersociety.org/10.1109/TPAMI.2019.2926463>.
- [13] Jiageng Mao, Niu Minzhe, ChenHan Jiang, hanxue liang, Jingheng Chen, Xiaodan Liang, Yamin Li, Chaoqiang Ye, Wei Zhang, Zhenguo Li, Jie Yu, Chunjing XU, and Hang Xu. One million scenes for autonomous driving: Once dataset. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/67c6a1e7ce56d3d6fa748ab6d9af3fd7-Paper-round1.pdf.
- [14] Yiyi Liao, Jun Xie, and Andreas Geiger. KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *Pattern Analysis and Machine Intelligence (PAMI)*, 2022.
- [15] Huijie Wang, Tianyu Li, Yang Li, Li Chen, Chonghao Sima, Zhenbo Liu, Bangjun Wang, Peijin Jia, Yuting Wang, Shengyin Jiang, et al. Openlane-v2: A topology reasoning benchmark for unified 3d hd mapping. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.
- [16] Qi Li, Yue Wang, Yilun Wang, and Hang Zhao. Hdmapnet: An online hd map construction and evaluation framework. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 4628–4634, 2022. doi: 10.1109/ICRA46639.2022.9812383.
- [17] Yicheng Liu, Tianyuan Yuan, Yue Wang, Yilun Wang, and Hang Zhao. VectorMapNet: End-to-end vectorized HD map learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 22352–22369. PMLR, 23–29 Jul 2023.
- [18] Bencheng Liao, Shaoyu Chen, Xinggang Wang, Tianheng Cheng, Qian Zhang, Wenyu Liu, and Chang Huang. Maptr: Structured modeling and learning for online vectorized hd map construction. *arXiv preprint arXiv:2208.14437*, 2022.
- [19] Bencheng Liao, Shaoyu Chen, Yunchi Zhang, Bo Jiang, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Maptrv2: An end-to-end framework for online vectorized hd map construction. *International Journal of Computer Vision*, Oct 2024. ISSN 1573-1405. doi: 10.1007/s11263-024-02235-z. URL <https://doi.org/10.1007/s11263-024-02235-z>.
- [20] Limeng Qiao, Wenjie Ding, Xi Qiu, and Chi Zhang. End-to-end vectorized hd-map construction with piecewise bezier curve. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13218–13228, June 2023.
- [21] Wenjie Ding, Limeng Qiao, Xi Qiu, and Chi Zhang. Pivotnet: Vectorized pivot learning for end-to-end hd map construction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3672–3682, 2023.
- [22] Shuo Wang, Fan Jia, Weixin Mao, Yingfei Liu, Yucheng Zhao, Zehui Chen, Tiancai Wang, Chi Zhang, Xiangyu Zhang, and Feng Zhao. Stream query denoising for vectorized hd-map construction. In *European Conference on Computer Vision*, pages 203–220. Springer, 2024.
- [23] Jiacheng Chen, Yuefan Wu, Jiaqi Tan, Hang Ma, and Yasutaka Furukawa. Maptracker: Tracking with strided memory fusion for consistent vector hd mapping. In *European Conference on Computer Vision*, pages 90–107. Springer, 2024.
- [24] Xiaoyu Zhang, Guangwei Liu, Zihao Liu, Ningyi Xu, Yunhui Liu, and Ji Zhao. Enhancing vectorized map perception with historical rasterized maps. In *European Conference on Computer Vision*, pages 422–439. Springer, 2024.
- [25] Anqi Shi, Yuze Cai, Xiangyu Chen, Jian Pu, Zeyu Fu, and Hong Lu. Globalmapnet: An online framework for vectorized global hd map construction. *arXiv preprint arXiv:2409.10063*, 2024.
- [26] Dapeng Zhang, Dayu Chen, Peng Zhi, Yinda Chen, Zhenlong Yuan, Chenyang Li, Rui Zhou, Qingguo Zhou, et al. Mapexpert: Online hd map construction with simple and efficient sparse map element expert. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 14745–14753, 2025.
- [27] Jing Yang, Sen Yang, Xiao Tan, and Hanli Wang. Histrackmap: Global vectorized high-definition map construction via history map tracking. *arXiv preprint arXiv:2503.07168*, 2025.

- [28] Fatih Erdoğan, Merve Rabia Barın, and Fatma Güney. Mapping like a skeptic: Probabilistic bev projection for online hd mapping. *arXiv preprint arXiv:2508.21689*, 2025.
- [29] Fabian Immel, Richard Fehler, Frank Bieder, and Christoph Stiller. Generation of training data from hd maps in the lanelet2 framework. *arXiv preprint arXiv:2407.17409*, 2024. URL <https://arxiv.org/abs/2407.17409>.
- [30] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [31] Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.
- [32] Girish Chandar Ganesan, Yuliang Guo, Liu Ren, and Xiaoming Liu. Unidac: Universal metric depth estimation for any camera, 2026. URL <https://arxiv.org/abs/2603.27105>.
- [33] Zirui Wu, Tianyu Liu, Liyi Luo, Zhide Zhong, Jianteng Chen, Hongmin Xiao, Chao Hou, Haozhe Lou, Yuantao Chen, Runyi Yang, Yuxin Huang, Xiaoyu Ye, Zike Yan, Yongliang Shi, Yiyi Liao, and Hao Zhao. Mars: An instance-aware, modular and realistic simulator for autonomous driving. In Lu Fang, Jian Pei, Guangtao Zhai, and Ruiping Wang, editors, *Artificial Intelligence*, pages 3–15, Singapore, 2024. Springer Nature Singapore. ISBN 978-981-99-8850-1.
- [34] 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. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=yvc2z8TYur>.
- [35] 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 *European Conference on Computer Vision*, pages 156–173. Springer, 2024.
- [36] Ziyu Chen, Jiawei Yang, Jiahui Huang, Riccardo de Lutio, Janick Martinez Esturo, Boris Ivanovic, Or Litany, Zan Gojic, Sanja Fidler, Marco Pavone, Li Song, and Yue Wang. Omnire: Omni urban scene reconstruction. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [37] Haibao Yu, Kuntao Xiao, Jiahang Wang, Ruiyang Hao, Guoran Hu, Yuxin Huang, Haifang Qin, Bowen Jing, Yuntian Bo, and Ping Luo. Recondrive: Fast feed-forward 4d gaussian splatting for autonomous driving scene reconstruction. In <https://arxiv.org/abs/2603.07552>, 2026.
- [38] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17853–17862, 2023.
- [39] Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Vad: Vectorized scene representation for efficient autonomous driving. *ICCV*, 2023.
- [40] Zoltan Szabadka, Martin Bruse, and Jyrki Alakuijala. Introducing Jpegli: A new JPEG coding library. Google Open Source Blog, April 2024. URL <https://opensource.googleblog.com/2024/04/introducing-jpegli-new-jpeg-coding-library.html>. Accessed: 2026-05-01.
- [41] K. Tan et al. H. Caesar, J. Kabzan. NuPlan: A closed-loop ml-based planning benchmark for autonomous vehicles. In *CVPR ADP3 workshop*, 2021.
- [42] John Lambert and James Hays. Trust, but verify: Cross-modality fusion for hd map change detection. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran, 2021.
- [43] Miriam Louise Carnot, Erik Fastermann, Jonas Kunze, Eric Peukert, André Ludwig, and Bogdan Franczyk. Gtsign-220: A crowd-sourced, stvo-aligned benchmark for fine-grained german traffic sign recognition. In *Intelligent Vehicles Symposium (IV)*, 2026.
- [44] Jan-Hendrik Pauls, Benjamin Schmidt, and Christoph Stiller. Automatic mapping of tailored landmark representations for automated driving and map learning. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6725–6731, 2021. doi: 10.1109/ICRA48506.2021.9561432.

