

Humanoid Balance and Step Recovery in Public Transport Scenarios using Reinforcement Learning

Master's thesis

Ruochen Shi
Wei Peng

DEPARTMENT OF MECHANICAL ENGINEERING
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2026
www.chalmers.se

Master's Thesis

Humanoid Balance and Step Recovery in Public Transport Scenarios using Reinforcement Learning

Ruochen Shi
Wei Peng



CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Mechanical Engineering
Division of Dynamics
Chalmers University of Technology
Gothenburg, Sweden 2026

Humanoid Balance and Step Recovery
in Public Transport Scenarios
using Reinforcement Learning
Ruochen Shi
Wei Peng

© Ruochen Shi, Wei Peng, 2026.

Supervisor: *Shivesh Kumar, Department of Mechanical Engineering*
Examiner: *Shivesh Kumar, Department of Mechanical Engineering*

Master's Thesis 2026
Department of Mechanical Engineering
Chalmers University of Technology
SE-412 96 Gothenburg
Sweden
Telephone +46 31 772 1000

Typeset in LaTeX
Gothenburg, Sweden 2026

Humanoid Balance and Step Recovery in Public Transport Scenarios using Reinforcement Learning

Ruochen Shi
Wei Peng
Department of Mechanical Engineering
Division of Dynamics
Chalmers University of Technology

Abstract

Humanoid robots that travel by public transport must remain standing while the vehicle floor accelerates, brakes, and turns beneath them. This thesis develops reinforcement-learning policies that keep a 29-degree-of-freedom Unitree G1 balanced on a finite moving floor, using only the robot's own inertial and joint measurements at deployment. The policies act directly on joint-position targets and are trained with proximal policy optimization and an asymmetric critic in Isaac Lab. Two support conditions form the main study. Without a handle, event-gated rewards allow the robot to step when its current stance can no longer arrest the motion. With a rigid handle, the reward keeps the feet as the primary support, limits the load on the hand, and returns the robot to a support-ready stance after a disturbance. A ceiling-pivoted suspended handle is studied in simulation as an extension.

In simulation, the no-handle policy completes almost all straight-line but only about half of the turning episodes of its training distribution, and it recovers by stepping with the disturbance. In a straight-line severity sweep, it leaves the usable floor area from a peak acceleration of about 2 m/s^2, whereas the rigid-handle policy does not step and exceeds the 10 N m hand-moment limit of its training from about 4 m/s^2. A sweep of the exported policies in MuJoCo shows the same limits. In a moving research van on a closed proving ground, both policies were compared with the manufacturer's standing controller. Across five paired straight-road maneuvers, the no-handle policy reduced the median torso-attitude variation from 6.6° to 1.9° and the median whole-body state deviation from 4.7 to 1.6; in turning, it reduced the attitude variation from 3.1° to 1.4° at the cost of more whole-body motion. The default controller lost balance in S4 and S5 (target speeds of 36 and 45 km/h), leaving 3/5 trials (60%) without tether intervention, compared with 5/5 (100%) for each learned configuration. The rigid-handle configuration produced the smallest median whole-body state deviation in both straight driving (0.039 versus 4.7 for the default condition) and turning (0.131 versus 0.95), with median torso-attitude variations of 1.3° and 1.7°, respectively. The results show that joint-level policies trained on a moving platform transfer to a real vehicle, and that the available support area or the admissible hand load, rather than upright balance itself, sets the limit of the tolerable disturbance.

Keywords: humanoid robot, balance control, reinforcement learning, moving platform, public transport, disturbance recovery, hand support, sim-to-real.

Acknowledgements

We would like to thank Shivesh Kumar, our supervisor and examiner, for giving us the opportunity to work on this topic and for his guidance and feedback throughout the project.
We thank the members of the RAIL lab for their support. In particular, we thank our friends Xiaoyang Chen and Zifeng Liao, who helped us with many practical problems along the way, from running quick tests and packing the robot for transport to fixing problems with the computer, and who also helped us a great deal in everyday life.
We are grateful to Daniel Alejandro Poveda Pi at the REVERE laboratory at Chalmers University of Technology for his help with setting up and carrying out the vehicle experiments, and to the REVERE laboratory for access to the vehicle and the test facilities.
Finally, we thank our families and friends for their support and encouragement throughout our studies.

Ruochen Shi and Wei Peng, Gothenburg, 2026

List of Acronyms

Below is the list of acronyms that have been used throughout this thesis listed in alphabetical order:

CAN	Controller Area Network
CoM	Center of Mass
CoP	Center of Pressure
DCM	Divergent Component of Motion
DoF	Degree of Freedom
GNSS	Global Navigation Satellite System
IMU	Inertial Measurement Unit
KL	Kullback–Leibler
MPC	Model Predictive Control
NH	No Handle
ONNX	Open Neural Network Exchange
PPO	Proximal Policy Optimization
QDD	Quasi-Direct Drive
QP	Quadratic Program
RH	Rigid Handle
RL	Reinforcement Learning
RMS	Root Mean Square
SDK	Software Development Kit
SH	Suspended Handle
URDF	Unified Robot Description Format
ZMP	Zero-Moment Point

Nomenclature

Below is the nomenclature of indices, sets, parameters, and variables that are used in more than one place in this thesis. Thresholds and weights that appear only in one table are defined there.

Indices, Sets, and Frames

t	Discrete control-time index
j	Actuated-joint index
i	Reward-term index; foot index $i \in \{l, r\}$
c	Support condition, $c \in \{\mathrm{NH}, \mathrm{RH}, \mathrm{SH}\}$
$\mathcal{C}_c$	Set of intended contacts of condition c
$\mathcal{R}^{(c)}$	Set of reward terms of condition c
$\mathcal{J}_{\mathrm{act}}$, $\mathcal{J}_{\mathrm{arm}}, \ldots$	Sets of actuated joints and of joint groups
$\mathcal{K}_{\mathrm{ankle}}$	Monitored ankle links
$\mathcal{F}_W,\ \mathcal{F}_P,\ \mathcal{F}_B$	World, platform, and robot-base frames
$\mathcal{F}_{F_l},\ \mathcal{F}_{F_r},\ \mathcal{F}_{H_r}$	Left-foot, right-foot, and right-hand frames

