HUMANO: A LARGE-SCALE BENCHMARK FOR GENERALIST HUMANOID LOCO-MANIPULATION

Anonymous authors
Paper under double-blind review

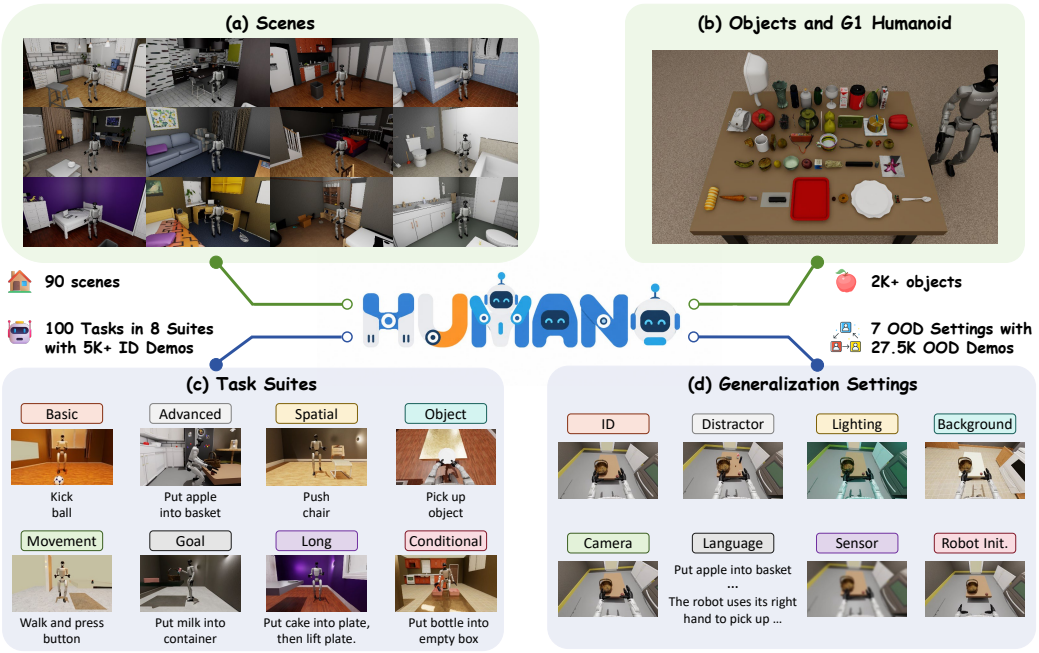


Figure 1: Overview of HUMANO. The benchmark contains 100 tasks across 8 capability suites, spanning 90 interactive scenes and over 2K object assets, together with 5.5K in-distribution demonstrations. It further provides seven controlled OOD settings with 27.5K additional demonstrations for systematic evaluation of multi-task capability and generalization in humanoid loco-manipulation.

ABSTRACT

Recent advances in vision-language-action models and whole-body control have rapidly expanded the capabilities of humanoid loco-manipulation, enabling generalist policies to perform increasingly diverse tasks with a single model. However, existing benchmarks often focus on individual tasks or small task suites, making it difficult to systematically evaluate multi-task capabilities and generalization. To fill this gap, we present HUMANO, a large-scale benchmark for generalist humanoid loco-manipulation. HUMANO contains 100 tasks organized into eight complementary suites, spanning 90 interactive scenes and over 2K object assets, with 5.5K in-distribution whole-body demonstrations. The benchmark covers fundamental skills, goal-directed loco-manipulation, generalization across spatial configurations, objects, robot configurations, and goals, as well as long-horizon skill composition and state-dependent reasoning. HUMANO further provides seven controlled OOD settings covering visual, linguistic, and robot-initialization shifts, together with 27.5K OOD demonstrations for future training and adaptation studies. All policies operate through a shared SONIC-based motion-latent interface, enabling standardized comparison under a common whole-body execution framework. Experiments with representative imitation and vision-language-action policies demonstrate the emerging potential of generalist models for multi-task humanoid loco-manipulation. Although multi-task learning remains more challenging than single-task training,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

performance continues to improve with increasing demonstration scale. Preliminary sim-to-real results further show that policies trained on HUMANO simulation data can transfer executable whole-body behaviors to a physical humanoid robot.

1 INTRODUCTION

Recent advances in vision-language-action models and whole-body control have substantially accelerated progress in humanoid loco-manipulation, enabling increasingly generalist policies to integrate perception, language understanding, locomotion, and manipulation within a single model (Bjorck et al., 2025; Intelligence et al., 2025; Wei et al., 2026a; Luo et al., 2026). This progress is moving humanoid robots beyond task-specific behaviors toward general-purpose systems capable of performing diverse whole-body tasks in complex environments. However, the evaluation of such systems still relies heavily on a relatively small number of real-robot tasks. While real-world evaluation is indispensable, it is costly, time-consuming, difficult to reproduce, and inherently limited in task and environment coverage. Simulation provides a complementary route for scalable and standardized evaluation, allowing policies to be tested efficiently across substantially larger task sets, diverse environments, and controlled distribution shifts. As humanoid policies become increasingly general, building such comprehensive simulation-based evaluation has therefore become an important prerequisite for measuring their true capabilities and generalization.

Existing humanoid simulation benchmarks have begun to address this need, but each captures only a subset of the capabilities required by generalist loco-manipulation policies. Some benchmarks primarily emphasize reinforcement learning (RL) and whole-body control (Sferrazza et al., 2024), while others expand task and scene diversity for imitation learning and generalist policies (Wei et al., 2026b). Recent efforts further study specialized aspects such as scalable demonstration generation (Lin et al., 2026) or coordination between high-level policies and low-level RL-based whole-body controllers (Wang et al., 2026). Nevertheless, existing benchmarks remain limited along one or more key dimensions, including task scale, capability diversity, systematic generalization, and standardized policy interfaces. As summarized in Table 1, a unified benchmark that brings these dimensions together is still missing, making it difficult to comprehensively compare the rapidly growing class of generalist humanoid loco-manipulation policies.

To bridge this gap, we introduce HUMANO, a large-scale benchmark for evaluating generalist humanoid loco-manipulation, as illustrated in Figure 1. HUMANO contains 100 tasks organized into eight complementary suites, spanning 90 interactive scenes and more than 2,000 object assets, together with 5.5k in-distribution whole-body demonstrations. The benchmark covers a broad spectrum of humanoid capabilities: *Basic* and *Advanced* evaluate fundamental whole-body skills and goal-directed loco-manipulation; *Spatial*, *Object*, *Movement*, and *Goal* probe generalization across spatial configurations, object variations, movement requirements, and task goals; and *Long-Horizon* and *Conditional* evaluate temporal skill composition and state-dependent decision making. Rather than increasing task count through minor variations, this structured design organizes distinct forms of whole-body behavior into a unified benchmark, enabling systematic evaluation of a single policy across diverse humanoid tasks.

Beyond task diversity, HUMANO provides seven controlled out-of-distribution (OOD) settings covering distractors, lighting, scene appearance, camera viewpoint, language variation, visual sensor corruption, and robot initialization. We additionally release 27.5K OOD demonstrations under these settings to support future training and adaptation studies. Under the standard HUMANO protocol, models are trained only on in-distribution data and evaluated directly under the corresponding OOD shifts. To further standardize policy comparison, HUMANO adopts the unified motion representation of SONIC (Luo et al., 2026) as the interface between high-level policy prediction and whole-body execution. Representative imitation policies and vision-language-action models therefore operate through the same motion-latent interface and share a common whole-body decoder, reducing confounding from heterogeneous low-level control stacks. We benchmark these models across all 100 tasks and seven OOD settings, and further study demonstration scaling, single-task versus multi-task learning, and sim-to-real transfer.

In summary, HUMANO provides a large-scale benchmark for generalist humanoid loco-manipulation, a systematic protocol for evaluating both multi-task competence and OOD generalization, and a unified motion-latent interface for controlled comparison of high-level policies.

Table 1: Comparison of representative humanoid loco-manipulation benchmarks in benchmark scale and generalization coverage. Generalization spans six categories: visual, language, spatial, semantic, sensing, and temporal. ✓ indicates explicit benchmark-level coverage, while – indicates that the corresponding aspect is not a primary component. For HUMANO, # Demos includes 5.5K in-distribution demonstrations and 27.5K additional OOD demonstrations.

					Generalization					
Benchmark	# Tasks	# Scenes	# Objects	# Demos	Visual	Lang.	Spatial	Semantic	Sensing	Temporal
HumanoidBench	27	–	20+	–	–	–	–	–	–	–
SIMPLE	60	50	1K+	6K	✓	✓	✓	✓	–	–
HumanoidMimicGen	9	–	–	9K	–	–	✓	✓	–	–
HumanoidArena	7	–	–	1.4K	✓	–	–	✓	–	✓
HUMANO (Ours)	**100**	**90**	**2K+**	**33K**	✓	✓	✓	✓	✓	✓

Through comprehensive evaluation and analysis, HUMANO establishes a standardized testbed for measuring progress toward more general and robust humanoid loco-manipulation.

2 Related Work

Humanoid Loco-Manipulation Models. Modern humanoid loco-manipulation systems commonly adopt a hierarchical architecture that separates high-level task-conditioned policy prediction from low-level whole-body execution. At the high level, generic imitation-learning methods such as ACT (Zhao et al., 2023) and Diffusion Policy (Chi et al., 2025) provide widely used action-prediction backbones, while generalist foundation policies such as $\pi_{0.5}$ (Intelligence et al., 2025) and GR00T N1 (Bjorck et al., 2025) scale policy learning across broader tasks and robot embodiments. More recent humanoid-specific models, including WholeBodyVLA (Jiang et al., 2026), Ψ_0 (Wei et al., 2026a), HEX (Bai et al., 2026), and OpenHLM (Hu et al., 2026), extend vision-language-action learning to coordinated locomotion and manipulation. In parallel, MotionWAM (Zheng et al., 2026), ω-0 (Li et al., 2026), and WholeBodyWAM Zhang et al. (2026) incorporate predictive motion or world dynamics into whole-body policy learning.

At the low level, whole-body control has progressed from motion imitation toward increasingly general motion tracking and interaction-aware execution. DeepMimic (Peng et al., 2018) and AMP (Peng et al., 2021) established influential learning paradigms for physics-based motion imitation, while BeyondMimic (Liao et al., 2026) further scales motion tracking and composition to diverse humanoid behaviors. HOMIE (Ben et al., 2025), AMO (Li et al., 2025a), and CLONE (Li et al., 2025b) extend whole-body control toward teleoperation and loco-manipulation, and SONIC (Luo et al., 2026) further scales motion tracking over large motion collections with a reusable motion representation. Despite progress at both levels, existing high-level policies are often coupled with different action representations and control stacks, making policy-level comparisons difficult to isolate. HUMANO therefore adopts the unified motion representation and whole-body decoder of SONIC as a common execution interface for heterogeneous high-level policies.

