

HANDOFF: Humanoid Agentic Task-Space Whole-Body Control via Distilled Complementary Teachers

Lizhi Yang¹ Junheng Li¹ Nehar Poddar² Yiling Hou¹ Gio Huh¹
Robert Griffin² Georgia Gkioxari¹ Aaron D. Ames¹

¹California Institute of Technology ²The Institute for Human & Machine Cognition
{lzyang, junhengl, yhou, ghuh, georgia, ames}@caltech.edu
{npoddar, rgriffin}@ihmc.org

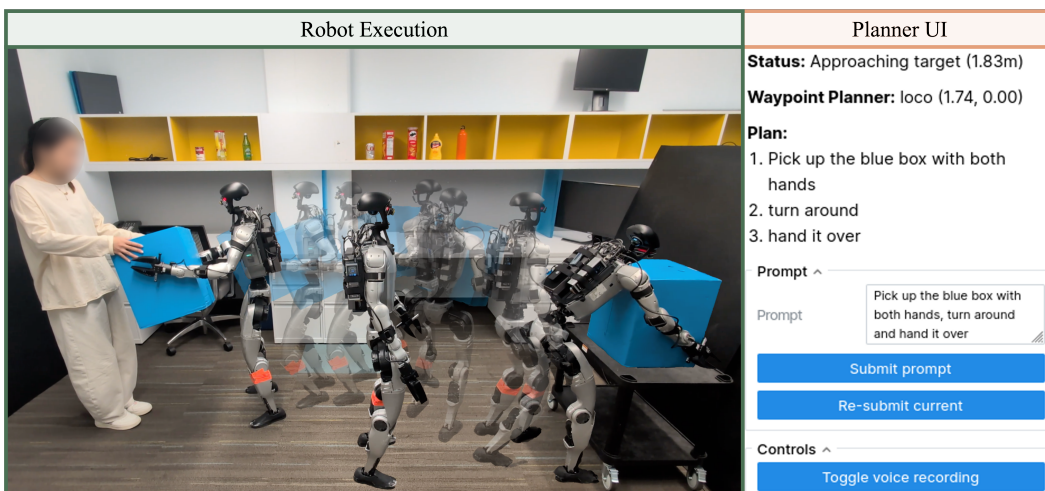


Figure 1: **HANDOFF** is a whole-body controller distilled from multiple teachers that accepts a compact, explicit 10-D planner-facing command. We demonstrate its effectiveness using a VLM-powered agentic planner that does not need extensive demonstration collection or model fine-tuning.

Abstract: For a humanoid robot to be deployed in the real world, the choice of command space (*i.e.*, the interface between task planning and whole-body control) is crucial. Existing whole-body controllers typically demand dense kinematic or spatial references that planners struggle to synthesize from task semantics. We instead propose a compact, explicit interface that is intuitive, general, modular, and expressive enough for diverse manipulation skills. To this end, we introduce **HANDOFF**, a single humanoid whole-body controller that follows this interface and is distilled via multi-teacher KL distillation under a context-conditioned gating scheme into a mixture-of-experts student from three complementary specialists: whole-body motion tracking with safety-filtered data, locomotion, and fall-recovery. On the Unitree G1, **HANDOFF** matches state-of-the-art velocity tracking and offers one of the largest robust manipulation workspaces. We further demonstrate hardware feasibility through multiple natural-language-driven task roll-outs, powered by a VLM-driven agentic planner with no task-specific data or controller fine-tuning.

Keywords: Reinforcement learning for physical robot control, Task and motion planning, Humanoid whole-body control, Loco-manipulation

1 Introduction

A humanoid robot fetching a cup of coffee for someone looks effortless in a video. The planning and control problem behind that loco-manipulation action is anything but: it has to coordinate object localization, sensing, motion planning, and whole-body control [1]. We ask the question: what does a planner actually want? As humans, we do not reason about exact joint positions; instead, we formulate sparse sub-goals (“find the coffee, walk near it, reach for it, grab it”), let lower-level motor reflexes handle each step of the walking gait, and recover from a stumble almost subconsciously. State-of-the-art whole-body controllers (WBC), however, mostly operate under the motion-tracking regime [2, 3, 4], requiring the planner to emit a dense, full-body kinematic stream at the controller rate. Producing such references requires collecting human teleoperation or motion-capture data, retargeting it to the specific humanoid embodiment, filtering for dynamic feasibility, and repeating that pipeline for every new skill or task variant. The planner is reduced to a data-replay engine tied to a particular library of demonstrations; a controller is only as useful as the commands a planner can realistically produce. What a planner truly wants, then, is a command space with four properties. It should be *intuitive* (a human, a geometric planner, or a VLM can each produce a valid executable command), *general* (one interface serves different loco-manipulation tasks), *modular* (planner, perception, and controller are decoupled and can be swapped independently), and *whole-body expressive* (compact commands still elicit coordinated full-body behavior).

We introduce HANDOFF, a task-space whole-body humanoid controller and an accompanying agent planner that targets a small, explicit command space:

$$c_t = [v_x, v_y, \omega_z, z, p_L^P, p_R^P],$$

where v_x, v_y, ω_z are planar base velocity commands, z is the commanded root height, and p_L^P, p_R^P are bilateral pelvis-frame wrist targets. Each component matches a planner family: locomotion stacks emit (v_x, v_y, ω_z) ; grasp planners emit a pelvis-frame end-effector target like $p_{L/R}^P$; any squat-or-reach heuristic sets z . Higher-level layers also operate at this abstraction: language-grounded task planners [5, 6] decompose long-horizon goals into atomic task-space subgoals; and VLAs [7, 8] emit end-effector actions that map onto our wrist-target slots, without any per-method retargeting or controller fine-tuning. At the same time, the same 10-D vector composes into whole-body actions: a low z paired with forward wrist targets induces a coordinated squat-and-reach, while an asymmetric wrist target during nonzero base velocity induces a one-handed reach-while-walking.

Such a controller must track task-space commands (base velocity, height, wrist-target reachability), produce coordinated whole-body manipulation behavior, and remain robust to disturbances such as recoverable falls; no single training regime delivers all three. We therefore pose this as a problem of *utilizing expert specialists*, each trained in its own regime on existing datasets, then composed into one deployable student. Concretely, we use a publicly available retargeted motion dataset to train a whole-body motion-tracking teacher [4, 9], a locomotion teacher with task-based velocity-tracking rewards [10], and a fall-recovery teacher trained with adversarial motion priors [11, 12]. Unlike WBC pipelines that often collect or retarget fresh full-body demonstrations for each new skill, our teachers are stand-alone modules that can be improved or replaced independently. We then distill these complementary teachers into a single deployable student using PPO [13] together with multi-teacher KL distillation [14] and mixture-of-experts training [15], under our context-conditioned gating scheme that routes each teacher’s supervision by a runtime regime signal. At runtime, the legs follow the locomotion teacher under nonzero commanded velocity, the arms follow the motion-tracking teacher throughout (enabling reaching, bi-manual coordination, and squatting), and the fall-recovery specialist assumes full supervision in niche situations. All three are distilled into one coordinated policy under the single 10-D interface, with no runtime controller switching.

The context-based distillation hook is also extensible: a new specialist (contact- or task-specific) plugs in as one new teacher head and one new context channel, with no changes to existing teachers or the command interface. By contrast, the existing landscape (Table 1) sits on a recurring trade-off: motion-tracking expressiveness has typically required dense, controller-specific interfaces; split-architecture controllers achieve a compact locomotion command but demand per-joint arm refer-

ences [16, 17]; and latent-action interfaces forfeit planner modularity [18]. HANDOFF targets all four properties at once through the distilled specialist mixture. We evaluate the controller in simulation on task-relevant physical metrics, deploy it through an agentic planner on a real Unitree G1 across multiple tasks, and will open source the full framework to facilitate future research.

2 Related Work

Whole-body motion-tracking with dense-reference interfaces. To enable diverse, expressive behaviors, some controllers expect a full-body kinematic reference. Motion-imitation training of physics-based characters [11, 19] now scales to humanoid robots via large-scale motion-tracking WBCs [2, 3, 4, 20, 21, 22] and masked multi-mode policies [23], with recent improvements on the residual [24], visual-hierarchical [25], data [26, 27], and real-time retargeting from human demonstrations [28, 29]. Such kinematic streams are hard for a planner to synthesize (Table 1); we trade some expressivity (like dancing) for an interface a simple planner can produce directly.

Task-command controllers and architectural decompositions. Other controllers take a compact mid-level command, differing in how the body is factored: HOMIE [16] drives the upper body from a cockpit with a separate RL lower body, FALCON [17] co-trains lower-body locomotion with a force-adaptive upper-body manipulator, and AMO [30] blends RL with dynamically optimized whole-body motion references. OmniH2O [31] and HOVER [23] expose a 3-point head-and-hands interface, but each point still needs a dense trajectory at controller rate; replacing the head with a planar base velocity removes that streaming requirement. Others keep the body holistic but still consume upper-body joint references [32], skill libraries [33], or per-task policies stacked on a learned controller [34, 35, 36, 37]; among those that scale across manipulation tasks [34, 35], adding a behavior requires fresh whole-body demonstrations and a new task policy. HANDOFF’s student consumes a 10-D command, so the layer above can be a planner, a VLM, or a learned end-effector policy whose action space is far smaller than a full-body trajectory.

Table 1: **Recent whole-body humanoid controllers by reference interface.** Each row shows what a method requires from the planner. The motion-tracking family needs a dense kinematic stream; the split-architecture velocity-command family takes a planar base command but still needs upper-body joint angles. *No ext. kin. ref.* marks whether the controller can be driven without an external joint-kinematic reference (e.g., without an IK solver or retargeter producing joint angles for any body part). In the frameworks listed here, HANDOFF is the only one that exposes a compact, planner-friendly interface.