- [45] Zihao Liu, Xiaoyu Zhang, Guangwei Liu, Ji Zhao, and Ningyi Xu. Leveraging enhanced queries of point sets for vectorized map construction. In *European Conference on Computer Vision*, 2024.
- [46] Fabian Immel, Jan-Hendrik Pauls, Richard Fehler, Frank Bieder, Jonas Merkert, and Christoph Stiller. SDTagnet: Leveraging text-annotated navigation maps for online HD map construction. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL <https://openreview.net/forum?id=N3E1cU8Cv3>.
- [47] Tianyuan Yuan, Yicheng Liu, Yue Wang, Yilun Wang, and Hang Zhao. Streammapnet: Streaming mapping network for vectorized online hd map construction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7356–7365, 2024.
- [48] Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, Jonathon Luiten, Manuel Lopez-Antequera, Samuel Rota Bulò, Christian Richardt, Deva Ramanan, Sebastian Scherer, and Peter Kotschieder. MapAnything: Universal feed-forward metric 3D reconstruction. In *International Conference on 3D Vision (3DV)*. IEEE, 2026.
- [49] 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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1389–1399, Los Alamitos, CA, USA, June 2023. IEEE Computer Society. doi: 10.1109/CVPR52729.2023.00140. URL <https://doi.ieeecomputersociety.org/10.1109/CVPR52729.2023.00140>.
- [50] Chaojun Ni, Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Wenkang Qin, Guan Huang, Chen Liu, Yuyin Chen, Yida Wang, Xueyang Zhang, et al. Recondreamer: Crafting world models for driving scene reconstruction via online restoration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1559–1569, 2025.
- [51] Alexandre Alahi, Kratharth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [52] Royden Wagner, Omer Sahin Tas, Jaime Villa, Felix Hauser, Yinzhe Shen, Marlon Steiner, Dominik Strutz, Carlos Fernandez, Christian Kinzig, Guillermo S. Guitierrez-Cabello, Hendrik Königshof, Fabian Immel, Richard Schwarzkopf, Nils Alexander Rack, Kevin Rösch, Kaiwen Wang, Jan-Hendrik Pauls, Martin Lauer, Igor Gilitschenski, Holger Caesar, and Christoph Stiller. Longtail driving scenarios with reasoning traces: The kitscenelongtail dataset, 2026. URL <https://arxiv.org/abs/2603.23607>.
- [53] Yinzhe Shen, Omer Şahin Tas, Kaiwen Wang, Royden Wagner, and Christoph Stiller. Divide and merge: Motion and semantic learning in end-to-end autonomous driving. *Transactions on Machine Learning Research*, 2025(11), 2025.
- [54] Peidong Li and Dixiao Cui. Navigation-guided sparse scene representation for end-to-end autonomous driving. In *International Conference on Learning Representations (ICLR)*, 2025.
- [55] Kaiwen Zhang, Zhenyu Tang, Xiaotao Hu, Xingang Pan, Xiaoyang Guo, Yuan Liu, Jingwei Huang, Li Yuan, Qian Zhang, Xiao-Xiao Long, Xun Cao, and Wei Yin. Epona: Autoregressive diffusion world model for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.
- [56] RawTherapee Development Team. RawTherapee: A powerful cross-platform raw photo processing program. URL <https://github.com/RawTherapee/RawTherapee>. Includes the AMaZE demosaicing algorithm and raw-domain chromatic aberration correction by E. J. Martinec.
- [57] Martin Bruse, Luca Versari, Zoltan Szabadka, and Jyrki Alakuijala. Users prefer jpegli over same-sized libjpeg-turbo or mozjpeg, 2024. URL <https://arxiv.org/abs/2403.18589>.
- [58] brighter AI Technologies. Face off: Privacy v progress — how deep natural anonymization protects privacy in the age of machine learning. White paper, brighter AI Technologies GmbH, Berlin, Germany, 2022. URL <https://ac-landing-pages-user-uploads-production.s3.amazonaws.com/0000122471/803bb7a7-de73-4596-9548-6d1ca3a80e32.pdf>.
- [59] Xiting Zhao and Sören Schwertfeger. 3dref: 3d dataset and benchmark for reflection detection in rgb and lidar data. In *2024 International Conference on 3D Vision (3DV)*, pages 225–234, 2024. doi: 10.1109/3DV62453.2024.00009.

- [60] Tiziano Guadagnino, Benedikt Mersch, Saurabh Gupta, Ignacio Vizzo, Giorgio Grisetti, and Cyrill Stachniss. Kiss-slam: A simple, robust, and accurate 3d lidar slam system with enhanced generalization capabilities. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5363–5370. IEEE, 2025.
- [61] Tobias Strauß, Julius Ziegler, and Johannes Beck. Calibrating multiple cameras with non-overlapping views using coded checkerboard targets. In *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, pages 2623–2628, 2014. doi: 10.1109/ITSC.2014.6958110.
- [62] Johannes Beck and Christoph Stiller. Generalized b-spline camera model. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 2137–2142, 2018. doi: 10.1109/IVS.2018.8500466.
- [63] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, RISHI HAZRA, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollar, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. In *The Fourteenth International Conference on Learning Representations*, 2026. URL <https://openreview.net/forum?id=r35c1VtGzw>.
- [64] JOSM. Java OpenStreetMap Editor. <https://josm.openstreetmap.de/>, 2026. Accessed: 01.05.2026.
- [65] Mapillary. Mapillary. <https://www.mapillary.com/app>, 2026. Street-level imagery platform. Accessed: 2026-05-04.
- [66] Steven Macenski, Tully Foote, Brian Gerkey, Chris Lalancette, and William Woodall. Robot operating system 2: Design, architecture, and uses in the wild. *Science Robotics*, 7(66):eabm6074, 2022. doi: 10.1126/scirobotics.abm6074. URL <https://www.science.org/doi/abs/10.1126/scirobotics.abm6074>.
- [67] Adam Lilja, Junsheng Fu, Erik Stenborg, and Lars Hammarstrand. Localization is all you evaluate: Data leakage in online mapping datasets and how to fix it. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22150–22159, 2024.
- [68] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. *Advances in Neural Information Processing Systems*, 36:72983–73007, 2023.

A Details on the Sensor Setup

Tables 7 to 10 describe our sensor setup in detail, with a real-world picture of it shown in Figure 10.

Table 7: **Camera setup.** All cameras are manufactured by Lucid Vision Labs and use low-distortion Fujinon CF8ZA-1S-23M lenses with 23 Mpx maximum resolution.

	Surround	Stereo (tilted)	Hi-Res/Long-range
Count	6	1 pair	1
Camera	ATL071S-CC	ATL071S-CC + ATL071S-MC	ATP162S-CC
Sensor	Sony IMX420, 1.1"	Sony IMX420, 1.1"	Sony IMX542, 1.1"
Resolution	3200 × 2200 (7.1 Mpx)	3200 × 2200 (7.1 Mpx)	5320 × 3032 (16.2 Mpx)
Pixel pitch	4.5 μm	4.5 μm	4.5 μm
FOV (H×V)	87.1°×63.3°	63.3°×86.9°	88.4°×54.4°

Table 8: **Lidar setup.** Per-unit specifications for the four lidar groups in the sensor suite.

	Top (–Dec. 2025)	Top (Jan. 2026–)	Corner (tilted)	Automotive
Count	1	1	2	4
Model	Velodyne VLS128-AP	Hesai OT128	Hesai XT32	Seyond Falcon K1
Channels	128	128	32	150 lines
FOV (H×V)	360°×40°	360°×40°	270°×31°	120°×25°
Resolution (H×V)	0.2°×0.1°	0.1°×0.125°	0.18°×1.3°	0.18°×0.24°
Range (max)	245 m	230 m	120 m	500 m
Range @ 10 % refl.	245 m	200 m	80 m	250 m
Wavelength	905 nm	905 nm	905 nm	1550 nm
Effective Points/s	2.19 M	6.91 M	864 k	900 k
Returns	strongest	last + strongest	last + strongest	strongest

The top lidar was improved in December 2025. FoV and effective points of corner lidars are intentionally limited.

Table 9: **Radar setup.** Specifications of the three Continental ARS548 RDI 4D imaging radars.

	Long-range
Count	3
Model	Continental ARS548 RDI
Frequency band	76 GHz to 77 GHz
FOV (H×V)	120°×28°
Beam width (3 dB, H×V)	1.2°×2.3°
Angular accuracy (H×V)	±0.1°×±0.1°
Range (max)	300 m
Range resolution	0.22 m
Velocity range	–400 to 200 km h ^{–1}
Output	4D detections + RCS

Table 10: **GNSS and GNSS/INS setup.** We combine two receivers with independent antennas.

