# VERTIFORMER: A Data-Efficient Multi-Task Transformer on Vertically Challenging Terrain

Mohammad Nazeri*, Anuj Pokhrel*, Alexandyr Card*, Aniket Datar*, Garrett Warnell†‡ and Xuesu Xiao*

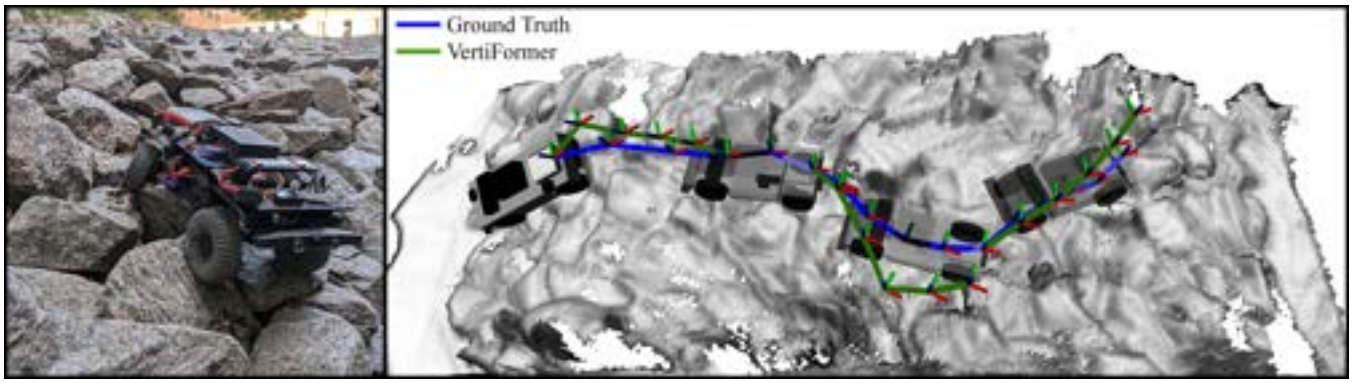



Fig. 1: VERTIFORMER is a data-efficient multi-task Transformer for off-road mobility on vertically challenging terrain. VERTIFORMER employs unified multi-modal latent representation, missing modality infilling, and non-autoregressive training to learn complex and nuanced vehicle-terrain interactions in $\mathbb{SE}(3)$ with only one hour of training data.

***Abstract*— We propose VERTIFORMER, a novel data-efficient multi-task Transformer trained with only one hour of multi-modal data to address the challenges of efficiently training Transformers for robot mobility on extremely rugged, vertically challenging, off-road terrain. With a Transformer encoder and decoder to predict the next robot pose, action, and terrain patch, VERTIFORMER employs a unified state representation and missing modality infilling to respectively enhance dynamics understanding and enable a variety of off-road mobility tasks simultaneously, e.g., forward and inverse kinodynamics modeling, and learning a navigation policy. By leveraging this unified representation alongside modality infilling, VERTIFORMER achieves real-time task switching during inference for improved fault tolerance and better generalization to unseen environments. Furthermore, VERTIFORMER's non-autoregressive design also mitigates computational bottlenecks and error propagation associated with autoregressive models. Our experiments offer insights into effectively utilizing Transformers for off-road robot mobility with limited data and demonstrate that VERTIFORMER can facilitate multiple off-road mobility tasks onboard a physical mobile robot with $\approx 20\%$ improvement in both success rate and traversal time compared to the Terrain-Attentive Learning (TAL ) baseline.** [1]

## I. INTRODUCTION

Autonomous mobile robots deployed in off-road environments face significant challenges posed by the underlying terrain. For example, irregular terrain topographies featuring vertical protrusions from the ground, i.e., vertically challenging terrain, pose extensive mobility risks [1], [2], [3], manifesting in several critical ways: compromised chassis stability, leading to potential rollover; increased wheel slippage, resulting in reduced traction and impaired locomotion; and unpredictable vehicle immobilization, causing the robot to get stuck, when interacting with vertically challenging terrain.

Precisely understanding the vehicle-terrain kinodynamic interactions is the key to mitigating such mobility challenges posed by off-road, vertically challenging terrain. Although data-driven approaches have shown promises in enabling off-road mobility in relatively flat environments [4], [5], [6], [7], [8], [1], [9], [10], [11], [12], [13], the intricate relationships between the robot chassis and vertically challenging terrain, e.g., suspension travel, tire deformation, changing normal and friction forces, and vehicle weight distribution and momentum, motivate the adoption of more sophisticated learning architectures to fully capture and represent the nuanced off-road kinodynamics [3].

Transformers are the preferred architectures to understand complex relationships, which show promise in Natural Language Processing (NLP) [14] and Computer Vision (CV) [15] with self-supervised pre-training emerging as a dominant methodology. This trend is now extending to robotics, impacting areas such as manipulation [16] and autonomous driving [17]. In addition to the advent of the well-studied Transformer architecture [18], [19], this progress is largely attributable to the availability of large-scale datasets [20], [21] as well as various Transformer training techniques including two primary pre-training paradigms: (i) Masked Modeling (MM) and (ii) Next-Token Prediction (NTP) [22].

