LYNXSCENES: A LARGE-SCALE REAL-WORLD BENCHMARK FOR OUTDOOR WHEEL-LEGGED ROBOT PERCEPTION, NAVIGATION, IMAGINATION AND RECONSTRUCTION

Anonymous authors
Paper under double-blind review



Figure 1: **Overview of LynxScenes.** The benchmark integrates large-scale real-world outdoor data collection, multimodal sensing, dense spatial, behavioral annotations, and four downstream tasks: instruction-conditioned trajectory prediction, BEV perception, future scene prediction, and multi-view scene reconstruction.

ABSTRACT

Wheel-legged robots operating in outdoor spaces shared with humans require accurate perception, socially compliant navigation, future-scene reasoning, and scene reconstruction. However, existing datasets typically support these capabilities in isolation and rarely provide dense spatial and behavior-level supervision on real wheel-legged platforms. We introduce LynxScenes, a large-scale real-world dataset and benchmark collected with a LYNX M20 wheel-legged robot. It contains approximately 100K 30-s navigation clips spanning 833.6 h and 3,294.7 km, with 3.0M annotated frames. The synchronized recordings include four fisheye cameras, dual LiDARs, IMU, odometry, and robot states. Annotations comprise 20.82M oriented 3D boxes across 27 categories, 13.89M road-structure annotations, and 20 navigation instructions organized into three layers and aligned with continuous future trajectories. LynxScenes supports four complementary tasks: instruction-conditioned trajectory prediction, BEV-based perception of 3D objects and road structures, future scene prediction, and multi-view scene reconstruction. For trajectory prediction, we further introduce a log-replay-based pseudo-closed-loop evaluator that measures safety, progress, and instruction compliance beyond trajectory imitation. Extensive evaluations across representative navigation, perception, world-model, and reconstruction methods reveal substantial remaining challenges in close-range outdoor robot navigation. The dataset and evaluation toolkit will be publicly released upon acceptance.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Wheel-legged robots are increasingly deployed in open outdoor environments shared with humans, including campuses, intersections, and parks. Combining efficient wheeled locomotion with the ability to traverse uneven terrain, they are promising platforms for delivery, inspection, and assistive applications. Unlike road vehicles operating within lane-structured environments or indoor robots navigating regular layouts, these robots move through loosely structured spaces from low, continuously changing viewpoints and often interact closely with pedestrians, cyclists, and other dynamic agents. Reliable navigation therefore requires not only safe motion but also socially compliant behavior in shared public spaces.

Such operation requires several tightly coupled capabilities: perceiving nearby agents and environmental structures, anticipating future scene evolution, generating continuous trajectories under platform-specific motion constraints, and reconstructing representative scenes for replay and analysis. Yet these capabilities are typically studied using separate datasets and evaluation protocols. Evaluating them on shared real-world observations enables consistent analysis under the same embodiment, environment, and interaction distribution. In autonomous driving, datasets such as KITTI-360 (Liao et al., 2022), nuScenes (Caesar et al., 2020), Waymo (Sun et al., 2020), and Argoverse 2 (Wilson et al., 2021) have enabled multiple autonomy problems to be studied on common real-world data. However, their vehicle-mounted viewpoints, lane-centric topology, and interaction regimes differ substantially from those of wheel-legged robots operating in pedestrian spaces.

Existing robot-navigation and language-conditioned navigation datasets provide only subsets of the required sensing and supervision, as summarized in Table 1. VLN benchmarks provide instruction-following supervision (Anderson et al., 2018; Ku et al., 2020; Krantz et al., 2020; Chen et al., 2019), but their simulated or graph-based interfaces largely abstract away physical robot dynamics and lack dense spatial annotations. Robot-centric datasets reduce this embodiment gap but remain fragmented across perception, social navigation, traversability, localization, and mapping (Martin-Martin et al., 2021; Karnan et al., 2022; Shah et al., 2021; Liang et al., 2025; Chaney et al., 2023; Jeong et al., 2024). Among closely related efforts, GrandTour (Frey et al., 2026) provides rich multimodal measurements for legged-robot perception and state estimation but lacks behavior-level language instructions, whereas CoT-VLNBench (Zhao et al., 2026) provides instructions, future trajectories, and 3D object boxes but lacks metric annotations of navigation-relevant road structures. The field therefore still lacks a large-scale real-world wheel-legged robot dataset that jointly provides multimodal observations, dense spatial annotations, behavior-level instructions, and continuous future trajectories.

To address this gap, we introduce LynxScenes (Fig. 1), a large-scale real-world dataset and benchmark for socially compliant outdoor wheel-legged robot navigation. Data are collected with a DEEP Robotics LYNX M20 operated by trained human operators following a predefined social-compliance protocol. LynxScenes contains approximately 100K non-overlapping 30-s navigation clips spanning 833.6 h and 3,294.7 km, with 3.0M annotated frames. It provides synchronized multimodal sensing and robot states, together with 20.82M oriented 3D boxes across 27 categories, 13.89M road-structure annotations, 20 behavior-level navigation instructions in three layers, and aligned continuous future trajectories.

Using the same sequences and training/evaluation splits, we benchmark four complementary tasks: instruction-conditioned trajectory prediction, BEV-based perception of 3D objects and road structures, future scene prediction, and multi-view scene reconstruction for visual replay and as a foundation for future simulation-based testing. For trajectory prediction, we further introduce a log-replay-based pseudo-closed-loop evaluator inspired by NAVSIM (Dauner et al., 2024). Its Quadruped Navigation Score (QNS) evaluates collision avoidance, walkable-area compliance, clearance, progress, and instruction adherence, complementing conventional trajectory-displacement errors. We evaluate 10 VLN/VLA policies, 4 BEV perception models, 6 world models, and 5 reconstruction methods. The results reveal substantial room for improvement in close-range interactions and complex outdoor environments. The dataset, standardized splits, and evaluation toolkit will be publicly released upon acceptance.

Our main contributions are summarized as follows: **(1)** We introduce LynxScenes, a large-scale real-world multimodal dataset collected on a wheel-legged robot under a standardized social-compliance protocol. It jointly provides socially compliant expert demonstrations, dense 3D object and road-

Table 1: Comparison with representative robot-navigation datasets and benchmarks. "Temp. ID" denotes persistent object identities across temporal observations; "Road/Trav. GT" denotes explicit road-structure, map, or traversability supervision; and "Cmd.→Traj." denotes direct conditioning of continuous trajectory or waypoint prediction on semantic navigation instructions. Scale is reported using the primary unit provided by each dataset. "N/R" indicates that a physical duration or traveled distance is not reported, while "–" denotes not applicable or unavailable.

Dataset	Real	Outdoor	Legged	LiDAR	3D Obj.	Temp. ID	Road/Trav.	Sem. Cmd.	Cmd.→Traj.	Scale	Duration	Distance
Real-world multimodal and robot-navigation datasets												
M2DGR (Yin et al., 2021)	✓	✓	–	✓	–	–	–	–	–	36 sequences	3.8 h	10.7 km
RECON (Shah et al., 2021)	✓	✓	–	✓	–	–	–	–	–	>5K trajectories	>40 h	N/R
SCAND (Karnan et al., 2022)	✓	✓	✓	✓	–	–	–	–	–	138 trajectories	8.7 h	40.2 km
GND (Liang et al., 2025)	✓	✓	–	✓	–	–	✓	–	–	10 campuses	11.1 h	53 km
M3ED (Chaney et al., 2023)	✓	✓	✓	✓	–	–	–	–	–	57 sequences	3.4 h	N/R
DiTer (Jeong et al., 2024)	✓	✓	✓	✓	–	–	✓	–	–	4 sequences	1.0 h	1.65 km
GrandTour (Frey et al., 2026)	✓	✓	✓	✓	–	–	–	–	–	49 sequences	>5 h	>10 km
Language-conditioned navigation benchmarks												
RxR (Ku et al., 2020)	–	–	–	–	–	–	–	✓	–	126K pairs	–	–
VLN-CE (Krantz et al., 2020)	–	–	–	–	–	–	–	✓	–	16.8K episodes	–	–
Touchdown (Chen et al., 2019)	–	✓	–	–	–	–	–	✓	–	9.3K examples	–	–
VLN-PE (Wang et al., 2025)	△	–	✓	–	–	–	–	✓	△	Physical benchmark	–	–
CoT-VLNBench (Zhao et al., 2026)	✓	✓	✓	✓	✓	✓	–	✓	✓	60 sequences / 175K frames	5 h	N/R
Ours	✓	✓	✓	✓	✓	✓	✓	✓	✓	**100K clips / 3.0M frames**	**833.6 h**	**3,294.7 km**

△ denotes partial or related support that is not directly equivalent to the corresponding task formulation.

structure annotations, structured behavior-level instructions, and continuous future trajectories. To the best of our knowledge, it is the largest publicly reported real-world wheel-legged robot navigation dataset with expert trajectories and dense 3D scene annotations. **(2)** We establish a unified benchmark spanning instruction-conditioned trajectory prediction, BEV perception, future scene prediction, and multi-view scene reconstruction, together with a log-replay-based pseudo-closed-loop evaluator that measures safety, navigation quality, and instruction adherence. **(3)** We systematically evaluate representative methods across all four tasks, establishing reproducible baselines, identifying persistent challenges in close-range outdoor navigation, and providing standardized data splits and a public evaluation toolkit.

2 RELATED WORK

2.1 REAL-WORLD ROBOT NAVIGATION DATASETS

Real-world robot datasets provide complementary supervision for localization, perception, and navigation. M2DGR (Yin et al., 2021) and M3ED (Chaney et al., 2023) emphasize multimodal sensing and state estimation, while RECON (Shah et al., 2021), SCAND (Karnan et al., 2022), GND (Liang et al., 2025), DiTer (Jeong et al., 2024), and GrandTour (Frey et al., 2026) provide various forms of trajectory, demonstration, traversability, or robot-state supervision. However, existing datasets typically expose only subsets of the spatial and behavioral information required to jointly study perception and navigation. In contrast, LynxScenes combines synchronized multimodal observations, persistent 3D objects, explicit road structures, structured navigation instructions, and continuous future trajectories within the same real-world wheel-legged robot dataset.

