Towards Capability-Aware Traversability Navigation for Unstructured Environments

Abstract— **Estimating traversability in unstructured environments requires conditioning on robot embodiment, as the same terrain can be traversable for one platform and unsafe for another. Yet, existing methods often struggle to transfer policies across distinct robot morphologies, relying on late-stage trajectory filtering that fails to intrinsically encode platform-specific constraints. In this paper, we propose Capability-Aware Traversability (CAT), a unified framework that embeds physical limits directly into the spatial feature space. CAT leverages an interactive zero-shot annotation pipeline grounded in physical trajectories to generate dense supervision masks. To align visual representations with specific robotic embodiments, the architecture dynamically modulates semantic terrain maps using robot-specific traversability vectors via Spatially-Adaptive Denormalization (SPADE) blocks. Extensive evaluations on both human-annotated and trajectory-aligned datasets demonstrate that CAT outperforms state-of-the-art baselines, achieving a 12.6% improvement in mean traversability and a 13.5% increase in Area Under the Curve (AUC). Real-world deployments on diverse platforms, including a legged quadruped and a wheeled skid-steer, further demonstrate the framework's ability to perform robust, real-time (5 Hz) embodiment-aware obstacle avoidance in unstructured environments.**

I. INTRODUCTION

Semantic perception provides an essential foundation for safe navigation in the real world. Spotting a muddy trail, one intuitively categorizes visual information from the environment to avoid slipping. When navigating unfamiliar terrain, humans intuitively look beyond geometry, assessing not only the environment itself but also how specific physical capabilities interact with the surroundings [1]. Similarly, robots are expected to replicate such capability awareness by mastering traversability prediction.

Traversability estimation projects environmental features into navigation costs to determine the feasibility of traversing a specific terrain for a given robot morphology [2]. However, merging interaction experiences across multiple platforms introduces conflicting ground-truth labels since distinct robots possess unique physical capabilities [3]. For instance, trajectories illustrating a quadruped navigating stairs are fundamentally unsuitable for a wheeled robot. Such embodiment-specific contradictions prevent the creation of a shared traversability representation while severely limiting policy transferability across different robotics platforms [3], [4], [5].

Recent approaches relying solely on prior robot experience are notable, but performance often degrades outside specific training domains [3], [4], [6]. Conversely, as Vision-Language Models (VLMs) advance, zero-shot traversability works have emerged as a natural alternative [7], [8], [9]. Yet, while VLMs exhibit impressive semantic reasoning, leveraging them solely for broad environmental context fails to inherently encode platform-specific constraints. Consequently, current pipelines often separate perception from physical validation by adding a secondary filtering stage to remove unfeasible paths rather than embedding these constraints directly into the core representation [10].

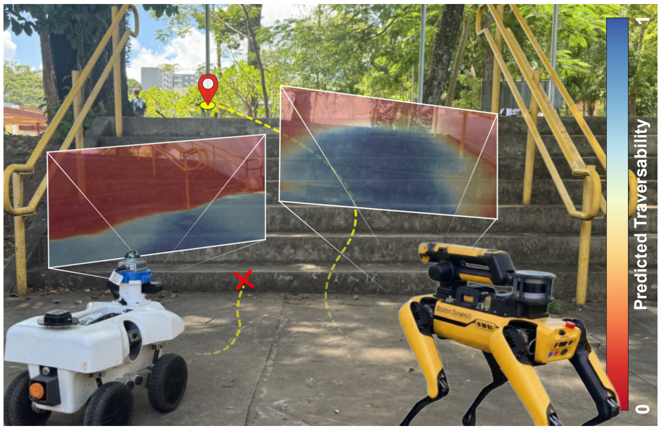

Fig. 1. **Visualization of Capability-Aware Traversability.** From a single observation of an unstructured outdoor scene, the model predicts a dense traversability map, with blue indicating higher traversability and red indicating lower traversability. Visually similar terrain can have different traversability outcomes depending on the robot's embodiment, motivating capability-conditioned traversability prediction rather than a single, robot-agnostic representation.

Even though robotics platforms operate under different embodiment profiles, the underlying relationship between semantic terrain classes and physical interaction admits a shared structure [11]. Rather than depending on late-stage trajectory filtering, perception models can internalize physical limits by actively modulating visual features based on robot-specific profiles, thereby integrating physical constraints directly into the spatial features to provide a unified foundation for multi-embodiment traversability estimation.

In this work, we propose **CAT** (Capability-Aware Traversability), a unified framework for predicting traversability directly within a capability-conditioned feature space. By using Spatially-Adaptive Denormalization (SPADE) blocks [12], CAT dynamically adapts visual embeddings to the robot's specific traversability profile, as illustrated in Fig. 1. Injecting categorical terrain maps into the network allows the model to ground semantic knowledge in spatial features.