The application of these paradigms to robotics is limited due to the inherent challenge associated with acquiring large-scale robotics datasets, especially for outdoor, off-road environments. The multi-modal nature of robotics data also

[1] https://github.com/mhnazeri/VertiFormer.

presents another significant challenge for Transformers to learn inter-modal relationships and understand the temporal progression of both the environment and the robot state at the same time. These two challenges of applying Transformers to robotics lead to our question: "*How can we train Transformers with limited multi-modal robotics data?*"

Motivated by this research question, this work presents VERTIFORMER, a novel data-efficient multi-task Transformer for robot mobility on extremely rugged, vertically challenging, off-road terrain that requires precisely understanding the kinodynamics in $\mathbb{SE}(3)$ to avoid getting stuck or rolling over. VERTIFORMER's novel unified latent representation of robot exteroception, proprioception, and action offers a stronger inductive bias, in the assumption of a shared physical manifold, and therefore off-loads the learning of inter-modality relationships from the Transformer. This consequently facilitates more effective learning with only one hour of data, contrasting current data-intensive methods in NLP, CV, and previous work in robotics [23], [24]. These works employ separate tokenization of modalities and depend solely on self-attention to capture complex inter-modal correlations within terabytes of data. Furthermore, VERTIFORMER's missing modality infilling enables various off-road mobility tasks within one model simultaneously without the need to retrain separate downstream tasks and mitigates the impact of missing modalities at inference time. Additionally, the non-autoregressive nature of VERTIFORMER avoids error propagation from earlier to later prediction steps and makes VERTIFORMER faster at inference because it does not require iterative queries for each step.

VERTIFORMER outperforms the navigation performance achieved by state-of-the-art kinodynamic modeling approaches specifically designed for vertically challenging terrain [25], [26] as well as general navigation models such as NoMaD [27], providing empirical evidence supporting the feasibility of training a Transformer on limited robotic datasets using effective training strategies. Our contributions can be summarized as follows:

- VERTIFORMER, a data-efficient, multi-task Transformer for off-road robot mobility on vertically challenging terrain in $\mathbb{SE}(3)$;
- a unified representation approach for modality-agnosticism to treat all modalities as a single distribution to off-load inter-modality relationship learning from the otherwise data-intensive Transformer;
- a missing modality infilling method that facilitates information sharing among multiple heads and therefore enables different off-road mobility tasks, i.e., forward/inverse kinodynamic learning (FKD/IKD) and emergent navigation policy (NP) via multi-task masking;
- an extensive evaluation of different Transformer designs, including MM, NTP, Encoder-only, and Decoder-only, for off-road kinodynamic representation learning; and
- physical on-robot experiments for different off-road mobility tasks on vertically challenging terrain and comparison against state-of-the-art methods.

## II. RELATED WORK

Transformers, initially proposed for language translation tasks, have demonstrated remarkable versatility across a spectrum of domains, including CV and robotics. This section provides an overview of existing work on Transformers in robotics and data-driven off-road mobility.

**Transformers in Robotics.** Recent years have witnessed a surge in the application of Transformers to robotics, encompassing both perception and planning: Generalist robot policies based on Transformers, e.g., Octo [28] and CrossFormer [23], with multi-modal sensory input [29] and action tokenization [30] aimed at handling diverse tasks such as manipulation and navigation; Studies in target-driven [31], [32], [33], [34] and image-goal navigation [35] have shown that Transformers significantly outperform traditional behavior cloning baselines [36], [37]; Reinforcement learning has been significantly enhanced by integrating the Transformer architecture, providing improved sequence modeling [38] and decision-making capabilities [39]; Transformers have also been used in motion planning to guide long-horizon navigation tasks [40] and reduce the search space for sampling-based motion planners [41].

A common characteristic of these models is their treatment of each sensor modality (e.g., vision, touch, and audio) as a distinct token, relying on the Transformer to learn the inter-modal correlations and their temporal dynamics. While this approach allows for flexible integration of diverse sensory information, it necessitates substantial amounts of training data to compensate. This data dependency poses a significant challenge, particularly in off-road robot mobility, where real-world, outdoor data acquisition can be expensive and time-consuming. Consequently, there remains a critical need for research focused on refining training methodologies and exploring architectural modifications specifically tailored to address the data scarcity and multi-modality often encountered in robotics.

**Learning Off-Road Mobility.** While most learning approaches for off-road autonomy focus on perception tasks [8], [42], [43], researchers have recently investigated off-road mobility to account for vehicle stability [2], [44], [12], wheel slippage [45], [46], [10], and terrain traversability [7], [9], [11], [47], [13]. A relevant work by [48] used Transformers to enable a universal forward kinodynamic model that can drive different ground vehicles. Most of these approaches have adopted specific techniques designed to address one particular off-road mobility task.

Focusing on multi-task kinodynamic representation for off-road mobility on vertically challenging terrain, our novel non-autoregressive VERTIFORMER employs unified modality latent representation and missing modality infilling to predict the next robot pose, action, and terrain patch in order to simultaneously enable a variety of off-road mobility tasks, i.e., FKD, IKD, NP, and terrain patch reconstruction, without a specific training procedure for each.