2.2 INSTRUCTION-CONDITIONED NAVIGATION AND PLANNING

Classical VLN benchmarks study instruction following primarily in simulated or image-based environments. R2R (Anderson et al., 2018), RxR (Ku et al., 2020), and VLN-CE (Krantz et al., 2020) focus on indoor navigation, whereas Touchdown (Chen et al., 2019) considers outdoor street-view imagery. Recent benchmarks move closer to physical navigation: VLN-PE (Wang et al., 2025) incorporates embodiment and realistic robot dynamics, while CoT-VLNBench (Zhao et al., 2026) provides real quadruped navigation data with language instructions, 3D object annotations, and continuous trajectories. However, CoT-VLNBench does not provide explicit road-structure geometry for trajectory-level spatial-safety evaluation. Our benchmark further aligns continuous future trajectories with a three-layer instruction taxonomy together with explicit object and road-structure annotations, enabling instruction-conditioned trajectory prediction and trajectory-level spatial-safety evaluation. Inspired by replay-based planning evaluation such as NAVSIM (Dauner et al., 2024),

we additionally evaluate predicted trajectories using safety, progress, and instruction-compliance criteria rather than trajectory similarity alone.

2.3 BEV Perception and Spatial Scene Understanding

Autonomous-driving datasets such as nuScenes (Caesar et al., 2020), Waymo (Sun et al., 2020), and Argoverse 2 (Wilson et al., 2021) have established large-scale benchmarks for 3D and BEV perception. Methods including BEVFormer (Li et al., 2022), BEVFusion (Liu et al., 2023), and MapTRv2 (Liao et al., 2023) further model surrounding objects and road geometry in BEV. These benchmarks are primarily designed around vehicle-scale traffic and driving-oriented map structures, whereas wheel-legged robots operate close to pedestrians, curbs, sidewalks, and crosswalks. LynxScenes therefore jointly evaluates nearby 3D objects and navigation-relevant road structures in a robot-centric BEV frame, providing spatial supervision that is directly aligned with the local constraints governing safe robot motion.

2.4 World Models and Neural Scene Reconstruction

Generative world models such as GAIA-1 (Hu et al., 2023), Navigation World Models (Bar et al., 2025), and Cosmos (NVIDIA, 2025; 2026) predict future visual observations from current scenes and motion or action conditions. In parallel, 3D Gaussian Splatting (Kerbl et al., 2023) and recent feed-forward reconstruction methods enable efficient scene reconstruction and novel-view synthesis. Most existing evaluation is conducted in driving or generic visual settings rather than low-viewpoint outdoor robot navigation. LynxScenes evaluates both future scene prediction and multi-view scene reconstruction on the same real-world robot data, providing complementary benchmarks for temporal scene evolution and viewpoint-dependent scene replay. This complementary design supports both predictive reasoning about future observations and scene-level analysis for outdoor robot navigation.

Detailed discussion of the related methods is provided in Appendix A.



Figure 2: Overview of the data-collection platform, representative navigation behaviors, and multimodal annotations in our benchmark. (a)–(e) show the DEEP Robotics LYNX M20 wheel-legged quadruped, its physical configuration, onboard sensing setup, and representative outdoor navigation scenarios. (f)–(g) illustrate robot behaviors together with temporally aligned semantic instructions covering primary navigation, special motion, and obstacle interaction. (h) visualizes synchronized four-view fisheye observations at different times and the corresponding bird's-eye-view annotations, including surrounding 3D objects, road structures, and the ego trajectory.

3 DATASET

LynxScenes is collected with a physical wheel-legged robot and organized into approximately 100K non-overlapping navigation clips of 30 s each, amounting to 3.0M annotated frames, 833.6 h of navigation, and 3,294.7 km of traveled distance. Each clip contains synchronized surround-view images, LiDAR point clouds, and robot states, while each annotated frame provides oriented 3D object boxes, road-structure annotations, structured navigation instructions, and a continuous future trajectory. The same recordings support instruction-conditioned trajectory prediction, BEV-based perception of 3D objects and road structures, future scene prediction, and multi-view scene reconstruction, as defined in Sec. 4.

3.1 PLATFORM AND SENSOR SETUP

The data-collection platform is a DEEPRobotics LYNX M20 wheel-legged quadruped (Fig. 2). It carries four fisheye cameras facing front, rear, left, and right, each recording 1920×1536 images at 10 Hz, and two 96-channel RoboSense Airy LiDARs mounted at the front and rear. Together, the LiDARs provide 360° horizontal and 90° vertical coverage to a range of 60 m at 10 Hz.

The platform additionally records IMU measurements, 6-DoF LiDAR-inertial odometry (LIO), and robot states. The robot-state stream provides the longitudinal velocity v_x, lateral velocity v_y, and yaw rate ω_z, and is used to construct the trajectory supervision described in Sec. 3.3. All geometric quantities are represented in the robot-centric Front–Left–Up (FLU) coordinate system, with x forward, y left, and z up. The complete sensor specifications are provided in Appendix B.2.

3.2 DATA COLLECTION AND SYNCHRONIZATION

All data are collected through human teleoperation of the physical robot in real outdoor environments, including sidewalks, streets, intersections, campuses, parks, narrow passages, slopes, open grounds, and rural roads. The recordings cover both structured and unstructured terrain, static obstacles, and frequent interactions with pedestrians, riders, and vehicles.

Annotated frames are sampled at 1 Hz from the synchronized multimodal recordings. Camera–LiDAR calibration, temporal alignment, and stream synchronization details are provided in Appendix B.2. Data collection is performed by trained operators following predefined socially compliant navigation guidelines; the collection protocol and quality-control procedure are detailed in Appendix B.5.

3.3 TRAJECTORY CONSTRUCTION

Reference future trajectories are constructed from the robot-state stream. Given the body-frame longitudinal velocity v_x, lateral velocity v_y, and yaw rate ω_z, the robot motion is temporally integrated and transformed into the robot coordinate frame at the current anchor timestamp. The resulting trajectory is resampled to 10 Hz, providing a sequence of future planar poses associated with each annotated frame. The complete integration formulation is given in Appendix B.3.

3.4 SPATIAL AND NAVIGATION ANNOTATIONS

Beyond trajectory supervision, the benchmark provides rich spatial and semantic annotations to support multiple key tasks in wheel-legged robot navigation. These annotations capture surrounding objects, road structures, and navigation intent, enabling both perception- and planning-oriented evaluation. The benchmark provides 27 object categories, four road-structure classes, and 20 navigation instructions organized into three layers. The complete annotation taxonomy is provided in Appendix B.4.

3D objects. Every object around the robot is annotated with an oriented 3D bounding box, one of 27 semantic categories, a persistent instance identity, a dynamic state, and a velocity. Identities are maintained across frames and temporally consecutive clips, which supports long-term association and per-object collision accounting. In total the dataset contains 20.82M boxes, 6.94 per annotated frame on average.

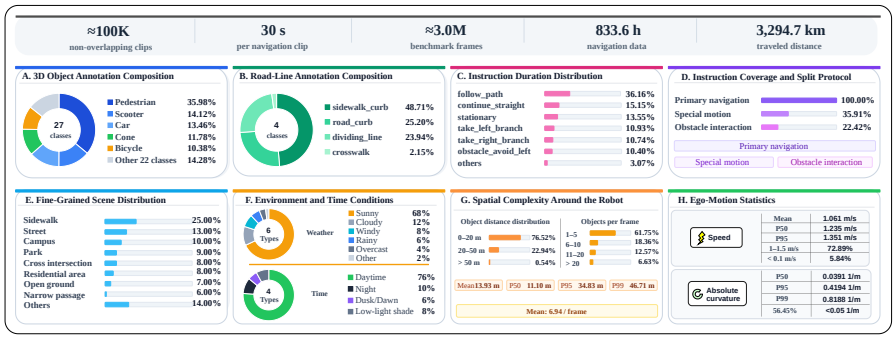


Figure 3: **Dataset statistics.** (A) Composition of 3D object annotations. (B) Composition of road-structure annotations. (C) Distribution of navigation instructions. (D) Coverage of the three instruction layers. (E) Scene-type distribution. (F) Environment and time-of-day conditions. (G) Object distance and per-frame object counts. (H) Ego speed and trajectory curvature.

Road structure. We annotate four road-structure classes, represented as polylines or closed crosswalk contours. These annotations encode the accessible area of the robot and are the basis of the Walkable-Area Compliance in Sec. 4. The dataset contains 13.89M road-structure annotations, 4.63 per frame on average.

Navigation instructions. Each frame is tagged with structured instructions from three independent layers. *Primary navigation* describes the dominant route intent and covers the entire timeline, *Special motion* marks robot-specific maneuvers and *Obstacle interaction* marks how the robot responds to an obstacle. Instructions within a layer are mutually exclusive, while instructions from different layers may overlap in time. Unlike joystick or velocity commands, these labels describe behavior-level intent, and their alignment with the future trajectory is what makes instruction-conditioned prediction and behavior-wise evaluation possible.

Annotation procedure and quality control. All annotations and recorded expert demonstrations undergo manual quality control for annotation completeness, consistency and compliance with the predefined socially compliant navigation guidelines. Detailed review and filtering procedures are provided in Appendix B.5.

3.5 Dataset Statistics and Splits

As summarized in Fig. 3, LynxScenes covers a wide range of real-world outdoor navigation conditions, including diverse scene types, environmental and illumination conditions, road structures, and human–robot interaction scenarios. Beyond environmental diversity, the dataset spans substantial variation in object categories, spatial complexity, semantic navigation behaviors, and ego-motion patterns, ranging from routine path following to special-motion and obstacle-interaction instructions. Such broad coverage supports systematic evaluation across heterogeneous navigation conditions and interaction complexities. The 30-s clip is used as the atomic split unit, with clips divided 8:1:1 into training, validation, and test sets while approximately preserving the navigation-instruction distribution. To avoid temporal and scene leakage, clips originating from the same continuous collection session are assigned to the same split.

4 Benchmark Task and Metrics

To validate the capability of our benchmark to support key wheel-legged robot navigation problems, we establish four representative benchmark tasks. These tasks span navigation prediction, spatial perception, future scene modeling, and scene reconstruction in real-world outdoor environments.

