TANGO: Humanoid Navigation in Cluttered Environments with a Whole-Body Vision-Language-Action Model

Author Names Omitted for Anonymous Review.
https://anonymous-proj-2026.github.io/

Fig. 1: **Long-Horizon Humanoid Navigation in Cluttered Scenes.** We propose TANGO, a whole-body foundation model for cluttered indoor scene navigation. TANGO demonstrates strong scene understanding and robust traversal capability, navigating a 30-meter route zero-shot in a real-world cluttered office scene.

***Abstract*—We study the problem of navigating cluttered indoor environments with a humanoid robot. Unlike conventional navigation methods that operate over simplified navigation abstractions, humanoid traversal in cluttered environments requires continuous geometry-aware whole-body adaptation, including coordinated arm placement, torso adjustment, and gait modulation for collision-free movement through complex 3D spaces. We introduce TANGO, the first whole-body vision-language navigation framework for language-conditioned humanoid traversal in cluttered environments. Given a natural-language instruction and egocentric RGB observations, TANGO predicts 29-DoF joint-space actions for whole-body humanoid control. We train TANGO entirely in simulation by synthesizing diverse collision-free traversal behaviors via global path planning, kinematic whole-body motion generation, obstacle-aware motion editing, and reinforcement-learning-based tracking. This pipeline provides dynamically feasible action supervision for learning language-conditioned whole-body policies. In extensive simulation experiments, TANGO demonstrates state-of-the-art performance on vision-language navigation, while outperforming strong modular baselines on navigating challenging scenes requiring obstacle negotiation. Lastly, we deploy TANGO zero-shot on a Unitree G1 humanoid robot, demonstrating robust language-guided traversal in cluttered real-world scenes without training on any real-world navigation data.**

I. Introduction

Language-guided traversal in complex 3D environments is a fundamental capability for domestic humanoid robots expected to assist with everyday tasks. Unlike wheeled or mobile-base robots, humanoids navigate with high-dimensional articulated bodies whose geometry changes continuously during motion, making the robot's body configuration an inherent part of the navigation problem. As illustrated in Figure 1, traversal feasibility in cluttered indoor spaces depends not only on the intended route but also on whether the robot can physically move through the surrounding scene geometry without collisions involving the arms, torso, or legs. This creates a tight coupling between navigation decisions and whole-body feasibility: *An action that appears valid at the planning level may still be infeasible for the embodied humanoid to execute.*

Despite rapid progress in humanoid control [1, 2, 3, 4], vision-language navigation (VLN) [5, 6, 7, 8], and collision-aware traversal [9, 10, 11], effectively integrating these capabilities remains largely unexplored.

Existing VLN methods typically formulate navigation as high-level decision making, where an agent predicts 2D waypoints or discrete actions from visual observations and language instructions [12, 13, 14]. While effective for mobile platforms and simplified embodied agents, their low-dimensional action spaces cannot explicitly represent the relationship between navigation decisions and whole-body feasibility, limiting their ability to handle spatially constrained traversal scenarios. Recent humanoid foundation models and whole-body VLA systems [15, 16, 17] have demonstrated impressive whole-body control capabilities. Nevertheless, navigation in these systems is typically represented through high-level locomotion commands and delegated to downstream controllers, preventing explicit reasoning about whole-body traversability during navigation. A complementary line of work explores collision-aware humanoid traversal through reinforcement learning [10, 18]. While effective in specific traversal scenarios, these approaches often rely on task-specific priors or training distributions, limiting their scalability to long-horizon language-guided navigation in diverse cluttered environments. Consequently, existing approaches remain unable to jointly reason about navigation intent and whole-body traversability, motivating the need for a unified framework for language-guided whole-body navigation.

To address these challenges, we present **T**raversal via Vision-Language **A**ction and **N**avigation for **G**eometry-Aware Whole-Body C**O**ntrol (**TANGO**), a unified VLA framework for humanoid navigation in cluttered environments. Given a language instruction and egocentric RGB observations, TANGO directly predicts 29-DoF joint-space actions for whole-body humanoid control, avoiding the need for separate navigation and control modules. A key challenge is obtaining large-scale training data that captures both semantic task diversity and physically plausible whole-body traversal behaviors. To this end, we develop a scalable simulation pipeline that automatically synthesizes collision-free humanoid traversal trajectories and provides dynamically feasible supervision for training the VLA model. At deployment, the learned policy is executed through a robust motion tracker with real-time action chunking [1], enabling reliable execution and zero-shot transfer to real humanoid hardware. We evaluate TANGO against state-of-the-art VLN and humanoid traversal baselines in both simulation and real-world environments. Across long-horizon navigation tasks requiring obstacle negotiation and geometry-aware whole-body adaptation, TANGO consistently achieves stronger collision-free traversal performance than existing methods.

II. Related Works

Large Models for Vision-Language Navigation. Recent large multi-modality models (LMMs) emerge with strong scene understanding and physical awareness, leading to extensive zero-shot navigation works that leverage off-the-shelf large models [19, 20, 21, 22]. Moreover, recent efforts have explored fine-tuning such models on simulated and real-world navigation samples, resulting in strong VLA models for navigating in diversified environments [7, 23, 24, 9, 6, 14]. Nevertheless, these methods take visual navigation as a pure planar trajectory planning task, omitting the physical gap [25] when deployed in real physical environments. In contrast, TANGO is trained with inherent physical awareness, mitigating the embodied gap while enabling explicit reasoning about whole-body traversability in cluttered environments. Recent efforts have also explored dual-system design for VLA models to achieve continuous, real-time navigation [9, 26]. In this work, we equip TANGO with a flow-matching-based action expert as *system-1*, trained with real-time chunking [27], to achieve continuous and responsive humanoid control.

Cluttered Environment Traversal. Traversal in cluttered scenes is critical for deploying embodied agents in complex real-world scenarios. Recent humanoid parkour works have demonstrated impressive traversal capabilities over challenging terrains and obstacles [28, 18, 29]. However, these methods mainly focus on short-horizon interactions with scene objects. In contrast, TANGO performs long-horizon navigation with collision avoidance, which requires excellence in physical and semantic understanding capabilities. HumanoidPF [10] introduces RL-based collision-free indoor traversal for humanoids and achieves high success rates in most cases, but remains difficult to scale, especially to long-horizon navigation and complex obstacle compositions. In this work, we synthesize collision-free and dynamically feasible humanoid motions through a scalable pipeline, collecting low-cost, high-quality datasets for 3D traversal policy training. Some VLN works [11, 30] also study traversal in cluttered scenes, but are fundamentally limited by their 2D problem formulation and primarily consider bypassing behaviors. In contrast, TANGO learns humanoid whole-body motions, enabling richer capabilities when facing complex obstacles.

Humanoid Whole-Body Control through Large-Scale Learning. Recent advances in humanoid motion tracking [31, 2, 1] have enabled large-scale learning of humanoid control policies. Representative works such as GR00T-N1.6 [16], Ψ_0 [15], and WholeBodyVLA [17] adopt a decoupled design, predicting upper-body motions while issuing high-level commands to a lower-body tracker. While this significantly simplifies loco-manipulation learning, it limits whole-body coordination required for tasks such as cluttered-scene traversal. In contrast, TANGO directly learns end-to-end whole-body motions and uses them as reference trajectories for a low-level tracker. Non-decoupled approaches, including LeVERB [32], HumanoidVLA [33], and PhysiFlow [34], learn latent motion representations decoded by specialized controllers. Instead, TANGO directly predicts executable whole-body actions and relies on a pre-trained general-purpose tracker for execution, avoiding controller co-training and task-specific motion decoders while enabling scalable deployment across diverse

humanoid platforms and traversal tasks.

III. METHOD

We present Traversal via Whole-Body Vision-Language-Action Model (TANGO), a whole-body VLA system for cluttered indoor vision-language navigation (Figure 2). In this section, we first provide a formal definition of whole-body vision-language navigation (Section III-A). Next, we introduce an automated data-generation pipeline for building a large-scale whole-body navigation dataset through scalable scene augmentation and motion editing in simulation environments (Section III-B). We then describe how TANGO generates whole-body motions from language instructions and RGB observations (Section III-C). Lastly, we describe how to deploy TANGO in both simulated environments and on a real humanoid robot (Section III-D).

A. Problem Formulation

Existing VLN works are fundamentally limited by their planar action spaces, failing to represent complex traversing motions in real-world environments. In this work, we study the problem of ***whole-body vision-language navigation***. Given a navigation language instruction ℓ, current observation $\mathbf{o}_t$ containing a temporal sequence of RGB images from front and downward cameras $\mathbf{I}_{1:t}^{\text{fr,dn}}$ and whole-body joint-angle proprioceptive state $\mathbf{q}_t$, our model learns to predict a *whole-body action chunk* $\mathbf{A}_t = \{\mathbf{a}_1, \cdots, \mathbf{a}_H\}$ over an action horizon H, where $\mathbf{a}_i = \{\mathbf{q}_{\text{d},i}, \mathbf{r}_{\text{b},i}\},\ i \in \{1, \cdots, H\}$, with $\mathbf{q}_{\text{d},i} \in \mathbb{R}^{29}$ and $\mathbf{r}_{\text{b},i} \in \mathbb{R}^6$ denoting the desired whole-body joint angles and base 6D rotation representation at the i-th step, respectively. The predicted action chunk is then streamed to a low-level motion tracker for physically grounded humanoid navigation.