Robot, Platform, and Support

$m,\ B$	Robot mass and weight, $B = mg$
$g,\ \boldsymbol{g}$	Gravitational acceleration and gravity vector
$\boldsymbol{q}_j,\ \dot{\boldsymbol{q}}_j$	Actuated-joint positions and velocities
$\boldsymbol{q}_j^0$	Nominal joint positions
$\boldsymbol{q}_{j,t}^{\mathrm{des}}$	Desired joint positions
$\boldsymbol{S}_a^{(c)}$	Diagonal matrix of joint-wise action scales
$K_p,\ K_d$	Joint stiffness and damping gains
τ_j	Applied joint torque
$\boldsymbol{p}_X^W,\ \boldsymbol{p}_X^P$	Position of point X in world and platform coordinates

$\boldsymbol{R}_{WP},\ \boldsymbol{R}_{WB}$	Orientation of the platform and of the base in the world frame
$x_B^P,\ y_B^P$	Platform-frame base position
$\bar{x},\ \bar{y}$	Limits of the usable floor area
$\kappa,\ \dot{\psi}_P,\ a_{\mathrm{n}}$	Path curvature, platform yaw rate, and normal acceleration
a_P	Horizontal platform acceleration
h_B	Base height above the platform origin
$\theta_B,\ \theta_T$	Base and torso tilt from the vertical
$\boldsymbol{f}_H,\ \boldsymbol{\tau}_H$	Force and moment of the rigid grasp, in $\mathcal{F}_P$
$\widehat{F}_h$	Estimated hand-force magnitude of the suspended handle
f_h	Strap tension
e_q	Root-mean-square joint deviation from $\boldsymbol{q}_j^0$
$s^{\mathrm{st}},\ e^{\mathrm{st}}$	Stance score and stance error
$\mathcal{H}(x; a, b)$	Smooth ramp from 0 at a to 1 at b
M_t	Rigid-handle recovery memory
g_t^{dist}	Disturbance gate of the suspended-handle termination
$t_g,\ T$	Grace period and duration of a termination check

Learning

Δt	Control interval, 0.02 s
$\pi_\theta^{(c)},\ V_\phi$	Actor and critic
$\boldsymbol{o}_t$	Single-frame actor observation
$\widetilde{\boldsymbol{o}}_t$	Two-frame actor observation history
$\boldsymbol{u}_t$	Normalized policy action
$\boldsymbol{\omega}_{B,t}$	Base angular velocity in the base frame
$\boldsymbol{\alpha}_{\mathrm{imu},t}$	Linear acceleration measured by the torso IMU
$\boldsymbol{g}_{B,t}$	Gravity direction projected into the base frame
$\Delta\boldsymbol{q}_{j,t}$	Joint offset from the nominal position
$J(\theta),\ L^{\mathrm{CLIP}}$	Expected discounted return and PPO surrogate objective
$\gamma,\ \lambda,\ \epsilon$	Discount factor, generalized-advantage parameter, and PPO clipping parameter
$\beta_t,\ \hat{A}_t$	PPO probability ratio and advantage estimate
N_{ep}	Number of control steps in an episode
$r_t^{(c)},\ \bar{r}_t^{(c)}$	Per-step reward of condition c and its unscaled sum
$w_i^{(c)}$	Weight of reward term i

Rewards

$g_t,\ \omega_t$	Base tilt and roll–pitch rate used by the no-handle recovery terms
$I_t^{\mathrm{u}},\ I_t^{\mathrm{s}}$	Instability and recovered-stability indicators
$\bar{U}_t,\ U_t,\ L_t,\ T_t$	Decaying memories of instability, liftoff, and touchdown
n_t	Number of loaded feet
F_{load}	Foot-loading threshold
$c^{\mathrm{slip}},\ c^{\mathrm{ctr}},\ c^{\mathrm{h}},\ c_t^{\mathrm{air}}$	Foot-slip, centering, low-height, and airborne-foot costs
$s_{q,t},\ s_t^{\mathrm{st}},\ s_{R,t}$	READY joint, stance, and orientation scores
$\widetilde{s}_{q,t},\ \widetilde{s}_t^{\mathrm{st}},\ \widetilde{s}_{R,t}$	Relaxed READY scores
$d_t,\ D_t$	Instantaneous recovery demand and early-step demand
$\lambda_{\mathrm{rec}},\ \lambda_{\mathrm{step}},\ \lambda_M$	Decay factors of the recovery memories
$F_{\Sigma,t},\ F_{\min,t}$	Total and lighter-foot load
μ_t	Lighter-foot multiplier of the rigid-handle support reward
Φ_t	Product of the stance and foot-yaw scores
Υ_t	Torso-upright factor of the suspended-handle support reward

Evaluation

$n,\ n_c,\ \hat{p}$	Number of episodes, completed episodes, and completion rate
$\alpha_t,\ \alpha_{\max}$	Platform-relative base tilt and its episode maximum
$d_t^{\mathrm{b}},\ d_{\min}^{\mathrm{b}}$	Boundary margin and its episode minimum
$n_s,\ T_{\mathrm{ep}}$	Number of corrective steps and episode duration
η	Foot-support fraction
$F_{H,95},\ \tau_{H,95}$	95th-percentile hand force and hand moment
$\phi,\ \theta$	Roll and pitch of the robot base
$J_x,\ J_x^{95}$	Whole-body state-deviation index and its 95th percentile
σ_{att}	Torso-attitude variation

Contents

List of Acronyms **ix**

Nomenclature **xi**

List of Figures **xix**

List of Tables **xxiii**

1 Introduction **1**
1.1 Motivation . . . 1
1.2 Problem and Research Gap . . . 2
1.3 Aim and Research Questions . . . 3
1.4 Contributions . . . 3
1.5 Scope . . . 4
1.6 Thesis Outline . . . 4

2 State of the Art **5**
2.1 Balance and Recovery . . . 5
2.2 Model-Based Control . . . 6
2.3 Reinforcement Learning Based Control . . . 7
2.4 Moving Support Surfaces . . . 8
2.4.1 Free Standing Balance . . . 9
2.4.2 Hand-Supported Balance . . . 9
2.5 Humanoid Hardware . . . 10
2.6 Research Gap . . . 11