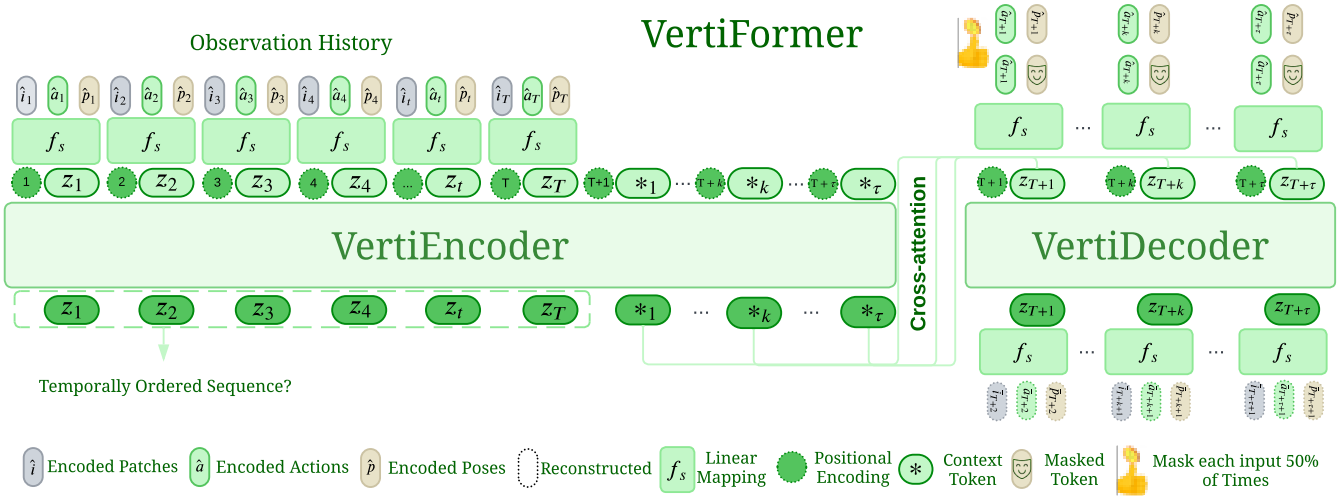



Fig. 2: **VERTIFORMER Architecture**. VERTIFORMER employs a TransformerEncoder (left) to receive a history of terrain patches, actions, and poses along with multiple context tokens. To predict future states, the model computes cross-attention between these context tokens and the upcoming actions or poses. VERTIDECODER uses causal masking to ensure that predictions are conditioned only on past and present information, preventing information leakage from future time steps.

## III. VERTIFORMER

We introduce VERTIFORMER, a data-efficient multi-task Transformer for kinodynamic representation and navigation on complex, vertically challenging, off-road terrain. We propose an efficient training methodology for training VERTIFORMER utilizing limited (one hour) robotics data, including unified multi-modal latent representation, missing modality infilling, and non-autoregressive training to improve data efficiency and enable multi-task learning.

### A. VERTIFORMER *Training*

VERTIFORMER consists of both TransformerEncoder (VERTIENCODER) and TransformerDecoder (VERTIDECODER), as illustrated in Fig. 2 left and right, respectively. Consistent with established practices [25], [26], VERTIFORMER receives a multi-modal sequence of actions $\mathbf{a}_{0:\mathrm{T}}$, robot poses $\mathbf{p}_{0:\mathrm{T}}$, and the underlying terrain patches $\mathbf{i}_{0:\mathrm{T}}$. Actions represent control input commands $(v, \omega) \in \mathbb{R}^2$, poses represent the full state of the robot $(x, y, z, \phi, \theta, \psi) \in \mathbb{SE}(3)$, and terrain patches represent a 40 x 40 area under the robot from a 2.5D elevation map. The VERTIENCODER first applies an independent linear mapping to each modality. Specifically, action commands $\mathbf{a}_{0:\mathrm{T}}$ are projected into an embedding space via a linear function $f_a$, yielding $\hat{\mathbf{a}}_{0:\mathrm{T}}$. Analogously, robot poses $\mathbf{p}_{0:\mathrm{T}}$ and terrain patches $\mathbf{i}_{0:\mathrm{T}}$ are transformed using linear mappings $f_p$ and $f_i$ respectively, producing a sequence of embeddings $\hat{\mathbf{p}}_{0:\mathrm{T}}$ and $\hat{\mathbf{i}}_{0:\mathrm{T}}$. This initial linear mapping can be formally expressed as:

$$\hat{a}_\mathrm{t} = f_a(a_\mathrm{t}) = W_a a_\mathrm{t} + b_a, a_\mathrm{t} \in \mathbf{a}_{0:\mathrm{T}}, \quad (1)$$

$$\hat{p}_\mathrm{t} = f_p(p_\mathrm{t}) = W_p p_\mathrm{t} + b_p, p_\mathrm{t} \in \mathbf{p}_{0:\mathrm{T}}, \quad (2)$$

$$\hat{i}_\mathrm{t} = f_i(i_\mathrm{t}) = W_i i_\mathrm{t} + b_i, i_\mathrm{t} \in \mathbf{i}_{0:\mathrm{T}}, \quad (3)$$

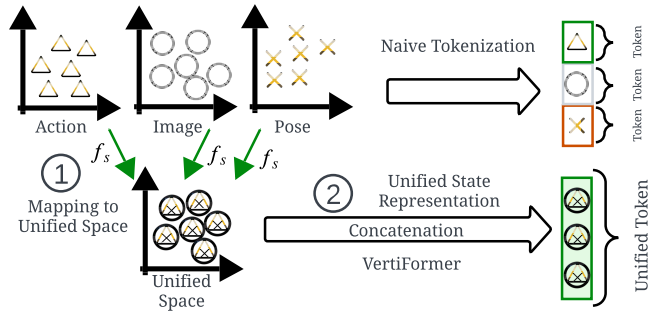
where $W_a$, $W_p$, and $W_i$ represent the weight matrices, and $b_a$, $b_p$, and $b_i$ denote the bias vectors for each respective modality.