Humanoid Data Generation and Simulation. Scaling humanoid loco-manipulation requires large amounts of whole-body interaction data, yet collecting such trajectories directly on physical robots remains costly and labor-intensive. Early systems therefore rely heavily on human teleoperation, motion capture, and human-to-humanoid retargeting, as demonstrated by HumanPlus (Fu et al., 2024), OmniH2O (He et al., 2024), and TWIST (Ze et al., 2025). More recent work increasingly turns to simulation and automated generation to improve data scalability. HumanoidMimicGen (Lin et al., 2026) generates loco-manipulation demonstrations through whole-body planning and demonstration adaptation, GRAIL (Xie et al., 2026) synthesizes humanoid-object interactions from 3D assets and video priors, and OASIS (Yu et al., 2026) combines realistic asset reconstruction, simulated data collection, and visual randomization for sim-to-real transfer. These advances substantially reduce the cost of acquiring humanoid training data, while HUMANO is complementary in focusing on standardized evaluation across diverse tasks and controlled distribution shifts.

Humanoid Loco-Manipulation Benchmarks. General-purpose simulators such as RoboVerse (Geng et al., 2025) and LocoMuJoCo (Al-Hafez et al., 2023) support humanoid learning but are not designed specifically for generalist loco-manipulation evaluation. HumanoidBench (Sferrazza et al.,

2024) provides 27 locomotion and manipulation tasks, SIMPLE (Wei et al., 2026b) expands to 60 tasks with progressive domain randomization, HumanoidMimicGen (Lin et al., 2026) focuses on scalable demonstration generation, and HumanoidArena (Wang et al., 2026) evaluates perturbation robustness and policy-tracker transfer. In contrast, HUMANO combines scale with structured capability evaluation, comprising 100 tasks across eight complementary suites, 90 scenes, over 2,000 objects, 5.5K in-distribution demonstrations, seven controlled OOD settings, and 27.5K additional OOD demonstrations. This design supports systematic evaluation of multi-task competence, capability generalization, and robustness.

3 THE HUMANO BENCHMARK

HUMANO contains 100 tasks across eight capability suites, 90 interactive scenes, over 2,000 object assets, 5.5k in-distribution demonstrations, and 27.5K OOD demonstrations. It further provides seven controlled OOD settings and a unified SONIC-based motion-latent interface. Section 3.1 introduces the environment assets, Section 3.2 describes the task-suite design, Section 3.3 presents the generalization settings, Section 3.4 details whole-body demonstration collection and the unified policy interface, and Section 3.5 describes dataset construction and statistics.

3.1 ENVIRONMENT ASSETS

HUMANO builds its environments on the iTHOR scene collection distributed through MolmoSpaces (Kim et al., 2026). iTHOR contains 120 hand-crafted interactive indoor scenes, evenly divided among kitchens, living rooms, bedrooms, and bathrooms. Since HUMANO does not target bathroom-specific activities, we exclude the 30 bathroom scenes and retain the remaining 90 scenes, comprising 30 kitchens, 30 living rooms, and 30 bedrooms. These scenes provide diverse indoor layouts, furniture arrangements, and interactive structures for constructing humanoid loco-manipulation tasks.

HUMANO further draws object assets from Synthesis, MolmoSpaces, and Objaverse, yielding more than 2K indoor objects with diverse geometries and semantic categories. The asset pool includes both manipulable objects and scene-level furniture, supporting tasks involving grasping, placement, rearrangement, articulated interaction, and whole-body movement in cluttered indoor environments. Together, the 90 scenes and 2K+ object assets provide the environment basis for the task suites and controlled generalization settings introduced below.

3.2 TASK SUITE DESIGN

As shown in Figure 2, HUMANO builds on the environment assets described above to construct 100 humanoid loco-manipulation tasks organized into eight complementary suites. Rather than relying on minor task variations, the suites are structured around distinct capability dimensions, spanning fundamental skills, goal-directed manipulation, generalization, temporal composition, and conditional decision making. Complete definitions of all 100 tasks are provided in Appendix A.

Basic Suite. The *Basic* suite contains 10 tasks and evaluates fundamental whole-body skills, including grasping, pushing, posture changes, reaching, and lower-body interaction.

Advanced Suite. The *Advanced* suite contains 10 tasks and extends these primitives to goal-directed loco-manipulation involving object transport, placement, and coordinated body-object interaction.

Spatial Suite. The *Spatial* suite contains 20 tasks and evaluates generalization to variations in target layout and interaction height relative to the robot.

Object Suite. The *Object* suite contains 10 tasks and evaluates whether manipulation skills transfer across different object categories and geometries while preserving the underlying task semantics.

Movement Suite. The *Movement* suite contains 10 tasks and evaluates robustness to variations in whole-body movement requirements while preserving the underlying manipulation objective.

Goal Suite. The *Goal* suite contains 10 tasks and evaluates goal-conditioned generalization by varying the desired target or destination while maintaining comparable task contexts.

Long-Horizon Suite. The *Long-Horizon* suite contains 20 tasks and evaluates temporal composition by chaining multiple locomotion and manipulation primitives into extended task sequences.

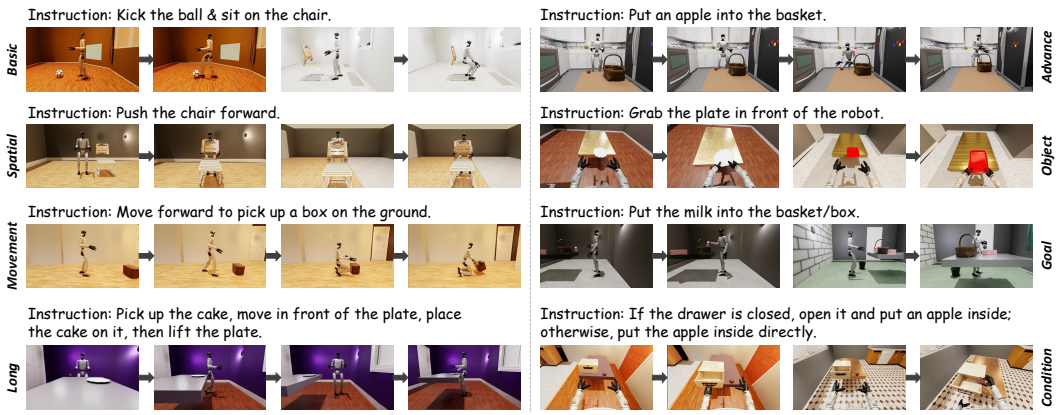


Figure 2: Representative tasks from the 8 HUMANO suites. Each row shows an example instruction and keyframes of the corresponding whole-body behavior, covering fundamental skills, loco-manipulation, generalization, long-horizon composition, and conditional reasoning.

Conditional Suite. The *Conditional* suite contains 10 tasks and evaluates state-dependent decision making, where the required action sequence changes according to the observed environment state.

3.3 Generalization Settings

As discussed in Section 3.2, all HUMANO task suites already incorporate generalization through variations in object spatial configuration, object identity, task goal, and temporal composition. To further strengthen generalization evaluation beyond these suite-specific variations, HUMANO introduces seven additional controlled out-of-distribution (OOD) settings that can be consistently applied across all eight task suites: *Distractor*, *Lighting*, *Background*, *Camera Viewpoint*, *Language*, *Sensor Noise*, and *Robot Initial State*. These settings probe complementary changes in visual appearance, linguistic expression, sensing quality, and robot initialization while preserving the underlying task objective. Detailed definitions and implementation details of the seven OOD settings are provided in Appendix B.

Distractor. We introduce one to five task-irrelevant objects near the interaction region while excluding objects that overlap semantically with task-relevant categories, testing robustness to visual and physical clutter.

Lighting. We vary illumination through changes in light color, spatial distribution, and source scale, evaluating robustness to diverse lighting conditions.

Background. We evaluate transfer to unseen scene layouts and appearance configurations, testing whether policies rely on task-relevant cues rather than memorized environments.

Camera Viewpoint. We perturb the extrinsics of the head-mounted cameras through controlled translations and rotations, measuring robustness to observation-view changes.

Language. We replace the original instruction with semantically equivalent expressions that vary wording and contextual description while preserving the intended task.

Sensor Noise. We perturb RGB observations using controlled Gaussian blur with different severity levels to evaluate robustness to degraded visual sensing.

Robot Initial State. We perturb the initial base position and orientation within a valid local region while rejecting configurations that introduce collisions or invalid task states.

Under the standard HUMANO protocol, the seven OOD settings are used only for evaluation: models are trained on in-distribution demonstrations and directly tested under each controlled shift. We additionally release 27.5K OOD demonstrations covering five settings for optional adaptation and robustness studies. Language variation does not require additional physical demonstrations, while Robot Initial State is excluded because corresponding demonstrations cannot be generated through the same rendering pipeline and related configuration variation is already covered by the Spatial suite.

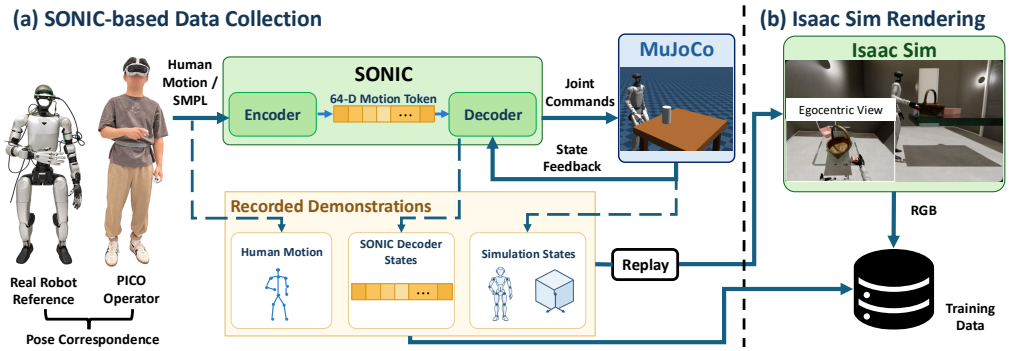


Figure 3: Overview of the HUMANO data-collection and rendering pipeline. (a) Human motion captured with a PICO VR system is encoded into SONIC motion-latent tokens and decoded into body joint commands for execution in MuJoCo, with hand commands provided separately by the PICO controls. Human motion, decoder states, and simulation states are recorded. (b) The saved robot and object states are replayed in Isaac Sim to render egocentric RGB observations. The rendered images and recorded demonstration data together form the training dataset.