3 System Description and Problem Formulation **13**
3.1 System Overview . . . 13
3.2 Robot and Control Interface . . . 13
3.3 Coordinate Frames . . . 15
3.4 Moving-Platform Environment . . . 16
3.5 Support Conditions . . . 17
3.5.1 No-Handle Support . . . 17
3.5.2 Rigid-Handle Support . . . 17
3.5.3 Suspended-Handle Support . . . 18
3.6 Balance and Recovery Task . . . 19
3.7 Assumptions and Proposed Workflow . . . 20

4 Reinforcement Learning Framework **23**
4.1 Framework Overview . . . 23
4.2 Training Disturbance Distributions . . . 24
4.3 Policy Interface . . . 25
4.3.1 Actor Observations . . . 25
4.3.2 Asymmetric Critic . . . 26
4.3.3 Joint-Position Actions . . . 27
4.3.4 Network Architecture . . . 27
4.4 Reward Structure . . . 27
4.5 No-Handle Balance and Recovery . . . 29
4.5.1 Stabilization and Floor Use . . . 30
4.5.2 Event-Gated Step Recovery . . . 31
4.6 Rigid-Handle Balance and Recovery . . . 33
4.6.1 Support-Ready Reference . . . 34
4.6.2 Recovery Demand and Relaxation . . . 34
4.6.3 Foot Support and Handle Load . . . 36
4.6.4 Recovery Step and Return to READY . . . 37
4.7 Suspended-Handle Extension . . . 39

5 Implementation and Evaluation Methodology **43**
5.1 Training . . . 43
5.1.1 Training Curriculum . . . 43
5.1.2 Policy Selection and Export . . . 43
5.1.3 Software and Model Versions . . . 45
5.2 Evaluation in Simulation . . . 45
5.2.1 Isaac Lab . . . 45
5.2.2 MuJoCo . . . 45
5.2.3 Metrics . . . 46
5.3 Evaluation in the Real World . . . 47
5.3.1 Physical Deployment . . . 47
5.3.2 Test Site and Vehicle . . . 48
5.3.3 Controller–Support Conditions . . . 49
5.3.4 Instrumentation and Synchronization . . . 49
5.3.5 Maneuver Matrix and Execution . . . 51
5.3.6 Data Processing and Metrics . . . 52
5.3.7 Trial Inclusion . . . 54

6 Simulation Results **55**
6.1 Training . . . 55
6.2 Isaac Lab Evaluation . . . 56
6.3 Recovery Behavior . . . 58
6.3.1 No-Handle Step Recovery . . . 58
6.3.2 Rigid-Handle Support . . . 59
6.4 MuJoCo Sim-to-Sim Evaluation . . . 60
6.5 Suspended-Handle Extension . . . 61
6.6 Summary . . . 64

7 Physical Experiments and Results **67**
7.1 Overview of Completed Trials 67
7.2 Paired Cross-Condition Comparison 68
7.3 Straight-Driving Results 69
7.4 Turning Results 72
7.5 Extended Straight-Road Sequence 74
7.6 Failure Cases and Safety Interventions 75
7.7 Summary and Key Physical Findings 76

8 Discussion **77**
8.1 Balance without Direct Platform-Motion Observation 77
8.2 Support-Specific Recovery Strategies 78
8.2.1 No-Handle Stepping 78
8.2.2 Rigid-Handle Support 78
8.2.3 Suspended Handle 78
8.3 Transfer from Isaac Lab through MuJoCo to Hardware 79
8.4 Comparison with the Manufacturer's Standing Controller 79
8.5 Implications for Humanoid Operation in Public Transport 80
8.6 Limitations 80

9 Conclusions and Outlook **83**
9.1 Summary of the Work 83
9.2 Main Findings 83
9.3 Outlook 84

Bibliography **87**

A Physical Trial Records **I**
A.1 Trial Inventory I
A.2 Additional Physical Results I
A.3 Time Histories of All Trials II

List of Figures

1.1 Conceptual overview of humanoid balance on a finite floor in a moving public-transport vehicle. Vehicle acceleration, braking, and turning can induce body sway and require a corrective step, while a nearby rigid handhold provides an additional support option. 2

2.1 Two complementary dimensions of the representative work reviewed in this chapter. The upper-left group distinguishes control approaches, whereas the upper-right group distinguishes problem and contact settings; an individual study may therefore relate to both groups. The lower band identifies the intersection addressed in this thesis. 6

2.2 Conceptual summary of balance recovery and moving-support disturbances. (a) The CoM projection, CoP/ZMP, and support polygon describe balance under the current contacts, while divergent motion may require a corrective step. (b) A short body push acts directly on the torso while the support remains stationary, whereas longitudinal acceleration, braking, and turning of a moving support act through the active foot and hand contacts. 9

3.1 Isaac Lab realizations of the finite moving-platform system under (a) no-handle, (b) rigid-handle, and (c) suspended-handle support, with the coordinate frames used in this chapter (x red, y green, z blue). The dimensions refer to the common rigid platform, and h_B is the height of the base above the platform origin. The dashed rectangle in (a) marks the usable floor area of Eq. (3.4). The suspended-handle condition is evaluated only in simulation. 14

3.2 Workflow from the moving-platform task to simulation and physical evaluation. 21

4.1 Overview of the learning framework. The loop in the upper row is the deployed controller, which runs identically in simulation and on the robot. The components in the shaded box are used only during training in simulation: the critic receives the actor observations together with privileged simulation state, and PPO updates both the actor and the critic. 23

4.2 Platform-motion families used for training the no-handle and rigid-handle policies. Each sketch shows the qualitative speed, path, or height profile of a family, and the percentage gives the probability with which it is sampled at an episode reset. The road overlay is sampled independently and superimposed on every rigid-handle family. The sketches use platform-fixed axes (x, y, z) (longitudinal, lateral, vertical), path speed v, acceleration components (a_x, a_y), path distance s, turn radius R, roll ϕ, and pitch θ. 26
4.3 Actor–critic structure used for PPO training. The red dashed outlines group the trainable networks, while the red arrows show their parameter updates. Blue arrows show the training data flow and the mean-action path used after deployment. The critic and PPO update are used only during training. 28
4.4 Event-gated no-handle recovery terms. Recent instability gates a stable-state term and, following liftoff and touchdown, a step-recovery term. The decay of the memories allows either term to contribute on successive eligible control steps. 32
4.5 Rigid-handle reward hierarchy. Nominal READY tracking and bilateral foot loading give way to relaxed tracking and an optional corrective step when recovery demand rises. The policy then restores foot support and READY posture. Feet-primary support remains rewarded throughout, while hand-load penalties remain active after their startup grace period. 34