*1) Unified Multi-Modal Latent Representation:* To offload cross-modal interaction learning from Transformer, it is crucial to establish a consistent distributional characteristic across the modality-specific embeddings. Instead of aligning different embeddings, VERTIFORMER treats all modalities as a single unified modality (modality-agnosticism). To achieve this, a subsequent linear transformation, denoted by $f_s$, is applied to the embeddings:

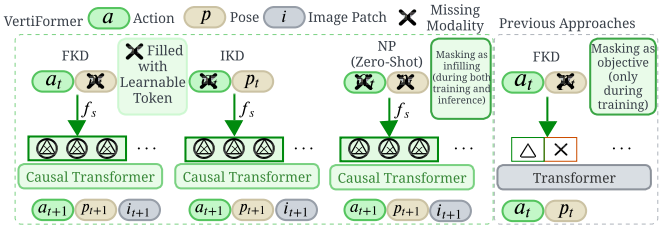
$$z_\mathrm{t} = f_s(\hat{a}_\mathrm{t}, \hat{p}_\mathrm{t}, \hat{i}_\mathrm{t}) = W_s(\hat{a}_\mathrm{t} \cdot \hat{p}_\mathrm{t} \cdot \hat{i}_\mathrm{t}) + b_s, t \in [0:T], \quad (4)$$

with $W_s$ and $b_s$ denoting the weight matrix and bias vector for $f_s$, respectively. This shared linear mapping $f_s$ aims to project all embeddings into a unified latent space, minimizing potential discrepancies in statistical properties. The resulting unified tokens, $\mathbf{z}_{0:\mathrm{T}}$, are then passed as input to the VERTIENCODER (Fig.2 top left). This procedure ensures a homogeneous input representation for the subsequent encoding layers, crucial for effective multi-modal fusion of robotic data (Fig. 3a). This new unified representation stems from the intuition that these input modalities represent the same scene and should therefore share a common representation space. To reinforce this, we also apply tied encoder-decoder weights, which further guide the modalities toward a shared distribution. This new modality unification approach results in a more coherent multi-modal representation, leading to improved kinodynamics understanding, particularly in *data-constrained* scenarios. Empirical results (Fig. 4a) supporting the importance of such unified representation, in contrast to the conventional individual modality representations, will be presented in Section IV.

*2) Missing Modality Infilling for Multi-Task Learning:* We propose stochastic modality infilling (Fig. 2, top right) to enable VERTIFORMER's multi-task prediction (i.e., pose, action, navigation, and terrain, Fig. 2, bottom right), aiming for enhanced data efficiency via shared representations. It




(a) Unified State Representation




(b) Missing Modality Infilling

Fig. 3: The integration of unified state representation and missing modality infilling enables VERTIFORMER to perform simultaneous temporal inference of FKD, IKD, and emergent NP.

allows one model with the same weight set to act as three different models (FKD, IKD, and NP) simultaneously.

After a warm-up phase, training involves replacing future ($\tau$ steps ahead) poses ($\mathbf{p}_{T+1:T+\tau}$) or actions ($\mathbf{a}_{T+1:T+\tau}$) with learnable vectors (50% probability each). This facilitates two tasks: Action-Conditioned Pose Prediction (given actions, predict poses) and Pose-Conditioned Action Prediction (given poses, predict actions), analogous to FKD and IKD respectively.

This strategy promotes a joint action-pose representation as the learnable tokens, processed by $f_s$ and thus aligned with the modality distributions. Consequently, the model supports dynamic task adaptation at inference and infers missing modalities through time (Fig. 3b).

Furthermore, by extending this infilling strategy to replace both future actions, $\mathbf{a}_{T+1:T+\tau}$, and future poses, $\mathbf{p}_{T+1:T+\tau}$, simultaneously, VERTIFORMER becomes a navigation policy, without being explicitly trained for. In this configuration, the model predicts both actions and poses solely based on the historical context, effectively mimicking the demonstrated behavior without requiring explicit information about future actions and poses from a planner. Notice that compared to masked modeling approaches [26] these learnable vectors are not masked as a learning objective, i.e., masked token reconstruction, instead they act as the modality representation and are present also during inference.

*3) Non-Autoregressive Training:* Building upon the work by [28] and [23], VERTIFORMER employs multiple context tokens to represent a distribution of plausible future states. These context tokens serve to inform VERTIDECODER in predicting both the future ego state and the evolution of the environment. Having multiple context tokens allows VERTIFORMER to predict the future non-autoregressively. The non-autoregressive nature of VERTIFORMER is motivated by the potential computational bottlenecks inherent in autoregressive models, which require querying the model multiple times and are subject to drifting due to error propagation from earlier steps. By learning multi-context representations, the non-autoregressive approach aims to improve both training efficiency and inference speed—a critical consideration for real-time robotic control applications.

We train VERTIFORMER by minimizing the Mean Squared Error between the model's predictions and the corresponding ground truth values. We evaluate the model by calculating the error rate between the model's predictions and the ground truth values on a held-out, unseen dataset.

### B. VERTIFORMER *Inference*

During FKD inference, VERTIENCODER receives the same historical input as training. VERTIDECODER receives sampled actions from an external sampling-based planner (e.g., MPPI [49]) while masking the corresponding poses, compelling the model to predict future poses based solely on the sampled actions (and the context tokens) so that the planner can choose the optimal trajectory to minimize a cost function. For IKD, a global planner generates desired future poses, and by masking the actions we encourage the model to predict future actions to achieve these globally planned poses. By masking both actions and poses, VERTIFORMER can still navigate by predicting actions as an emergent task.

## IV. TRAINING ON ONE HOUR OF DATA

VERTIFORMER is trained on one hour of data, as a low-data regime stress test to show the efficiency of VERTIFORMER's architectural design choices. The data come from a human-teleoperated demonstration of driving an open-source four-wheeled ground vehicle [3], Verti-4-Wheeler (V4W), on a custom-built off-road testbed composed of hundreds of rocks and boulders. The demonstrator mostly aims to drive the robot to safely and stably traverse the vertically challenging terrain, but still occasionally encounters dangerous situations such as large roll angles and getting stuck between rocks. VERTIFORMER leverages these situations to explore and understand a wider range of kinodynamic interactions. In contrast, TAL (CNN-based) lacks this adaptability; when trained on the same data, it cannot generalize to unseen terrain dynamics as VERTIFORMER does (see Sec. V). A primary challenge here is that standard Transformer methodologies from NLP and CV do not translate easily to small robotics datasets, as the lack of inductive bias in Transformers [19] usually demands much larger data scales. However, our experiments suggest that VERTIFORMER's judicious modifications to established MM and NTP training paradigms can facilitate effective Transformer training even with limited robotics data.