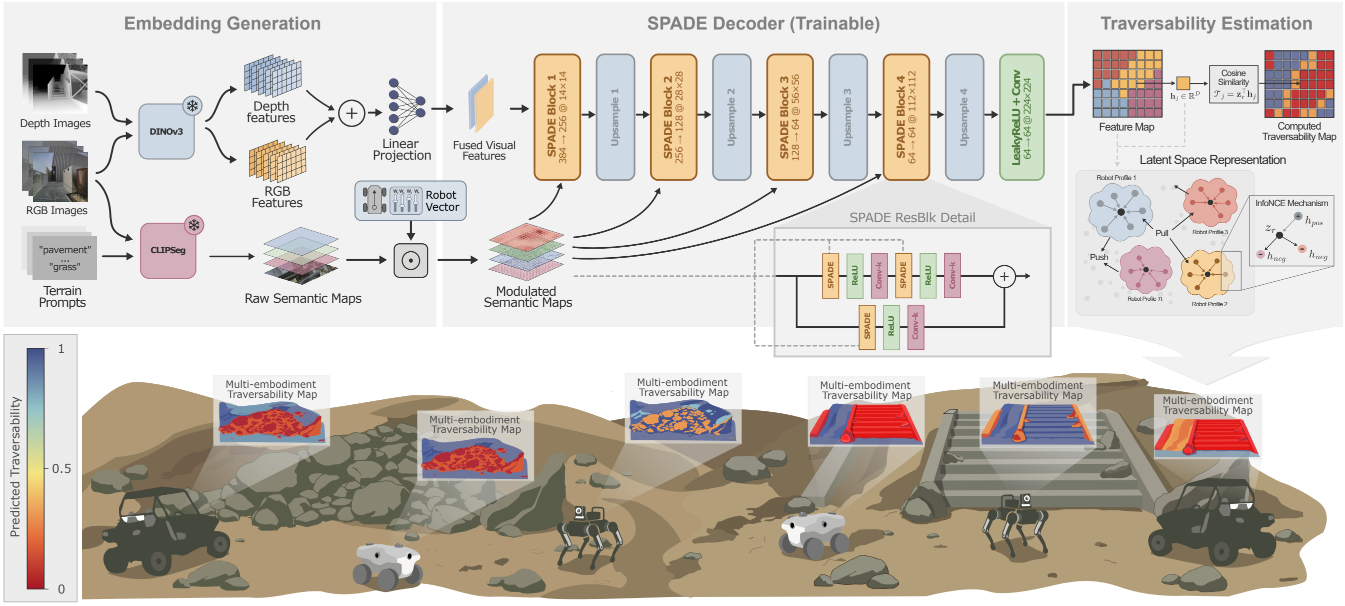

Fig. 2. **Overview of Capability-Aware Traversability**. In the Embedding Generation stage, DINOv3 [13] encodes RGB and depth images into fused visual features, while CLIPSeg [14] processes terrain prompts to produce raw semantic maps. To account for distinct physical constraints, a robot vector scales the raw semantic outputs to generate modulated semantic maps. The modulated semantic representation conditions a trainable SPADE decoder that upsamples and fuses semantics with visual features to produce a capability-conditioned spatial feature map. Traversability is computed via similarity to robot-specific prototypes in the learned latent space, resulting in embodiment-specific traversability maps from the same scene input.

II. RELATED WORK

a) Semantic-based Traversability: Incorporating semantic understanding into navigation stacks is a standard approach to reduce the gap between geometric sensing and environmental context [15]. Recent works process semantic maps either as direct inputs to provide contextual features to the perception module or as a distinct output layer to guide the downstream path planner [16], [17], [18]. While effective for unstructured outdoor navigation, treating semantics as a direct input forces the network to learn a static mapping from visual features to traversability costs. Rather than relying on semantics as a rigid input feature, our work uses semantic representations to condition the traversability estimation. This conditioning enables the model to dynamically weigh visual cues based on the scene's semantic context.

b) Self-Supervised Traversability Learning: Using the physical interaction of the robot with the environment has been widely applied to estimate traversability. To map safe regions, previous works combine information ranging from traction [3], [4], IMUs [6], RGB-D cameras [19], [20], and LiDARs [21], [22]. Although robots' experience provides a highly reliable ground truth, the resulting trajectory data creates an inherently sparse supervision signal. To extract more information from limited interaction data, self-supervised models correlate physically safe trajectories with dense visual representations, propagating traversability labels from the narrow path to visually similar regions across the scene [23], [24]. Prior methods are able to reconstruct well-defined structures that possess clear trails or sidewalks, but in truly unstructured environments, purely visual similarity often correlates poorly with actual traversability. Our approach uses an interactive zero-shot process to refine propagated labels for distinct robots navigating unstructured environments.

c) Conditioned-Aware Navigation: Recently, navigation frameworks have been increasingly leveraged by VLMs and vision-language-action models to achieve zero-shot generalization in novel environments [7], [8], [9], [10]. Such models typically work either as high-level scene interpreters or as end-to-end path generators [25]. Parallel to foundation models, weakly-supervised frameworks estimate relative traversability by relying solely on sparse pairwise human annotations [26]. However, both paradigms struggle to directly internalize the mobility realities of specific robotic embodiments. Using human judgment introduces a bias, as human visual discernment may not accurately reflect the embodiment limits of specific platforms. In contrast, integrating VLMs into real-time navigation remains a challenge. Problems such as inference latency and high computational costs are the most common issues that often prove impractical for time-critical tasks such as collision avoidance [27]. To address the challenge of incorporating robot-specific mobility characteristics, current architectures rely on human-ranked labels to assess terrain [26], or a separate filtering step to evaluate generated paths [10], [28]. In contrast, our approach tightly couples the semantic context with the perception layer, internalizing physical constraints directly within the model.