5.1 Onboard controller on the G1. Left: operating modes of the finite-state machine. Right: the policy thread builds the observation history from the low state, evaluates the actor, and converts its output into joint targets at 50 Hz; the state-machine loop sends the latest targets with the joint gains to the robot at 1 kHz. 48
5.2 Physical test platform. (a) The Chalmers REVERE research van. (b) The Unitree G1 in the prepared cargo compartment, with the approximate length and width of the usable floor. The flat central floor, covered wheel-arch structures, overhead support frame, and rear safety tether define the robot's physical test environment. (c) The rigid-handle installation and a CAD view of the replacement hand component; the red box marks the connection to the vehicle-fixed support. 49
5.3 Top view of the test vehicle with the measurement equipment. 50
5.4 Measured vehicle speed of the straight-road runs, aligned at the start of motion. The dashed horizontal lines mark the target speeds of Table 5.7. 52

6.1 Final training stage of the no-handle and rigid-handle policies. Top: mean episode reward. Bottom: share of episodes ending by the time limit and by the most frequent termination conditions. Lines show the mean over consecutive blocks of 500 (no handle) and 160 (rigid handle) iterations; the shaded band in the top row is the interquartile range within each block. 56
6.2 Straight-line severity sweep in Isaac Lab, 200 episodes per level. Top: completed episodes with 95% Wilson intervals. Bottom: corrective steps per 10 s (no handle) and 95th-percentile hand torque (rigid handle), as median and interquartile range. Dark shading marks the main training range of the peak acceleration and light shading the rarely sampled tails. 58
6.3 No-handle episode at a peak acceleration and braking of 1.5 m/s^2 in Isaac Lab. Top: frames (a) at rest, (b) during acceleration, (c) during braking, and (d) after recovery. Bottom: platform acceleration, platform-relative base pitch and roll, vertical foot forces normalized by the body weight B, and longitudinal position of the feet and the base on the platform; the dotted line marks the edge of the usable floor area. Shaded intervals mark acceleration and braking, and the thin red lines mark the approximate times t of the frames. 59
6.4 Rigid-handle episode at a peak acceleration and braking of 3 m/s^2 in Isaac Lab. Top: frames (a) at rest, (b) during acceleration, (c) during braking, and (d) after recovery. Bottom: platform acceleration, platform-relative base pitch and roll, vertical foot forces, hand force, and hand torque; dashed lines mark the termination limits and shaded intervals mark acceleration and braking; the thin red lines mark the approximate times t of the frames. 60
6.5 MuJoCo sim-to-sim evaluation of the no-handle and rigid-handle policies under (a) straight-line acceleration and (b) turning. The frames show both policies at the instants t_0–t_3 marked below them, starting from rest at $t_0 = 0$. The lower panels show the platform input: speed v, magnitude of the horizontal platform acceleration $\|\boldsymbol{a}\|$, and, for the turn, the yaw rate $|\omega_z|$. 62
6.6 Straight-line severity sweep in MuJoCo, 30 episodes per level. Top: completed episodes with 95% Wilson intervals. Bottom: corrective steps per 10 s (no handle) and 95th-percentile hand torque (rigid handle), as median and interquartile range. Shading as in Figure 6.2. 63
6.7 Suspended-handle episode on a fixed straight-line trajectory in Isaac Lab. Top: frames (a) at rest, (b) during acceleration, (c) during braking, and (d) recovery after the platform has stopped. Bottom: platform acceleration, platform-relative base pitch and roll, vertical foot forces, handle swing angle, and estimated hand force; forces are normalized by the body weight B. Shaded intervals mark acceleration and braking, and the thin red lines mark the approximate times t of the frames. 64

7.1 Paired physical-response metrics for the common straight and turning identifiers. Each circle is one recorded trial, thin lines connect equal target identifiers, and diamonds denote condition medians across the tested identifiers. J_x^{95} is displayed on a logarithmic scale. 68
7.2 S4 straight-road case. Top: robot-view frames before (a), during (b), and after (c) the main response in each condition. Bottom: time histories over the complete overlapping analysis interval, showing vehicle speed, longitudinal acceleration a_x, torso roll and pitch relative to their median over the first 2 s, and the state-deviation index J_x on a logarithmic scale. Vertical lines mark the frame times (a)–(c). . . . 71
7.3 T12 turning case. Top: robot-view frames (a)–(c) showing the observed motion within each steady-turn window. Bottom: time histories over the complete trial, showing vehicle speed, lateral acceleration a_y with the sign chosen positive in the turn, torso roll and pitch relative to their median over the first 2 s, and J_x on a logarithmic scale. The shaded band marks the automatically selected steady-turn window, and the vertical lines mark the frame times (a)–(c). 73
7.4 Extended straight-road comparison between the no-handle and rigid-handle learned configurations. Circles show the individual S6–S10 results, colored lines connect equal identifiers, and diamonds denote condition medians. The J_x^{95} panel uses a logarithmic scale. 75
7.5 Video sequences of two visible loss-of-balance cases under the default controller. In both S4 and S5, the robot moves toward the right-side wheel-arch or cabin interior and the rear safety tether becomes visibly taut, limiting further descent. 75

A.1 Time histories of straight-road trial S1. III
A.2 Time histories of straight-road trial S2. III
A.3 Time histories of straight-road trial S3. IV
A.4 Time histories of straight-road trial S5. IV
A.5 Time histories of straight-road trial S6. V
A.6 Time histories of straight-road trial S7. V
A.7 Time histories of straight-road trial S8. VI
A.8 Time histories of straight-road trial S9. VI
A.9 Time histories of straight-road trial S10. VII
A.10 Time histories of turning trial T1 (nominal radius 8 m). VII
A.11 Time histories of turning trial T3 (nominal radius 8 m). VIII
A.12 Time histories of turning trial T4 (nominal radius 10 m). VIII
A.13 Time histories of turning trial T6 (nominal radius 10 m). IX
A.14 Time histories of turning trial T7 (nominal radius 12 m). IX
A.15 Time histories of turning trial T9 (nominal radius 12 m). X
A.16 Time histories of turning trial T10 (nominal radius 20 m). X
A.17 Time histories of the rigid-handle turning trials T2, T5, and T8 (nominal radii 8, 10, and 12 m). XI
A.18 Time histories of the rigid-handle turning trials T11 and T13 (nominal radius 20 m). XI

