Verti-Bench: A General and Scalable Off-Road Mobility Benchmark for Vertically Challenging Terrain

Tong Xu, Chenhui Pan, Madhan B. Rao, Aniket Datar, Anuj Pokhrel, Yuanjie Lu, and Xuesu Xiao
George Mason University

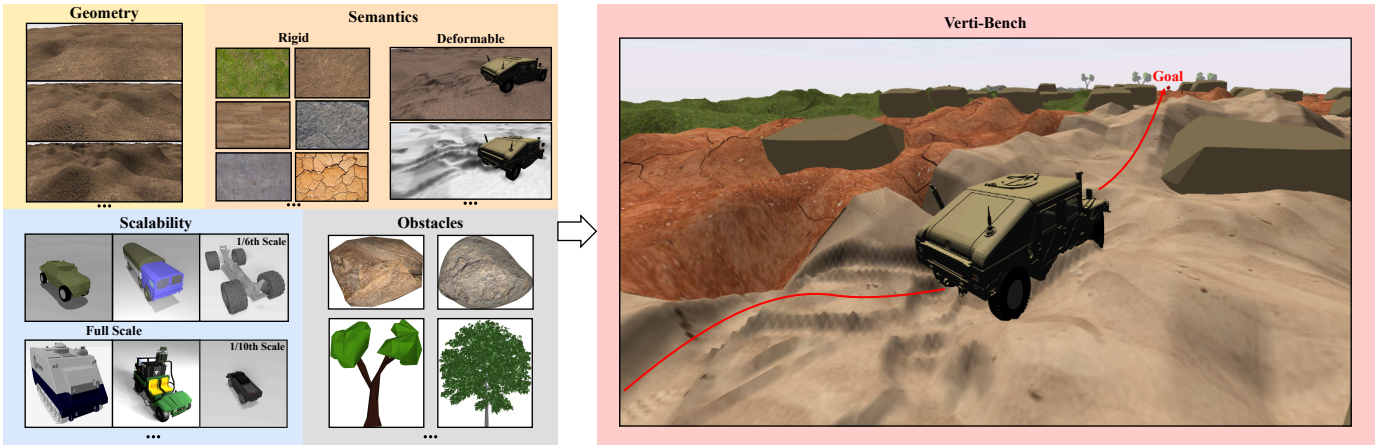

Fig. 1: Based on high-fidelity multi-physics simulation, Verti-Bench encapsulates a variety of off-road features, i.e., geometry, semantics, and obstacles, and vehicles and can be scaled to different sizes. 100 off-road environments and 1000 navigation tasks with millions of off-road features can objectively and quantitatively evaluate off-road vehicle mobility on extremely rugged, vertically challenging terrain.

***Abstract*—Recent advancement in off-road autonomy has shown promises in deploying autonomous mobile robots in outdoor off-road environments. Encouraging results have been reported from both simulated and real-world experiments. However, unlike evaluating off-road perception tasks on static datasets, benchmarking off-road mobility still faces significant challenges due to a variety of factors, including variations in vehicle platforms and terrain properties. Furthermore, different vehicle-terrain interactions need to be unfolded during mobility evaluation, which requires the mobility systems to interact with the environments instead of comparing against a pre-collected dataset. In this paper, we present Verti-Bench, a mobility benchmark that focuses on extremely rugged, vertically challenging off-road environments. 100 unique off-road environments and 1000 distinct navigation tasks with millions of off-road terrain properties, including a variety of geometry and semantics, rigid and deformable surfaces, and large natural obstacles, provide standardized and objective evaluation in high-fidelity multi-physics simulation. Verti-Bench is also scalable to various vehicle platforms with different scales and actuation mechanisms. We also provide datasets from expert demonstration, random exploration, failure cases (rolling over and getting stuck), as well as a gym-like interface for reinforcement learning. We use Verti-Bench to benchmark ten off-road mobility systems, present our findings, and identify future off-road mobility research directions. Verti-Bench project website can be found at https://cs.gmu.edu/~xiao/Research/Verti-Bench.**

I. Introduction

Off-road autonomous mobile robots present unique opportunities in search and rescue, environment monitoring, scientific exploration, and other application domains. However, off-road autonomy presents unique challenges to both robot perception and vehicle mobility that distinguish from structured on-road environments [1, 2, 3], such as variable geometry, deformable surfaces, and natural obstacles. Considering recent advancement, standardized benchmarks for off-road autonomy are necessary to objectively and quantitatively evaluate and compare the progress of the off-road robotics community.

Unlike off-road perception, evaluating off-road mobility is more difficult. A plethora of off-road perception datasets [4, 5, 6, 7, 8, 9, 10, 11] are available to provide static ground truth labels to evaluate against perception systems' outputs, since the robot actions can be recorded and fed into the perception systems, not generated by them (if actions are necessary at all). However, mobility systems produce new actions to drive robots to different states than those collected in the dataset, i.e., distribution shift, and thus cannot be evaluated by comparing against a static dataset. Therefore, vehicle-terrain interactions need to be unfolded during off-road mobility evaluation, which creates difficulty in standardization across research groups.

Considering a lack of standard off-road mobility evaluation, researchers currently develop their own benchmarks to evaluate their mobility systems and re-implement previous systems to compare against. Such a practice, however, leads to ad-hoc evaluation in a few aspects: The evaluation platforms, either a physical or a simulated robot, vary in terms of robot size, weight, actuation mechanism, and/or levels of off-road

physics simulation fidelity; The evaluation environments range from handcrafted off-road simulation features [12, 13, 14, 15, 16, 17], small-scale indoor testbeds [18, 19, 20, 21], enclosed outdoor tracks [22, 23, 24, 25, 26], and large-scale real-world testing facilities [6, 27, 5]; Re-implementation of previous approaches on one's own robot is not only laborious, but also subject to misinterpretation of implementation details. Therefore, a standard off-road mobility benchmark which is general and scalable to all these aspects is desired for the off-road mobility research community.

In this work, we present Verti-Bench (Fig. 1), a general and scalable off-road mobility benchmark that focuses on extremely rugged, vertically challenging terrain with a variety of unstructured off-road features. The main goal of Verti-Bench is to directly compare the performances of different off-road mobility systems. Based on a high-fidelity multi-physics dynamics simulator, Chrono [28], Verti-Bench encapsulates variations in four orthogonal dimensions: Using the Sliced Wasserstein Autoencoder (SWAE) [29] and real-world off-road terrain data, off-road geometry is represented as a diverse set of 2.5D elevation maps; Ten terrain semantics classes, including seven rigid and three deformable, are designed with different distributions of physics parameters, e.g., friction coefficient, cohesive effect, and soil stiffness; Different types of natural obstacles, e.g., boulders and vegetations, are randomly distributed based on different densities; A set of off-road vehicles with different scales (from 1/10th to full scale) and actuation mechanisms (4-, 6-, and 8-wheeled and 2-tracked chassis, single- and double-wishbone, multilink, toebar leaf-spring, and special tensioning suspensions, as well as pitman-arm, rack-and-pinion, toebar, bellcrank/rotary arm, and differential steering) are provided, with the possibility of adding customized vehicles. Using Verti-Bench, we also provide datasets from expert demonstration, random exploration, failure cases (rolling over and getting stuck), as well as a gym-like interface for Reinforcement Learning (RL). We use Verti-Bench to benchmark ten off-road mobility systems, present our findings, and identify future off-road mobility research directions. To summarize, our contributions are:

- A general off-road mobility benchmark on vertically challenging terrain with 100 off-road environments and 1000 navigation tasks scalable to various vehicle types;
- Millions of off-road terrain features including geometry, semantics (rigid and deformable), and obstacles;
- Various datasets and a RL interface to facilitate data-driven off-road mobility;
- Findings and future research directions based on benchmark results of various off-road mobility systems.

