© 2026 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Learning a Whole-Body Pivot-Walking Gait for Heavy-Box Transport with a Humanoid

Rikuto Hidaka[1], Tsung-Yuan Chen[2] and Kento Kawaharazuka[3]

***Abstract*—Through pivot walking, humans can transport objects that are too heavy to lift. Previous humanoid studies have relied on planned sequences in which object manipulation and stepping are executed separately and the object is brought to rest between successive rotations. However, few studies have developed a single learned controller for humanoid pivot walking. In this paper, we propose a single reinforcement-learning policy that jointly coordinates stepping, balance control, and object manipulation. The policy is trained using rewards that promote alternating pivot motions and a phase reference that progresses according to the displacement of the object. In simulation, we confirmed that the humanoid can pivot a box. For deployment, we fine-tuned the policy into an object-blind version that follows a periodic clock instead of the progress reference. On the real robot, the deployed policy transported the box forward by more than $1\,\mathrm{m}$ in all five trials; each trial ended at a base-tilt safety stop. These results demonstrate that coordinated humanoid pivot walking can be learned with reinforcement learning and transferred to a physical humanoid robot without direct object-state sensing.**

I. Introduction

Humanoid robots are becoming increasingly capable of dynamic locomotion, yet many useful tasks require moving objects that are too large or heavy to lift. Direct lifting may exceed the robot's arm capacity, while pushing or dragging can require large sustained forces and substantial sliding at the floor. Humans can instead exploit ground support through pivot walking: the object is rocked between its bottom corners and rotated forward while most of its weight remains supported by the ground. Pivot walking may therefore provide humanoids with a useful transport mode for heavy objects (Fig. 1).

Pivoting has been studied extensively as a nonprehensile manipulation primitive, including for humanoid transport (Section II). Existing humanoid systems, however, rely on explicit contact models and planned quasi-static sequences. Object rotation and robot stepping are scheduled as separate phases, and the object is nearly stationary between rotations. These methods demonstrate reliable task-level planning, yet they do not realize pivot walking as a continuously executed learned gait in which stepping, balance, and object rocking are generated jointly.



Fig. 1. **Pivot walking experiment.** A Unitree G1 humanoid transports a $14.9\,\mathrm{kg}$ box by tilting it onto a bottom corner and rotating it about alternating corners. Representative frames from one alternating lean cycle in a hardware trial. Background blurred.

In this work, we propose such a whole-body pivot-walking gait with reinforcement learning. In the six separately trained routes we tested, a basic transport objective produces tilted dragging or unstable pushing rather than repeated pivoting. We introduce box-level rewards that encourage alternating rocking and a spatial reference indexed by realized box progress. This reference remains synchronized with the object when motion slows or stalls. We then transfer the learned gait to a proprioceptive policy by removing box-state observations and continuing reinforcement learning with clock conditioning (Section IV-E). Progress pacing is thus used during training, while the deployed policy follows an internal clock.

In simulation, π_{path} transports a $14.9\,\mathrm{kg}$ box $4.12\,\mathrm{m}$ in $20\,\mathrm{s}$ with zero observed terminations across 128 initial

[1]College of Arts and Sciences, The University of Tokyo, Tokyo, Japan. rhdkreal@gmail.com
[2]Next Generation Artificial Intelligence Research Center, Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, Japan. chen.tsungyuan@g.ecc.u-tokyo.ac.jp
[3]Department of Mechano-Informatics, Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, Japan. kawaharazuka@jsk.imi.i.u-tokyo.ac.jp

box orientations (Section V). In real robot experiments, the object-blind policy reproduces repeated lean and yaw reversals and achieves more than $1\,\mathrm{m}$ of forward transport in all five trials. All trials terminate at the robot's base-tilt safety trigger, so the present result demonstrates repeatable short-horizon gait transfer rather than sustained autonomous delivery (Section V-G).

Our contributions are:

- A whole-body humanoid policy that walks, balances, and pivots a free-standing box with both hands in a single gait, learned without an explicit object-contact model or a planned settling phase.
- Box-level pivoting rewards and a progress-paced reference with which the alternating gait is learned; in our tests, a transport reward alone led to dragging or unstable pushing.
- Repeatable short-horizon transfer to the physical G1 without box-state sensing, achieving repeated lean and yaw reversals and transporting the box forward by more than $1\,\mathrm{m}$ in all five trials.

II. Related Work

A. Reinforcement Learning for Legged Locomotion

Reinforcement learning has enabled robust sim-to-real locomotion for legged robots [1], [2], while massively parallel simulation has made the training of such policies practical [3]. Periodic conditioning and reward composition have also been used to learn multiple bipedal gaits [4]. We build on this general training framework and consider a coupled loco-manipulation problem rather than command-tracking locomotion. The policy must coordinate foot placement, balance, and upper-body forces with the rocking dynamics of a heavy external object.

B. Humanoid Loco-Manipulation

Humanoid systems have demonstrated constraint-based loco-manipulation [5], whole-body interaction with bulky objects [6], and transport of irregular loads using locomotion priors [7]. These approaches primarily lift, carry, embrace, place, or continuously push objects. In contrast, pivot walking repeatedly changes the object's supporting ground corner while the robot continues to step. The policy must therefore coordinate locomotion with both contact forces and the object's alternating rocking motion.

C. Pivoting and Nonprehensile Manipulation

Pivoting is a classical nonprehensile manipulation primitive in which an object rotates about a supporting environmental contact rather than being fully supported by the robot [8]. Prior work has studied the mechanics and planning of pushing [9], whole-body humanoid pushing of large objects [10], [11], and pivot-based reorientation [12].

