Zero-Shot Adaptation to Robot Structural Damage via Natural Language-Informed Kinodynamics Modeling

Anuj Pokhrel[1], Aniket Datar[1], Mohammad Nazeri[1], Francesco Cancelliere[2], and Xuesu Xiao[1]

[1]George Mason University
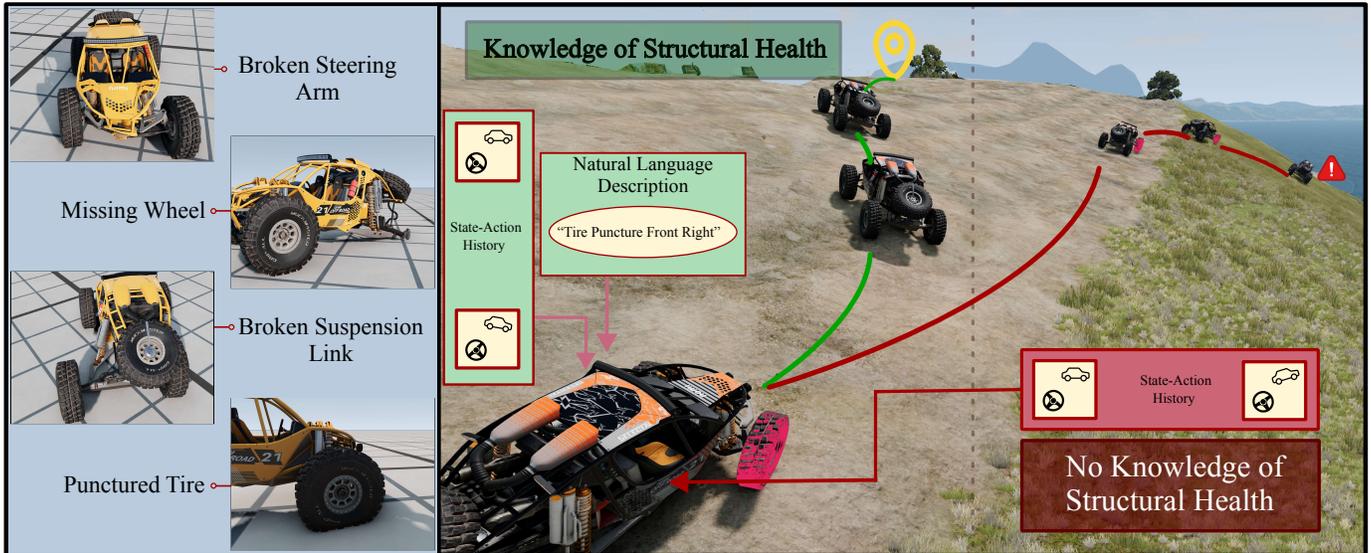[2] University of Catania
Code: ⦿ GitHub   Video: ▶ YouTube

Fig. 1: High-performance mobile robots may experience a variety of structural damages during in-the-wild operations (left), which significantly alter vehicle kinodynamics (right). When such kinodynamics discrepancies are not timely accounted for, catastrophic consequences may occur, e.g., falling off a cliff on the right due to a punctured front right tire (right).

Abstract—High-performance autonomous mobile robots endure significant mechanical stress during in-the-wild operations, e.g., driving at high speeds or over rugged terrain. Although these platforms are engineered to withstand such conditions, mechanical degradation is inevitable. Structural damage manifests as consistent and notable changes in kinodynamic behavior compared to a healthy vehicle. Given the heterogeneous nature of structural failures, quantifying various damages to inform kinodynamics is challenging. We posit that natural language can describe and thus capture this variety of damages. Therefore, we propose Zero-shot Language Informed Kinodynamics (ZLIK), which employs self-supervised learning to ground semantic information of damage descriptions in kinodynamic behaviors to learn a forward kinodynamics model in a data-driven manner. Using the high-fidelity soft-body physics simulator BeamNG.tech, we collect data from a variety of structurally compromised vehicles. Our learned model achieves zero-shot adaptation to different damages with up to 81% reduction in kinodynamics error and generalizes across the sim-to-real and full-to-1/10th scale gaps.

I. INTRODUCTION

High-performance autonomous mobile robots deployed in the wild must contend with impact forces, torsional stress, and vibration caused by rapid and complex vehicle-terrain interactions. While modern platforms are engineered with robust chassis and suspension systems to mitigate these forces, the harsh nature of real-world environments makes mechanical degradation inevitable. Consequently, robots frequently sustain structural damage ranging from punctured tires to bent steering columns and compromised suspensions (Fig. 1 right). Nevertheless, mission-critical deployments frequently demand that robots tolerate and adapt to such damage to ensure timely task completion.

Such structural damage, however, drastically alters the vehicle's kinodynamics by changing how forces are distributed and how the vehicle responds to control inputs. For example, a punctured tire isn't able to generate as much traction as compared to other wheels, resulting in significant drag and rotational bias even when commanded with straight-line velocities. If the navigation system operates using a clean kinodynamics model while the physical platform is damaged, the discrepancy leads to tracking errors, instability, and potentially catastrophic control failure like collision with obstacles, vehicle rollover, or falling off a cliff (as shown in Fig. 1 right).

Accurately compensating for such kinodynamics mismatch

poses a significant representation challenge. Structural degradation manifests heterogeneously: it ranges from scalar parametric deviations (e.g., tire pressure loss) to complex geometric deformations (e.g., bent or broken suspension linkages) that fundamentally alter the robot's physical constraints. Current state-of-the-art approaches to resolve the changing kinodynamics problem [55] rely on adaptive control strategies or fine-tuning a previously learned kinodynamics model. However, these methods depend on an n-shot adaptation phase, requiring data collection through further vehicle-terrain interactions after the damage has occurred. This delayed process necessitates driving a structurally compromised robot with an incorrect kinodynamics model to collect training data, which is inherently risky and can worsen mechanical failure during data collection.

To bridge this gap without requiring dangerous online data collection and retraining, we propose a novel framework that leverages natural language to instantly adapt to vehicle structural damage in a zero-shot manner, i.e., Zero-shot Language Informed Kinodynamics (ZLIK). We argue that, although the variety of structural damages cannot be represented by a single metric, they are semantically distinct and can be easily described using language. Therefore, we incorporate vehicle damage into kinodynamics modeling through language descriptions. We utilize a Sentence Transformer to map natural language descriptions into a semantic embedding space, where damages exhibiting similar kinodynamic behaviors are clustered closer. These embeddings condition a kinodynamics model with spatiotemporal attention to predict the robot's future states based on the current robot health. ZLIK allows the robot to instantly switch its internal kinodynamics model based on a diagnosis provided by an external vehicle health monitoring system, enabling zero-shot adaptation to different damages. We validate ZLIK using a high-fidelity soft-body physics simulator, BeamNG.tech [3] and deploy the system on a real-world robotic platform. Our contributions are as follows:

1) A novel self-supervised learning approach that aligns language embeddings with kinodynamic responses in the presence of various vehicle structural damages;
2) An accurate kinodynamics model that utilizes the aligned language embeddings to represent heterogeneous structural damages for instant model adaptation;
3) Rigorous experiments including detailed studies on damage embeddings, model architecture, semantic grounding, and transfer from a simulated full-size vehicle to a physical 1/10th scale robot with different damages.

II. RELATED WORK

We review related work in kinodynamics modeling, online adaptation, and robotic foundation models.