B. Simulation Data Generation

Environment Augmentation. Existing navigation datasets [5, 35, 36] primarily capture standard room layouts with loose spatial constraints, rarely requires complex whole-body behaviors like sideways walking, stepping over, or crouching for humanoid traversal. To create more challenging scenarios, we augment over 1,300 indoor scenes from VLNVerse [35] and SAGE-3D [36] with realistic assets. Following HumanoidPF [10], we introduce three obstacle categories: lateral obstacles to construct narrow passages, ground-level obstacles to necessitate stepping, and overhead obstacles to enforce upper-body clearance. To ensure visual and semantic consistency, we curate category-specific templates from existing assets within the source datasets and automatically instantiate these semantically matched objects via $\mathrm{SO}(3)$ transformations. See Appendix VII-A for further details.

Collision-Free Motion Generation. Existing motion generation pipelines often rely on labor-intensive human motion capture followed by retargeting, which introduces motion degradation and embodiment mismatch. To generate scalable, human-like traversal data, we propose *Plan, Edit, Track* (PET), an automatic pipeline that synthesizes collision-free, whole-body motions for humanoid cluttered indoor navigation.

Given a start and goal location in an augmented scene, PET first *plans* a collision-aware planar reference path using A* with an obstacle-aware heuristic that biases the search toward safer regions. A heading adjustment module detects narrow passages and inserts 90-degree heading changes, inducing sideways walking when frontal traversal is spatially constrained. The resulting trajectory is converted into velocity commands for SONIC [1] to produce natural humanoid walking motions. We then replay these trajectories in IsaacSim [37] to render egocentric RGB observations at 2 Hz, which Gemini 2.5 Flash [38] uses to generate dense navigation instructions.

PET subsequently *edits* the synthesized motions to incorporate whole-body obstacle interactions. For obstacles within a spatial corridor around the planned path, we refine nearby body motions via potential-field guidance. Inspired by HumanoidPF [10], we apply sampled repulsive forces to key body links through SoftMimic-style pseudo-forces [39], while a gait adaptation module adjusts locomotion around ground-level obstacles. This stage yields reference motions with explicit whole-body avoidance behaviors, such as arm clearance, stepping over, and crouching.

Finally, PET employs a SONIC *tracker* as an executability filter. The edited motions are tracked in simulation to verify physical feasibility and collision freedom. Failing or colliding trajectories are discarded. Crucially, instead of using the tracked trajectories as training supervision, which might degrade the human-like quality, we utilize the collision-free reference motions from the planning and editing stages as the action supervision signal. This preserves human-like structures while ensuring physical executability. Based on the verified trajectories, we regenerate the RGB observations and update the language instructions accordingly. Full details are discussed in Appendix VII-B.

C. TANGO Architecture and Training

Whole-body VLA navigation in complex environments demands physical world understanding, continuous action prediction, and robust action execution. To address these requirements, TANGO adopts a *triple-system* architecture [40, 41, 15] integrating a Vision-Language (VL) backbone (*system-2*), a multi-modal diffusion transformer (MM-DiT) action expert with real-time chunking [27] (*system-1*), and an off-the-shelf motion tracker (*system-0*), as shown in Figure 2. We jointly train the VL backbone and the action expert. During deployment, the predicted action chunks are streamed to the low-level tracker to generate continuous, high-frequency humanoid control signals.

System-2: Vision-Language Perception. We instantiate *system-2* using Qwen2.5VL-7B [42], warm-started with InternVLA-N1 [9] weights to inherit strong navigation priors. At timestep t, front and downward camera views $(\mathbf{I}_t^{\text{fr}}, \mathbf{I}_t^{\text{dn}})$ are vertically stacked into a single frame $\mathbf{I}_t$. To manage long-horizon video history $\mathbf{I}_{1:T}$ without token explosion, we apply Budget-Aware Token Sampling (BATS) [6]. Frames are

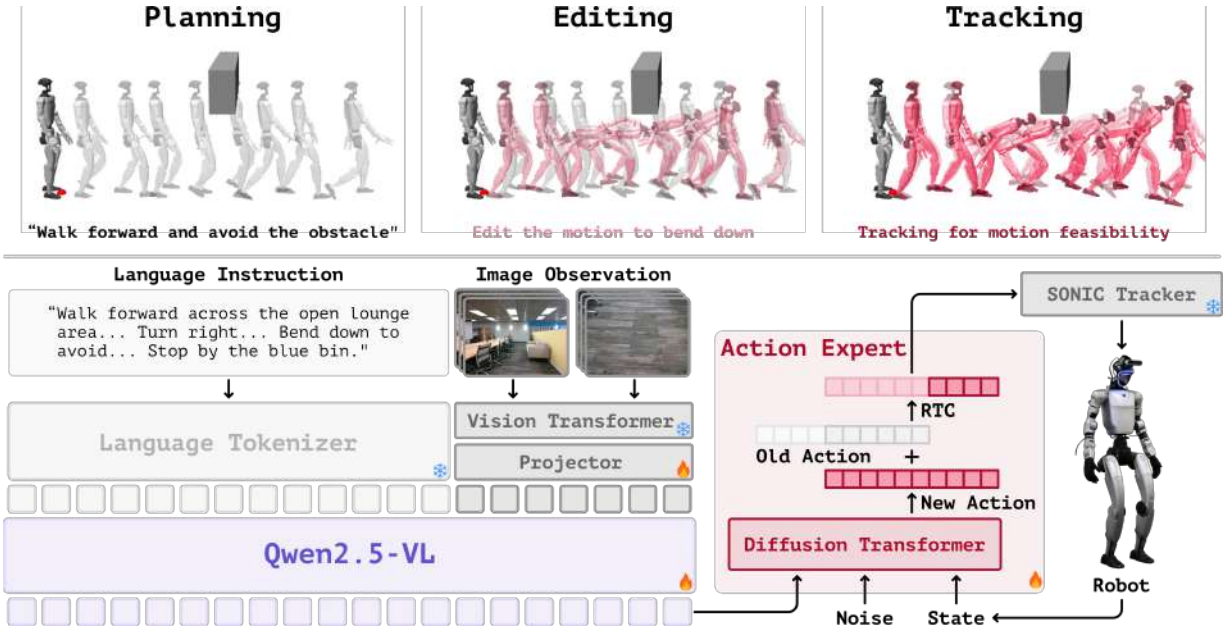

Fig. 2: **TANGO Architecture and Data Pipeline.** Top: the proposed Plan–Edit–Track (PET) pipeline automatically synthesizes collision-free whole-body traversal data. Bottom: **TANGO** combines a vision-language backbone, a diffusion-based action expert, and a low-level tracker for language-guided whole-body humanoid control.

sampled according to the pruning probability $P(t) = (1-\epsilon)e^{k(t-T)/T}+\epsilon,\ t \in [1,T]$, where ϵ and k regulate temporal intensity. The visual features $v_i = \text{VisionEncoder}(\mathbf{I}_i) \in \mathbb{R}^{n\times p\times c}$ are further spatial-grid pooled via $\tilde{v}_i = \text{GridPool}(v_i, g_i) \in \mathbb{R}^{g_i\times c}$ [43], allocating finer grids to recent observations and coarser grids to history. Finally, the sampled visual tokens $\tilde{v}_{\text{sampled}}$ and language instruction ℓ are fed into the VLM to produce a latent context token $z = \text{VLM}(\tilde{v}_{\text{sampled}}, \ell)$.

System-1 & System-0: Action Prediction and Execution. Conditioned on the latent z and current proprioception $\mathbf{q}_{d,t}$, the *system-1* action expert predicts a future whole-body reference chunk. While standard actions are defined as joint angles and base poses $\mathbf{a}_i = \{\mathbf{q}_{d,i}, \mathbf{r}_{b,i}\}$, we formulate a stabilized training target to facilitate regression $\tilde{\mathbf{a}}_i = \{\mathbf{q}_{d,i}, \tilde{\mathbf{r}}_{b,i}, \Delta x_i, \Delta y_i, \Delta\psi_i\}$, where the base yaw in $\tilde{\mathbf{r}}_{b,i}$ is parameterized relative to the first frame of the chunk, and the auxiliary deltas $(\Delta x_i, \Delta y_i, \Delta\psi_i)$ explicitly encode chunk-level planar displacement and heading changes. We implement *system-1* using a flow-based MM-DiT [44] trained via flow-matching to generate the horizon $\tilde{\mathbf{A}}_{t:t+H}$. To align offline training with online streaming execution, we apply training-time RTC [27], conditioning the model on a randomized committed prefix of d actions to inpaint the remaining horizon. The generated chunk is subsequently recovered to $\mathbf{A}_{t:t+H}$ and streamed to the SONIC tracker (*system-0*), which tracks the reference against robot proprioception to provide high-frequency joint commands.