Most closely related, Yoshida et al. [13], [14] demonstrated humanoid transport through planned quasi-static pivot sequences, with object manipulation and robot stepping scheduled as separate phases. Murooka et al. [15] subsequently added online object-state and force estimation to adapt planned humanoid pivoting to motion error and unknown physical parameters. Zhang et al. [16] formulated pivoting-gait modes with graph model predictive control, though on a fixed industrial dual-arm manipulator rather than a walking humanoid. Reinforcement learning has also produced *in-hand* pivoting [17], where a gripper supports the object and the mobile base remains fixed.

In contrast, we propose a single whole-body humanoid policy that jointly generates stepping, balance, bimanual interaction, and alternating object rocking, without an explicit contact-mode schedule or a planned settling phase. The contribution is therefore the realization of pivoting as a continuous learned humanoid gait rather than a new pivoting primitive.

D. Reward Shaping and State-Indexed Motion References

Complex locomotion and manipulation strategies can emerge from sparse task objectives [18], [19]. In our setting, however, the outcome-only training routes tested in Section V-B converge to tilted dragging or unstable pushing rather than alternating pivoting. We therefore use skill-level reward shaping to make the desired object motion reachable during exploration. Unlike potential-based shaping [20], this shaping intentionally changes the task optimum by specifying an alternating box-rocking pattern, while leaving joint trajectories, foot placement, contact timing, and interaction forces to be learned.

Phase-conditioned locomotion policies commonly use elapsed time or a periodic clock [4]. Conceptually related state-based phase variables appear in virtual-constraint control, where motion is indexed by robot configuration rather than time [21]. We instead index the box-yaw reference by the object's realized forward progress. The reference therefore pauses when the box stalls and reverses when it moves backward, coupling the desired rocking phase to task progress rather than elapsed time.

III. Problem Formulation

A. Task Definition and Physical Setup

We consider whole-body transport of a free-standing box by a 29-DoF Unitree G1 EDU humanoid (legs, waist, arms, wrists). The goal is to learn a policy that observes the robot state and, during the main training stage, the box state, and outputs whole-body joint targets that move the box forward. The transport axis is fixed at episode initialization, and the box yaw should stay close to its initial heading. The robot receives no walking velocity command, no footstep sequence, and no contact schedule. The desired strategy keeps both hands on the two side faces of the box and repeatedly rocks

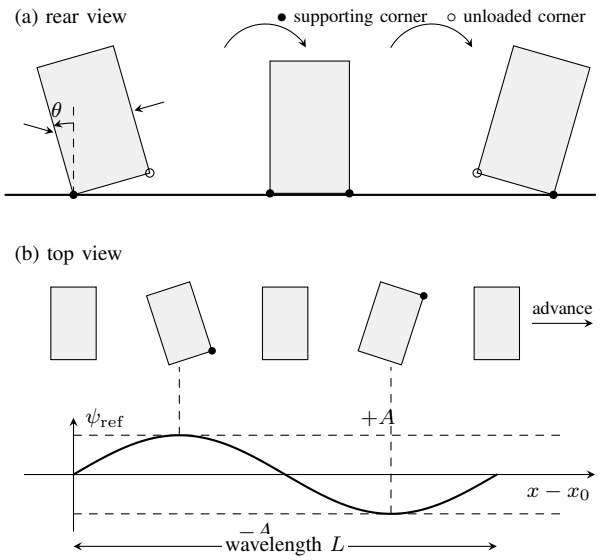


Fig. 2. Pivot-walking schematic (not to scale). (a) Rear view: the box rocks between its two bottom corners, so one corner supports the box while the opposite corner is unloaded, and the hands press on the side faces to steer the rotation. (b) Top view: box yaw alternates with amplitude A as the box advances, transferring support once per half wavelength. The yaw reference ψ_{ref} used for training (Section IV-C) is indexed by box progress x, so it halts whenever the box halts.

it between its two bottom corners while alternating its yaw, which produces forward motion (Fig. 2).

Simulation uses mjlab, built on MuJoCo Warp for GPU-parallel training. The physical object is a modified G1 shipping case. It measures $0.475 \times 0.35 \times 0.82\,\mathrm{m}$ (width $\times$ depth $\times$ height) and weighs $14.9\,\mathrm{kg}$. Simulation models it as a rigid free-standing cuboid (hardware preparation in Section V-G). The robot starts in a crouched carry pose with a nominal base height of $0.765\,\mathrm{m}$, and each palm touches one side face of the box near mid-height. By pressing the two side faces with different forces, the hands can rotate the box in yaw while the ground carries most of its weight. Hand contact is encouraged by the reward and is not enforced as a constraint.

If the box were lifted with the load shared equally, each arm would carry $7.45\,\mathrm{kg}$. This is more than twice the $3\,\mathrm{kg}$ per-arm payload of the G1 EDU [22]. Dragging, in turn, requires greater force as friction increases and can abrade the object and the floor. We therefore study ground-supported pivot transport for this robot and object. Box mass and box–ground friction vary across training and evaluation, and the exact values are given in Sections IV and V.