A. Kinodynamics Modeling

Kinodynamics modeling has historically relied on analytical formulations [38, 42], such as bicycle or Ackermann steering models. These models offer computational efficiency but often

oversimplify complex vehicle-terrain interactions [9]. Therefore, researchers have predominantly shifted towards data-driven solutions [54, 14]. To satisfy the high data requirements of these approaches without risking hardware, researchers frequently leverage high-fidelity simulators [48, 3, 28, 44, 47]. To improve predictions beyond the training distributions [37], recent works have incorporated uncertainty quantification using Gaussian Processes [36, 32, 22, 35] or enforced kinematic constraints via Physics-Informed Neural Networks [1, 27, 62, 5]. Regardless of the model architecture, accurate prediction relies on information provided by the input modalities.

Input modalities in existing works have primarily focused on visual semantics [36, 53], elevation mapping [9, 7, 33, 34], terrain geometry [29, 12], and inertial measurements [56, 19, 39] to perceive the external environment and inform kinodynamics. However, changes in robot's internal physical state such as structural damages also play a critical role in determining kinodynamics. Structural damages can take a variety of forms, which can or cannot be measured by commonly used sensor suites and manifest by both quantitative measurements and qualitative descriptions. Therefore, we posit natural language descriptions provide the necessary semantic signal to represent a robot's internal health conditions and can cover a variety of heterogeneous vehicle damages that also affect kinodynamics.

B. Online Adaptation

Traditional online system identification methods are designed to handle parametric variations in complex environments, such as estimating friction coefficients [40, 32] or slip ratios [59]. However, these approaches assume a fixed vehicle structure and treat deviations as scalar parameter changes within a known physics model. In contrast, structural damages or mechanical failures induce large kinodynamics shifts that cannot be captured by simple parameter variations [6].

To address non-linear dynamics shifts, researchers have relied on meta-learning and online adaptation methods [52]. Approaches such as Model-Agnostic Meta Learning [11], Meta-Reinforcement Learning [43], Function Encoders [17], and CaRoL [16] treat dynamics shifts as new tasks, updating the base model's weights through gradient-based updates using recent state-action pairs. These approaches require new physical interactions to infer new dynamics and introduce adaptation delays in an n-shot fashion [31, 26, 23]. Such continuous environment interactions using an incorrect kinodynamics model caused by delayed adaptation can lead to unsafe maneuvers on a mechanically compromised platform.

In contrast, ZLIK entirely abandons the paradigm of online retraining with new data. Instead, we frame the problem as conditional modeling. By training our model on a diverse manifold of damage descriptions, ZLIK learns a fixed set of weights that can generalize across different structural health states. During deployment, the dynamics model's weights do not need to be updated through dangerous data collection and delayed weight adaptation, rather it simply switches its predictive context based on the natural language descriptions of the damage. This zero-shot adaptability allows the robot

to respect its new physical limits the moment a damage description is provided.

C. Robotic Foundation Models

Using pre-trained models to accelerate learning is a well-established concept in robotics. Data-driven approaches have heavily relied on visual backbones pre-trained on large-scale datasets to extract robust environmental features [18]. Architectures such as ResNet [15] and SegFormer [57] have become standard encoders for tasks ranging from semantic segmentation [45] to traversability estimation [49].

The introduction of the Transformer architecture [50] has shifted the paradigm from simple feature extraction to holistic foundation models [21, 58]. In the context of dynamics modeling, this architecture is especially efficient in capturing temporal context and long-horizon dependencies [25, 33, 55]. However, standard robotic implementations typically tokenize the entire state tuple (x, y, z, roll, pitch, yaw) as a single input embedding [34]. This monolithic approach obscures the independence of state dimensions. To better capture the nuances of structural damages where specific degrees of freedom may be affected disproportionately, we decompose the state vector into dimension-specific tokens, allowing the attention mechanism to isolate compromised dynamic channels [13, 61].

While Transformers have facilitated the use of Vision-Language Models [41, 24] and Vision-Language Action Models [4, 46, 20], the use of language in these frameworks is primarily for high-level reasoning [30, 10, 60]. To the best of our knowledge, language hasn't been utilized to describe the intrinsic physical state of the robot itself to inform low-level kinodynamics modeling. Our work bridges this gap by leveraging the semantic richness of pre-trained language models to encode physical variations into robot kinodynamics.

III. APPROACH

Our objective is to learn a kinodynamics model that accurately predicts the future state of a mechanically compromised robot. We formulate this as learning a kinodynamics function conditioned on the robot's state-action history and a textual description of the damage.

A. Problem Formulation

We define the robot's state at time t as $\mathbf{s}_t \in \mathcal{S} \subset \mathbb{SE}(3)$, representing the 6-Degree of Freedom (DoF) pose: $\mathbf{s}_t = [x_t, y_t, z_t, \text{roll}_t, \text{pitch}_t, \text{yaw}_t]^\top$. The control input applied to the robot, i.e., linear and angular velocities, is denoted by $\mathbf{u}_t \in \mathcal{U}$: $\mathbf{u}_t = [v_t, \omega_t]^\top$. Under nominal conditions, the evolution of the robot's state on uniform terrain is governed by a nominal forward kinodynamics function $f_{\text{nominal}} : \mathcal{S} \times \mathcal{U} \to \mathcal{S}$:

$$\mathbf{s}_{t+1} = f_{\text{nominal}}(\mathbf{s}_t, \mathbf{u}_t). \tag{1}$$

Physically, this function encapsulates how the robot's chassis, suspension, and powertrain distribute forces. The chassis distributes external impact forces across a mesh of beams and nodes, while the suspension dampens vertical impulses to maintain consistent tire contact. In a healthy state, the mapping

from $\mathbf{u}_t$ to the resultant forces acting on the center of mass remains unbiased and follows standard rigid body mechanics.

However, when a robot incurs structural damage, this physical mapping is fundamentally altered. A specific damage instance, denoted as $d \in \mathcal{D}$, breaks the nominal force distribution pathways in the chassis or energy transfer pathways in the powertrain. For example, a snapped half-shaft prevents the powertrain from transferring torque to a specific wheel, while a compromised suspension linkage alters how ground reaction forces are resolved into the chassis. These mechanical failures introduce a non-linear disturbance to the system dynamics.

We posit that while these disturbances are complex, they manifest as consistent, distinct patterns in robot's kinodynamic behavior. For example, a burst front left tire will cause vertical impulses in the front left side of the vehicle along with lateral movement towards the left. To account for these deviations, we extend the kinodynamics formulation to depend explicitly on the damage instance d_t at time t. Furthermore, because the dynamic effects of damage often exhibit temporal dependencies that cannot be easily captured in a single state transition, we condition our model on a history length H. We formulate the damage-aware forward kinodynamics as a function $f_{\text{damage-aware}} : \mathcal{S}^H \times \mathcal{U}^H \times \mathcal{D} \to S$:

$$\mathbf{s}_{t+1} = f_{\text{damage-aware}}(\mathbf{s}_{t-H+1:t}, \mathbf{u}_{t-H+1:t}, d_t). \tag{2}$$

Here, d_t acts as a latent variable that effectively switches the context of the prediction based on the current damage. The core challenge then becomes how to define and use d_t for kinodynamic predictions. In our approach, we aim to construct a representation space for damage that is both semantically descriptive and kinodynamically grounded (see Fig. 2a).

B. Damage Representation

To enable zero-shot adaptation, our damage representation space aims to bridge the gap between natural language descriptions and kinodynamic behaviors. Since characterizing heterogeneous structural damages via a single scalar metric is difficult, we use a natural language description of the robot's health state at time t, denoted as x_t. We employ a pre-trained sentence Transformer ϕ to encode this x_t into a semantic embedding. Sentence Transformers are optimized such that semantically similar damage descriptions result in spatially proximal embedding vectors, preserving the linguistic meaning of the damage description.