II. Related Work

In this section, we review the field of off-road autonomy and current evaluation practices to motivate Verti-Bench.

A. Off-Road Autonomy

Starting from the DARPA LAGR program [1], roboticists have started developing autonomous systems to operate in off-road enviornments. Compared to indoor or on-road operations, off-road environments pose significant challenges during the entire sense-plan-act loop: Robot state estimation systems, such as visual inertial odometry [30] and simultaneous localization and mapping [31], are easily affected by the unstructured visual features from the off-road environments as well as noisy inertial signals caused by the extensive vehicle vibrations; Perception is more than a geometric problem, i.e., free vs. obstacle spaces, as in indoor or on-road settings, and requires the consideration of semantics, e.g., gravel vs. grass. Both terrain geometry and semantics need to be represented [32, 33, 34, 35] for downstream planning and control tasks; Off-road planners and controllers need to reason beyond collision avoidance and consider factors such as vehicle stability [36, 37], wheel slippage [38, 39, 40], and terrain traversability [41, 42, 43, 44], oftentimes in $\mathbb{SE}(3)$ instead of $\mathbb{SE}(2)$ [19] due to the uneven off-road terrain surfaces [45, 19]. The recent increase in interest in off-road autonomy [46] necessitates standard benchmarks to objectively and quantitatively evaluate research progress from the entire community, which is the motivation behind Verti-Bench.

B. Evaluating Off-Road Perception

Off-road perception research still dominates the majority of the body of off-road autonomy work [46]. Fortunately, evaluating off-road perception can be mostly carried out on pre-collected, static datasets in a vehicle-agnostic fashion. For both evaluation and training of perception systems in a data-driven manner, a variety of off-road perception datasets are available for research in semantics segmentation [4, 5, 8], freespace detection [7], place recognition [9], traversability estimation [10], and map reconstruction [11]. Given a new off-road perception system taking the recorded sensor data and robot actions as inputs, its outputs can be simply compared against the ground truth labels provided by those datasets. A performance metric, e.g., segmentation accuracy, detection rate, recognition count, traversability correctness, and reconstruction precision, can then be computed to quantify how well the new perception system compares against others. Unfortunately, off-road mobility evaluation cannot be conducted on such pre-collected, static datasets. Notice that while off-road dynamics datasets [6] can be used to learn off-road navigation models, they cannot be used to evaluate off-road mobility.

C. Evaluating Off-Road Mobility

To the best of our knowledge, no standard off-road mobility benchmarks currently exist in the literature. Unlike off-road perception evaluation, mobility evaluation requires the vehicle-terrain interactions to be unfolded, since a different action will lead the robot into a different state absent from the dataset. Such a distribution shift necessitates mobility evaluation to be based on the unfolded, not pre-collected, vehicle trajectories.

To unfold vehicle-terrain interactions for objective off-road mobility evaluation, the vehicle platforms and terrain properties are required to be standardized, which, unfortunately, vary significantly across different research groups: Robots of

different sizes, weights, actuation mechanisms (e.g., wheeled, tracked, steered, and differential-driven) are used to compare new mobility systems with existing ones. The latter is usually customized to fit for a different robot, potentially causing misinterpretation of implementation details; Different simulators, e.g., Gazebo [12], IsaacGym [13], Unreal [14], and Unity [17], are leveraged to train and evaluate off-road mobility systems. To improve simulation speed, most of those simulators do not focus on physics fidelity, which is crucial for off-road mobility on extremely rugged and deformable surfaces; In the real world, small-scale indoor testbeds have been set up with foam, rocks, plywood, and pipes to emulate vertically challenging terrain in the wild [18, 19, 20, 21]; Enclosed outdoor off-road tracks have been constructed to evaluate aggressive autonomous driving [22, 23, 24, 25, 26]; A few research groups have access to large-scale off-road testing facilities with full-size vehicles [6, 27, 5].

With such diverse setups, an objective evaluation and comparison across different off-road mobility systems become infeasible. Verti-Bench, based on a high-fidelity multi-physics dynamics simulator, is hence motivated to fill such a gap of missing standard off-road mobility benchmarks. Notice that Verti-Bench is not meant to replace existing evaluation setups, but to complement them with a new standardized option to facilitate fair comparison across off-road mobility systems.

III. Verti-Bench

We present Verti-Bench's core high-fidelity multi-physics dynamics engine, diverse set of off-road features, including wide-ranging geometry, physics-grounded semantics, and natural obstacles, scalability to a variety of vehicle platforms, and standardized metrics to quantify off-road mobility performance. We also discuss various datasets we collect using Verti-Bench to complement real-world off-road mobility data to develop data-driven systems.

A. Simulation

Verti-Bench is based on Project Chrono [28], a high-fidelity multi-physics dynamics engine with a platform-independent open-source design implemented in C++ with a Python version, PyChrono. Compared to other commonly used robotics simulators (e.g., Gazebo, Unreal, Unity, PyBullet, MuJoCo, and IsaacGym with well-known physics limitations especially for differential-drive mobile robots [47]), Chrono is especially suitable to simulate complex off-road vehicle-terrain interactions involving suspension, tire, track, and terrain deformation, varying terrain contact friction, vehicle weight distribution and momentum, motor, powertrain, transmission, and wheel torque characteristics, aggressive vehicle poses with all six Degrees of Freedom (DoFs), etc. In Chrono, vehicle systems and terrain properties are made of rigid and flexible/compliant parts with constraints, motors and contacts, along with three-dimensional shapes for collision detection.

One point worth noting is that Verti-Bench's choice of Chrono as its core simulator is primarily due to its high-fidelity multi-physics dynamics, a vital aspect for off-road mobility evaluation. However, Chrono is not the best simulator for photorealism, one focus of off-road perception simulation, which is out of scope of Verti-Bench. For the perception components, Verti-Bench has standard interfaces to provide ground truth elevation and semantics maps and obstacle occupancy grids. Another point is that Chrono is not yet GPU-accelerated. Combined with the high computation required for high physics fidelity, Chrono can only provide slightly faster-than-real-time simulation, depending on the complexity of the simulated environments (e.g., areas of deformable terrain and number, size, and number of mesh vertices of obstacles). Therefore, despite its intended efficient usage in off-road mobility evaluation, learning off-road mobility is expected to take a significant amount of training time with Verti-Bench (e.g., using our provided gym-like RL interface).

In Chrono, each of the 100 Verti-Bench full-scale environments is constructed as a 129m$\times$129m world, with a resolution of 1m per pixel. Each environment can be down-scaled to cater vehicles of different sizes, e.g., 1/6th or 1/10th scale. Each pixel contains geometry, semantics, and obstacle information (details below). For each of the 100 environments, ten pairs of start and goal locations separated by 120 m are distributed in a circular manner, leading to a total of 1000 navigation tasks.