Method	Locomotion ref.	Height ref.	Arm ref.	No dense streaming	No ext. kin. ref.	Single policy
Motion-tracking WBCs [2, 3, 4, 20, 21, 22]	kin. ref. (29D)	from kin.	full upper kin. (14D)	×	×	✓
ResMimic [24]	kin. ref. (29D)	from kin.	full upper kin. (14D)	×	×	×
VisualMimic [25]	keypoint stream (18D)	from kin.	keypoint stream (18D)	×	✓	✓
HOVER [23]	multi-mode kin. (~80D + mask)	from kin.	from kin.	×	×	✓
HOMIE / FALCON [16, 17]	velocity (2–6D)	scalar h (1D)	full upper kin. (14D)	✓	×	×
AMO [30]	velocity (3D)	torso pose (4D)	full upper kin. (14D)	✓	×	×
HUMI [35]	pelvis traj. (6D)	from pelvis	EE traj. (12D)	×	✓	✓
HERO [32]	velocity (6D)	scalar h (1D)	upper kin. (17D) + IK residual (6D)	✓	×	×
HANDOFF (Ours)	velocity (3D)	scalar h (1D)	wrist target (6D)	✓	✓	✓

VLA- and VLM-driven systems. Early planner–controller systems couple language models to skill libraries [5], and vision-language-action models train a single transformer to emit low-level actions [6, 7, 8]. Humanoid variants pre-train on egocentric video and post-train an action expert above a low-level controller [38, 39]; Being-0 [40] stacks a VLM connector on a modular RL/ACT skill library; LeVERB [18] rejects explicit interfaces and proposes a latent “verb” vocabulary. We do not replace this layer; the 10-D base-and-hand interface is a reasonable target for such planners.

Specialist-to-generalist distillation and expert routing. Recent humanoid work combines modes, locomotion behaviors, motion clusters, embodiment-specific specialists, parkour skills, and motion-tracking teachers into a single student [23, 41, 42, 43, 44]; closest to ours, GMT [22] also gates a Motion MoE, but over clusters of a single motion-tracking manifold driven by a dense kinematic reference, and TeleGate [45] keeps experts intact and trains a gating network at inference rather than collapsing them into one student. Earlier, MaskedMimic [46] distilled a single dense-

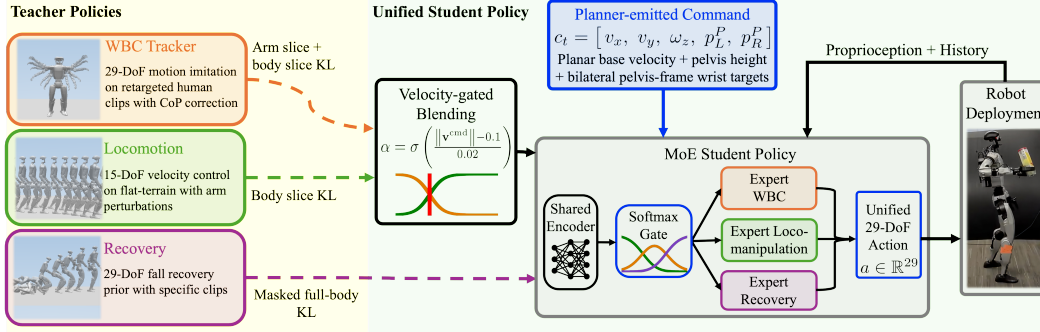


Figure 3: **System overview.** We train 3 teachers separately: a 29-DoF WBC motion-tracking teacher on CoP-filtered retargeted clips, a 15-DoF body-slice locomotion teacher on flat terrain with curriculum-blended arm perturbations, and a 29-DoF fall-recovery teacher on locomotion + paired fall-recovery clips. The MoE student maps a 10-D command and 11-frame proprioception history to 29-DoF actions. Context-based action-sliced KL pulls the body slice toward a velocity-gated blend of WBC and locomotion teachers, the arm slice toward WBC, and the fall-recovery teacher onto the full action when recovery is active; routing is shaped with load-balancing and recovery losses.

reference motion-tracking teacher into a partial-observation student via DAgger in simulated character animation, addressing heterogeneity across input modalities rather than across teacher regimes. We address a different source of heterogeneity: a regime conflict created by the compact interface itself; under fixed $[v_x, v_y, \omega_z, z, p_L^P, p_R^P]$, a whole-body human-data teacher is expressive but coverage-limited, while locomotion and fall-recovery teachers are reliable but specialized.

3 Whole-Body Control

We train our whole-body controller as seen in Fig.3 with a context-based multi-teacher distillation pipeline. First, we train the teachers independently with PPO in their respective regimes: a whole-body motion-tracking teacher on retargeted human motion clips, a locomotion teacher on flat terrain with curriculum-blended motion data, and a fall-recovery teacher on a curated mix of locomotion and paired fall-and-recovery sequences. We then distill them into a single student under a context-based KL scheme. The “context” is a per-step regime signal $\mathbf{x}_t = (\|c_t^{vel}\|, \text{recover}_t)$ (the commanded base-velocity magnitude and a binary recovery flag) that determines which teacher supervises which action slice (detailed in Section 3.4). Mixture-of-experts routing supervision and a load-balancing loss keep one expert active per regime. We first describe the three teachers that supply complementary supervision, then the student that fuses them via context-based, action-sliced KL distillation, and finally, the agentic planner that emits the commands the controller consumes.

3.1 Whole-Body Control Teacher

The whole-body teacher is a 29-DoF motion-tracking specialist trained with asymmetric actor-critic PPO [47] on retargeted human motion clips [9], following the data-driven tracking philosophy of recent motion-imitation works [3, 4, 20, 21]. The actor sees an 11-frame deployment-ready proprioceptive history together with the full 29-D reference joint angles from the current clip frame; the critic additionally consumes privileged signals unavailable at deployment (measured base linear velocity, reference root pose, key-body positions, and domain-randomization parameters). Full observation tables, reward weights, network sizes, and PPO hyperparameters are in Appendices A, B, and C.4. The trained teacher is a strong prior on posture, reach, squat, and bilateral coordination, but its locomotion behavior is

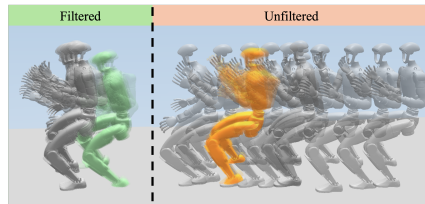


Figure 2: **CoP filtering.** The raw retargeted motion dataset contains dynamically infeasible frames, which we correct with a closed-form CBF projection on the static-CoP margin before training. An example is shown here where the corrected reference (left) stays within the support polygon and the unfiltered version (right) drifts.

anchored to the motion data, which makes it unreliable as a pure velocity tracker, as demonstrated in a velocity-tracking ablation shown in Table 2a. Furthermore, the motion data contains some dynamically infeasible squat frames that would be unsafe to track on the real robot, so we correct them with a closed-form CBF projection on the static-center-of-pressure (CoP) margin (Appendix C.1) before training the teacher. Fig. 2 illustrates the correction on a squat clip; the same filter is also applied at inference time to improve squat behavior.

3.2 Locomotion Teacher

The locomotion teacher is a 15-DoF body-slice (legs and waist) policy trained on flat terrain, with arms driven by curriculum-blended motion-data samples so that the policy is robust to the arm-induced CoM shifts during downstream distillation. The actor sees commanded velocity, projected gravity, base angular velocity, joint state, last action, and a 4-D gait-phase block; the critic additionally consumes privileged measured base linear velocity. The reward stack combines linear and angular velocity tracking with gait/stance shaping; full observation tables, the curricula for the arm motion and velocity, and PPO hyperparameters are in Appendices A, B, and C.4.

3.3 Fall-recovery Teacher

The mid-fall and just-recovered regime is supplied by an Adversarial Motion Prior (AMP) [11] teacher trained on a curated mix of locomotion and paired fall-and-recovery sequences, full-body 29-DoF, with the standard AMP discriminator + small torso-anchor task reward [12]. A fraction of environments are spawned at reset in a delayed fallen state to keep the recovery distribution well-represented. The discriminator architecture, dataset composition, recovery-reset curriculum, observation tables, and full reward weights and hyperparameters are in Appendices A, B, and C.4.