**Unified latent space representation** facilitates FKD, IKD, and NP by decoupling inter-modality learning from

temporal progression modeling, with only the latter handled by the Transformer, which otherwise becomes data-intensive.

To evaluate VERTIFORMER's dynamics understanding, we use a sequence order prediction task following BERT [14], in which the model classifies sequences as chronologically ordered (50%) or randomly shuffled (50%). This probes the model's grasp of temporal dependencies and kinodynamics evolution.

As illustrated (Fig. 4a), non-unified tokens result in poor kinodynamic understanding (minimal loss decrease) and suggest that fragmented processing hinders capturing temporal relationships, i.e., the model cannot understand that a sequence $t = \{1, 2, 3, 5, 4\}$ is not physically feasible. While larger datasets might compensate, they are often unavailable in robotics and contradict the goal of this work.

Conversely, the unified representation significantly improves the model's ability to discern temporal order and understand system dynamics by consolidating information cohesively. This underscores the importance of unified representations for learning complex dynamics effectively from limited robotics data, unlike in data-rich NLP/CV domains.

**Longer prediction horizons** in navigation planning improve foresight but increase uncertainty via error accumulation, especially in autoregressive models like VERTIDECODER where errors propagate. We compare the autoregressive VERTIDECODER with the non-autoregressive VERTIFORMER on long-horizon accuracy. Results (Fig. 4b) show VERTIFORMER predicts longer (2s) with less drift than VERTIDECODER predicting shorter (1s), highlighting the advantage of non-autoregressive models for reducing compounding errors in long-term predictions.

**MM vs NTP vs End2End** are currently the prominent approaches in CV, NLP, and robotics respectively. We compare MM, NTP, and End2End for off-road mobility tasks. We analyze a MM encoder (VERTICODER [26]), an autoregressive NTP decoder (VERTIDECODER, Fig. 2 right trained alone without cross-attention), a non-Transformer End2End model [26], and VERTIFORMER, our non-autoregressive Transformer (Fig. 2). VERTICODER and VERTIDECODER use the unified representation (Fig. 4a). The End2End model employs ResNet-18 for computational balance.

Evaluations (Fig. 4c, 1s horizon) show VERTIFORMER achieves superior performance on FKD, IKD, and NP error rates. Its non-autoregressive prediction leads to better accuracy than the autoregressive VERTIDECODER (which cannot directly perform NP, as it has access to both action and pose at each step). VERTIFORMER's joint multi-task training also surpasses VERTICODER's separate training [26] (except Z prediction). The End2End model exhibits the highest errors, highlighting Transformer's benefits for kinodynamics learning.

Beyond accuracy, VERTIFORMER supports concurrent multi-task execution during inference, vital for real-time robotics, especially with missing modalities (e.g., sensor degradation and planner failure).

## V. ROBOT EXPERIMENTS

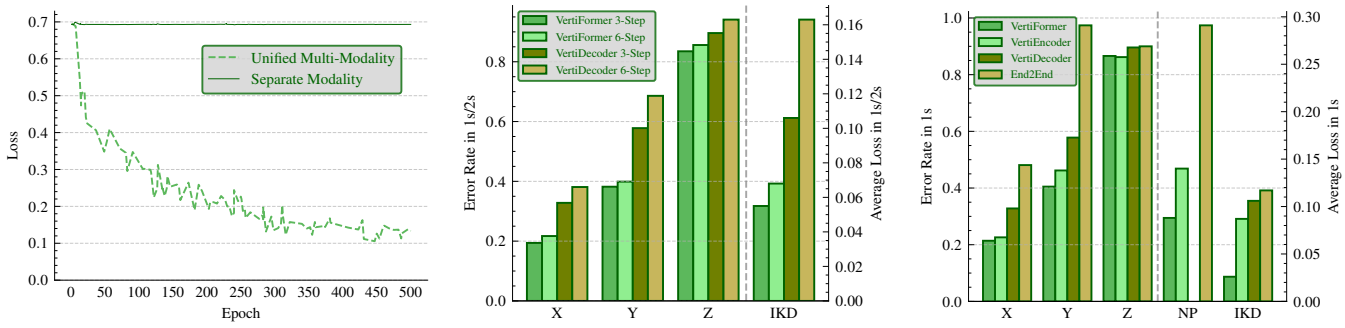
We implement VERTIFORMER's FKD, IKD, and NP on the V4W ground robot platform. The experiments are carried out on a *never*-seen-before, 4 m $\times$ 2.5 m testbed made of rocks/boulders, wooden planks, AstroTurf with crumpled cardboard boxes underneath creating loose ground, and modular 0.8 m $\times$ 0.75 m expanding foam to represent different types of vertically challenging terrain with different friction coefficients and varying deformability (Fig. 5). The modular foam and rocks/boulders do not deform, while the rocks may shift positions under the weight of the robot. On the other hand, the wooden planks and AstroTurf are completely deformable and change the terrain topography during wheel-terrain interactions. The one-hour training dataset used only consists of robot teleoperation on the rigid rock/boulder testbed and hence the experiment testbed is an unseen environment, posing generalization challenges for VERTIFORMER. To ensure interaction diversity, the rock arrangement on the testbed was significantly reconfigured before each trial.