4.1 Instruction-Conditioned Trajectory Prediction

We formulate the primary benchmark task as instruction-conditioned continuous trajectory prediction. For each anchor t, a model receives four historical timestamps $\{-3, -2, -1, 0\}$ s, each with

synchronized front, rear, left, and right fisheye images (16 images total), the current motion state $\mathbf{s}_t = [v_x, v_y, \omega_z]$, and the three-layer semantic instruction. It predicts the complete future trajectory $\hat{\mathbf{T}}_t \in \mathbb{R}^{10\times 3}$ over 1 s at 10 Hz, where each pose is $[x_{\text{forward}}, y_{\text{left}}, \psi]$ in the anchor Body-FLU frame.

We report imitation accuracy using ADE, FDE, AHE, FHE, and complement it with a pseudo-closed-loop replay inspired by NAVSIM (Dauner et al., 2024). Annotated agents are replayed against the predicted ego trajectory and six interpretable sub-scores measure No Collision (NC), Walkable-Area Compliance (WAC), Time-to-Collision (TTC), Clearance (CL), Ego Progress (EP), and Instruction Compliance (IC). They are combined as

$$\text{QNS} = \text{NC} \times \text{WAC} \times \frac{w_{\text{EP}}\text{EP} + w_{\text{TTC}}\text{TTC} + w_{\text{CL}}\text{CL} + w_{\text{IC}}\text{IC}}{w_{\text{EP}} + w_{\text{TTC}} + w_{\text{CL}} + w_{\text{IC}}}. \tag{1}$$

The replay construction, footprint geometry, interpolation rules, thresholds, and instruction-specific compliance criteria are detailed in Appendix C.

4.2 BEV-Based Perception of 3D Objects and Road Structures

To evaluate spatial understanding for wheel-legged robot navigation, we consider BEV-based perception of surrounding 3D objects and navigation-relevant road structures.

For 3D object detection, models predict confidence-scored oriented boxes in the current robot frame and are evaluated on Car and Pedestrian using center-distance-based mAP and HDS (Caesar et al., 2020). **For road-structure detection**, Road Curb, Sidewalk Curb, and Road–Sidewalk Dividing Line form an ordinary-line class, while Crosswalk is represented as a closed-contour class and evaluated using Chamfer-distance-based AP. Detailed matching rules, thresholds, and metric definitions are provided in Appendix E.

4.3 Future Scene Prediction

To evaluate future scene prediction, we assess whether models can generate the future egocentric visual evolution of outdoor robot navigation scenes. Given an initial rectified front-view image $\mathbf{I}_0$, models generate a future RGB video and, when supported, additionally receive future motion conditions; text prompts are used when required by the native model interface.

We evaluate complementary aspects of generation quality, including perceptual quality, temporal quality, appearance consistency, and distributional realism, using Q-Align-I/V (Wu et al., 2024), CLIP-IQA+ (Wang et al., 2023), LAION-Aes (LAION-AI), DOVER-T (Wu et al., 2023), VBench temporal metrics (Huang et al., 2024), CLIP-I (Cherti et al., 2023), and FVD (Unterthiner et al., 2018). Model details, temporal normalization, and metric definitions are provided in Appendix F.

4.4 Multi-View Scene Reconstruction

To evaluate multi-view scene reconstruction, models reconstruct the observed environment from multiple context views and render RGB images at held-out viewpoints.

We benchmark representative Gaussian-splatting reconstruction methods and evaluate novel-view rendering using complementary pixel-, structural-, and perceptual-quality metrics. Detailed context-view settings, evaluation protocols, and metric definitions are provided in Appendix G.

5 EXPERIMENTS

5.1 Trajectory-Prediction Baselines and Results

Baselines. We evaluate five VLN models: InternNav (Wei et al., 2026), NaVILA (Cheng et al., 2025), OmniNav (Xue et al., 2026), StreamVLN (Wei et al., 2025), and SocialNav (Qwen3-VL-4B) (Chen et al., 2026); and five VLA models: LingBot-VLA-v2 (Wu et al., 2026), DepthVLA (Yuan et al., 2026), OpenPI π0.5 (Physical Intelligence et al., 2025), OpenVLA (Kim et al., 2025), and Isaac-GR00T-N1.7 (NVIDIA et al., 2025). All models are evaluated under the common trajectory-prediction protocol; model-specific adaptations are detailed in Appendix D.

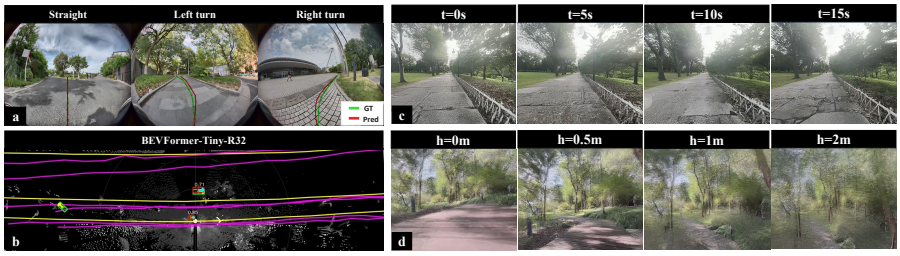


Figure 4: **Qualitative visualizations of the four benchmark tasks in LynxScenes. (a) Instruction-conditioned trajectory prediction:** examples of predicted future trajectories under three navigation commands. **(b) BEV-based perception of 3D objects and road structures. (c) Future scene prediction:** generated future egocentric observations at different rollout times. **(d) Multi-view scene reconstruction:** novel-view rendering results under different viewpoint changes.

Table 2: Results on instruction-conditioned trajectory prediction and future scene prediction. Best and second-best results are shown in **bold** and underlined, respectively.

(1) Instruction-conditioned trajectory prediction

Model	ADE↓	FDE↓	AHE↓	FHE↓	NC↑	WAC↑	TTC↑	CL↑	EP↑	IC↑	QNS↑
VLN-based policies											
InternNav	0.079	0.098	0.038	0.062	0.945	0.955	0.871	0.815	0.893	0.818	0.764
NaVILA	0.030	**0.059**	**0.028**	**0.048**	**0.968**	**0.975**	**0.912**	0.838	**0.941**	**0.893**	**0.843**
OmniNav	0.090	0.152	0.033	0.054	0.939	0.949	0.863	0.811	0.866	0.803	0.743
StreamVLN	0.036	0.067	0.032	0.053	0.964	0.972	0.905	**0.843**	0.933	0.884	0.833
SocialNav Qwen3-VL-4B	0.295	0.538	0.061	0.106	0.843	0.862	0.811	0.825	0.851	0.725	0.757
VLA-based policies											
LingBot-VLA-v2	0.044	0.068	0.038	0.054	0.958	0.966	0.893	0.829	0.926	0.863	0.809
DepthVLA	0.056	0.074	0.032	0.049	0.952	0.961	0.882	0.825	0.915	0.842	0.789
OpenPI π0.5	0.036	0.075	0.032	0.052	0.961	0.969	0.899	0.834	0.922	0.874	0.819
OpenVLA	0.041	0.068	0.030	0.051	0.956	0.964	0.888	0.823	0.928	0.855	0.802
Isaac-GR00T-N1.7	**0.027**	0.062	0.029	0.051	0.936	0.967	0.869	0.834	0.875	0.891	0.820
Reference trajectories											
Human trajectory	0	0	0	0	1.000	1.000	1.000	0.851	1.000	0.978	0.946
Constant velocity	0.112	0.213	0.041	0.079	0.942	0.951	0.858	0.806	0.884	0.672	0.716
Stationary	0.561	1.032	0.052	0.101	1.000	1.000	1.000	0.496	0.017	0.152	0.149

(3) Future scene prediction

Model	Q-Align-I↑	Q-Align-V↑	CLIP-I↑	CLIP-IQA+↑	LAION-Aes↑	DOVER-T↑	Smooth.↑	Dynamic↑	FVD↓
Matrix-Game-3	2.759	0.474	76.72	0.380	4.374	0.066	**0.978**	**1.000**	964.1
WorldStereo	3.310	0.633	85.73	0.474	4.786	0.050	0.896	0.995	**386.2**
Cosmos3-Nano	3.340	0.619	**87.85**	**0.588**	4.770	0.088	0.969	0.895	411.3
LingBot-World	**3.629**	**0.717**	83.11	0.575	**4.853**	0.078	0.942	**1.000**	764.7
Wan2.2-TI2V-5B	3.445	0.660	87.73	0.475	4.769	**0.094**	0.955	**1.000**	616.3
CogVideoX1.5-5B-I2V	2.734	0.479	85.79	0.386	4.629	0.051	0.921	0.985	692.7

Results. Table 2(1) reports the trajectory-prediction results. Among learned policies with pseudo-closed-loop evaluation, NaVILA achieves the highest QNS of 0.843, followed by StreamVLN at 0.833. For imitation accuracy, Isaac-GR00T-N1.7 obtains the lowest ADE (0.027), while NaVILA achieves the lowest FDE, AHE, and FHE. The constant-velocity and stationary references further show that collision avoidance alone is insufficient, as useful navigation also requires progress and instruction compliance.

5.2 BEV-Perception Baselines and Results

Baselines. We evaluate BEVFormer-Tiny (Li et al., 2022), MapTRv2 (Liao et al., 2023), BEVFusion (Liu et al., 2023), and SparseDrive (Sun et al., 2025). Their sensor inputs, training configurations, and checkpoint-selection protocols are detailed in Appendix E.

Results. Table 3(2) shows that SparseDrive achieves the highest road-structure mAP, whereas BEVFusion obtains the highest box mAP and HDS. The latter result should be interpreted together with BEVFusion's additional LiDAR input. Additional qualitative results are provided in Appendix E.

Table 3: Results on BEV-based perception and multi-view scene reconstruction. Best and second-best results are shown in **bold** and underlined, respectively. † denotes early-stopped training.