3.4 Planner-Friendly Whole-Body Controller

The student observes a planner-emitted 10-D command $c_t = [v_x, v_y, \omega_z, z, p_L^P, p_R^P]$ (planar base velocity, yaw rate, root height, and bilateral pelvis-frame wrist targets) and an 11-frame proprioception history, and emits 29-D actions through a soft Mixture-of-Experts head with one expert per teacher ($N = 3$), gated by a routing network over a shared encoder latent (Appendix C.3); soft routing keeps the policy differentiable and avoids the bimodal artifacts of hard top- k routing. The KL supervision is *context-conditioned*: a regime signal $\mathbf{x}_t = (\|c_t^{\text{vel}}\|, \text{recover}_t)$ drives both a continuous gate α that blends body-slice supervision between WBC and locomotion teachers and a binary mask $\mathbb{1}[\text{recover}_t]$ that routes the fall-recovery teacher in — soft blend and hard mask are two instances of the same context-driven gating. We split the action into a body slice $a^B = a_{0:15}$ (legs+waist) and arm slice $a^A = a_{15:29}$ (arms): the WBC teacher π_{wbc} and fall-recovery teacher π_{amp} define targets over both slices, while the locomotion teacher π_{loco} covers a^B only. Concretely, the body slice is supervised by a continuous-context convex blend of WBC and locomotion KL with sigmoid weight $\alpha = \sigma((\|c_t^{\text{vel}}\| - 0.1)/0.02)$, pulling toward the WBC teacher below 0.1 m/s commanded velocity and toward the locomotion teacher above it. The arm slice is supervised by the WBC teacher, except in recovery-active environments where the fall-recovery teacher assumes full supervision over the whole action space. The student is trained on $\mathcal{L} = \mathcal{L}_{\text{PPO}} + \lambda_B \mathcal{L}_{\text{KL}}^B + \lambda_A \mathcal{L}_{\text{KL}}^A + \lambda_{\text{AMP}} \mathcal{L}_{\text{KL}}^{\text{AMP}} + \beta_{\text{LB}} \mathcal{L}_{\text{LB}} + \beta_{\text{R}} \mathcal{L}_{\text{R}}$, where \mathcal{L}_{LB} is a standard MoE load-balancing loss [15] and \mathcal{L}_{R} pulls gate mass to a designated recovery expert; the framework extends to additional teachers by adding one expert head and gated-KL term per specialist. \mathcal{L}_{PPO} optionally absorbs a whole-body *stability reward stack* [48] (CoM- and LIPM capture-point-in-support-polygon, ankle/hip/step hierarchy, and linear/angular momentum-change penalties), shared across the WBC and locomotion teachers and the student; the AMP teacher is left untouched, and the stack composes with the recovery branch since it shapes the non-recovery distribution. Per-teacher inheritance and weights are in Appendix B. Full per-term observation tables for the student actor and critic are in Appendix A; reward and KL details are in Appendix B and Appendix C.2; velocity-tracking ablation is in Table 2a.

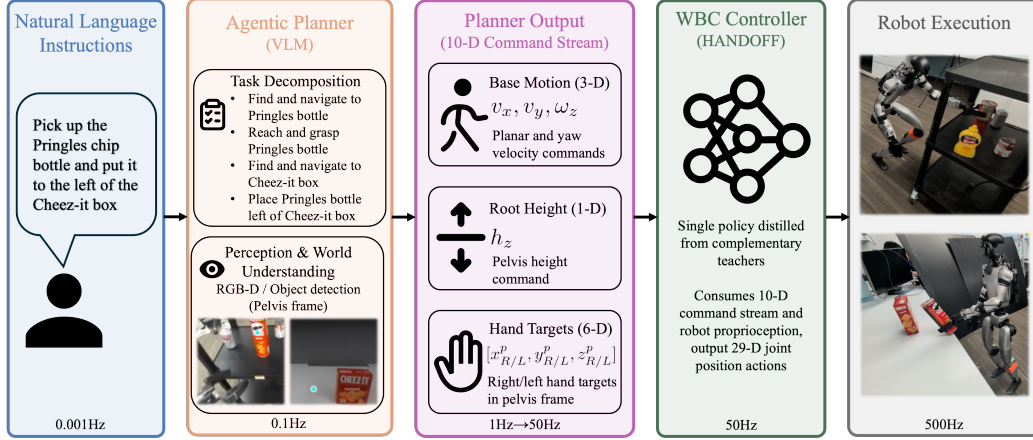


Figure 4: **Agentic deployment pipeline.** A natural-language instruction (0.001 Hz) is decomposed into atomic tasks by a high-level reasoner (regex + LLM fallback). A VLM (0.1 Hz) projects 2D detections onto the RGB-D point cloud to emit pelvis-frame waypoints; a waypoint tracker produces (v_x, v_y, ω_z) , and near the target a skill selector emits root height z and bilateral wrist targets $p_{R/L}^P$ (1–50 Hz) with a kinematics-based wrist correction. The resulting 10-D stream feeds the distilled controller (50 Hz), whose 29-DoF actions are tracked on hardware at 500 Hz.

3.5 Agentic planner

While everything above the controller is swappable as long as the planner emits the required 10-D command, we implement one concrete loco-manipulation stack (Fig.4). A high-level reasoner decomposes natural-language instructions into atomic tasks (regex parsing with an LLM fallback for novel instructions); a VLM then projects predicted 2D points and bounding boxes onto the RGB-D point cloud to emit pelvis-frame waypoints, from which a waypoint tracker derives (v_x, v_y, ω_z) . Once near the target, a skill selector emits root height z and bilateral wrist targets p_L^P, p_R^P from the target waypoint, with a simple kinematics-based wrist correction that aligns the gripper horizontally with the grasping surface.

4 Experiments

We conduct all experiments using a Unitree G1 (29 DoF) humanoid with an external payload (Jetson Thor + Dex1-1 grippers) on mjlabs/MuJoCo [49], and train all policies with the Rsl-rl framework [50]. We first ablate the design choices that drive velocity tracking and the manipulation workspace (Section 4.1), then compare against state-of-the-art whole-body controllers (Section 4.2), and finally demonstrate that the 10-D command space can be driven end-to-end by a modular agentic planner, showing the same controller driving multiple tasks in simulation and on hardware (Section 4.3) with zero data collection or model fine-tuning in the loop.

We evaluate the controller on two quantitative axes. For velocity tracking, we sweep per-axis velocity commands along $[-1, 1]$ and report the mean absolute error $|\Delta v| = \mathbb{E}_t[|v^{\text{cmd}}(t) - v^{\text{base}}(t)|]$ between the commanded and the realized base-twist component. For the manipulation workspace, we sample bilateral wrist targets uniformly inside $[-0.6, 0.6]^3$ m in the pelvis frame, with a 1s default-command warm-up, a 2s settle, and a 4s measurement window per target. We deem a trial feasible when both wrists stay within 15 cm of the target during measurement, the policy does not fall, and pelvis horizontal drift stays under 25 cm; we then report the robust workspace volume, defined as $\text{hull_vol} \times \text{feasible_frac}$, which is the effective volume the policy can both reach and stay feasible in, and restrict it to the forward half-space $\text{target}_x \geq 0$. The forward-half restriction is deliberate: every loco-manipulation task we care about, e.g., picking from a table, handing off, or placing onto a shelf, happens in front of the robot. Table 2b reports the quantitative results; full velocity sweeps (Fig.6 and Fig.7) and per-axis hull-volume comparisons (Fig.8) are in Appendix D.



Figure 5: **Agentic deployment snapshots with VLM text prompt.** The same 10-D controller, driven by the agentic planner, executing a range of loco-manipulation tasks on the Unitree G1 hardware and in simulation: pick-and-place, pick-transport-place, squat-pick, bimanual-pick-and-hand-off, bilateral pick-and-place, and task continuation after fall recovery. No controller-side change, data collection or model fine-tuning is required between tasks.

Table 2: **Velocity tracking and robust bilateral workspace.** Velocity errors are mean $|\text{cmd} - \text{realized}|$ over 20 commands per axis spanning the $[-1, 1]$ sweep. *Feas.* is the fraction of bilateral wrist-target trials with wrist error under 15 cm, no fall, and pelvis drift under 25 cm (Appendix D); *Robust WS* is the wrist-workspace hull volume scaled by feasibility in the forward half-space $x \geq 0$ (pelvis frame). Feasibility and robust workspace are computed over 2000 discovery and 400 accuracy bilateral wrist targets per controller (seed 42).

(a) Ablation progression.					(b) SOTA comparison.					
Ablation	$ \Delta v_x $	$ \Delta v_y $	$ \Delta \omega_z $	Robust WS	Method	$ \Delta v_x $	$ \Delta v_y $	$ \Delta \omega_z $	Feas.	Robust WS
	(m/s, rad/s) ↓			(m ³) ↑		(m/s, rad/s) ↓			(%) ↑	(m ³) ↑
Direct	0.29	0.43	0.08	0.20	FALCON [17]	0.07	0.10	0.07	35.5	0.06
+ Dual teacher	0.14	0.25	0.09	0.31	OpenHomie [16]	0.06	0.12	0.10	62.9	0.15
+ RandCmd	0.14	0.13	0.05	0.29	AMO [30]	0.06	0.19	0.01	80.9	0.22
+ Split KL + MoE (Ours)	0.07	0.14	0.04	0.27	SONIC [2]	0.03	0.15	0.02	89.8	0.26
+ AMP recovery (Ours + Rec.)	0.07	0.15	0.05	0.26	Ours	0.07	0.14	0.04	87.1	0.27
+ Stability (Ours + Stab.)	0.07	0.18	0.04	0.31	Ours + Rec.	0.07	0.15	0.05	92.0	0.26
+ Stab. + Rec. (Ours + Stab. + Rec.)	0.06	0.18	0.06	0.31	Ours + Stab.	0.07	0.18	0.04	97.7	0.31
					Ours + Stab. + Rec.	0.06	0.18	0.06	90.8	0.31

4.1 Ablation studies

The progression also shows why each specialist is necessary, since “Direct” (the motion-tracking teacher alone) is weakest on every axis: adding the standalone locomotion teacher (Direct \rightarrow + Dual teacher) drives the largest velocity-tracking jump; randomized commands (+ Dual teacher \rightarrow + RandCmd) close the lateral-velocity gap; context-based split-KL + MoE (Ours) closes the remaining v_x gap; the fall-recovery teacher (Ours+Rec.) adds survival, a binary capability absent unless distilled in; stability rewards (Ours+Stab.) push the robust workspace to 0.31 m³; and stacking both (Ours+Stab.+Rec.) matches that workspace while regaining the best $|\Delta v_x|$. The + Dual teacher attains a large workspace but lags Ours on v_x and ω_z , whereas Ours retains both.