However, linguistic proximity does not always guarantee kinodynamic proximity. For example, a "broken axle" and a "snapped half-shaft" are linguistically different but kinodynamically identical. To bridge this gap, we must align the linguistic semantic space with the physical behavioral space. We posit that damages manifest as consistent, distinct patterns in the robot's state-action trajectory. We define the robot's kinodynamic behavior at time t as its state-action trajectory over the history length H:

$$\tau_{t,H} = ((s_{t-H+1}, u_{t-H+1}), ..., (s_t, u_t)) \in \mathbb{R}^{(6+2) \times H}.$$

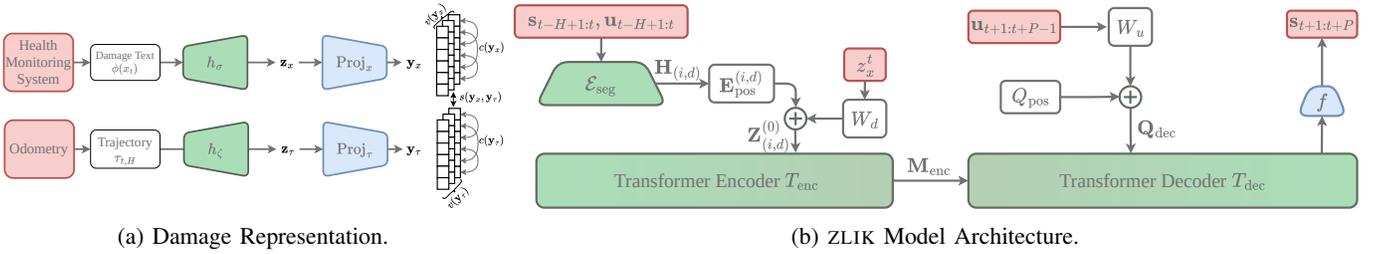(a) Damage Representation.　　　　　(b) ZLIK Model Architecture.

Fig. 2: (a) Constructing a representation space for damage that is both semantically descriptive and kinodynamically grounded. (b) ZLIK leverages a Transformer Encoder-Decoder structure to approximate damaged kinodynamics.

Our objective is to construct a shared representation space $\mathcal{Z}$ where the semantic encoding of the damage is structurally aligned with its induced kinodynamic behavior. We use Multi-Layer Perceptrons (MLP) to map the natural language description of the health state x and kinodynamic behavior τ into this shared space, resulting in $\mathbf{z}_x = h_\sigma(\phi(x))$ and $\mathbf{z}_\tau = h_\zeta(\tau)$.

As shown in Fig. 2a, we align the projection of these representations, denoted as $\mathbf{y}$, using Variance-Invariance-Covariance Regularization (VICReg) [2]. This loss function forces the projected text embedding $\mathbf{z}_x$ to be predictive of the kinodynamic behavior $\mathbf{z}_\tau$. The loss is defined as a weighted sum of three terms:

$$\mathcal{L}_{\text{align}}(\mathbf{y}_x, \mathbf{y}_\tau) = \overbrace{\lambda s(\mathbf{y}_x, \mathbf{y}_\tau)}^{\text{Invariance}} + \overbrace{\mu[v(\mathbf{y}_x) + v(\mathbf{y}_\tau)]}^{\text{Variance}} + \overbrace{\nu[c(\mathbf{y}_x) + c(\mathbf{y}_\tau)]}^{\text{Covariance}},$$

where λ, μ, and ν are hyperparameters controlling the importance of each term. Here, the invariance term minimizes the Mean Squared Error (MSE) between the semantic damage description and the observed kinodynamic behavior. This forces the language embedding to be a proxy for the physical reality of the damage:

$$s(\mathbf{y}_x, \mathbf{y}_\tau) = \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{y}_x^{(i)} - \mathbf{y}_\tau^{(i)}\|_2^2,$$

where N is the batch size. Furthermore, the variance term ensures that the damage embeddings maintain diversity across the batch, capturing heterogeneity of structural failures. This prevents the model from collapsing to a trivial solution where all the embeddings map to a single point with zero variance. Variance of each embedding dimension is forced to be above a threshold γ by a hinge loss on the standard deviation σ:

$$v(\mathbf{y}) = \frac{1}{K} \sum_{j=1}^{K} \max(0, \gamma - \sigma(\mathbf{y}^j, \epsilon)),$$

where $\sigma(x, \epsilon) = \sqrt{\text{Var}(x) + \epsilon}$, ϵ is a small scalar value to prevent numerical instabilities, and K is the total number of dimensions in the embedding vector $\mathbf{y}$. Finally, the covariance term ensures the damage embedding captures independent factors of the dynamics rather than redundant correlations. We define the covariance matrix $C(\mathbf{y})$ over the batch dimension N to capture the correlation between embedding features:

$$C(\mathbf{y}) = \frac{1}{N-1} \sum_{i=1}^{N} (\mathbf{y}^{(i)} - \bar{\mathbf{y}})(\mathbf{y}^{(i)} - \bar{\mathbf{y}})^T, \text{ where } \bar{\mathbf{y}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{y}^i.$$

The sum of squared off-diagonal coefficients are minimized to de-correlate the dimensions:

$$c(\mathbf{y}) = \frac{1}{K} \sum_{i \neq j} [C(\mathbf{y})]_{i,j}^2.$$

By forcing the off-diagonal coefficients to zero, we maximize the information content of the damage representation, allowing the downstream kinodynamics model to leverage independent axes of variation. The scale factor of $1/K$ ensures the covariance criterion scales linearly with the dimensionality.

Finally, given a dataset of N paired samples $\{(x_i, \tau_i)\}_{i=1}^{N}$, we seek the optimal encoder parameters σ^* and ζ^* to minimize this loss expectation:

$$\sigma^*, \zeta^* = \underset{\sigma, \zeta}{\arg\min} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{\text{align}}(h_\sigma(\phi(x_i)), h_\zeta(\tau_i)). \quad (3)$$

This formulation ensures that the learned embedding space $\mathcal{Z}$ is robust across diverse structural damages.

C. Downstream Kinodynamics Modeling

After the self-supervised training to correlate the damage representation and kinodynamic behavior, we freeze the projection head h_σ. We replace the damage instance d_t in Eqn. (2) with encoded representation of the damage instance's language description, $\mathbf{z}_x^t$. We assume language description of damage x_t is provided by a vehicle health monitoring system. By utilizing the frozen projection head h_σ, we learn to approximate forward kinodynamic function $f_\theta(\cdot)$ as a downstream task. The forward kinodynamic function is given by:

$$\underbrace{\mathbf{s}_{t+1:t+P}}_{\text{Predicted Future States}} = f_\theta(\underbrace{\mathbf{s}_{t-H+1:t}, \mathbf{u}_{t-H+1:t}}_{\text{History}}, \underbrace{\mathbf{u}_{t+1:t+P-1}}_{\text{Future Actions}}, \mathbf{z}_x^t),$$

$$(4)$$

where θ is learnable parameters and P is the prediction horizon. We learn the optimal parameters θ^* in a supervised fashion using our dataset of damaged robot operation in BeamNG.tech, $\{\mathbf{s}_{t+1:t+P}^j, \mathbf{s}_{t-H+1:t}^j, \mathbf{u}_{t-H+1:t}^j, \mathbf{u}_{t+1:t+P-1}^j, \mathbf{z}_x^{t,j}\}_{j=1}^{N}$. We seek the optimal kinodynamics parameters θ^*:
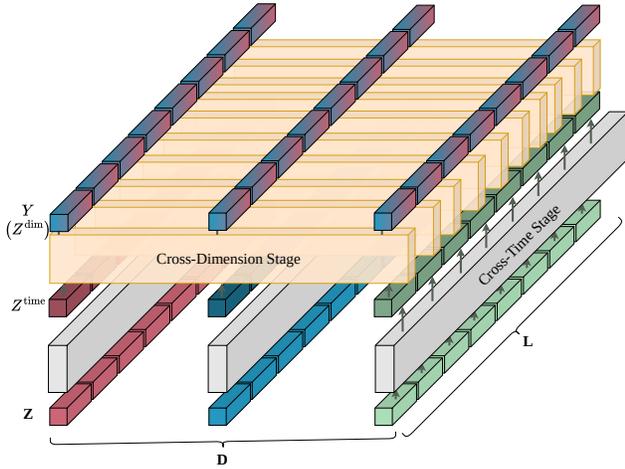
Fig. 3: ZLIK's Transformer Encoder uses a two-stage attention architecture for both time (gray) and state dimension (orange).