	GNSS	GNSS/INS
Model	Septentrio mosaic-X5	Septentrio AsteRx SBI3 Pro+
Antennas	1	2
Hardware channels	448	544
RTK accuracy (H/V)	0.6 cm / 1.0 cm	0.6 cm / 1.0 cm
Standalone accuracy (H/V)	1.2 m / 1.9 m	1.2 m / 1.9 m
Heading accuracy (RTK)	—	0.2°
Pitch/roll accuracy (RTK)	—	0.02°
Velocity accuracy	3 cm s ^{–1}	2 cm s ^{–1} (RTK)
Position update rate	100 Hz	10 Hz (integrated)
Integrated IMU	—	ADIS16500



Figure 10: The KITScenes recording vehicle with the sensor setup as roofmount.

A.1 Sensor Data Processing and Privacy

To enable both long-range perception and neural rendering applications, it is crucial to preserve the high raw image fidelity. We record raw Bayer images and employ high-quality offline debayering pipeline using AMAZE with chromatic aberration correction [56]. Images are then compressed using JPEGLI [40], a JPEG-compatible codec, with 4:4:4 chroma subsampling at Q95, yielding visually lossless image quality at manageable file sizes [57]. To comply with European privacy regulations, all faces and license plates are anonymized using DNAT, a state-of-the-art inpainting method that preserves photometric realism better than conventional blurring approaches [58].

For the 360° main lidar data, we deliberately preserve non-return information to serve as additional information for occupancy tasks. Additionally, all Hesai lidars return dual echoes, which carry valuable information about reflective surfaces and in adverse-weather conditions [59]. The four Seyond lidars furthermore provide elongation information of each return.

Radar and GNSS. Three 4D imaging radars complement the lidar suite, providing Doppler velocity measurements and resilience under adverse weather conditions where lidar performance degrades. A redundant combination of one GNSS receiver and one combined GNSS-INS unit provides a high-accuracy localization reference used for map validation and as a SLAM reference.

Localization / SLAM. To achieve reprojection level accuracy of georeferenced 6-DoF poses, we fuse the position data of the redundant RTK GNSS sensors into a modified version of KISS-SLAM [60] which we plan to publish.

B Calibration Details

Camera calibration is significantly facilitated by using hardware-triggered global shutter cameras with low-distortion lenses. Intrinsic and extrinsic camera calibration is performed using checkerboard targets [61] and a reference camera model [62]. A pinhole model is fitted to this reference model at subpixel accuracy.

From the mechanical construction, translation and orientation are known all sensors up to few degrees and centimeters. The remaining refinement hence focuses on the angular error that dominates at the long perception ranges that we tackle and which is best observed using far-range natural surroundings rather than close-by targets.

To avoid motion artifacts, we select one reference frame per standstill phase. Lidar-to-lidar calibration is then formulated as joint ICP problem across all sensors and frames. Using the same standstill frames, radar points are registered to the joint lidar point cloud using ICP.

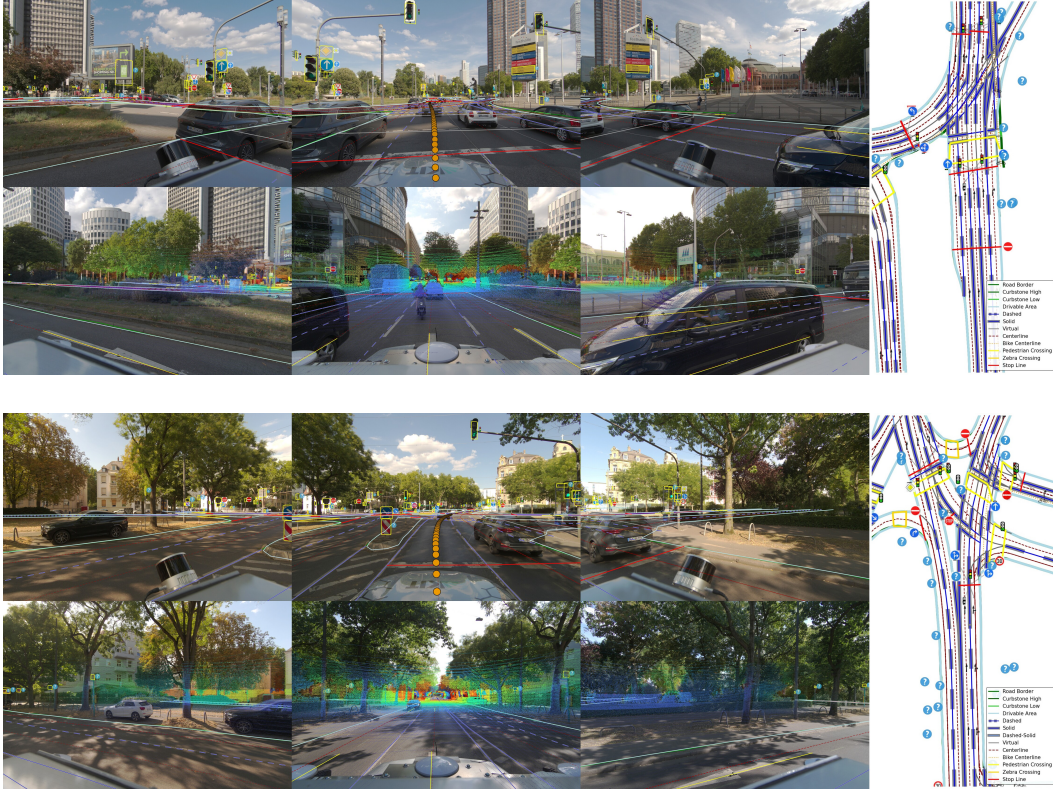


Figure 11: Example frames of a scenario in Frankfurt. The reprojected traffic lights and signs can be seen from multiple camera reprojections, highlighting the 3D nature of our HD map. The lidar data is reprojected with a transparency fade-out. The reprojected map element colors encode different class types. Icons representing class labels are transparently overlaid near the reprojected map element.

Finally, we align the lidar and camera rigs by maximizing the reprojection of retro-reflective lidar points on semantically segmented [63] traffic signs and license plates using differentiable splatting. This lidar-to-camera calibration framework will be made available as open source post submission.

C Additional Sample Data Visualization

To give further insights into the sensor and annotation data we provide additional samples visualizations in Figure 11

D Data Collection Routes and Conditions

The vehicle (Figure 10) was human-driven by trained operators. All scenes are manually selected to ensure high annotation and localization quality while capturing diverse traffic scenarios and map layouts.

E Annotation Protocol and Quality Control

Annotation Protocol and Quality Control. Annotation is performed by an in-house team of workers within 10 000 total working hours ("approx. 160 hours per km²). Annotation tooling extends the Java OpenStreetMap Editor [64] with tooling we developed for Lanelet2-native primitive creation, spline interpolation, routing-graph visualization, topology editing; this tooling and our class label presets will be released open-source. Aerial imagery is sourced from municipal mapping authorities with up to 6 cm ground sampling resolution, exceeding the resolution of public products; imagery

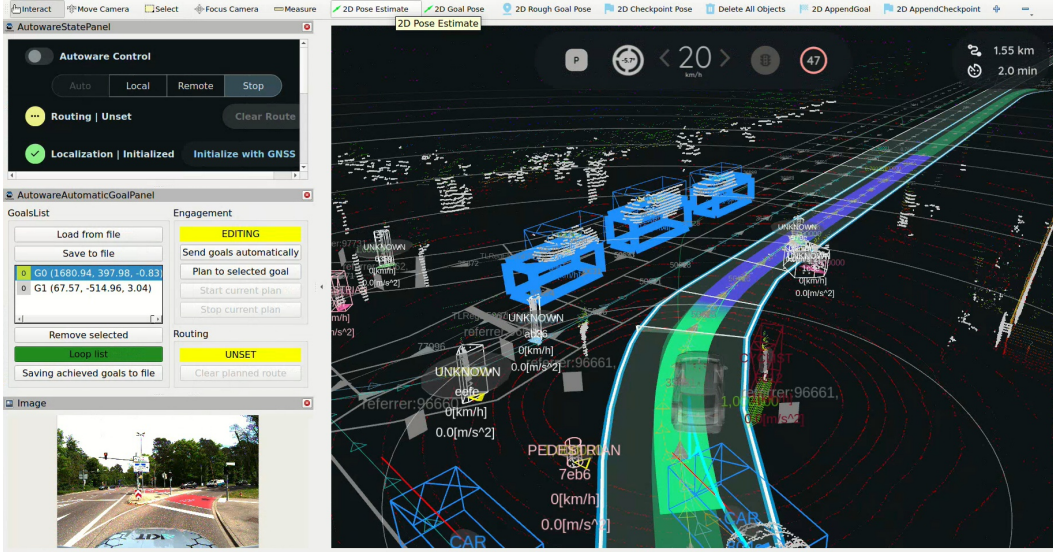


Figure 12: The open-source autonomous driving stack Autoware [10] driving closed-loop on one of the KITScenes Lanelet2 [9] HD maps

from 2023-2024 is used and is validated frame-by-frame against the 2025/26 sensor recordings to detect map changes and adapt or exclude them for our dataset release.