B. Geometry

Real-world off-road terrain is characterized by various geometry in terms of elevation changes, e.g., slopes, hills, ditches, gullies, ravines, and other form of undulations. Some of such terrain can be traversed by certain types of off-road vehicles, while others cannot. Autonomous off-road mobility systems need to decide which of them can be attempted with what vehicle maneuvers. For example, a steep slope with low friction cannot be traversed at low speeds, but large vehicle momentum by high speeds at the bottom can help the vehicle ascend the top; Approaching a deep ditch quickly may get the vehicle stuck due to extensive suspension depression at the bottom, but slowly negotiating through is possible to mitigate suspension travel in order to maximize clearance.

Therefore, the geometry of Verti-Bench environments is represented as 2.5D elevation maps created by SWAE [29] and real-world elevation data. To be specific, we physically construct vertically challenging terrain with boulders and rocks and use a Microsoft Azure Kinect RGB-D camera to create elevation maps of different real-world terrain surfaces [48]. We then use SWAE [29], a scalable generative model that captures the rich and often nonlinear distribution of high-dimensional data, as a feature extractor to reduce the dimension of the real-world elevation maps while preserving the original elevation information in a latent space, from which samples can be drawn to generate new elevation maps that resemble real-world vertically challenging terrain. To further introduce diversity and quantification of Verti-Bench geometry, we scale the output of the trained SWAE to 30%, 60%, and 100% and denote them as low, medium, and high elevation level (Fig. 2 top). Each Verti-Bench environment is generated with 1/3 probability of each elevation level. Fig. 2 bottom

shows the histogram of elevation values of all three levels of terrain geometry. High elevation environments also have the largest variance (most rugged terrain), while low elevation environments are smoother.

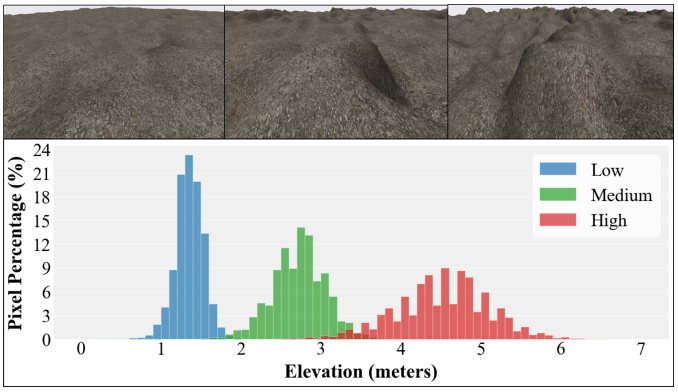

Fig. 2: Top: Low, Medium, and High Elevation Maps; Bottom: Elevation Histograms across Three Elevation Levels.

C. Semantics

In addition to geometry, off-road terrain also presents challenges in terms of semantics and its associated physical vehicle-terrain contact features, such as friction, slip, and deformability. For example, an autonomous off-road mobility system should be aware that when driving through an icy or sandy laterally inclined slope with low friction or high deformability, sideway sliding downhill or wheel sinkage due to imbalanced load and then rollover is possible, respectively.

Therefore, we also add ten different semantics classes to the terrain elevation. We design seven rigid and three deformable semantics classes with different textures and distributions of physics parameters (Fig. 3). To be specific, the seven rigid semantics classes, i.e., grass, wood, gravel, dirt, clay, rock, and concrete, associate with a normal distribution of friction coefficient. When a pixel is sampled to be a certain terrain type, its friction coefficient is sampled from the corresponding distribution. We fix the restitution coefficient to 0.01 for all rigid semantics classes. For the three deformable terrain classes, i.e., snow, mud, and sand, we adopt the deformable Soil Contact Model (SCM) based on the Bekker-Wong model [49] to simulate terrain deformation after wheel interaction: SCM presents the underlying terrain by a 2D grid and assumes each cell can only be displaced vertically and does not maintain any history other than the current vertical displacement. We hard-code three sets of physics parameters, including cohesive effect, soil stiffness, and hardening effect, for three different deformability levels, i.e., soft, medium, and hard. Verti-Bench also provides terrain with granular materials. But due to the slow simulation speed when simulating thousands of particles, it is only reserved for special evaluation circumstances where granular materials must be simulated and simulation speed is not of concern. All statistics of the ten terrain semantics classes can be found in Fig. 3.

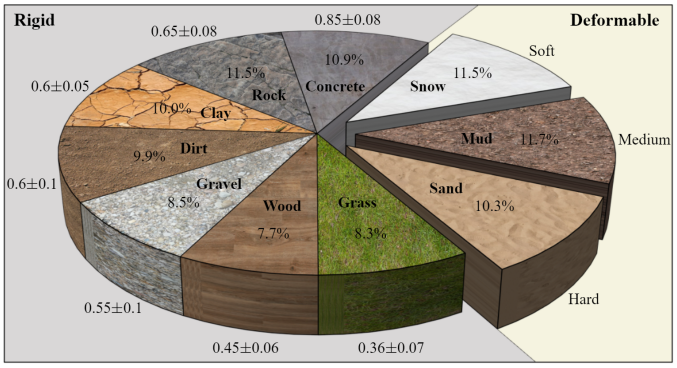

Fig. 3: Seven Rigid (percentage, mean and variance of friction coefficient, and texture) and Three Deformable (percentage, deformability, and texture) Terrain Semantics.

To create various terrain semantics while maintaining simulation efficiency, each 129×129 Verti-Bench environment is first partitioned into a 16×16 grid, with each grid cell as a 9×9 patch (one overlapping pixel between every pair of patches to assure connectivity). To emulate real-world continuous terrain patches with same semantics and similar physical properties, we employ a cluster-based approach, where cluster centers are sampled from the environment. Each patch is then associated with its nearest cluster center using Euclidean distance. For all patches associated with a cluster center, the same semantics class is sampled and corresponding texture assigned, with each patch's physical property sampled from a predetermined distribution (Fig. 3). This approach creates natural physical variations within every region of the same semantics class while maintaining semantics diversity across regions.

D. Obstacles

Undulating geometry and varying semantics require off-road mobility systems to understand fine-grained vehicle-terrain interactions when driving on them. Off-road obstacles, like large boulders or trees, exist in real-world off-road environments, which are simply beyond vehicles' mechanical capabilities and hence need to be avoided. We also include natural obstacles in Verti-Bench to pose challenges to obstacle avoidance systems. For example, a large boulder triple the size of the vehicle is completely non-traversable, while a steep hill as part of the terrain may or may not be ascended with the right maneuver. We add natural obstacles as instances of the former. To further promote variation, we randomly sample the locations and types (different sizes of boulders or trees) of 10, 20, and 40 obstacles to place on each 129×129 Verti-Bench environment, denoted as sparse, medium, and dense for obstacle distribution. We resample a new obstacle if the old one is within 10 m of another obstacle, a start, or a goal. Assuming a holonomic point-mass vehicle, we also provide pre-planned global paths leading from start to goal locations and avoiding obstacles. Fig. 4 shows three examples of sparse, medium, and dense obstacle distributions and their corresponding global paths.