3.4 Demonstration Collection

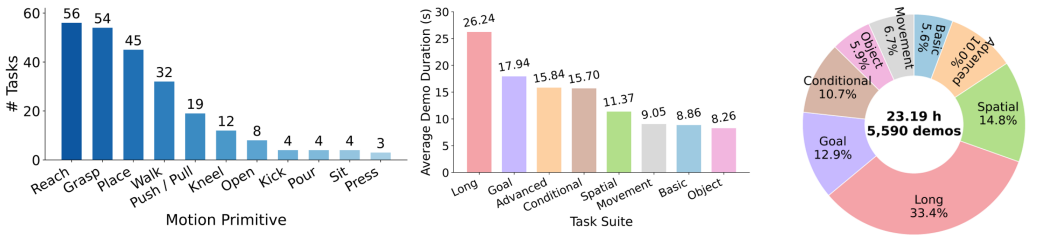
Whole-Body Teleoperation. As shown in Figure 3, HUMANO collects demonstrations through a PICO-based VR teleoperation system coupled with the SONIC whole-body controller (Luo et al., 2026). During teleoperation, we record human motion, SONIC decoder states, and simulation states. The simulation states capture the robot's base pose and joint configurations, together with the states of target objects. Unlike teleoperation pipelines that separately specify upper-body manipulation and base locomotion commands, our system represents operator motion within a unified whole-body motion space. The captured motion is encoded into SONIC motion-latent tokens and decoded into joint commands for whole-body execution, enabling coordinated locomotion and manipulation through a single control interface. This design supports complex behaviors such as walking, turning, bending, reaching, carrying, and pushing without manually decomposing them into separate locomotion and manipulation commands.

Simulation Replay and Rendering. HUMANO separates interactive demonstration collection in MuJoCo from offline visual rendering in Isaac Sim. For rendering, we reconstruct the corresponding scene in Isaac Sim and replay these saved states frame by frame. Scenes and task placements are selected to accommodate the recorded trajectories, rejecting conflicts with fixed structures and selectively hiding conflicting background assets. Egocentric RGB observations are then rendered from the robot-mounted camera and combined with the corresponding recorded demonstration data to construct the training dataset. This workflow preserves the collected motion sequence while allowing visual observations to be generated independently of the teleoperation process.

Unified Policy Interface. A central design choice of HUMANO is to use a common SONIC-based whole-body representation throughout demonstration collection, policy learning, and benchmark evaluation. Heterogeneous high-level policies, including imitation-learning methods and vision-language-action models, are adapted to predict actions in the same motion-latent space. At execution time, these predictions are decoded by a shared SONIC whole-body controller into executable robot motions. This unified interface provides a consistent action representation and execution pipeline across model families, facilitating policy adaptation and standardized evaluation.

3.5 Dataset

Dataset Construction. Using the above teleoperation system, we collect 5.5k in-distribution demonstrations covering all 100 tasks across the eight HUMANO suites. We also provide 27.5K demonstrations under the seven controlled OOD settings introduced in Section 3.3. These OOD trajectories are kept separate from the default training set and are intended for optional adaptation, robustness, and data-scaling studies. Under the standard HUMANO protocol, policies are trained only on the in-distribution demonstrations and evaluated directly under the corresponding distribution shifts.



(a) Atomic Motion Distribution (b) Average Demonstration Duration (c) Trajectory Duration Distribution

Figure 4: Dataset statistics of HUMANO. (a) Coverage of representative atomic motion primitives across the 100 tasks. (b) Average demonstration duration for each task suite. (c) Distribution of total trajectory duration across the eight task suites.

Dataset Format. Each demonstration is stored as a synchronized multimodal trajectory containing egocentric RGB observations, language instructions, robot proprioceptive states, and SONIC-compatible whole-body action targets. We additionally record task and environment metadata, including the task identifier, suite, scene, object configuration, robot initialization, and generalization setting when applicable. This unified representation allows different policy architectures to consume the same observations and predict actions under a common whole-body control interface.

Dataset Statistics. As shown in Figure 4, we characterize the HUMANO dataset in terms of motion-primitive coverage and trajectory duration. Figure 4(a) reports the number of tasks involving representative whole-body motion primitives, including reaching, grasping, placing, walking, pushing/pulling, kneeling, opening, kicking, pouring, sitting, and pressing, demonstrating broad coverage of fundamental humanoid behaviors. Figure 4(b) shows the average demonstration duration across the eight task suites, with longer durations generally corresponding to tasks requiring more extended behavioral sequences. Figure 4(c) further summarizes the distribution of total trajectory duration across suites. Overall, HUMANO contains 5,590 demonstrations totaling 23.19 hours, with the Long suite accounting for the largest fraction of trajectory time.

4 Experiments

4.1 Experimental Setup

Baselines. We evaluate six representative policy architectures spanning imitation learning and recent generalist robot policies: Diffusion Policy (Chi et al., 2025), Flow Matching (Dynamics & Team, 2025), ACT (Zhao et al., 2023), $\pi_{0.5}$ (Intelligence et al., 2025), GR00T N1.7 (Bjorck et al., 2025), and Ψ_0 (Wei et al., 2026a). All baselines are adapted to the unified SONIC-based motion-latent interface described in Section 3.4. Detailed model configurations and training hyperparameters are provided in the appendix.

Detailed model configurations and training settings are provided in Appendix C, with the shared evaluation protocol described in Appendix C.3.

4.2 Main Results

Table 2 reports the main in-distribution results on six HUMANO suites, covering 70 tasks in total. Each baseline is trained independently on each suite using the corresponding in-distribution demonstrations and evaluated with 25 rollouts per task. We report the average success rate for each suite and the task-weighted average across all evaluated tasks.

Overall, the evaluated policies exhibit substantial differences in both absolute performance and capability profiles. GR00T N1.7 achieves the highest overall success rate of 29.3%, followed by $\pi_{0.5}$ at 27.1%, substantially outperforming the remaining baselines. On the *Basic* suite, $\pi_{0.5}$ achieves the highest success rate of 46.0%, followed by GR00T N1.7 at 39.2%. In contrast, GR00T N1.7 performs best on the *Spatial*, *Object*, and *Goal* suites. These results demonstrate the promising potential of generalist policies for multi-task humanoid loco-manipulation, while also revealing distinct strengths across capability dimensions.

Table 2: Main in-distribution results on six HUMANO suites. We report success rate (%) averaged over 25 rollouts per task. **Bold** denotes the best result.

Method	Basic	Advanced	Spatial	Object	Goal	Conditional	Avg.
Diffusion Policy	14.8	0.0	9.8	11.6	0.0	0.8	6.7
Flow Matching	3.2	4.0	19.0	17.2	3.2	4.0	9.9
ACT	16.0	0.8	12.6	21.6	0.0	4.8	9.8
$\pi_{0.5}$	**46.0**	**32.0**	29.8	32.4	2.0	**18.0**	27.1
GR00T N1.7	39.2	18.0	**42.0**	**39.2**	**8.4**	16.0	**29.3**
Ψ_0	12.8	12.4	20.8	22.8	0.8	3.2	13.4

Table 3: Out-of-distribution generalization on HUMANO basic suite. We report success rate (%) under each controlled OOD setting, averaged over 10 rollouts per task. **Bold** denotes the best result.

Method	Distractor	Lighting	Background	Camera	Language	Sensor	Robot Init.	OOD Avg.
Diffusion Policy	8.0	18.0	18.0	14.0	12.0	2.0	13.0	12.1
Flow Matching	8.0	6.0	13.0	7.0	15.0	2.0	7.0	8.3
ACT	14.0	14.0	19.0	10.0	20.0	3.0	10.0	12.9
$\pi_{0.5}$	**47.0**	**49.0**	**46.0**	**46.0**	**38.0**	**40.0**	**44.0**	**44.3**
GR00T N1.7	20.0	35.0	42.0	40.0	35.0	25.0	34.0	33.0
Ψ_0	11.0	11.0	8.0	8.0	33.0	7.0	10.0	12.6

Performance varies considerably across task suites. The *Basic* suite is relatively more tractable for the strongest policies, whereas success rates are substantially lower on suites requiring goal conditioning or state-dependent decision making, particularly *Goal* and *Conditional*. Moreover, no single policy dominates across all evaluated suites: $\pi_{0.5}$ performs best on *Basic*, *Advanced*, and *Conditional*, while GR00T N1.7 leads on *Spatial*, *Object*, and *Goal*. This cross-suite variation highlights the importance of evaluating humanoid policies across multiple capability dimensions rather than relying solely on an aggregate success rate.

4.3 Out-of-Distribution Generalization Study

We next evaluate the robustness of all baselines under the seven controlled OOD settings introduced in Section 3.3: *Distractor*, *Lighting*, *Background*, *Camera Viewpoint*, *Language*, *Sensor Noise*, and *Robot Initial State*. For each task-shift combination, we perform 10 rollouts. Models are trained only on the corresponding in-distribution demonstrations, without using the released OOD demonstrations.

Table 3 summarizes the results on the Basic suite. $\pi_{0.5}$ achieves the strongest OOD performance across all seven settings, with an average success rate of 44.3%, followed by GR00T N1.7 at 33.0%. In contrast, Diffusion Policy, Flow Matching, and ACT achieve average success rates of only 12.1%, 8.3%, and 12.9%, respectively, revealing a substantial robustness gap between the strongest generalist policies and conventional imitation-learning baselines.

The difficulty also varies across OOD settings. Averaged over all six baselines, *Sensor Noise* is the most challenging shift, with a success rate of 13.2%, while *Language* and *Background* are less disruptive. The stronger performance under background variation may partly benefit from the scene diversity already present in the in-distribution training data. Overall, OOD robustness remains highly sensitive to the type of distribution shift, especially visual degradation.

4.4 Analysis Study

Multi-Task Scaling. We first study how multi-task performance scales with the amount of demonstration data. Specifically, we train the same policy using 10%, 20%, 50%, and 100% of the available demonstrations while keeping the model architecture and training configuration fixed. As shown in Figure 5(a), training with only 5 demonstrations per task fails to produce a functional multi-task policy, resulting in a 0.0% success rate. Increasing the data budget to 10 demonstrations per task substantially improves performance to 29.6%, which remains similar at 30.0% with 25 demonstrations and further increases to 39.2% with the full 50-demonstration budget. These results indicate that sufficient demonstration coverage is critical for stable multi-task learning, while the continued improvement at larger data scales suggests further potential from increasing demonstration data.

Table 4: Single-task versus suite-level training of GR00T N1.7 on the Basic suite. We report success rate (%) for each task and the average across all 10 tasks.

Setting	Ball Kick	Chair Push	Chair Sit	Dolly Push	Ground Basket Pick	Ground Bottle Pick	Ground Box Pick	Table Apple Pick	Table Book Pick	Table Plate Pick	Avg.
Task-Level	76	52	88	68	48	32	56	8	52	12	49.2
Suite-Level	24	100	56	8	76	4	20	12	48	44	39.2

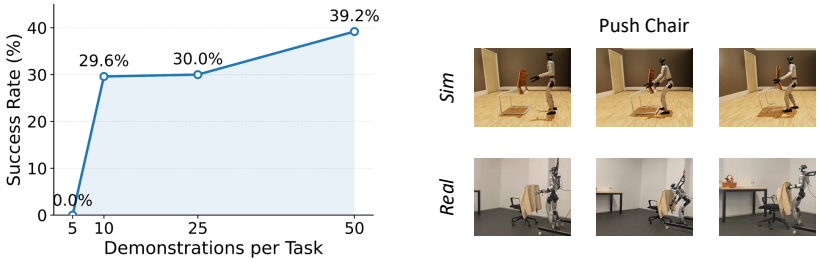


Figure 5: Scaling and sim-to-real evaluation of HUMANO. (a) Multi-task success rate with increasing demonstration budgets. (b) Sim-to-real transfer on the Push Chair task, with representative simulation and real-robot rollouts.

