Learning Humanoid Navigation from Human Data

Anonymous Submission

***Abstract*—We present EgoNav, a system that enables a humanoid robot to traverse diverse, unseen environments by learning entirely from 5 hours of human walking data, with no robot data or finetuning. A diffusion model predicts distributions of plausible future trajectories conditioned on past trajectory, a 360° visual memory fusing color, depth, and semantics, and video features from a frozen DINOv3 backbone that capture appearance cues invisible to depth sensors. A hybrid sampling scheme achieves real-time inference in 10 denoising steps, and a receding-horizon controller selects paths from the predicted distribution. We validate EgoNav through offline evaluations, where it outperforms baselines in collision avoidance and multi-modal coverage, and through zero-shot deployment on a Unitree G1 humanoid across unseen indoor and outdoor environments. Behaviors such as waiting for doors to open, navigating around pedestrians, and avoiding glass walls emerge naturally from the learned prior. We release the dataset and trained models. Our website: https://egonav-submit.github.io/project-page/**

I. Introduction

How can a humanoid robot learn to navigate diverse, unseen environments without explicitly building a map? Previous work on learning-based robot navigation collected data directly on robot platforms, which is costly and difficult to scale. Cross-embodiment approaches such as NoMaD [1] train navigation policies on hundreds of hours of robot driving data and demonstrate transfer across wheeled and legged platforms, but each new domain still requires in-domain training data, and large-scale humanoid navigation datasets do not yet exist.

Vision-language models such as NaVILA [2] sidestep robot data by using a VLM to translate language instructions into locomotion commands. However, they require step-by-step human instructions, emit discrete unimodal commands without spatial grounding, and are too slow for closed-loop control.

In manipulation, a promising alternative has emerged: learning from human demonstrations. Systems such as UMI [3] and DexCap [4] train policies entirely on human data and transfer them to robots without any robot-specific data collection. Navigation is a natural next frontier for this paradigm. Human walking data is cheap and scalable: a single person with body-mounted cameras can gather hours of diverse navigation data without any robot hardware. Human walking data also encodes rich commonsense about navigation: which paths are physically plausible, how to avoid obstacles, and where alternative routes exist. Despite this promise, recent attempts to learn navigation from human data [5] still struggle to work without robot data and show limited generalization.

What prevents this paradigm from working in real world? We identify some key challenges that must be addressed:

- **Insufficient scene coverage:** Scene coverage comes two fold, one is the field of view (FOV) of the camera, the other is the richness of the scene description. Most trajectory prediction approaches assume ground truth scenes [6]–[8] or predict from past trajectories alone [9]–[11]. The egocentric perspective is more practical, but a stereo depth camera covers only ∼90° FOV, is blind to transparent or reflective surfaces such as glass walls, and carries no semantic understanding of the scene.
- **Unimodal predictions:** Human motion is inherently multi-modal: at any decision point, multiple future trajectories are plausible. Yet most methods produce single-trajectory estimates [12]–[15], providing incomplete information for downstream path selection. A useful navigation prior must capture the diverse distribution of plausible futures, not a single guess.
- **Interface representation:** Some prior work uses end-to-end methods that couple terrain traversal and locomotion, making it hard to generalize in real world. Others predict hand and eye trajectories in robot frame feeding to a decoupled locomotion policy, but despite careful engi-

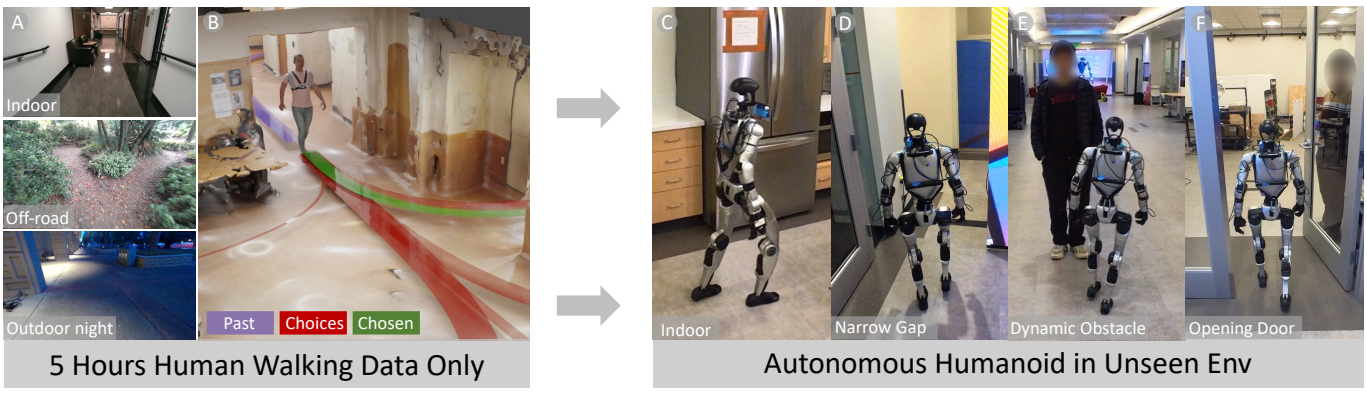

Fig. 1. **EgoNav** learns a navigation prior from human walking data: given past trajectory (purple) and a 360° visual context of the scene, a diffusion model generates a distribution of plausible future paths (red ribbons), with ribbon width indicating likelihood. The learned prior transfers directly to a Unitree G1 humanoid with zero robot data.

neering to bridge the embodiment gap, find that human data alone is insufficient and must be supplemented with robot demonstrations. A successful human-to-robot transfer requires an embodiment-agnostic interface between navigation and locomotion, one that conveys not only *where* to go but also *what the path looks like* semantically, so that downstream controllers can adapt their behavior to the terrain.

The key is to learn an embodiment-agnostic *navigation prior*: a distribution over plausible future trajectories in world coordinates conditioned on the surrounding scene. In this paper, we present EgoNav, a system that addresses these challenges and learns a generalizable navigation prior from human walking data for humanoid deployment (Fig. 1):

- We construct a 360° panoramic **visual memory** (VM) fusing color, depth, and semantic channels from a rolling buffer of egocentric observations, augmented with video features from a frozen DINOv3 backbone [16] that capture appearance-level cues invisible to depth sensors, such as glass walls and dynamic agents.
- We train a **conditional diffusion model** on the VM and video features that generates diverse future trajectory samples, naturally capturing the multi-modal distribution of plausible paths. To overcome the latency barrier of iterative diffusion sampling for real-time deployment [8], [17], we introduce a hybrid DDIM–DDPM sampling scheme that achieves near-full quality in only 10 steps.
- We build a complete **pipeline from human data to humanoid deployment** on a Unitree G1: a receding-horizon controller selects from the predicted distribution with latency compensation and mode-consistency.