Map creation is split into two complementary annotation passes. Road-level content, like lane geometry, road markings, lane topology, crosswalks, as well as BEV traffic light and sign positions, are annotated from aerial imagery, which provides a geo-referenced, occlusion-free top-down view. Additionally, we leverage crowdsourced streetlevel imagery [65] to resolve ambiguities. Elevated objects, like traffic lights, road signs, and poles, are additionally localized directly from geo-referenced onboard sensor data, yielding 3D shapes including orientation that are reprojection-accurate to the calibrated camera images [44]. Both annotation layers are fused into a unified map representation and manually reviewed for geometric and semantic correctness. The resulting maps are further validated by automated structural-consistency checks, including Lanelet2-based topology checks and application tests in Autoware’s planning simulation.

No region is annotated by a single person across all stages: annotators rotate between geometric drafting, attribute classification, and topology linking, providing implicit cross-validation and reducing systematic per-annotator artifacts. Quality control combines (i) Lanelet2 core-logic validators, (ii) geometric and point/line integrity checks, (iii) relational and topological completeness checks including a routing-graph orphan check, and (iv) an in-house aerial-image polyline-attribute classifier that is run on every annotated polyline at QA time to flag outliers and likely tag errors against the source aerial imagery.

F Closed-loop autonomous driving map verification trials

Following formal test-suite validation and simulation-based verification, closed-loop driving trials constitute the final stage of HD map qualification. To facilitate adoption by the research community and to demonstrate the operational readiness of our maps, we validate their compatibility with Autoware [10], an internationally adopted open-source autonomous driving stack that serves as the reference software platform for robotaxi deployments in Japan and beyond. A representative closed-loop trial scene, with the Lanelet2 map visualized in RViz within a ROS 2 [66] environment, is depicted in Figure 12.

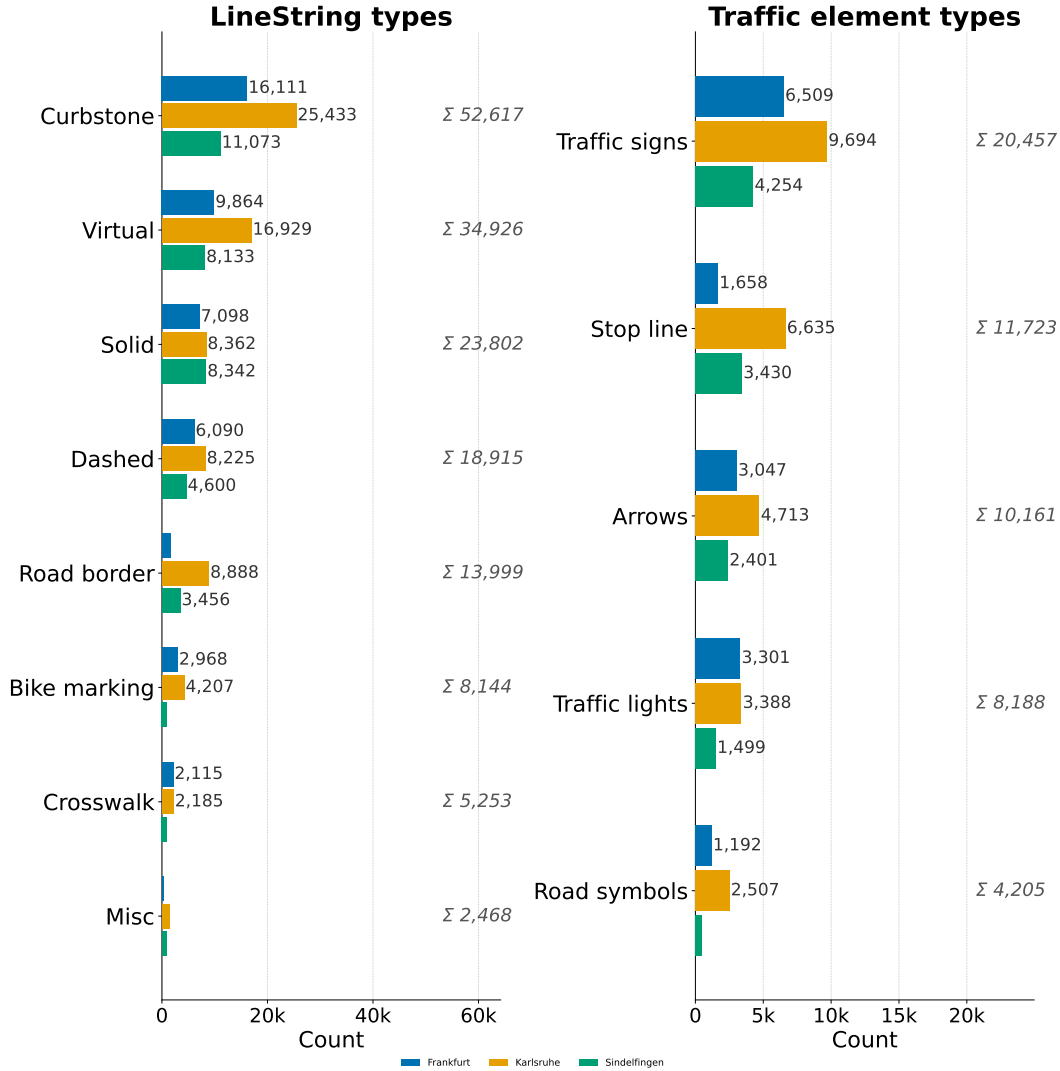


Figure 13: Binned label category statistics over Lanelet2 map elements. The left plot covers elements on the road surface as 3D polylines and the right side covers 3D traffic lights and signs as well as other lines usually not part of a lanelet border.

G Extended Dataset Statistics

G.1 Splits

A persistent weakness of previous HD map construction method evaluation protocols is that training and validation regions often overlap geographically, allowing models to implicitly memorize map priors and inflate reported performance [67, 47].

To close this loophole, we adopt a manually selected geographic split via specifically selected polygon regions of complex road layout areas for validation and test split. We greedily assign scenarios with poses in these polygon areas to the test and consecutively validation split and compute the distance between each pose of each scenario, allowing no overlap between scenario pairs including a test scenario up to 100 m and pairs including a val scenario up to 70 m distance. The result is a strict geo-disjoint train/val/test boundary with no map overlap guaranteed across all three splits, visualized in Figure 14. A unique aspect of our benchmark strategy is to withhold all map data in the test split. This is the first such held-out test set in the online HD map perception space and allows for truly trustworthy method comparisons in leaderboard challenges.

Table 11: Default Splits of the KITScenes Multimodal dataset. A distance threshold of 100 m to test scenarios and 70 m to val scenarios guarantees geographically separate evaluation.