III. METHOD

CAT, illustrated in Fig. 2, is an embodiment-conditioned traversability estimation framework capable of perceiving semantic risks in unstructured outdoor environments. During inference, CAT predicts dense traversability scores directly from observed RGB-D images. To align the predicted traversability map with specific robot constraints, the architecture processes frozen visual embeddings alongside semantic terrain probabilities from CLIPSeg [14] using a SPADE [12] decoder. The following subsections detail the core inference pipeline, the capability-conditioning mechanism, and the interactive zero-shot annotation recipe required to train the network.

A. Generating Traversability Labels

To build a training set across diverse robotic navigation datasets, CAT uses the robot's past physical trajectories as the primary supervision signal. The ground-truth poses from the available state estimation are used to project the robot's trajectories into the image space of the aligned RGB-D camera, accounting for both the robot's footprint and the sensor mounting configuration. Let the 3D trajectory over a horizon H be defined as the sequence $\mathcal{P} = \{P_i\}_{i=1}^{H}$, where each point $P_i \in \mathbb{R}^3$ represents a position in the world frame at time step i. Assuming homogeneous coordinates, each 3D point projects to a corresponding 2D pixel coordinate $p_i \in \mathbb{R}^2$, as described in Eq. 1.

$$p_i = KTP_i \qquad (1)$$

where K is the camera intrinsic matrix, and T denotes the extrinsic transformation matrix from the world frame to the camera frame. By applying Eq. 1 to the entire sequence $\mathcal{P}$ and adding a depth filter to remove occluded points generates a sparse collection of positive trajectory pixels.

However, relying solely on sparse annotations limits the environment's spatial understanding. To densify the supervision, the trajectory labels are expanded into dense positive masks. Directly mapping off-the-shelf semantic segmentation to binary traversability labels is often insufficient, because a generic semantic class such as "grass" or "dirt" does not strictly correlate with safe passage, as physical traversability depends on the specific robot's capabilities and local geometric variations. Therefore, instead of treating semantic classes as rigid ground truth, the proposal pipeline generates labels through an interactive zero-shot process, as illustrated in Fig. 3.

Specifically, given an initial RGB image, GroundingDINO [29] extracts bounding boxes that define potential traversable areas using textual prompts. The system jointly leverages the physically driven trajectory by sampling positive point prompts from the robot's projected traversal area. Both the textual bounding boxes and the physical trajectory points condition Segment Anything Model 2 (SAM 2) [30], grounding the visual segmentation in the robot's true capabilities [31]. The output of the SAM 2 inference provides a dense traversability mask, which generates a large set of positive masks for training. Once the initial mask is generated, the model propagates the segmentation through video sequences. To ensure label integrity during propagation, the pipeline continuously monitors SAM 2 confidence scores and temporal consistency, measured by the intersection-over-union (IoU) between two consecutive masks. Whenever either metric drops below a predefined threshold, the current image is sent to a human for refinement.

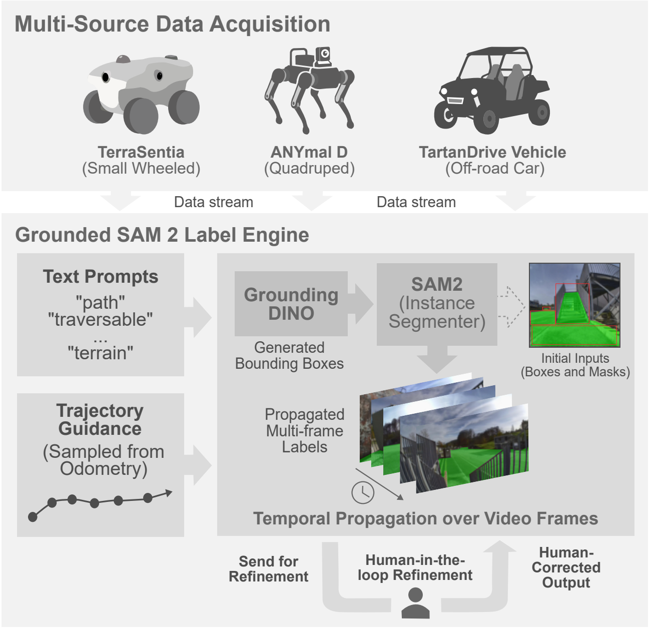

Fig. 3. **Overview of the trajectory-grounded label generation pipeline.** Data streams from the WayFASTER [4], GrandTour [32], and TartanDrive [33] datasets are processed through an interactive label engine, combining textual bounding boxes from GroundingDINO [29] with physical trajectory guidance to condition SAM 2 [30]. The pipeline temporally propagates the resulting dense masks across video frames and routes low-confidence segmentations to a human-in-the-loop.

B. Semantic Terrain Mapping and Traversability Costs

Robots can traverse different terrains depending on their embodiment, meaning the same scene may produce distinct traversability predictions for each platform. To account for this, we decompose terrain-dependent traversability into two complementary components: a spatial semantic map of the terrain types present and a robot-specific vector encoding how each terrain type relates to a given platform's capabilities.

For semantic representation, a frozen CLIPSeg [14] produces per-pixel terrain class probabilities. Given a set of K terrain prompts, CLIPSeg outputs logits for each prompt-image pair. Applying a softmax through the prompts at each pixel results in a semantic map $\mathbf{S} \in [0,1]^{H \times W \times K}$, where each channel encodes the pixel-wise probability of the corresponding terrain class.