We emphasize that EgoNav learns a *navigation prior*: a scene-informed distribution of plausible paths from which a downstream controller selects and executes. High-level reasoning (VLMs) and low-level locomotion are advancing rapidly, but the middle layer, knowing *where one can walk*, remains missing. A robust, generalizable navigation prior fills this gap, onto which goals or task planning can be layered. The resulting navigation prior, trained entirely on 5 hours of human walking data with no robot data or finetuning, transfers directly to autonomous humanoid navigation in unseen environments. We validate EgoNav in offline evaluations, demonstrating superior collision avoidance and multi-modal trajectory coverage over baselines. Through real-world deployment on the G1 humanoid in previously unseen environments, we show that behaviors such as waiting for doors to open, navigating around pedestrians, and avoiding glass walls emerge from the learned prior without explicit programming. The dataset and trained models will be publicly released.

II. Related Work

Our work learns humanoid navigation from human walking data, drawing on advances in scene understanding, multi-modal generative modeling, and human-to-robot transfer for deployment. We review related work across these areas.

Trajectory prediction for autonomous systems. In autonomous driving, trajectory prediction is essential for safe

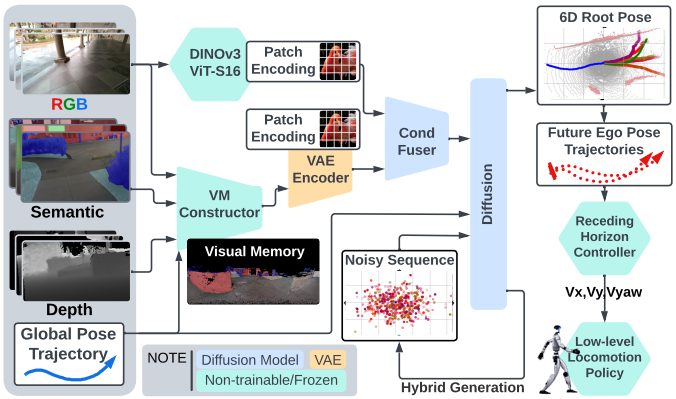

Fig. 2. **Overview of the proposed method:** A rolling buffer of 32 segmented RGB frames and cleaned depth frames are combined to construct a single visual memory (VM). The VM is encoded into a 64-dimensional embedding, concatenated with 6D pose as input to the diffusion model. Then future trajectory along with expected VM is denoised. All input and output of the prediction module are in the egocentric frame.

planning [8]. Early recurrent approaches [18] have given way to transformer-based models [19], [20] that better handle diverse sensor streams and long-range dependencies. AV methods commonly represent the environment in bird's-eye-view (BEV) or occupancy grids [21]–[23], which simplify prediction under structured traffic rules but struggle in unstructured settings [24]. For pedestrian trajectory prediction, classical approaches such as Social Force [12], Social GAN [25], and Social LSTM [26] model interactions between agents but largely ignore environmental context, predicting solely from past trajectories [9]–[11]. Prominent datasets such as UCY/ETH [27] and the Stanford Drone Dataset [28] provide only BEV observations, reducing the problem to two dimensions. While effective offline, the lack of scene awareness and reliance on BEV limit real-world deployment.

Egocentric scene representations. Recent work has shifted toward richer scene representations for trajectory prediction. While AV methods [6], [7], [29], [30] rely on synthesized BEV, walking scenarios demand more flexible egocentric representations. Early approaches use first-person RGB(D) directly [31], [32], but single frames provide limited spatial coverage and lose context outside the camera frustum. Wang et al. [13] address this with a depth panorama built from a rolling buffer of egocentric observations, extending the effective field of view to 360°, though without color or semantic information. More recently, vision foundation models such as DINOv3 [16] have emerged as powerful feature extractors, capturing appearance and semantic cues that complement geometric depth. Our visual memory builds on the panoramic paradigm by fusing color, depth, and semantic channels into a single compact representation, augmented with DINOv3 features.

Diffusion models for trajectory prediction. Denoising diffusion probabilistic models (DDPM) [33] and their accelerated variant DDIM [34] have become a standard framework for multi-modal generation. In the trajectory domain, TRACE [8] applies guided diffusion to generate diverse future human motions in 3D scenes, demonstrating clear advantages over deterministic predictions in capturing multi-modal distributions, but operates offline without real-time constraints. Jia et

al. [17] use DDPM to predict egocentric upper-body motion including head pose and gaze, though their focus is on visuomotor coordination rather than locomotion trajectories. A common limitation of diffusion-based prediction methods is inference speed: the iterative denoising process is typically too slow for real-time robotic applications. Our method addresses this through a hybrid DDIM–DDPM sampling scheme that achieves real-time performance with minimal degradation in as few as 10 steps, while using classifier-free guidance to condition generation on scene context and video features.

Egocentric trajectory prediction. A growing body of work addresses motion prediction from the egocentric perspective. LookOut [14] lifts egocentric DINOv2 features into a 3D voxel grid, collapses it to a bird's-eye view, and regresses future 6D head poses, but produces only deterministic single-trajectory estimates. EgoCogNav [15] jointly predicts egocentric body-frame trajectories and perceived path uncertainty using DINOv2 features fused with gaze and motion cues, introducing a cognition-aware dimension; however, it also produces single-trajectory estimates and does not deploy on a robot. HEAD [5] learns deterministic humanoid goal-conditioned navigation on a Unitree G1 robot by predicting hand and eye trajectories. Despite careful engineering (image undistortion, homography alignment, temporal subsampling) to address the embodiment gap, the authors find that human data alone is insufficient and must be supplemented with robot demonstrations, showing limited generalization within room scale. The primary issue being hand and eye trajectories remain deeply coupled to the robot embodiment.