Split	Scenarios	Hours	Path (km)	Released Data
train	534	3.00	81.0	all
val	117	0.60	14.5	all
overlap-train-val	23	0.13	3.3	all
test	206	1.17	30.1	no maps
test-e2e	127	0.76	33.1	no: maps, geo-pose, data after keyframe
total	1007	5.66	162.0	

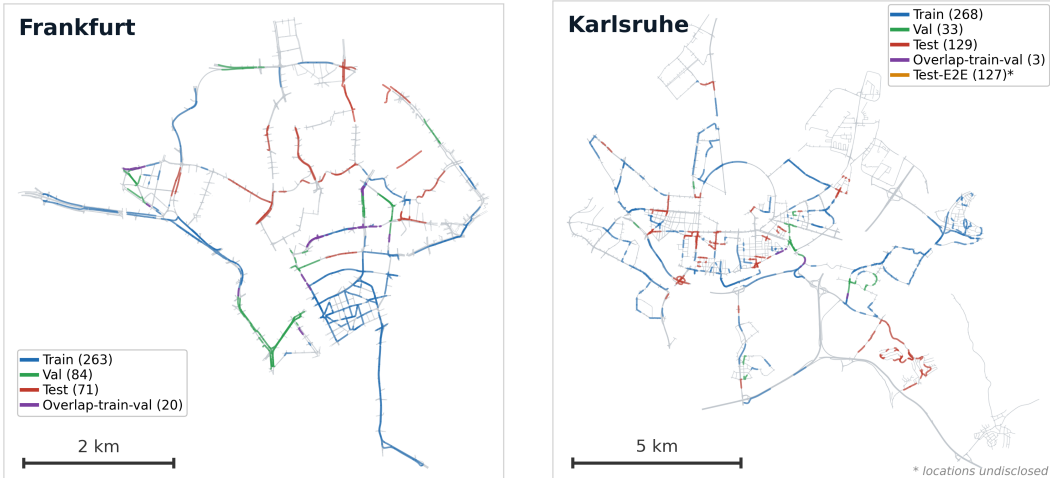


Figure 14: Visualization of the split Definition for two KITScenes cities. The color indicates the split bucket of a scenario.

The additional test-e2e split provides only local-frame poses, no maps, and no future poses and sensor data after the keyframe of end-to-end predictions and is intended for held-out end-to-end driving evaluation.

A minor additional split called overlap-train-val is published comprising scenes with geographic overlap with the train and val sets. It joins our validation benchmark protocol for the long-range depth, novel view synthesis, and end-to-end benchmarks (Sections 4.2–4.4), but is excluded from online HD map perception evaluation (Section 4.1), which requires strict train/val geo-separation.

H Per-Benchmark Details

H.1 Online HD Map Construction

MapQR-Topo architecture. The MapQR-Topo head proposed in Section 4.1 consumes the map element tokens from the decoder and predicts pairwise relations between all predicted map elements. Predicted relations are evaluated by matching each predicted map element to its nearest ground-truth counterpart via the Hungarian algorithm, with Chamfer distance as the cost function, and reporting a topology AP score (AP_{Topo}) computed over the predicted and ground-truth edges. Contrary to the topology metric proposed by OpenLane-v2 [15] ($TOP_{ll} + TOP_{lt}$), this metric design allows untangled evaluation of detection and topology prediction performance since we directly utilize the Hungarian matching algorithm already used for assigning predicted elements to GT elements. This avoids chamfer-distance thresholds for positive matching pairs, but instead computes the globally optimal one-to-one matching. Due to this change, even elements that fall out of the typical detection thresholds of 0.5, 1.0 and 1.5 meters can be successfully evaluated with respect to their topology, when connections are predicted in agreement with the ground truth.



Figure 15: HD map outlines for our maps in the four cities of Frankfurt, Karlsruhe, Sindelfingen and Stuttgart.

An overview of the adapted architecture is shown in Figure 16 and a qualitative example of the predictions in Figure 17.

Setup. We employ the split described in Section G.1, ensuring evaluation on previously unseen map areas. An included converter [29] translates Lanelet2 maps to and from the polyline instance graph representation used by state-of-the-art map perception models [19], making it straightforward to apply and evaluate existing methods on our benchmark. For our benchmark, we use a subset of 120 out of 220 available traffic sign classes, retaining only the most common and semantically relevant signs. All methods are trained for 6 epochs in line with standard evaluation settings on comparable dataset sizes [19].

Metrics. We follow the standard protocol of [19], reporting Average Precision (AP) with Chamfer distance thresholds of 0.5 m, 1.0 m, and 1.5 m, averaged across all map element classes. To preserve readability, we group map element classes into six categories: Lane Markings (LM), Lane Centerlines (LC), Road Infrastructure (RI), Traffic Lights (TL), Traffic Signs (TS), and Road Markings (RM); the

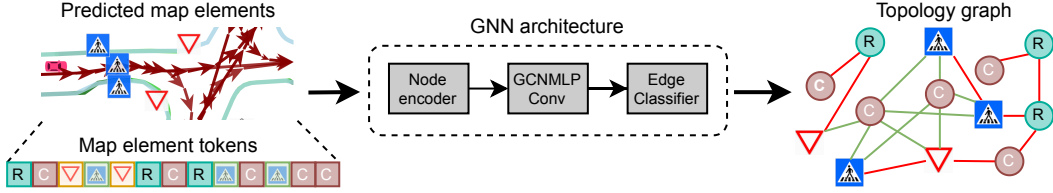


Figure 16: Schematic overview of the topology prediction with a GNN for the map elements road border (R), centerline (C) and traffic signs. Green links indicate positive link predictions, while red edges show sampled negatives, included to counteract the class imbalance introduced by the predominance of missing links.

full assignment is given in Table 16. For topology evaluation, we match each predicted map element to its nearest ground-truth counterpart via the Hungarian algorithm with Chamfer distance as the cost function, and report a topology AP score (AP_{Topo}) computed over the predicted and ground-truth edges.

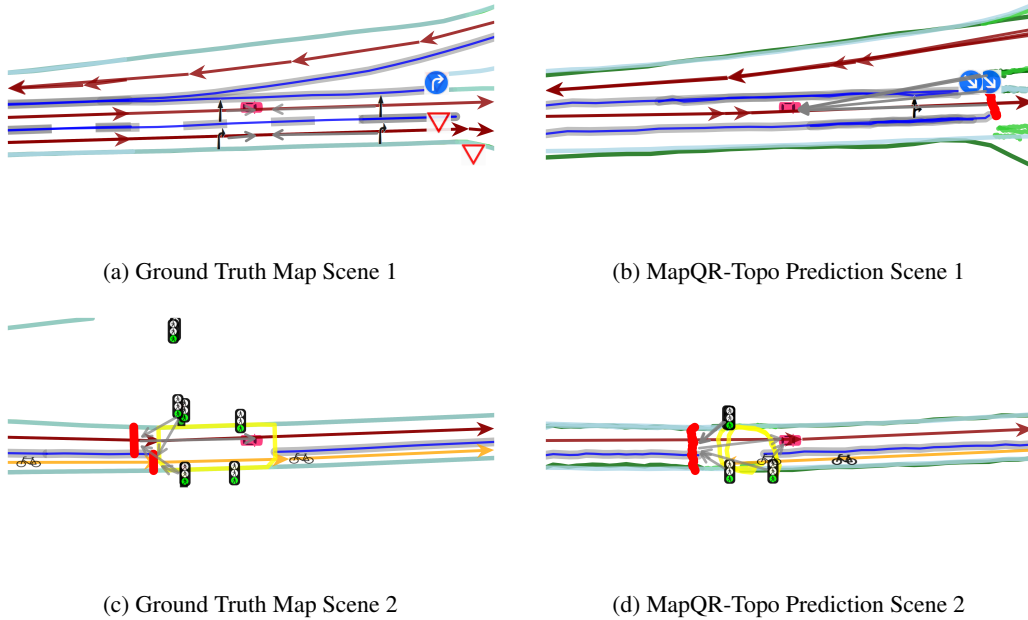


Figure 17: Ground truth HD maps (left) and MapQR-Topo predictions (right) for two scenes. Grey edges denote topology links.