B. Policy Observations and Actions

At each control step the policy maps an observation to an action, $a_t = \pi_\theta(o_t)$. The action sets joint-position targets for all 29 actuated joints as offsets from a nominal pose. The policy runs at $50\,\mathrm{Hz}$ and a low-level PD controller tracks the targets. The box-observing training actor receives robot proprioception (joint positions and velocities, base angular velocity, and projected gravity), the previous action, and the box state relative to the robot base. The box state contains the relative position, the sine and cosine of the yaw deviation from the initial heading, and gravity projected into the box frame. Skill-reward training adds a three-dimensional reference-phase observation (Section IV-B). No walking velocity command is given. Uniform noise is added to each actor observation term in raw units, for example ± 0.05 on the box terms. The critic receives the noise-free state, including variables that the deployment actor cannot observe. We study two actor settings. The box-observing policy above is used for gait learning. The object-blind deployment actor removes all box observations and instead receives a periodic phase signal and the robot yaw from gait onset (Section IV-E).

C. Episode and Termination

Each simulation episode lasts $20\,\mathrm{s}$ and terminates early if the box tilt exceeds $35°$, the robot falls (base orientation more than $70°$ from upright), or a foot contacts the box. Foot–box contact is treated as failure because the task is defined as bimanual pivot transport rather than leg-assisted pushing. Episode completion measures consistency across initial box orientations ($\pm 5°$, one box). Transport distance and strategy-level metrics are evaluated separately, because forward progress alone cannot distinguish alternating-corner pivoting from tilted dragging. The corresponding gait-validity metrics are defined in Section V-A.

IV. Method

A. Overview

Our goal is a single whole-body policy that advances a ground-supported box through repeated alternating pivot motions. The policy controls all 29 robot joints and must coordinate box rocking, bimanual side contact, balance, and stepping without an explicit contact-mode schedule. Fig. 3 summarizes the training pipeline.

The method has four stages. First, the robot learns to balance while keeping side contact with a tilted, ground-supported box. Second, a temporal yaw reference and box-level skill rewards let it discover an alternating pivot gait at low box mass. Third, the temporal reference is replaced by a spatial reference indexed by realized box progress, and the gait is adapted through a mass curriculum. Finally, box-state observations are removed and the policy is trained further with an internal clock for hardware deployment.

B. Pivot-Gait Reward

The basic transport objective rewards achieved forward displacement while penalizing persistent box-heading deviation:

$$r_{\mathrm{fwd}} = \mathrm{clip}\Big(\tfrac{x^{\mathrm{box}} - x_0}{x^* - x_0}, 0, 1\Big) \cdot \exp\big(-(\psi^{\mathrm{box}}/\sigma_\psi)^2\big), \quad (1)$$

where x^{box} is the box position along the fixed transport axis, x_0 its initial value, x^* the nominal progress target, ψ^{box} the

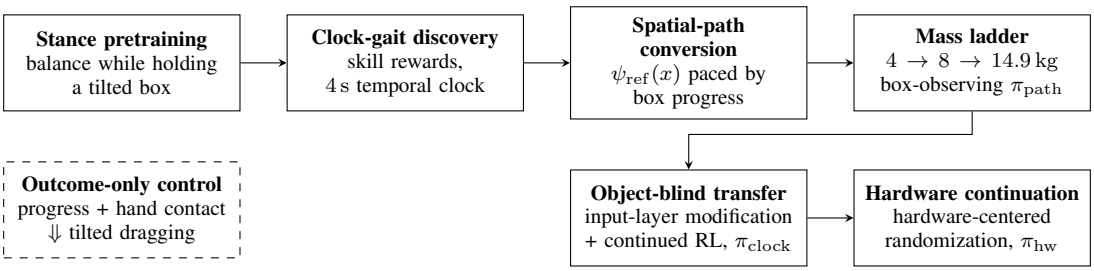


Fig. 3. Training pipeline. The main route (solid) pretrains a stance, discovers the alternating gait under a temporal clock, converts the reference to a progress-paced spatial path, and adapts the gait up the mass ladder. The resulting box-observing policy π_{path} is transferred to an object-blind actor π_{clock}, and a hardware-centered continuation π_{hw} runs in the physical experiments (Section V-G). The outcome-only control (dashed) is trained independently for comparison, not used as an initialization.

box yaw deviation, and $\sigma_\psi = 30^\circ$. This state reward is paid at every step, so earlier progress yields a larger return. The yaw factor discourages persistent heading drift while permitting the alternating yaw excursions needed for pivot walking. A hand–box contact reward and regularization terms for posture, action rate, and upright base orientation complete the basic objective, and early termination incurs a penalty of -200. In the outcome-only routes of Section V-B, this objective converges to tilted dragging or unstable pushing instead of repeated pivoting. We therefore add box-level skill rewards.

For gait discovery, a Gaussian reward tracks a temporal yaw reference, a simple periodic target for exploration:

$$r_{\mathrm{osc}} = \exp\Big(-\big(\tfrac{\psi^{\mathrm{box}}-\psi_{\mathrm{ref}}}{\sigma_{\mathrm{osc}}}\big)^2\Big), \qquad \psi_{\mathrm{ref}}(t) = A\sin\big(\tfrac{2\pi t}{T}\big), \quad (2)$$

with period $T = 4\,\mathrm{s}$ and width $\sigma_{\mathrm{osc}} = \max(0.6A, 6^\circ)$. The amplitude A grows from 5° to 25° over roughly the first 2000 training iterations, and the actor observes the sine and cosine of the reference phase together with the current amplitude factor. A binary tilt reward encourages configurations associated with bottom-corner support:

$$r_{\mathrm{tilt}} = \begin{cases} 1, & 12^\circ < \theta < 30^\circ, \\ 0, & \text{otherwise}, \end{cases} \quad (3)$$

where θ is the angle between the box up-axis and the world vertical. The lower bound discourages flat-ground rotation, and the upper bound leaves margin before the 35° termination threshold. This term does not prescribe which corner is loaded. Alternating support arises from its combination with the yaw reference. An additional reward keeps each hand near its assigned side face, and a small torso–box contact penalty prevents the robot from leaning on the object. All terms are combined additively, because a multiplicative combination gives zero reward in the initial upright state and removes the incentive to start moving. The reference specifies only the desired object motion. Joint trajectories, foot placements, contact timing, and hand forces are all learned.