Robot navigation and human-to-robot transfer. In manipulation, recent works such as UMI [3] and DexCap [4] have demonstrated that policies trained on human demonstration data can transfer directly to robots without robot-specific data collection. In navigation, NoMaD [1] trains a diffusion-based policy on over 100 hours of pure robot driving data and demonstrates cross-platform transfer, but outputs only short-horizon 2D waypoints from a single forward-facing camera with no semantic scene understanding. More fundamentally, it still requires extensive robot data collection. The paradigm of learning navigation from human data remains largely unexplored. VP-Nav [35] develops a point-goal navigation system for quadrupeds that couples vision with proprioceptive feedback, enabling the robot to detect obstacles invisible to depth sensors such as glass walls. Its classical planning pipeline (occupancy map + FMM) is environment-agnostic but purely geometric, with no semantic scene understanding. ANYmal Parkour [36] and Locomotion Beyond Feet [37] demonstrate agile legged locomotion through libraries of terrain-specific skills (jumping, climbing, crouching), but only generalize to novel arrangements of known obstacle categories. Concurrent advances in general humanoid locomotion [38], [39] have made real-world deployment increasingly feasible, but these controllers solve *how* to walk without addressing *where* to walk. Our system fills this gap: trained entirely on human walking data with no robot data or finetuning, it provides a generalizable navigation prior that transfers directly to a humanoid robot.

Our work differs from the above in a fundamental way: we require *no robot data and no finetuning*. While NoMaD depends on large-scale purely robot-collected datasets and HEAD must supplement human data with robot demonstrations, our model predicts embodiment-agnostic, 6d trajectory distributions, trained entirely on human walking demonstrations, and transfers directly to a Unitree G1 humanoid. Beyond this data paradigm shift, we provide (1) a 360° panoramic visual memory augmented with DINOv3 video features for richer scene understanding than forward-facing cameras alone, (2) long-horizon multi-modal trajectory distributions (5 s, 20Hz, 6DoF) rather than short-horizon action predictions, and (3) real-world humanoid deployment validated across multiple unseen environments.

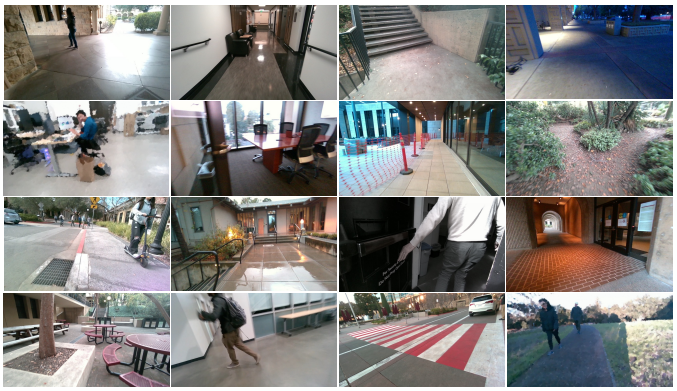

Fig. 3. **Dataset:** The dataset has a mix of weather, road, lighting, and traffic.

III. Method

EgoNav learns a navigation prior entirely from human walking data and deploys it zero-shot on a humanoid robot. The system is designed around four principles:

- **Human-native.** The system is capable of learning entirely from human walking demonstrations, requiring no robot data, simulation, or task-specific engineering. Human data is cheap, scalable, and encodes rich commonsense about navigation.
- **Scene-aware.** The system derives sufficient scene understanding from egocentric observations alone, without requiring pre-built maps, prior environment models, or any external localization.
- **Distributional.** The system models the full distribution of plausible future paths, providing downstream components with diverse candidates for informed selection, rather than committing to a single trajectory estimate.
- **Robot-ready.** The learned prior is deployable on physical robots, not merely performant in offline evaluation. This requires addressing real-time inference speed, system latency compensation, closed-loop mode consistency, and robust obstacle avoidance.

Formally, let a trajectory τ be a sequence of 6D poses in the 3D world. At time t, we model the multi-modal distribution of future trajectories conditioned on past trajectory, a visual memory encoding S, and egocentric video features F:

$$\hat{\tau}_{t:t+T} \sim p_\theta(\cdot|\tau_{1:t}, S, F) \tag{1}$$

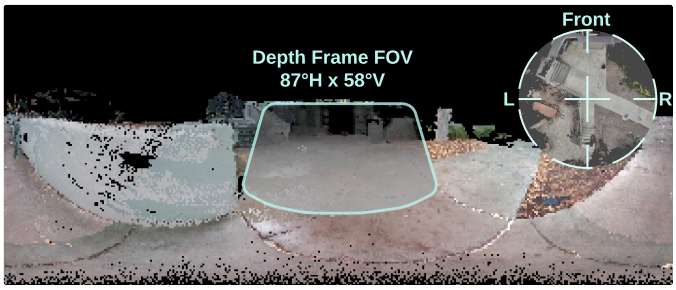

Fig. 4. **Comparing depth frame with visual memory:** A raw depth frame has only ∼90° of FOV and misses important scene information. The depth frame sees only the open space ahead and does not capture the stairs, the right turn path, or the wall to the left. Black regions are areas not yet observed.

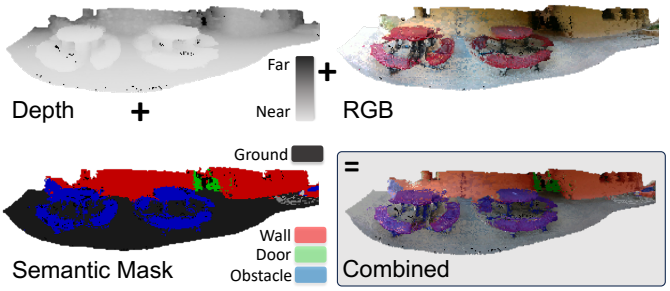

Fig. 5. **Channels in Visual Memory:** The visual memory integrates past frames into a single panorama. It consists of a depth, color, and intensity-encoded 8-class semantic channel. 4 out of 8 channels are shown in the figure.

Each sample from p_θ corresponds to a distinct plausible future path. The following sections describe the design decisions we make to realize these goals (Fig. 2).

A. Data Collection

Collection Setup. A single person wearing body-mounted cameras can gather hours of navigation data simply by walking through diverse environments, with no robot hardware or teleoperation required. Our dataset, collected on a university campus under institutional IRB approval, consists of 44 sequences (∼7 minutes, 600 meters each) recorded at 20 Hz using an Intel RealSense T265 for SLAM-based 6-DoF localization and a RealSense D455 stereo camera for aligned RGBD (Fig. 3). The full dataset totals 300 minutes and over 25 km of walking across diverse weather, surfaces, and traffic conditions. Compared to existing egocentric datasets such as TISS [40] and Aria Everyday Activities [41], ours provides dense ground-truth depth from stereo sensors at higher logging frequency.