**FKD:** VERTIFORMER's FKD task is integrated with the MPPI planner [49] with 1000 samples and a horizon of 18 steps. We sample across a range of control sequences centered around the last optimal control sequence selected by the robot. The first three actions in a sampled control sequence are passed to VERTIFORMER along with six past poses, actions, and terrain patches at 3 Hz consisting of one second. The model is repeated six times and outputs 18 future poses of the robot, which are combined to create one candidate trajectory. All 1000 candidate trajectories are then evaluated by a cost function, which calculates the cost of each trajectory based on the Euclidean distance to the goal and roll and pitch angles of the robot. Higher distance, roll, and pitch values are penalized with higher cost. Based on the cost function, MPPI outputs the best control sequence moving the robot forward at 3 Hz. The V4W executes the first action and replans.

**IKD:** We integrate VERTIFORMER's IKD task with a global planner based on Dijkstra's algorithm [50], which minimizes traversability cost on a traversability map [51]. The global planner generates three desired future poses with the lowest cost and passes them to VERTIFORMER, which also has access to six past poses, actions, and terrain patches. VERTIFORMER then produces three future actions to drive the robot to the three desired future poses. Similarly to FKD, the V4W executes the first action and then replans at 3 Hz.

**NP:** We implement VERTIFORMER's NP by passing in six past poses, actions, and terrain patches to VERTIFORMER. The model outputs three future actions to take. Similarly to FKD and IKD, the first action is executed by V4W and the replanning of NP runs at 3 Hz.

For FKD and IKD, a trial is deemed successful if the robot reaches the defined goal without rolling over or getting stuck. For NP without explicit goal information, a trial is considered successful if the robot successfully traverses the entire testbed.




(a) (b) (c)

Fig. 4: (a) Without unified latent representation, the model cannot capture temporal dependencies and understand kinodynamic transitions, resulting in an almost flat learning curve. (b) VERTIFORMER is capable of predicting a longer horizon without losing much accuracy due to its non-autoregressive nature. (c) VERTIFORMER achieves the best accuracy across FKD, IKD, and NP compared to VERTICODER (MM), VERTIDECODER (NTP), and End2End.

**Results and Discussions:** The results of the three methods are then compared to MPPI using TAL [25], a highly accurate forward kinodynamic model specifically designed for vertically challenging terrain, and NoMaD [27], a state-of-the-art general navigation model based on diffusion policy. We added Unified State Representation to VERTICODER (VERTICODER–*unified*) to show the applicability of this contribution to other approaches and how the improved kinodynamic understanding contributes to the downstream navigation task. We train NoMaD from scratch (NoMaD–*scratch* in Table I) to illustrate the difficulty of learning from our limited (one hour) data, while comparison with pre-trained NoMaD highlights the inadequacy of 2D assumptions for challenging terrain, necessitating an understanding of 3D robot-terrain interactions. Since NoMaD tackles a different problem than VERTIFORMER, fine-tuning its pre-trained weights with data from an unrelated task would negatively impact its performance. We report success rate (SR), traversal time (TT), and roll and pitch angles in Table I.

Our experiments reveal an obvious distinction between VERTICODER and VERTICODER–*unified* where the only difference is the mapping of modalities to the unified representation before passing to the Transformer (see Sec. III-A.1). However, there is a nuanced performance difference between VERTICODER–*unified* and VERTIFORMER, particularly concerning NP and IKD. VERTICODER–*unified* leverages MM pre-training to learn a general kinodynamic representation. Then it trains separate downstream task heads with the learned representation, providing VERTICODER–*unified* with privileged information for each task. This specialized training allows VERTICODER–*unified* to effectively leverage the provided data for NP. In contrast, VERTIFORMER's NP is an emergent behavior. It is not explicitly trained on NP, relying instead on its modality infilling strategy. This infilling effectively handles missing modalities by replacing them with a trained mask, enabling the model to infer behavior without direct NP training. While this approach allows VERTIFORMER to perform NP without specialized training, it also explains why VERTICODER–*unified*, with its dedicated head, achieves a higher success rate. A similar trend is observed with IKD. VERTIDECODER has access to both predicted and actual actions and poses at each time step, providing richer guidance for the IKD process. This richer information stream in VERTIDECODER is the reason for achieving a higher success rate, especially considering the inherent difficulty of IKD compared to FKD. VERTIFORMER, however, faces a challenge in IKD and takes longer to finish the traversal. The infilling strategy, while effective for missing modality, is not as accurate as the actual modality.

Regarding FKD, the architectural difference between VERTIFORMER and VERTICODER–*unified* causes different navigation behaviors. VERTICODER–*unified*'s specialized task head for FKD treats each future step independently without any attention weights between steps. While this approach facilitates faster MPPI initial convergence due to a lack of cross attention, it can also lead to drift, causing inconsistencies between predicted steps and ultimately resulting in a larger standard deviation of traversal time across trials. While VERTICODER–*unified*'s MPPI converges quickly, it struggles with long-term consistency. VERTIFORMER takes a different approach. By employing attention and cross-attention mechanisms between historical and future steps, it dynamically incorporates past information into future predictions. VERTIFORMER also decouples the perception module (encoder), with a strong physical understanding, from the prediction/policy module (decoder) and links them via cross-attention, significantly improving generalization. This allows VERTIFORMER to consider the historical context through cross-attention and causal masking when predicting future states, leading to more coherent and consistent predictions. Consequently, although MPPI might require more time to converge on a path with VERTIFORMER, once it does, the resulting behavior is more robust and less variable across trials, reflected in a smaller traversal time standard deviation.