H.2 Long-range Monocular Depth Estimation

Setup. As RGB input, all methods use our front-facing 16.2 Mpx high-resolution long-range camera detailed in Section 3.1. We used the best self-reported pretrained weights for [32, 31, 48]. Outputs are upsampled to camera resolution where necessary. Ground-truth depth maps are produced by fusing motion-compensated lidar point clouds (all sensors from Section 3.1) over a ± 1 s temporal window. The merged cloud is projected onto the camera plane at $2\times$ super-resolution using point splatting, and outliers are removed via MAD-based consistency rejection and edge erosion at depth discontinuities. The super-resolved map is then block-median downsampled to native resolution. Only pixels with valid depth values are used for evaluation. Frames are sampled at 0.1 Hz (one per 10 s), each placed at the temporal midpoint of its 10 s window.

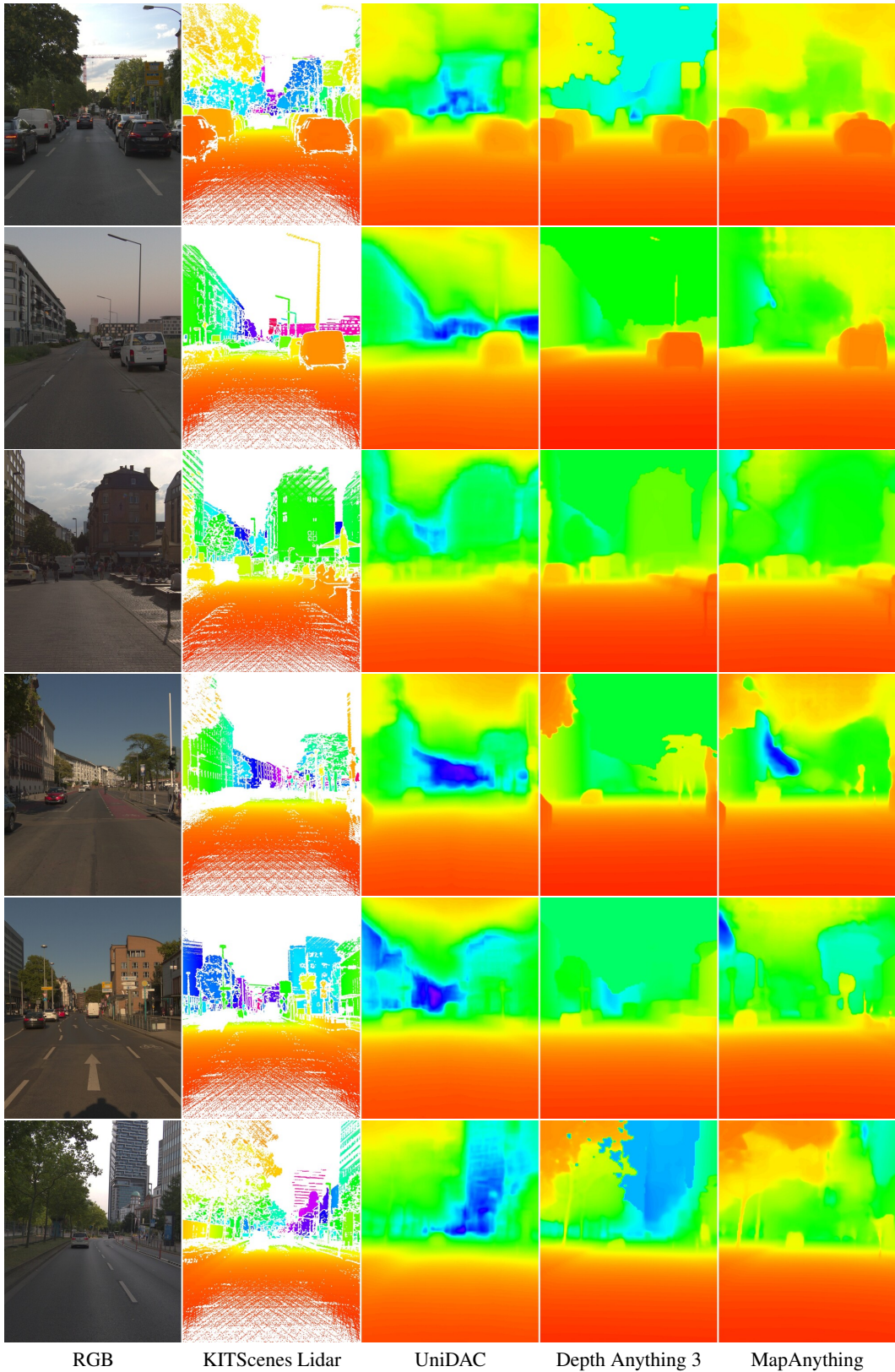


Figure 18: Qualitative comparison of monocular depth estimation methods. The corresponding non-linear depth scale is introduced in Figure 6.

Table 12: Monocular depth estimation — 0–100 m

Method	Abs.Rel ↓	Sq.Rel ↓	MAE [m] ↓	RMSE [m] ↓	IMAE ↓	IRMSE ↓	LogMAE ↓	LogRMSE ↓	SILog ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
UniDAC	0.386	0.220	6.331	8.867	0.030	0.037	0.330	0.365	0.203	24.12	85.24	96.19
DA3	0.278	0.223	6.000	10.546	0.030	0.047	0.290	0.362	0.251	48.64	80.36	92.95
MapAny	0.149	0.103	3.525	7.628	0.014	0.033	0.149	0.245	0.227	83.04	92.26	96.90

Table 13: Monocular depth estimation — 100–200 m

Method	Abs.Rel ↓	Sq.Rel ↓	MAE [m] ↓	RMSE [m] ↓	IMAE ↓	IRMSE ↓	LogMAE ↓	LogRMSE ↓	SILog ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
UniDAC	0.302	0.136	40.743	46.288	0.006	0.007	0.423	0.477	0.215	40.17	63.92	79.02
DA3	0.472	0.267	63.226	67.389	0.010	0.011	0.728	0.770	0.222	12.32	30.16	49.64
MapAny	0.485	0.286	65.447	71.390	0.011	0.014	0.768	0.847	0.302	16.34	27.24	42.23

Table 14: Monocular depth estimation — > 200 m

Method	Abs.Rel ↓	Sq.Rel ↓	MAE [m] ↓	RMSE [m] ↓	IMAE ↓	IRMSE ↓	LogMAE ↓	LogRMSE ↓	SILog ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
UniDAC	0.540	0.320	139.066	145.433	0.007	0.007	0.864	0.894	0.205	1.78	14.90	37.70
DA3	0.689	0.497	175.246	179.841	0.012	0.013	1.284	1.302	0.188	0.86	2.73	12.70
MapAny	0.772	0.607	195.620	199.949	0.018	0.019	1.578	1.601	0.241	0.03	1.47	1.91

Table 15: Monocular depth estimation — Overall

Method	Abs.Rel ↓	Sq.Rel ↓	MAE [m] ↓	RMSE [m] ↓	IMAE ↓	IRMSE ↓	LogMAE ↓	LogRMSE ↓	SILog ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
UniDAC	0.384	0.218	7.194	12.240	0.029	0.037	0.332	0.371	0.226	24.36	84.73	95.77
DA3	0.282	0.225	7.336	15.300	0.029	0.046	0.300	0.378	0.267	47.91	79.32	92.03
MapAny	0.156	0.107	4.988	14.265	0.014	0.033	0.163	0.278	0.260	81.70	90.94	95.75

H.3 Novel View Synthesis

Setup. The reconstructed scene is re-rendered from the front-facing camera at seven lateral offsets $\Delta y \in \{-3, -2, -1, 0, +1, +2, +3\}$ m in the ego frame. Ground-truth traffic signs are projected from the scenario’s Lanelet2 HD map into each shifted viewpoint. To ensure a fair comparison, we apply lidar-based occlusion filtering: a sign is considered visible only if the density of lidar points (laterally shifted by the same offset) does not indicate a foreground occlusion. Detections in the rendered views are obtained with OWLv2 [68] at a confidence threshold of 0.15 and matched against the projected GT bounding boxes using an IoU threshold of 0.5.

Metrics. We report traffic sign recall, defined as the ratio of detected visible GT signs to the total number of visible GT signs. Evaluation is performed at two scales: a *low* resolution (280×518) matching the model’s typical output scale, and a *high* resolution (1600×2844) corresponding to the native sensor imagery after cropping out the ego-vehicle and sensor hardware. Since the model renders at the lower scale, the *high* evaluation uses its output bilinearly upsampled to the cropped sensor resolution. The detector’s performance on the real photograph (“Photo”) at each scale serves as the upper bound, and we evaluate a single frame every 10 s following the protocol of Section 4.2.