Preprocessing. Color frames are semantically labeled by DINOv2 [42] with a Mask2Former head into 8 classes (ground, stair, door, wall, obstacle, movable, rough ground, unlabeled), and depth frames are preprocessed with a Canny edge filter to remove stereo artifacts along object boundaries. All past poses are transformed to an egocentric coordinate frame. We extract overlapping sub-trajectories for data augmentation, yielding over 320,000 training samples.

B. Scene Representation

Visual Memory (VM). A single stereo frame covers only ∼90° and misses critical context such as side paths, nearby obstacles, and the space behind the wearer (Fig. 4). The visual memory addresses this by accumulating a rolling buffer of RGBD frames into a single 360° ego-centric panorama of size $180\times360\times5$ (R, G, B, depth, semantic; 1 pixel per degree) via 3D reprojection (Fig. 5). The VM is encoded by a pretrained spatial VAE (trained with InfoLoss [43], L1 loss on RGBD channels, and cross-entropy loss on semantic channels) into an $8\times20\times8$ latent feature map that preserves the geometric layout of the panorama. This spatial encoding is then compressed to a 64-dimensional embedding by a lightweight adapter shared with the DINOv3 branch (Section III-B), which aggregates spatial features via learned attention. The VAE remains frozen during diffusion training. VM construction runs on the Jetson Orin NX CPU within 30 ms. Removing all visual input (VM and video features) produces the worst collision avoidance in our offline evaluation (Section IV), confirming the importance of scene context.

Egocentric Video Features. While the VM provides a rich spatial summary, it has three blind spots: transparent surfaces such as glass walls are invisible to stereo depth, sensor inaccuracies produce artifacts along object boundaries, and dynamic obstacles that enter or leave the scene between buffer updates are not reliably captured. To complement the VM with appearance-level cues, we extract patch-level features from the most recent egocentric RGB frame using a frozen DINOv3 ViT-S/16 backbone [16], yielding a 15×27 spatial grid of 384-dimensional patch tokens. A lightweight adapter aggregates these via learned attention: first along the horizontal axis (left–right context), then along the vertical axis (near–far context) with an intermediate 1D convolution, producing a 64-dimensional embedding. At deployment, the smallest ViT-S/16 runs ∼6 ms per frame, adding minimal latency. The impact of these features is validated in real-world deployment (Section IV-C), where they are critical for avoiding obstacles invisible to depth alone.

C. Trajectory Diffusion Model

Architecture. The prediction module is a non-autoregressive UNet diffusion model that predicts all 100 future steps (5 seconds at 20 Hz) simultaneously, ensuring temporal coherence across the full trajectory. Multi-head self-attention (MHSA) layers between downsample and upsample blocks relate conditioning embeddings to different parts of the trajectory. The past trajectory is represented as 100 steps of x-y-z position and ortho6d [44] orientation (9 channels) at 20 Hz. All three conditions: VM encoding, video features, and past trajectory, are fused by an adapter module and injected into each UNet block via classifier-free guidance, dropping each condition 10% of the time during training.

D. Deployment on Unitree G1

Hardware. The system is deployed on a Unitree G1 humanoid. We mount the data collection cameras directly on the chest of the robot. VM construction, VAE encoding, and trajectory controller run locally on the onboard Jetson Orin NX. Due to constrained compute on edge device, the

diffusion model and semantic segmentation run on a Jetson Thor connected via low-latency local network, which can be carried on the back of the robot for long-range trials. The complete prediction loop runs in real time at $\sim$2 Hz.

Hybrid Generation. To achieve real-time inference, we introduce a hybrid sampling scheme that initiates with DDIM steps to quickly approximate the trajectory distribution, followed by DDPM refinement steps to recover fine-grained details. This retains the multi-modal structure of full DDPM while being 100$\times$ faster. The model is trained with a linear noise schedule; at inference, DDIM uses evenly spaced skip steps to rapidly reach a low-noise regime, then DDPM walks the consecutive final steps to recover fine details. All three choices:linear schedule, skip-step DDIM, and consecutive-step DDPM finishing, are necessary: replacing the linear schedule (e.g. with cosine) breaks the skip-step assumption, and omitting either phase degrades quality significantly. We evaluate step combinations in Section IV; the optimal configuration of 5 DDIM + 5 DDPM steps achieves near-full-DDPM quality at real-time rates. On Jetson Thor, the model generates 110 trajectories per second; with a batch size of 64, inference runs at approximately 1.7 Hz, sufficient for closed-loop control.

Receding Horizon Controller. The navigation prior outputs 6-DoF waypoints in a local egocentric frame. Because this representation specifies *where* to walk without dictating *how*, it serves as an embodiment-agnostic interface: the same model trained on human data transfers to the G1 without finetuning. At each prediction cycle, the diffusion model generates 64 candidate trajectories spanning a 5-second horizon. Trajectories that collide with obstacles in the VM point cloud are filtered out via KD-tree queries. The remaining candidates are clustered via K-Means (k=3) based on their positions at a 2-second horizon. Each cluster receives two scores: a popularity score (fraction of trajectories in the cluster) and a momentum penalty (distance from the cluster center to the previously selected intention point). The combined score favors large, consistent clusters, preventing erratic switching between modes across cycles. The medoid of the highest-scoring cluster is selected for execution. The controller dynamically estimates the current system latency, discards the corresponding initial segment of the trajectory, and executes the plan with smooth blending against the previous trajectory until the next prediction batch arrives.

IV. Evaluation

We evaluate EgoNav to answer three questions:

- **Q1**: Does each design component contribute to prediction quality? (Section IV-B)
- **Q2**: Does the multi-modal approach outperform unimodal and non-panoramic baselines? (Section IV-B)
- **Q3**: Can the learned prior, trained entirely on human data, deploy on a real humanoid in unseen environments with zero robot data? (Section IV-C)

A. Experimental Setup

Metrics. We evaluate predictions using 3 metrics (Fig. 6).

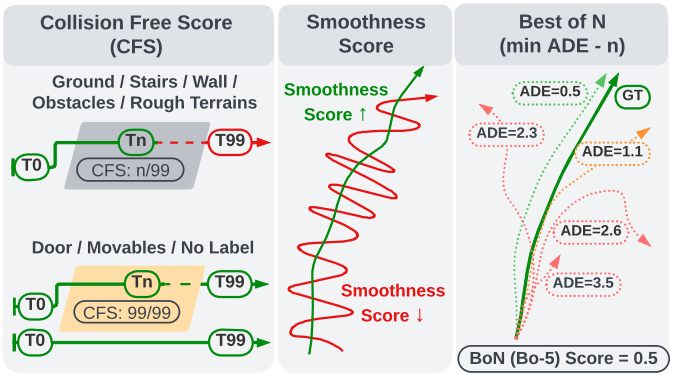

Fig. 6. **Metrics overview:** Collision-free score (CFS) uses selected semantics. All subsequent trajectories will be marked as collided; higher CFS and Smoothness are better, lower Best of N is better.