## VI. CONCLUSIONS

In this work, we introduce VERTIFORMER, a novel data-efficient multi-task Transformer designed for learning kinodynamic representations on vertically challenging, off-road

Fig. 5: Unseen Test Environments with Rocks/Boulders, Wooden Planks, Astro-Turf, and Expanding Foam.

| Task | Model | SR ↑ | TT ↓ | Roll ↓ | Pitch ↓ |
|---|---|---|---|---|---|
| FKD | TAL [25] | 8/10 | $11.80 \pm 0.87$ | $0.198 \pm 0.38$ | $\mathbf{0.086} \pm 0.07$ |
| | VERTIDECODER | 6/10 | $15.12 \pm 1.78$ | $0.180 \pm 0.30$ | $0.114 \pm 0.09$ |
| | VERTICODER [26] | 6/10 | $14.32 \pm 1.23$ | $0.178 \pm 0.25$ | $0.112 \pm 0.09$ |
| | VERTICODER–*unified* | **10**/10 | $\mathbf{8.58} \pm 1.54$ | $0.189 \pm 0.23$ | $0.116 \pm 0.08$ |
| | VERTIFORMER | **10**/10 | $9.42 \pm 0.61$ | $\mathbf{0.169} \pm 0.17$ | $0.096 \pm 0.08$ |
| IKD | VERTIDECODER | **10**/10 | $15.92 \pm 1.08$ | $0.181 \pm 0.23$ | $0.125 \pm 0.08$ |
| | VERTICODER [26] | 4/10 | $17.35 \pm 2.68$ | $0.186 \pm 0.21$ | $0.097 \pm 0.08$ |
| | VERTICODER–*unified* | 7/10 | $\mathbf{13.99} \pm 3.27$ | $\mathbf{0.136} \pm 0.14$ | $\mathbf{0.069} \pm 0.07$ |
| | VERTIFORMER | 8/10 | $17.16 \pm 6.10$ | $\mathbf{0.136} \pm 0.10$ | $0.077 \pm 0.07$ |
| NP | NoMaD [27] | 1/10 | 22.3 | 0.187 | 0.09 |
| | NoMaD–*scratch* | 0/10 | - | - | - |
| | VERTICODER [26] | 5/10 | $18.21 \pm 3.68$ | $0.184 \pm 0.43$ | $0.090 \pm 0.09$ |
| | VERTICODER–*unified* | **9**/10 | $13.49 \pm 3.33$ | $0.175 \pm 0.37$ | $\mathbf{0.089} \pm 0.09$ |
| | VERTIFORMER | 8/10 | $\mathbf{12.64} \pm 3.89$ | $\mathbf{0.154} \pm 0.11$ | $0.099 \pm 0.08$ |

TABLE I: Physical results with VERTIFORMER, VERTICODER, VERTICODER with Unified Representation (VERTICODER–*unified*), VERTIDECODER, NoMaD, and TAL .

terrain. VERTIFORMER demonstrates the capacity to simultaneously address forward kinodynamics learning, inverse kinodynamics learning, and navigation policy learning tasks, only using one hour of training data. Key contributions include a unified latent space representation enhancing temporal understanding, learned modality infilling facilitating multiple off-road mobility tasks simultaneously and acting as a proxy for missing modalities during inference, and multi-context tokens enabling multi-step prediction without autoregressive feedback. All three contributions improve robustness and generalization of VERTIFORMER to out-of-distribution environments. We provide extensive experiment results and empirical guidelines for training Transformers under extreme data scarcity. Our evaluations across all three downstream tasks demonstrate that VERTIFORMER outperforms baseline models, including TAL [25], VERTICODER [26], VERTIDECODER, and NoMaD [27], while exhibiting reduced overfitting and improved generalization and highlighting the efficacy of the proposed architecture and training methodology for learning kinodynamic representations in data-constrained settings. Physical experiments also demonstrate that VERTIFORMER can enable superior off-road robot mobility on vertically challenging terrain. We leave extending this work to general navigation as future work.

It is crucial to acknowledge that our observations are primarily associated with the challenges inherent in wheeled locomotion on complex, vertically challenging, off-road terrain that requires an understanding of the robot-terrain interactions in 3D and may not be applicable to other robotic domains, such as manipulation, without further investigation.

## ACKNOWLEDGMENT

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 Laboratory (ARL, W911NF2220242, W911NF2320004, W911NF2420027, W911NF2520011), Air Force Research Laboratory (AFRL) and US Air Forces Central (AFCENT, GS00Q14OADU309), Google DeepMind (GDM), Clearpath Robotics, and Raytheon Technologies (RTX).

## REFERENCES

[1] P. Borges, T. Peynot, S. Liang, B. Arain, M. Wildie, M. Minareci, S. Lichman, G. Samvedi, I. Sa, N. Hudson, M. Milford, P. Moghadam, and P. Corke, "A survey on terrain traversability analysis for autonomous ground vehicles: Methods, sensors, and challenges," *Field Robotics*, vol. 2, pp. 1567–1627, 2022.

[2] 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.

[3] 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)*, May 2024, pp. 16 322–16 329.

[4] T. Overbye and S. Saripalli, "Fast local planning and mapping in unknown off-road terrain," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 5912–5918.

[5] 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*, 2020.

[6] 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.

[7] 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.

[8] 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, Jun. 2022.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[15] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, and et al., "DINOv2: Learning Robust Visual Features without Supervision," Apr. 2023.

[16] Y. Seo, D. Hafner, H. Liu, F. Liu, S. James, K. Lee, and P. Abbeel, "Masked World Models for Visual Control," May 2023.

[17] A. Bar, G. Zhou, D. Tran, T. Darrell, and Y. LeCun, "Navigation World Models," Dec. 2024.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," Jun. 2021.