Joint Training Objectives. Following dual-branch designs for VL decoding [45, 6], we append a text-decoding branch to System-2 and co-tune navigation tasks alongside VideoQA samples [46] to preserve generalized world knowledge. The joint optimization objective is defined as $\mathcal{L} = \mathcal{L}_{\text{CE}} + w_{\text{FM}} \cdot \mathcal{L}_{\text{FM}}$, where $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss for VideoQA, $\mathcal{L}_{\text{FM}}$ denotes the flow-matching loss, and $w_{\text{FM}} = 20$. **TANGO** is trained end-to-end for a single epoch with a learning rate of 1×10^{-5}.

D. Deployment

We design the deployment system of **TANGO** on a Unitree G1 in both simulation and real world, to ensure the effectiveness and robustness of our *triple-system* design in both scenarios.

Simulation Deployment. The task of whole-body navigation requires both high rendering quality and physical authenticity. To achieve both, we employ a digital twin teleportation system in simulation, where we leverage MuJoCo [47] for low-level tracker deployment and physical simulation, and teleport a humanoid digital twin in IsaacSim [37] to acquire high-fidelity visual observation from designated camera pose solved from a given humanoid robot pose using forward kinematics (FK), and use it as VLA input.

Real-World Deployment. As shown in Figure 3, **TANGO** adopts a cloud-edge deployment architecture that separates the compute-intensive VLA module from the high-frequency WBC module. The VLA (*system-2* and *system-1*) runs on a server equipped with an RTX 4090, while the SONIC tracker (*system-0*) runs on an onboard Jetson Orin NX. The humanoid captures front- and downward-facing RGB observations using RealSense D455 and D435i cameras and streams them, together with proprioceptive states, to the server with approximately 20ms latency. The server serves as a global clock and performs VLA inference every 0.5s, matching an execution horizon of $s = 15$ actions at 30Hz. Predicted motion chunks are streamed back to the robot, resampled to 50Hz, and

TABLE I: **VLNVerse benchmark result.** We evaluate TANGO on VLNVerse benchmark against strong baselines with different action representations. Among all methods, only TANGO is enabled with low-level physical control, while others are evaluated in a teleportation setting.

Methods	Low-level Control	Val Seen				Val Unseen			
		NE↓	OSR↑	SR↑	SPL↑	NE↓	OSR↑	SR↑	SPL↑
CMA [48]	✗	5.36	59.81	37.35	33.36	5.16	62.79	31.15	27.92
RDP [25]	✗	4.02	68.09	47.28	**41.69**	3.75	**71.93**	48.60	**42.72**
Seq2Seq [49]	✗	4.78	44.68	32.62	30.39	4.36	49.58	35.03	33.37
Ours	✓	**3.90**	**70.31**	**54.69**	40.18	**3.72**	71.07	**52.89**	40.18

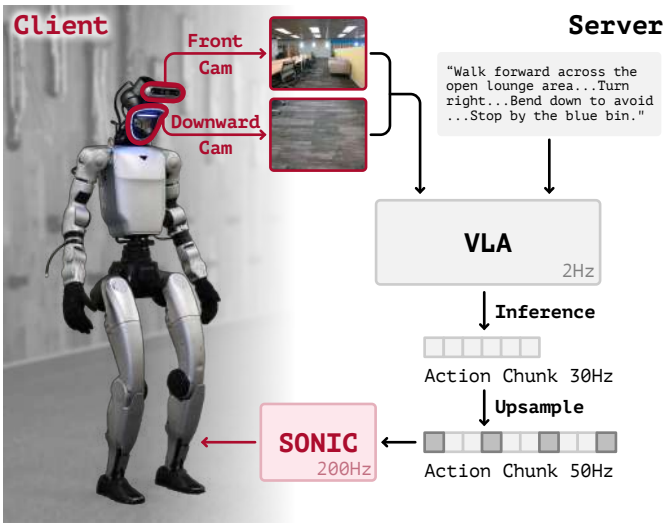

Fig. 3: **TANGO real-world deployment system.** We adopt a server-client design to separately run low-frequency VLA model and high-frequency whole-body tracker.

executed by the SONIC tracker, which closes the low-level control loop at approximately 200Hz.

IV. Experiments

To evaluate the effectiveness of our method, we conduct extensive experiments and ablation studies to answer three key questions: 1) Can TANGO perform well on VLN tasks compared to state-of-the-art baselines? 2) Can TANGO effectively learn to traverse through cluttered indoor environment without collision? and 3) Is the key design of our method effective?

A. VLN Performance

We evaluate TANGO on VLNVerse, a newly established VLN benchmark with photorealistic indoor scenes in IsaacSim. We compare against strong baselines in the benchmark, including discrete action model CMA [48] and Seq2Seq [49], and continuous action model RDP [25]. We evaluate all the methods on fine-grained task's validation split. We report Success Rate (SR), Success weighted by Path Length (SPL), Net Error (NE), and Oracle Success Rate (OSR). To ensure fair comparison, all the methods are trained on 3963 trajectories in fine-grained training split, and evaluated on 423/825 trajectories in seen/unseen validation splits. For metric calculation and other experiment setting details, please refer to the Appendix.

TABLE II: Performance comparison on cluttered VLNVerse scenes.

Methods	NE↓	SR↑	SPL↑	CR↓
InternVLA-N1†	5.34	26.67	11.92	19.03
InternVLA-N1‡	**3.19**	35.29	24.35	16.11
Ours	4.01	**43.75**	**31.83**	**9.90**

† Equipped with Unitree low-level controller and MPC.
‡ Tracked by HumanoidPF generalist policy

Since all baselines lack low-level control modules, we evaluate them in a teleportation setting following the original VLNVerse protocol. In contrast, TANGO is the only method equipped with low-level control and operates under realistic physical constraints. Nevertheless, as shown in Table I, TANGO consistently outperforms three strong baselines in terms of SR and NE on both the seen and unseen validation sets, while achieving SPL and OSR comparable to the state-of-the-art RDP baseline. These results highlight the potential of whole-body navigation methods in cluttered indoor environment traversal. The relatively lower SPL suggests reduced navigation efficiency, likely due to the conservative behavior of the low-level tracker, which may favor safer but less direct trajectories around obstacles.

B. Cluttered Environment Traversal Performance

We evaluate our method on the unseen split of our augmented VLNVerse scenes (Section III-B) to verify its obstacle-avoidance and spatial understanding capability in 3D environments. We compare our method with InternVLA-N1 in zero-shot setting, with two kinds of low-level executors: Unitree official RL controller [50] with a model predictive control (MPC) module, which executes basic movement according to planar velocity command; and HumanoidPF generalist policy [10] which performs obstacle-avoidance motions in complex 3D scenes. Note that while HumanoidPF leverages LiDAR as additional input for 3D scene information, our method uses pure RGB input.

To better quantify navigation safety in cluttered environments, we introduce Collision Rate (CR) [35], defined as the percentage of predicted trajectories that result in collision. As shown in Table II, TANGO achieves the highest SR and SPL while maintaining the lowest CR across all methods. In detail, TANGO reduces CR from 16.11% to 9.90% compared to the strongest baseline, despite relying solely on RGB observations,