Collision-free score measures trajectory feasibility against a point cloud re-projected from the VM: at each time step, we query the 20 nearest points via a KD-tree; a collision is detected if more than 10 points fall within a 16 cm radius, and all subsequent steps are marked as collided. The score counts consecutive collision-free steps against static obstacles (ground, stairs, walls, rough terrain), excluding doors and movable objects.

Smoothness is the reciprocal of the mean absolute error in speed and acceleration relative to ground truth:

$$Smoothness = 1/\frac{\sum_{i=1}^{n}(|V_i - \hat{V}_i| + |a_i - \hat{a}_i|)}{n} \quad (2)$$

Best of N (minADE-K) selects the closest of K predicted trajectories to ground truth, measuring multi-modal coverage. While *Best of 1* (ADE) evaluates single-prediction accuracy for comparison with unimodal baselines.

Dataset and Training. We report the full system trained on the complete dataset at the top of Table I. The w/o DINOv3 variant uses the same checkpoint with video features zeroed out at inference (enabled by classifier-free guidance). The diffusion model has 46M parameters. Component ablations that require retraining (architecture or input changes) and baselines are trained on the pilot subset for faster iteration, sharing the same pretrained VAE encoder for fair comparison. The ground truth (GT) collision score (97.6) is not perfect due to SLAM drift and depth artifacts.

B. Offline Evaluation

We ablate key components and compare baselines in Table I.

Scene Understanding. Complete removal of visual input yields the worst-performing model (collision 82.5), confirming the importance of scene context; without the VM, generated samples largely memorize training trajectories and produce incorrect mode distributions. Removing the semantic channel causes a significant drop in collision avoidance (-5.1): without semantics, the model cannot differentiate between doors and walls, and conflicting training signals cause it to ignore geometric constraints in the VM.

Model Architecture. Adding multi-head self-attention (MHSA) layers between down and up blocks improves collision avoidance by $+2.6$ with negligible impact on inference

TABLE I
OFFLINE EVALUATION ACROSS ABLATIONS AND BASELINES. **BEST**, *2nd*.

Full System (Full Data)	Collision ↑	Smoothness ↑	Best of 1 ↓	Best of 15 ↓
GT	97.6	∞	0.0	0.0
Ours	**91.4**	**4.82**	**0.76**	**0.39**
w/o DINOv3 features	*90.6*	*4.76*	*0.81*	*0.41*
Ours - Pilot	89.2	2.04	0.87	0.47
Component Ablations (Pilot Data)				
w/o attention	86.6	2.78	1.00	0.49
w/o semantic	84.1	4.17	0.91	0.53
w/o VM or DINO (Traj only)	82.5	2.04	1.19	0.48
Baselines (Pilot Data)				
Ours - Pilot	**89.2**	2.04	**0.87**	**0.47**
VAE-LSTM	*84.5*	**9.09**	*1.01*	N/A
CXA Transformer	80.3	*3.70*	1.25	N/A

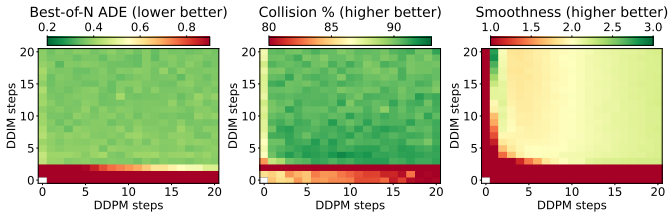

Fig. 7. **Hybrid generation step search:** BoN, Collision, Smoothness score for different DDIM/DDPM step combinations. The optimal configuration is 5 DDIM + 5 DDPM. All sweeps are performed on the same checkpoint. Combination 0+0 is pure noise therefore omitted.

speed, confirming that attention helps relate VM context to trajectory samples.

DINOv3 Video Features. Zeroing out DINOv3 features at inference causes a modest offline drop (91.4 → 90.6 collision). This is expected: the offline collision metric relies on the same depth-based point cloud which does not see glass or dynamic obstacles where the DINOv3 variant excels. The critical role of DINOv3 emerges in real-world deployment, as we analyzed later in Section IV-C.

Data Scaling. Training on the full dataset improves all metrics over the pilot subset (collision 89.2 → 91.4, Best of 15 0.47 → 0.39), confirming that our navigation prior is scalable and can benefit from more diverse walking data.

Hybrid Generation. We systematically evaluate different combinations of DDIM and DDPM steps (Fig. 7). The optimal configuration of 5 DDIM + 5 DDPM steps achieves near-full-DDPM quality with 100× fewer steps. Fewer than 7 total steps degrades smoothness significantly. Pure DDIM with the same total steps performs notably worse in both collision and smoothness, confirming the value of the DDPM refinement.

Comparison with Baselines.

VAE-LSTM [13] auto-regressively predicts future states, producing deterministic, unimodal outputs. Its step-by-step generation excels at smooth motion (+7.05), and it demonstrates surprisingly robust obstacle avoidance despite fewer than 1M parameters, likely due to its panorama depth inputs providing broad environmental context similar to our VM. However, it cannot represent multi-modal distributions and requires one forward pass per predicted step, making it the slowest model out of three.

CXA-Transformer [40] employs cascaded cross-attention to fuse pedestrian poses, semantic segmentation, and past trajectories. Its collision score is significantly lower (−8.9) than ours, and its ADE is worse (+0.38), though it achieves higher smoothness (+1.66). The original method uses 2 Hz RGBD directly without a panoramic representation, which limits situational awareness; without the structured scene context that a VM provides, the model likely requires significantly more data to learn robust obstacle avoidance. Like LSTM, it produces deterministic predictions. Overall, our method achieves the best collision avoidance and mode coverage among all compared methods.

TABLE II
REAL-WORLD DEPLOYMENT STATISTICS ON THE UNITREE G1.

	Static	Corridor	Glass	Dynamic
Duration (min)	16	7.5	6.5 (5)	7.5
Distance (m)	471	240	225 (120)	201
Interventions	9	1	5 (11)	6
Interventions/min	0.56	0.13	0.77 (2.2)	0.80
Autonomous time%	97.2	99.3	96.2 (89.0)	96.0

Glass column: values in parentheses are w/o DINOv3.