C. Progress-Paced Spatial Reference

A temporal reference continues to advance even when a heavy box stalls, so the clock demands yaw reversals that the realized box motion has not reached. We therefore index the reference by box progress instead of time:

$$\psi_{\mathrm{ref}}(x) = A\sin\Big(\tfrac{2\pi\,(x-x_0)}{L}\Big), \qquad A = 18^\circ,\ L = 0.20\,\mathrm{m}. \quad (4)$$

The desired yaw changes sign every $L/2 = 0.10\,\mathrm{m}$ of box progress. When the box stops, the reference phase also stops, and backward motion reverses the phase.

A stationary box makes the spatial reference constant, so yaw tracking alone can be maximized without transport. We therefore multiply the tracking term by capped positive box velocity:

$$r_{\mathrm{path}} = \exp\Big(-\big(\tfrac{\psi^{\mathrm{box}}-\psi_{\mathrm{ref}}(x)}{6^\circ}\big)^2\Big)\cdot \mathrm{clip}\left(\dot{x}^{\mathrm{box}}, 0, v_{\max}\right), \quad (5)$$

where $v_{\max} = 0.15\,\mathrm{m/s}$ and $\dot{x}^{\mathrm{box}}$ is the box velocity along the transport axis. The cap saturates the reward at higher speeds and is not a hard speed limit. The velocity factor is left unnormalized, and its scale is absorbed into the reward weight. This term replaces the temporal tracking term, while the other rewards remain active.

For heavy-box training we add an overspeed penalty on a 2 s moving average of box speed and a quadratic penalty on base-height deviation. The average tolerates the short velocity peaks of the rocking gait, and the height term maintains the nominal crouch.

The actor observes the sine and cosine of $2\pi(x - x_0)/L$, the same two-dimensional phase embedding as the clock, so training resumes directly across the switch. We denote the resulting box-observing policy π_{path}.

D. Training Curriculum

Training uses PPO with 4096 parallel environments. The actor and critic are multilayer perceptrons with hidden dimensions $(512, 256, 128)$ and ELU activations. Each iteration collects 24 policy steps per environment and performs five optimization epochs, with $\gamma = 0.99$, GAE $\lambda = 0.95$, and a KL-adaptive learning rate. The policy is trained through four stages, each resuming the previous one.

1) **Stance pretraining.** At $4\,\mathrm{kg}$, the robot learns to stay balanced while keeping side contact with a tilted,

ground-supported box, with progress rewards disabled and box tilt and yaw randomized at reset.

2) **Pivot-gait discovery.** The temporal reference and skill rewards are activated at $4\,\mathrm{kg}$ together with the amplitude curriculum. Alternating yaw swings appear within approximately 2000 iterations.
3) **Spatial-reference conversion.** The temporal phase is replaced by the progress-paced reference of Eq. (4) and training continues.
4) **Mass adaptation.** The gait is adapted from 4 to 8 and then $14.9\,\mathrm{kg}$, with a separate continuation to $24\,\mathrm{kg}$. The rewards and reference stay fixed apart from the heavy-box regularization of Section IV-C.

We use this staged initialization because direct training at the heavier masses did not produce a stable gait in our runs.

E. Object-Blind Deployment Transfer

The policy π_{path} requires the box pose and the box progress that indexes its phase. Neither is available on our hardware, which has no motion capture or box-mounted sensing. We therefore transfer the gait to an actor that uses onboard measurements only, by modifying the actor input layer and continuing RL training.

The box-state columns of the actor's first layer are removed and all remaining weights are kept. The critic retains the box state as privileged information during continued training. We also append a zero-initialized input carrying the sine and cosine of the robot yaw from gait onset, which reduces slow heading drift.

Because box progress is no longer observed, the spatial phase is replaced by a periodic clock, and the spatial-path term is replaced by a capped forward-speed reward. The box velocity in this reward is computed by the simulator, so removing box sensing from the actor does not restrict training. The learned gait has a full yaw cycle of about $1.0\,\mathrm{s}$ at $14.9\,\mathrm{kg}$, and for hardware-oriented training we choose a slower $1.4\,\mathrm{s}$ clock period to temper the motion. The resulting actor π_{clock} observes joint positions and velocities (29 each), base angular velocity and projected gravity (3 each), the previous action (29), the three-dimensional clock signal, and the robot-yaw pair, 98 scalars in total, all available onboard.

Continued training randomizes box mass, box–ground and hand friction, box center of mass, initial box placement, PD gains, and hand-contact geometry. A slow-start factor ramps the rocking amplitude and the speed reward over the first $6\,\mathrm{s}$ of each episode. A final continuation adds symmetry and heading-drift regularization and re-centers the friction randomization on the measured floor. This policy, denoted π_{hw}, is used in the physical experiments (Section V-G).

V. Experiments

A. Evaluation Protocol and Metrics

Unless stated otherwise, each simulation result evaluates the deterministic actions of one trained checkpoint for $20\,\mathrm{s}$ in 256 parallel environments (128 for the $14.9\,\mathrm{kg}$ path policy). The episodes randomize only the initial box orientation, with yaw, roll, and pitch sampled within $\pm 5^{\circ}$, not the training seed or the physical parameters. Reported forward travel is accumulated positive travel and can exceed net displacement, and "terminations" counts robot falls, box falls, and foot–box events. Training rewards progress toward a nominal $2\,\mathrm{m}$ target (Eq. 1). Evaluation instead runs for $20\,\mathrm{s}$ without a stop command, so travel beyond $2\,\mathrm{m}$ shows gait continuation rather than controlled delivery.