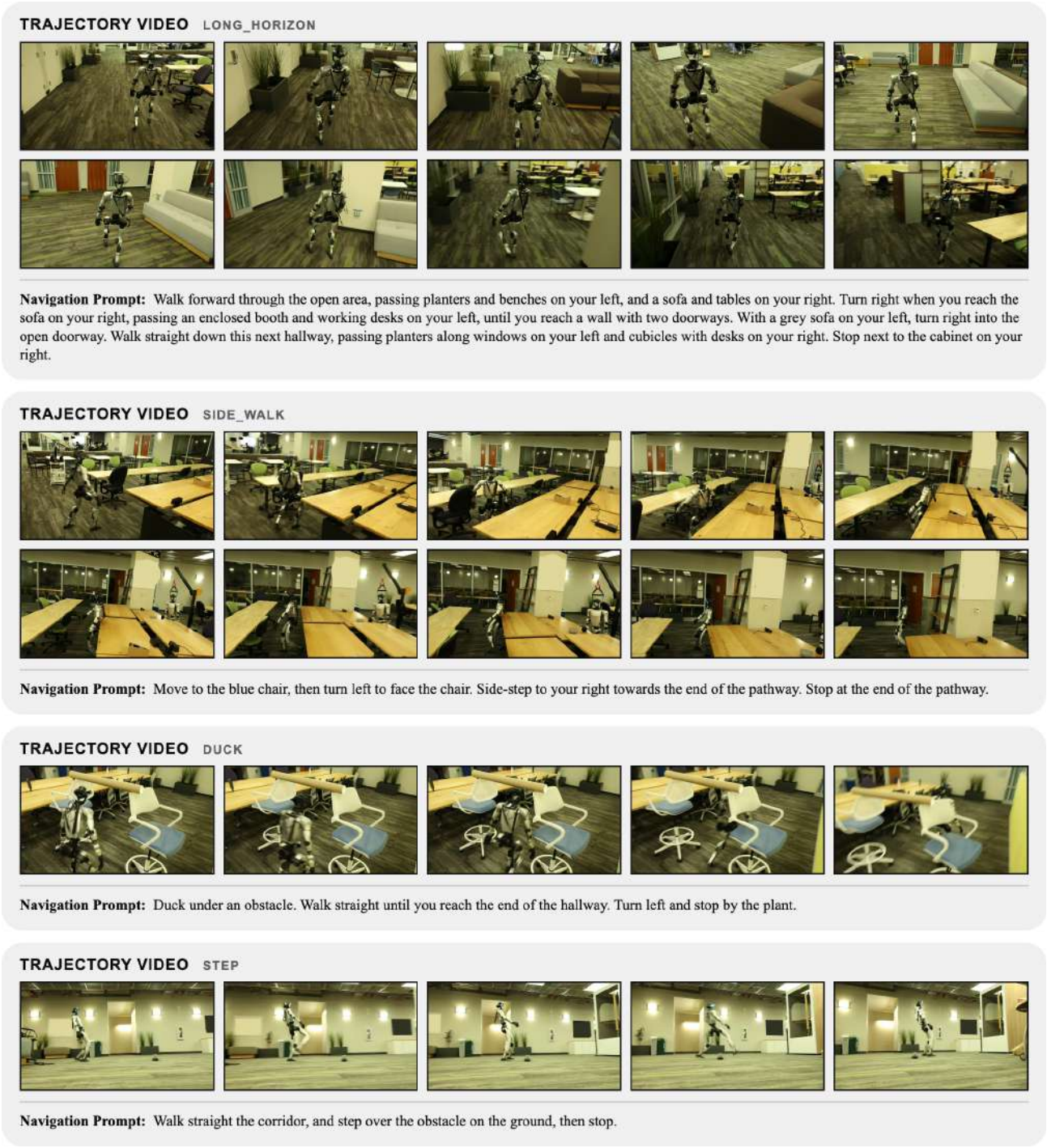

Fig. 4: **Real-world deployment of TANGO.** We show real-world qualitative results of TANGO in long-horizon navigation, side-stepping through narrow pathways, bending down to avoid overhead obstacles, and stepping over obstacles on the ground, demonstrating zero-shot sim-to-real transfer without any real-world training.

TABLE III: Comparison of navigation methods under fine-grained tasks.

Methods	Low-level Control	Val Unseen	
		SR↑	SPL↑
InternVLA-N1	✗	43.69	35.74
Ours-2D	✗	45.74	35.44
InternVLA-N1†	✓	42.37	22.50
Ours-2D†	✓	26.67	8.34
Ours	✓	**52.89**	**40.18**

† Equipped with Unitree low-level controller and MPC.

whereas the HumanoidPF-tracked baseline additionally has access to LiDAR-based geometric perception. These results suggest that end-to-end whole-body action generation is a more effective solution for cluttered-scene traversal than coupling a conventional navigation policy with a downstream obstacle-avoidance controller.

C. Real-World Experiment

We conduct real-world experiments to evaluate whether TANGO can transfer from simulation to a physical humanoid platform without any real-world training. In particular, we focus on scenarios that require simultaneous language-guided navigation, obstacle avoidance, and whole-body motion adaptation, which jointly test the key capabilities targeted by our approach.

As shown in Figure 4, we evaluate TANGO in four representative real-world scenarios: long-horizon navigation, side-stepping through a narrow pathway, bending down to avoid overhead obstacles, and stepping over obstacles on the ground. In all cases, TANGO demonstrates robust scene understanding and spatial traversal capability. The robot executes continuous whole-body motions while following natural-language instructions, adapts its body configuration to negotiate obstacles, and maintains progress toward the navigation goal, demonstrating zero-shot sim-to-real transfer in cluttered physical environments.

D. Ablation Studies

We study the impact of action representation and low-level execution on navigation performance to validate our 29-DoF whole-body action space. We compare against InternVLA-N1 in a zero-shot setting and a planar variant of our method, denoted as *Ours-2D*. Since planar trajectory prediction is an auxiliary objective in our formulation (Section III-C), *Ours-2D* isolates the effect of whole-body action generation. For both baselines, we evaluate under two execution settings: teleportation and physical execution using the Unitree low-level controller with MPC.

As shown in Table III, both planar policies experience a substantial performance drop when moving from teleportation to physical execution, highlighting the difficulty of transferring conventional navigation policies to embodied settings. In contrast, TANGO consistently outperforms all baselines and remains robust under low-level control constraints. These results demonstrate that whole-body action generation is critical for bridging high-level navigation and physical execution, and can be effectively learned through an end-to-end training pipeline.

V. Conclusion

This work introduces **TANGO**, to our best knowledge, the first whole-body vision-language navigation framework that directly predicts 29-DoF joint-space actions. We build up a diverse large-scale dataset covering not only indoor navigation patterns but also whole-body collision avoidance prior. Given language instruction, image observations and robot proprioception, we train a Qwen2.5VL-7B [42] based backbone with flow matching action expert. To match real-time execution on humanoid robot, training-time RTC [27] and high-frequency general tracker [1] are leveraged.

Our experiment results indicate both state-of-the-art vision-language navigation capability and decent traversal performance in cluttered environments. Through ablation study, we specifically prove the effectiveness of learning whole-body action space end-to-end.

Overall, **TANGO** represents a meaningful step toward practical whole-body large planning models by demonstrating the feasibility of directly predicting 29-DoF actions. Building on **TANGO** as a foundation model, future work can further adapt and scale this framework toward more general loco-manipulation foundation models capable of tackling challenging tasks that require coordinated and active use of the robot's entire body.

Limitation. Despite the promising results, the capability of low-level tracker shows up as a key constraint for further deployment in more complex environments, *e.g.* walking up stairs. Another limitation lies in our vision input, which relies solely on RGB images. This may restrict the model's ability to fully understand complex scenes, particularly in cluttered, visually ambiguous, or low-light environments, where depth camera and LiDAR are expected to help. We leave these for future work to explore.

References

[1] Z. Luo, Y. Yuan, T. Wang, C. Li, S. Chen, F. Castañeda, Z.-A. Cao, J. Li, D. Minor, Q. Ben, X. Da, R. Ding, C. Hogg, L. Song, E. Lim, E. Jeong, T. He, H. Xue, W. Xiao, Z. Wang, S. Yuen, J. Kautz, Y. Chang, U. Iqbal, L. J. Fan, and Y. Zhu, "Sonic: Supersizing motion tracking for natural humanoid whole-body control," *arXiv preprint arXiv:2511.07820*, 2025.

[2] Y. Ze, S. Zhao, W. Wang, A. Kanazawa, R. Duan, P. Abbeel, G. Shi, J. Wu, and C. K. Liu, "Twist2: Scalable, portable, and holistic humanoid data collection system," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2026.

[3] Q. Liao, T. E. Truong, X. Huang, Y. Gao, G. Tevet, K. Sreenath, and C. K. Liu, "Beyondmimic: From motion tracking to versatile humanoid control via guided diffusion," *arXiv preprint arXiv:2508.08241*, 2025.

[4] A. Allshire, H. Choi, J. Zhang, D. McAllister, A. Zhang, C. M. Kim, T. Darrell, P. Abbeel, J. Malik, and A. Kanazawa, "Visual imitation enables contextual humanoid control," in *Proceedings of the Conference on Robot Learning (CoRL)*, 2025.

[5] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. van den Hengel, "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[6] J. Zhang, A. Li, Y. Qi, M. Li, J. Liu, S. Wang, H. Liu, G. Zhou, Y. Wu, X. Li *et al.*, "Embodied navigation foundation model," in *The 14th International Conference on Learning Representations*, 2026.

[7] J. Zhang, K. Wang, R. Xu, G. Zhou, Y. Hong, X. Fang, Q. Wu, Z. Zhang, and H. Wang, "Navid: Video-based vlm plans the next step for vision-and-language navigation," *Robotics: Science and Systems*, 2024.

[8] N. Hirose, C. Glossop, A. Sridhar, D. Shah, O. Mees, and S. Levine, "Lelan: Learning a language-conditioned navigation policy from in-the-wild video," in *Conference on Robot Learning*, 2024.