To couple the semantic map with the capabilities of a specific platform, each robot r is characterized by a traversability vector, which quantifies how suitable each terrain class is for that specific platform. Instead of manually engineering these cost tables, we leverage a VLM to generate them automatically. We prompt the VLM with a set of uniformly sampled RGB images from the dataset to provide visual context, along with a natural language description of

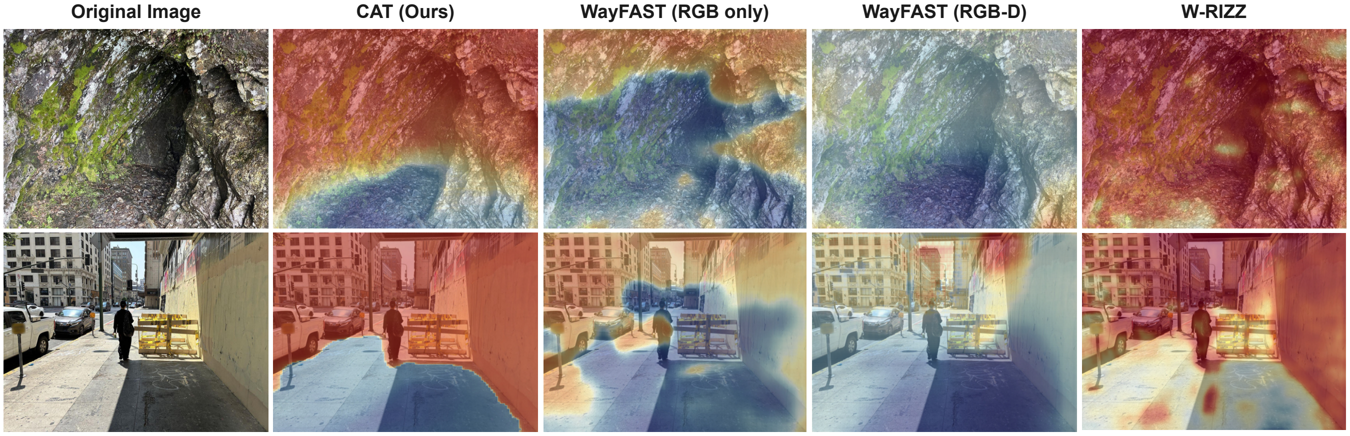

Fig. 4. **Qualitative comparison of traversability prediction in unstructured environments from NaviTrace Dataset [34].** While baselines like W-RIZZ and WayFAST struggle to differentiate between visually similar but physically distinct terrains, CAT successfully grounds the spatial features using the semantic terrain map, resulting in higher confidence bounds on traversable areas.

the robot's physical profile. Given this context, the model outputs a traversability score for each terrain-robot pair. This process is executed once per data–robot combination and requires no domain-specific supervision.

C. Capability-Conditioned Traversability

In this subsection, we present the module that generates robot-specific dense traversability maps from RGB and depth images. To achieve robot-specific outputs, the architecture merges frozen visual features with a robot-modulated semantic representation.

We employ DINOv3 [13] as the frozen feature encoder f, leveraging its backbone's capacity to capture rich patch-level semantics. The encoder f maps the RGB image $\mathbf{x} \in \mathbb{R}^{H\times W\times 3}$ into a dense spatial feature map $\mathbf{F} \in \mathbb{R}^{h\times w\times d}$. To incorporate 3D geometric information, we replicate the single-channel depth image across three channels to match the expected input dimensions of the frozen backbone. Then, a separate inference step through f extracts depth features $\mathbf{F}_{\text{depth}} \in \mathbb{R}^{h\times w\times d}$. To fuse the modalities, we first concatenate the two feature maps along the channel dimension to form a unified visual representation, as defined by Eq. 2.

$$\mathbf{F}_{\text{vis}} = \text{Concat}(\mathbf{F}, \mathbf{F}_{\text{depth}}) \in \mathbb{R}^{h\times w\times 2d} \tag{2}$$

As shown in Eq. 3, a learnable linear layer then projects the combined feature representation down to a fused feature dimension of d'.

$$\mathbf{F}_{\text{vis}} = \text{Linear}(\mathbf{F}_{\text{vis}}) \in \mathbb{R}^{h\times w\times d'} \tag{3}$$

Finally, the resulting fused feature map $\mathbf{F}_{\text{vis}}$ serves as the initial input to the decoder for generating traversability maps.

Simultaneously, the network modulates the semantic map using the robot's traversability vector via element-wise multiplication. The resulting operation suppresses semantic channels corresponding to impassable terrains for the robot while preserving traversable channels.

The visual features and the modulated map $\hat{\mathbf{S}}_r$ then converge in a progressive SPADE [12] decoder, where the modulated map acts as a spatially-varying semantic condition. The decoder consists of four SPADE residual blocks upsampling the feature map from 14×14 to 224×224. Inside each block, SPADE replaces standard batch normalization for the intermediate feature map h, as defined by Eq. 4.

$$\text{SPADE}(\mathbf{h}, \hat{\mathbf{S}}_r) = \gamma(\hat{\mathbf{S}}_r) \odot \text{BN}(\mathbf{h}) + \beta(\hat{\mathbf{S}}_r) \tag{4}$$

where γ and β are learned convolutional functions of $\hat{\mathbf{S}}_r$, resized to match the current feature resolution. The decoder produces a D-dimensional feature map, L_2-normalized along the channel dimension. Because different robots generate distinct maps, the SPADE layers produce distinct modulations and, consequently, distinct output features from the same visual input.