Because scalar task metrics cannot separate pivot walking from tilted dragging, we also count amplitude-segmented yaw swings (excursions over 10° between sign changes) and report the lowest-corner motion proxy, the integrated planar speed of the lowest box corner divided by forward travel. The proxy is near zero for rotation about a stationary corner and near one for pure translation; it does not identify the active contact. For selected checkpoints we therefore replay deterministic trajectories in MuJoCo and measure per-corner box–ground contact forces, from which we report left–right support alternations, the fraction of time one corner carries over 60% of the normal force, and the loaded-corner slide per forward travel. Swings, support alternations, and hardware yaw reversals (Section V-G) use different detectors and are not interchangeable counts.

B. Does Outcome-Only Training Discover Pivot Walking?

None of the six outcome-only routes (one at $8\,\mathrm{kg}$, one at $14.9\,\mathrm{kg}$, four at $24\,\mathrm{kg}$, one training seed each) produces repeated alternating yaw swings. Five routes converge to stable constant-tilt dragging, holding a tilt of 25°–31° and advancing only 0.18–$0.33\,\mathrm{m}$ per episode. The sixth is an unstable deep-tilt push that repeatedly topples the box. These routes show that the gait did not emerge under the tested settings, not that it cannot emerge, and they motivate the explicit pivoting objective. To test this within our pipeline, we train two branches from the same $4\,\mathrm{kg}$ stance-pretrained policy for 6000 updates. The only difference is the box-level skill rewards. With them, the robot learns repeated yaw swings (11.4 per episode), although it still often fails at this early stage. Without them, it learns no swings and often ends with a foot touching the box. In this single-seed test, only the branch with skill rewards learned the gait.

C. Does the Learned Policy Actually Pivot the Box?

Main result. Table I evaluates the path-conditioned policy π_{path} across the mass curriculum. At the physical mass of $14.9\,\mathrm{kg}$ it completes all 128 episodes without termination and accumulates $4.12\,\mathrm{m}$ of forward travel in $20\,\mathrm{s}$ ($0.21\,\mathrm{m/s}$), with 41.2 yaw swings per episode (mean amplitude 32.6°). Nearly all of this travel (98.7%) occurs during the swings. The full yaw-cycle period lengthens from $0.68\,\mathrm{s}$ at $4\,\mathrm{kg}$ to $0.97\,\mathrm{s}$ at $14.9\,\mathrm{kg}$, although the reference carries no time schedule. With a fixed spatial wavelength, the slowing follows the

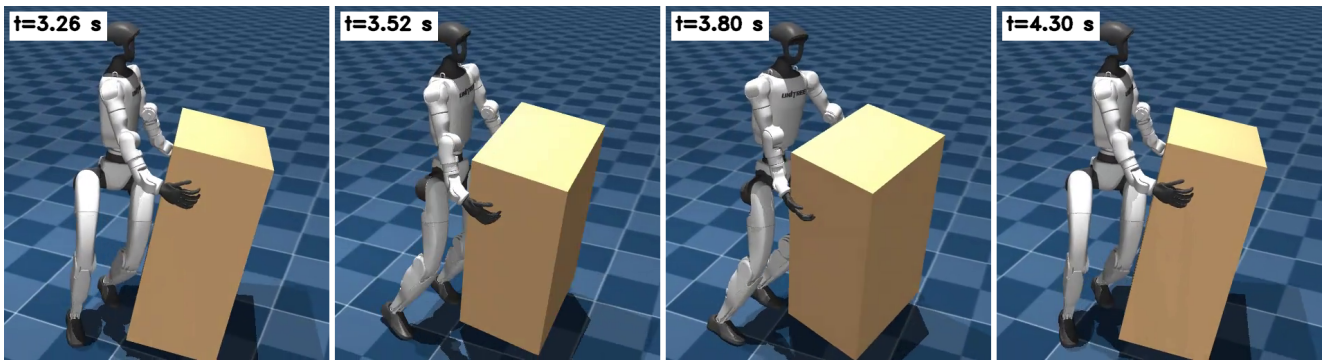


Fig. 4. π_{post}, a posture-refined sibling of π_{path}, at 14.9 kg (a separate checkpoint from Table I). Four frames span one full alternating yaw cycle ($+17°$ peak, transition between lean directions near zero yaw, $-17°$ peak, return) while the box stays leaned onto its near edge; evaluation records 3.68 m/20 s and zero terminations over 256 environments.

TABLE I

PATH-CONDITIONED PIVOT GAIT ACROSS THE MASS CURRICULUM (μ_g=2, 20 s). VALUES ARE MEANS ACROSS 256 EVALUATED INITIAL CONDITIONS (128 AT 14.9 kg); $T_{\text{cycle}} = 2(20\,\text{s})/N_{\text{swing}}$; LC MOTION IS THE NORMALIZED LOWEST-CORNER MOTION PROXY. ALL ROWS COMPLETE THE EPISODE WITH ZERO TERMINATIONS.

Mass [kg]	Travel [m]	Swings [/env]	T_{cycle} [s]	LC motion
4	5.88	58.8	0.68	0.57
8	5.45	54.5	0.73	0.36
14.9	**4.12**	**41.2**	**0.97**	**0.059**

lower transport speed. The lowest-corner proxy falls to 0.059, consistent with a nearly stationary supporting corner.