List of Tables

3.1 Nominal joint positions and joint gains. The right-arm entries in the last column give the support-ready posture of the hand-supported policies. 15
3.2 Common robot and platform model. 16
3.3 Scope of the support conditions. 17
3.4 Termination thresholds of the three support conditions. Grace periods t_g are measured from the start of the episode; durations T are the times for which a condition must hold continuously. 20

4.1 Moving-platform disturbance parameters used for training. a_x and a_y denote longitudinal and lateral platform acceleration, v path speed, $v_{\max}$ the speed cap of a straight trajectory, R the turning radius, and A_z and a_z the vertical displacement and acceleration of the road overlay. 25
4.2 Single-frame actor observation shared by all policies. Noise is uniform and applied during training only. 26
4.3 Policy observation interfaces. 27
4.4 Joint-position action scales in radians per unit policy output. Each value sets the joint-target change produced by one unit of action. . . 28
4.5 Functional comparison of the three policy objectives. Values in parentheses are signed weights $w_i^{(c)}$ in Eq. (4.4). 29
4.6 No-handle event and floor-use parameters. The projected-gravity thresholds are dimensionless; angular-rate thresholds are in rad/s. . . 31
4.7 Complete no-handle reward inventory. The w_i values are signed weights. Shoulder and elbow pose costs supplement the whole-arm cost; the waist row combines two equal contributions. 33
4.8 Rigid-handle READY, recovery-memory, and foot-support parameters. 35
4.9 Signals forming the instantaneous rigid-handle recovery demand d_t. Each gate rises from zero at the deadband to one at full demand; for the lighter-foot load and the stance score, demand rises as the quantity falls. 36
4.10 Parameters of the rigid-handle step and return terms. 38
4.11 Complete rigid-handle reward inventory; w_i denotes the signed weight of each term. 39
4.12 Suspended-handle foot-support, torso-alignment, and hand-load reward parameters. 41

4.13 Complete suspended-handle reward inventory; w_i denotes the signed weight of each term. 41

5.1 PPO settings of the final training stage that differ between the policies. 44
5.2 Shared PPO and network settings. 44
5.3 Training stages of increasing platform-motion difficulty. 44
5.4 Software and hardware. 45
5.5 Platform motions in training, in the simulation evaluations, and in the vehicle tests. Straight-line motion is characterized by the peak longitudinal acceleration $|a_x|$ and the top speed, turning by the speed v, the radius R, and the lateral acceleration $|a_y|$. Physical values are the nominal targets given to the driver. 46
5.6 Instrumentation used in the physical experiments. 50
5.7 Nominal physical-test matrix. The target values guided the driver; the measured vehicle motion was retained for the quantitative analysis. 51
5.8 Target and realized turning maneuvers. The realized radius $\hat{R}$ is the median of $v/|\dot{\psi}|$ and the realized speed $\bar{v}$ the mean GNSS speed over the steady part of each run, in which the yaw rate exceeds 80% of its 95th percentile. A dash marks a maneuver that was not driven in that condition. 52
5.9 Quantities used in the physical-data analysis. 53
5.10 Fixed parameters used to process the physical vehicle and robot records. 54

6.1 Share of episodes ending by each condition, averaged over the last 100 iterations of the final no-handle (25,000 iterations) and rigid-handle (8,000 iterations) training stages. An episode can meet more than one condition at the same step, so a column can sum to slightly more than 100%. 55
6.2 Isaac Lab evaluation within the training distribution, 200 episodes per motion family. Peak tilt is the median of the per-episode maximum; the fourth column gives the median number of corrective steps per episode (no handle) or the median 95th-percentile hand force as a fraction of body weight (rigid handle). 57
6.3 Suspended-handle policy in Isaac Lab, 200 episodes per training family. Tilt, swing angle, hand force, and wrist torque are medians over episodes of the per-episode maximum or 95th percentile. 61
6.4 End of the suspended-handle evaluation episodes, 200 per family. . . 65

7.1 Coverage of the analysis-ready physical dataset. The rigid-handle turning count includes the exploratory T13 run; the three-way subset contains only target identifiers available in every condition. 67
7.2 Median [interquartile range] across the tested identifiers for the two physical-response metrics. Each trial metric is calculated over the central 80% of its overlapping vehicle–robot interval. 68
7.3 S4 case-study input and response summaries. The acceleration is given as its 1st and 99th percentiles. 70
7.4 T12 steady-turn case-study input and response summaries. 72

7.5 Extended S6–S10 straight-road results: median [interquartile range]. . 74

A.1 Retained physical trials by controller–support condition. I
A.2 Per-trial response metrics for the retained straight-road tests. I
A.3 Per-trial response metrics for the retained turning tests. The radius is the nominal skidpad radius specified by the test plan. II

1
Introduction

1.1 Motivation

A humanoid robot deployed at different sites may need to travel between them. Like human commuters, it could use shared public transport instead of a dedicated vehicle. The vehicle can also become the workplace itself. A humanoid on board could check tickets, answer passenger questions, or assist passengers who travel in a wheelchair or with a pram as they board, find a place, and secure it before departure. In either role, the robot must remain standing while the vehicle moves. Acceleration, braking, turning, road irregularities, and vehicle attitude changes then move the floor beneath it. Unlike a short push against the torso, these disturbances act through every active contact, can persist for several seconds, and reverse direction between acceleration and braking. The floor is also finite: a corrective step that prevents an immediate fall consumes part of the space available for the next one.