C. Real-world Deployment

To validate EgoNav beyond offline metrics, we deploy the full system zero-shot on a Unitree G1 humanoid robot in unseen environments. We do not compare against baselines, as they either lack a real-time system (LookOut, CXA, VAE-LSTM) or require robot-collected data (NoMaD, HEAD). We evaluate across four categories (Table II): static indoor scenes (kitchen, lab), corridors, areas with glass walls, and dynamic scenes (pedestrians, moving objects). Over 37.5 minutes of autonomous operation covering 1,137 m, the system achieves 96–99% autonomous time, with corridors reaching 99% and the more challenging glass and dynamic scenes slightly lower.

During static scene evaluation, we randomly rearrange objects to test generalization. The system handles the kitchen even with all cabinets pulled out (Fig. 8B). The supplementary video shows robustness in more scenes where furniture is moved to random locations. At a T-junction (Fig. 8C), EgoNav predicts a bimodal distribution and momentum matching commits to turning left. The system also exhibits surprising dynamic scene behaviors. In Fig. 8D, it waits at a closed door until it opens, then proceeds when the path is clear. In a crowd gathering scene (Fig. 8E), it finds narrow gaps between pedestrians and navigates through.

DINOv3 features prove critical for glass environments: in Fig. 8F the system turns away from a glass wall invisible to depth sensors. Without DINOv3, the intervention rate in glass environments nearly triples (0.77 → 2.2/min, Table II).

Failure Mode Analysis. We also deployed ablation versions to isolate each component. Without DINOv3, glass walls invisible to depth cause the intervention rate to nearly triple (0.77 → 2.2/min, Table II), confirming that the modest offline delta masks a critical real-world contribution. Without semantic labels, geometrically similar doors and walls become indistinguishable, and the model frequently predicts paths through solid walls. Without the momentum penalty, trajectory clusters alternate between modes across cycles, producing oscillatory behavior at decision points.

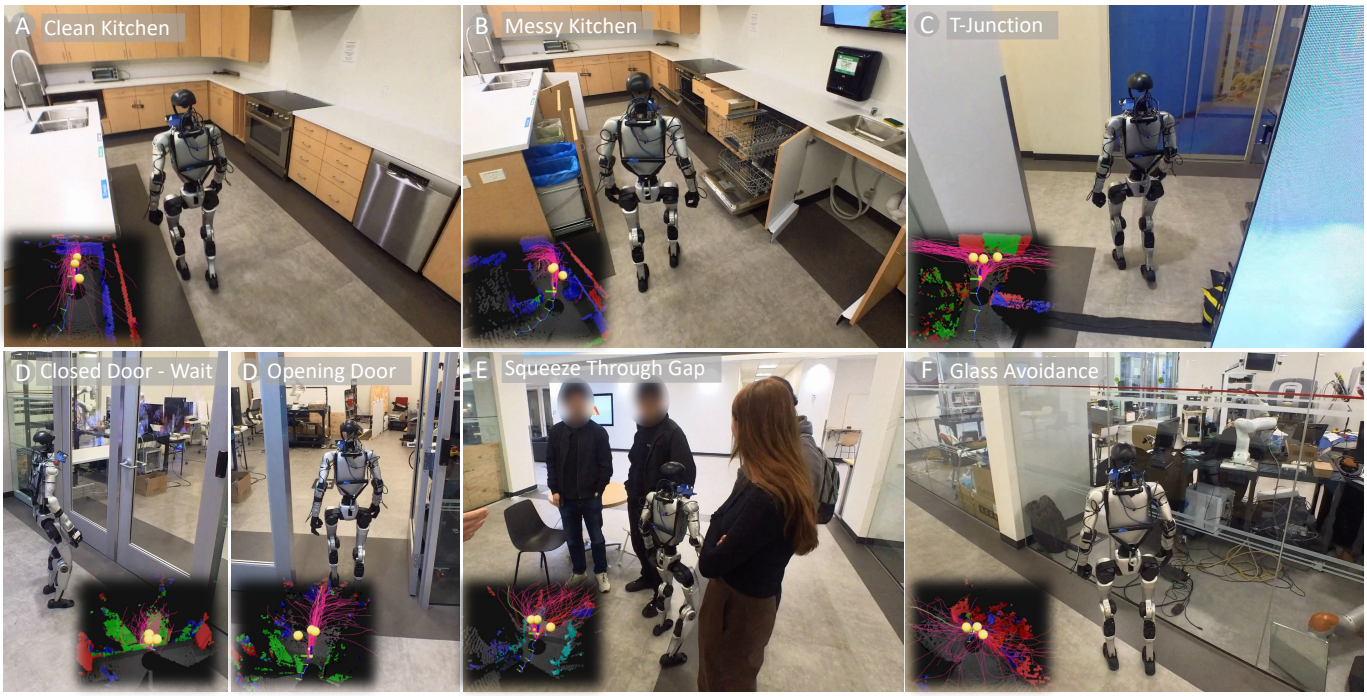

Fig. 8. **Real-world deployment on Unitree G1.** The humanoid navigates diverse unseen environments using EgoNav with zero robot training data. Insets show predicted trajectory distributions overlaid on the semantic point cloud. (A,B) Static indoor scenes with different clutter levels. (C) T-junction with multi-modal trajectory predictions. (D) The robot waits at a closed door and proceeds when it opens, demonstrating semantic understanding. (E) Navigating through a gap between pedestrians. (F) Avoiding a glass wall invisible to depth sensors, enabled by DINOv3 video features.

V. Conclusion

We presented EgoNav, a system that learns humanoid navigation entirely from human walking data, requiring no robot data collection or task-specific finetuning. By training a conditional diffusion model on egocentric observations of human walking, the system internalizes commonsense navigation behavior: which paths are traversable, how to avoid obstacles, and where plausible routes exist. A 360° visual memory and DINOv3 video features provide rich scene understanding, while a hybrid sampling scheme enables real-time inference. We validated EgoNav through offline evaluations and real-world deployment on a Unitree G1 humanoid robot navigating unseen indoor and outdoor environments, demonstrating that human walking data alone can produce a navigation prior capable of semantic reasoning: waiting at closed doors, navigating around pedestrians, and avoiding transparent obstacles, all without explicit programming.

Future work includes goal-conditioned trajectory selection for directed navigation, monocular depth estimation to relax the stereo camera requirement, and scaling to more diverse environments. More broadly, a navigation prior can serve as the missing middle layer between high-level task planning and low-level locomotion: the predicted paths tell a controller where to go, while the semantic visual memory along those paths informs which locomotion skill to use.