4.2 Comparison with state-of-the-art whole-body controllers

Prior controllers do not natively expose pelvis-frame wrist targets, so this is an *adapted-interface* comparison: we equip each baseline with a differential-IK head [51] that maps (p_L^P, p_R^P) to arm joint targets while freezing non-arm DoFs, leaving the policy in full control of walking and balance (details in Appendix D). Differential IK is a *favorable* adapter — precise for instantaneous Cartesian targeting whenever a solution exists — yet it still adds an external reference-generation layer and does not guarantee dynamic whole-body feasibility. Even under this strong adapter, our velocity tracking sits within the SOTA cluster on every axis while delivering the largest robust workspace (0.31 m^3); Ours+Stab. takes the best feasibility (97.7%), and Ours+Stab.+Rec. regains the best $|\Delta v_x|$ among our variants. HANDOFF stays competitive because it learns wrist-target, locomotion, and balance coupling natively rather than through an external translation layer.

4.3 Agentic deployment in sim and real

The controller is planner-agnostic as long as the planner follows the 10-D command interface, be it traditional task planning, agentic planning, or even vision-language-action models. We demonstrate this end-to-end in two settings. (1) Sim: Loco-manipulation task continuation after fall recovery, under the same 10-D stream with no controller-side change, possible only because the recovery specialist was distilled in. (2) Real: on a 29-DoF Unitree G1 with onboard sensing (ZED-M RGB-D camera) and compute (Nvidia Jetson Thor with both local VLM and ChatGPT APIs) where the controller consumes the command stream produced by the planner. We test the robot in multiple tasks, including pick-and-place, pick-transport-place, squat-pick, bimanual-pick-and-hand-off, and bilateral pick-and-place. Fig. 5 and the video show representative rollouts.

Our quantitative claims concern the controller and its interface. The agentic planner is one representative implementation of this interface, and the hardware rollouts show it can be instantiated in an untethered real-robot stack without task-specific controller retraining.

5 Conclusion

We propose HANDOFF, a planner-friendly humanoid loco-manipulation controller built around a compact 10-D task-space interface that is expressive enough to encode loco-manipulation, yet small enough for diverse planners to drive without emitting full-body joint references. Single teacher-student distillation does not produce this cleanly: a motion-tracking teacher gives expressive posture priors but degenerate velocity tracking, while a locomotion teacher tracks velocity but loses whole-body coordination. We reconcile them via context-based distillation inside a mixture-of-experts student: the body slice follows a velocity-gated convex blend of the WBC and locomotion teachers, the arm slice anchors to the WBC teacher, and a recovery-masked KL term routes a third expert to a fall-recovery teacher. The resulting controller matches state-of-the-art velocity tracking and can be instantiated in a modular agentic planner for diverse simulation and real-robot task rollouts.

6 Limitations

Wrist-position targets. The interface exposes 3-D pelvis-frame wrist positions, not full 6-D gripper poses; a runtime kinematic correction (Section 3.5) handles tool-frame residuals, and direct 6-D tracking is future work.

Limited perception. The hardware uses a single fixed-pose head-mounted RGB-D camera, restricting perception to the forward field of view; gimballed head and wrist cameras are future work.

Specialist coverage. The teacher set is broad but not exhaustive; further specialists (terrain, contact, heavy load, etc.) will plug in as future work.

Acknowledgments

This research is supported in part by the Technology Innovation Institute (TII) and by The Dow Chemical Company project #227027AW.

References

- [1] Z. Gu, J. Li, W. Shen, W. Yu, Z. Xie, S. McCrory, X. Cheng, A. Shamsah, R. Griffin, C. K. Liu, et al. Humanoid locomotion and manipulation: Current progress and challenges in control, planning, and learning. *IEEE/ASME Transactions on Mechatronics*, 31(2):2300–2330, 2026.
- [2] Z. Luo, Y. Yuan, T. Wang, C. Li, S. Chen, F. Castaneda, Z.-A. Cao, J. Li, D. Minor, Q. Ben, et al. Sonic: Supersizing motion tracking for natural humanoid whole-body control. *arXiv preprint arXiv:2511.07820*, 2025.
- [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] 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. *arXiv preprint arXiv:2511.02832*, 2025.
- [5] B. Ichter, A. Brohan, Y. Chebotar, C. Finn, K. Hausman, A. Herzog, D. Ho, J. Ibarz, A. Irpan, E. Jang, et al. Do as i can, not as i say: Grounding language in robotic affordances. In K. Liu, D. Kulic, and J. Ichnowski, editors, *Proceedings of The 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pages 287–318. PMLR, 2023.
- [6] D. Driess, F. Xia, M. S. M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, et al. PaLM-e: An embodied multimodal language model. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 8469–8488. PMLR, 23–29 Jul 2023.
- [7] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In J. Tan, M. Toussaint, and K. Darvish, editors, *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pages 2165–2183. PMLR, 2023.
- [8] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. P. Foster, P. R. Sanketi, Q. Vuong, et al. Openvla: An open-source vision-language-action model. In P. Agrawal, O. Kroemer, and W. Burgard, editors, *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pages 2679–2713. PMLR, 06–09 Nov 2025.
- [9] Bones Studio. BONES-SEED: Skeletal everyday embodiment dataset. <https://bones.studio/datasets/seed>, 2026.
- [10] L. Yang, B. Werner, M. de Sa, and A. D. Ames. Cbf-rl: Safety filtering reinforcement learning in training with control barrier functions. *arXiv preprint arXiv:2510.14959*, 2025.
- [11] X. B. Peng, Z. Ma, P. Abbeel, S. Levine, and A. Kanazawa. Amp: Adversarial motion priors for stylized physics-based character control. *ACM Transactions on Graphics (ToG)*, 40(4): 1–20, 2021.
- [12] ccrpRepo. AMP_mjlab: G1 AMP motion control on mjlab + rsl_rl. https://github.com/ccrpRepo/AMP_mjlab, 2025.

- [13] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- [14] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *stat*, 1050: 9, 2015.
- [15] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.
- [16] Q. Ben, F. Jia, J. Zeng, J. Dong, D. Lin, and J. Pang. HOMIE: Humanoid Loco-Manipulation with Isomorphic Exoskeleton Cockpit. In *Proceedings of Robotics: Science and Systems*, Los Angeles, CA, USA, June 2025.
- [17] Y. Zhang, Y. Yuan, P. Gurunath, I. Gupta, S. Omidshafiei, A.-a. Agha-mohammadi, M. Vazquez-Chanlatte, L. Pedersen, T. He, and G. Shi. Falcon: Learning force-adaptive humanoid loco-manipulation. *arXiv preprint arXiv:2505.06776*, 2025.
- [18] H. Xue, X. Huang, D. Niu, Q. Liao, T. Kragerud, J. T. Gravdahl, X. B. Peng, G. Shi, T. Darrell, K. Sreenath, et al. Leverb: Humanoid whole-body control with latent vision-language instruction. *arXiv preprint arXiv:2506.13751*, 2025.
- [19] X. B. Peng, P. Abbeel, S. Levine, and M. Van de Panne. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions On Graphics (TOG)*, 37(4):1–14, 2018.
- [20] Z. Fu, Q. Zhao, Q. Wu, G. Wetzstein, and C. Finn. Humanplus: Humanoid shadowing and imitation from humans. In *Conference on Robot Learning*, pages 2828–2844. PMLR, 2025.
- [21] M. Ji, X. Peng, F. Liu, J. Li, G. Yang, X. Cheng, and X. Wang. Exbody2: Advanced expressive humanoid whole-body control. *arXiv preprint arXiv:2412.13196*, 2024.
- [22] Z. Chen, M. Ji, X. Cheng, X. Peng, X. B. Peng, and X. Wang. Gmt: General motion tracking for humanoid whole-body control. *arXiv preprint arXiv:2506.14770*, 2025.
- [23] T. He, W. Xiao, T. Lin, Z. Luo, Z. Xu, Z. Jiang, J. Kautz, C. Liu, G. Shi, X. Wang, et al. Hover: Versatile neural whole-body controller for humanoid robots. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9989–9996. IEEE, 2025.
- [24] S. Zhao, Y. Ze, Y. Wang, C. K. Liu, P. Abbeel, G. Shi, and R. Duan. Resmimic: From general motion tracking to humanoid whole-body loco-manipulation via residual learning. *arXiv preprint arXiv:2510.05070*, 2025.
- [25] S. Yin, Y. Ze, H.-X. Yu, C. K. Liu, and J. Wu. Visualmimic: Visual humanoid loco-manipulation via motion tracking and generation. *arXiv preprint arXiv:2509.20322*, 2025.
- [26] F. Liu, Z. Gu, Y. Cai, Z. Zhou, H. Jung, J. Jang, S. Zhao, S. Ha, Y. Chen, D. Xu, et al. Opt2skill: Imitating dynamically-feasible whole-body trajectories for versatile humanoid loco-manipulation. *IEEE Robotics and Automation Letters*, 2025.
- [27] L. Yang, X. Huang, Z. Wu, A. Kanazawa, P. Abbeel, C. Sferrazza, C. K. Liu, R. Duan, and G. Shi. Omniretarget: Interaction-preserving data generation for humanoid whole-body loco-manipulation and scene interaction. *arXiv preprint arXiv:2509.26633*, 2025.
- [28] L. Penco, B. Clément, V. Modugno, E. M. Hoffman, G. Nava, D. Pucci, N. G. Tsagarakis, J.-B. Mouret, and S. Ivaldi. Robust real-time whole-body motion retargeting from human to humanoid. In *2018 IEEE-RAS 18th International Conference on Humanoid Robots (Humanoids)*, pages 425–432. IEEE, 2018.