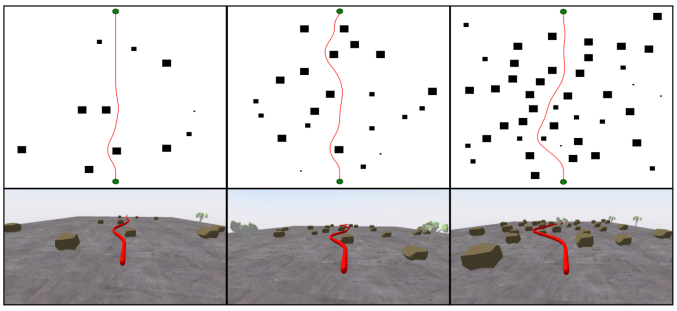

Fig. 4: Top: Sparse, Medium, and Dense Obstacles (black) and Global Paths (red) between Start and Goal (green). Bottom: Corresponding simulation scenario in Verti-Bench (elevation and semantics are removed for obstacle clarity).

E. Vehicles

We also provide a set of vehicle platforms in Verti-Bench, with the possibility of adding new customized ones in the future, so that different off-road mobility systems can be evaluated on standardized vehicles. Compared to simplified vehicles in existing simulators, the Verti-Bench vehicles are more sophisticated and articulated, including engine/motor, drivetrain, transmission, suspension, steering mechanism, and wheel torque, whose responses to complex terrain interactions are simulated. To be specific, Verti-Bench provides nine types of off-road vehicles, which are sourced from Project Chrono [28], open-source real and simulated research platforms [50], and custom-created vehicles using 3D scanning and modeling (with a Creality CR-Scan Raptor 3D scanner) of real-world scaled vehicles (Fig. 5). Those vehicles vary in terms of scale (1/10th, 1/6th, and full scale), chassis (4-, 6-, and 8-wheeled and 2-tracked), suspension (single- and double-wishbone, multilink, toebar leaf-spring, and special tensioning), steering (pitman-arm, rack-and-pinion, toebar, bellcrank/rotary arm, and differential), and tires (rigid and handling, excluding FEA-based models due to significantly reduced simulation speed). All vehicles, regardless of their sources, are implemented as native C++ classes in Chrono's C++ framework. The C++ implementations are then compiled and exposed to Python through SWIG-generated bindings.

F. Metrics

Verti-Bench automatically computes a set of standard metrics to quantify off-road mobility performance, while additional metrics can be customized as needed: A successful completion is defined as the robot reaching within 10 meters of the goal in less than 60 seconds. A successful trial may include contacts with obstacles. Success Rate captures the percentage of successful trials over total number of attempts; Traversal Time indicates how long it takes to finish a successful traversal; Roll and Pitch describes the stability of the vehicle during traversal, whose raw and absolute values can be used to derive mean, variance, and maximum; Vehicle actions, such as throttle and steering, can be used as metrics to quantify energy consumption, planning confidence, and path smoothness, along with other metrics. Verti-Bench provides infrastructure to save raw vehicle-terrain interaction data as well as to compute performance metrics.

G. Datasets

Using Verti-Bench, we also collect a few datasets to facilitate future data-driven off-road mobility research. Current datasets are collected on the High Mobility Multipurpose Wheeled Vehicle (HMMWV, Fig. 1 right), while future data collection can expand to different vehicles.

1) Expert Demonstration: A team of four human operators collect 4 hours of expert demonstration of successfully driving the off-road vehicle in different Verti-Bench environments. We filter out all failure cases, e.g., vehicle rollover and getting-stuck, to maintain high demonstration quality.

2) Random Exploration: To facilitate off-road kinodynamics learning [6], we collect a random exploration dataset on different off-road terrain by driving the off-road vehicle with sinusoidal steering and 2 m/s speed commands for ten hours. Each data collection trial terminates if the vehicle rolls over, gets stuck, or reaches the environment boundaries.

3) Failure Cases: To enable future data-driven off-road mobility by preventing vehicle failures, we also curate a dataset of failure cases by providing the last ten seconds of trajectory before the vehicle rolls over or gets stuck. We divide all failure cases into two categories: (1) Rollover: the robot has a $> 30°$ roll angle at the end of a failed trial; (2) Stuck: the robot does not move more than 1 m during the last 10 seconds of a failed trial (with $\leq 30°$ roll). The failure cases can be used to learn high-cost regions to be avoided in data-driven mobility systems.

4) RL Interface: Although not being a main purpose of Verti-Bench, we also provide a gym-like RL interface so that vehicles can learn off-road mobility through trial-and-error experiences in Verti-Bench. Vehicle states and actions, as well as reward functions, can be customized by our interface, which communicates with existing RL algorithm implementations. Verti-Bench provides interfaces of different inputs to the mobility systems, such as elevation and semantic maps, robot states, and raw sensor data. Users can choose appropriate inputs based on their evaluation needs.

IV. Evaluation and Discussions

We evaluate ten off-road mobility systems using Verti-Bench, ranging from purely classical, end-to-end learning, and hybrid systems. We present and discuss our evaluation results and point out future research directions.

A. Off-Road Mobility Systems for Evaluation

The three classical off-road mobility systems include

- PID: A controller that takes a local goal 10 m away from the robot on the global path and minimizes the error angle between the desired and vehicle heading by regulating the steering and maintaining a 3 m/s speed;
- Elevation Heuristics (EH): A controller that splits the elevation map in front of the current robot pose to five

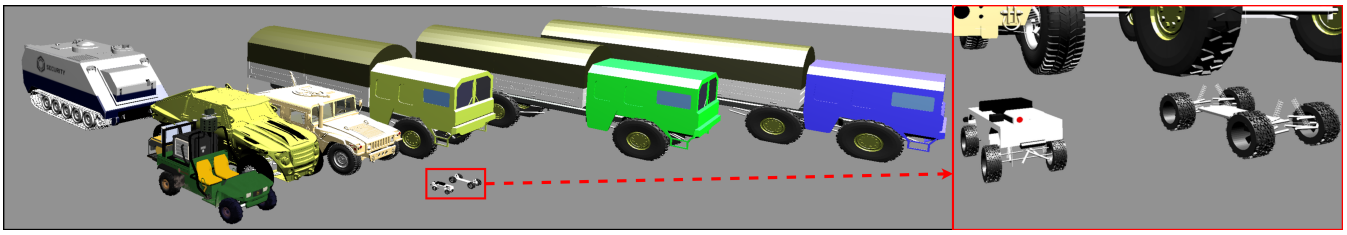

Fig. 5: Verti-Bench Vehicles with Different Scale (1/10th, 1/6th, and full scale), Chassis (4-, 6-, and 8-wheeled and 2-tracked), Steering (pitman-arm, rack-and-pinion, toebar, bellcrank/rotary arm, and differential), and Tires (rigid and handling).

regions and drives toward the region with the most similar mean to the current terrain patch and lowest variance;
- MPPI: An MPPI-based [51] planner that uses a 2D bicycle model for trajectory rollout and obstacle avoidance.

The three systems based on end-to-end learning include
- RL: A RL policy learned from trial and error [15];
- MCL: A RL policy learned from a manually designed curriculum [15];
- ACL: A RL policy learned using Automatic Curriculum Learning [52].