References

[1] A. Sridhar, D. Shah, C. Glossop, and S. Levine, "NoMaD: Goal Masked Diffusion Policies for Navigation and Exploration," Oct. 2023, arXiv:2310.07896 [cs]. [Online]. Available: http://arxiv.org/abs/2310.07896

[2] A.-C. Cheng, Y. Ji, Z. Yang, Z. Gongye, X. Zou, J. Kautz, E. Bıyık, H. Yin, S. Liu, and X. Wang, "NaVILA: Legged Robot Vision-Language-Action Model for Navigation," Feb. 2025, arXiv:2412.04453 [cs]. [Online]. Available: http://arxiv.org/abs/2412.04453

[3] C. Chi, Z. Xu, C. Pan, E. Cousineau, B. Burchfiel, S. Feng, R. Tedrake, and S. Song, "Universal Manipulation Interface: In-The-Wild Robot Teaching Without In-The-Wild Robots," Mar. 2024, arXiv:2402.10329 [cs]. [Online]. Available: http://arxiv.org/abs/2402.10329

[4] C. Wang, H. Shi, W. Wang, R. Zhang, L. Fei-Fei, and C. K. Liu, "DexCap: Scalable and Portable Mocap Data Collection System for Dexterous Manipulation," Jul. 2024, arXiv:2403.07788 [cs]. [Online]. Available: http://arxiv.org/abs/2403.07788

[5] S. Chen, Y. Ye, Z.-A. Cao, J. Lew, P. Xu, and C. K. Liu, "Hand-Eye Autonomous Delivery: Learning Humanoid Navigation, Locomotion and Reaching," Aug. 2025, arXiv:2508.03068 [cs]. [Online]. Available: http://arxiv.org/abs/2508.03068

[6] H. Cui, V. Radosavljevic, F.-C. Chou, T.-H. Lin, T. Nguyen, T.-K. Huang, J. Schneider, and N. Djuric, "Multimodal Trajectory Predictions for Autonomous Driving using Deep Convolutional Networks," Mar. 2019, arXiv:1809.10732 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1809.10732

[7] B. Varadarajan, A. Hefny, A. Srivastava, K. S. Refaat, N. Nayakanti, A. Cornman, K. Chen, B. Douillard, C. P. Lam, D. Anguelov, and B. Sapp, "MultiPath++: Efficient Information Fusion and Trajectory Aggregation for Behavior Prediction," in *2022 International Conference on Robotics and Automation (ICRA)*. Philadelphia, PA, USA: IEEE, May 2022, pp. 7814–7821. [Online]. Available: https://ieeexplore.ieee.org/document/9812107/

[8] D. Rempe, Z. Luo, X. B. Peng, Y. Yuan, K. Kitani, K. Kreis, S. Fidler, and O. Litany, "Trace and Pace: Controllable Pedestrian Animation via Guided Trajectory Diffusion," Apr. 2023, arXiv:2304.01893 [cs]. [Online]. Available: http://arxiv.org/abs/2304.01893

[9] J. Wiest, M. Höffken, U. Kreßel, and K. Dietmayer, "Probabilistic trajectory prediction with Gaussian mixture models," in *2012 IEEE Intelligent Vehicles Symposium*, Jun. 2012, pp. 141–146, iSSN: 1931-0587.

[10] T. Gu, G. Chen, J. Li, C. Lin, Y. Rao, J. Zhou, and J. Lu, "Stochastic Trajectory Prediction via Motion Indeterminacy Diffusion," 2022, pp. 17 113–17 122. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2022/html/Gu_Stochastic_Trajectory_Prediction_via_Motion_Indeterminacy_Diffusion_CVPR_2022_paper.html

[11] K. Lv, L. Yuan, and X. Ni, "Learning Autoencoder Diffusion Models of Pedestrian Group Relationships for Multimodal Trajectory Prediction," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–12, 2024, conference Name: IEEE Transactions on Instrumentation and Measurement. [Online]. Available: https://ieeexplore.ieee.org/document/10466609

[12] D. Helbing and P. Molnar, "Social Force Model for Pedestrian Dynamics," *Physical Review E*, vol. 51, no. 5, pp. 4282–4286, May 1995, arXiv:cond-mat/9805244. [Online]. Available: http://arxiv.org/abs/cond-mat/9805244

[13] W. Wang, M. Raitor, S. Collins, C. K. Liu, and M. Kennedy III, "Trajectory and Sway Prediction Towards Fall Prevention," Mar. 2023, arXiv:2209.11886 [cs]. [Online]. Available: http://arxiv.org/abs/2209.11886

[14] B. Pan, A. W. Harley, C. K. Liu, and L. J. Guibas, "LookOut: Real-World Humanoid Egocentric Navigation," Aug. 2025, arXiv:2508.14466 [cs]. [Online]. Available: http://arxiv.org/abs/2508.14466

[15] Z. Qiu, Z. Liu, W. Niu, T. Bhattacharjee, and S. Kalantari, "EgoCogNav: Cognition-aware Human Egocentric Navigation," Nov. 2025, arXiv:2511.17581 [cs]. [Online]. Available: http://arxiv.org/abs/2511.17581

[16] 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," Aug. 2025, arXiv:2508.10104 [cs]. [Online]. Available: http://arxiv.org/abs/2508.10104

[17] W. Jia, B. Lai, M. Liu, D. Xu, and J. M. Rehg, "Learning Predictive Visuomotor Coordination," Mar. 2025, arXiv:2503.23300 [cs]. [Online]. Available: http://arxiv.org/abs/2503.23300

[18] Y. C. Tang and R. Salakhutdinov, "Multiple Futures Prediction," Dec. 2019, arXiv:1911.00997 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1911.00997

[19] J. Mercat, T. Gilles, N. E. Zoghby, G. Sandou, D. Beauvois, and G. P. Gil, "Multi-Head Attention for Multi-Modal Joint Vehicle Motion Forecasting," Dec. 2019, arXiv:1910.03650 [cs]. [Online]. Available: http://arxiv.org/abs/1910.03650

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," Aug. 2023, arXiv:1706.03762 [cs]. [Online]. Available: http://arxiv.org/abs/1706.03762

[21] J. Strohbeck, V. Belagiannis, J. Muller, M. Schreiber, M. Herrmann, D. Wolf, and M. Buchholz, "Multiple Trajectory Prediction with Deep Temporal and Spatial Convolutional Neural Networks," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Las Vegas, NV, USA: IEEE, Oct. 2020, pp. 1992–1998. [Online]. Available: https://ieeexplore.ieee.org/document/9341327/