Single-Task vs. Multi-Task Learning. We next compare single-task and multi-task learning under the same per-task demonstration budget. In the single-task setting, an independent policy is trained for each task, whereas in the multi-task setting, a single shared policy is jointly trained on all tasks in the suite. As shown in Table 4, single-task training achieves an average success rate of 49.2%, compared with 39.2% for suite-level multi-task training. This overall gap is expected, as a single-task policy only needs to optimize for one task, whereas a shared policy must jointly model multiple behaviors. Nevertheless, multi-task training does not degrade every task: Chair Push, Ground Basket Pick, Table Apple Pick, and Table Plate Pick achieve higher success rates under joint training. This suggests that related tasks may share reusable motion primitives and interaction patterns that provide beneficial transfer. Overall, these results highlight both the greater difficulty of multi-task humanoid learning and the potential benefit of shared behaviors across tasks.

4.5 Sim-to-Real Evaluation

Finally, we evaluate whether a policy trained in HUMANO can transfer to a physical humanoid robot. We select the representative *Push Chair Forward* task, which requires sustained physical interaction and coordinated whole-body movement. The policy is trained using only simulated demonstrations and deployed on the real robot through the same SONIC-based motion-latent interface and whole-body controller used in simulation. Table 5(b) reports the results in simulation and on the real robot. GR00T N1.7 achieves a success rate of 66.7% in the real world, compared with 100% in simulation. Despite the remaining sim-to-real gap, these results provide preliminary evidence that policies trained with HUMANO can transfer executable whole-body behaviors to the physical robot under the same motion-control interface.

5 Conclusion

We introduced HUMANO, a large-scale benchmark for generalist humanoid loco-manipulation with 100 tasks across eight capability suites, 90 interactive scenes, and over 2,000 object assets. HUMANO provides controlled OOD evaluation and a unified SONIC-based motion-latent interface for standardized policy comparison. Experiments show promising multi-task capability in generalist robot policies, continued gains from increased demonstration scale, and preliminary sim-to-real transfer, while substantial capability and robustness gaps remain. We hope HUMANO serves as a scalable platform for advancing general and robust humanoid loco-manipulation.

AI USE STATEMENT

In this work, we used generative AI tools for language editing, including grammar correction, sentence refinement, and improving the clarity and readability of the manuscript. We have not used generative AI tools for generating research ideas, designing experiments, collecting or analyzing data, deriving scientific claims, or producing experimental results. Additionally, we used generative AI tools for recommended disclosure tasks related to writing assistance. All AI-assisted edits were carefully reviewed and revised by the authors. The authors verified the accuracy, consistency, and originality of the final manuscript, and take full responsibility for all content, including text, claims, and artifacts presented in this work.

REPRODUCIBILITY STATEMENT

We provide detailed information to facilitate reproduction of HUMANO and the reported experiments. The benchmark construction, task-suite design, generalization settings, demonstration collection pipeline, and evaluation protocol are described in the main paper, while the appendix provides complete definitions of all 100 tasks, detailed configurations of the seven OOD settings, dataset statistics, baseline implementations, training configurations, and task-level evaluation results. We will open-source the complete HUMANO benchmark, including environment and task configurations, data-collection code, demonstration data, training and inference pipelines, evaluation scripts, and baseline implementations, to support reproducible evaluation and future research.

REFERENCES

Firas Al-Hafez, Guoping Zhao, Jan Peters, and Davide Tateo. Locomujoco: A comprehensive imitation learning benchmark for locomotion. *arXiv preprint arXiv:2311.02496*, 2023.

Shuanghao Bai, Meng Li, Xinyuan Lv, Jiawei Wang, Xinhua Wang, Fei Liao, Chengkai Hou, Langzhe Gu, Wanqi Zhou, Kun Wu, et al. Hex: Humanoid-aligned experts for cross-embodiment whole-body manipulation. *arXiv preprint arXiv:2604.07993*, 2026.

Qingwei Ben, Feiyu Jia, Jia Zeng, Junting Dong, Dahua Lin, and Jiangmiao Pang. Homie: Humanoid loco-manipulation with isomorphic exoskeleton cockpit. In *Robotics: Science and Systems*, 2025.

Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, et al. RoboTwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11):1684–1704, 2025.

Boston Dynamics and TRI Research Team. Large behavior models and atlas find new footing, 2025. URL `https://bostondynamics.com/blog/large-behavior-models-atlas-find-new-footing/`. Blog post.

Senyu Fei, Siyin Wang, Junhao Shi, Zihao Dai, Jikun Cai, Pengfang Qian, Li Ji, Xinzhe He, Shiduo Zhang, Zhaoye Fei, Jinlan Fu, Jingjing Gong, and Xipeng Qiu. LIBERO-Plus: A progressive robustness benchmark for visual-language-action models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38574–38583, June 2026.

Zipeng Fu, Qingqing Zhao, Qi Wu, Gordon Wetzstein, and Chelsea Finn. Humanplus: Humanoid shadowing and imitation from humans. In *8th Conference on Robot Learning*, 2024.

Connor Gäde, Matthias Kerzel, Erik Strahl, and Stefan Wermter. Sim-to-real neural learning with domain randomisation for humanoid robot grasping. In *Artificial Neural Networks and Machine*

Learning – ICANN 2022, volume 13529 of *Lecture Notes in Computer Science*, pp. 342–354. Springer, 2022. doi: 10.1007/978-3-031-15919-0_29.

Haoran Geng, Feishi Wang, Songlin Wei, Yuyang Li, Bangjun Wang, Boshi An, Charlie Tianyue Cheng, Haozhe Lou, Peihao Li, Yen-Jen Wang, et al. Roboverse: Towards a unified platform, dataset and benchmark for scalable and generalizable robot learning. In *Robotics: Science and Systems*, 2025.

Tairan He, Zhengyi Luo, Xialin He, Wenli Xiao, Chong Zhang, Weinan Zhang, Kris M. Kitani, Changliu Liu, and Guanya Shi. Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning. In *8th Conference on Robot Learning*, 2024.

Yingdong Hu, Haodong Zhu, Boyuan Zheng, Yihang Hu, Tong Zhang, Zunhao Chen, Junming Zhao, Ruiqian Nai, and Yang Gao. Openhlm: An empirical recipe for whole-body humanoid loco-manipulation. *arXiv preprint arXiv:2606.22174*, 2026.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi_{0.5}$: A vision-language-action model with open-world generalization. In *Conference on Robot Learning*, 2025.

Haoran Jiang, Jin Chen, Qingwen Bu, Li Chen, Modi Shi, Yanjie Zhang, Delong Li, Chuanzhe Suo, Chuang Wang, Zhihui Peng, et al. Wholebodyvla: Towards unified latent vla for whole-body loco-manipulation control. In *The Fourteenth International Conference on Learning Representations*, 2026.

Yejin Kim, Wilbert Pumacay, Omar Rayyan, Max Argus, Winson Han, Eli Vanderbilt, Jordi Salvador, Abhay Deshpande, Rose Hendrix, Snehal Jauhri, et al. MolmoSpaces: Large-scale open ecosystem for robot manipulation and navigation. In *Proceedings of Robotics: Science and Systems*, Sydney, Australia, July 2026. doi: 10.15607/RSS.2026.XXII.091.

Jialong Li, Xuxin Cheng, Tianshu Huang, Shiqi Yang, Ri-Zhao Qiu, and Xiaolong Wang. Amo: Adaptive motion optimization for hyper-dexterous humanoid whole-body control. *arXiv preprint arXiv:2505.03738*, 2025a.

Yixuan Li, Yutang Lin, Jieming Cui, Tengyu Liu, Wei Liang, Yixin Zhu, and Siyuan Huang. Clone: Closed-loop whole-body humanoid teleoperation for long-horizon tasks. In *9th Annual Conference on Robot Learning*, 2025b.

Zhe Li, Zhenzhe Zhang, Yangyang Wei, Wenjie Zhang, Xichen Yuan, Peiyuan Zhi, Gen Li, Xinying Guo, Fengjie Gao, Jianfei Yang, et al. ω_0: A latent predictive world action model for concurrent humanoid loco-manipulation. *arXiv preprint arXiv:2608.06375*, 2026.

Qiayuan Liao, Takara E Truong, Xiaoyu Huang, Yuman Gao, Guy Tevet, Koushil Sreenath, and C Karen Liu. Beyondmimic: From motion tracking to versatile humanoid control via guided diffusion. *Science Robotics*, 11(117):eadx8924, 2026.

Kevin Lin, Ajay Mandlekar, Caelan Reed Garrett, Nikita Chernyadev, Yu Fang, Runyu Ding, Yuqi Xie, Linxi Fan, and Yuke Zhu. Humanoidmimicgen: Data generation for loco-manipulation via whole-body planning and adaptation. In *ICRA 2026 Workshop on Synthetic Data for Robot Learning*, 2026.

Zhengyi Luo, Ye Yuan, Tingwu Wang, Chenran Li, Fernando Castañeda, Sirui Chen, Zi-Ang Cao, Jiefeng Li, David Minor, Qingwei Ben, et al. Sonic: Supersizing motion tracking for natural humanoid whole-body control. *Science Robotics*, 11(117):eaed4592, 2026.

Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel Van de Panne. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions On Graphics (TOG)*, 37(4):1–14, 2018.

Xue Bin Peng, Ze Ma, Pieter Abbeel, Sergey Levine, and Angjoo Kanazawa. Amp: Adversarial motion priors for stylized physics-based character control. *ACM Transactions on Graphics (ToG)*, 40(4):1–20, 2021.

Carmelo Sferrazza, Dun-Ming Huang, Xingyu Lin, Youngwoon Lee, and Pieter Abbeel. Humanoidbench: Simulated humanoid benchmark for whole-body locomotion and manipulation. In *Robotics: Science and Systems*, 2024.

Taowen Wang, Zikang Xie, Bin Yang, Yunheng Wang, Zizhao Yuan, Yuetong Fang, Yixiao Feng, Yichi Wang, Xingyu Chen, Haodong Chen, et al. Humanoidarena: Benchmarking egocentric hierarchical whole-body learning. *arXiv preprint arXiv:2606.17833*, 2026.

Songlin Wei, Hongyi Jing, Boqian Li, Zhenyu Zhao, Jiageng Mao, Zhenhao Ni, Sicheng He, Jie Liu, Xiawei Liu, Kaidi Kang, et al. Ψ_0: An open foundation model towards universal humanoid loco-manipulation. In *Robotics: Science and Systems*, 2026a.