The four hybrid (classical and learning) systems include
- WMVCT: A planner based on a decomposed 6-DoF kinodynamic model (bicycle model for x, y, and yaw, elevation map for z, and neural network prediction for roll and pitch) [53];
- MPPI-6: An MPPI-based planner with a learned full 6-DoF kinodynamic model for trajectory rollout [45];
- TAL: An MPPI-based planner with a 6-DoF kinodynamic model that learns to attend to specific terrain patches [54];
- TNT: An MPPI-based planner that samples based on traversability and then unfolds 6-DoF kinodynamics [55].

We reach out to the authors of the original papers for their implementations of their mobility systems. Considering Verti-Bench's focus on off-road mobility evaluation, we make minimal modifications to their implementations to interface with Verti-Bench so that their mobility systems are no longer dependent on any perception system. For example, visual odometry inputs are replaced with ground truth vehicle states from Verti-Bench; Real-world elevation mapping systems are skipped by directly providing their systems with ground truth Verti-Bench elevation maps; All learning systems and components in the evaluation are not trained in Verti-Bench.

B. Evaluation Results and Discussions

The evaluation results are shown in Fig. 6, including percentage of Succuss Rate and mean and variance of Traversal Time, Roll, and Pitch with respect to three types of elevation level, terrain semantics, and obstacle density for all systems and tasks. We also present an analysis of failure cases shown in Table I. All 1000 Verti-Bench navigation tasks have been used to evaluate each mobility system (no task has been used for training). A complete evaluation of each system requires approximately 10 hours. All task configurations have been documented in YAML files, which can be used by external research teams to replicate and expand these evaluations. Additionally, we provide our terrain generation pipeline, allowing researchers to extend or customize environmental parameters according to their specific research objectives and experiment requirements.

TABLE I: Failure Case Analysis of 1000 tasks with Ten Mobility Systems

	PID	EH	MPPI	RL	MCL	ACL	WMVCT	MPPI-6	TAL	TNT
Rollover (%)	14.7	16.3	17.9	14.3	16.9	15.4	18.9	18.0	17.0	17.5
Stuck (%)	36.5	39.4	40.9	75.7	41.3	48.0	39.6	35.0	24.6	25.1
Success (%)	48.8	44.3	41.2	10.0	41.8	36.6	41.5	47.0	58.4	57.4

In general, navigation performance significantly declines with increasing elevation levels, deformable surfaces, and obstacle densities, including reduced Success Rate and increased Traversal Time, Roll, and Pitch. In addition to mean, the variance of Roll and Pitch also drastically increases, indicating much less stable vehicle chassis when traversing high elevation, deformable, and obstacle-dense environments. Among the three categories, end-to-end learned mobility systems achieve the worst performance, while hybrid systems outperform the other two in general. While ACL is expected to outperform MCL, the results suggest otherwise. Such results indicate that end-to-end learning methods, trained from other sources, still have much room for improvement in terms of generalization in Verti-Bench. Considering failure cases, the systems fail more frequently due to getting stuck than rolling over. While the failure rates due to rollover are relatively consistent across all systems, failure rates due to getting stuck differ significantly, with RL getting stuck 75.7% of time, compared to TAL and TNT with the lowest getting-stuck rates (24.6% and 25.1% respectively).

In terms of elevation, for hybrid systems, TAL and TNT are the two top performing planners among all systems overall, achieving the highest Success Rate in all cases and lowest Roll and Pitch angle in most cases. WMVCT and MPPI-6 achieve good Success Rate in high elevation environments, but with large Roll and Pitch. Classical planners perform in between their end-to-end and hybrid counterparts. PID, due to its simplicity and robustness, performs very well in low elevation environments, with EH catching up on Success Rate when facing higher elevation. MPPI does not perform well in

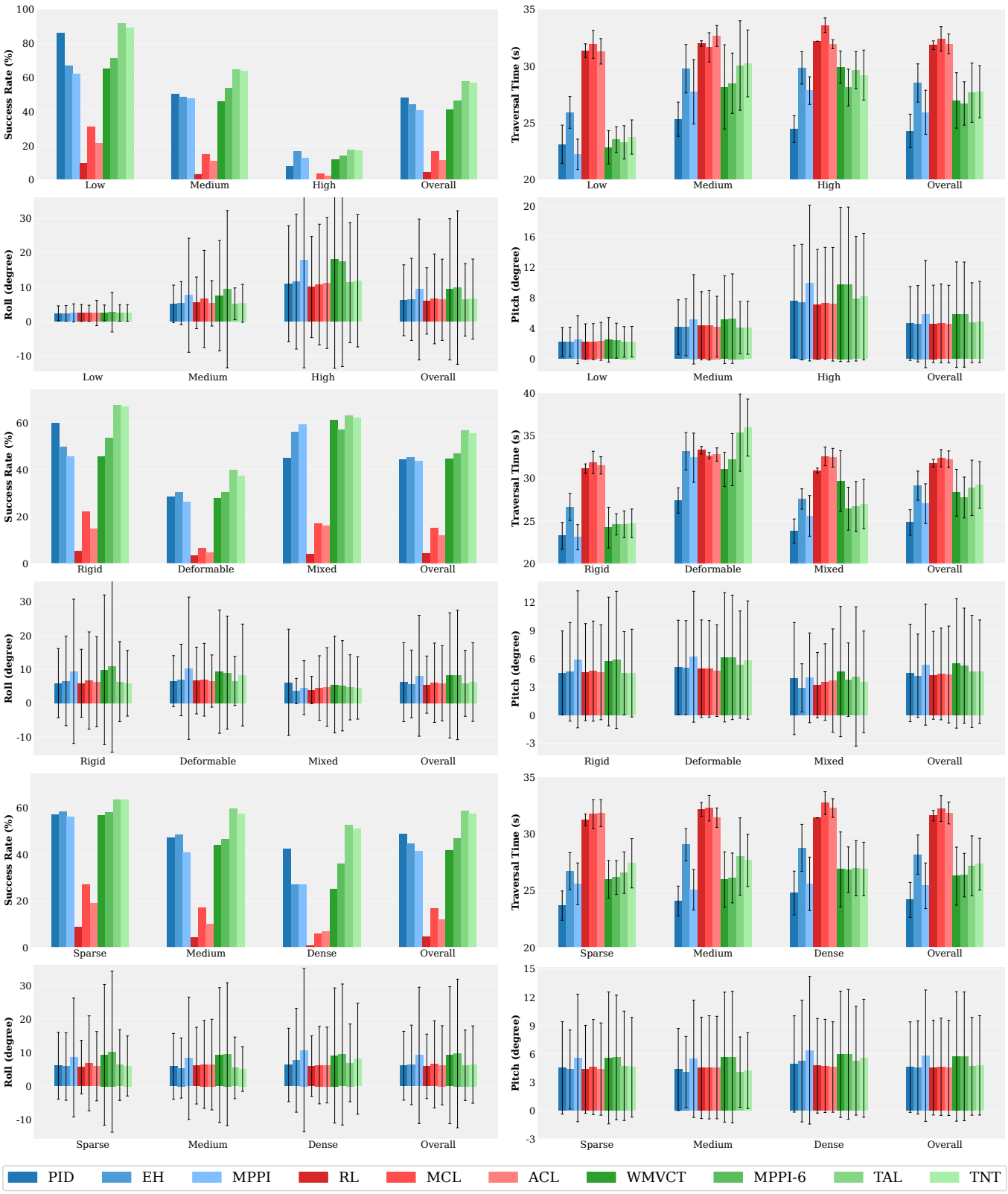

Fig. 6: Success Rate, Traversal Time, Roll, and Pitch with respect to Elevation Level (top), Terrain Semantics (middle), and Obstacle Density (bottom) of Ten Off-Road Mobility Systems on 1000 Navigation Tasks.

most cases and only outperforms PID in terms of Success Rate in high elevation environments. Notice that Traversal Time is only averaged over successful trials and thus only indicates how fast a mobility system is given navigation success.