[9] M. Wei, C. Wan, J. Peng, X. Yu, Y. Yang, D. Feng, W. Cai, C. Zhu, T. Wang, J. Pang, and X. Liu, "Ground slow, move fast: A dual-system foundation model for generalizable vision-language navigation," in *The 14th International Conference on Learning Representations*, 2026.

[10] H. Xue, S. Liang, Z. Zhang, Z. Zeng, Y. Liu, Y. Lian, J. Wang, Q. Liu, X. Shi, and Y. Li, "Collision-free humanoid traversal in cluttered indoor scenes," *arXiv preprint arXiv:2601.16035*, 2026.

[11] T. Xu, J. Chen, J. Zhang, W. Zhang, Z. Qi, M. Li, Z. Zhang, and H. Wang, "Mm-nav: Multi-view vla model for robust visual navigation via multi-expert learning," *arXiv preprint arXiv:2510.03142*, 2025.

[12] D. Shah, A. Sridhar, N. Dashora, K. Stachowicz, K. Black, N. Hirose, and S. Levine, "Vint: A foundation model for visual navigation," in *Conference on Robot Learning (CoRL)*, 2023.

[13] D. Shah, B. Osinski, B. Ichter, and S. Levine, "Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action," in *Conference on Robot Learning (CoRL)*, 2023.

[14] N. Hirose, C. Glossop, D. Shah, and S. Levine, "Omnivla: An omni-modal vision-language-action model for robot navigation," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2026.

[15] S. Wei, H. Jing, B. Li, Z. Zhao, J. Mao, Z. Ni, S. He, J. Liu, X. Liu, K. Kang, S. Zang, W. Yuan, M. Pavone, D. Huang, and Y. Wang, "ψ_0: An open foundation model towards universal humanoid loco-manipulation," *Robotics: Science and Systems*, 2026.

[16] GEAR Team, "GR00T N1.6: An improved open foundation model for generalist humanoid robots," Dec. 2025, nVIDIA Research Blog.

[17] H. Jiang, J. Chen, Q. Bu, L. Chen, M. Shi, Y. Zhang, D. Li, C. Suo, C. Wang, Z. Peng, and H. Li, "Wholebodyvla: Towards unified latent vla for whole-body loco-manipulation control," in *The 14th International Conference on Learning Representations*, 2026.

[18] Z. Wu, X. Huang, L. Yang, Y. Zhang, X. Chen, P. Abbeel, R. Duan, A. Kanazawa, C. Sferrazza, G. Shi, and C. K. Liu, "Perceptive humanoid parkour: Chaining dynamic human skills via motion matching," *Robotics: Science and Systems*, 2026.

[19] G. Zhou, Y. Hong, and Q. Wu, "Navgpt: Explicit reasoning in vision-and-language navigation with large language models," in *The AAAI Conference on Artificial Intelligence*, 2024.

[20] G. Zhou, Y. Hong, Z. Wang, X. E. Wang, and Q. Wu, "Navgpt-2: Unleashing navigational reasoning capability for large vision-language models," in *European Conference on Computer Vision (ECCV)*, 2024.

[21] B. Chandaka, G. X. Wang, H. Chen, H. Che, A. J. Zhai, and S. Wang, "Human-like navigation in a world built for humans," in *Conference on Robot Learning (CoRL)*, 2025.

[22] N. Rajabi and J. Kosecka, "Travel: Training-free retrieval and alignment for vision-and-language navigation," *arXiv preprint arXiv:2502.07306*, 2025.

[23] 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," in *Robotics: Science and Systems*, 2025.

[24] A. Li, Z. Wang, J. Zhang, M. Li, Y. Qi, Z. Chen, Z. Zhang, and H. Wang, "Urbanvla: A vision-language-action model for urban micromobility," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2026.

[25] L. Wang, X. Xia, H. Zhao, H. Wang, T. Wang, Y. Chen, C. Liu, Q. Chen, and J. Pang, "Rethinking the embodied gap in vision-and-language navigation: A holistic study

of physical and visual disparities," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

[26] N. Hirose, C. Glossop, D. Shah, and S. Levine, "Asyncvla: An asynchronous vla for fast and robust navigation on the edge," *arXiv preprint arXiv:2602.13476*, 2026.

[27] K. Black, A. Z. Ren, M. Equi, and S. Levine, "Training-time action conditioning for efficient real-time chunking," *arXiv preprint arXiv:2512.05964*, 2025.

[28] Z. Zhuang, S. Yao, and H. Zhao, "Humanoid parkour learning," in *Conference on Robot Learning (CoRL)*, 2024.

[29] Y. Chen, J. Ma, Z. Luo, Y. Han, Y. Dong, B. Xu, and P. Lu, "Learning autonomous and safe quadruped traversal of complex terrains using multi-layer elevation maps," *IEEE Robotics and Automation Letters*, vol. 10, no. 10, pp. 9606–9613, 2025.

[30] J. Liu, T. Xu, J. Chen, L. Yue, J. Zhang, Z. Wang, M. Li, Q. Zhao, A. Li, Q. Su, Z. Zhang, and H. Wang, "Spannav: Generalized spatial awareness for versatile vision-language navigation," *arXiv preprint arXiv:2603.09163*, 2026.

[31] J. Li, X. Cheng, T. Huang, S. Yang, R.-Z. Qiu, and X. Wang, "Amo: Adaptive motion optimization for hyper-dexterous humanoid whole-body control," *Robotics: Science and Systems*, 2025.

[32] H. Xue, X. Huang, D. Niu, Q. Liao, T. Kragerud, J. T. Gravdahl, X. B. Peng, G. Shi, T. Darrell, K. Sreenath, and S. Sastry, "Leverb: Humanoid whole-body control with latent vision-language instruction," *arXiv preprint arXiv:2506.13751*, 2025.

[33] P. Ding, J. Ma, X. Tong, B. Zou, X. Luo, Y. Fan, T. Wang, H. Lu, P. Mo, J. Liu, Y. Wang, H. Zhou, W. Feng, J. Liu, S. Huang, and D. Wang, "Humanoid-vla: Towards universal humanoid control with visual integration," *arXiv preprint arXiv:2502.14795*, 2025.

[34] W. Qin, S. Wu, C. Chen, M. Liu, L. Feng, X. Cui, H. Han, and H. Wang, "Physiflow: Physics-aware humanoid whole-body vla via multi-brain latent flow matching and robust tracking," *arXiv preprint arXiv:2603.05410*, 2026.

[35] S. Lin, Z. Li, X. Zhao, G. Zhou, L. Wang, R. Wei, R. Tang, J. Li, H. Wang, J. Pang, A. van den Hengel, J. Liu, and Q. Wu, "Vlnverse: A benchmark for vision-language navigation with versatile, embodied, realistic simulation and evaluation," *arXiv preprint arXiv:2512.19021*, 2025.

[36] B. Miao, R. Wei, Z. Ge, X. sun, S. Gao, J. Zhu, R. Wang, S. Tang, J. Xiao, R. Tang, and J. Li, "Towards physically executable 3d gaussian for embodied navigation," in *The 14th International Conference on Learning Representations*, 2026.

[37] NVIDIA, "Isaac sim," 2025. [Online]. Available: https://github.com/isaac-sim/IsaacSim

[38] G. Comanici *et al.*, "Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities," *arXiv preprint arXiv:2507.06261*, 2025.

[39] G. B. Margolis, M. Wang, N. Fey, and P. Agrawal, "Softmimic: Learning compliant whole-body control from examples," *arXiv preprint arXiv:2510.17792*, 2025.

[40] P. Intelligence, K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, M. Y. Galliker, D. Ghosh, L. Groom, K. Hausman, B. Ichter, S. Jakubczak, T. Jones, L. Ke, D. LeBlanc, S. Levine, A. Li-Bell, M. Mothukuri, S. Nair, K. Pertsch, A. Z. Ren, L. X. Shi, L. Smith, J. T. Springenberg, K. Stachowicz, J. Tanner, Q. Vuong, H. Walke, A. Walling, H. Wang, L. Yu, and U. Zhilinsky, "$\pi_{0.5}$: a vision-language-action model with open-world generalization," *arXiv preprint arXiv:2504.16054*, 2025.

[41] NVIDIA, :, J. Bjorck, F. Castañeda, N. Cherniadev, X. Da, R. Ding, L. J. Fan, Y. Fang, D. Fox, F. Hu, S. Huang, J. Jang, Z. Jiang, J. Kautz, K. Kundalia, L. Lao, Z. Li, Z. Lin, K. Lin, G. Liu, E. Llontop, L. Magne, A. Mandlekar, A. Narayan, S. Nasiriany, S. Reed, Y. L. Tan, G. Wang, Z. Wang, J. Wang, Q. Wang, J. Xiang, Y. Xie, Y. Xu, Z. Xu, S. Ye, Z. Yu, A. Zhang, H. Zhang, Y. Zhao, R. Zheng, and Y. Zhu, "Gr00t n1: An open foundation model for generalist humanoid robots," *arXiv preprint arXiv:2503.14734*, 2025.