In simulation, the feet stay low: the soles are within 15 mm of the floor in 99.0%/98.4% of left/right samples and rise above this height only 1.4/2.9 times per episode, indicating shuffling rather than distinct steps.

Contact-force validation. Replaying π_{post}, a posture-refined sibling of π_{path} (Fig. 4), in deterministic MuJoCo, we measure box–ground contact forces directly. Over 20 s the dominant support alternates 34 times between the left and right bottom corners, a single corner is dominant during 99.5% of the rollout, and the loaded-corner slide is 0.09 of forward travel (Fig. 5). The learned motion is therefore alternating-corner pivoting rather than tilted translation, for the replayed checkpoint and its contact model. Replaying π_{path} itself gives 32 alternations, 95.7% single-corner dominance, and slide 0.123, although it travels less in this replay (2.32 m) than in training simulation.

Seed replication. Re-running the pipeline with two additional training seeds also reproduces the gait. All three seeds discover the gait at 4 kg, complete the curriculum with zero terminations, transport 4.12–4.32 m with 41–43 swings, and slow with mass in the same way. Only the proxy is seed-dependent (0.059, 0.29, 0.054); the reproducible outcome across these seeds is the alternating gait, not consistently low corner motion.

TABLE II

COMPLETE REFERENCE-CONDITIONING ROUTES AT 14.9 kg (20 s). (*COMPUTED OVER ROLLOUTS WITH REPEATED BOX-FALL RESETS).

Reference	Init	Travel [m]	Swings [/env]	Term.	LC
clock 4 s (hand-chosen)	mass ladder	0.84	10.4	0	0.23
clock 1.1 s	4 s-gait policy	1.51	36.0	0	0.18
clock 1.1 s	scratch + ladder	4.83	35.9	963	0.08*
path (progress-paced)	mass ladder	**4.12**	**41.2**	**0**	**0.059**

D. How Do Temporal and Progress-Paced Routes Compare?

Table II compares four complete training routes at 14.9 kg. The hand-chosen 4 s clock is stable yet slow (0.84 m). Retuned to 1.1 s, the clock route reaches 1.51 m, and a from-scratch 1.1 s clock reaches 4.83 m while toppling the box 963 times across 256 environments. The progress-paced route transports 4.12 m with zero terminations. The results are consistent with progress coupling aiding synchronization for the heavier box. To control for initialization, we continued π_{path} for 1200 updates under its path phase or a 0.97 s clock; both branches lost the gait, with no yaw swings, so this comparison does not separate the two references.

E. How Does Performance Change with Box Mass?

At 24 kg, under the same side-contact configuration and μ_g=2, the four outcome-only routes of Section V-B either drag (0.18–0.33 m) or topple the box (0.42 m with 1003 box falls). Adapting the pivot policy with the skill objective instead transports 0.76–0.81 m with 10–13 swings and zero terminations across two seeds, versus at most 0.33 m for the failure-free draggers. Reducing the skill-reward weights below half lets the gait decay toward dragging. A probe task that rewards only rotating the box to a 30° yaw succeeds at 24 kg (0.99 tracking) and fails at 28 kg, bracketing an empirical boundary of the tested training procedure and contact configuration, not a physical limit.

F. Can the Gait Be Transferred without Box-State Sensing?

In the training simulator, the clock-conditioned actor π_{clock} (Section IV-E) travels 4.67 m in the 20 s evaluation with zero

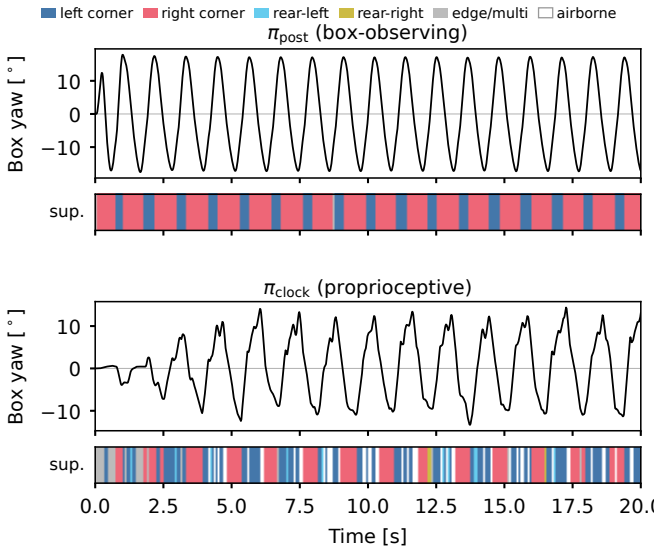


Fig. 5. Box yaw and dominant support corner from measured box–ground contact forces in deterministic replay

terminations. In the same simulator, the deployed policy π_{hw} travels $4.44\,\text{m}$ with about 26 swings, and the box falls in 4 of 512 episodes. Its corner proxy is high (0.66), so we cannot confirm alternating support for this policy. Its feet stay within $15\,\text{mm}$ of the floor in $86\%/91\%$ of left/right samples. Foot motion was not measured on hardware. The exported π_{path} and π_{clock} policies also run unchanged in an independent MuJoCo deployment stack with the C++ controller used on the robot. With simulated box pose, π_{path} travels $10.2\,\text{m}$ in $55\,\text{s}$ without falls, confirming backend transfer. π_{clock}, which receives no box pose or progress signal, travels $9.9\,\text{m}$ in its $73.8\,\text{s}$ active segment with zero falls, leaving the box upright at the commanded stop. It drifts $1.70\,\text{m}$ laterally, consistent with uncorrected phase or heading error.