Humans face the same problem whenever they stand in a moving vehicle. For small disturbances, a standing person keeps the feet in place and rotates mainly about the ankles; as the disturbance grows and the center of pressure approaches the edge of the feet, the hips take an increasing share of the correction [1]. Larger disturbances require a change of support, either a compensatory step or a reach to a handhold [2]. Volunteer tests on a moving platform that reproduces bus braking and acceleration show this progression for passengers standing without a handhold. Only the mildest braking pulse of 1.0 $\mathrm{m/s^2}$ was recovered without stepping, the most effective responses combined steps with a staggered stance and a lowered center of mass, and in the most severe acceleration pulse of 3.0 $\mathrm{m/s^2}$ most backward-facing volunteers had to be caught by the safety harness [3, 4].

Measurements from city-bus operation show that the severity of a disturbance depends on both the magnitude and the duration of the acceleration pulse [5], and studies of standing passengers identify acceleration, jerk, stance, and access to a handhold as the factors that determine whether a passenger remains stable [6, 7]. Without holding on, a passenger keeps balance up to a longitudinal acceleration of roughly 0.11–0.15g [7]. A hard start, sharp braking, or a tight turn exceeds this range, and passengers then hold on to a handrail or a strap instead of relying on their feet alone.

A humanoid faces the same ingredients (Figure 1.1): support motion that it cannot measure in advance, a bounded standing area, and the option of holding on to the vehicle.

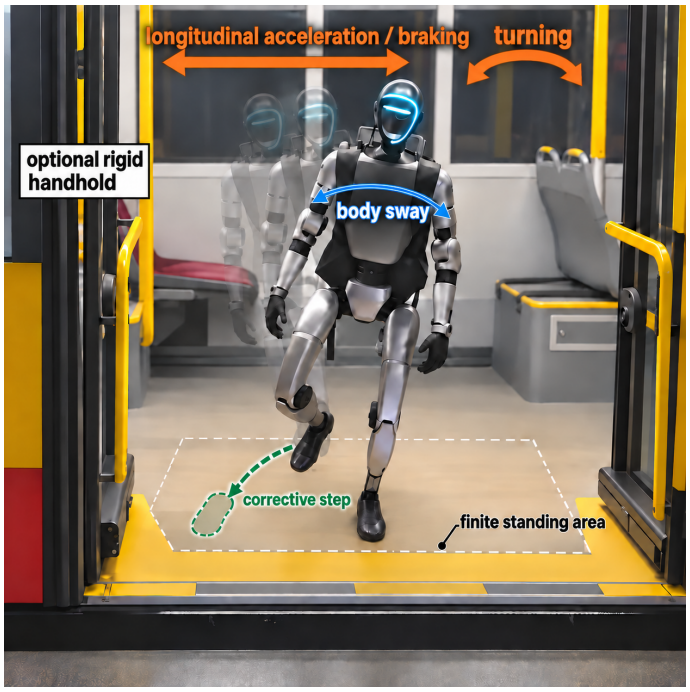


Figure 1.1: Conceptual overview of humanoid balance on a finite floor in a moving public-transport vehicle. Vehicle acceleration, braking, and turning can induce body sway and require a corrective step, while a nearby rigid handhold provides an additional support option.

1.2 Problem and Research Gap

Without hand support, the robot can shift the center of pressure under its feet, use upper-body momentum, or take a corrective step. A rigid handle adds a second load path but closes a kinematic chain through the arm, torso, legs, and vehicle, so that the robot can in principle remain upright by hanging from the handle instead of standing on its feet. The two conditions therefore call for different recovery objectives, even when they share the same onboard sensing and joint-position interface.
Reduced-order models such as the zero-moment point and the capture point explain when a humanoid can balance on its current contacts and where it must step otherwise [8, 9]. Learned policies have transferred standing push recovery to a physical humanoid [10] and produced whole-body motion on full-size humanoids [11], but these results were obtained on stationary ground with disturbances applied to the robot. Studies of moving supports, such as a humanoid on a seesaw [12], a small biped on a rotating platform [13], and planar walking on a swaying surface [14], show that support motion changes the balance problem, but they consider a single motion type and no hand support. How a full-size humanoid can learn recovery under combined vehicle-like motion, how the objective should change when a handle is available, and whether the learned response carries over to a moving vehicle are questions that these works leave open.
We therefore seek a feedback policy that infers the disturbance from its own inertial and joint measurements, steps when the feet alone cannot restore balance, loads

the feet rather than the handle when a rigid handle is available, and transfers from simulation to hardware. Reward design is the main mechanism by which we induce these behaviors; the vehicle experiments test whether they are present on the physical robot.

1.3 Aim and Research Questions

The aim of this thesis is to develop and evaluate reinforcement-learning (RL) policies for humanoid balance and recovery on a finite moving floor representative of public transport, using only robot-mounted inertial and joint measurements at deployment. We address three research questions:

1. **RQ1:** What balance and recovery behavior can a direct joint-position policy achieve on a moving platform whose motion it does not observe, and how does this behavior differ between support-specific objectives for no-handle and rigid-handle support?
2. **RQ2:** To what extent do the learned behaviors transfer from Isaac Lab through MuJoCo to the physical G1 in a moving vehicle?
3. **RQ3:** How does the physical response compare with that of the manufacturer's standing controller under measured straight-driving and turning disturbances?

1.4 Contributions

In this thesis, we train RL policies that keep a 29-degree-of-freedom (DoF) Unitree G1 standing on a moving vehicle floor, and we evaluate them in a moving vehicle. We follow the same progression as a passenger. We first study balance without hand support, where the robot can use only its feet, its whole-body posture, and corrective steps, and determine where this support reaches its limit. We then add a rigid handle for disturbances beyond that limit, where the question becomes how the robot should share the load between its feet and its hand. The central argument of the thesis is that a direct joint-level policy, trained on a finite moving platform with a reward designed for the available support, produces balance and recovery behavior that transfers to a real vehicle and regulates body attitude more closely than the manufacturer's standing controller. Simulation serves to train the policies, to analyze their recovery mechanisms, and to check the deployment interface; the physical vehicle experiments provide the main evidence.

The thesis makes three contributions:

1. a moving-platform balance task in which longitudinal, turning, grade, and road-induced heave, roll, and pitch disturbances reach a 29-DoF humanoid through a finite support surface, while the actor observes only onboard measurements;
2. support-specific reward formulations: event-gated stepping recovery for no-handle support, and a combination of posture recovery, foot support, and handle-load limits for rigid-handle support; and
3. a physical evaluation of both policies in a moving vehicle on a closed proving ground, in which robot state, vehicle motion, and video are recorded syn-

chronously and the learned policies are compared with the manufacturer's standing controller under the same nominal maneuvers.