[42] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin, "Qwen2.5-vl technical report," *arXiv preprint arXiv:2502.13923*, 2025.

[43] J. Zhang, K. Wang, S. Wang, M. Li, H. Liu, S. Wei, Z. Wang, Z. Zhang, and H. Wang, "Uni-navid: A video-based vision-language-action model for unifying embodied navigation tasks," *Robotics: Science and Systems*, 2025.

[44] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, D. Podell, T. Dockhorn, Z. English, K. Lacey, A. Goodwin, Y. Marek, and R. Rombach, "Scaling rectified flow transformers for high-resolution image synthesis," *arXiv preprint arXiv:2403.03206*, 2024.

[45] S. Wang, J. Zhang, M. Li, J. Liu, A. Li, K. Wu, F. Zhong, J. Yu, Z. Zhang, and H. Wang, "Trackvla: Embodied visual tracking in the wild," in *Conference on Robot Learning (CoRL)*, 2025.

[46] X. Shen, Y. Xiong, C. Zhao, L. Wu, J. Chen, C. Zhu, Z. Liu, F. Xiao, B. Varadarajan, F. Bordes, Z. Liu, H. Xu, H. J. Kim, B. Soran, R. Krishnamoorthi, M. Elhoseiny, and V. Chandra, "Longvu: Spatiotemporal adaptive compression for long video-language understanding," in *The 42nd International Conference on Machine Learning*, 2025.

[47] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Sys-*

tems, 2012, pp. 5026–5033.

[48] X. Wang, Q. Huang, A. Celikyilmaz, J. Gao, D. Shen, Y.-F. Wang, W. Y. Wang, and L. Zhang, "Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation," *arXiv preprint arXiv:1811.10092*, 2018.

[49] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *arXiv preprint arXiv:1409.3215*, 2014.

[50] Unitree Robotics, "Unitree RL Gym: Reinforcement learning implementation for unitree robots," https://github.com/unitreerobotics/unitree_rl_gym, 2025, accessed: 2026-05-29; commit 276801e46c5d433564f24658bac64f254b7d2d4b.

CONTENTS

VI. Project Page and Supplementary Results

The main paper includes representative real-world trajectory galleries showing that TANGO can perform zero-shot deployment on a Unitree G1 humanoid, including long-horizon language-guided navigation, cluttered-scene traversal, narrow-passage side-stepping, ducking, and stepping-over behaviors. This appendix provides additional details on environment augmentation, motion generation, evaluation metrics, and training. More real-world videos and qualitative examples are available on our project page.

We also provide a fully anonymous project page link here: https://anonymous-proj-2026.github.io/. It contains video demonstrations of the real-world experiments, additional qualitative examples, and visualizations of TANGO's whole-body traversal behaviors. These materials are intended to complement the quantitative results in the main paper and provide a more comprehensive view of how TANGO transfers from simulation to real humanoid deployment.

VII. Environment Augmentation and Motion Generation Details

A. Environment Augmentation Details

To better evaluate whole-body planning, we augment VLNVerse scenes with trajectory-conditioned obstacles rather than randomly sampled clutter. Given an original scene, its A* navigation trajectories, and the corresponding occupancy map extracted from the scene, we insert obstacles along the paths at the locations where whole-body behaviors are likely required. We consider three targeted interaction types: *stride*, where a low obstacle is placed on the path to encourage stepping over; *sidle*, where a pair of side obstacles forms a narrow passage; and *squat*, where an overhead or torso-height obstacle encourages ducking or lowering the body.

For each scene, candidate placements are sampled along valid A* trajectories while avoiding the start and goal regions. Each candidate is aligned with the local path direction so that the inserted obstacle naturally interacts with the robot's intended route. The obstacle assets are selected according to the target behavior: low objects such as rugs, cushions, or stools for *stride*; chairs, plants, shelves, or similar objects for *sidle*; and ceiling lights, lamps, curtains, or wall-mounted objects for *squat*. This design makes the augmented scenes semantically plausible while explicitly inducing whole-body navigation challenges.

To preserve scene validity, each placement is checked against the occupancy map and all available trajectories in the same scene. In particular, obstacles are prevented from blocking unrelated paths or appearing too close to trajectory endpoints, while the source trajectory is allowed to be affected by the inserted obstacle. After placement, the augmented obstacles are also written into the occupancy representation so that downstream planners observe the same geometry as the policy. Using this pipeline, we generate augmented scenes for all 263 VLNVerse scenes. Please refer to Figure 5 for a visualization of the augmented scenes.

To better align instructions and action labels to facilitate training, we further augment the language prompts from VLNVerse with the knowledge of added obstacles and their corresponding target behavior types via Gemini 2.5 Flash [38].

B. Motion Generation Details

This section describes the Plan, Edit, Track (PET) pipeline used to generate the trajectories that supervise TANGO. Given an augmented indoor scene and a start–goal pair, PET produces a 29-DoF whole-body reference motion, then re-renders egocentric observations and regenerates language instructions. The pipeline consists of three stages, described in Appendix VII-B (Plan), Appendix VII-B (Edit, which contains the motion editing deferred from the main text), and Appendix VII-B (Track).

***Plan*: Path Planning and Reference-Gait Synthesis** Given the start s, the goal g, and the scene occupancy map, which already includes the augmented obstacles , we plan a planar path with A* on the 2D floor grid. To favor safer routes with whole-body clearance, we add an obstacle-aware term to the step cost. This term is computed from the 2D signed distance to occupied cells, $\Phi_{\mathrm{2D}}(\mathbf{x})$:

$$c(\mathbf{x}) \;=\; c_{\text{step}} \;+\; \lambda \exp\!\big(-\Phi_{\mathrm{2D}}(\mathbf{x})/d_0\big), \qquad (1)$$

so cells near obstacles receive a soft penalty that decays with clearance d_0, while the admissible Euclidean-to-goal heuristic

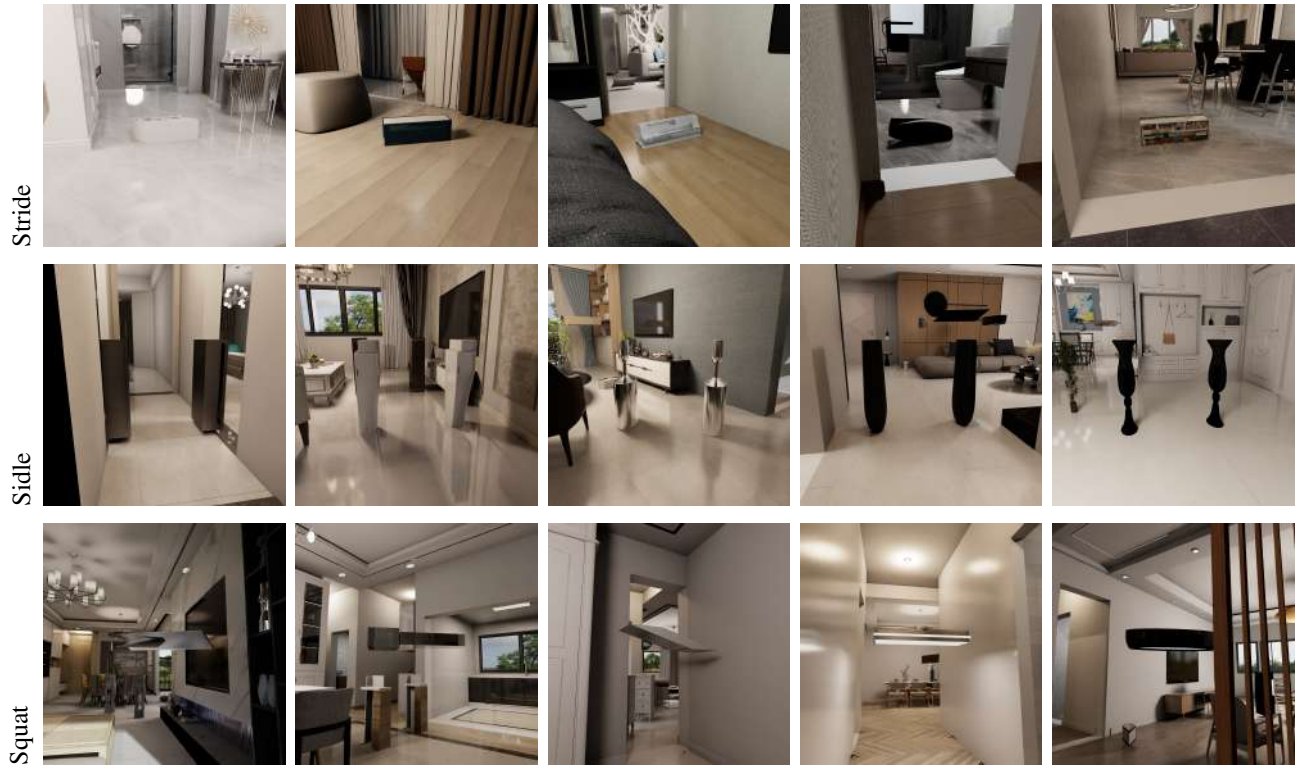

Fig. 5: **Examples of targeted environment augmentation.** Examples of augmented VLNVerse scenes with three obstacle types: low obstacles for stepping over, side obstacles for sidling through narrow passages, and overhead obstacles for squatting.

preserves optimality under this cost. The output is a planar polyline $\mathcal{P} = \{(x_k, y_k)\}_{k=1}^{K}$ with a tangent heading $\psi_k = \mathrm{atan2}(y_{k+1} - y_k,\, x_{k+1} - x_k)$.