$$\theta^* = \underset{\theta}{\arg\min} \sum_{j=1}^{N} \left\| \overbrace{\mathbf{s}_{t+1:t+P}^{j}}^{\text{Actual Future States}} - \right.$$

$$\left. f_\theta\big(\underbrace{\mathbf{s}_{t-H+1:t}^{j}, \mathbf{u}_{t-H+1:t}^{j}}_{\text{History}}, \underbrace{\mathbf{u}_{t+1:t+P-1}^{j}}_{\text{Future Actions}}, \mathbf{z}_x^{t,j}\big) \right\|^2 .$$

D. Spatiotemporal Attention

We approximate the kinodynamic function in Eqn. (4) using a Transformer Encoder-Decoder architecture (Fig. 2b), similar to Crossformer [61], designed to capture spatiotemporal dependencies as shown in Fig. 3. To tokenize the robot's state history for this model, we employ a Dimension-Segment Embedding, denoted as $\mathcal{E}_{\text{seg}}$. This embedding transforms the raw state-action history $\tau_{t,H}$ into a structured feature array $\mathbf{H}$ by preserving the independence of state dimensions.

Specifically, $\mathcal{E}_{\text{seg}}$ partitions the temporal history of each dimension into local windows of length L_{seg}, which are then projected into the Transformer's latent dimension d_{model}. This results in a 2D feature array $\mathbf{H} \in \mathbb{R}^{L \times D \times d_{\text{model}}}$, where D represents the state dimensions and $L = H/L_{\text{seg}}$ denotes the number of resulting time segments (Fig. 3 bottom).

To condition the dynamics on the robot's structural health, as shown in Fig. 2b, we inject the semantic damage embedding $\mathbf{z}_x \in \mathbb{R}^{d_{\text{damage}}}$ directly into this latent space along with learnable positional embedding to create the input to the encoder layer,

$$\mathbf{Z}_{i,d}^{(0)} = \mathbf{H}_{i,d} + \mathbf{E}_{\text{pos}}^{(i,d)} + W_d \mathbf{z}_x.$$

The indices (i,d) correspond to the time segment and state dimension. $\mathbf{E}_{\text{pos}} \in \mathbb{R}^{d_{\text{model}}}$ is a learnable positional embedding, and $W_d \in \mathbb{R}^{d_{\text{model}} \times d_{\text{damage}}}$ is a learnable projection matrix. By adding the damage context $W_d \mathbf{z}_x$ globally to every segment (i,d) we bias the robot's entire history with semantic signature of the damage.

The Transformer Encoder T_{enc} processes $\mathbf{Z}^{(0)}$ through a stack of two-stage attention layers to capture the dependencies along the spatial and temporal dimensions. First, the model uses Multi-head Self-Attention (MSA) across time segments i to capture the temporal evolution of each dimension d (Fig. 3 gray). Second, the model captures inter-variate dependencies by applying MSA across dimensions d for each time segment i (Fig. 3 orange). This disentangled attention mechanism enables the network to learn how the injected damage $\mathbf{z}_x$ specifically alters the correlations between physical dimensions. We denote the output of the encoder as the damage-informed memory, $\mathbf{M}_{\text{enc}} = T_{\text{enc}}(\mathbf{Z}^{(0)})$.

To predict the future trajectory, we employ a non-auto-regressive Transformer decoder T_{dec} conditioned on the sequence of future actions $\mathbf{u}_{t+1:t+P-1}$. We project the future actions into a latent space using a learnable projection matrix W_u and then add to a learnable positional embedding $\mathbf{Q}_{\text{pos}}$, generating the decoder query $\mathbf{Q}_{\text{dec}} = \mathbf{Q}_{\text{pos}} + W_u \mathbf{u}_{t+1:t+P-1}$. The decoder performs cross-attention on the given query to retrieve relevant dynamics from the encoder's damage-informed memory $\mathbf{M}_{\text{enc}}$. The final decoder output is projected to the state dimension via a linear head $\hat{\mathbf{s}}_{t+1:t+P} = f(T_{\text{dec}}(\mathbf{Q}_{\text{dec}}, \mathbf{M}_{\text{enc}})) \in \mathbb{R}^{P \times 6}$.

IV. IMPLEMENTATIONS

In this section, we present implementation details of our approach and experiments.

A. Simulation, Dataset, Robot, and Damages

For data generation and experimentation, we use a high-fidelity soft-body physics simulator, BeamNG.tech, which has been extensively used for dynamics simulation in off-road environments. The vehicles in BeamNG.tech are built as a network of nodes and meshes, while also simulating the elasticity and strength of each mechanical part. In our experiment, we use a full-size vehicle as our robot. The simulator maintains an internal robot health state accessible in JSON format. We convert the relevant JSON keys and values into natural language descriptions x_t.

Our dataset includes six damage classes including single and a combination of broken part(s):

- **Tire Puncture**: a single punctured tire;
- **Tire & Spring**: a single punctured tire along with a broken suspension on the same wheel;
- **Multiple Tires Punctured & Suspensions Broken (MTP & SB)**: damages to adjacent tires and suspensions;
- **Broken Axle**: either of the axles broken;
- **Fall**: Drop the robot from a random height in $[5, 15]$ meters at a random orientation to incur multiple damages;
- **No Damage**: healthy vehicle.

The mechanically compromised robots are then operated using a random walk algorithm. We collect vehicle state and action to generate training data for both damage representation pre-training and downstream kinodynamics learning. The dataset is collected uniformly across all classes. In total, we

collect 150K data points, which are divided into an 80-20 train-validation split. Another unseen dataset of 30K is collected for testing, which are used to report all experiment results.

For real-world experiments, we use an open-source, 1/10$^{\text{th}}$ scale Verti-4-Wheeler platform, V4W [8]. Real-time state information is provided at 100Hz by a motion capture setup. The V4W is damaged by manually removing a wheel or the foam material inside a tire. All kinodynamics training and inference is conducted on an NVIDIA A6000 GPU with 48GB of RAM wirelessly connected with the robot.

B. Damage Representation

To encode the natural language damage descriptions, we utilize the EmbeddingGemma (300M) [51] model as our sentence Transformer backbone ϕ. The model is chosen for its balance of lightweight inference and good semantic understanding.

The kinodynamic trajectory encoder h_ζ is implemented as a Transformer Encoder. The encoder processes a history window of $H = 200$ steps (10 seconds), where the input at each step is the concatenation of the relative pose in the robot frame and executed action. We utilize a learnable positional encoding to preserve temporal order. The encoder consists of 3 Transformer layers with 8 attention heads, a hidden dimension of 128, and a feed-forward dimension of 512. We apply a dropout of 0.2 for regularization. The output sequence is aggregated via global average pooling to extract a fixed-size feature vector.