Contact replay shows that the alternating support survives the loss of sensing. π_{clock} produces 30 left–right support alternations in $20\,\text{s}$, close to the $\sim$29 its $1.4\,\text{s}$ clock predicts, with one corner dominant 78% of the time and a loaded-corner slide of 0.24 of forward travel, against 0.09 for the box-observing π_{post} (per-policy values, not a head-to-head comparison; Fig. 5). For π_{clock}, the gait survives the loss of sensing, with more sliding and drift.

G. Does the Gait Transfer to Hardware?

The hardware experiment evaluates transfer of the alternating gait, not delivery or set-down. We ran five filmed trials on the physical G1 with the $14.9\,\text{kg}$ case of Section III, prepared with casters removed, handle covers flattened, and rubber sheets at the hand-contact regions, and report all five. The deployed actor π_{hw} is the hardware continuation of Section IV-E and runs on the onboard inputs only, with the same observations and $1.4\,\text{s}$ clock as π_{clock}. Section V-F reports π_{hw} in simulation; this campaign is its hardware

TABLE III
REPEATED HARDWARE TRIALS (n=5).

Alternating gait reproduced	5/5
Observed gait duration [s]	16.2 ± 3.0
Active-gait marker speed [m/s]	0.11 ± 0.01
Detected yaw reversals [/trial]	15.6 ± 3.0
Start–final-rest advance [m]	1.93 ± 0.41
Base-tilt safety trigger ended trial	5/5
Robot fell after Passive transition	5/5
Box toppled during triggered ending	2/5

evaluation. Box motion is estimated offline from two front-face markers and is used only for gait timing and relative progress. The one physically measured distance is the tape-measured advance in Table III.

The alternating gait is reproduced in all five trials (Table III, Figs. 1, 6, and 7). The box advances through repeated lean and yaw reversals, with tilt swings averaging $16°$ (peaks 22–$34°$) at a cadence commensurate with the $1.4\,\text{s}$ clock; the motion is qualitatively consistent with the alternating-corner support measured in simulation, though the contact corner is not instrumented. All five trials end 12–$22\,\text{s}$ after gait onset, when base tilt exceeds $1.0\,\text{rad}$ and the controller drops to Passive, after which the robot collapses; no box fall occurred during the preceding gait, and the box toppled in two trials during this triggered ending. The endings are systematic (a same-signed final yaw excursion in all five tracks and the same backward crouching collapse in the side views). The yaw drift is consistent with accumulated phase or heading error in the object-blind controller. However, we did not log joint and IMU data, so we could not identify what makes the base tilt grow. The campaign demonstrates repeatable short-horizon transfer of the gait; sustained delivery, direct measurement of support-corner alternation, and a controlled stop remain undemonstrated.

VI. CONCLUSION

In summary, we presented a learned whole-body pivot-walking gait that jointly coordinates humanoid stepping, balance, and bimanual manipulation of a ground-supported box. A spatial reference paced by the realized box progress couples the alternating box motion to actual transport during training. The deployed policy instead follows a periodic clock.

In the six outcome-only routes we tested, a transport reward alone does not produce the gait. With box-level pivoting rewards, the policy transports a $14.9\,\text{kg}$ box by an average of $4.12\,\text{m}$ in $20\,\text{s}$ over 128 initial box orientations ($\pm 5°$, one box) with no observed terminations, and contact-force replay of a posture-refined continuation verifies repeated alternation of the dominant support corner. On hardware, an object-blind continuation using only onboard observations reproduces the lean and yaw reversals and advances the box by more than $1\,\text{m}$ in all five trials, although every trial ends at the robot's base-tilt safety trigger.

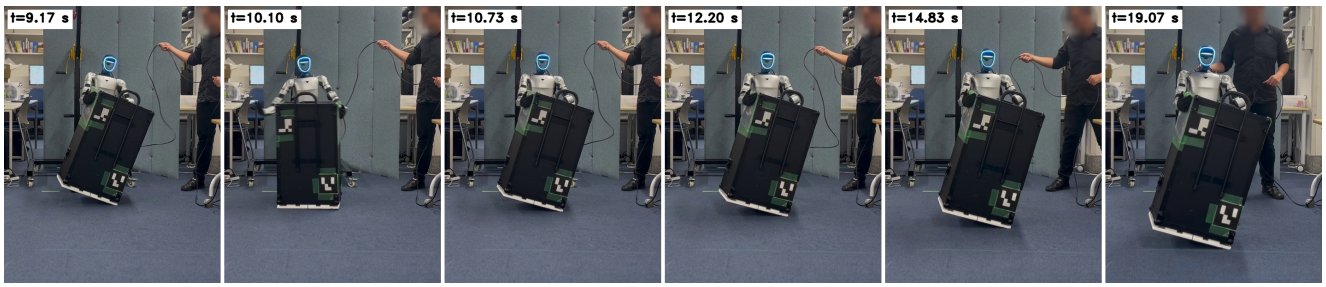


Fig. 6. Transfer of real robot experiment of the alternating gait by π_{hw} (trial 3, fixed frontal camera; timestamps from the marker track): a lean peak (t=9.17 s), the return through near-upright at corner transfer (10.10 s), the opposite lean (10.73 s), and continued alternation as the box advances toward the camera; the active contact corner is not directly measured, and the uncut recordings include the safety-triggered trial ending.