To structure the latent space produced by the decoder, during training, the architecture relies on a contrastive learning approach, a technique already proven highly effective for terrain traversability analysis [23], [35]. For each robot profile, the network maintains a unit-norm vector traversability prototype $\mathbf{z}_r \in \mathbb{R}^D$ that is updated via an exponential moving average. At each spatial location, the network evaluates traversability by computing the cosine similarity between the local feature and the robot-specific prototype, defined as $\mathcal{T}_j = \mathbf{z}_r^\top \mathbf{h}_j$, where the variable $\mathbf{h}_j \in \mathbb{R}^D$ denotes the feature vector from the decoder at pixel index j. We use this similarity both to produce the traversability score map and to guide the training objective.

The network minimizes a combined loss, defined in Eq. 5, consisting of a trajectory-based term and a mask-based contrastive term.

$$\mathcal{L} = (1-\omega)\,\mathcal{L}_{\text{traj}} + \omega\,\mathcal{L}_{\text{mask}} \tag{5}$$

Both $\mathcal{L}_{\text{traj}}$ and $\mathcal{L}_{\text{mask}}$ operate as pixel-level InfoNCE losses. The trajectory term samples positive features exclusively from the robot path pixels, whereas the mask term samples from the dense, trajectory-grounded mask pixels. The InfoNCE objective pulls positive spatial features $\mathbf{h}_j$ closer to the prototype $\mathbf{z}_r$ while pushing negative spatial features

TABLE I
CAT TRAINING HYPERPARAMETERS

Parameter	**Value**	**Parameter**	**Value**
Encoder Dim d	768	Projected Dim d'	384
Optimizer	AdamW	Batch Size	128
Learning Rate	$3{\times}10^{-4}$	Weight Decay	10^{-2}
Feature Dim D	64	InfoNCE τ	0.07
Loss Weight ω	0.05	EMA Momentum	0.99
Terrain Classes K	10	Trav. Threshold	0.4

away. To maintain label consistency with the specific physical constraints of the target robot, the pipeline dynamically refines the positive mask. If the expected traversability of a pixel falls below a predefined threshold, the system reclassifies the specific pixel as a negative example.

IV. EXPERIMENTAL RESULTS

We evaluate our method across two distinct data distributions to assess both human-alignment and physical trajectory-alignment. To isolate the spatial traversability map's contributions independently of a downstream path planner, we construct a positive ground-truth mask in both datasets by applying a safety buffer around the target path.

First, we establish performance using NaviTrace [34], a high-quality, human-annotated dataset. By masking the human-provided paths, we quantify how well the high-confidence regions of our predicted maps correlate with human-validated safe zones. Second, to overcome the limitations of a potentially human-biased dataset, we evaluate on a held-out 20% split of our self-supervised training data. In this uncontaminated test split, we apply the safety buffer to raw physical trajectory masks. This measures how effectively the model recognizes actually executed trajectories as feasible.

To contextualize our performance, we benchmark CAT against three state-of-the-art methods: WayFAST (RGB and RGB-D) [3] and W-RIZZ [26].

A. Training Details

CAT's trainable component is exclusively the SPADE decoder, which has ~9M parameters. Both DINOv3 and CLIPSeg backbones remain fully frozen, with embeddings pre-computed to disk before training. The decoder is conditioned on modulated semantic maps from $K{=}10$ terrain prompts, producing an L_2-normalized feature map of dimension $D{=}64$. Traversability scores are cosine similarities in $[-1, 1]$, normalized to $[0, 1]$ for visualization, where blue denotes traversable regions and red denotes obstacles. We jointly train the models for four robot profiles (wheeled, legged, differential, and ATV). The training process optimizes an InfoNCE loss averaged across all profiles per batch by sampling 256 positive and 1024 negative pixels for the trajectory mask, along with 512 positive and 1024 negative pixels for the generated mask.

All traversability score values were generated locally with Qwen3-VL using samples from the dataset to provide more useful information during training. During inference, we use fixed scores. The model was trained for 35 epochs on a single NVIDIA L40S GPU (~8 h), using a 70/30 train/validation split. Full hyperparameters are reported in Table I.

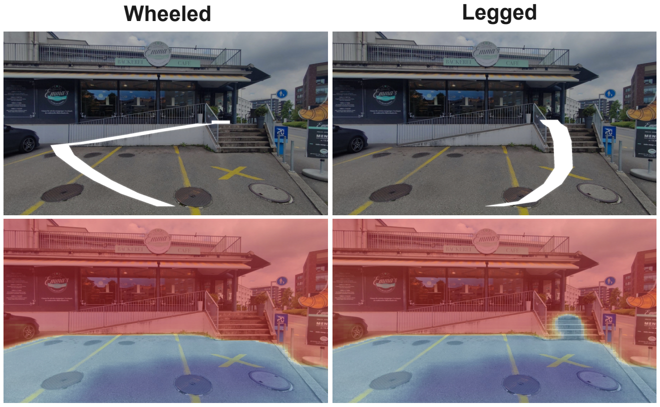

Fig. 5. **Capability-Conditioned Modulation.** From a single RGB-D input, CAT generates embodiment-specific traversability maps. The Wheeled profile (left) restricts navigation to flat geometry, whereas the Legged profile (right) permits traversal over stairs and rough terrain. NaviTrace [34] human annotations are overlaid as ground truth.