Songlin Wei, Zhenhao Ni, Jie Liu, Zhenyu Zhao, Junjie Ye, Hongyi Jing, Junkai Xia, Xiawei Liu, Michael Leong, Liang Heng, Di Huang, and Yue Wang. SIMPLE: Simulation-based policy learning and evaluation for humanoid loco-manipulation. In *Conference on Robot Learning*, 2026b.

Tianyi Xie, Haotian Zhang, Jinhyung Park, Zi Wang, Bowen Wen, Jiefeng Li, Xueting Li, Qingwei Ben, Haoyang Weng, Yufei Ye, et al. Grail: Generating humanoid loco-manipulation from 3d assets and video priors. *arXiv preprint arXiv:2606.05160*, 2026.

Zehao Yu, Jiakun Zheng, Weiji Xie, Jiyuan Shi, Chenyun Zhang, Chenjia Bai, and Xuelong Li. Oasis: From simulation data collection to real-world humanoid loco-manipulation. *arXiv preprint arXiv:2606.08548*, 2026.

Yanjie Ze, Zixuan Chen, Joao Pedro Araujo, Zi-ang Cao, Xue Bin Peng, Jiajun Wu, and Karen Liu. Twist: Teleoperated whole-body imitation system. In *Conference on Robot Learning*, pp. 2143–2154, 2025.

Bowei Zhang, Qiyao Zhang, Shuanghao Bai, Xinhua Wang, Meng Li, Yilei Wang, Leiwang Zhang, Jian Tang, Lu Zhou, Lei Sun, et al. Wholebodywam: Learning whole-body world action models with scalable motion priors. *arXiv preprint arXiv:2609.18197*, 2026.

Tony Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Robotics: Science and Systems XIX*, 2023.

Jia Zheng, Teli Ma, Yudong Fan, Zifan Wang, Shuo Yang, and Junwei Liang. Motionwam: Towards foundation world action models for real-time humanoid loco-manipulation. *arXiv preprint arXiv:2606.09215*, 2026.

A Full Task Definitions

This section provides the complete task definitions of HUMANO. The benchmark contains 100 tasks organized into eight suites: *Basic*, *Advanced*, *Spatial*, *Object*, *Movement*, *Goal*, *Long-Horizon*, and *Conditional*. These suites are designed to separate different sources of difficulty in humanoid loco-manipulation. Rather than constructing a large task set through minor scene or object perturbations, HUMANO varies task semantics, spatial configuration, manipulated objects, target goals, temporal composition, and environment-dependent decision rules in a controlled manner. For suites involving structured task variants, the tables below explicitly specify the corresponding object choices, target choices, spatial configurations, or conditional branches.

Basic Suite. The Basic suite serves as the foundation for the remaining suites. It contains 10 tasks covering representative whole-body interaction primitives, including tabletop grasping, floor-level grasping, pushing, sitting, button pressing, and kicking. These tasks span different interaction heights and body configurations: tabletop tasks primarily require reaching and grasping, floor-level tasks involve larger posture changes, while pushing, sitting, and kicking require coordinated whole-body control and balance. Each task has a single primary objective, providing a basic assessment of fundamental humanoid interaction skills before additional generalization, composition, or conditional reasoning is introduced.

Table 5: Full task definitions of the Basic and Advanced suites.

Basic Suite		**Advanced Suite**	
Task ID	Task Definition	Task ID	Task Definition
B-1	Pick up an apple from a table.	A-1	Put an apple into a basket.
B-2	Pick up a plate from a table.	A-2	Place a plate on a cabinet.
B-3	Push a chair.	A-3	Throw a milk carton into a trash bin.
B-4	Push a cart.	A-4	Put bread onto a tray.
B-5	Sit on a chair.	A-5	Pick up a water bottle from the floor and place it on a table.
B-6	Press a button.	A-6	Lift a box and place it on a table.
B-7	Pick up a basket from the floor.	A-7	Pick up a basket from the floor and place it on a table.
B-8	Pick up a water bottle from the floor.	A-8	Push a chair to the front of a table.
B-9	Pick up a box from the floor.	A-9	Push a cart to the front of a table.
B-10	Kick a ball.	A-10	Pour an eraser out of a pencil holder.

Table 6: Full task definitions of the Spatial suite.

Task ID	Task Definition
S-1	Pick up an apple from five tabletop locations: two corners, two edges, and the center.
S-2	Pick up a plate from five tabletop locations: two corners, two edges, and the center.
S-3	Pick up a bottle from three heights: a high table, a low table, and the floor.
S-4	Pick up a box from three heights: a high table, a low table, and the floor.
S-5	Push a cart with the robot initialized on four different sides of the cart.
S-6	Push a chair with the robot initialized on four different sides of the chair.
S-7	Kick a soccer ball with the robot initialized on four different sides of the ball.
S-8	Press a button at five tabletop locations: two corners, two edges, and the center.
S-9	Pick up a basket from three heights: the floor, a low table, and a high table.
S-10	Put an apple into a basket under five spatial configurations.
S-11	Place a plate on a cabinet under five spatial configurations.
S-12	Throw a milk carton into a trash bin under five spatial configurations.
S-13	Put bread onto a tray under five spatial configurations.
S-14	Pick up a water bottle from the floor and place it on a table under five spatial configurations.
S-15	Pick up a box from the floor and place it on a table under five spatial configurations.
S-16	Push a chair to the front of a table under five spatial configurations.
S-17	Move a cart to the front of a table under five spatial configurations.
S-18	Pour an eraser out of a pencil holder under five spatial configurations.
S-19	Place a book on three different levels of a bookshelf.
S-20	Pick up a basket from the floor and place it on a table under five spatial configurations.

Advanced Suite. The Advanced suite extends the Basic suite from primitive execution to goal-directed loco-manipulation. Its 10 tasks require the robot not only to interact with an object, but also to move or transform it toward a specified target state, such as placing an object into a receptacle, moving a floor-level object onto a table, pushing furniture toward a destination, or pouring an object from a container. Compared with the Basic suite, these tasks require longer coordination between perception, locomotion, manipulation, and object placement while retaining relatively simple task semantics. The suite therefore evaluates whether individually acquired whole-body skills can be integrated into complete goal-directed behaviors.

Spatial Suite. The Spatial suite evaluates whether a policy can preserve the same task semantics under changes in spatial configuration. It contains 20 tasks derived from representative interaction patterns in the Basic and Advanced suites, while systematically varying object locations, interaction heights, robot-object relations, or target locations. For example, tabletop objects may appear at different corners, edges, or the center of a surface; grasp targets may be located on a high table, a low table, or the floor; and the robot may be initialized on different sides of the object. Other tasks use multiple spatial arrangements of the same source and target objects. By keeping the manipulation

Table 7: Full task definitions of the Object and Movement suites.

Object Suite		Movement Suite	
Task ID	Task Definition	Task ID	Task Definition
O-1	Single-handed tabletop grasping of an apple, banana, or cake.	M-1	Pick up an apple from a table.
O-2	Two-handed tabletop grasping of a plate or tray.	M-2	Pick up a plate from a table.
O-3	Pick up an apple, banana, plate, or tray from a table.	M-3	Push a chair.
O-4	Single-handed floor-level grasping of a water bottle or basket.	M-4	Push a cart.
O-5	Two-handed floor-level grasping of a box or storage box.	M-5	Sit on a chair.
O-6	Pick up a water bottle, basket, box, or storage box from the floor.	M-6	Press a button.
O-7	Put a small object—an apple, banana, or cake—into a basket.	M-7	Pick up a basket from the floor.
O-8	Put a large object—a plate or tray—into a cabinet.	M-8	Pick up a water bottle from the floor.
O-9	Throw a tabletop object—milk, bread, or cake—into a trash bin.	M-9	Pick up a box from the floor.
O-10	Put food—bread or cake—onto a tray.	M-10	Kick a ball.

objective fixed while changing geometry, this suite separates spatial generalization from changes in object identity or task goal. The complete configurations used by each task are listed in Table 6.

Object Suite. The Object suite focuses on generalization across manipulated objects while preserving the underlying interaction pattern. Each task defines a shared manipulation objective and instantiates it with multiple object categories or geometries. The suite includes both single-handed and two-handed grasping, tabletop and floor-level interaction, and goal-directed placement into baskets, cabinets, trays, or trash bins. For example, the same grasping behavior may be instantiated with an apple, banana, or cake, while larger-object tasks vary between plates, trays, boxes, and storage boxes. This construction tests whether a policy learns reusable manipulation behavior rather than overfitting to the visual appearance or geometry of a particular object instance.

Movement Suite. The Movement suite contains 10 tasks whose high-level objectives mirror the fundamental interaction patterns in the Basic suite, including grasping, pushing, sitting, pressing, and kicking. Keeping the task semantics aligned with the Basic suite provides a controlled basis for studying movement-related variation without simultaneously changing the underlying task objective. Accordingly, the table reports the common high-level task definitions, while movement-specific configurations are treated separately from task semantics.

Goal Suite. The Goal suite evaluates whether a policy can condition its behavior on the specified task goal when multiple valid interaction targets are simultaneously present. Each of the 10 tasks fixes the manipulated object and surrounding context but varies the desired destination or receptacle. For example, an apple may need to be placed into a basket, box, or tray, while a plate may be assigned to a basket, box, or cabinet. Because the alternative targets remain available in the scene, successful execution requires the policy to identify the instructed goal rather than simply applying a fixed object-to-target association. This suite therefore isolates goal-conditioned generalization from changes in the basic manipulation skill itself.

Long-Horizon Suite. The Long-Horizon suite contains 20 tasks that compose multiple locomotion and manipulation primitives into extended behavioral sequences. Unlike the Basic and Advanced suites, successful completion requires several intermediate states to be reached in the correct order. Representative tasks include opening a drawer before placing an object inside, placing an object onto a receptacle before lifting it, moving a container before inserting another object, and sequentially manipulating multiple objects. These tasks introduce longer temporal dependencies and expose

Table 8: Full task definitions of the Goal suite.

Task ID	Task Definition
G-1	Put an apple into the specified target among a basket, box, and tray.
G-2	Put a plate into the specified target among a basket, box, and cabinet.
G-3	Put bread into the specified target among a basket, tray, and trash bin.
G-4	Put milk into the specified target among a basket, box, and trash bin.
G-5	Put a stationery item into the specified target among a pencil case, pencil holder, and tray.
G-6	Put a book into the specified target among a bookshelf and cabinet.
G-7	Pick up a bottle from the floor and place it on the specified target among a table and cabinet.
G-8	Pour an eraser from a pencil holder onto the specified target among a table and cabinet.
G-9	Push a cart to the specified destination in front of a table or cabinet.
G-10	Place a soccer ball at the specified target on the floor or on a cabinet.

Table 9: Full task definitions of the Long-Horizon suite.