For terrain semantics, all systems perform best on rigid terrain, with TAL and TNT achieving highest Success Rates above 60%. Performance drops significantly on deformable surfaces, with even the highest only achieving around 40% Success Rates while end-to-end systems struggle below 10%. Mixed terrain results fall between rigid and deformable terrain. Deformable surfaces also show increased roll and pitch variance across all systems, indicating less stable navigation and highlighting the challenge of modeling vehicle dynamics on unpredictable surfaces.

As obstacle density increases, Success Rate declines and Traversal Time increases across all systems, with the performance gap between hybrid and other systems widening in dense environments. Obstacle density does not directly affect vehicle stability, showing similar Roll and Pitch mean and variance.

Our evaluation results indicate the potential of hybrid mobility systems to tackle vertically challenging terrain by combining the best of both worlds of classical and learning approaches. The overall success of TAL and TNT indicates the importance of an accurate 6-DoF kinodynamic model enabled by sophisticated learning techniques in conjunction with a sampling-based motion planner. MPPI-6, with a 6-DoF kinodynamic model based on a simplistic neural network, underperforms TAL and TNT, while the inaccuracies introduced by WMVCT's efficient 6-DoF decomposition lead to the worst mobility performance among hybrid systems. The degraded performance of all hybrid systems when facing high elevation, deformable surfaces, and dense obstacles motivates further research, potentially to both increase the kinodynamic modeling accuracy and improve the sampling-based motion planner. On the other hand, it is surprising to see the superior performance of the simple PID planner in low elevation environments compared to the more sophisticated EH, whose advantage only starts to slightly emerge in high elevation environments. This observation reveals a tradeoff between system complexity and performance when facing simple environments. One potential future research direction is to develop off-road mobility systems composed of multiple planners with different complexities and specialties to fit different environments [56]. Lastly, research of end-to-end learning approaches, despite their recent success in relatively benign indoor or on-road enviornments, still needs to focus on robustly generalizing to out-of-distribution scenarios, which are very common to encounter in off-road environments.

V. Real-World Validation

To validate the Verti-Bench evaluation results, we deploy one representative mobility system from each of the three classes, i.e., PID for classical, ACL for end-to-end, and TNT for hybrid, on a physical 1/10th scale open-source Verti-4-Wheeler robot [19] on an off-road mobility testbed constructed by rocks, foam, grass, and wood, presenting different geometry, semantics, and obstacle features (Fig. 7). The testbed is constructed in two different configurations to include low and high elevation in order to validate the cross-elevation evaluation results from Verti-Bench.

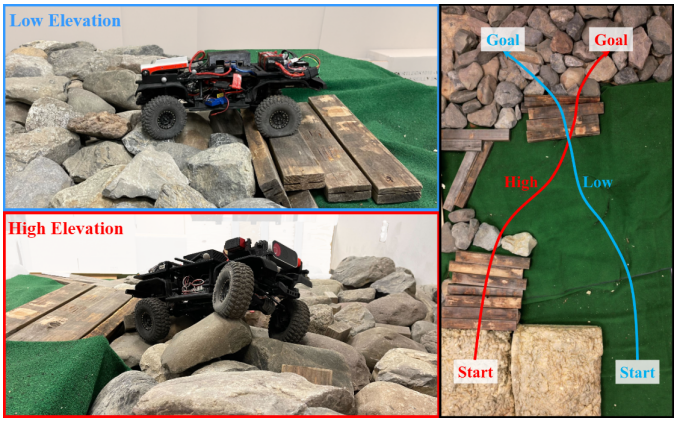

Fig. 7: Physical Off-Road Testbed Similar to Verti-Bench.

TABLE II: **Physical Validation of PID, ACL, and TNT:** Success Rate, Traversal Time, Roll, and Pitch.

Low Elevation	PID	ACL	TNT
Success Rate ↑	**5/5**	3/5	**5/5**
Traversal Time ↓	**6.64s±1.09s**	7.03s±0.47s	8.90s±0.76s
Roll ↓	6.45°±5.20°	6.20°±4.13°	**6.02°±4.92°**
Pitch ↓	5.45°±3.37°	6.34°±3.32°	**4.74°±3.23°**
High Elevation	**PID**	**ACL**	**TNT**
Success Rate ↑	3/5	2/5	**5/5**
Traversal Time ↓	**11.00s**±1.00s	16.00s±0.72s	17.50s±1.95s
Roll ↓	10.14°±8.96°	13.30°±12.72°	**7.12°±6.65°**
Pitch ↓	**7.61°±5.46°**	9.78°±6.76°	9.26°±8.41°

Table II shows the physical validation results of the three mobility systems on both low and high elevation testbeds. In general, the trend of the physical experiment results matches with that of Verti-Bench evaluation results: On the low elevation testbed, both PID and TNT are able to finish all five trials, while ACL still suffers from poor generalization. PID is still the fastest due to a lack of consideration of elevation. The roll and pitch angles are small and stable for all systems due to the lower and smoother terrain elevation; On the high elevation testbed, the difference between TNT and PID starts to emerge. TNT succeeds all five trials, while PID fails two. TNT and PID also exhibit smaller roll and pitch angle respectively. PID is still the fastest. ACL, similar to all previous cases, fails three trials and experiences the largest roll and pitch angles. Notice that due to the difficulty in conducting physical experiments, we only limit to physically evaluating three systems and five trials each, totaling 30 trials. This contrast against our 10000 trials (ten systems, 1000 trials each) in Verti-Bench, which can achieve much better statistical significance, further suggests the utility of Verti-Bench to evaluate off-road mobility.

VI. Limitations

Despite being a general and scalable benchmark, Verti-Bench still has a few limitations. Due to the high requirement to compute high-fidelity physics and a lack of GPU acceleration, Verti-Bench can only achieve near-real-time simulation speed, with real time factor ranging between 0.4 and 1.5 (faster and slower than real time respectively) depending on simulation complexity. Integrating with GPU accelerators to increase benchmarking efficiency is an important next step. Verti-Bench aims to evaluate off-road mobility and assumes ground truth perception is available to the mobility system. However, such an assumption does not hold in the real world. Future work will add realistic perception noises and test the robustness of mobility systems when facing imperfect vehicle state estimation, elevation and semantics mapping, and obstacle detection. Another direction of expanding the current Verti-Bench is to create more complex real-world counterparts than the current small-scale physical testbed so that the sim-to-real gap can be more extensively studied to further validate the efficacy and improve the fidelity of Verti-Bench evaluation.