B. Traversability Prediction

We evaluate predictive performance within the safety-buffered positive masks using two key metrics. First, *Mean Traversability* quantifies the model's average prediction confidence inside the safety-buffer regions. Second, we compute the Area Under the Curve (AUC) to assess the network's fundamental ability to rank safe terrain strictly higher than hazards.

As shown in Table II, CAT outperforms all baselines across both metrics. By utilizing the modulated semantic map to actively ground visual features within the SPADE decoder, CAT achieves a 12.6% improvement in mean traversability over WayFAST (RGB-D) [3] on NaviTrace. This performance delta holds in our unannotated test split, demonstrating generalization to raw physical interaction data. In terms of discriminative power, CAT achieves a 13.5% AUC improvement over W-RIZZ [26]. Qualitatively, Fig. 4 shows that explicitly conditioning visual features on semantic terrain data resolves ambiguity in highly unstructured environments, yielding clearer traversable corridors than baseline models.

C. Capability-Conditioned Modulation

To evaluate embodiment awareness, we filtered the NaviTrace dataset to include scenes with distinct, non-overlapping feasible paths for both *Legged* and *Wheeled* platforms.

Because the legged robot intrinsically possesses superior mobility on rough terrain, the wheeled platform naturally underperforms when evaluated on legged-specific paths, as detailed in Table III. Given that human annotations inherently point toward safe paths, the nearly 0.1 performance gap between the profiles on these terrains is highly significant, confirming our network's adaptive capability shown in Fig. 5.

TABLE II

QUANTITATIVE COMPARISON OF TRAVERSABILITY PREDICTION ON NAVITRACE AND TRAJECTORY TEST SPLIT

Method	NaviTrace (Human-Aligned)		Ours Test Split (Trajectory-Aligned)	
	Mean Traversability ↑	AUC ↑	Mean Traversability ↑	AUC ↑
WayFAST (RGB) [3]	0.451	0.417	0.461	0.283
WayFAST (RGB-D) [3]	0.572	0.374	0.500	0.290
W-RIZZ (RGB) [26]	0.561	0.465	0.387	0.335
CAT (Ours)	**0.644**	**0.528**	**0.501**	**0.348**

TABLE III

MODULATION PERFORMANCE ON EMBODIMENT-SPECIFIC REGIONS

Embodiment	Wheeled Path Mean ↑	Legged Path Mean ↑
Wheeled Profile	0.708	0.639
Legged Profile	0.728	0.730

TABLE IV

REAL-WORLD NAVIGATION SUCCESS RATE

Scenario	Robot (Profile)	Success Rate
Forested Area	Spot (Legged)	10/10
Staircase Setting	TerraSentia (Wheeled)	7/10

D. Real-World Performance

Each experiment was repeated 10 times to account for environmental and system variability. We deployed CAT on Boston Dynamics' Spot (legged, main RGB-D camera) and the TerraSentia skid-steer (wheeled, ZED 2i). Operating on a Jetson Orin Nano, CAT achieved a stable inference rate of 5 Hz. Traversability predictions were coupled with a greedy pure-pursuit local planner directed toward the furthest traversable point in view.

Figure 6 highlights two primary test cases of embodiment-aware obstacle avoidance. In the left scenario, the Spot robot accurately segmented a tree as non-traversable, generating a safe bypass trajectory through a forested area. In the right scenario, CAT correctly applied wheeled geometric constraints to mask the abrupt elevation drop of a staircase as a severe hazard, successfully forcing the TerraSentia to calculate an evasive route rather than attempting an unfeasible traversal. Table IV reports the quantitative results for both deployments. The recorded wheeled failures were due to the greedy behavior of the pure-pursuit planner rather than perceptual errors, where the planner aggressively targeted a distant traversable waypoint, causing the robot to cut through a correctly masked hazard zone to reach the distant goal.

E. Discussion and Limitations

While CAT demonstrates improvements in multi-embodiment navigation, the system also achieves zero-shot identification of unrepresented dynamic obstacles like humans, as shown in the left panel of Fig. 7. However, CAT's performance remains tightly coupled to the underlying semantic segmentation map. Although empirically rare, if the segmentation fails to accurately categorize the terrain, such as misclassifying a heterogeneous ground, the network inherits the resulting perceptual error detailed in the right panel of Fig. 7. Because the capability vector directly scales the semantic logits, a fundamental failure in the semantic map causes the SPADE decoder to improperly modulate geometric embeddings, resulting in lost resolution or the grouping of unrelated categories. Future work will explore tightly coupling the segmentation and traversability layers in an end-to-end differentiable manner to recover from upstream semantic noise.