[22] M. Schreiber, S. Hoermann, and K. Dietmayer, "Long-Term Occupancy Grid Prediction Using Recurrent Neural Networks," Jun. 2019, arXiv:1809.03782 [cs]. [Online]. Available: http://arxiv.org/abs/1809.03782

[23] Y. Chai, B. Sapp, M. Bansal, and D. Anguelov, "MultiPath: Multiple Probabilistic Anchor Trajectory Hypotheses for Behavior Prediction," Oct. 2019, arXiv:1910.05449 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1910.05449

[24] D. Rempe, J. Philion, L. J. Guibas, S. Fidler, and O. Litany, "Generating Useful Accident-Prone Driving Scenarios via a Learned Traffic Prior," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans, LA, USA: IEEE, Jun. 2022, pp. 17 284–17 294. [Online]. Available: https://ieeexplore.ieee.org/document/9880074/

[25] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social GAN: Socially Acceptable Trajectories with Generative Adversarial Networks," Mar. 2018, arXiv:1803.10892 [cs]. [Online]. Available: http://arxiv.org/abs/1803.10892

[26] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social LSTM: Human Trajectory Prediction in Crowded Spaces," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 961–971. [Online]. Available: http://ieeexplore.ieee.org/document/7780479/

[27] A. Lerner, Y. Chrysanthou, and D. Lischinski, "Crowds by Example," *Computer Graphics Forum*, vol. 26, no. 3, pp. 655–664, 2007, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-8659.2007.01089.x. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-8659.2007.01089.x

[28] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese, "Learning Social Etiquette: Human Trajectory Understanding In Crowded Scenes," in *Computer Vision – ECCV 2016*, ser. Lecture Notes in Computer Science, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 549–565.

[29] Y. Chen, B. Ivanovic, and M. Pavone, "ScePT: Scene-consistent, Policy-based Trajectory Predictions for Planning," Jun. 2022, arXiv:2206.13387 [cs]. [Online]. Available: http://arxiv.org/abs/2206.13387

[30] Q. Sun, S. Zhang, D. Ma, J. Shi, D. Li, S. Luo, Y. Wang, N. Xu, G. Cao, and H. Zhao, "Large Trajectory Models are Scalable Motion Predictors and Planners," Feb. 2024, arXiv:2310.19620. [Online]. Available: http://arxiv.org/abs/2310.19620

[31] K. K. Singh, K. Fatahalian, and A. A. Efros, "KrishnaCam: Using a longitudinal, single-person, egocentric dataset for scene understanding tasks," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. Lake Placid, NY, USA: IEEE, Mar. 2016, pp. 1–9. [Online]. Available: http://ieeexplore.ieee.org/document/7477717/

[32] H. S. Park, J.-J. Hwang, Y. Niu, and J. Shi, "Egocentric Future Localization," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 4697–4705. [Online]. Available: http://ieeexplore.ieee.org/document/7780877/

[33] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," Dec. 2020, arXiv:2006.11239 [cs, stat]. [Online]. Available: http://arxiv.org/abs/2006.11239

[34] J. Song, C. Meng, and S. Ermon, "Denoising Diffusion Implicit Models," Oct. 2022, arXiv:2010.02502 [cs]. [Online]. Available: http://arxiv.org/abs/2010.02502

[35] Z. Fu, A. Kumar, A. Agarwal, H. Qi, J. Malik, and D. Pathak, "Coupling Vision and Proprioception for Navigation of Legged Robots," Jul. 2022, arXiv:2112.02094 [cs]. [Online]. Available: http://arxiv.org/abs/2112.02094

[36] D. Hoeller, N. Rudin, D. Sako, and M. Hutter, "ANYmal Parkour: Learning Agile Navigation for Quadrupedal Robots," Jun. 2023, arXiv:2306.14874 [cs]. [Online]. Available: http://arxiv.org/abs/2306.14874

[37] T. H. Yang, H. Shi, J. Hu, Z. Zhang, D. Jiang, W. Wang, Y. He, Z. Wu, Y. Chen, Y. Hou, M. Kennedy, S. Song, and C. K. Liu, "Locomotion Beyond Feet," Jan. 2026, arXiv:2601.03607 [cs]. [Online]. Available: http://arxiv.org/abs/2601.03607

[38] I. Radosavovic, B. Zhang, B. Shi, J. Rajasegaran, S. Kamat, T. Darrell, K. Sreenath, and J. Malik, "Humanoid Locomotion as Next Token Prediction," Feb. 2024, arXiv:2402.19469 [cs]. [Online]. Available: http://arxiv.org/abs/2402.19469

[39] K. Hu, H. Shi, Y. He, W. Wang, C. K. Liu, and S. Song, "Robot Trains Robot: Automatic Real-World Policy Adaptation and Learning for Humanoids," Aug. 2025, arXiv:2508.12252 [cs]. [Online]. Available: http://arxiv.org/abs/2508.12252

[40] J. Qiu, L. Chen, X. Gu, F. P.-W. Lo, Y.-Y. Tsai, J. Sun, J. Liu, and B. Lo, "Egocentric Human Trajectory Forecasting With a Wearable Camera and Multi-Modal Fusion," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 8799–8806, Oct. 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9813561/

[41] Z. Lv, N. Charron, P. Moulon, A. Gamino, C. Peng, C. Sweeney, E. Miller, H. Tang, J. Meissner, J. Dong, K. Somasundaram, L. Pesqueira, M. Schwesinger, O. Parkhi, Q. Gu, R. De Nardi, S. Cheng, S. Saarinen, V. Baiyya, Y. Zou, R. Newcombe, J. J. Engel, X. Pan, and C. Ren, "Aria Everyday Activities Dataset," Feb. 2024, arXiv:2402.13349 [cs]. [Online]. Available: http://arxiv.org/abs/2402.13349

[42] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "DINOv2: Learning Robust Visual Features without Supervision," Feb. 2024, arXiv:2304.07193 [cs]. [Online]. Available: http://arxiv.org/abs/2304.07193

[43] S. Zhao, J. Song, and S. Ermon, "InfoVAE: Information Maximizing Variational Autoencoders," May 2018, arXiv:1706.02262 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1706.02262

[44] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li, "On the Continuity of Rotation Representations in Neural Networks," Jun. 2020, arXiv:1812.07035 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1812.07035