Both the text embedding and the trajectory encoding are projected into a shared latent space of dimension K = 128 using three-layer MLPs with 256, 256, 128 hidden dimensions, Batch Normalization, and ReLU activations.

We align the semantic and kinodynamic representations, $\mathbf{z}_x$ and $\mathbf{z}_\tau$, using Eqn (3). The hyperparameters for the loss function are set to $\lambda = 25.0, \mu = 10.0$, and $\nu = 0.1$ after hyperparameter tuning.

C. Damaged Kinodynamics Model

Our Transformer implementation takes as input the state-action pairs over a history $H = 40$ (2 seconds), a sequence of future actions of length $P = 10$ (0.5 seconds), and the semantic damage embedding $\mathbf{z}_x^t$ as the input.

The input state history is segmented into patches of length $L_{\text{seg}} = 4$, resulting in 10 segments per state dimension. These segments are projected into a latent dimension of $d_{\text{model}} = 256$. The learnable positional embedding added to these segments preserves both temporal ordering and state dimension identity. The 128-dimensional damage embedding $\mathbf{z}_x$ is projected via a linear layer to match d_{model} and is added element-wise to the position-encoded segments to condition the global context. The Transformer Encoder consists of $N = 3$ layers utilizing two-stage attention. We implement the temporal attention phase using a local window size of $W_{\text{size}} = 2$ to efficiently capture local dependencies. The Transformer Decoder consists of $N = 4$ layers. To condition the generation on control inputs, the future action sequence $\mathbf{u}_{t+1:t+P-1}$ is segmented, projected to d_{model} via a linear layer, and added directly to the Decoder's

learnable positional query embeddings. The Decoder then attends to the Encoder memory via cross-attention.

The model is implemented in PyTorch. We set $h = 4$ attention heads, a routing factor of $c = 10$, and a feed-forward dimension of $d_{\text{ff}} = 512$ for both Encoder and Decoder layers. We apply a dropout of 0.2 throughout the network. The final Decoder output is projected to the state dimension via a linear head, and the network is trained to minimize MSE between the ground truth and predicted trajectories in the next 0.5 seconds.

V. EXPERIMENTS

We design our experiments to answer four core research questions regarding the efficacy of ZLIK.

A. Research Question 1: Zero-Shot Performance

In this experiment, we demonstrate ZLIK's ability to achieve zero-shot adaptation to structural failures and validate the necessity of explicit damage representation in kinodynamics modeling for structurally compromised robots. We compare ZLIK's trajectory prediction performance against a non-adaptive and a state-of-the-art adaptive baseline, ANYCAR [55]. The non-adaptive model employs a Transformer Encoder-Decoder architecture similar to our backbone but is trained exclusively on state-action data from a structurally healthy robot. This model does not receive damage embedding as input, thus named CleanTransformer. We utilize the ANYCAR model [55] as a baseline for kinodynamics adaptability. Like the CleanTransformer, the ANYCAR model does not utilize damage embedding and is also trained on state-action data of the healthy robot.

To evaluate the efficacy of online adaptation versus our zero-shot approach, we fine-tune the ANYCAR model on new data collected on damaged robots. For a fair comparison with ZLIK's adaptability, we fine-tune ANYCAR with only 20 seconds of data. To show ANYCAR's full capability with larger amount of new iteration data collected after damage has occurred, we also fine-tune it on 5 and 10 minutes of new data.

Results presented in Table I show ZLIK significantly outperforms all baselines across every damage class. Most notably, ZLIK achieves superior prediction accuracy in a zero-shot manner, effectively generalizing to unseen damages without requiring any online data collection or weight updates.

In contrast, ANYCAR struggles to generalize to the altered dynamics in zero-shot. While the fine-tuned ANYCAR variants show progressive improvement as the duration of online interaction increases from 20 seconds to 10 minutes, they fail to match the precision of ZLIK. Even after 10 minutes of data collection, which involves significant risk when operating a mechanically compromised platform, the fine-tuned models still yield higher prediction errors than our instant, language-conditioned approach. Furthermore, CleanTransformer performs poorly across all scenarios. Despite possessing a similar backbone as our method, its inability to account for the structural health state leads to substantial prediction errors.

TABLE I: Comparison of Trajectory Prediction Error across All Damage Classes. ZLIK achieves the lowest error in zero-shot in every scenario, even outperforming models with 10 minutes of environment-specific fine-tuning.

Damage Class	Model	MSE $\pm$ Std
Overall	**ZLIK (Ours)**	**0.50 $\pm$ 1.29**
	CleanTransformer	1.59 $\pm$ 1.55
	ANYCAR (Zero-Shot)	2.59 $\pm$ 2.44
Fall	**ZLIK (Ours)**	**1.52 $\pm$ 2.48**
	CleanTransformer	3.85 $\pm$ 2.46
	ANYCAR (Zero-Shot)	5.24 $\pm$ 3.66
	ANYCAR (20s Fine-tuning)	5.20 $\pm$ 3.37
	ANYCAR (5m Fine-tuning)	3.15 $\pm$ 3.33
	ANYCAR (10m Fine-tuning)	2.76 $\pm$ 3.13
Broken Axle	**ZLIK (Ours)**	**0.26 $\pm$ 0.80**
	CleanTransformer	0.97 $\pm$ 0.74
	ANYCAR (Zero-Shot)	2.02 $\pm$ 1.40
	ANYCAR (20s Fine-tuning)	2.51 $\pm$ 1.90
	ANYCAR (5m Fine-tuning)	1.89 $\pm$ 1.67
	ANYCAR (10m Fine-tuning)	1.73 $\pm$ 1.71
Multiple Tires Punctured & Suspensions Broken	**ZLIK (Ours)**	**0.32 $\pm$ 0.13**
	CleanTransformer	1.57 $\pm$ 0.32
	ANYCAR (Zero-Shot)	3.39 $\pm$ 2.08
Punctured Tire	**ZLIK (Ours)**	**0.19 $\pm$ 0.16**
	CleanTransformer	0.91 $\pm$ 0.13
	ANYCAR (Zero-Shot)	1.32 $\pm$ 0.96
Broken Suspension	**ZLIK (Ours)**	**0.21 $\pm$ 0.20**
	CleanTransformer	0.92 $\pm$ 0.15
	ANYCAR (Zero-Shot)	1.67 $\pm$ 1.57
No Damage	**ZLIK (Ours)**	**0.45 $\pm$ 1.14**
	CleanTransformer	1.38 $\pm$ 0.80
	ANYCAR (Zero-Shot)	2.08 $\pm$ 1.61

B. Research Question 2: Architectural Necessity

To validate the architectural necessity of our 6-DoF state decomposition, we compare ZLIK against a Monolithic Transformer baseline. We use a Transformer Encoder-Decoder architecture that takes damage embedding as input but processes the robot's state as a single token, obscuring the independent relationships between dimensions.

Table II shows the comparison of MSE and its standard deviation across all classes. The Monolithic Transformer yields larger prediction error. This trend is also evident in Fig. 4, where we compare the errors in the decomposed states in damage classes with multiple mechanical failures (Fall and Multiple Tires Punctured & Suspensions Broken). This highlights that without state decomposition, the model cannot effectively disentangle the complex, non-linear disturbances induced by vehicle damage.

TABLE II: Impact of State Decomposition on Trajectory Prediction Accuracy for all Damage Classes. The Monolithic Transformer without state decomposition, yields larger error.