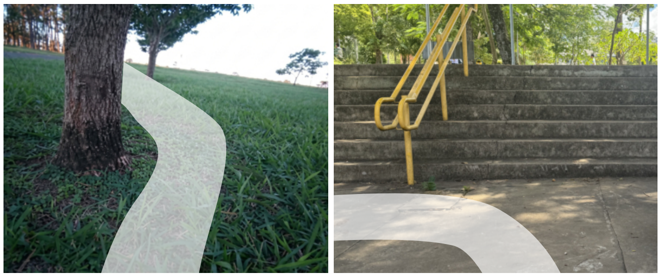

Fig. 6. **Real-world embodiment-aware navigation using CAT on heterogeneous platforms.** Spot (left), representing the legged profile in a green area, uses CAT to predict a dense traversability map from onboard RGB-D. The network correctly marks the tree trunk as non-traversable, producing a safe bypass trajectory toward the goal. TerraSentia (right), representing the wheeled profile facing a staircase, applies wheeled mobility constraints via CAT to assign low traversability to the abrupt elevation change. The assignment prevents an unsafe attempt to traverse the stairs and forces an evasive route.

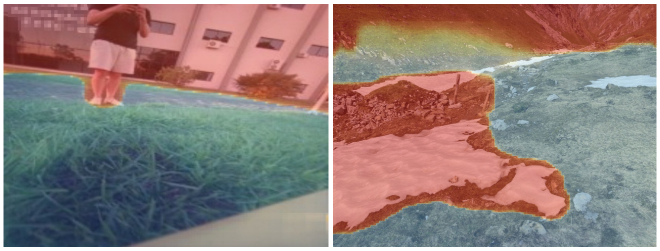

Fig. 7. **CAT Findings and Limitations.** During real-world deployments, CAT consistently identifies correctly unlabeled forms (left), as human beings, presenting zero-shot capabilities for all experiments. On the other hand, in scenarios (right) where the segmentation map incorrectly groups a terrain feature, the final predictions lose resolution or group unrelated categories in the same traversability map.

V. CONCLUSION

We proposed CAT, a framework that reformulates traversability estimation as a capability-conditioned representation learning problem. By modulating visual features through robot-specific traversability vectors and semantically adaptive SPADE blocks, CAT embeds physical constraints directly into the spatial feature space, removing the need for late-stage trajectory filtering. Our results demonstrate that CAT outperforms existing self-supervised and weakly-supervised baselines across both human-aligned and trajectory-aligned evaluation protocols, while exhibiting robust generalization to unseen scene elements at deployment time. Our work establishes embodiment-aware feature conditioning as a scalable path toward transferable traversability representations for unstructured field robotics, supporting future research on this topic.

References

[1] M. Kim and C. F. Doeller, "Cognitive Maps for a Non-Euclidean Environment: Path Integration and Spatial Memory on a Sphere," *Psychological Science*, vol. 35, pp. 1217 – 1230, 2024.

[2] J. J. Gibson, *The Ecological Approach to Visual Perception*. Houghton Mifflin, 1979.

[3] M. V. Gasparino, A. N. Sivakumar, Y. Liu, A. E. Velasquez, V. A. Higuti, J. Rogers, H. Tran, and G. Chowdhary, "WayFAST: Navigation with Predictive Traversability in the Field," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10 651–10 658, 2022.

[4] M. V. Gasparino, A. N. Sivakumar, and G. Chowdhary, "WayFASTER: a Self-Supervised Traversability Prediction for Increased Navigation Awareness," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 8486–8492.

[5] M. Eder and G. Steinbauer-Wagner, "Robot-Dependent Traversability Estimation for Outdoor Environments using Deep Multimodal Variational Autoencoders," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 12 635–12 642.

[6] L. Wellhausen, A. Dosovitskiy, R. Ranftl, K. Walas, C. Cadena, and M. Hutter, "Where Should I Walk? Predicting Terrain Properties From Images Via Self-Supervised Learning," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1509–1516, 2019.

[7] V. S. Dorbala, G. A. Sigurdsson, J. Thomason, R. Piramuthu, and G. S. Sukhatme, "CLIP-nav: Using CLIP for Zero-Shot Vision-and-Language Navigation," in *Workshop on Language and Robotics at CoRL*, 2022.

[8] K. Weerakoon, M. Elnoor, G. Seneviratne, V. Rajagopal, S. H. Arul, J. Liang, M. K. M. Jaffar, and D. Manocha, "BehAV: Behavioral Rule Guided Autonomy Using VLMs for Robot Navigation in Outdoor Scenes," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 7044–7051.

[9] S. Gummadi, M. V. Gasparino, G. Capezzuto, M. Becker, and G. Chowdhary, "ZeST: an LLM-based Zero-Shot Traversability Navigation for Unknown Environments," *arXiv preprint arXiv:2508.19131*, 2025.

[10] M. G. Castro, S. Rajagopal, D. Gorbatov, M. Schmittle, R. Baijal, O. Zhang, R. Scalise, S. Talia, E. Romig, C. de Melo, B. Boots, and A. Gupta, "VAMOS: A Hierarchical Vision-Language-Action Model for Capability-Modulated and Steerable Navigation," *Under Review*, 2025.

[11] M. J. Miles, H. Biggie, and C. Heckman, "Terrain-aware semantic mapping for cooperative subterranean exploration," *Frontiers in Robotics and AI*, vol. 10, p. 1249586, 2023.

[12] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic Image Synthesis with Spatially-Adaptive Normalization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[13] O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa, F. Massa, D. Haziza, L. Wehrstedt, J. Wang, T. Darcet, T. Moutakanni, L. Sentana, C. Roberts, A. Vedaldi, J. Tolan, J. Brandt, C. Couprie, J. Mairal, H. Jégou, P. Labatut, and P. Bojanowski, "DINOv3," 2025.