Task ID	Task Definition
L-1	Pick up an apple from a table, open a drawer, and place the apple inside.
L-2	Open a drawer, pick up a plate from a table, and place the plate inside.
L-3	Put both bread and cake onto a tray.
L-4	Put bread onto a tray and then lift the tray.
L-5	Put cake onto a plate and then lift the plate.
L-6	Pick up a box from the floor, place it on a cart, and push the cart forward.
L-7	Push a chair and then sit on it.
L-8	Move a soccer ball from a table to the floor and then kick it.
L-9	Move a basket from the floor to a table and then put a tabletop apple into the basket.
L-10	Move a box from the floor to a table and then put a floor-level water bottle into the box.
L-11	Put both an apple and a banana into a basket.
L-12	Move an apple from a basket onto a table and throw a banana into a trash bin.
L-13	Pour two peanuts from a container onto a plate.
L-14	Arrange two books on a cabinet.
L-15	Put an apple into a basket, lift the basket, and place it on the floor.
L-16	Move an apple from a high table into a drawer of a low table.
L-17	Move a plate from a low table into a drawer of a high table.
L-18	Move a cart away and then push a chair to the front of a table.
L-19	Put a book into a bookshelf.
L-20	Move a book from a bookshelf into a tabletop drawer.

policies to error accumulation: an early failure in grasping, navigation, placement, or articulated interaction can prevent all subsequent stages from being completed. The suite is therefore designed to evaluate temporal skill composition rather than isolated primitive success.

Conditional Suite. The Conditional suite evaluates state-dependent decision making. Each of its 10 tasks admits more than one valid action sequence, and the correct sequence is determined by the observed environment state. Some tasks explicitly require branching on articulated-object state, such as opening a cabinet or drawer only when it is closed. Others require selecting an empty container, choosing a basket according to its contents, or deciding whether an obstructing object must first be moved. Consequently, success cannot be achieved by executing a single fixed trajectory for every episode; the policy must first identify the relevant state and then select the corresponding action sequence. This suite complements the Long-Horizon suite by introducing conditional branching in addition to temporal composition.

B OOD Setting Details

B.1 Protocol Overview

This section provides the implementation details of the seven controlled OOD settings introduced in Section 3.3: *Distractor*, *Lighting*, *Background*, *Camera Viewpoint*, *Language*, *Sensor Noise*,

Table 10: Full task definitions of the Conditional suite.

Task ID	Task Definition
C-1	If the cabinet door is closed, open it and put the plate inside; otherwise, directly put the plate inside.
C-2	If the drawer is closed, open it and put the pencil inside; otherwise, directly put the pencil inside.
C-3	Put a water bottle into the empty container.
C-4	Put an apple into the basket that contains no other fruit.
C-5	If the chair is not in front of the table, push it to the table and sit down; otherwise, directly sit down.
C-6	If the basket is on the floor, first move it onto the table and then put the apple inside; otherwise, directly put the apple inside.
C-7	Pour peanuts into the empty container among the plate and tray.
C-8	Put bread into the empty container among the plate and tray.
C-9	If a chair is in front of the table, move it away before pushing the cart; otherwise, directly push the cart.
C-10	If the lower bookshelf level contains a book, place the tabletop book on the lower level; if the upper level contains a book, place it on the upper level.

and *Robot Initial State*. The settings are designed as factorized diagnostic tests and are evaluated independently rather than composed within the same episode.

Except for *Background*, each OOD episode is constructed from a corresponding in-distribution configuration by modifying only the designated factor. The task objective, success criterion, task-relevant objects, and all other environment states are kept unchanged. Sampled perturbation parameters and simulator states are stored and reused across evaluated policies, ensuring matched evaluation conditions. The *Background* setting instead evaluates transfer to globally held-out scenes and therefore changes the underlying scene layout and appearance while preserving the task semantics.

B.2 OOD Settings

Distractor. The Distractor condition evaluates robustness to task-irrelevant visual and physical clutter by introducing one to five additional objects near the interaction region (Wei et al., 2026b; Chen et al., 2026; Fei et al., 2026). To preserve the original task semantics, objects that are semantically identical or closely related to the manipulation targets, destination containers, and other task-relevant entities are excluded from the candidate pool. Eligible semantic categories are sampled before individual assets to reduce sampling bias toward categories containing more object instances.

Distractors are initialized from stable poses and placed on support surfaces already present in the task. Candidate placements are checked for physical support, collision avoidance, scene-boundary compliance, and visibility from the head-mounted camera. The task-relevant objects, robot initial state, task objective, and success criterion remain unchanged. Accepted object identities, poses, and sampling seeds are stored for deterministic replay.

Lighting. The *Lighting* condition evaluates robustness to illumination changes while preserving the scene geometry, object configuration, robot state, and task objective. Related benchmarks consider variations in light color, intensity, position, direction, and shadow configuration (Wei et al., 2026b; Chen et al., 2026; Fei et al., 2026).

The ID configuration contains six cylindrical lights arranged in two rows of three, with color temperature sampled from $[2001, 8001]\,\mathrm{K}$ and light radius sampled from $[0.08, 0.12]\,\mathrm{m}$. OOD episodes instantiate one of three lighting subtypes: *RGB color* applies a green-, magenta-, or cyan-dominant emission profile to all lights; *Local illumination* switches off one row of lights without compensating for the reduced illumination; and *Source size* samples the light radius from either $[0.03, 0.06]\,\mathrm{m}$ or $[0.16, 0.24]\,\mathrm{m}$, with intensity adjusted to approximately preserve nominal emitted power. Only one subtype is applied per episode, and the resulting lighting configuration remains fixed throughout the rollout.

Background. Existing benchmarks define background variation at different granularities. RoboTwin 2.0 and LIBERO-Plus primarily perturb scene textures and surface materials within fixed task environments, whereas SIMPLE exposes policies to diverse complete indoor scenes but

does not impose a globally disjoint room-level split for evaluating the same tasks (Wei et al., 2026b; Chen et al., 2026; Fei et al., 2026). HUMANO instead defines *Background* OOD using FloorPlans excluded from all training data, thereby evaluating task transfer to unseen room appearance, geometry, furniture configurations, and obstacle layouts.

The global scene split covers kitchens, living rooms, and bedrooms. For each room category, 20 FloorPlans are assigned to the training pool and 10 to the held-out pool, yielding 60 training and 30 held-out FloorPlans. All scene variants derived from the same underlying FloorPlan remain on the same side of the split, preventing an evaluation layout from appearing in the training data of another task. Unlike the other OOD conditions, which are derived from corresponding ID configurations, Background configurations are matched at the task level and instantiated exclusively in held-out FloorPlans. The task definition, task-relevant objects, objective, and success criterion are preserved.

To ground a task in a candidate FloorPlan, HUMANO constructs a protected spatial footprint from the training demonstration trajectories. For table-based tasks, a compatible table or counter in the background scene is used as a spatial anchor when available; the background anchor is subsequently hidden so that the task-native support and object configuration remain unchanged. Otherwise, the task is placed in a valid free-floor region. The background is aligned using only planar translation and yaw, without scaling or modifying the task geometry. A candidate placement is rejected if the complete task footprint falls outside the room or intersects immutable scene structures. Removable background assets that conflict with the task workspace or trajectory are hidden together with their linked components and dependent supported objects, avoiding partial furniture and unsupported scene fragments. The selected alignment and asset-visibility state are recorded for deterministic replay.

For tasks containing a task table, Background additionally samples from material families disjoint from those used for training and ID evaluation. These appearance changes do not modify the table's collision geometry or physical properties. Because the held-out FloorPlan and table appearance may change jointly, Background is evaluated as a compound scene-level shift rather than a texture-only perturbation; its results should therefore not be interpreted as isolating the effect of room geometry or appearance alone.

Camera Viewpoint. Camera-pose variation has been considered in both manipulation robustness benchmarks and humanoid sim-to-real data generation. LIBERO-Plus perturbs the position, orientation, and field of view of an external scene camera, whereas OASIS randomizes camera extrinsics when rendering G1 training data (Fei et al., 2026; Yu et al., 2026). HUMANO instead evaluates perturbations of the egocentric head-mounted stereo rig while keeping the robot state, physical task layout, camera intrinsics, and image resolution unchanged.

The perturbation ranges are adapted from humanoid-camera settings in OASIS and NICO (Yu et al., 2026; Gäde et al., 2022). OASIS provides the centimeter-scale translation and small-angle rotation ranges used for G1 rendering, while the larger pitch and yaw ranges follow the viewing-angle envelope used for NICO. These references are used to determine the perturbation scale rather than to replicate their protocols: OASIS applies camera randomization as a training augmentation, and NICO changes the robot's head joints, whereas HUMANO applies evaluation-time offsets to the camera-to-head extrinsic transformation.

We define six single-axis perturbation subtypes: x, y, z, roll, pitch, and yaw. For each Camera Viewpoint episode, exactly one subtype is selected and its offset is sampled as

$$\Delta x, \Delta y, \Delta z \sim \mathcal{U}(-0.01, 0.01)\ \mathrm{m}, \tag{1}$$

$$\Delta\mathrm{roll} \sim \mathcal{U}(-1.5^\circ, 1.5^\circ), \tag{2}$$

$$\Delta\mathrm{pitch} \sim \mathcal{U}(-15^\circ, 5^\circ), \tag{3}$$

$$\Delta\mathrm{yaw} \sim \mathcal{U}(-10^\circ, 10^\circ), \tag{4}$$

where all offsets associated with the unselected axes are set to zero. The translation is expressed in the nominal camera coordinate frame, and the rotation is composed using static XYZ roll–pitch–yaw angles. This single-axis construction separates sensitivity to translational and rotational changes and avoids confounding multiple viewpoint shifts within the same episode.

The sampled transformation is applied to the stereo rig before deriving the left- and right-eye poses. Consequently, both cameras undergo the same rigid extrinsic perturbation while their relative orientation and stereo baseline remain unchanged. The perturbed configuration is fixed throughout the episode and stored with the episode state for deterministic replay.

Language. Following prior robustness benchmarks that vary linguistic expressions while preserving task intent (Chen et al., 2026; Fei et al., 2026), the *Language* setting associates each task with one canonical training instruction and three evaluation variants held out from the HUMANO training annotations:

- **T1—Action-oriented:** a direct description of the required action and intended outcome;
- **T2—Context-oriented:** a description that first establishes a task-relevant scene or object relation and then states the required action; and
- **T3—Paraphrased:** a semantically equivalent reformulation of the canonical instruction using alternative wording or sentence structure.

T1 and T2 contain between 15 and 35 English words, whereas T3 has no fixed length constraint. All variants preserve the task-relevant entities and relations, subgoal order, terminal condition, and hand specification when explicitly required by the task. They do not disclose sampled initial states, numerical success thresholds, unobserved scene properties, or additional execution strategies.

Only the instruction is replaced relative to the corresponding ID configuration. The three variants are constructed and stored offline and are assigned deterministically across evaluation episodes by cycling through T1, T2, and T3.