(2) BEV-based perception

Method	Input	Line↑	Box↑	HDS↑
BEVFormer-Tiny†	Cam.+Pose	16.92	52.57	54.39
MapTRv2†	Camera	13.63	45.90	47.57
BEVFusion	Cam.+LiDAR	19.29	**67.11**	**64.48**
SparseDrive	Cam.+Pose	**21.08**	58.70	61.12

(4) Multi-view scene reconstruction

Method	K	PSNR↑	SSIM↑	LPIPS↓	MS-S↑	DISTS↓	FSIMc↑	GMSD↓
ReCoSplat	8–16	**15.937**	0.2425	**0.5113**	**0.3901**	0.2133	0.6364	0.2436
YoNoSplat	8–16	15.390	**0.2736**	0.6721	0.3096	0.3986	0.5113	0.2928
AnySplat	8–16	14.544	0.2096	0.5119	0.3359	**0.2128**	**0.6429**	**0.2393**
FreeSplatter-S	2	14.537	0.2373	0.6461	0.3099	0.3277	0.5687	0.2656
MVSplat	2	11.777	0.1938	0.6307	0.3008	0.3312	0.5807	0.2774

5.3 Future-Scene-Prediction Baselines and Results

Baselines. We evaluate six representative world and video-generation models: Matrix-Game-3 (Skywork AI Matrix-Game Team, 2026), WorldStereo (Zhang et al., 2026), Cosmos3-Nano (NVIDIA, 2026), LingBot-World (Robbyant Team et al., 2026), Wan2.2-TI2V-5B (Wan Team et al., 2025), and CogVideoX1.5-5B-I2V (Yang et al., 2024). All methods are evaluated on the same 200-clip subset; model details and temporal-normalization protocols are provided in Appendix F.

Results. Table 2(3) shows that no single model dominates all evaluation dimensions. LingBot-World achieves the highest Q-Align-I, Q-Align-V, and LAION-Aes scores; Cosmos3-Nano performs best on CLIP-I and CLIP-IQA+; Wan2.2-TI2V-5B obtains the highest DOVER-T; and WorldStereo achieves the lowest FVD. The results reveal trade-offs among perceptual quality, temporal quality, appearance consistency, and distributional realism.

5.4 Multi-View-Reconstruction Baselines and Results

Baselines. We evaluate five pretrained Gaussian-splatting methods without additional fine-tuning: ReCoSplat (Cheng et al., 2026), YoNoSplat (Ye et al., 2026), AnySplat (Jiang et al., 2025), FreeSplatter-S (Xu et al., 2025), and MVSplat (Chen et al., 2024). Model-specific reconstruction interfaces and evaluation settings are provided in Appendix G.

Results. Table 3(4) shows complementary strengths across methods. ReCoSplat achieves the best PSNR, LPIPS, and MS-SSIM, YoNoSplat obtains the highest SSIM, and AnySplat performs best on DISTS, FSIMc, and GMSD. No single method dominates all reconstruction metrics.

Figure 4 provides representative qualitative examples across the four benchmark tasks. More extensive qualitative results for trajectory prediction, BEV perception, future scene prediction, and multi-view scene reconstruction are provided in Appendices D, E, F, and G, respectively.

6 CONCLUSION

We present LynxScenes, a large-scale real-world benchmark for outdoor wheel-legged robot autonomy that combines synchronized multimodal sensing, dense spatial annotations, behavior-level navigation instructions, and continuous future trajectories. Built on the same real-world robot data, the benchmark supports four complementary tasks: instruction-conditioned trajectory prediction, BEV-based perception of 3D objects and road structures, future scene prediction, and multi-view scene reconstruction. For trajectory prediction, we further introduce a log-replay-based pseudo-closed-loop evaluator that complements imitation metrics with safety, progress, walkable-area compliance, and instruction adherence.

Across representative VLN/VLA policies, BEV perception models, world models, and reconstruction methods, our experiments reveal substantial remaining challenges in close-range interaction, spatial understanding, future-scene modeling, and viewpoint-dependent reconstruction. These results suggest that outdoor wheel-legged robot autonomy remains difficult even when perception, navigation, prediction, and reconstruction are evaluated on the same embodiment and scene distribution. We hope LynxScenes provides a common real-world foundation for studying these capabilities together and supports future progress toward more integrated and socially compliant outdoor robot autonomy.

AI USE STATEMENT

In this work, we used generative AI tools only to assist with language polishing, grammatical correction, and improving the clarity and organization of the manuscript. Generative AI was not used to develop the research ideas, design the dataset or benchmark, generate annotations, implement the reported methods, conduct experiments, produce experimental results, or create figures. All technical contributions, methodological choices, experimental designs, analyses, and conclusions are the authors' original work.

All AI-assisted edits were manually reviewed and revised by the authors to ensure technical accuracy, consistency with the underlying research, and appropriate attribution. The authors take full responsibility for the final content of this work, including all text, claims, results, and artifacts.

ETHICS STATEMENT

LynxScenes is collected using a physical wheel-legged robot operating in real outdoor environments shared with pedestrians, riders, and vehicles. Data collection is performed by trained human operators following predefined socially compliant navigation guidelines, including yielding, maintaining appropriate clearance, and safely negotiating interactions with nearby agents. Recorded demonstrations additionally undergo manual behavior-level review, and segments containing unsafe, socially inappropriate, or inconsistent operator behavior are excluded from the benchmark.

Because the dataset is collected in real-world shared spaces, it may contain incidental observations of members of the public. To protect privacy, personally identifiable visual information is anonymized before public release. In particular, visible faces and vehicle license plates are automatically detected and blurred, followed by manual inspection to identify and correct missed or insufficient anonymization. Other privacy-sensitive information is also reviewed during data processing, and samples that cannot be adequately anonymized are excluded from the public release. These measures are intended to reduce the risk of exposing individual identities while preserving the utility of the data for robotics research.

The released dataset will follow applicable institutional, legal, and privacy requirements, together with documentation describing its intended research use and limitations. The benchmark is intended for research on robot perception, navigation, prediction, and reconstruction, and should not be interpreted as validating deployment without task-specific safety assessment.

REPRODUCIBILITY STATEMENT

We provide detailed documentation of the dataset, benchmark protocols, and baseline evaluations to support reproducibility. Section 3 describes the data collection, multimodal sensing, trajectory construction, annotation design, and dataset splits, while Section 4 defines the four benchmark tasks and their evaluation metrics. Section 5 reports the evaluated baselines and main quantitative results.

Additional implementation details are provided in the appendix: Appendix B describes sensor configuration, calibration, synchronization, trajectory construction, annotation taxonomy, and data-quality procedures; Appendix C specifies the pseudo-closed-loop navigation evaluator and QNS components; Appendix D documents the adaptation of trajectory-prediction baselines; and Appendices E–G provide the evaluation protocols and implementation details for BEV perception, future scene prediction, and multi-view scene reconstruction, respectively. Upon acceptance, we plan to release the dataset and evaluation toolkit to support reproducible comparison with future methods.

REFERENCES

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

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)*, pp. 11621–11631, 2020.

Kenneth Chaney, Fernando Cladera, Ziyun Wang, Anthony Bisulco, M. Ani Hsieh, Christopher Korpela, Vijay Kumar, Camillo J. Taylor, and Kostas Daniilidis. M3ed: Multi-robot, multi-sensor, multi-environment event dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 4016–4023, June 2023.

Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. MVSplat: Efficient 3D gaussian splatting from sparse multi-view images. In *European Conference on Computer Vision*, 2024. URL https://arxiv.org/abs/2403.14627.

Ziyi Chen, Yingnan Guo, Zedong Chu, Minghua Luo, Yanfen Shen, Mingchao Sun, Junjun Hu, Shichao Xie, Yang Kuan, Pei Shi, Zhining Gu, Lu Liu, Honglin Han, Xiaolong Wu, Mu Xu, and Yu Zhang. Socialnav: Training human-inspired foundation model for socially-aware embodied navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28796–28806, 2026.

An-Chieh Cheng, Yandong Ji, Zhaojing Yang, Zaitian Gongye, Xueyan Zou, Jan Kautz, Erdem Biyik, Hongxu Yin, Sifei Liu, and Xiaolong Wang. NaVILA: Legged Robot Vision-Language-Action Model for Navigation. In *Proceedings of Robotics: Science and Systems*, LosAngeles, CA, USA, June 2025. doi: 10.15607/RSS.2025.XXI.018.

Freeman Cheng, Botao Ye, Xueting Li, Junqi You, Fangneng Zhan, and Ming-Hsuan Yang. ReCoSplat: Online feed-forward gaussian splatting via render-and-compare. *arXiv preprint arXiv:2603.09968*, 2026. URL https://freemancheng.com/ReCoSplat/.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2829, 2023. doi: 10.1109/CVPR52729.2023.00276.

Daniel Dauner, Marcel Hallgarten, Tianyu Li, Xinshuo Wen g, Zhiyu Huang, Zetong Yang, Hongyang Li, Igor Gili tschenski, Boris Ivanovic, Marco Pavone, Andr eas Geiger, and Kashyap Chitta. NAVSIM: Data-driven non-reactive autonomous vehicle simulat ion and benchmarking. In *Advances in Neural Information Processing Systems (NeurIP S), Datasets and Benchmarks Track*, 2024.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *arXiv preprint arXiv:2004.07728*, 2020. URL https://arxiv.org/abs/2004.07728.

Jonas Frey, Turcan Tuna, Frank Fu, Katharine Patterson, Tianao Xu, Maurice Fallon, Cesar Cadena, and Marco Hutter. Grandtour: A legged robotics dataset in the wild for multi-modal perception and state estimation, 2026. URL https://arxiv.org/abs/2602.18164. *Equal contribution (Turcan Tuna and Jonas Frey).

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. GAIA-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

Seokhwan Jeong, Hogyun Kim, and Younggun Cho. DiTer: Diverse terrain and multimodal dataset for field robot navigation in outdoor environments. *IEEE Sensors Letters*, 8(3):1–4, March 2024. doi: 10.1109/LSENS.2024.3356870. Art. no. 5500704.

Lihan Jiang, Yucheng Mao, Linning Xu, Tao Lu, Kerui Ren, Yichen Jin, Xudong Xu, Mulin Yu, Jiangmiao Pang, Feng Zhao, Dahua Lin, and Bo Dai. AnySplat: Feed-forward 3D gaussian splatting from unconstrained views. *ACM Transactions on Graphics*, 44(6):1–16, 2025. URL https://arxiv.org/abs/2505.23716.

Haresh Karnan, Anirudh Nair, Xuesu Xiao, Garrett Warnell, Sören Pirk, Alexander Toshev, Justin Hart, Joydeep Biswas, and Peter Stone. Socially compliant navigation dataset (scand): A large-scale dataset of demonstrations for social navigation. *IEEE Robotics and Automation Letters*, 2022.