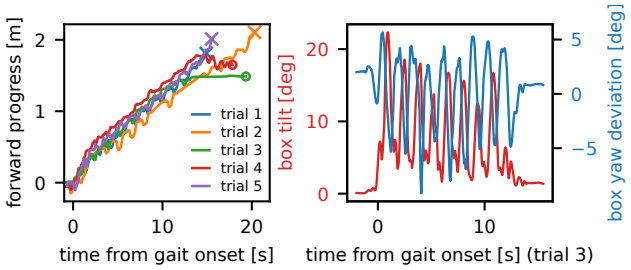


Fig. 7. Marker-tracked real robot trials. Left: estimated forward progress from gait onset for all five trials (crosses mark trials ending with box tilt above 30°, circles mark lower final tilt). Right: estimated box tilt and yaw for trial 3, showing repeated lean and yaw reversals during forward motion. Marker estimates are approximate and do not identify the active ground-contact corner; the relative progress shown is not the tape-measured advance of Table III.

The results demonstrate repeatable short-horizon transfer of a learned humanoid pivot gait without direct box-state sensing. The deployed policy cannot sense or correct its drift relative to the box. Sustained transport will therefore require estimating phase and heading drift from onboard sensing and adding controlled stopping.

Acknowledgment

This work was partially supported by JST Moonshot R&D Program Japan Grant Number JPMJMS263D.

References

[1] J. Hwangbo *et al.*, "Learning agile and dynamic motor skills for legged robots," *Science Robotics*, vol. 4, no. 26, 2019, art. no. eaau5872.

[2] J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter, "Learning quadrupedal locomotion over challenging terrain," *Science Robotics*, vol. 5, no. 47, 2020, art. no. eabc5986.

[3] N. Rudin, D. Hoeller, P. Reist, and M. Hutter, "Learning to walk in minutes using massively parallel deep reinforcement learning," in *Proc. 5th Conf. Robot Learning (CoRL)*, 2022, pp. 91–100.

[4] J. Siekmann, Y. Godse, A. Fern, and J. Hurst, "Sim-to-real learning of all common bipedal gaits via periodic reward composition," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2021, pp. 7309–7315.

[5] P.-A. Léziart, M. Morisawa, and F. Kanehiro, "Learning humanoid loco-manipulation with constraints as terminations," in *Proc. IEEE/SICE Int. Symp. Syst. Integr. (SII)*, 2026, pp. 325–332.

[6] C. Zheng *et al.*, "Embracing bulky objects with humanoid robots: Whole-body manipulation with reinforcement learning," *arXiv preprint arXiv:2509.13534*, 2025.

[7] Z. Pan, Y. Yu, and Z. Zou, "Learning humanoid loco-manipulation for transporting unwieldy construction objects," in *Proc. 43rd Int. Symp. Autom. Robot. Construction (ISARC)*, 2026, pp. 9–16.

[8] Y. Aiyama, M. Inaba, and H. Inoue, "Pivoting: A new method of graspless manipulation of object by robot fingers," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 1993, pp. 136–143.

[9] M. T. Mason, "Mechanics and planning of manipulator pushing operations," *Int. J. Robot. Res.*, vol. 5, no. 3, pp. 53–71, 1986.

[10] K. Harada, S. Kajita, K. Kaneko, and H. Hirukawa, "Pushing manipulation by humanoid considering two-kinds of ZMPs," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2003, pp. 1627–1632.

[11] M. Murooka, S. Nozawa, Y. Kakiuchi, K. Okada, and M. Inaba, "Whole-body pushing manipulation with contact posture planning of large and heavy object for humanoid robot," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2015, pp. 5682–5689.

[12] Y. Hou, Z. Jia, and M. T. Mason, "Fast planning for 3d any-pose-reorienting using pivoting," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2018, pp. 1631–1638.

[13] E. Yoshida *et al.*, "Pivoting based manipulation by a humanoid robot," *Autonomous Robots*, vol. 28, no. 1, pp. 77–88, 2010.

[14] ——, "Planning whole-body humanoid locomotion, reaching, and manipulation," in *Motion Planning for Humanoid Robots*. Springer, 2010, pp. 99–128.

[15] M. Murooka, S. Noda, S. Nozawa, Y. Kakiuchi, K. Okada, and M. Inaba, "Achievement of pivoting large and heavy objects by life-sized humanoid robot based on online estimation control method of object state and manipulation force," *J. Robot. Soc. Japan*, vol. 32, no. 7, pp. 595–602, 2014.

[16] A. Zhang, K. Koyama, W. Wan, and K. Harada, "Controlling pivoting gait using graph model predictive control," *IEEE Access*, vol. 9, pp. 73 757–73 770, 2021.

[17] R. Antonova, S. Cruciani, C. Smith, and D. Kragic, "Reinforcement learning for pivoting task," *arXiv preprint arXiv:1703.00472*, 2017.

[18] N. Heess *et al.*, "Emergence of locomotion behaviours in rich environments," *arXiv preprint arXiv:1707.02286*, 2017.

[19] M. Andrychowicz *et al.*, "Learning dexterous in-hand manipulation," *Int. J. Robot. Res.*, vol. 39, no. 1, pp. 3–20, 2020.

[20] A. Y. Ng, D. Harada, and S. Russell, "Policy invariance under reward transformations: Theory and application to reward shaping," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 1999, pp. 278–287.

[21] E. R. Westervelt, G. Buche, and J. W. Grizzle, "Experimental validation of a framework for the design of controllers that induce stable walking in planar bipeds," *Int. J. Robot. Res.*, vol. 23, no. 6, pp. 559–582, 2004.

[22] Unitree Robotics, "G1 humanoid robot – product specifications," https://www.unitree.com/g1/, 2024, accessed 2026-07-20.