Sensor Noise. Following the Gaussian-blur evaluation in LIBERO-Plus (Fei et al., 2026), the *Sensor Noise* setting applies Gaussian blur to both RGB views of the head-mounted stereo camera. The blur standard deviation is sampled uniformly from the discrete set

$$\sigma \in \{1, 2, \ldots, 10\} \text{ pixels}. \tag{5}$$

The same severity is used for both views and remains fixed throughout an episode. Blur is applied independently to the RGB channels before any policy-specific resizing or normalization. All non-RGB observations and the underlying simulator state remain unchanged. Whereas LIBERO-Plus applies the corruption to its agent-view observation, HUMANO applies it consistently to both egocentric head-camera views. Thus, this condition specifically measures robustness to controlled RGB blur rather than to general sensor corruption.

Robot Initial State. Prior benchmarks evaluate initial-state robustness through variations in manipulator starting poses or task-specific humanoid initialization regions (Fei et al., 2026; Wei et al., 2026b). HUMANO isolates changes in the humanoid floating-base configuration while preserving the task objects and non-root joint states. We consider two subtypes:

1. **Translation.** A planar offset is sampled uniformly by area within a disk of radius $R = 0.1\,\mathrm{m}$:
$$r = R\sqrt{u}, \qquad u \sim \mathcal{U}(0, 1), \tag{6}$$
$$\theta \sim \mathcal{U}(0, 2\pi), \tag{7}$$
$$(\Delta x, \Delta y) = (r\cos\theta, r\sin\theta). \tag{8}$$
2. **Yaw.** The base position is retained, and its heading is perturbed according to
$$\Delta\psi \sim \mathcal{U}(-10^\circ, 10^\circ). \tag{9}$$

Translation and yaw are evaluated separately. The base height, non-root joint states, task-object poses, and remaining scene state are inherited from the corresponding ID configuration. A sampled pose is accepted only if the robot remains within the valid scene region, satisfies its joint limits, introduces no new or deeper self-collision or external collision, and retains valid foot–ground support without non-foot ground contact. Up to 100 candidates are sampled; configurations for which no valid candidate is found are rejected rather than replaced by the original ID pose.

For each accepted candidate, the complete simulator state is stored and reused for deterministic evaluation. The validity test is based on the frozen initial configuration and does not use policy success or certify subsequent dynamic stability or task reachability.

C Implementation Details of Baselines

C.1 Common Training Setup

We evaluate six representative policies on the eight HUMANO suites: ACT (Zhao et al., 2023), Diffusion Policy (Chi et al., 2025), Flow Matching (Dynamics & Team, 2025), $\pi_{0.5}$ (Intelligence et al., 2025), GR00T N1.7 (Bjorck et al., 2025), and Ψ_0 (Wei et al., 2026a). For each baseline, we train a separate policy for each suite using all training demonstrations from the corresponding tasks. All baselines share the same suite splits, observation and action interfaces, and optimization-step budget, while retaining their native architectures, preprocessing pipelines, and training objectives.

All experiments use eight Nvidia A100 GPUs with a per-GPU batch size of 64, resulting in a global batch size of 512 without gradient accumulation. Each model is trained for 20,000 optimization steps. ACT, Diffusion Policy, Flow Matching, and $\pi_{0.5}$ use the LeRobot v0.6.0 codebase, whereas GR00T N1.7 and Ψ_0 use their respective native implementations with HUMANO-specific data and embodiment adapters.

Each observation contains an egocentric RGB image, $I_t \in [0, 255]^{360\times640\times3}$, and a 46-dimensional proprioceptive state,

$$s_t = \left[s_t^{\text{joint}}, g_t^{\text{proj}}\right] \in \mathbb{R}^{46}, \tag{10}$$

where $s_t^{\text{joint}} \in \mathbb{R}^{43}$ denotes the joint state and $g_t^{\text{proj}} \in \mathbb{R}^3$ is the projected gravity vector. Images are stored at their original resolution and processed using each model's native visual preprocessing pipeline.

All policies predict the same 78-dimensional action representation,

$$a_t = \left[a_t^{\text{motion}}, a_t^{\text{left}}, a_t^{\text{right}}\right] \in \mathbb{R}^{78}, \tag{11}$$

where $a_t^{\text{motion}} \in \mathbb{R}^{64}$ is the SONIC motion token and $a_t^{\text{left}}, a_t^{\text{right}} \in \mathbb{R}^7$ are the left- and right-hand commands. All baselines use one observation timestep and predict an action chunk of length $H = 40$. Language-conditioned policies additionally receive the corresponding task instruction.

Training data are drawn exclusively from the HUMANO training split. For LeRobot-based baselines, the source datasets are converted to the LeRobot v3.0 format with explicit `observation.state` and `action` fields. GR00T and Ψ_0 use model-specific adapters that expose the same 46-dimensional state and 78-dimensional action interface to their native data loaders. Normalization statistics are computed from the training data, while model-specific normalization, padding, and validity masks are handled by the corresponding native processors.

Table 11 summarizes the main training configurations of the six baselines.

Table 11: Baseline training configurations. GR00T uses 40 valid action positions within its internally padded action representation. The $\pi_{0.5}$ learning rate follows the configured pretrained checkpoint.

Policy	Obs. Horizon	Action Horizon	Learning Rate
ACT	1	40	1×10^{-5}
Diffusion Policy	1	40	1×10^{-4}
Flow Matching	1	40	1×10^{-4}
$\pi_{0.5}$	1	40	2×10^{-4}
GR00T N1.7	1	40	2×10^{-4}
Ψ_0	1	40	2×10^{-4}

C.2 Baseline Configurations

ACT. We use the LeRobot `ACTPolicy` implementation with an ImageNet-pretrained ResNet-18 visual encoder and a Transformer encoder–decoder. The Transformer uses a hidden dimension of 512, eight attention heads, four encoder layers, one decoder layer, a feed-forward dimension of 3,200, and dropout of 0.1. We retain the conditional variational formulation with a 32-dimensional latent variable and a four-layer latent encoder. Training minimizes an ℓ_1 action reconstruction loss together with a KL regularization term weighted by 10. We set `n_obs_steps=1`, `chunk_size=40`, and `n_action_steps=40`, with the state projection and action head adapted to the 46-dimensional state and 78-dimensional action spaces.

Diffusion Policy. Diffusion Policy uses an ImageNet-pretrained ResNet-18 visual encoder and a conditional one-dimensional U-Net. Visual features and proprioceptive states form the global condition for action-sequence denoising. The U-Net uses channel widths of 512, 1,024, and 2,048, a kernel size of five, and a 128-dimensional diffusion-timestep embedding. Training follows the DDPM noise-prediction objective with 100 diffusion timesteps and a squared-cosine noise schedule. We use one observation timestep, an action horizon of 40, and 10 reverse-diffusion steps at inference. States and actions are normalized using min–max statistics.

Flow Matching. The Flow Matching baseline uses LeRobot's `MultiTaskDiTPolicy` with the `flow_matching` objective. Pretrained CLIP ViT-B/16 encoders provide visual and language features, which condition the action Transformer together with the proprioceptive state. The Transformer contains six layers with a hidden dimension of 512, eight attention heads, rotary positional embeddings, and dropout of 0.1; the CLIP text encoder remains frozen. The model learns the velocity field between Gaussian noise and the target action chunk using the native beta timestep sampler with $\alpha = 1.5$, $\beta = 1.0$, and $s = 0.999$. Images are processed at 224×224, and inference uses 100-step Euler integration to generate a 40-step action sequence.

$\pi_{0.5}$. We use the LeRobot `PI05Policy` implementation initialized from a pretrained $\pi_{0.5}$ checkpoint. The PaliGemma vision-language backbone and Gemma action expert retain their native flow-matching objective, tokenizer, and prompt construction. The vision-language backbone is frozen, while the action expert and action projection layers are fine-tuned. We configure `max_state_dim=46`, `max_action_dim=78`, `chunk_size=40`, and `n_action_steps=40`. Proprioceptive states are discretized using the native state tokenizer and incorporated into the language prompt, while images are resized with padding to 224×224. Quantile-based state and action normalization is retained.

The pretrained action interface has 32 dimensions, whereas HUMANO uses 78-dimensional actions. We therefore expand the action input projection, output projection, and output bias to the required dimensionality. Existing pretrained entries are preserved, while additional dimensions are initialized by cyclically copying the pretrained entries along the expanded dimension.

GR00T N1.7. GR00T N1.7 uses a Cosmos-Reason2-2B vision-language backbone followed by an embodiment-conditioned flow-matching action head. We initialize from the pretrained N1.7-3B checkpoint and register the `UNITREE_G1_SONIC` embodiment. The 46-dimensional state is organized into leg, waist, arm, hand, and projected-gravity fields, while the 78-dimensional action is divided into the SONIC motion token and bilateral hand commands. The native processor places these fields within GR00T's internally padded 132-dimensional state and action representations and constructs the corresponding validity masks. The action modality contains 40 valid future timesteps.

Fine-tuning updates the projectors and action head while keeping the language and visual backbones frozen. We use AdamW with a learning rate of 2×10^{-4}, a 5% warm-up ratio, and weight decay of 10^{-5}. Color jitter uses brightness 0.3, contrast 0.4, saturation 0.5, and hue 0.08.

Ψ_0. Ψ_0 uses a Qwen3-VL-2B-based vision-language backbone and an ActionTransformer action head. The action head contains six Transformer blocks with a hidden dimension of 1,536, 24 attention heads, and a feed-forward dimension of 2,048. We configure a 46-dimensional state, a 78-dimensional action, one observation timestep, and a 40-step action chunk. Training uses the native flow objective

Table 12: Per-task in-distribution success rates (%) on the HUMANO Basic suite, evaluated over 25 rollouts per task.

Task	Diffusion Policy	Flow Matching	ACT	$\pi_{0.5}$	GR00T N1.7	Ψ_0
BallKick	8.0	0.0	12.0	96.0	24.0	8.0
ChairPush	40.0	0.0	32.0	8.0	100.0	72.0
ChairSit	88.0	20.0	52.0	100.0	56.0	0.0
DollyPush	4.0	0.0	4.0	60.0	8.0	16.0
GroundBasketPick	4.0	0.0	0.0	8.0	76.0	20.0
GroundBottlePick	0.0	0.0	12.0	24.0	4.0	0.0
GroundBoxPick	4.0	12.0	40.0	84.0	20.0	0.0
TableApplePick	0.0	0.0	0.0	8.0	12.0	0.0
TableBookPick	0.0	0.0	4.0	72.0	48.0	0.0
TablePlatePick	0.0	0.0	4.0	0.0	44.0	12.0
Avg.	14.8	3.2	16.0	46.0	39.2	12.8

Table 13: Per-task in-distribution success rates (%) on the HUMANO Advanced suite, evaluated over 25 rollouts per task.