1.5 Scope

The thesis addresses standing and short-horizon recovery of a robot that is already in place on the vehicle floor; boarding, navigation, and grasp acquisition lie outside its scope. The simulated platform trajectories represent classes of vehicle motion rather than one specific route or vehicle. No-handle and rigid-handle support form the principal study and are evaluated both in simulation and on hardware. A ceiling-pivoted suspended handle is studied in simulation as an extension. The manufacturer's standing controller serves as the physical reference because it is the controller the G1 would otherwise use in this situation.

1.6 Thesis Outline

Chapter 2 reviews the balance, learning, moving-support, and hand-support concepts needed by the method and the choice of humanoid hardware. Chapter 3 defines the robot, platform, support mechanics, and recovery task. Chapter 4 presents the policy interface and reward design, and Chapter 5 gives the implementation and evaluation protocol. Chapters 6 and 7 report simulation and physical results. Chapter 8 interprets the observed behavior and transfer, and Chapter 9 concludes the thesis.

2

State of the Art

This chapter introduces the balance, learning, moving-support, and hand-support concepts on which the method builds, and positions the thesis with respect to prior work in each area. The chapter first distinguishes stabilization on the current contacts from recovery by stepping, then reviews model-based and reinforcement learning based control, reviews how moving floors and hand support change the balance problem, and finally discusses how the size and actuation of a humanoid affect it. Figure 2.1 groups the representative studies along two dimensions: how the balance behavior is generated and in which support setting it is required.

2.1 Balance and Recovery

For links with masses m_i and world-frame center-of-mass positions $\boldsymbol{p}_i \in \mathbb{R}^3$, the whole-body center of mass (CoM) is

$$\boldsymbol{p}_{\mathrm{CoM}} = \frac{1}{\sum_i m_i} \sum_i m_i \boldsymbol{p}_i. \tag{2.1}$$

On a stationary horizontal surface, the vertical projection of the CoM and the center of pressure (CoP) describe how the current foot contacts are used. The support polygon is the convex hull of the active contact regions; a CoP outside this polygon cannot be generated while the assumed flat contacts are maintained.

These geometric quantities are necessary but not sufficient conditions for balance, because they neglect momentum, friction limits, actuator limits, and future contact changes. The zero-moment point (ZMP) adds inertial effects under reduced-order assumptions. For a linear inverted pendulum with constant CoM height z_c,

$$\ddot{x} = \omega_0^2 (x - p), \qquad \omega_0 = \sqrt{g/z_c}, \tag{2.2}$$

where x is the horizontal CoM position, p is the ZMP position, and g is the gravitational acceleration. ZMP preview control is a foundation of biped walking [8]; it assumes a stationary, planar support.

Stepping becomes necessary when the current contacts cannot arrest the divergent part of this motion. For the same model, the divergent component of motion (DCM) is $\xi = x + \dot{x}/\omega_0$, and capture-point analysis identifies the foot placement that brings this component to rest [9, 23]. The concept explains why a controller must eventually unload a foot and step once ankle, hip, trunk, and arm motion are no longer sufficient, even when, as in this thesis, no foothold is commanded explicitly.

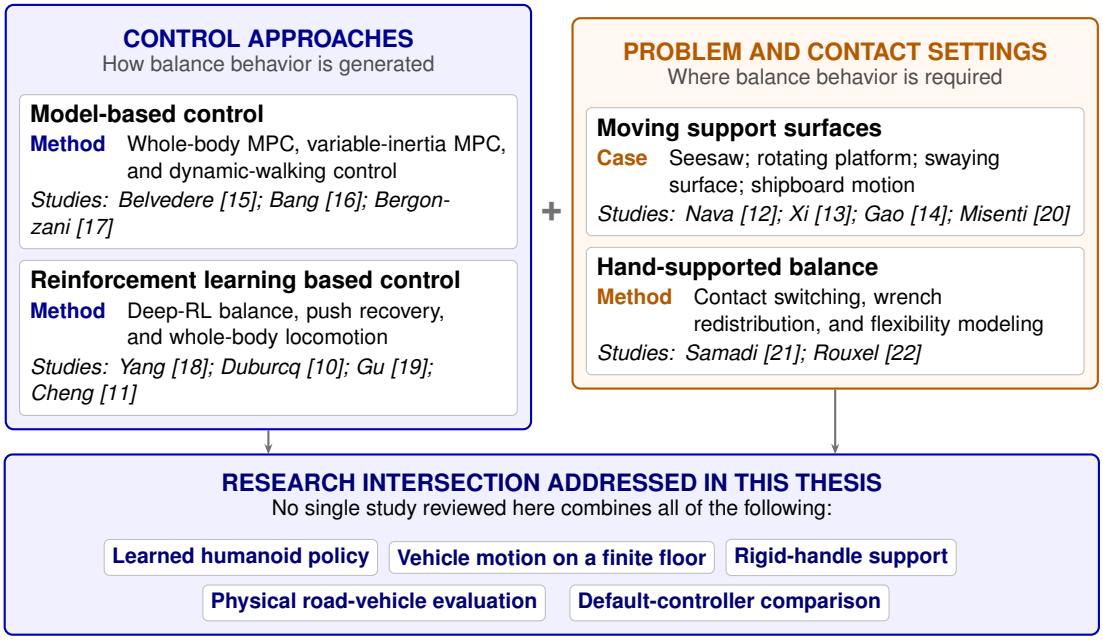


Figure 2.1: Two complementary dimensions of the representative work reviewed in this chapter. The upper-left group distinguishes control approaches, whereas the upper-right group distinguishes problem and contact settings; an individual study may therefore relate to both groups. The lower band identifies the intersection addressed in this thesis.

2.2 Model-Based Control

Model-based controllers formulate humanoid balance and locomotion as an optimal control problem over the robot state, the joint torques, and the contact forces. The problem is constrained by the whole-body dynamics, by unilateral contact and friction limits, and by joint and actuator limits. Wensing et al. identify three difficulties in solving it online: contact makes the dynamics non-smooth, the constraints are non-convex, and the state of a legged robot is high-dimensional [24]. Practical controllers therefore split the problem into smaller parts.