To enable side walking through narrow passages, we used a rule-based yaw rewriter. Along $\mathcal{P}$, we detect narrow passages by probing the lateral free width $w_\perp(s)$ on both sides of the path at arc length s. When $w_\perp(s)$ falls below a threshold $w_{\min}$, the body heading is rewritten to be 90° offset from the travel tangent, $\psi_k^{\text{body}} = \psi_k \pm \frac{\pi}{2}$, with the sign chosen so that the leading shoulder remains on the wider side. Smooth $\pm 90^\circ$ transitions into and out of the passage produce a crab-walking, or side-stepping, gait in which the direction of translation and the facing direction are decoupled.

The rewritten planar trajectory $\{(x_k, y_k, \psi_k^{\text{body}})\}$ is converted into a stream of base velocity commands $\mathbf{u}_k = (v_x, v_y, \omega)_k$, where the forward and lateral linear velocities and yaw rates are obtained by finite-differencing the path under the rewritten heading. These commands are fed to SONIC [1], a motion-tracking model trained on $\sim$800 hours of high-quality motion-capture data, which produces a natural whole-body gait. A PD controller corrects drift between the realized base pose and $\mathcal{P}$, yielding a dynamically feasible reference walk $\mathcal{M}^{\text{ref}} = \{\mathbf{q}_t^{\text{ref}}\}_{t=1}^{T}$ with $\mathbf{q}_t^{\text{ref}} \in \mathbb{R}^{29}$ joint angles plus the floating-base pose. This gait is smooth, but above the floor it remains obstacle-agnostic: it follows the planar route and side-walks through narrow passages, but it does not yet step over, duck under, or otherwise adapt the body to 3D obstacle geometry. These adaptations are introduced in the Edit stage.

***Edit*: Whole-Body Obstacle-Avoidance Editing** The Edit stage transforms the reference motion $\mathcal{M}^{\text{ref}}$ into a whole-body obstacle-avoidance motion. The reference gait is kept as a locomotion prior: its timing, gait phase, and motion style are preserved, and edits are applied locally only where the path corridor contains obstacles. At each frame, we solve a whole-body inverse-kinematics (IK) problem with four objectives: reference posture tracking, foot contact and landing targets, a center-of-mass (CoM) target, and potential-field (PF) link forces. These objectives cover the three behavior families introduced above. A potential-field crouch controller handles squat obstacles (Appendix VII-B and Appendix VII-B), a gait-adaptation module handles stride obstacles (Appendix VII-B), and crab-walking from the Plan stage, together with lateral PF guidance, handles sidle passages.

The scene voxel occupancy is converted into a 3D signed distance field $\Phi(\mathbf{p})$ using the fast marching method, which provides distance values and gradients $\nabla\Phi$ throughout the workspace. Thin structural obstacles, such as bars, are also represented as exact oriented boxes with an analytic SDF, so that clearance queries remain continuous. Editing is restricted to a spatial corridor, defined as a tube of half-width w_{corr} ($\approx 0.5\,\text{m}$) around the planned path. Every field sample whose

horizontal position lies outside the tube is treated as free space,

$$\tilde{\Phi}(\mathbf{p}) = \begin{cases} \Phi(\mathbf{p}), & \mathrm{dist}_{xy}(\mathbf{p}, \mathcal{P}) \le w_{\mathrm{corr}}, \\ +\infty, & \text{otherwise}, \end{cases} \tag{2}$$

and the guidance vector outside the tube is zeroed. As a result, the editor ignores clutter that A* has already routed around, and only on-path obstacles that the robot must negotiate affect the edit.

Following HumanoidPF [10], we use a sampled repulsive field for avoidance and pass this guidance to the IK through SoftMimic-style pseudo-forces [39]. Instead of imposing hard collision constraints, each force is converted into a bounded soft displacement target that the whole-body IK balances against posture and stability terms. We compute a guidance field $\mathbf{g}(\mathbf{p})$ by combining the path tangent $\hat{\boldsymbol{\tau}}$, which encourages forward progress, with the SDF repulsion $\nabla\tilde{\Phi}$, which encourages obstacle avoidance. The weights are chosen so that repulsion dominates near obstacles:

$$\mathbf{g}(\mathbf{p}) \;=\; \hat{\boldsymbol{\tau}}(\mathbf{p}) \;+\; \beta\, e^{-\tilde{\Phi}(\mathbf{p})/\sigma}\, \frac{\nabla\tilde{\Phi}(\mathbf{p})}{\|\nabla\tilde{\Phi}(\mathbf{p})\|}. \tag{3}$$

We apply forces to a set of key body links $\mathcal{L}$, including the shoulders, elbows, wrists, torso, and optionally the knees and pelvis. Feet and ankles are excluded so that the contact pattern is still determined by the reference gait and the gait-adaptation module. For each link $\ell \in \mathcal{L}$ at world position $\mathbf{p}_\ell$, we sample a repulsive force $\mathbf{f}_\ell = F_{\max}\,\mathbf{g}(\mathbf{p}_\ell)$ and convert it into a displacement under a link stiffness κ, clipped to a per-link cap $\delta_{\max}$:

$$\Delta\mathbf{p}_\ell = \mathrm{clip}\Big(\tfrac{1}{\kappa}\,\mathbf{f}_\ell,\; \delta_{\max}\Big), \qquad \tilde{\mathbf{f}}_\ell = \kappa\,\Delta\mathbf{p}_\ell. \tag{4}$$

The bounded pseudo-forces $\{\tilde{\mathbf{f}}_\ell\}$ are added as soft tasks to the whole-body IK and are temporally low-pass filtered to remove discontinuities caused by field changes at gait transitions. For squat obstacles, we keep only the vertical component of each force, so the field provides a body-lowering cue for ducking rather than a lateral push. For sidle passages, the horizontal components push the trunk and arms away from the side walls in coordination with the crab-walking heading. The per-link forces can also be aggregated into a common CoM offset, with only the residual deviation applied to each link, so that the body moves away from obstacles as a coherent whole rather than having each limb react independently.

Two additional terms make ducking under squat obstacles start early and remain stable. The look-ahead term addresses the fact that the raw field can stay near zero until the body is already under a ceiling. We probe $\tilde{\Phi}$ ahead of the upper body at several head and shoulder heights and at several forward distances. If the minimum probe lies within a margin η of the geometry, a downward force proportional to the violation is added, so the body begins to lower before reaching the obstacle. The virtual head barrier protects a point $\mathbf{p}_{\text{head}} = \mathbf{p}_{\text{torso}} + h\hat{z}$ with clearance $b = \tilde{\Phi}(\mathbf{p}_{\text{head}}) - r$. When $b < b_m$, it adds a strictly downward force $k_s(b_m - b) + k_h \max(0, -b)$, corresponding to soft and hard barrier components, clipped to a maximum trunk displacement. The resulting downward displacement of the upper body is converted into a coordinated squat by coupling the crouch activation $\alpha = \mathrm{clip}(\Delta z/\delta_{\max}, 0, 1)$ to posture: a forward waist-pitch target $\theta^{\text{waist}} \leftarrow \theta^{\text{waist}} + \alpha\,\theta_0$, a proportional hip-pitch bias, and a downward shift of the CoM target $\Delta z_{\text{CoM}} = \alpha\, c_z$. This CoM shift lets the legs lower the body rather than letting the CoM task pull the body back upright. The implementation supports three coordinated presets: waist-only, which bends at the waist while the tracker controls the legs; com-drop, which lowers the body vertically; and full-squat, which combines waist motion, pelvis tilt, hip flexion, knee bending, ankle dorsiflexion, and CoM lowering. During an active crouch, the upper-body reference-tracking terms are relaxed so that the field can reshape the trunk instead of competing with the upright reference.