VII. Conclusions

We present Verti-Bench, a general and scalable off-road mobility benchmark for vertically challenging terrain. Based on Chrono, a high-fidelity multi-physics dynamics engine, Verti-Bench includes 100 off-road enviornments and 1000 navigation tasks with millions of off-road terrain features covering geometry, semantics, and obstacles. Verti-Bench can also scale to off-road vehicles of different sizes, weights, chassis, suspensions, and steering mechanisms. Standardized metrics and datasets are provided to quantify off-road mobility performance and facilitate data-driven mobility. Ten off-road mobility systems are evaluated in Verti-Bench, whose results are further validated on a physical testbed. We also point out future research directions to improve off-road mobility.

Acknowledgments

This work has taken place in the RobotiXX Laboratory at George Mason University. RobotiXX research is supported by National Science Foundation (NSF, 2350352), Army Research Office (ARO, W911NF2320004, W911NF2420027, W911NF2520011), Air Force Research Laboratory (AFRL), US Air Forces Central (AFCENT), Google DeepMind (GDM), Clearpath Robotics, Raytheon Technologies (RTX), Tangenta, Mason Innovation Exchange (MIX), and Walmart.

References

[1] L. D. Jackel, E. Krotkov, M. Perschbacher, J. Pippine, and C. Sullivan, "The darpa lagr program: Goals, challenges, methodology, and phase i results," *Journal of Field robotics*, vol. 23, no. 11-12, pp. 945–973, 2006.

[2] J. E. Naranjo, M. Clavijo, F. Jiménez, O. Gomez, J. L. Rivera, and M. Anguita, "Autonomous vehicle for surveillance missions in off-road environment," in *2016 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2016, pp. 98–103.

[3] S. R. Price, H. B. Land, S. S. Carley, S. R. Price, S. J. Price, and J. R. Fairley, "Expanding ground vehicle autonomy into unstructured, off-road environments: Dataset challenges." *Applied Sciences (2076-3417)*, vol. 14, no. 18, 2024.

[4] M. Wigness, S. Eum, J. G. Rogers, D. Han, and H. Kwon, "A rugd dataset for autonomous navigation and visual perception in unstructured outdoor environments," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 5000–5007.

[5] P. Jiang, P. Osteen, M. Wigness, and S. Saripalli, "Rellis-3d dataset: Data, benchmarks and analysis," in *2021 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2021, pp. 1110–1116.

[6] S. Triest, M. Sivaprakasam, S. J. Wang, W. Wang, A. M. Johnson, and S. Scherer, "Tartandrive: A large-scale dataset for learning off-road dynamics models," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 2546–2552.

[7] C. Min, W. Jiang, D. Zhao, J. Xu, L. Xiao, Y. Nie, and B. Dai, "Orfd: A dataset and benchmark for off-road freespace detection," in *2022 international conference on robotics and automation (ICRA)*. IEEE, 2022, pp. 2532–2538.

[8] P. Mortimer, R. Hagmanns, M. Granero, T. Luettel, J. Petereit, and H.-J. Wuensche, "The goose dataset for perception in unstructured environments," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 14 838–14 844.

[9] J. Knights, K. Vidanapathirana, M. Ramezani, S. Sridharan, C. Fookes, and P. Moghadam, "Wild-places: A large-scale dataset for lidar place recognition in unstructured natural environments," in *2023 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2023, pp. 11 322–11 328.

[10] S. Sharma, L. Dabbiru, T. Hannis, G. Mason, D. W. Carruth, M. Doude, C. Goodin, C. Hudson, S. Ozier, J. E. Ball *et al.*, "Cat: Cavs traversability dataset for off-road autonomous driving," *IEEE Access*, vol. 10, pp. 24 759–24 768, 2022.

[11] Y. Liu, Y. Fu, M. Qin, Y. Xu, B. Xu, F. Chen, B. Goossens, P. Z. Sun, H. Yu, C. Liu *et al.*, "Botanicgarden: A high-quality dataset for robot navigation in unstructured natural environments," *IEEE Robotics and Automation Letters*, 2024.

[12] A. Rana, A. Petitti, A. Ugenti, R. Galati, G. Reina, and A. Milella, "Towards digital twin of off-road vehicles using robot simulation frameworks," *IEEE Access*, 2024.

[13] Y. Yu, J. Xu, and L. Liu, "Adaptive diffusion terrain generator for autonomous uneven terrain navigation," *arXiv preprint arXiv:2410.10766*, 2024.

[14] P. Young, S. Kysar, and J. P. Bos, "Unreal as a simulation environment for off-road autonomy," in *Autonomous Systems: Sensors, Processing, and Security for Vehicles and Infrastructure 2020*, vol. 11415. SPIE, 2020, pp.

113–120.

[15] T. Xu, C. Pan, and X. Xiao, "Reinforcement learning for wheeled mobility on vertically challenging terrain," in *2024 IEEE International Symposium on Safety Security Rescue Robotics (SSRR)*. IEEE, 2024, pp. 125–130.

[16] X. Cai, J. Queeney, T. Xu, A. Datar, C. Pan, M. Miller, A. Flather, P. R. Osteen, N. Roy, X. Xiao *et al.*, "Pietra: Physics-informed evidential learning for traversing out-of-distribution terrain," *IEEE Robotics and Automation Letters*, 2025.

[17] J. So, A. Xie, S. Jung, J. Edlund, R. Thakker, A. Aghamohammadi, P. Abbeel, and S. James, "Sim-to-real via sim-to-seg: End-to-end off-road autonomous driving without real data," *arXiv preprint arXiv:2210.14721*, 2022.

[18] L. Xu, K. Chai, Z. Han, H. Liu, C. Xu, Y. Cao, and F. Gao, "An efficient trajectory planner for car-like robots on uneven terrain," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 2853–2860.

[19] A. Datar, C. Pan, M. Nazeri, and X. Xiao, "Toward wheeled mobility on vertically challenging terrain: Platforms, datasets, and algorithms," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 16 322–16 329.

[20] X. Xiao and R. Murphy, "A review on snake robot testbeds in granular and restricted maneuverability spaces," *Robotics and Autonomous Systems*, vol. 110, pp. 160–172, 2018.

[21] X. Xiao, E. Cappo, W. Zhen, J. Dai, K. Sun, C. Gong, M. J. Travers, and H. Choset, "Locomotive reduction for snake robots," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2015, pp. 3735–3740.

[22] B. Goldfain, P. Drews, C. You, M. Barulic, O. Velev, P. Tsiotras, and J. M. Rehg, "Autorally: An open platform for aggressive autonomous driving," *IEEE Control Systems Magazine*, vol. 39, no. 1, pp. 26–55, 2019.

[23] X. Xiao, J. Biswas, and P. Stone, "Learning inverse kinodynamics for accurate high-speed off-road navigation on unstructured terrain," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 6054–6060, 2021.

[24] P. Atreya, H. Karnan, K. S. Sikand, X. Xiao, S. Rabiee, and J. Biswas, "High-speed accurate robot control using learned forward kinodynamics and non-linear least squares optimization," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 11 789–11 795.

[25] H. Karnan, K. S. Sikand, P. Atreya, S. Rabiee, X. Xiao, G. Warnell, P. Stone, and J. Biswas, "Vi-ikd: High-speed accurate off-road navigation using learned visual-inertial inverse kinodynamics," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 3294–3301.