Qualitative Lateral NVS Results. We further showcase qualitative results on the map-based NVS evaluation benchmark proposed in Section 4.3. Figure 19 illustrates the impact of lateral translation on structural fidelity. We observe that while the reconstruction at the driven trajectory (Figure 19e, Figure 19m) shows reasonable alignment with the HD map projections, lateral shifts reveal significant geometric inaccuracies. Specifically, the traffic signs rendered in the novel views synthesized by ReconDrive do not fully align with the ground-truth projections. This misalignment, which becomes more pronounced as the lateral offset increases, suggests that the underlying geometry lacks the precision required for consistent projection at novel viewpoints. These artifacts degrade the object’s visual signature, causing the detector to miss signs that are clearly visible in the ground-truth photograph. This highlights that photometric consistency on the training distribution does not guarantee geometric accuracy in novel spatial views, a gap that is critical for safety-oriented simulation.

Photometric NVS Benchmarks. We complement the map-based evaluation with two photometric NVS benchmarks that measure complementary aspects of reconstruction quality. Full quantitative results for these benchmarks are summarized in Table 17. The first benchmark, *Held-out cross-mount NVS*, uses the six ring cameras as model inputs, while the independently mounted 18 Mpx camera_base_front_center provides the novel-view target at the same instant. This benchmark measures spatial extrapolation to a viewpoint the method never observes as input. The second

Table 16: Assignment of the map element classes to the six aggregated reporting categories used in Table 3.

Lane Markings (LM)	Lane Centerlines (LC)	Road Infrastructure (RI)
road_border	centerline	curbstone_high
dashed	bike_centerline	curbstone_low
solid		fence
solid_solid		building
solid_dashed		wall
dashed_solid		guard_rail
bike_marking_dashed		drivable_area
bike_marking_solid		divider
pedestrian_crossing		
zebra_crossing		
te_stop_line		

Traffic Lights (TL)	Traffic Signs (TS)	Road Markings (RM)
te_tl_car	te_ts_stop	te_arrow_go_straight
te_tl_bike	te_ts_yield	te_arrow_turn_left
te_tl_pedestrian	te_ts_no_entry	te_arrow_turn_right
te_tl_misc	te_ts_right_of_way	te_arrow_go_straight_or_left
	te_ts_priority_road	te_arrow_go_straight_or_right
	te_ts_one_way_street	te_arrow_turn_left_or_right
	te_ts_roundabout	te_bike_symbol
	te_ts_speed_limit	te_bus_symbol
	te_ts_pedestrian_crossing	te_symbol30
	te_ts_turn_right	te_symbol50
	te_ts_turn_left	te_symbol70
	te_ts_go_straight	
	te_ts_go_straight_or_right	
	te_ts_go_straight_or_left	
	te_ts_turn_left_or_right	
	te_ts_pass_right	
	te_ts_pass_left	

Table 17: ReconDrive evaluated on the KITScenes NVS benchmark (140 sequences, 216 windows). **Top:** photometric quality on three protocols. **Bottom:** traffic-sign recall on the front camera at seven lateral offsets; “Photo” is the detector’s recall on the real photograph (upper bound). \uparrow/\downarrow denote higher/lower is better.

Held-out Cam NVS			Ego Recon			Ego NVS		
PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
23.51	0.783	0.318	32.42	0.951	0.073	22.61	0.678	0.352

benchmark, *Ego-trajectory NVS*, uses frames t and $t + 6$ as context inputs, and frames 1–5 as interpolation targets, measuring temporal NVS quality at training-distribution viewpoints following [37]. We distinguish between *recon* (frame 0, reconstruction fidelity) and *nvs* (frames 1–5, temporal interpolation).

H.4 End-to-End Driving

Evaluation sample construction. The 200 e2e samples are drawn from the $\text{val} \cup \text{overlap-train-val}$ scenes; each is a non-overlapping 9s window of 4s past observation and up to 5s of future trajectory anchored at a keyframe. Poses are released in a local frame only, preventing geo-referenced retrieval against external maps. Headline numbers use the 3s horizon for parity with nuScenes- and nuPlan-trained baselines; the full 5s protocol is offered as the long-horizon challenge.

Map-grounded metric definitions. *Drivable-surface survival* is the fraction of predicted waypoints lying inside the union of drivable Lanelet2 polygons. *Centerline distance* is the mean lateral offset

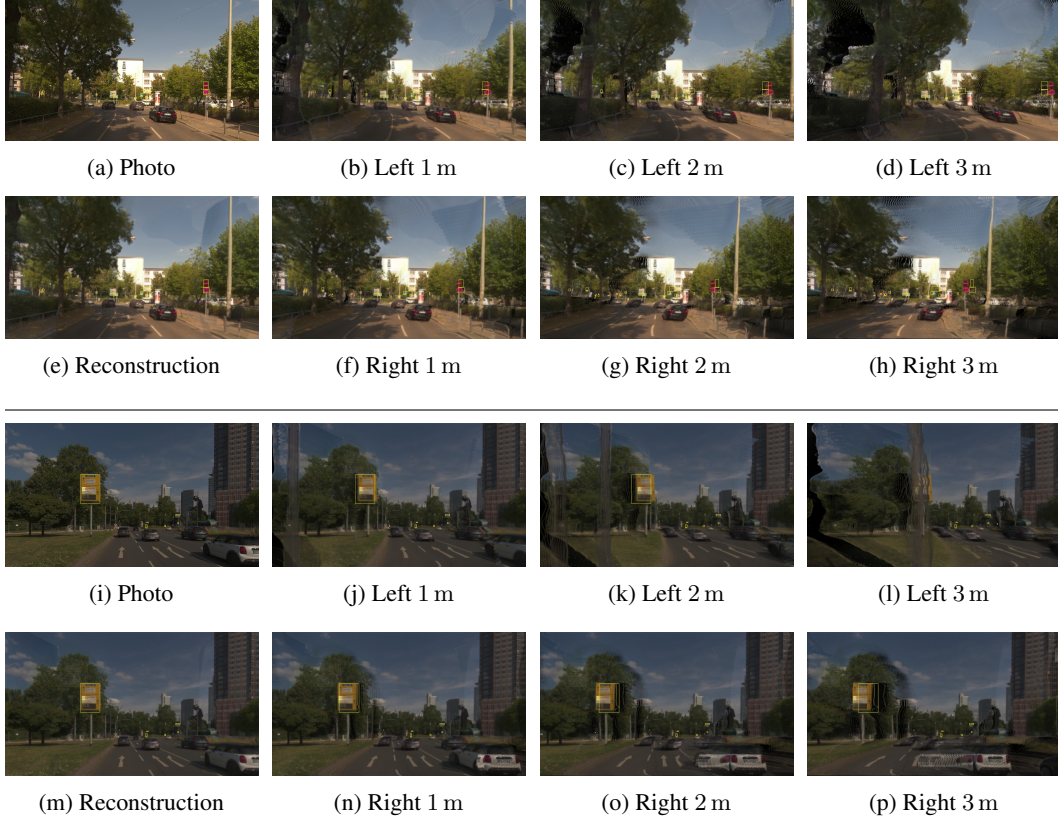


Figure 19: Qualitative comparison of traffic-sign recall under lateral viewpoint shifts. (a), (i) Real photograph with projected GT annotations (yellow). (b–h), (j–p) Lateral NVS rendered by ReconDrive [37] with detections. Note the deviation between rendered traffic signs and GT projections in the lateral views.

from each waypoint to the closest drivable centerline. For *collision-free rate*, the ego footprint is checked against a lidar-derived occupancy layer and logged dynamic-agent bounding boxes; a trajectory is collision-free if no waypoint intersects either set. *Multi-Maneuver Score (MMS)* [52] scores a prediction against the best of at least three human-annotated admissible 5 s reference maneuvers under joint similarity, comfort, instruction-following, and collision criteria; predicted 3 s trajectories are linearly extrapolated to 5 s.