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.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model. In *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 2679–2713. PMLR, 06–09 Nov 2025. URL https://proceedings.mlr.press/v270/kim25c.html.

Jacob Krantz, Erik Wijmans, Arjun Majundar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph: Vision and language navigation in continuous environments. In *European Conference on Computer Vision (ECCV)*, 2020.

Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-Across-Room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In *Conference on Empirical Methods for Natural Language Processing (EMNLP)*, 2020.

LAION-AI. LAION-Aesthetics Predictor V1. https://github.com/LAION-AI/aesthetic-predictor. A linear aesthetic-quality predictor built on CLIP image embeddings.

Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. *arXiv preprint arXiv:2203.17270*, 2022.

Jing Liang, Dibyendu Das, Daeun Song, Md Nahid Hasan Shuvo, Mohammad Durrani, Karthik Taranath, Ivan Penskiy, Dinesh Manocha, and Xuesu Xiao. Gnd: Global navigation dataset with multi-modal perception and multi-category traversability in outdoor campus environments, 2025.

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. In *arXiv preprint arXiv:2308.05736*, 2023.

Yiyi Liao, Jun Xie, and Andreas Geiger. KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Zhijian Liu, Haotian Tang, Alexander Amini, Xingyu Yang, Huizi Mao, Daniela Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

Roberto Martin-Martin, Mihir Patel, Hamid Rezatofighi, Abhijeet Shenoi, JunYoung Gwak, Eric Frankel, Amir Sadeghian, and Silvio Savarese. Jrdb: A dataset and benchmark of egocentric robot visual perception of humans in built environments. *IEEE transactions on pattern analysis and machine intelligence*, 2021.

NVIDIA. Cosmos world foundation model platform for physical AI. *arXiv preprint arXiv:2501.03575*, 2025.

NVIDIA. Cosmos 3: Omnimodal world models for physical ai. *arXiv preprint arXiv:2606.02800*, 2026. URL https://arxiv.org/abs/2606.02800.

NVIDIA, Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Jim Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, Joel Jang, Zhenyu Jiang, Jan Kautz, Kaushil Kundalia, Lawrence Lao, Zhiqi Li, Zongyu Lin, Kevin Lin, Guilin Liu, Edith Llontop, Loic Magne, Ajay Mandlekar, Avnish Narayan, Soroush Nasiriany, Scott Reed, You Liang Tan, Guanzhi Wang, Zu Wang, Jing Wang, Qi Wang, Jiannan Xiang, Yuqi Xie, Yinzhen Xu, Zhenjia Xu, Seonghyeon Ye, Zhiding Yu, Ao Zhang, Hao Zhang, Yizhou Zhao, Ruijie Zheng, and Yuke Zhu. GR00T N1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, et al. $\pi_{0.5}$: A vision-language-action model with open-world generalization, 2025. URL https://arxiv.org/abs/2504.16054.

Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, Hanlin Wang, Yinghao Xu, Shuailei Ma, Yihang Chen, Jie Liu, Yansong Cheng, Yao Yao, Jiayi Zhu, Yihao Meng, Kecheng Zheng, Qingyan Bai, Jingye Chen, Zehong Shen, Yue Yu, Xing Zhu, Yujun Shen, and Hao Ouyang. Advancing open-source world models. *arXiv preprint arXiv:2601.20540*, 2026.

Dhruv Shah, Benjamin Eysenbach, Nicholas Rhinehart, and Sergey Levine. Rapid Exploration for Open-World Navigation with Latent Goal Models. In *5th Annual Conference on Robot Learning*, 2021. URL https://openreview.net/forum?id=d_SWJhyKfVw.

Skywork AI Matrix-Game Team. Matrix-Game 3.0: Real-time and streaming interactive world model with long-horizon memory. Technical report, 2026. URL https://github.com/SkyworkAI/Matrix-Game/blob/main/Matrix-Game-3/assets/pdf/report.pdf.

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)*, pp. 2446–2454, 2020.

Wenchao Sun, Xuewu Lin, Yining Shi, Chuang Zhang, Haoran Wu, and Sifa Zheng. Sparsedrive: End-to-end autonomous driving via sparse scene representation. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8795–8801, 2025.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric and challenges. *arXiv preprint arXiv:1812.01717*, 2018. URL https://arxiv.org/abs/1812.01717.

Wan Team, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Jianyi Wang, Kelvin C. K. Chan, and Chen Change Loy. Exploring CLIP for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 2555–2563, 2023. doi: 10.1609/aaai.v37i2.25353.

Liuyi Wang, Xinyuan Xia, Hui Zhao, Hanqing Wang, Tai Wang, Yilun Chen, Chengju Liu, Qijun Chen, and Jiangmiao Pang. Rethinking the embodied gap in vision-and-language navigation: A holistic study of physical and visual disparities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9455–9465, October 2025. doi: 10.1109/ICCV51701.2025.00882.

Zhou Wang, Eero P. Simoncelli, and Alan C. Bovik. Multiscale structural similarity for image quality assessment. In *Proceedings of the 37th Asilomar Conference on Signals, Systems and Computers*, volume 2, pp. 1398–1402, 2003. doi: 10.1109/ACSSC.2003.1292216.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. doi: 10.1109/TIP.2003.819861.

Meng Wei, Chenyang Wan, Xiqian Yu, Tai Wang, Yuqiang Yang, Xiaohan Mao, Chenming Zhu, Wenzhe Cai, Hanqing Wang, Yilun Chen, Xihui Liu, and Jiangmiao Pang. StreamVLN: Streaming vision-and-language navigation via SlowFast context modeling, 2025. URL https://arxiv.org/abs/2507.05240.

Meng Wei, Chenyang Wan, Jiaqi Peng, Xiqian Yu, Yuqiang Yang, Delin Feng, Wenzhe Cai, Chenming Zhu, Tai Wang, Jiangmiao Pang, and Xihui Liu. Ground slow, move fast: A dual-system foundation model for generalizable vision-language navigation. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=GK4rznYwhn.

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*, 2021.

Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20144–20154, 2023.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-Align: Teaching LMMs for visual scoring via discrete text-defined levels. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54015–54029. PMLR, 2024. URL https://proceedings.mlr.press/v235/wu24ah.html.

Wei Wu, Fangjing Wang, Fan Lu, He Sun, Shi Liu, Yunnan Wang, Yibin Yan, Yong Wang, Shuailei Ma, Xinyang Wang, Yibin Liu, Shuai Yang, Tianxiang Zhou, Kejia Zhang, Lei Zhou, Cheng Su, Nan Xue, Bin Tan, Han Zhang, Youchao Zhang, Fei Liao, Xing Zhu, Yujun Shen, and Kecheng Zheng. From foundation to application: Improving VLA models in practice, 2026. URL https://arxiv.org/abs/2607.06403. LingBot-VLA 2.0 technical report.

Jiale Xu, Shenghua Gao, and Ying Shan. FreeSplatter: Pose-free gaussian splatting for sparse-view 3D reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/papers/Xu_FreeSplatter_Pose-free_Gaussian_Splatting_for_Sparse-view_3D_Reconstruction_ICCV_2025_paper.pdf.

Wufeng Xue, Lei Zhang, Xuanqin Mou, and Alan C. Bovik. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *IEEE Transactions on Image Processing*, 23(2):684–695, 2014. doi: 10.1109/TIP.2013.2293423.

Xinda Xue, Junjun Hu, Minghua Luo, Shichao Xie, Jintao Chen, Zixun Xie, Kuichen Quan, Wei Guo, Mu Xu, and Zedong Chu. OmniNav: A unified framework for prospective exploration and visual-language navigation. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2509.25687.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. CogVideoX: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Botao Ye, Boqi Chen, Haofei Xu, Daniel Barath, and Marc Pollefeys. YoNoSplat: You only need one model for feedforward 3D gaussian splatting. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2511.07321.

Jie Yin, Ang Li, Tao Li, Wenxian Yu, and Danping Zou. M2dgr: A multi-sensor and multi-scenario slam dataset for ground robots. *IEEE Robotics and Automation Letters*, 7(2):2266–2273, 2021.

Tianyuan Yuan, Yicheng Liu, Chenhao Lu, Zhuoguang Chen, Tao Jiang, and Hang Zhao. Depthvla: Enhancing vision-language-action models with depth-aware spatial reasoning. In *2026 IEEE International Conference on Robotics and Automation (ICRA)*, 2026.

Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. FSIM: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20(8):2378–2386, 2011. doi: 10.1109/TIP.2011.2109730.

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*, 2018. URL https://richzhang.github.io/PerceptualSimilarity/.

Yisu Zhang, Chenjie Cao, Tengfei Wang, Xuhui Zuo, Junta Wu, Jianke Zhu, and Chunchao Guo. Worldstereo: Bridging camera-guided video generation and scene reconstruction via 3d geometric memories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 40327–40339, 2026.

Xiao Zhao, Chang Liu, Ruiteng Ji, Zheyuan Zhang, Mingxu Zhu, Linna Song, Zhe Ren, Qingliang Luo, Yuhang Gao, Zhaolong Du, Chufan Guo, and Kuifeng Su. CoT-VLNBench: A benchmark for visual chain-of-thought reasoning in vision-language-navigation robots. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 36573–36581, 2026. doi: 10.1609/aaai.v40i43.40980.

Hongyu Zhou, Longzhong Lin, Jiabao Wang, Yichong Lu, Dongfeng Bai, Bingbing Liu, Yue Wang, Andreas Geiger, and Yiyi Liao. HUGSIM: A real-time, photo-realistic and closed-loop simulator for autonomous driving. *arXiv preprint arXiv:2412.01718*, 2024. URL https://arxiv.org/abs/2412.01718.

A Additional Related Work

A.1 Real-World Multimodal Robot Datasets

Existing real-world robot datasets provide complementary forms of multimodal sensing and navigation supervision. M2DGR (Yin et al., 2021) provides synchronized fisheye cameras, LiDAR, inertial sensing, GNSS, and odometry for multi-sensor SLAM, while M3ED (Chaney et al., 2023) extends multimodal sensing to multiple robotic platforms and challenging environments. GrandTour (Frey et al., 2026) further provides synchronized multimodal observations, proprioception, and high-precision trajectories for legged-robot perception and state estimation.