- [29] J. P. Araujo, Y. Ze, P. Xu, J. Wu, and C. K. Liu. Retargeting matters: General motion retargeting for humanoid motion tracking. *arXiv preprint arXiv:2510.02252*, 2025.
- [30] 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. In *Proceedings of Robotics: Science and Systems*, Los Angeles, CA, USA, June 2025. doi:10.15607/RSS.2025.XXI.061.
- [31] T. He, Z. Luo, X. He, W. Xiao, C. Zhang, W. Zhang, K. M. Kitani, C. Liu, and G. Shi. Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning. In *Conference on Robot Learning*, pages 1516–1540. PMLR, 2025.
- [32] R. Dong, Z. Li, X. He, and S. Gupta. Learning humanoid end-effector control for open-vocabulary visual loco-manipulation. *arXiv preprint arXiv:2602.16705*, 2026.
- [33] Z. Zhang, C. Chen, H. Xue, J. Wang, S. Liang, Y. Liu, Z. Zhang, H. Wang, and L. Yi. Unleashing humanoid reaching potential via real-world-ready skill space. *IEEE Robotics and Automation Letters*, 11(2):2082–2089, 2025.
- [34] Y. Fu, F. Xie, C. Xu, J. Xiong, H. Yuan, and Z. Lu. Demohlm: From one demonstration to generalizable humanoid loco-manipulation. *arXiv preprint arXiv:2510.11258*, 2025.
- [35] R. Nai, B. Zheng, J. Zhao, H. Zhu, S. Dai, Z. Chen, Y. Hu, Y. Hu, T. Zhang, C. Wen, et al. Humanoid manipulation interface: Humanoid whole-body manipulation from robot-free demonstrations. *arXiv preprint arXiv:2602.06643*, 2026.
- [36] Z. Su, B. Zhang, N. Rahmanian, Y. Gao, Q. Liao, C. Regan, K. Sreenath, and S. S. Sastry. Hitter: A humanoid table tennis robot via hierarchical planning and learning. *arXiv preprint arXiv:2508.21043*, 2025.
- [37] J. Dao, H. Duan, and A. Fern. Sim-to-real learning for humanoid box loco-manipulation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 16930–16936. IEEE, 2024.
- [38] H. Jiang, J. Chen, Q. Bu, L. Chen, M. Shi, Y. Zhang, D. Li, C. Suo, C. Wang, Z. Peng, et al. Wholebodyvla: Towards unified latent vla for whole-body loco-manipulation control. *arXiv preprint arXiv:2512.11047*, 2025.
- [39] S. Wei, H. Jing, B. Li, Z. Zhao, J. Mao, Z. Ni, S. He, J. Liu, X. Liu, K. Kang, et al. Ψ_0 : An open foundation model towards universal humanoid loco-manipulation. *arXiv preprint arXiv:2603.12263*, 2026.
- [40] H. Yuan, Y. Bai, Y. Fu, B. Zhou, Y. Feng, X. Xu, Y. Zhan, B. F. Karlsson, and Z. Lu. Being-0: A humanoid robotic agent with vision-language models and modular skills. *arXiv preprint arXiv:2503.12533*, 2025.
- [41] Y. Zhao, X. Wang, D. Wang, X. Liu, D. Lu, Q. Han, P. Liu, and C. Bai. Towards adaptive humanoid control via multi-behavior distillation and reinforced fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 18818–18826, 2026.
- [42] Y. Wang, M. Yang, G. Ding, Y. Zhang, W. Zeng, X. Xu, H. Jiang, and Z. Lu. From experts to a generalist: Toward general whole-body control for humanoid robots. *Advances in Neural Information Processing Systems*, 38:147748–147772, 2026.
- [43] Q. Peng, Y. Lin, Y. Xue, J. Pang, and W. Zhang. Embodiment-aware generalist specialist distillation for unified humanoid whole-body control. *arXiv preprint arXiv:2602.02960*, 2026.
- [44] Z. Wu, X. Huang, L. Yang, Y. Zhang, K. Sreenath, X. Chen, P. Abbeel, R. Duan, A. Kanazawa, C. Sferrazza, et al. Perceptive humanoid parkour: Chaining dynamic human skills via motion matching. *arXiv preprint arXiv:2602.15827*, 2026.

- [45] J. Li, B. Tang, and F. Wu. Telegate: Whole-body humanoid teleoperation via gated expert selection with motion prior. *arXiv preprint arXiv:2602.09628*, 2026.
- [46] C. Tessler, Y. Guo, O. Nabati, G. Chechik, and X. B. Peng. Maskedmimic: Unified physics-based character control through masked motion inpainting. *ACM Transactions On Graphics (TOG)*, 43(6):1–21, 2024.
- [47] L. Pinto, M. Andrychowicz, P. Welinder, W. Zaremba, and P. Abbeel. Asymmetric actor critic for image-based robot learning. *Robotics: Science and Systems XIV*, 2018.
- [48] N. Poddar, S. McCrory, L. Penco, G. Clark, H. E. Svitl, and R. Griffin. Embedding classical balance control principles in reinforcement learning for humanoid recovery. *arXiv preprint arXiv:2603.08619*, 2026.
- [49] K. Zakka, Q. Liao, B. Yi, L. L. Lay, K. Sreenath, and P. Abbeel. mjlab: A lightweight framework for gpu-accelerated robot learning. *arXiv preprint arXiv:2601.22074*, 2026.
- [50] C. Schwarke, M. Mittal, N. Rudin, D. Hoeller, and M. Hutter. Rsl-rl: A learning library for robotics research. *arXiv preprint arXiv:2509.10771*, 2025.
- [51] K. Zakka. mink: Python inverse kinematics based on MuJoCo. <https://github.com/kevinzakka/mink>, 2024.

A Observations

This section enumerates the actor and critic observation groups used by each policy. Asymmetric actor-critic is in force throughout: anything in a *critic-only* group is privileged information used to fit the value function and is unavailable to the deployed actor.

A.1 Whole-Body Motion-Tracking Teacher

Group	Term	Dim	Description
actor (current)	joint position (relative)	29	per-joint angle relative to default
	joint velocity	29	per-joint angular velocity
	IMU roll/pitch	2	roll and pitch from IMU
	base angular velocity	3	IMU body-frame ω
	last action	29	previous policy action
	ref. root xy-velocity (body)	2	reference root linear velocity in body frame
	ref. root height	1	reference root z from motion clip
	ref. root roll/pitch	2	reference root orientation (roll, pitch)
	ref. root yaw ang. vel.	1	reference root yaw angular velocity (body frame)
	ref. joint angles	29	full 29-D reference joint angles from current clip frame
actor (history)	11-frame stack of <i>actor (current)</i>	64	encoded by 1-D temporal-conv into a 64-D latent
critic (extras)	measured base linear velocity	3	not available to actor
	reference root pose	7	reference root position (3) + quaternion (4) from motion clip
	reference key-body positions	27	9 key bodies (wrists, knees, ankles, elbows, torso) $\times 3$ in body frame
	foot friction, motor strength scales, base mass perturbation, encoder bias, applied hand forces	95	domain-randomization parameters

A.2 Locomotion Teacher

Group	Term	Dim	Description
actor (current)	projected gravity	3	gravity in IMU frame
	base angular velocity	3	IMU body-frame ω
	twist command	3	$[v_x^{\text{cmd}}, v_y^{\text{cmd}}, \omega_z^{\text{cmd}}]$
	joint position (relative)	29	all 29 joints
	joint velocity	29	
	last action	15	body-slice action only
	gait-phase features	4	$[\sin \phi_L, \cos \phi_L, \sin \phi_R, \cos \phi_R]$
critic (extras)	measured base linear velocity	3	not available to actor
	foot friction, motor strength scales, base mass perturbation, encoder bias, applied hand forces	95	domain-randomization parameters

A.3 Fall-recovery Teacher

Group	Term	Dim	Description
actor	base angular velocity	3	IMU body-frame ω
	projected gravity	3	gravity in IMU frame
	twist command	3	$[v_x^{\text{cmd}}, v_y^{\text{cmd}}, \omega_z^{\text{cmd}}]$

Group	Term	Dim	Description
	joint position (relative)	29	
	joint velocity	29	
	last action	29	full body
	(per-frame total 96; history length 4, flattened)	$\times 4$	no temporal encoder; MLP input grows $4\times$ (total 384)
critic	all actor terms	96	same per-frame terms as actor; flattened over the 4-frame history
	measured base linear velocity	3	privileged
	body positions in body frame	39	13 reference bodies $\times 3$
	body orientations in body frame	78	13 bodies $\times 6$ (6-D rotation, two rows of rotation matrix)
discriminator	body positions, orientations, linear and angular velocities in body frame, single frame	195	13 bodies $\times (3 + 6 + 3 + 3)$; input to the AMP discriminator on (s_t, s_{t+1}) pairs (total 390)

A.4 Student Policy

Group	Term	Dim	Description
actor (current)	joint position (relative)	29	
	joint velocity	29	
	projected gravity	3	
	base angular velocity	3	
	last action	29	
	planner command	10	10-D command
			$[v_x, v_y, \omega_z, z, p_L^P, p_R^P]$
	gait-phase features	4	$[\sin \phi_L, \cos \phi_L, \sin \phi_R, \cos \phi_R]$ (internal block, not part of the planner interface)
actor (history)	11-frame stack of <i>actor (current)</i>	64	encoded by 1-D temporal-conv into a 64-D latent
critic (extras)	measured base linear velocity	3	privileged information
	foot friction, motor strength scales, base mass perturbation, encoder bias, applied hand forces	95	domain-randomization parameters
recovery flag	per-env binary recovery indicator	1	used to gate the AMP-recovery KL term and the recovery-routing loss
blend weight	$\alpha = \sigma((\ c^{\text{vel}}\ - 0.1)/0.02)$	1	used to interpolate body-slice KL between WBC and locomotion teachers

B Rewards

This section lists every reward term and its weight, exactly as instantiated in the canonical training configuration. Where possible, the reward kernel is given in compact form.