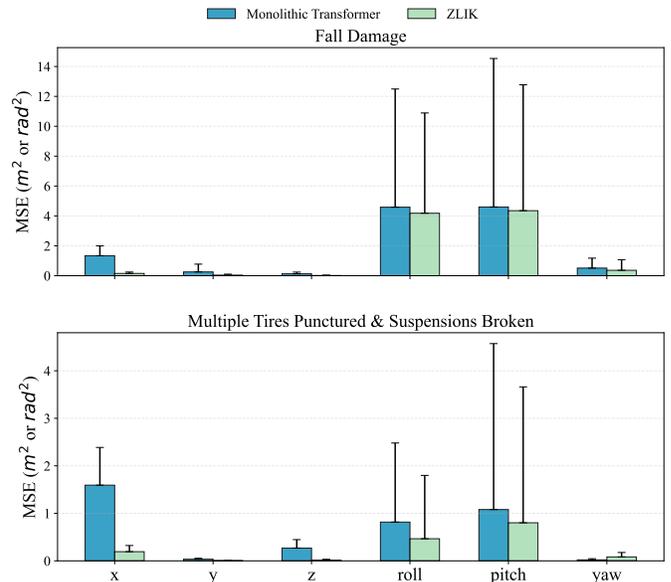| Model Architecture	State Decomposition	MSE $\pm$ Std
Monolithic Transformer	$\times$	0.81 $\pm$ 1.11
ZLIK (Ours)	$\checkmark$	**0.50 $\pm$ 1.09**

Fig. 4: ZLIK outperforms Monolithic Transformer in terms of MSE and standard deviation across state dimensions for Fall and Multiple Tires Punctured & Suspensions Broken.

C. Research Question 3: Semantic Grounding

We structure this experiment to evaluate ZLIK's ability to successfully align linguistic semantics and physical behaviors. Specifically, we test the hypothesis that the model uses semantic description to achieve high prediction accuracy. We choose trials of four distinct damage classes from the 30K unseen test set. Trials with matching descriptions are compared against trials where an incorrect description is randomly assigned.

The confusion matrix (Table III) shows a clear diagonal dominance, indicating that prediction error is minimized when the input language description matches the physical state of the robot. When the model is provided with an incorrect description for a given damage, the prediction error increases significantly. This validates that the model is effectively using the language-informed semantic embedding $\mathbf{z}_x$ to condition its kinodynamics prediction.

The fourth column of the confusion matrix highlights an interesting observation. Due to the severity of failures in the Multiple Tires Punctured & Suspensions Broken class, the maneuverability of such a damaged robot is significantly restricted. As the vehicle's range of possible motion is reduced, kinodynamic prediction becomes inherently easier, resulting in lower overall errors, considering the representation of state-action history can provide sufficient prior to bias the future
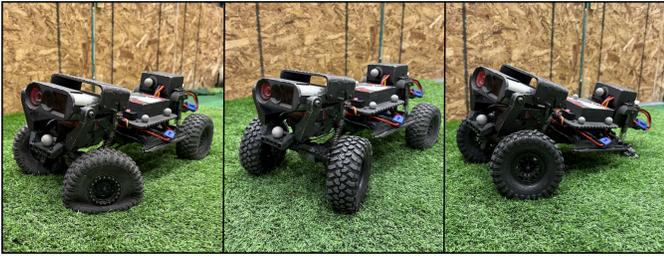
Fig. 5: Healthy Physical Robot Platform, V4W (middle), Used in the Cross-Embodiment Generalization Experiment, Compared with the Damaged V4W Variants, Front Left Tire Punctured (left) and Rear Left Wheel Removed (right).

state predictions. However, even in classes with low prediction errors, the inclusion of the correct damage embedding provides a measurable positive impact. This demonstrates that even when the dynamics are constrained by severe damages, semantic grounding allows the model to capture subtle independent nuances in independent DoFs.

TABLE III: The confusion matrix evaluates ZLIK's ability to ground linguistic semantics in physical behaviors. The diagonal dominance demonstrates that the model successfully utilizes the semantic embedding $\mathbf{z}_x^t$ to condition kinodynamics.

Input Language Description	Actual Physical State			
	Broken Axle	Fall	No Damage	MTP & SB
Broken Axle Description	**0.28** (±**0.82**)	0.38 (±0.88)	0.29 (±0.61)	0.21 (±0.41)
Fall Description	0.34 (±0.84)	**0.31** (±**0.85**)	0.36 (±0.63)	0.23 (±0.48)
No Damage Description	0.29 (±0.83)	0.39 (±0.88)	**0.28** (±**0.61**)	0.22 (±0.47)
MTP & SB Description	0.29 (±0.82)	0.37 (±0.88)	0.29 (±0.60)	**0.19** (±**0.39**)

D. Research Question 4: Cross-Embodiment Generalization

This experiment introduces a large domain shift as the test robot differs from the training robot not only in terms of the deployment environment (simulation vs real world), but also fundamental physical parameters like mass, dimensions, wheelbase, and actuation limits. To validate ZLIK, we test our model on a 1/10th scale physical V4W platform (Fig. 5) [7]. We alter the robot's health state by removing the foam inside the front left tire to simulate a punctured tire and removing the rear left wheel to simulate a detached wheel. The V4W with altered health states suffers from non-nominal kinodynamic behaviors compared to a healthy V4W platform, exhibiting the problem we aim to address with ZLIK. We compare the prediction error of ZLIK with the CleanTransformer, the Monolithic Transformer, and the ANYCAR model, as shown in Fig. 6.
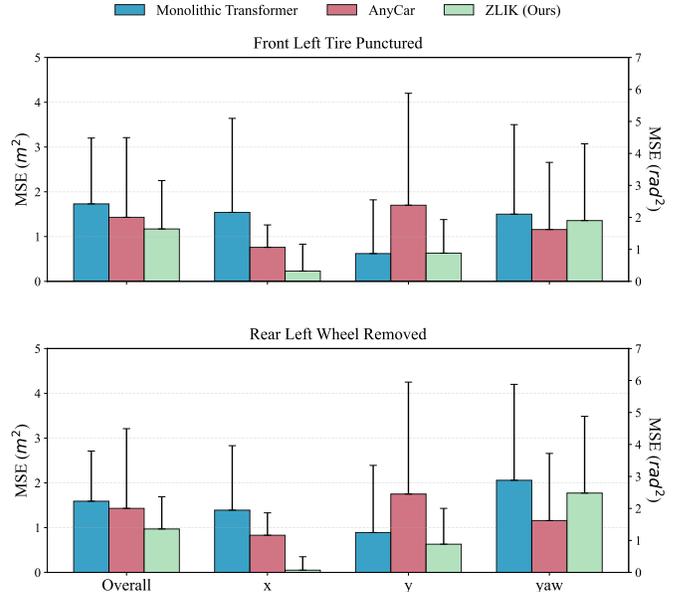

Fig. 6: Cross-Embodiment Evaluation on a Physical 1/10th Scale V4W Robot. ZLIK achieves the best performance in zero-shot on two unseen damage classes in most dimensions.

Transitioning from a full-sized simulated vehicle to a 1/10th scale physical robot inevitably introduces a performance gap. We observe that the relative error rate increases for all models on the physical platform compared to simulation. This increase is expected due to unmodeled physical factors, such as sensor noise, communication latency, and friction disparity inherent to the sim-to-real and full-to-1/10th scale gaps.