Several datasets focus more directly on navigation. RECON (Shah et al., 2021) provides real-robot trajectories for goal-conditioned open-world navigation, while SCAND (Karnan et al., 2022) collects human-teleoperated demonstrations for socially compliant navigation. GND (Liang et al., 2025) introduces semantic traversability supervision for outdoor global navigation, and DiTer (Jeong et al., 2024) combines multimodal quadruped observations with reference trajectories and global maps across diverse outdoor terrains.

These datasets emphasize different subsets of sensing, localization, trajectory, or traversability supervision. LynxScenes instead integrates temporally aligned multimodal sensing, persistent 3D objects, road-structure geometry, structured navigation instructions, and future trajectories in a single wheel-legged robot dataset, allowing the same recordings to support multiple perception, navigation, prediction, and reconstruction tasks.

A.2 Navigation Policies and Planning Evaluation

Recent navigation models increasingly map multimodal observations and semantic instructions directly to continuous robot behavior. NaVILA (Cheng et al., 2025) adapts vision-language-action modeling to legged robot navigation, while StreamVLN (Wei et al., 2025) introduces streaming visual-language reasoning for continuous navigation. InternNav (Wei et al., 2026) further explores generalizable vision-language navigation with a dual-system architecture.

General-purpose VLA models provide another relevant policy family. OpenVLA (Kim et al., 2025) learns continuous robot actions from vision-language inputs, while π0.5 (Physical Intelligence et al., 2025) uses a flow-matching action expert to generate continuous action chunks. These approaches are originally developed under different embodiments, action spaces, and task definitions. In our benchmark, they are adapted to a common observation, instruction, robot-state, and continuous-trajectory interface, enabling direct comparison under the same real-world navigation data.

Planning evaluation also increasingly extends beyond open-loop trajectory displacement. NAVSIM (Dauner et al., 2024) evaluates predicted driving trajectories through non-reactive scene replay and combines multiple safety- and progress-related criteria. Our pseudo-closed-loop evaluator follows this general replay-based principle but adapts it to wheel-legged robot navigation, where nearby pedestrians and riders, curb and sidewalk boundaries, and behavior-level instructions directly constrain acceptable trajectories.

A.3 BEV Perception for Driving and Robot Navigation

Large-scale autonomous-driving datasets including nuScenes (Caesar et al., 2020), Waymo Open Dataset (Sun et al., 2020), and Argoverse 2 (Wilson et al., 2021) have established widely used benchmarks for 3D object detection and spatial scene understanding.

BEV-based perception methods provide a natural representation for combining surrounding observations and geometric reasoning. BEVFormer (Li et al., 2022) constructs spatiotemporal BEV representations from multi-camera images. BEVFusion (Liu et al., 2023) combines camera and LiDAR information in a unified BEV representation, while MapTRv2 (Liao et al., 2023) represents road and map elements using vectorized geometric predictions.

These approaches motivate the perception baselines used in our benchmark, but the spatial constraints differ from conventional autonomous driving. Vehicle-oriented benchmarks primarily emphasize traffic participants, lanes, and HD-map elements, whereas outdoor wheel-legged robots frequently operate near pedestrians, riders, curbs, sidewalk boundaries, road–sidewalk dividing lines, and crosswalks. Our BEV task therefore evaluates both nearby 3D objects and navigation-relevant road structures in the robot-centric coordinate frame.

A.4 World Models for Future Scene Prediction

World models predict how future observations evolve under actions, motion conditions, or other control signals. GAIA-1 (Hu et al., 2023) demonstrates generative world modeling for autonomous driving, while Navigation World Models (Bar et al., 2025) generate future egocentric observations conditioned on navigation actions. Cosmos (NVIDIA, 2025) further explores large-scale world foundation models for physical-AI applications.

These works demonstrate the potential of generative prediction for reasoning about future scene evolution, but are commonly evaluated in autonomous-driving, generic video-generation, or other embodiment settings. Outdoor wheel-legged robot navigation introduces a distinct visual regime, including low camera height, close-range pedestrians and riders, vegetation, curbs, and frequent viewpoint changes. Our benchmark evaluates whether existing world models can generate future robot-centric observations under these conditions using perceptual, temporal, conditioning-consistency, and distributional metrics.

A.5 Neural Scene Reconstruction and Scenario Replay

3D Gaussian Splatting (Kerbl et al., 2023) provides an efficient explicit scene representation for novel-view synthesis. Recent methods extend Gaussian-based reconstruction to sparse, unposed, and online observations. MVSplat (Chen et al., 2024) uses posed multi-view cost volumes, FreeSplatter (Xu et al., 2025) targets sparse-view pose-free reconstruction, and AnySplat (Jiang et al., 2025) reconstructs Gaussian scene representations from unconstrained views. ReCoSplat (Cheng et al., 2026) performs online reconstruction through render-and-compare, while YoNoSplat (Ye et al., 2026) supports both posed and unposed feed-forward reconstruction.

Beyond novel-view synthesis, reconstructed Gaussian scenes have also been explored for scenario replay and simulation. HUGSIM (Zhou et al., 2024), for example, integrates Gaussian-splatting-

based scene reconstruction into a photorealistic closed-loop driving simulator. Our current reconstruction benchmark is more limited in scope: it evaluates observation-grounded novel-view rendering on real robot navigation sequences rather than claiming full closed-loop simulation capability. Together with the world-model task, it provides complementary evaluation of future visual prediction and explicit scene reconstruction.

B Dataset Details

B.1 Benchmark Task Pipeline

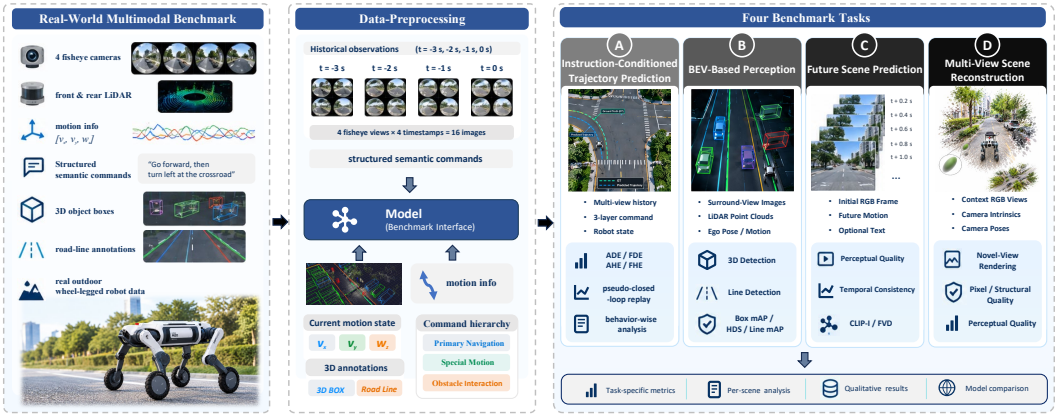


Figure 5: **Benchmark task pipeline.** The same real-world multimodal recordings and structured annotations support four tasks: (A) instruction-conditioned trajectory prediction, (B) BEV-based perception of 3D objects and road structures, (C) future scene prediction, and (D) multi-view scene reconstruction. Task-specific inputs are constructed from the shared recordings; LiDAR, ego-motion, instruction, and camera-pose conditions are used when required by the corresponding model. Outputs are evaluated using task-specific quantitative metrics and qualitative comparisons.

B.2 Sensor Configuration and Calibration

Table 4 summarizes the sensing and state modalities of the data-collection platform.

Calibration. The four fisheye cameras are individually calibrated for camera intrinsics, fisheye distortion, and extrinsics with respect to the robot body frame. The front and rear LiDARs are registered to the same robot-centric frame. As shown in Fig. 6, cross-modal calibration is qualitatively verified by projecting LiDAR point clouds onto the calibrated fisheye images. Projected points are color-coded by depth along the camera axis, with red indicating nearer points and blue indicating farther points.

Temporal synchronization. Annotated frames are sampled at 1 Hz from the 10 Hz camera and LiDAR recordings. For each annotation timestamp, the temporally closest camera images and LiDAR sweeps are selected, while LIO and robot-state measurements are aligned to the same anchor timestamp. The resulting anchor therefore provides temporally aligned visual, geometric, ego-motion, and robot-state observations for downstream annotation and benchmark construction.

B.3 Trajectory Construction Details

Reference future trajectory supervision is constructed from the recorded motion-state stream. At each time step, the motion state provides the longitudinal velocity v_x, lateral velocity v_y, and yaw rate ω_z in the robot-centric FLU coordinate system.

Given the body-frame velocities $(v_x^k, v_y^k, \omega_z^k)$ at sampling interval Δt, the planar pose (x_k, y_k, ψ_k) is integrated according to

Table 4: Sensing and state modalities of the data-collection platform.

Modality	Number	Specification	Coverage	Native Rate
Fisheye Camera	4	1920×1536	Front/Rear/Left/Right	10 Hz
3D LiDAR	2	RoboSense Airy, 96 ch.	Front/Rear	10 Hz
IMU	1	Inertial measurements	Ego state	200 Hz
LIO/Odometry	1	6-DoF pose	Ego motion	$\sim$200 Hz
Motion State	1	v_x, v_y, ω_z	Ego motion	$\sim$20 Hz
GPS	1	Global positioning	Reference only	$\sim$2 Hz



Figure 6: **Camera–LiDAR calibration validation.** LiDAR point clouds are projected onto the calibrated fisheye camera views to qualitatively verify cross-modal alignment. Projected points are color-coded by depth along the camera axis, with red indicating nearer points and blue indicating farther points.

$$
\begin{aligned}
x_{k+1} &= x_k + \left(v_x^k \cos\psi_k - v_y^k \sin\psi_k\right) \Delta t, \\
y_{k+1} &= y_k + \left(v_x^k \sin\psi_k + v_y^k \cos\psi_k\right) \Delta t, \\
\psi_{k+1} &= \psi_k + \omega_z^k \Delta t.
\end{aligned} \tag{2}
$$

The integrated sequence is resampled to 10 Hz and transformed into the robot coordinate frame at the prediction anchor. Each annotated frame is therefore paired with the subsequent robot motion represented as a sequence of planar poses. LIO trajectories are retained separately for tasks requiring metric camera motion.