B.1 Whole-Body Motion-Tracking Teacher

Term	Kernel	Weight
<i>Tracking (TWIST2)</i>		
tracking_joint_dof	$\exp(-0.15 \sum_i w_i^2 (q_i^{\text{ref}} - q_i)^2)$	+2.0
tracking_joint_vel	$\exp(-0.01 \sum_i w_i^2 (\dot{q}_i^{\text{ref}} - \dot{q}_i)^2)$	+0.2
tracking_root_translation_z	$\exp(-5 (z^{\text{ref}} - z)^2)$	+1.0

Term	Kernel	Weight
tracking_root_rotation	$\exp(-5 \ q_{\text{err}}\ ^2)$ on roll/pitch/yaw quaternion error	+1.0
tracking_root_linear_vel	$\exp(-\ v^{\text{ref}} - v\ ^2/\sigma^2)$, $\sigma = 1.0$	+1.0
tracking_root_angular_vel	$\exp(-\ \omega^{\text{ref}} - \omega\ ^2/\sigma^2)$, $\sigma \approx \pi$	+1.0
tracking_keybody_pos	$\exp(-10 \sum_b \ p_b^{\text{ref}} - p_b\ ^2)$ in body-local frame	+2.0
tracking_keybody_pos_global	same kernel, world-frame body positions	+2.0
<i>Regularization</i>		
alive	1[not terminated]	+0.5
feet_slip	exp-cost on tangential foot velocity during contact	-0.1
feet_contact_forces	penalty on contact force magnitude, clipped at 500 N	-5×10^{-4}
feet_stumble	contact transitions outside expected stance	-1.25
dof_pos_limits	Gaussian-kernel penalty near joint limits	-5.0
dof_torque_limits	penalty for torque above $0.95\tau_{\text{max}}$	-1.0
dof_vel	$\sum_i \dot{q}_i^2$	-10^{-4}
dof_acc	$\sum_i \ddot{q}_i^2$	-5×10^{-8}
action_rate_l2	$\ a_t - a_{t-1}\ ^2$	-0.1
joint_limit	duplicate joint-limit term, stronger weight	-10.0
self_collisions	contact-force exceedance (> 10 N) on collision sensor	-10.0
feet_air_time	target-swing-time bonus (target 0.5 s, gait-gated)	+5.0
ang_vel_xy	$\omega_x^2 + \omega_y^2$	-0.01
ankle_dof_acc	ankle-only acceleration penalty	-10^{-7}
ankle_dof_vel	ankle-only velocity penalty	-2×10^{-4}

B.2 Locomotion Teacher

Term	Kernel	Weight
<i>Tracking</i>		
lin_vel_tracking	$\exp(-((v_x - v_x^{\text{cmd}})^2 + (v_y - v_y^{\text{cmd}})^2)/\sigma^2)$, $\sigma = 1$	+1.0
ang_vel_tracking	$\exp(-(\omega_z - \omega_z^{\text{cmd}})^2/\sigma^2)$, $\sigma \approx \pi$	+1.0
<i>Gait and stance shaping</i>		
pose	body-joint deviation from default, gait-conditioned σ	$-0.15 \rightarrow -0.5$
foot_clearance	swing-foot height penalty below 0.05 m	-6.0
foot_swing_height	swing-foot height penalty below 0.08 m	-0.75
stand_pose	default-pose penalty when $\ c^{\text{vel}}\ < 0.1$	-5.0
flat_foot	foot-tilt penalty on contact: tangential gravity component squared	-0.5
gait_phase_contact	contact-vs-phase mismatch reward (stance ratio 0.6)	+0.5
feet_distance_lateral	penalty if lateral foot distance leaves $[0.2, 0.35]$ m	+0.5
knee_distance_lateral	penalty if lateral knee distance leaves $[0.2, 0.35]$ m	+1.0
<i>Regularization</i>		
foot_slip	tangential foot velocity during contact	-0.25
action_rate_l2	$\ a_t - a_{t-1}\ ^2$	-0.01
joint_acc_l2	$\sum_i \ddot{q}_i^2$	-2.5×10^{-7}
joint_pos_limits	Gaussian-kernel penalty near joint limits	-10.0
self_collisions	contact-force exceedance on collision sensor	-10.0

Term	Kernel	Weight
termination_penalty	1[terminated]	-200

B.3 fall-recovery teacher

The fall-recovery teacher mixes a discriminator-based motion-prior reward with a small anchor-tracking task reward.

Term	Kernel	Weight
<i>Anchor-tracking task</i>		
track_anchor_linear_velocity	$\exp(-\ v^{\text{anch}} - v^{\text{cmd}}\ ^2/1.0^2)$ on torso anchor	+1.0
track_anchor_angular_velocity	$\exp(-(\omega_z^{\text{anch}} - \omega_z^{\text{cmd}})^2/3.14^2)$	+1.0
track_root_height	$\exp(-(z - z^{\text{cmd}})^2/0.3^2)$ with delay masking	+1.0
body_ang_vel_xy_l2	$(\omega_x^2 + \omega_y^2)/3.14^2$ on root body	+0.5
<i>Regularization</i>		
is_terminated	1[terminated]	-200
joint_acc_l2	$\sum_i \ddot{q}_i^2$	-2.5×10^{-7}
joint_pos_limits	Gaussian-kernel penalty near joint limits	-10.0
action_rate_l2	$\ a_t - a_{t-1}\ ^2$	-0.01
foot_slip	tangential foot velocity on contact	-0.25
self_collisions	contact-force exceedance on collision sensor	-0.1

The discriminator is trained on (state, next-state) pairs drawn from the AMP motion buffer. The total reward delivered to PPO is

$$r_{\text{AMP}} = (1 - \alpha) r_{\text{disc}} + \alpha r_{\text{task}}, \quad \alpha = 0.75,$$

with discriminator loss coefficient 1.0 and gradient-penalty coefficient $\lambda_{\text{gp}} = 10$. Up to 40% of envs are spawned in delayed (fallen) reset states with a maximum delay of 250 steps to keep the recovery distribution well-represented.

B.4 Student

The student keeps a small task-reward stack on top of the distillation losses described in Section 3.4.

Term	Kernel	Weight
<i>Task</i>		
tracking_hand_pos	exp-kernel on bilateral pelvis-frame wrist-position error	+6.0
tracking_root_translation.z	root-height tracking	+1.0
tracking_root_rotation	root-orientation tracking	+1.0
tracking_root_linear_vel	exp-kernel on commanded base-velocity error	+1.0
tracking_root_angular_vel	exp-kernel on commanded yaw-rate error	+1.0
gait_phase_contact (when on)	motion-conditioned gait-phase contact reward	+0.5
<i>Inherited TWIST2 regularization</i>		
All regularization terms from the WBC motion-tracking teacher above except joint_limit and self_collisions, which are dropped		—

Distillation losses. On top of the task rewards, the actor receives the three context-conditioned KL terms detailed in Section 3.4: a body-slice KL convex-blended between WBC and locomotion teachers under continuous velocity context, an arm-slice KL anchored to the WBC teacher, and a discrete-context AMP KL masked to recovery-active samples on the full action vector. In the canonical run, the cosine-annealed coefficients are

$$\lambda_B : 0.4 \rightarrow 0.2, \quad \lambda_A : 0.1 \rightarrow 0.05, \quad \lambda_{AMP} : 0.4 \rightarrow 0.2,$$

over the first 60,000 update steps. The two MoE routing-shaping terms add a load-balance loss with coefficient 0.01 and a recovery-routing loss with coefficient 0.5. Full PPO hyperparameters and the AMP-teacher discriminator settings are in Appendix C.4.

Stability reward stack (Ours+Stab., Ours+Stab. +Rec.). Both the Ours+Stab. and Ours+Stab. +Rec. variants add a shared whole-body stability stack to the regularization rewards used by the WBC motion-tracking teacher, the locomotion teacher, and the student — i.e., all three policies are retrained with the stack enabled, since the terms shape balance behavior at the source. The Ours+Stab. +Rec. variant additionally pulls in the AMP fall-recovery teacher and the full 3-teacher MoE / recovery-routing loss machinery exactly as in Ours+Rec.; the stability stack composes cleanly with the recovery branch because it shapes the non-recovery distribution, while the AMP teacher continues to dominate the recovery-flagged samples through the discrete-context KL of Section 3.4. The AMP teacher itself is left untouched (it has its own anchor-tracking and discriminator reward). The stack is omitted from the main-text per-teacher reward subsections to keep them aligned with the canonical (Ours / Ours+Rec.) runs; the methods section (Section 3) refers to it only as a PPO add-on, and the full per-teacher mapping is given here. All terms are computed from the entity-level mass-weighted whole-body CoM and key kinematic / contact signals, ported from a public IsaacLab-IHMC stability suite [48] and tuned against the G1’s 29-DoF biped morphology.