Task	Diffusion Policy	Flow Matching	ACT	$\pi_{0.5}$	GR00T N1.7	Ψ_0
TableAppleToBasket	0.0	0.0	0.0	4.0	0.0	0.0
TablePlateToShelf	0.0	0.0	0.0	0.0	0.0	0.0
TableMilkToTrashCan	0.0	0.0	0.0	20.0	8.0	0.0
TableBreadToTray	0.0	0.0	0.0	0.0	0.0	0.0
GroundBottleToTable	0.0	4.0	0.0	20.0	8.0	4.0
GroundBoxToTable	0.0	0.0	4.0	36.0	16.0	4.0
GroundBasketToTable	0.0	28.0	4.0	96.0	60.0	72.0
ChairPushToTable	0.0	8.0	0.0	92.0	56.0	40.0
DollyPushToTable	0.0	0.0	0.0	20.0	12.0	4.0
TableEraserOutOfPenHolder	0.0	0.0	0.0	32.0	20.0	0.0
Avg.	0.0	4.0	0.8	32.0	18.0	12.4

with 1,000 training timesteps, while inference uses 10 sampling steps. The vision-language backbone remains frozen and only the action head is optimized.

Pretrained action-head tensors with incompatible dimensions are loaded by copying the overlapping entries, while newly introduced entries retain their default initialization. Training uses bfloat16 mixed precision, gradient checkpointing, and AdamW with a learning rate of 2×10^{-4}, 1,000 warm-up steps, cosine decay to 5×10^{-7}, and weight decay of 10^{-8}. We use the native state and action normalization schemes and image augmentation at 240×320 resolution.

C.3 Evaluation Protocol

All six baselines are evaluated using the same simulation and control pipeline. MuJoCo simulates robot and object dynamics, while Isaac Sim renders the corresponding visual observations. Each episode is initialized from a saved evaluation configuration containing the robot state, task objects, and scene configuration. Policy states and action buffers are reset between episodes, and language-conditioned policies receive the corresponding task instruction.

Each predicted 78-dimensional action is decomposed into a 64-dimensional SONIC motion token and two 7-dimensional hand commands. The SONIC controller converts the motion token into whole-body motor commands, while the hand commands control the corresponding hand joints. Predicted action chunks are stored in an execution queue and consumed sequentially at the action rate defined by the dataset. When the queue is exhausted, the policy is queried again using the latest

Table 14: Per-task in-distribution success rates (%) on the HUMANO Spatial suite, evaluated over 25 rollouts per task.

Task	Diffusion Policy	Flow Matching	ACT	$\pi_{0.5}$	GR00T N1.7	Ψ_0
BallKick	52.0	44.0	16.0	16.0	96.0	48.0
ChairPush	20.0	100.0	52.0	96.0	100.0	92.0
ChairPushToTable	24.0	76.0	60.0	88.0	80.0	60.0
ChairSit	0.0	0.0	0.0	0.0	4.0	4.0
DollyPush	76.0	12.0	48.0	64.0	76.0	12.0
DollyPushToTable	8.0	8.0	4.0	64.0	20.0	0.0
GroundBasketPick	8.0	60.0	24.0	84.0	96.0	68.0
GroundBasketToTable	0.0	8.0	0.0	64.0	92.0	40.0
GroundBottlePick	0.0	0.0	0.0	0.0	0.0	0.0
GroundBottleToTable	0.0	8.0	4.0	12.0	60.0	8.0
GroundBoxPick	0.0	16.0	8.0	0.0	24.0	24.0
GroundBoxToTable	0.0	24.0	12.0	36.0	84.0	32.0
TableApplePick	0.0	0.0	0.0	8.0	0.0	0.0
TableAppleToBasket	0.0	0.0	0.0	0.0	0.0	0.0
TableBreadToTray	0.0	0.0	0.0	0.0	0.0	0.0
TableButtonPress	0.0	0.0	4.0	0.0	20.0	0.0
TableEraserOutOfPenHolder	0.0	0.0	0.0	0.0	4.0	0.0
TableMilkToTrashCan	0.0	0.0	0.0	4.0	0.0	0.0
TablePlatePick	8.0	24.0	20.0	60.0	40.0	28.0
TablePlateToShelf	0.0	0.0	0.0	0.0	44.0	0.0
Avg.	9.8	19.0	12.6	29.8	42.0	20.8

Table 15: Per-task in-distribution success rates (%) on the HUMANO Object suite, evaluated over 25 rollouts per task.

Task	Diffusion Policy	Flow Matching	ACT	$\pi_{0.5}$	GR00T N1.7	Ψ_0
TableSmallObjectPick	12.0	8.0	12.0	40.0	32.0	4.0
TableLargeObjectPick	0.0	0.0	28.0	68.0	100.0	56.0
TableObjectPick	12.0	12.0	20.0	44.0	44.0	28.0
GroundSmallObjectPick	52.0	52.0	44.0	56.0	44.0	44.0
GroundLargeObjectPick	4.0	28.0	52.0	48.0	44.0	28.0
GroundObjectPick	12.0	48.0	36.0	32.0	8.0	24.0
TableSmallObjectToBasket	24.0	4.0	4.0	24.0	56.0	24.0
TableLargeObjectToCabinet	0.0	0.0	0.0	0.0	0.0	0.0
TableObjectToTrashCan	0.0	0.0	0.0	4.0	16.0	12.0
FoodToTray	0.0	20.0	20.0	8.0	48.0	8.0
Avg.	11.6	17.2	21.6	32.4	39.2	22.8

observation. The low-level controller continues executing the current motion command between policy updates.

The rollout horizon depends on the task suite. Long-Horizon tasks are allowed at most 7,000 environment steps, while tasks in the other seven suites are limited to 3,000:

$$T_{\max}^{(s)} = \begin{cases} 7{,}000, & s = \text{Long-Horizon}, \\ 3{,}000, & \text{otherwise}. \end{cases} \tag{12}$$

These limits count environment interactions during policy execution and exclude robot initialization, stabilization, controller warm-up, and action prefetching. An episode terminates when the task-specific success condition is satisfied, the environment terminates or truncates, the step budget is exhausted, or an execution-ending failure occurs. Episodes that reach the maximum horizon without completing the task are counted as failures.

Table 16: Per-task in-distribution success rates (%) on the HUMANO Goal suite, evaluated over 25 rollouts per task.

Task	Diffusion Policy	Flow Matching	ACT	$\pi_{0.5}$	GR00T N1.7	Ψ_0
GoalApplePlacement	0.0	12.0	0.0	4.0	20.0	0.0
GoalPlatePlacement	0.0	0.0	0.0	0.0	20.0	0.0
GoalBreadPlacement	0.0	12.0	0.0	12.0	24.0	0.0
GoalMilkPlacement	0.0	4.0	0.0	0.0	4.0	0.0
GoalObjectToBasket	0.0	0.0	0.0	0.0	0.0	0.0
GoalBookPlacement	0.0	0.0	0.0	0.0	0.0	0.0
GoalGroundBottlePlacement	0.0	0.0	0.0	0.0	0.0	0.0
GoalEraserPour	0.0	4.0	0.0	4.0	16.0	4.0
GoalStorageBoxPlacement	0.0	0.0	0.0	0.0	0.0	0.0
GoalFootballPlacement	0.0	0.0	0.0	0.0	0.0	4.0
Avg.	0.0	3.2	0.0	2.0	8.4	0.8

Task completion is determined by the task-specific success checker implemented in the environment. The final outcome additionally accounts for robot falls and detected scene-wall collisions. A fall is detected when a core body part—the pelvis, waist, or torso—remains in contact with the ground for at least 0.5 s. An episode is counted as successful only when the task success condition is satisfied without a recorded fall or execution-ending collision.

The primary evaluation metric is task success rate. For a suite containing K tasks, we compute

$$\mathrm{SR}_{\mathrm{suite}} = \frac{1}{K} \sum_{k=1}^{K} \left(\frac{1}{N_k} \sum_{i=1}^{N_k} y_{k,i} \right), \tag{13}$$

where N_k is the number of evaluation episodes for task k and $y_{k,i} \in \{0, 1\}$ denotes the final outcome of episode i. Under the standard HUMANO protocol, all tasks within the same evaluation setting use the same number of rollouts, making this equivalent to pooling episode outcomes within the suite.

All compared policies use the same evaluation configurations, execution interface, task success criteria, and suite-specific rollout budgets. Model configurations, preprocessing statistics, and training states required for inference are retained in the corresponding checkpoints. The HUMANO-specific modifications are limited to dataset conversion, observation and action interfaces, embodiment registration, checkpoint compatibility, and launcher configuration; the core architectures and learning objectives remain those of the original baseline implementations.

D Additional Experiments

D.1 Task-wise Results of In-distribution Setting

As shown in Tables 12–17, task-level performance varies both within and across suites. In the Basic suite, the strongest policies reach near-perfect performance on several tasks, including *BallKick*, *ChairPush*, and *ChairSit*, while tasks such as tabletop grasping and some floor-level interactions remain considerably harder. Similar variation is observed in the Advanced, Spatial, and Object suites, where high success on individual tasks coexists with complete failure on others. The contrast is most pronounced in the Goal and Conditional suites: several tasks remain unsolved by all evaluated policies, whereas a small subset achieves moderate to high success. Moreover, the best-performing policy varies across tasks, indicating complementary rather than uniform capabilities across generalist models. Overall, these results show that suite-level averages can obscure substantial task-specific differences and motivate fine-grained evaluation across diverse humanoid behaviors.

E Limitations

Despite its scale, HUMANO has several limitations. First, the current benchmark primarily focuses on indoor household environments and does not yet cover broader outdoor, industrial, or human-

Table 17: Per-task in-distribution success rates (%) on the HUMANO Conditional suite, evaluated over 25 rollouts per task.

Task	Diffusion Policy	Flow Matching	ACT	$\pi_{0.5}$	GR00T N1.7	Ψ_0
PlateToDrawer	0.0	0.0	0.0	0.0	0.0	0.0
AppleToDrawer	0.0	4.0	8.0	4.0	8.0	0.0
BottleToEmptyBox	0.0	0.0	0.0	8.0	12.0	0.0
AppleToFruitFreeBasket	4.0	12.0	12.0	56.0	44.0	0.0
ChairToTableAndSit	0.0	0.0	0.0	0.0	0.0	0.0
BasketToTableAppleToBasket	4.0	0.0	4.0	40.0	16.0	12.0
PeanutsToEmptyPlateOrTray	0.0	0.0	0.0	4.0	16.0	8.0
BreadToPlateOrTray	0.0	4.0	4.0	8.0	16.0	0.0
ClearChairAndPushDolly	0.0	20.0	20.0	60.0	48.0	12.0
BookToOccupiedShelfLayer	0.0	0.0	0.0	0.0	0.0	0.0
Avg.	0.8	4.0	4.8	18.0	16.0	3.2

interactive scenarios. Second, although the seven OOD settings capture several important sources of distribution shift, future extensions could further consider variations in physical properties, contact dynamics, terrain conditions, and robot hardware. Third, real-robot evaluation is currently limited to a small number of representative tasks, and broader physical validation remains an important direction for future work.