B.4 Annotation Taxonomy

B.5 Data Collection and Quality Control

Socially compliant data collection. Data collection is performed by trained human operators following predefined socially compliant navigation guidelines. The protocol covers common interactions in shared outdoor spaces, including yielding to pedestrians, maintaining safe clearance, passing obstacles on an appropriate side, and negotiating encounters with pedestrians, riders, and vehicles. These guidelines are used to improve the behavioral consistency and safety of the recorded expert trajectories.

Table 5: Annotation taxonomy: 27 object categories, 4 road-structure classes, and 20 navigation instructions in three layers.

Group	Categories
3D objects (27)	
Human	Pedestrian, Pedestrian Accessories
Mobility	Bicycle, Bicycle Rider, Scooter, Scooter Rider, Motorcycle, Motorcycle Rider, Tricycle, Tricycle Rider
Vehicle	Car, SUV, Van, Truck, Bus
Traffic Facility	Cone, Traffic Barrier, Road Barrel, Reflective Triangle, Fire Hydrant
Other	Robot Dog, Platform Cart, Construction Cart, Excavator, Forklift, Dog, Cat
Road structure Line (4)	
	Sidewalk Curb, Road Curb, Road–Sidewalk Dividing Line, Crosswalk
Navigation instructions (20)	
Primary Navigation	`follow_path`, `continue_straight_at_junction`, `take_left_branch`, `take_right_branch`, `u_turn`, `unknown_navigation`
Special Motion	`stationary`, `reversing`, `rotate_in_place_left/right`, `lane_left/right`, `lateral_left/right`, `mixed_motion`
Obstacle Interaction	`obstacle_wait`, `obstacle_avoid_left/right`, `obstacle_reroute`, `unknown_obstacle_interaction`

Annotation quality control. All annotations are created manually and pass through a dedicated review loop. Reviewers inspect annotated frames for completeness and consistency, and frames containing missing, misaligned, or inconsistent labels are returned for correction before acceptance.

Demonstration quality control. Recorded expert demonstrations additionally undergo behavior-level review. Human–robot interaction segments are inspected for consistency with the predefined socially compliant navigation guidelines. Segments containing unsafe, socially inappropriate, or inconsistent operator behavior are manually identified and excluded from the benchmark.

C Navigation Evaluation Details

Trajectory imitation metrics. Given predicted and reference trajectories, ADE and FDE measure the mean and final Euclidean position errors, respectively. AHE and FHE measure the corresponding mean and final absolute heading errors, with angular differences wrapped to $[-\pi, \pi)$.

Pseudo-closed-loop replay. The replay uses the predicted pose $\hat{\mathbf{p}}_k = (\hat{x}_k, \hat{y}_k, \hat{\psi}_k)$ at 0.1-s intervals and a rectangular ego footprint. Annotated objects are replayed as BEV rectangles by interpolating box centers and yaw between 1-Hz keyframes using persistent instance identities; an object present in only one neighboring keyframe is held fixed. Road-structure polylines from the anchor keyframe define the local walkable-area constraints. The replay is non-reactive.

- **No Collision (NC):** 0 for an at-fault collision with a person, rider, vehicle, animal, or robot; 0.5 for an inanimate object; and 1 otherwise. A collision is not at fault when the ego is stationary and another object moves into it.
- **Walkable-Area Compliance (WAC):** 0 if the ego footprint crosses a curb or a road–sidewalk dividing line outside a crosswalk, unless the recorded trajectory crosses the same boundary within the horizon; 1 otherwise.
- **Time-to-Collision (TTC):** 0 if 1-s constant-velocity propagation from any predicted step intersects a replayed agent; 1 otherwise.
- **Clearance (CL):** $\min(1, d_{\min}/0.5\,\mathrm{m})$, where $d_{\min}$ is the minimum distance to any person, rider, or animal.
- **Ego Progress (EP):** predicted progress projected onto the recorded path, normalized by the recorded-path length and clipped to $[0, 1]$; set to 1 when the recorded motion is below 0.2 m.
- **Instruction Compliance (IC):** checks instruction-specific trajectory geometry: at least 20° signed heading change for branches, U-turns, and in-place rotation; signed dominant lateral velocity for lateral or lane moves; negative longitudinal velocity for reversing; displacement below 0.2 m for stationary or waiting; passage on the instructed side for obstacle avoidance; and continued motion when the operator moved for path following or continuing straight.

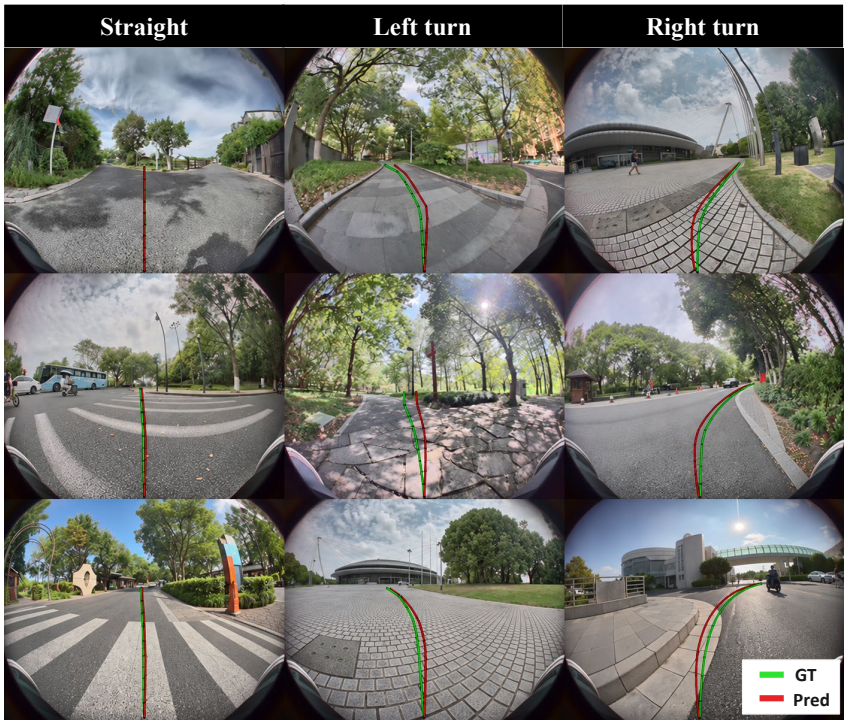


Figure 7: **Qualitative trajectory predictions with NaVILA.** Columns correspond to three navigation instructions: continue straight at junction, take left branch, and take right branch. Red curves denote predicted trajectories, while green curves denote ground-truth trajectories.

D TRAJECTORY-PREDICTION BASELINE ADAPTATION DETAILS

All trajectory-prediction experiments use the 8:1:1 clip-level train/validation/test split described in Sec. 3.5. Each model receives the same observation history, four-camera visual input, current robot state, semantic navigation instruction, and future trajectory target, and predicts the complete 10×3 Body-FLU trajectory over the future 1 s. Ground-truth 3D object and road-structure annotations are used only by the evaluator and are not provided as test-time model inputs.

InternNav and StreamVLN use four-layer Transformer trajectory decoders with Flow Matching to jointly generate the complete trajectory. NaVILA and OmniNav use two-layer Transformer decoders for parallel regression of all 10×3 trajectory variables. LingBot-VLA-v2 and OpenPI $\pi 0.5$ retain their native Flow-Matching action experts; DepthVLA retains its Flow-Matching action head adapted to the 10×3 output; and OpenVLA uses a two-layer Transformer decoder for parallel continuous trajectory regression.

SocialNav builds on Qwen3-VL-4B and augments its visual tokens with camera-identity, temporal-slot, and continuous timestamp embeddings. A four-layer ActionFormer and motion-conditioned Flow-Matching head then generate the complete trajectory.

Isaac-GR00T-N1.7 replaces the original manipulation diffusion head with a two-layer, eight-head Transformer trajectory decoder conditioned on learned horizon queries and the current motion state. The decoder predicts normalized residuals over a constant-twist prior for all 10×3 trajectory variables.

All future trajectory variables are predicted jointly rather than waypoint-wise autoregressively. Representative trajectory predictions are shown in Fig. 7.

E BEV-BASED PERCEPTION DETAILS

3D object metrics. Confidence-ranked predictions are matched one-to-one by ground-plane center distance. Box mAP averages AP over Car and Pedestrian and center-distance thresholds of $\{0.5, 1, 2, 4\}$ m. mATE, mASE, and mAOE use 2-m matching, with mASE defined as one minus box-size IoU after center and yaw alignment. We report

$$\mathrm{HDS} = \frac{5\,\mathrm{mAP} + \sum_{k\in\{\mathrm{ATE,ASE,AOE}\}} \max(0, 1 - m_k)}{8}, \tag{3}$$

where m_k denotes the corresponding detection error. HDS is computed from unscaled values; the aggregate scores in Table 3(2) are reported as percentages.

Road-structure metrics. Road Curb, Sidewalk Curb, and Road–Sidewalk Dividing Line are merged into the ordinary-line class, while Crosswalk forms the closed-contour class. Predictions and references are clipped to $[-30, 30] \times [-15, 15]$ m, sampled along arc length, and matched one-to-one using symmetric Chamfer distance. Each class AP averages the $\{0.5, 1, 1.5\}$ m matching thresholds, and line-detection mAP is the equal-weight mean of ordinary-line and crosswalk AP.

Experimental settings. The BEV split contains 46,111 training, 5,917 validation, and 5,950 test frames. All four baselines use the same split, evaluator, and common optimization schedule. The adapted configurations use a ResNet-34/FPN initialized from the same public ImageNet weights, an effective batch size of 192, and validation ordinary-line AP@0.5 m for checkpoint selection. MapTRv2 is framewise and camera-only; BEVFormer-Tiny and SparseDrive use cameras and ego pose for temporal processing; BEVFusion uses cameras and LiDAR. All reported runs use seed 42. BEVFormer-Tiny and MapTRv2 were early-stopped after 92.7% and 89.0% of their scheduled iterations, respectively.

Additional qualitative results are shown in Fig. 8.