Despite the increase in relative error rate, ZLIK demonstrates improvement over the baselines in a zero-shot fashion. This suggests that our semantic embeddings capture the fundamental nature of the damage (e.g., "loss of a wheel leads to drag on that side") rather than overfitting to the full-size training vehicle in simulation. Even for the unseen 1/10th scale physical test robot, the semantic grounding provides a prior that is robust enough to outperform the three baselines, effectively bridging the gap across different vehicle embodiments.

VI. CONCLUSIONS

We present ZLIK, a natural language-informed kinodynamics modeling approach that can adapt to robot structural damage in a zero-shot manner. By grounding linguistic semantics in physical behaviors, ZLIK precisely predicts kinodynamics when facing a variety of vehicle damages without any online data collection and retraining, even crossing the sim-to-real as well as full-to-1/10th scale gaps. One potential avenue for future work is to consider terrain geometry and semantics in the kinodynamics model, as such environmental features also play a vital role in determining the changes in kinodynamics when facing structural damages.

REFERENCES

[1] Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378:686–707, 2019.

[2] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

[3] BeamNG GmbH. BeamNG.tech. URL https://www.beamng.tech/.

[4] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alexander Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *Robotics: Science and Systems XIX*, 2023.

[5] Xiaoyi Cai, James Queeney, Tong Xu, Aniket Datar, Chenhui Pan, Max Miller, Ashton Flather, Philip R Osteen, Nicholas Roy, Xuesu Xiao, et al. Pietra: Physics-informed evidential learning for traversing out-of-distribution terrain. *IEEE Robotics and Automation Letters*, 2025.

[6] Antoine Cully, Jeff Clune, Danesh Tarapore, and Jean-Baptiste Mouret. Robots that can adapt like animals. *Nature*, 521(7553):503–507, 2015.

[7] Aniket Datar, Chenhui Pan, Mohammad Nazeri, Anuj Pokhrel, and Xuesu 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)*, pages 5438–5443. IEEE, 2024.

[8] Aniket Datar, Chenhui Pan, Mohammad Nazeri, and Xuesu Xiao. Toward wheeled mobility on vertically challenging terrain: Platforms, datasets, and algorithms. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 16322–16329. IEEE, 2024.

[9] Aniket Datar, Chenhui Pan, and Xuesu Xiao. Learning to model and plan for wheeled mobility on vertically challenging terrain. *IEEE Robotics and Automation Letters*, 2024.

[10] Mohamed Elnoor, Kasun Weerakoon, Gershom Seneviratne, Ruiqi Xian, Tianrui Guan, Mohamed Khalid M Jaffar, Vignesh Rajagopal, and Dinesh Manocha. Vlm-gronav: Robot navigation using physically grounded vision-language models in outdoor environments. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2391–2398. IEEE, 2025.

[11] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[12] Jason Gibson, Anoushka Alavilli, Erica Tevere, Evangelos A Theodorou, and Patrick Spieler. Dynamics modeling using visual terrain features for high-speed autonomous off-road driving. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9809–9815. IEEE, 2025.

[13] Jake Grigsby, Zhe Wang, Nam Nguyen, and Yanjun Qi. Long-range transformers for dynamic spatiotemporal forecasting. *arXiv preprint arXiv:2109.12218*, 2021.

[14] Tyler Han, Alex Liu, Anqi Li, Alex Spitzer, Guanya Shi, and Byron Boots. Model predictive control for aggressive driving over uneven terrain. *arXiv preprint arXiv:2311.12284*, 2023.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

[16] Zechen Hu, Tong Xu, Xuesu Xiao, and Xuan Wang. Carol: Context-aware adaptation for robot learning. *IEEE Robotics and Automation Letters*, 10(11):12063–12070, 2025. doi: 10.1109/LRA.2025.3619779.

[17] Tyler Ingebrand, Adam J Thorpe, and Ufuk Topcu. Function encoders: A principled approach to transfer learning in hilbert spaces. *arXiv preprint arXiv:2501.18373*, 2025.

[18] Sanghun Jung, JoonHo Lee, Xiangyun Meng, Byron Boots, and Alexander Lambert. V-strong: Visual self-supervised traversability learning for off-road navigation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1766–1773. IEEE, 2024.

[19] Haresh Karnan, Kavan Singh Sikand, Pranav Atreya, Sadegh Rabiee, Xuesu Xiao, Garrett Warnell, Peter Stone, and Joydeep 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)*, pages 3294–3301. IEEE, 2022.

[20] Kento Kawaharazuka, Jihoon Oh, Jun Yamada, Ingmar Posner, and Yuke Zhu. Vision-language-action models for robotics: A review towards real-world applications. *IEEE Access*, 2025.

[21] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[22] Hojin Lee, Junsung Kwon, and Cheolhyeon Kwon. Learning-based uncertainty-aware navigation in 3d off-road terrains. In *2023 IEEE International Conference on*

Robotics and Automation (ICRA), pages 10061–10068. IEEE, 2023.

[23] Jacob Levy, Jason Gibson, Bogdan Vlahov, Erica Tevere, Evangelos Theodorou, David Fridovich-Keil, and Patrick Spieler. Meta-learning online dynamics model adaptation in off-road autonomous driving. *arXiv preprint arXiv:2504.16923*, 2025.

[24] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[25] Faraz Lotfi, Khalil Virji, Farnoosh Faraji, Lucas Berry, Andrew Holliday, David Meger, and Gregory Dudek. Uncertainty-aware hybrid paradigm of nonlinear mpc and model-based rl for offroad navigation: Exploration of transformers in the predictive model. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2925–2931, 2024. doi: 10.1109/ICRA57147.2024.10610452.

[26] Elena Sorina Lupu, Fengze Xie, James Alan Preiss, Jedidiah Alindogan, Matthew Anderson, and Soon-Jo Chung. Magic vfm-meta-learning adaptation for ground interaction control with visual foundation models. *IEEE Transactions on Robotics*, 41:180–199, 2024.

[27] Parv Maheshwari, Wenshan Wang, Samuel Triest, Matthew Sivaprakasam, Shubhra Aich, John G Rogers III, Jason M Gregory, and Sebastian Scherer. Piaug–physics informed augmentation for learning vehicle dynamics for off-road navigation. *arXiv preprint arXiv:2311.00815*, 2023.

[28] Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, et al. Isaac gym: High performance gpu-based physics simulation for robot learning. *arXiv preprint arXiv:2108.10470*, 2021.

[29] Xiangyun Meng, Nathan Hatch, Alexander Lambert, Anqi Li, Nolan Wagener, Matthew Schmittle, JoonHo Lee, Wentao Yuan, Zoey Chen, Samuel Deng, et al. Terrainnet: Visual modeling of complex terrain for high-speed, off-road navigation. *arXiv preprint arXiv:2303.15771*, 2023.

[30] Chen Min, Jilin Mei, Heng Zhai, Shuai Wang, Tong Sun, Fanjie Kong, Haoyang Li, Fangyuan Mao, Fuyang Liu, Shuo Wang, et al. Advancing off-road autonomous driving: The large-scale orad-3d dataset and comprehensive benchmarks. *arXiv preprint arXiv:2510.16500*, 2025.

[31] Anusha Nagabandi, Ignasi Clavera, Simin Liu, Ronald S Fearing, Pieter Abbeel, Sergey Levine, and Chelsea Finn. Learning to adapt in dynamic, real-world environments through meta-reinforcement learning. *arXiv preprint arXiv:1803.11347*, 2018.

[32] Tomáš Nagy, Ahmad Amine, Truong X Nghiem, Ugo Rosolia, Zirui Zang, and Rahul Mangharam. Ensemble gaussian processes for adaptive autonomous driving on multi-friction surfaces. *IFAC-PapersOnLine*, 56(2):494–500, 2023.