[20] A. O'Neill, A. Rehman, A. Maddukuri, A. Gupta, A. Padalkar, A. Lee, A. Pooley, A. Gupta, A. Mandlekar, A. Jain *et al.*, "Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 6892–6903.

[21] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2446–2454.

[22] L. Chen, Z. Wang, S. Ren, L. Li, H. Zhao, Y. Li, Z. Cai, H. Guo, L. Zhang, Y. Xiong, Y. Zhang, R. Wu, Q. Dong, G. Zhang, J. Yang, L. Meng, S. Hu, Y. Chen, J. Lin, S. Bai, A. Vlachos, X. Tan, M. Zhang, W. Xiao, A. Yee, T. Liu, and B. Chang, "Next Token Prediction Towards Multimodal Intelligence: A Comprehensive Survey," Dec. 2024.

[23] R. Doshi, H. Walke, O. Mees, S. Dasari, and S. Levine, "Scaling Cross-Embodied Learning: One Policy for Manipulation, Navigation, Locomotion and Aviation," Aug. 2024.

[24] J. Jones, O. Mees, C. Sferrazza, K. Stachowicz, P. Abbeel, and S. Levine, "Beyond sight: Finetuning generalist robot policies with heterogeneous sensors via language grounding," *arXiv preprint arXiv:2501.04693*, 2025.

[25] 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)*. Abu Dhabi, United Arab Emirates: IEEE, Oct. 2024, pp. 5438–5443.

[26] M. Nazeri, A. Datar, A. Pokhrel, C. Pan, G. Warnell, and X. Xiao, "VertiCoder: Self-Supervised Kinodynamic Representation Learning on Vertically Challenging Terrain," Sep. 2024.

[27] A. Sridhar, D. Shah, C. Glossop, and S. Levine, "NoMaD: Goal Masked Diffusion Policies for Navigation and Exploration," Oct. 2023.

[28] Octo Model Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, T. Kreiman, C. Xu, J. Luo, Y. L. Tan, P. Sanketi, Q. Vuong, T. Xiao, D. Sadigh, C. Finn, and S. Levine, "Octo: An Open-Source Generalist Robot Policy," May 2024.

[29] J. Jones, O. Mees, C. Sferrazza, K. Stachowicz, P. Abbeel, and S. Levine, "Beyond Sight: Finetuning Generalist Robot Policies with Heterogeneous Sensors via Language Grounding," Jan. 2025.

[30] K. Pertsch, K. Stachowicz, B. Ichter, D. Driess, S. Nair, Q. Vuong, O. Mees, C. Finn, and S. Levine, "FAST: Efficient Action Tokenization for Vision-Language-Action Models," Jan. 2025.

[31] H. Du, X. Yu, and L. Zheng, "VTNet: Visual Transformer Network for Object Goal Navigation," May 2021.

[32] H. Wang, A. H. Tan, and G. Nejat, "NavFormer: A Transformer Architecture for Robot Target-Driven Navigation in Unknown and Dynamic Environments," 2024.

[33] M. Nazeri, J. Wang, A. Payandeh, and X. Xiao, "VANP: Learning Where to See for Navigation with Self-Supervised Vision-Action Pre-Training," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Abu Dhabi, United Arab Emirates: IEEE, Oct. 2024, pp. 2741–2746.

[34] W. Huang, Y. Zhou, X. He, and C. Lv, "Goal-Guided Transformer-Enabled Reinforcement Learning for Efficient Autonomous Navigation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 2, pp. 1832–1845, Feb. 2024.

[35] N. Pelluri, "Transformers for Image-Goal Navigation," May 2024.

[36] D. A. Pomerleau, "ALVINN: An autonomous land vehicle in a neural network," in *Advances in Neural Information Processing Systems*, D. Touretzky, Ed., vol. 1. Morgan-Kaufmann, 1988.

[37] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, X. Zhang, J. Zhao, and K. Zieba, "End to End Learning for Self-Driving Cars," Apr. 2016.

[38] X. Zhang, Z. Feng, Q. Qiu, Y. Chen, B. Hua, and J. Ji, "NaviFormer: A Data-Driven Robot Navigation Approach via Sequence Modeling and Path Planning with Safety Verification," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, May 2024, pp. 14 756–14 762.

[39] L. Chen, K. Lu, A. Rajeswaran, K. Lee, A. Grover, M. Laskin, P. Abbeel, A. Srinivas, and I. Mordatch, "Decision Transformer: Reinforcement Learning via Sequence Modeling," *arXiv:2106.01345 [cs]*, Jun. 2021.

[40] D. Lawson and A. H. Qureshi, "Control Transformer: Robot Navigation in Unknown Environments Through PRM-Guided Return-Conditioned Sequence Modeling," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct. 2023, pp. 9324–9331.

[41] J. J. Johnson, U. S. Kalra, A. Bhatia, L. Li, A. H. Qureshi, and M. C. Yip, "Motion Planning Transformers: A Motion Planning Framework for Mobile Robots," Nov. 2022.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] J. Seo, S. Sim, and I. Shim, "Learning Off-Road Terrain Traversability with Self-Supervisions Only," May 2023.

[48] W. Xiao, H. Xue, T. Tao, D. Kalaria, J. M. Dolan, and G. Shi, "AnyCar to Anywhere: Learning Universal Dynamics Model for Agile and Adaptive Mobility," Sep. 2024.

[49] G. Williams, A. Aldrich, and E. A. Theodorou, "Model predictive path integral control: From theory to parallel computation," *Journal of Guidance, Control, and Dynamics*, 2017.

[50] E. W. Dijkstra, "A note on two problems in connexion with graphs," *Numerische mathematik*, vol. 1, no. 1, pp. 269–271, 1959.

[51] 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," Sep. 2024.