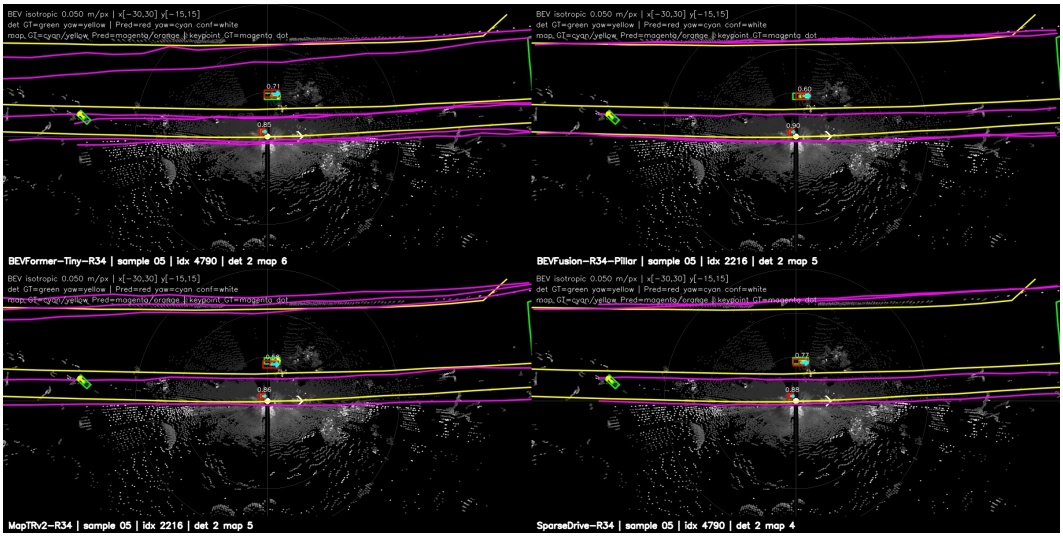


Figure 8: Qualitative line detection comparison on a simple straight-road test scene. The panels show BEVFormer-Tiny (top left), BEVFusion (top right), MapTRv2 (bottom left), and SparseDrive (bottom right). Ground-truth road geometry is shown in cyan, yellow, green, and blue; predictions are shown in magenta and orange. Predictions with confidence at least 0.25 are displayed. The candidate pool was constructed using ground-truth content only.

Table 6: World-model baselines and their evaluation interfaces. I_0 denotes the first rectified front-view frame. Native frame rates and conditioning formats are retained during generation.

Model	Model Type	Visual Input	Motion / Text Condition	FPS
Matrix-Game-3	Action-conditioned world model	I_0	Trajectory-derived keyboard/mouse actions	17
WorldStereo	Camera-controlled world model	I_0	Camera intrinsics + W2C trajectory	16
Cosmos3-Nano	Action-conditioned world model	I_0	Resampled 9-D motion sequence	10
LingBot-World	Camera-controlled world model	I_0	Interpolated camera trajectory	16
Wan2.2-TI2V-5B	Generic image-to-video	I_0	Forward-motion text prompt	24
CogVideoX1.5-5B-I2V	Generic image-to-video	I_0	Forward-motion text prompt	16

F Future Scene Prediction Details

F.1 Task Formulation and Conditioning

Given an initial rectified front-view image $\mathbf{I}_0$, a future motion condition $\mathbf{U}_{1:T}$, and an optional text condition $\mathbf{c}$, a model generates a future egocentric RGB video:

$$\hat{\mathbf{V}}_{1:T} = f_\theta\left(\mathbf{I}_0, \mathbf{U}_{1:T}, \mathbf{c}\right), \tag{4}$$

where $\hat{\mathbf{V}}_{1:T} = \{\hat{\mathbf{I}}_1, \ldots, \hat{\mathbf{I}}_T\}$.

For camera- or action-conditioned world models, $\mathbf{U}_{1:T}$ contains the recorded future ego-motion after being converted to the representation and temporal convention required by the corresponding model. For generic image-to-video models without an explicit motion-control interface, $\mathbf{U}_{1:T}$ is omitted and a fixed forward-motion text prompt is provided. These methods therefore serve primarily as visual-generation baselines rather than directly controllable world models.

F.2 Evaluation Subset

The evaluation subset contains 200 deterministic navigation clips selected from the test split according to front-camera video and trajectory synchronization requirements. Each sample contains a rectified front-view pinhole video with a horizontal field of view of 90°, the first RGB observation frame, camera intrinsics, and synchronized camera and body poses.

F.3 Model-Specific Conditioning Interfaces

We preserve the native conditioning interface of each evaluated model. Table 6 summarizes the visual input, motion or text condition, and native generation frame rate.

F.4 Temporal Normalization

The evaluated models generate videos with different native frame rates, resolutions, and rollout lengths. To ensure comparable evaluation, each generated video is divided into 16 equal temporal intervals, and the midpoint frame of each interval is deterministically selected.

Q-Align-I, CLIP-I, CLIP-IQA+, and LAION-Aes are averaged over the sampled frames. FVD uses the same 16-frame sampling procedure for both generated and real videos. Q-Align-V, DOVER-T, Motion Smoothness, and Dynamic Degree follow the temporal sampling procedures of their official implementations.

F.5 World-Model Metric Definitions

We evaluate four complementary aspects of generated-video quality.

Perceptual quality. Q-Align-I (Wu et al., 2024) evaluates frame-level perceptual quality. CLIP-IQA+ (Wang et al., 2023) provides a semantic-aware no-reference image-quality score, while LAION-Aes (LAION-AI) measures visual aesthetic quality.

Temporal quality. Q-Align-V (Wu et al., 2024) evaluates overall video quality and temporal coherence. DOVER-T (Wu et al., 2023) measures technical video quality, while Motion Smoothness and

Dynamic Degree (Huang et al., 2024) characterize temporal continuity and the amount of temporal variation.

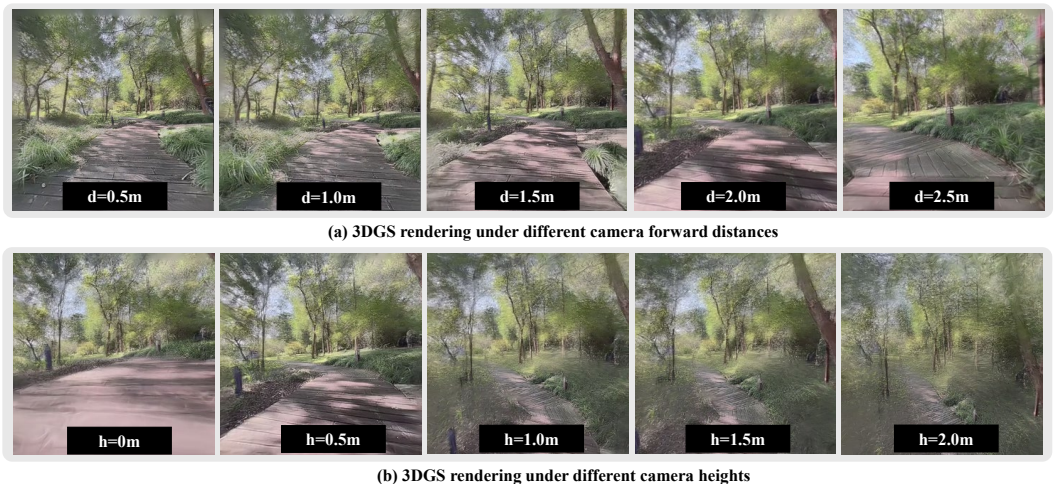
Appearance consistency. CLIP-I measures appearance consistency between generated frames and the initial observation. For each sampled generated frame, cosine similarity with the first frame of the corresponding real video is computed using an OpenCLIP ViT-B/32 image encoder (Cherti et al., 2023) and multiplied by 100.

Distributional realism. Fréchet Video Distance (FVD) (Unterthiner et al., 2018) measures the distance between generated and real video distributions using I3D video features. FVD is computed jointly over all real and generated videos for each evaluated model; lower values indicate a generated distribution closer to the real-world robot observations.

F.6 Benchmark Scope

The current evaluation primarily measures visual fidelity, temporal quality, appearance consistency, and distributional realism. It does not directly evaluate whether the generated motion exactly follows the intended future motion condition. Explicit motion-conditioning accuracy therefore remains an important extension of the benchmark.

G Multi-View Scene Reconstruction Details



(a) 3DGS rendering under different camera forward distances

(b) 3DGS rendering under different camera heights

Figure 9: **Qualitative 3DGS multi-view scene reconstrucion under viewpoint changes.** (a) Renderings from increasing forward camera offsets and (b) renderings at different camera heights within the same reconstructed scene. d and h denote the relative camera translation and height with respect to the reference viewpoint.

Baseline configurations. All five methods are evaluated using pretrained checkpoints without additional fine-tuning. ReCoSplat updates a scene through render-and-compare (Cheng et al., 2026), while YoNoSplat supports both posed and unposed reconstruction (Ye et al., 2026). AnySplat and FreeSplatter-S jointly infer camera geometry and Gaussian scene representations from unposed images (Jiang et al., 2025; Xu et al., 2025). MVSplat uses posed multi-view cost volumes (Chen et al., 2024); in our evaluation, it is supplied with motion-integrated camera trajectories.

Evaluation protocol. All methods are evaluated on rectified front-camera images from 200 `data_test` clips. Rendering quality is measured using 8-bit RGB images at 224×224 resolution. ReCoSplat, YoNoSplat, and AnySplat use 8–16 context views and are evaluated on 2,495 held-out views in total. FreeSplatter-S and MVSplat use two context views and are evaluated on 3,162 held-out views. Metric values are first averaged over held-out views within each clip and then equally averaged across clips.

Because the methods differ in context-view count, pose requirements, native reconstruction interfaces, and held-out-view counts, the reported results should be interpreted as system-level benchmark performance rather than a controlled architecture-only comparison.

Metric definitions. PSNR measures pixel-level reconstruction fidelity. SSIM (Wang et al., 2004) and MS-SSIM (Wang et al., 2003) measure local and multi-scale structural similarity, respectively. LPIPS-VGG (Zhang et al., 2018) measures learned perceptual distance, while DISTS (Ding et al., 2020) measures deep-feature structure and texture differences. FSIMc (Zhang et al., 2011) combines phase, gradient, and color similarity, whereas GMSD (Xue et al., 2014) measures spatial variation in gradient-magnitude similarity. Together, these metrics cover complementary pixel-, structural-, and perceptual-quality dimensions.

Benchmark scope. The reconstruction task evaluates observation-grounded novel-view rendering under viewpoint changes. The reported metrics do not directly measure geometric reconstruction accuracy, future scene dynamics, or closed-loop navigation safety.