Term	Kernel	Weight
<code>com.in.support.polygon</code>	$\exp(-d^2/\sigma^2)$, d = horizontal distance from whole-body CoM to the active feet-contact AABB (tol. 5 cm, $\sigma=0.1$); polygon extended to active hand contacts when present (fall-recovery)	+0.2
<code>capture.point.in.support.polygon</code>	$\exp(-d^2/\sigma^2)$ with d = LIPM capture-point distance to feet polygon, $CP = CoM_{xy} + CoM_{xy}/\sqrt{g/h_{CoM}}$; full reward when leaning against a wall/table sensor	+0.3
<code>ankle.hip.step</code>	Ankle→Hip→Step strategy hierarchy: ankle torque opposes CoM drift (gated by friction-cone CoP saturation), hip activates when ankle saturates, swing-foot velocity toward capture point activates when both saturate	+0.2
<code>linear.momentum.change</code>	$-\ \dot{p}\ ^2$, $p = \sum_i m_i v_i$ whole-body linear momentum	-1×10^{-6}
<code>angular.momentum.change</code>	$-\ \dot{L}\ ^2$, $L = \sum_i (r_i - r_{CoM}) \times m_i v_i$ orbital angular momentum (spin term omitted)	-1×10^{-5}

The CoM / capture-point terms use an axis-aligned bounding box of active foot contacts (force threshold 20 N for the CoM term, 50 N for the capture-point term) as a fast, GPU-friendly proxy for the convex-hull support polygon; the approximation is exact in double stance and incurs negligible error in single-stance walking. The ankle/hip/step strategy reads ankle and hip torques from `qfrc.actuator`, normalizes them by the per-joint torque limit, and additionally penalizes CoP saturation against a $\mu = 0.7$ friction cone, so that ankle/hip torque rewards saturate exactly when the foot is about to slip and the step term takes over.

C Implementation Details

C.1 Motion-Data Curation

This appendix gives the full per-frame projection used by the CoP-feasibility filter described in Section 3.1.

Static capture-point definition. In the quasi-static regime ($\dot{c}_{xy} \approx 0$) the LIP capture point reduces to the planar projection of the CoM, $\xi_{\text{stat}} = c_{xy}$. For each frame, forward kinematics through Pinocchio gives the 29-DoF body pose; the support polygon is the axis-aligned bounding box of the eight foot-contact corners (four per ankle-roll link), shrunk inward by a safety margin $\delta = 0.10$ m to give \mathcal{S}_δ . The signed distance from the static CoP to \mathcal{S}_δ is

$$h(r, r_q, q; \delta) = \min\{c_x - x_{\min}, x_{\max} - c_x, c_y - y_{\min}, y_{\max} - c_y\},$$

positive when the projected CoM is strictly inside the polygon, zero on its boundary, and negative when the frame is unbalanced.

Squat detection. The filter triggers only on frames flagged as quasi-static squats: root height below 0.65 m, root XY-speed below 0.2 m/s. A 7-frame morphological erosion removes isolated detections, and clips with fewer than 15 consecutive squat frames are skipped entirely. Frames already satisfying $h \geq h_{\text{tgt}}$ with $h_{\text{tgt}} = 0.04$ m pass through unchanged.

Joint-correction subspace. Unsafe frames are projected onto the safe set in a 7-D joint-correction subspace spanned by bilateral hip pitch, ankle pitch, ankle roll, and waist pitch — the joints with the most kinematic authority over CoM displacement. Treating the discrete CBF condition $h(q + Eu) \geq h_{\text{tgt}}$ as an affine inequality after a first-order expansion, the minimum-effort correction admits a closed-form half-space projection,

$$u^*(q) = \frac{\max(0, h_{\text{tgt}} - h(q))}{\|a(q)\|^2} a(q), \quad a(q) = E^\top [J_{c,xy}^{\text{cc}}]^\top \nabla_\xi h,$$

where $J_{c,xy}^{\text{cc}}$ is the contact-consistent CoM Jacobian under the rigid double-support constraint $J^{\text{feet}} \dot{q} = 0$. We iterate this Newton step inside an Armijo backtracking line search to handle the polygon’s piecewise-linear geometry, then re-anchor the floating base so that the mid-feet pose is invariant under the correction. A 10-frame linear ramp blends corrected and raw trajectories at the boundaries of each squat segment.

Deployment-time velocity-space filter. The same barrier h , gradient ∇h , and contact-consistent CoM Jacobian power a deployment-time filter applied on top of the student’s commands at inference, so that the squat behavior learned from filtered data is preserved on the real platform. It shares the analytic geometry of the offline filter described above and acts on the same 7-DoF subspace (bilateral hip pitch, ankle pitch, ankle roll, and waist pitch), differing only in that it projects on the velocity command \dot{q} scaled by the LIP lookahead $\Delta t + 1/\omega_0$, with $\omega_0 = \sqrt{g/h_{\text{CoM}}}$.

C.2 Distillation Internals

The student loss is *context-conditioned* on $\mathbf{x}_t = (\|c_t^{\text{vel}}\|, \text{recover}_t)$: a continuous component (velocity magnitude) drives the body-slice blend, and a binary component (recovery flag) drives the AMP mask. The first two paragraphs below are the soft and hard instances of the context-dependent KL; the third gives the per-dimension reduction shared by all KL terms.

Continuous-context body-slice KL. The body-slice supervision is a per-step convex blend of WBC and locomotion KL,

$$\mathcal{L}_{\text{KL}}^B = (1 - \alpha) D_{\text{KL}}(\pi_\theta^B \parallel \pi_{\text{wbc}}^B) + \alpha D_{\text{KL}}(\pi_\theta^B \parallel \pi_{\text{loco}}^B), \quad \alpha = \sigma\left(\frac{\|c_t^{\text{vel}}\| - 0.1}{0.02}\right),$$

with $\|c_t^{\text{vel}}\| = \sqrt{v_x^2 + v_y^2 + \omega_z^2}$. Below the 0.1 m/s threshold the body slice is supervised almost entirely by the WBC teacher; above it the student inherits from the locomotion teacher; the 0.02 width gives a sharp but differentiable transition.

Discrete-context AMP KL. The AMP KL is masked to recovery-active samples,

$$\mathcal{L}_{\text{KL}}^{\text{AMP}} = \frac{\sum_t \mathbf{1}[\text{recover}_t] D_{\text{KL}}(\pi_\theta \| \pi_{\text{amp}})_t}{\sum_t \mathbf{1}[\text{recover}_t] d},$$

where d is the number of action dimensions. The recovery flag is set on a fixed fraction of envs (20%) at reset to put the agent in fallen poses with delay-masked rewards.

Per-dimension KL. On diagonal Gaussians, every KL above is computed per-dimension and reduced by an active-subset mean,

$$D_{\text{KL}}^{(i)} = \log \frac{\sigma_t^{(i)}}{\sigma_s^{(i)}} + \frac{(\sigma_s^{(i)})^2 + (\mu_s^{(i)} - \mu_t^{(i)})^2}{2(\sigma_t^{(i)})^2} - \frac{1}{2}.$$

For the body and arm slices, dimensions corresponding to other slices and envs flagged “in recovery” are excluded from the mean so that the documented coefficients describe the per-non-recovery-sample weight.

Coefficient schedule. The KL coefficients are cosine-annealed over the first 60,000 update steps with the values listed in Appendix C.4 under *Student DAgger losses*.

C.3 Mixture-of-Experts Architecture

The MoE student uses three experts that share a 64-D latent produced by the proprio-history temporal-conv encoder. The gating network is a small MLP with hidden sizes (128, 64) that maps the latent to a 3-way softmax. Each expert is a (256, 128) MLP producing the action mean; the log-std is shared across experts. All three experts are evaluated at every step and their action-mean outputs are blended by the gate; soft routing keeps the policy fully differentiable and avoids the bimodal artifacts that hard top- k routing introduces when the gate is uncertain.

Two routing-shaping terms are added to the PPO objective (coefficients in Appendix B). A subset-aware load-balancing loss penalizes deviation of the renormalized non-recovery gate weights from a uniform target over the WBC and locomotion experts; concretely, on non-recovery samples the gate is restricted to the WBC and locomotion experts and renormalized to sum to one, and the standard load-balancing penalty is applied to that two-way distribution. The recovery expert is excluded from this term and supervised separately by a recovery-routing loss that pushes gate mass onto it on recovery-active samples and leaves it alone otherwise. The remaining two experts thus divide locomotion and manipulation between themselves without an explicit regime label.

C.4 Training Setup

PPO + DAgger. All policies are trained with the Rsl-rl framework inside mjlabs using an asymmetric actor-critic, obs normalization on both heads, and a scalar (shared) Gaussian action distribution initialized at $\sigma = 1.0$. Each iteration collects 24 steps from 4,096 parallel envs on a single GPU ($\approx 98\text{K}$ transitions/iter), then runs 5 epochs of Adam SGD with 4 mini-batches at learning rate 1×10^{-3} under an adaptive KL schedule (target $D_{\text{KL}} = 0.01$). PPO clip 0.2 with clipped value loss, value-loss coefficient 1.0, entropy coefficient 0.005, $\gamma = 0.99$, GAE $\lambda = 0.95$, and global gradient-norm clip 1.0. The WBC teacher and the student each train for up to 30,000 updates; the locomotion teacher for 20,000; the fall-recovery teacher for 100,000.

Network sizes. The WBC, locomotion, and fall-recovery teacher actors and critics are (512, 256, 128) MLPs with ELU activation. The student backbone is a wider (512, 512, 256, 128)

MLP with Swish activation and LayerNorm; the 11-frame proprio history is encoded by a 1-D temporal-conv into a 64-D latent, and the planner-interface and reference-motion blocks enter through a 128-D motion latent. The MoE expert and gate sizes are listed in Appendix C.3.

Student Dagger losses. The student extends the same PPO core with the context-conditioned KL terms of Section 3.4 plus two MoE routing-shaping terms. In the canonical 3-teacher recovery run, the cosine-annealed KL coefficients are

$$\lambda_B : 0.4 \rightarrow 0.2, \quad \lambda_A : 0.1 \rightarrow 0.05, \quad \lambda_{AMP} : 0.4 \rightarrow 0.2,$$

over the first 60,000 update steps; the asymmetric arm coefficient ($\lambda_A < \lambda_B$) deliberately loosens the WBC pull on the arms so the body teachers stay in charge of locomotion stability. The MoE routing-shaping terms (load-balance, 0.01; recovery-routing, 0.5) are defined in Appendix C.3.