[33] Mohammad Nazeri, Aniket Datar, Anuj Pokhrel, Chenhui Pan, Garrett Warnell, and Xuesu Xiao. Verticoder: Self-supervised kinodynamic representation learning on vertically challenging terrain. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6536–6543. IEEE, 2025.

[34] Mohammad Nazeri, Anuj Pokhrel, Alexandyr Card, Aniket Datar, Garrett Warnell, and Xuesu Xiao. Vertiformer: A data-efficient multi-task transformer for off-road robot mobility. *arXiv preprint arXiv:2502.00543*, 2025.

[35] Jingyun Ning and Madhur Behl. DKMGP: A gaussian process approach to multi-task and multi-step vehicle dynamics modeling in autonomous racing. In Necmiye Ozay, Laura Balzano, Dimitra Panagou, and Alessandro Abate, editors, *7th Annual Learning for Dynamics & Control Conference*, volume 283 of *Proceedings of Machine Learning Research*, pages 59–71. PMLR, 2025.

[36] Chris J. Ostafew, Angela P. Schoellig, and Timothy D. Barfoot. Robust constrained learning-based nmpc enabling reliable mobile robot path tracking. *The International Journal of Robotics Research*, 35(13):1547–1563, 2016. doi: 10.1177/0278364916645661. URL https://doi.org/10.1177/0278364916645661.

[37] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

[38] Hans B. Pacejka and Egbert Bakker. The magic formula tyre model. *Vehicle System Dynamics*, 21(sup001):1–18, 1992. doi: 10.1080/00423119208969994. URL https://doi.org/10.1080/00423119208969994.

[39] Anuj Pokhrel, Mohammad Nazeri, Aniket Datar, and Xuesu Xiao. CAHSOR: Competence-aware high-speed off-road ground navigation in SE(3). *IEEE Robotics and Automation Letters*, 2024.

[40] Sadegh Rabiee and Joydeep Biswas. A friction-based kinematic model for skid-steer wheeled mobile robots. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8563–8569. IEEE, 2019.

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[42] Rajesh Rajamani. *Vehicle dynamics and control*. Springer, 2006.

[43] Kate Rakelly, Aurick Zhou, Chelsea Finn, Sergey Levine, and Deirdre Quillen. Efficient off-policy meta-reinforcement learning via probabilistic context variables. In *International conference on machine learning*, pages 5331–5340. PMLR, 2019.

[44] Adrian Remonda, Nicklas Hansen, Ayoub Raji, Nicola

Musiu, Marko Bertogna, Eduardo Veas, and Xiaolong Wang. A simulation benchmark for autonomous racing with large-scale human data. *Advances in Neural Information Processing Systems*, 37:102078–102100, 2024.

[45] Amirreza Shaban, Xiangyun Meng, JoonHo Lee, Byron Boots, and Dieter Fox. Semantic terrain classification for off-road autonomous driving. In *Conference on Robot Learning*, pages 619–629. PMLR, 2022.

[46] Dhruv Shah, Błażej Osiński, Sergey Levine, et al. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. In *Conference on robot learning*, pages 492–504. PMLR, 2023.

[47] Alessandro Tasora, Radu Serban, Hammad Mazhar, Arman Pazouki, Daniel Melanz, Jonathan Fleischmann, Michael Taylor, Hiroyuki Sugiyama, and Dan 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*, pages 19–49. Springer, 2016.

[48] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033, 2012. doi: 10.1109/IROS.2012.6386109.

[49] Samuel Triest, Matthew Sivaprakasam, Shubhra Aich, David Fan, Wenshan Wang, and Sebastian Scherer. Velociraptor: Leveraging visual foundation models for label-free, risk-aware off-road navigation. In *8th Annual Conference on Robot Learning*, 2024.

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[51] Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, et al. Embeddinggemma: Powerful and lightweight text representations. *arXiv preprint arXiv:2509.20354*, 2025.

[52] Sean J Wang, Honghao Zhu, and Aaron M Johnson. Pay attention to how you drive: Safe and adaptive model-based reinforcement learning for off-road driving. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 16954–16960. IEEE, 2024.

[53] Maggie Wigness, Sungmin Eum, John G. Rogers, David Han, and Heesung 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)*, pages 5000–5007, 2019. doi: 10.1109/IROS40897.2019.8968283.

[54] Grady Williams, Paul Drews, Brian Goldfain, James M Rehg, and Evangelos A Theodorou. Information-theoretic model predictive control: Theory and applications to autonomous driving. *IEEE Transactions on Robotics*, 34(6):1603–1622, 2018.

[55] Wenli Xiao, Haoru Xue, Tony Tao, Dvij Kalaria, John M Dolan, and Guanya Shi. Anycar to anywhere: Learning universal dynamics model for agile and adaptive mobility. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8819–8825. IEEE, 2025.

[56] Xuesu Xiao, Joydeep Biswas, and Peter Stone. Learning inverse kinodynamics for accurate high-speed off-road navigation on unstructured terrain. *IEEE Robotics and Automation Letters*, 6(3):6054–6060, 2021.

[57] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

[58] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024.

[59] Xihang Yu, Sangli Teng, Theodor Chakhachiro, Wenzhe Tong, Tingjun Li, Tzu-Yuan Lin, Sarah Koehler, Manuel Ahumada, Jeffrey M. Walls, and Maani Ghaffari. Fully proprioceptive slip-velocity-aware state estimation for mobile robots via invariant kalman filtering and disturbance observer. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8096–8103, 2023. doi: 10.1109/IROS55552.2023.10342519.

[60] Liangdong Zhang, Yiming Nie, Haoyang Li, Fanjie Kong, Baobao Zhang, Shunxin Huang, Kai Fu, Chen Min, and Liang Xiao. A vision-language-action model with visual prompt for off-road autonomous driving. *arXiv preprint arXiv:2601.03519*, 2026.

[61] Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=vSVLM2j9eie.

[62] Zhipeng Zhao, Bowen Li, Yi Du, Taimeng Fu, and Chen Wang. Physord: a neuro-symbolic approach for physics-infused motion prediction in off-road driving. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 11670–11677. IEEE, 2024.

A. Training Dataset

We collect the training dataset across six damage classes in the BeamNG.tech simulator. The total dataset comprises 153,387 trajectory data points, distributed as follows:

TABLE IV: Training Dataset Trajectory Distribution

Damage Class	No. of Trajectories
No Damage	17,955
Tire Puncture	20,007
Broken Suspension	25,308
Multiple Tires Punctured & Suspensions Broken	27,873
Fall	46,854
Broken Axle	15,390
Total	**153,387**

B. Simulation Experiment Dataset

For **Research Questions 1 and 2**, we evaluate the model using a different robot configuration to test zero-shot adaptability:

- **Robot Mass:** 1315 kg
- **Drivetrain:** Front-Wheel Drive (FWD)
- **Test Data:** 15,390 trajectory data points with uniform class distribution.

For **Research Question 3**, we evaluate the model in a test set generated on the same robot used for training data generation.

- **Robot Mass:** 1888 kg
- **Drivetrain:** 4-Wheel Drive (4WD)
- **Test Data:** Uniformly sampled 15,390 trajectories to match the size of previous experiments.

C. V4W Dataset

The real-world validation is conducted on the V4W platform at a sampling rate of 20Hz. The data collected for the two damage classes is summarized in Table V.

TABLE V: Real-World Dataset (V4W Platform)

Damage Class	Trajectories	Approx. Duration (min)
Front-Left Puncture	4,705	3.92
Rear-Left Removed	4,355	3.63