[26] Y. Pan, C.-A. Cheng, K. Saigol, K. Lee, X. Yan, E. A. Theodorou, and B. Boots, "Imitation learning for agile autonomous driving," *The International Journal of Robotics Research*, vol. 39, no. 2-3, pp. 286–302, 2020.

[27] T. Han, A. Liu, A. Li, A. Spitzer, G. Shi, and B. Boots, "Model predictive control for aggressive driving over uneven terrain," *arXiv preprint arXiv:2311.12284*, 2023.

[28] A. Tasora, R. Serban, H. Mazhar, A. Pazouki, D. Melanz, J. Fleischmann, M. Taylor, H. Sugiyama, and D. Negrut, "Chrono: An open source multi-physics dynamics engine," in *High Performance Computing in Science and Engineering: Second International Conference, HPCSE 2015, Soláň, Czech Republic, May 25-28, 2015, Revised Selected Papers 2*. Springer, 2016, pp. 19–49.

[29] S. Kolouri, P. E. Pope, C. E. Martin, and G. K. Rohde, "Sliced-wasserstein autoencoder: An embarrassingly simple generative model," *arXiv preprint arXiv:1804.01947*, 2018.

[30] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza, "On-manifold preintegration for real-time visual–inertial odometry," *IEEE Transactions on Robotics*, vol. 33, no. 1, pp. 1–21, 2016.

[31] C. Stachniss, J. J. Leonard, and S. Thrun, "Simultaneous localization and mapping," *Springer Handbook of Robotics*, pp. 1153–1176, 2016.

[32] T. Miki, L. Wellhausen, R. Grandia, F. Jenelten, T. Homberger, and M. Hutter, "Elevation mapping for locomotion and navigation using gpu," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 2273–2280.

[33] P. Ewen, A. Li, Y. Chen, S. Hong, and R. Vasudevan, "These maps are made for walking: Real-time terrain property estimation for mobile robots," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 7083–7090, 2022.

[34] N. Dashora, D. Shin, D. Shah, H. Leopold, D. Fan, A. Agha-Mohammadi, N. Rhinehart, and S. Levine, "Hybrid imitative planning with geometric and predictive costs in off-road environments," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 4452–4458.

[35] K. S. Sikand, S. Rabiee, A. Uccello, X. Xiao, G. Warnell, and J. Biswas, "Visual representation learning for preference-aware path planning," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 11 303–11 309.

[36] A. Pokhrel, A. Datar, M. Nazeri, and X. Xiao, "CAHSOR: Competence-aware high-speed off-road ground navigation in SE (3)," *IEEE Robotics and Automation Letters*, vol. 9, no. 11, pp. 9653–9660, 2024.

[37] J. Bae, T. Kim, W. Lee, and I. Shim, "Curriculum learning for vehicle lateral stability estimations," *IEEE Access*, vol. 9, pp. 89 249–89 262, 2021.

[38] S. Siva, M. Wigness, J. Rogers, and H. Zhang, "Robot adaptation to unstructured terrains by joint representation and apprenticeship learning," in *Robotics: Science and Systems (RSS)*, 2019.

[39] L. Sharma, M. Everett, D. Lee, X. Cai, P. Osteen, and

J. P. How, "Ramp: A risk-aware mapping and planning pipeline for fast off-road ground robot navigation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 5730–5736.

[40] S. Siva, M. Wigness, J. G. Rogers, L. Quang, and H. Zhang, "Nauts: Negotiation for adaptation to unstructured terrain surfaces," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 1733–1740.

[41] M. G. Castro, S. Triest, W. Wang, J. M. Gregory, F. Sanchez, J. G. Rogers, and S. Scherer, "How does it feel? self-supervised costmap learning for off-road vehicle traversability," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 931–938.

[42] X. Cai, S. Ancha, L. Sharma, P. R. Osteen, B. Bucher, S. Phillips, J. Wang, M. Everett, N. Roy, and J. P. How, "Evora: Deep evidential traversability learning for risk-aware off-road autonomy," *IEEE Transactions on Robotics*, 2024.

[43] J. Seo, S. Sim, and I. Shim, "Learning off-road terrain traversability with self-supervisions only," *IEEE Robotics and Automation Letters*, vol. 8, no. 8, pp. 4617–4624, 2023.

[44] D. D. Fan, K. Otsu, Y. Kubo, A. Dixit, J. Burdick, and A.-A. Agha-Mohammadi, "Step: Stochastic traversability evaluation and planning for risk-aware off-road navigation," in *Robotics: Science and Systems (RSS)*, 2021.

[45] H. Lee, T. Kim, J. Mun, and W. Lee, "Learning terrain-aware kinodynamic model for autonomous off-road rally driving with model predictive path integral control," *IEEE Robotics and Automation Letters*, 2023.

[46] X. Xiao, B. Liu, G. Warnell, and P. Stone, "Motion planning and control for mobile robot navigation using machine learning: a survey," *Autonomous Robots*, vol. 46, no. 5, pp. 569–597, 2022.

[47] Isaac Gym - NVIDIA Developer Forums, "Using isaacgym to simulate differential drive of a four-wheel robot - robotics - isaac," https://forums.developer.nvidia.com/t/using-isaacgym-to-simulate-differential-drive-of-a-four-wheel-robot/235455, 2025.

[48] T. Miki, L. Wellhausen, R. Grandia, F. Jenelten, T. Homberger, and M. Hutter, "Elevation mapping for locomotion and navigation using gpu," 2022.

[49] S. Laughery, G. R. Gerhart, and R. Goetz, *Bekker's terramechanics model for off-road vehicle research*. US Army TARDEC Hammond, WI, USA, 1990.

[50] A. Elmquist, A. Young, T. Hansen, S. Ashokkumar, S. Caldararu, A. Dashora, I. Mahajan, H. Zhang, L. Fang, H. Shen *et al.*, "Art/atk: A research platform for assessing and mitigating the sim-to-real gap in robotics and autonomous vehicle engineering," *arXiv preprint arXiv:2211.04886*, 2022.

[51] G. Williams, A. Aldrich, and E. A. Theodorou, "Model predictive path integral control: From theory to parallel computation," *Journal of Guidance, Control, and Dynamics*, vol. 40, no. 2, pp. 344–357, 2017.

[52] T. Xu, C. Pan, and X. Xiao, "Verti-selector: Automatic curriculum learning for wheeled mobility on vertically challenging terrain," *arXiv preprint arXiv:2409.17469*, 2024.

[53] A. Datar, C. Pan, and X. Xiao, "Learning to model and plan for wheeled mobility on vertically challenging terrain," *IEEE Robotics and Automation Letters*, vol. 10, no. 2, pp. 1505–1512, 2025.

[54] A. Datar, C. Pan, M. Nazeri, A. Pokhrel, and X. Xiao, "Terrain-attentive learning for efficient 6-dof kinodynamic modeling on vertically challenging terrain," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 5438–5443.

[55] C. Pan, A. Datar, A. Pokhrel, M. Choulas, M. Nazeri, and X. Xiao, "Traverse the non-traversable: Estimating traversability for wheeled mobility on vertically challenging terrain," *arXiv preprint arXiv:2409.17479*, 2024.

[56] S. Choudhury, S. Arora, and S. Scherer, "The planner ensemble: Motion planning by executing diverse algorithms," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2015, pp. 2389–2395.