Fall-recovery teacher AMP losses. The discriminator MLP is (1024, 512, 256) trained with discriminator-loss coefficient 1.0, WGAN-style gradient-penalty coefficient $\lambda_{gp} = 10$, and weight decay 1×10^{-3} on the trunk and 1×10^{-2} on the head; expert transitions are buffered in a 100,000-entry on-policy AMP replay. The PPO reward delivered to the teacher is $r = (1 - \ell) r_{disc} + \ell r_{task}$ with $\ell = 0.75$ and a global 0.1 scale on r_{disc} .

Domain randomization. Randomization spans base mass, foot friction, PD gains, motor strength scales, encoder bias, applied hand forces, and per-env actuation delay. Randomized values are exposed to the critic but not to the actor.

Checkpoint selection. Every reported run is exported from the final checkpoint.

D Extended Experimental Results

Figures 6–8 report the per-axis breakdowns underlying the aggregate $|\Delta v|$ and workspace convex hulls; the evaluation protocol is described in Section 4.

Velocity-tracking sweep curves. The ablation panel (Fig. 6) traces how each ingredient — adding the locomotion teacher, randomized commands, asymmetric split-KL with MoE, and the AMP recovery teacher — tightens the realized-versus-commanded curve toward the unit-slope ideal on all three axes. The SOTA panel (Fig. 7) overlays our final controller against AMO, OpenHomie, FALCON, and SONIC on the same $[-1, 1]$ sweep; our curves sit within the SOTA cluster on every axis, with the largest gap to perfect tracking occurring near the saturation regions where the legged platform’s velocity budget is binding.

Bimanual wrist-workspace details. The workspace panel (Fig. 8) shows the bimanual wrist-reach hulls in the pelvis frame from three orthogonal views (XY top-down, YZ front, XZ side), restricted to the forward half-space $x \geq 0$ that drives the *Robust WS* column. Our controller, with and without the fall-recovery teacher, covers the largest forward-half workspace at feasibility comparable to SONIC, while FALCON’s hull stops well short of the side and top reaches. The reported feasibility and robust-workspace numbers are aggregated over 2000 discovery targets and 400 accuracy targets per controller, all sampled with a fixed seed (42) so every method sees the same target set; a trial counts as feasible when both wrists stay within 15 cm of the target throughout the measurement window, the policy does not fall, and pelvis horizontal drift stays under 25 cm.

Differential-IK baseline adapter. For the comparison in Section 4.2, every baseline that does not natively expose pelvis-frame wrist targets is driven through a shared differential-IK head built on `mink` [51]. The two wrist targets enter as `RelativeFrameTasks` rooted at the pelvis body, so the Cartesian targets are interpreted directly in the pelvis frame (the same frame the controllers consume); only position is tracked (orientation cost 0). All non-arm DoFs are hard-frozen via an equality constraint (`DofFreezingTask`) rather than a soft penalty, so the IK only moves the arms

and the policy retains full authority over the lower body. The QP uses a wrist position cost of 10, a posture regularizer of 10^{-4} , Levenberg–Marquardt damping 10^{-3} , and solver damping 10^{-1} , run for up to 30 iterations per call at the 50 Hz control rate (daqp solver, with a quadprog fallback and a hold-last-target fallback on the rare infeasible QP). The solver warm-starts from a fixed bent-elbow ready pose to avoid the straight-down elbow singularity and persists the arm configuration across calls; on FALCON the resulting reference is additionally low-pass filtered with a ~ 200 ms time constant to match its teleop pipeline and suppress reference jitter. Joint limits are enforced by the underlying MJCF, and all methods receive the identical wrist-target distribution used for our own controller.

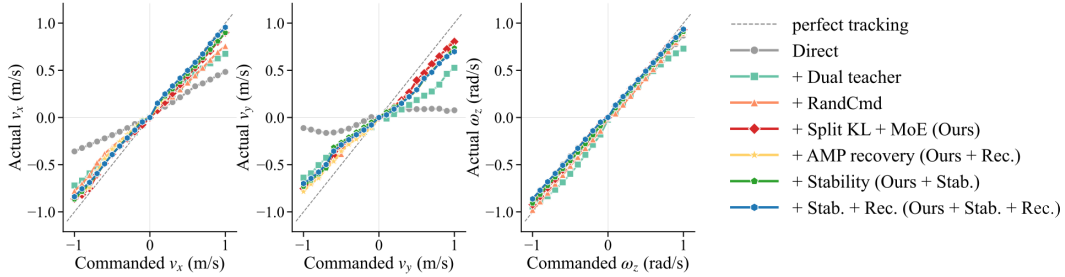


Figure 6: Ablation progression: per-axis realized-versus-commanded velocity sweep across $[-1, 1]$.

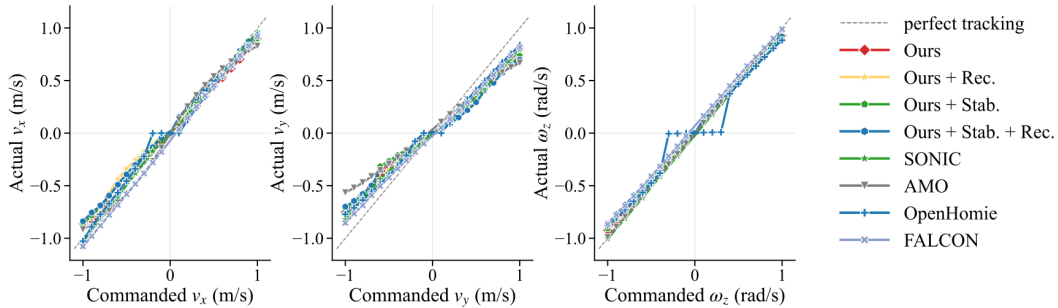


Figure 7: SOTA comparison: per-axis realized-versus-commanded velocity sweep across $[-1, 1]$.

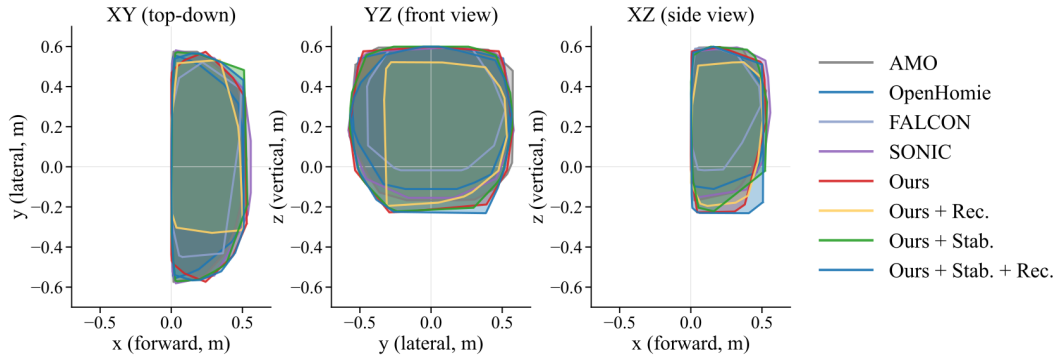
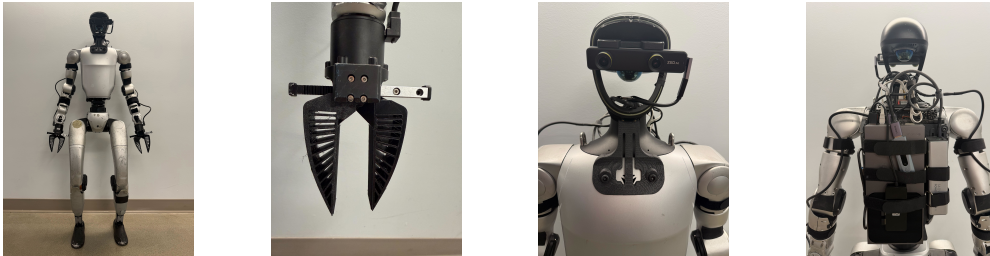


Figure 8: Bimanual wrist-workspace hulls in three orthogonal pelvis-frame views, forward half ($x \geq 0$).

E Hardware Setup

All hardware experiments run on a Unitree G1 humanoid (29 DoF) with stock 3-finger hands replaced by bilateral Dex1-1 anthropomorphic grippers and a head-mounted ZED-M stereo RGB-D camera that supplies frames to the VLM and the waypoint-projection step. A back-mounted Nvidia Jetson Thor runs the full onboard stack end to end — the 50 Hz RL controller, the agentic planner of Section 3.5, and local VLM inference (with an optional ChatGPT-API fallback over Wi-Fi) — powered together with the Dex1-1 grippers by a single 140 W USB-PD powerbank for a fully untethered envelope. The Jetson + powerbank + gripper payload is modeled as rigid masses on the simulator’s G1 and included in the domain randomization of Appendix C.4, keeping the deployed policy within its training mass distribution. The assembled platform, gripper, camera, and compute stack are shown in Fig. 9.



(a) Assembled G1 platform. (b) Dex1-1 end-effector. (c) ZED-M stereo RGB-D camera. (d) Back-mounted compute stack.

Figure 9: **Real-robot deployment platform.** (a) Unitree G1 with bilateral Dex1-1 grippers and a head-mounted ZED-M stereo RGB-D camera. (b) Close-up of one Dex1-1 gripper, replacing the stock 3-finger hand. (c) Head-mounted ZED-M stereo RGB-D camera providing the RGB and depth frames consumed by the VLM. (d) Back-mounted Nvidia Jetson Thor and 140 W USB-PD powerbank that together drive the onboard RL controller, the agentic planner, and local VLM inference (if needed).