[14] T. Lüddecke and A. Ecker, "Image Segmentation Using Text and Image Prompts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022.

[15] A. Shaban, X. Meng, J. Lee, B. Boots, and D. Fox, "Semantic Terrain Classification for Off-Road Autonomous Driving," in *Proceedings of the 5th Conference on Robot Learning*, A. Faust, D. Hsu, and G. Neumann, Eds. PMLR, 2022.

[16] P. Roth, J. Nubert, F. Yang, M. Mittal, and M. Hutter, "ViPlanner: Visual Semantic Imperative Learning for Local Navigation," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 5243–5249.

[17] 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*, vol. 40, pp. 3756–3777, 2024.

[18] S. Ægidius, D. Hadjivelichkov, J. Jiao, J. Embley-Riches, and D. Kanoulas, "Watch Your STEPP: Semantic Traversability Estimation using Pose Projected Features," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

[19] M. Mattamala, J. Frey, P. Libera, N. Chebrolu, G. Martius, C. Cadena, M. Hutter, and M. Fallon, "Wild Visual Navigation: Fast Traversability Learning via Pre-Trained Models and Online Self-Supervision," *Autonomous Robots*, vol. 49, no. 3, p. 19.

[20] G. Kahn, P. Abbeel, and S. Levine, "BADGR: An Autonomous Self-Supervised Learning-Based Navigation System," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 1312–1319, 2021.

[21] F. Affonso, F. A. G. Tommaselli, G. Capezzuto, M. V. Gasparino, G. Chowdhary, and M. Becker, "CROW: A Self-Supervised Crop Row Navigation Algorithm for Agricultural Fields," *Journal of Intelligent & Robotic Systems*, vol. 111, no. 1, p. 28, 2025.

[22] J. Seo, T. Kim, K. Kwak, J. Min, and I. Shim, "ScaTE: A Scalable Framework for Self-Supervised Traversability Estimation in Unstructured Environments," *IEEE Robotics and Automation Letters*, 2023.

[23] S. Jung, J. Lee, X. Meng, B. Boots, and A. Lambert, "V-STRONG: Visual Self-Supervised Traversability Learning for Off-road Navigation," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 1766–1773.

[24] Y. Kim, J. Lee, C. Lee, J. Mun, D. Youm, J. Park, and J. Hwangbo, "Learning Semantic Traversability With Egocentric Video and Automated Annotation strategy," *IEEE Robotics and Automation Letters*, vol. 9, no. 11, pp. 10 423–10 430, 2024.

[25] A.-C. Cheng, Y. Ji, Z. Yang, X. Zou, J. Kautz, E. Biyik, H. Yin, S. Liu, and X. Wang, "NaVILA: Legged Robot Vision-Language-Action Model for Navigation," in *RSS*, 2025.

[26] A. Schreiber, A. N. Sivakumar, P. Du, M. V. Gasparino, G. Chowdhary, and K. Driggs-Campbell, "W-RIZZ: A Weakly-Supervised Framework for Relative Traversability Estimation in Mobile Robotics," *IEEE Robotics and Automation Letters*, vol. 9, no. 6, pp. 5623–5630, 2024.

[27] J. Park, P. Kim, and D. Ko, "Real-time open-vocabulary perception for mobile robots on edge devices: a systematic analysis of the accuracy-latency trade-off," *Frontiers in Robotics and AI*, 2025.

[28] K. Lee and K. Lee, "Terrain-aware path planning via semantic segmentation and uncertainty rejection filter with adversarial noise for mobile robots," *Journal of Field Robotics*, vol. 42, no. 1, pp. 287–301, 2025.

[29] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su *et al.*, "Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection," in *European Conference on Computer Vision*. Springer, 2024, pp. 38–55.

[30] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer, "SAM 2: Segment Anything in Images and Videos," *arXiv preprint arXiv:2408.00714*, 2024.

[31] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan, Z. Zeng, H. Zhang, F. Li, J. Yang, H. Li, Q. Jiang, and L. Zhang, "Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks," 2024.

[32] J. Frey, T. Tuna, F. Fu, K. Patterson, T. Xu, M. Fallon, C. Cadena, and M. Hutter, "GrandTour: A Legged Robotics Dataset in the Wild for Multi-Modal Perception and State Estimation," 2026.

[33] M. Sivaprakasam, P. Maheshwari, M. G. Castro, S. Triest, M. Nye, S. Willits, A. Saba, W. Wang, and S. Scherer, "TartanDrive 2.0: More Modalities and Better Infrastructure to Further Self-Supervised Learning Research in Off-Road Driving Tasks," 2024.

[34] T. Windecker, M. Patel, M. Reuss, R. Schwarzkopf, C. Cadena, R. Lioutikov, M. Hutter, and J. Frey, "NaviTrace: Evaluating Embodied Navigation of Vision-Language Models," *Preprint submitted to arXiv*, October 2025.

[35] J. Seo, S. Sim, and I. Shim, "Learning Off-Road Terrain Traversability with Self-Supervisions Only," *IEEE Robotics and Automation Letters*, vol. 8, no. 8, pp. 4617–4624, 2023.