For stride obstacles, upper-body guidance is not enough because the swing feet must step over the object. The gait-adaptation module retargets footsteps online while preserving the phase and timing of the reference gait. At each foot lift-off, the base is projected onto the planned path, a short arc-length look-ahead is queried, and the landing target is specified in the path-yaw frame. This target is clamped to a maximum displacement from the reference touchdown, which prevents the IK solver from pursuing unreachable foot placements. A footprint-aware SDF scan detects a low obstacle ahead of the swing foot and returns the interval $[d_{\text{near}}, d_{\text{far}}]$ spanned by the obstacle, together with its top height z_{top}. We classify obstacles by height: only $z_{\text{top}} \le z_{\text{bar}}$, corresponding to a steppable bar, triggers a step-over motion. Taller geometry is treated as a wall or ceiling and is handled by the lateral and upper-body layers, so the robot does not attempt to step over a wall. For a steppable bar, the landing target is pushed past the far edge by a margin, $d^{\text{land}} = d_{\text{far}} + m$. The landing objective is evaluated with respect to this target rather than the reference touchdown on the bar. This both clears the obstacle and compensates for the forward-reach undershoot of the downstream tracker, keeping the realized landing past the bar.

The swing trajectory is formed by warping the reference foot trajectory from motion capture or RL data onto the retargeted endpoints and adding a vertical clearance arc. This keeps the natural swing shape and leaves ordinary steps unchanged. With swing phase $u \in [0, 1]$, after smoothstep interpolation to $\bar{u}$, and reference foot path $\mathbf{p}^{\text{ref}}(u)$,

$$\mathbf{p}^{\text{swing}}(u) = \mathbf{p}^{\text{ref}}(u) + (1 - \bar{u})\,\Delta\mathbf{p}_{\text{lift}} + \bar{u}\,\Delta\mathbf{p}_{\text{land}} + h_{\text{arc}}\,\rho(u)\,\hat{z}, \tag{5}$$

where $\Delta\mathbf{p}_{\text{lift}}, \Delta\mathbf{p}_{\text{land}}$ are the offsets between the retargeted and reference endpoints, and $\rho(u)$ is an asymmetric clearance profile with a fast quarter-sine rise to an early peak followed by a quarter-cosine fall. The asymmetric profile keeps the foot high early in the swing, which is needed for the trailing leg that crosses the bar shortly after lift-off. A symmetric $\sin(\pi u)$ arc would be too low at this stage. The peak height h_{arc} is solved for each step so that the full foot footprint clears the

obstacle box at every sampled phase,

$$z_{\text{ground}} + h_{\text{arc}}\,\rho(u) \geq z_{\text{top}}(u) + \epsilon, \quad \forall\, u : \text{footprint}(u) \cap \text{obstacle} \neq \varnothing, \tag{6}$$

using the same profile ρ in both the solver and the executed arc, so that the planned and executed motions remain consistent. A swing-knee bend bias encourages knee flexion rather than straight-leg extension for obstacle clearance. When several low obstacles appear in sequence, a short foothold sequence places natural-length steps through the gaps instead of requiring a single long step.

Because the foot retargeter moves the support polygon while the reference pelvis still follows the original gait, the body can lag behind the planned support. During a forward step-over, this mismatch can make the body lean in place rather than translate forward. Two small offsets, smoothed with EMA, correct this effect: root redirection shifts the pelvis reference toward the planned support, and CoM redirection shifts the IK CoM target accordingly. During a bar crossing, these offsets switch to a low-latency update rate so that the base translates with the crossing leg. This produces a forward step-over rather than an in-place lunge, with a ramped transition to avoid abrupt motion.

Each frame is solved in two passes. The first pass performs a lower-body projection that edits the reference to match the foot, support, and CoM targets. The second pass performs full compliant whole-body IK, which tracks the edited reference while applying the PF link forces. Strong foot-anchor costs are used so that the feet behave as contacts and posture edits do not induce foot sliding. The output is the edited 29-DoF whole-body trajectory $\mathcal{M}^{\text{edit}}$.

***Track*: Feasibility Restoration and Collision Filtering** The edited trajectory $\mathcal{M}^{\text{edit}}$ is kinematic and may violate dynamic feasibility, for example through overly fast swings or aggressive crouches. We restore feasibility by passing it through the SONIC tracker [1] in closed loop. The edited whole-body reference, including joint positions, joint velocities, and base orientation, is streamed frame by frame to the RL tracking policy, which tracks it under physics in MuJoCo and outputs a dynamically feasible motion $\mathcal{M}^{\text{track}}$. The same tracker is used at deployment (system-0), making the synthesized supervision consistent with execution. The known forward undershoot of the tracker is also why the Edit stage uses generous step-over landing margins.

VIII. Experiment Metrics

We evaluate navigation performance with standard vision-language navigation (VLN) metrics, including Success Rate (SR), Success weighted by Path Length (SPL), Navigation Error (NE), and Oracle Success Rate (OSR). For cluttered-scene traversal, we additionally report Collision Rate (CR) to quantify physical safety.

Let $\mathcal{E}$ denote the evaluation set and $|\mathcal{E}|$ its number of episodes. For each episode $e \in \mathcal{E}$, the policy receives a language instruction ℓ_e and sequential observations as defined in Section III-A, and executes whole-body action chunks through the low-level tracker. The resulting robot-base trajectory is denoted as

$$\Gamma_e = (\mathbf{x}_{e,0}, \mathbf{x}_{e,1}, \ldots, \mathbf{x}_{e,K_e}), \tag{7}$$

where $\mathbf{x}_{e,0}$ is the start position, $\mathbf{x}_{e,K_e}$ is the final stopping position, and K_e is the number of sampled trajectory points in episode e. Let $\mathbf{g}_e$ denote the target goal position, and let $d_{\mathcal{G}}(\cdot,\cdot)$ denote geodesic distance in the navigation environment. We use δ as the success threshold, set to 3 meters following standard VLN evaluation. We further denote the executed path length as

$$P_e = \sum_{k=1}^{K_e} \|\mathbf{x}_{e,k} - \mathbf{x}_{e,k-1}\|_2\,, \tag{8}$$

and the shortest-path distance from the start position to the goal as

$$L_e = d_{\mathcal{G}}(\mathbf{x}_{e,0}, \mathbf{g}_e). \tag{9}$$

a) Navigation Error (NE).: Navigation Error measures the average geodesic distance between the final stopping position and the target goal:

$$\text{NE} = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} d_{\mathcal{G}}(\mathbf{x}_{e,K_e}, \mathbf{g}_e). \tag{10}$$

Lower NE indicates that the agent stops closer to the target goal.

b) Success Rate (SR).: An episode is considered successful if the final stopping position is within the success threshold δ of the target goal. The success indicator is

$$S_e = \mathbf{1}\left[d_{\mathcal{G}}(\mathbf{x}_{e,K_e}, \mathbf{g}_e) \leq \delta\right], \tag{11}$$

and SR is computed as

$$\text{SR} = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} S_e. \tag{12}$$

In our tables, SR is reported as a percentage.

c) Oracle Success Rate (OSR).: Following common VLN evaluation, Oracle Success Rate measures whether the executed trajectory ever enters the goal region, regardless of the final stopping position. The oracle success indicator is

$$O_e = \mathbf{1}\left[\min_{0 \leq k \leq K_e} d_{\mathcal{G}}(\mathbf{x}_{e,k}, \mathbf{g}_e) \leq \delta\right], \tag{13}$$

and OSR is computed as

$$\text{OSR} = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} O_e. \tag{14}$$

OSR therefore evaluates whether the trajectory reaches the goal neighborhood at least once, rather than whether the agent stops there.

d) Success weighted by Path Length (SPL).: SPL jointly measures task completion and path efficiency:

$$\text{SPL} = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} S_e \frac{L_e}{\max(P_e, L_e)}. \tag{15}$$

Failed episodes contribute zero to SPL, while successful but unnecessarily long trajectories are penalized by the path-length ratio.

e) Collision Rate (CR).: For cluttered-scene traversal, we report Collision Rate as an episode-level safety metric. Let

$$C_e = \mathbf{1}\left[\Gamma_e \text{ results in at least one collision}\right]. \tag{16}$$

CR is then defined as

$$\mathrm{CR} = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} C_e. \tag{17}$$

Lower CR indicates safer whole-body traversal in cluttered environments. In our tables, CR is reported as a percentage.

IX. Training Details

For our training procedure, we emphasize rare whole-body behaviors through both data-level rebalancing and loss-level weighting. After converting rollouts into action chunks, we rebalance the training set by upsampling behaviorally important rows, including large turns, sideways motion, squat, stride, and mixed squat-stride segments; and especially, stride and sideways motions are boosted most strongly to compensate for their lower frequency and weaker action signal. We further apply motion-conditioned per-dimension loss weights on the action vector, assigning higher weights to the joint-position dimensions most associated with each behavior, such as hip, spine, and knee joints for squat, spine and shoulder joints for stride, their union for mixed motions, and shoulder plus body-frame motion dimensions for sideways motion.

The model is trained with full-parameter tuning on 16 nodes of 8 $\times$ NVIDIA A100 GPUs, reflecting a substantial training effort targeted specifically at improving whole-body loco-manipulation behaviors.