Setup details. The 200 e2e samples are drawn from the $\text{val} \cup \text{overlap-train-val}$ scenes of Table 11, i.e., clips of 10 to 60 s recorded at 10 Hz; each sample is a non-overlapping 9 s window of 4 s past observation and up to 5 s of future trajectory anchored at a keyframe. Training is unrestricted on $\text{train} \cup \text{test}$, totalling over 100 km; the *test* split withholds maps but retains all sensor and trajectory data and is therefore usable for sensor-conditioned e2e training. The held-out *test-e2e* split contains 127 scenes and 33 km with future sensor data and global pose withheld after the keyframe, and is reserved for a future leaderboard. Headline numbers in Table 6 use the 3 s horizon for parity with nuScenes- and nuPlan-trained baselines; the full 5 s protocol is offered as the long-horizon benchmark for future work.

HD Map and occupancy map grounded metric definitions. All map-grounded metrics are evaluated on the predicted ego trajectory sampled at 10 Hz, with the ego footprint oriented along the predicted heading at every timestamp. *Drivable-surface survival* at a given horizon is the fraction of e2e samples for which all four corners of the ego footprint stayed inside the union of drivable Lanelet2 polygons of the scene’s local map at every timestamp up to that horizon. *Collision-free rate* is defined analogously: a sample is counted at a given horizon if the ego footprint never intersects the lidar-derived occupancy layer at any timestamp up to that horizon. *Centerline distance* is the mean

Table 18: Multi-Maneuver Score [52] on the 200 nine-second e2e samples drawn from the val \cup overlap-train-val scenes of KITScenes Multimodal, broken down by scene category. MMS is higher-is-better; best values are bold, second-best underlined.

Model	MMS \uparrow						
	avg	selec.	constr.	overt.	inters.	night	nom.
UniAD [38]	3.44	2.67	1.75	3.50	3.32	2.88	3.71
DMAD [53]	3.43	2.90	3.00	3.11	3.12	3.25	3.90
SSR [54] (non-temp.)	3.99	2.78	9.50	3.59	3.80	3.00	4.37
SSR [54] (temporal)	<u>3.90</u>	2.61	<u>6.25</u>	4.34	<u>3.73</u>	3.38	<u>4.08</u>
Epona [55] (AR, 10)	2.70	2.29	0.50	2.39	2.41	3.00	3.11
Epona [55] (AR, 100)	3.20	3.60	3.00	<u>3.70</u>	2.60	3.21	3.56
Epona [55] (SS, 10)	3.29	<u>3.09</u>	4.50	3.01	2.50	<u>4.65</u>	3.93
Epona [55] (SS, 100)	3.38	2.67	4.50	2.88	2.84	4.67	3.94

Epona is evaluated with single-step (SS) or autoregressive (AR) rollouts; 10 and 100 denote the number of diffusion denoising steps. Predicted 3 s trajectories are linearly extrapolated to match the 5 s evaluation horizon of [52].

lateral offset from the ego centre to the closest drivable centerline, averaged over all waypoints up to the evaluation horizon. Per-horizon profiles for all three metrics are plotted over a 3 s prediction range in Figure 20.

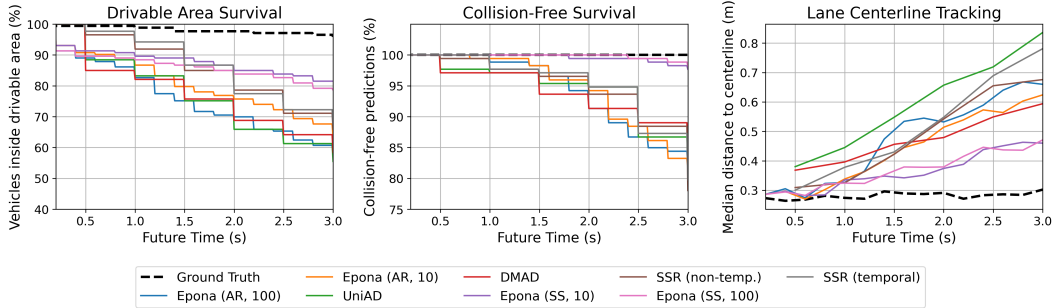


Figure 20: Per-horizon profiles of map-grounded safety and lane-compliance metrics for all evaluated end-to-end models on KITScenes Multimodal, plotted over the prediction horizon. From left to right: drivable-surface survival, lane-membership under strict topological lane definitions, collision-free rate against static and dynamic obstacles, and centerline-tracking error. All metrics degrade sharply beyond the 3 s headline horizon.

Baselines. We zero-shot evaluate the public checkpoints of UniAD [38], DMAD [53], SSR [54], and Epona [55]; no fine-tuning on KITScenes Multimodal is performed. UniAD, DMAD, and SSR consume the six ring cameras and are conditioned on a discrete navigation command, namely turn left, turn right, or go straight, derived from the ground-truth future trajectory. Epona consumes the front-view camera only and runs without navigation commands. SSR is evaluated in two configurations: *non-temp.* uses only the current keyframe whereas *temporal* aggregates past BEV features. Epona is evaluated both as single-step (SS) prediction and autoregressive (AR) rollout, each at 10 and 100 diffusion denoising steps.

Multi-Maneuver Score and Results. We follow the protocol of [52]: each scene is annotated with at least three admissible 5 s reference maneuvers spanning the recorded path together with alternative valid paths and a comfort variant. MMS scores a prediction against the best-matching reference under the joint similarity, comfort, instruction-following, and collision criteria of [52]. Predicted 3 s trajectories are linearly extrapolated to 5 s to match the evaluation horizon. We report MMS overall and per scene category in Table 18.

Under MMS in Table 18, the navigation-conditioned models (UniAD, DMAD, SSR) generally rank ahead of the navigation-free Epona variants, indicating that the explicit instruction signal

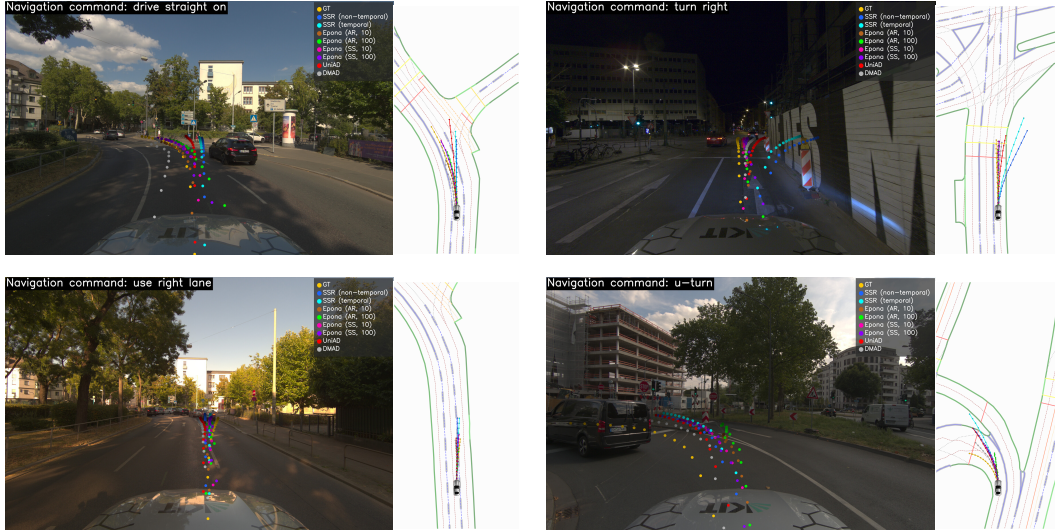


Figure 21: Additional qualitative end-to-end predictions on KITScenes Multimodal, complementing Figure 9. Each scene pairs the front-view camera image with a top-down view of all model trajectories overlaid on the HD map and ground truth.

yields more reliable alignment with at least one of the admissible maneuvers. Epona, by contrast, achieves the lowest positional errors in Table 6 owing to its larger pretraining corpus and stronger kinematic prior, but pays for it under MMS, where the multi-maneuver criterion explicitly rewards instruction-following over geometric proximity to the single recorded trajectory.

Additional qualitative examples. Figure 21 shows four further scenes complementing the two highlighted in Figure 9.

I Compute Resources

All models were trained and evaluated on 16 Nvidia A6000 Ada GPUs.