The first split replaces the whole-body dynamics by a reduced-order model for planning. The linear inverted pendulum of Eq. (2.2) and the centroidal dynamics, which relate the total contact wrench to the rate of change of the linear and angular momentum, are small enough for real-time optimization. Their assumptions, such as a constant CoM height or negligible limb inertia, also restrict the motions that can be planned. The second split separates planning from feedback. A model-predictive controller (MPC) plans CoM motion, contact forces, and footholds over a receding horizon, and a whole-body controller solves a small quadratic program (QP) at each control step to turn these references into joint torques that respect the contact and joint constraints [24].

Recent work narrows the gap between the two levels. Whole-body MPC couples centroidal feasibility to joint-level decisions and has been used for disturbance rejection with upper-body motion [15], while variable-inertia MPC reduces the mismatch between reduced-order and full-body dynamics during fast motion [16]. Such methods

have enabled dynamic walking on full-size humanoids [17].
The guarantees of these controllers hold only as far as their models are accurate [24]. Contact friction, actuator dynamics, and joint compliance must be identified or bounded, and on a moving floor the controller additionally needs an estimate of the support motion, which the robot cannot measure independently of its own response. Wensing et al. note that learning-based methods handle effects that are hard to model, and they argue for combining the two approaches [24].

2.3 Reinforcement Learning Based Control

Learning-based control replaces part of the online optimization with a policy trained from interaction. Deep RL has produced human-comparable balancing behavior in simulation [18], standing push recovery on a physical humanoid [10], and whole-body locomotion and expressive motion on real humanoids [19, 11]. Massively parallel simulation makes it practical to train such policies over broad disturbance distributions [25]. A direct policy uses one action space for quiet standing, upper-body compensation, foot unloading, and stepping, without a separate foothold planner.
This flexibility shifts the burden to the training objective. A policy rewarded mainly for nominal posture resists corrective motion, whereas a policy rewarded mainly for survival can exploit simulator contacts or move excessively. For this reason, we evaluate the reward design through the resulting behavior, its failure cases, and its transfer to hardware, and not through training return alone.
The robot state is only partially observed through joint encoders and an inertial measurement unit (IMU), so the task is a partially observable Markov decision process. A policy π_θ with parameters θ maps an observation $\boldsymbol{o}_t$ to a normalized action $\boldsymbol{u}_t$ and is trained to maximize the expected discounted return

$$J(\theta) = \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{N_{\text{ep}}-1} \gamma^t r_t\right], \tag{2.3}$$

where r_t is the reward at control step t, $\gamma \in (0,1)$ is the discount factor, and N_{ep} is the number of control steps in an episode. A short observation history lets the actor infer quantities, such as accelerations of the support, that cannot be identified from a single sample.
We use proximal policy optimization (PPO) [26], which maximizes the clipped surrogate

$$L^{\text{CLIP}}(\theta) = \mathbb{E}_t\Big[\min\Big(\beta_t(\theta)\hat{A}_t, \text{clip}(\beta_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\Big)\Big], \tag{2.4}$$

where $\beta_t(\theta) = \pi_\theta(\boldsymbol{u}_t \mid \boldsymbol{o}_t)/\pi_{\theta_{\text{old}}}(\boldsymbol{u}_t \mid \boldsymbol{o}_t)$ is the probability ratio between the updated and the previous policy, $\hat{A}_t$ is an advantage estimate, and ϵ bounds the change of the policy in one update. The advantage is computed with generalized advantage estimation [27] from a learned value function V_ϕ, the critic. Because the critic is used only during training, it may receive simulator-only quantities that the deployed actor cannot measure; such an asymmetric critic [28] improves value estimation without changing the information available to the actor. The contribution of this thesis lies in the task, the reward, and the evaluation; the optimization algorithm is used unchanged.

The actor outputs desired joint-position offsets that a lower-level proportional–derivative controller tracks. Training and deployment therefore share the joint ordering, nominal pose, action scaling, update rate, and actuator gains. Two forms of randomization play different roles: randomizing the platform motion defines the range of tasks the policy must solve, whereas randomizing quantities such as friction, masses, delays, or gains addresses uncertainty in the simulation model. Evaluation in a second simulator exposes sensitivity to one numerical contact model, and hardware evaluation tests the remaining modeling gap.

2.4 Moving Support Surfaces

A push and a moving floor can cause similar initial body motion but differ dynamically. A push acts on selected links while the contact frame stays fixed; platform motion drives every active contact and may persist, reverse during braking, or combine translation with rotation. For a point mass m expressed in a frame attached to the platform,

$$\begin{aligned} m\boldsymbol{a}_{\mathrm{rel}} = \sum \boldsymbol{f} + m\boldsymbol{g} - m\boldsymbol{a}_P - 2m\boldsymbol{\Omega} \times \boldsymbol{v}_{\mathrm{rel}} \\ - m\dot{\boldsymbol{\Omega}} \times \boldsymbol{r} - m\boldsymbol{\Omega} \times (\boldsymbol{\Omega} \times \boldsymbol{r}), \end{aligned} \tag{2.5}$$

where $\boldsymbol{r}$, $\boldsymbol{v}_{\mathrm{rel}}$, and $\boldsymbol{a}_{\mathrm{rel}}$ are the position, velocity, and acceleration of the mass relative to the platform origin, $\sum \boldsymbol{f}$ is the sum of contact forces, $\boldsymbol{g}$ is the gravity vector, $\boldsymbol{a}_P$ is the acceleration of the platform origin, and $\boldsymbol{\Omega}$ is the platform angular velocity. The last four terms show that linear acceleration, turning, and changes of turning rate each require different contact forces, even when the robot holds the same posture relative to the floor. Figure 2.2 summarizes the balance quantities and contrasts a push with motion of the support.

Public-transport disturbances comprise longitudinal acceleration and braking, lateral acceleration and yaw during turns, grade, and road- or track-induced vertical and rotational motion. Measured events are time histories rather than single peak values [5, 6], and access to a handhold changes passenger stability [7]. These observations lead us to train on complete platform trajectories and to report the measured vehicle motion alongside every physical result.

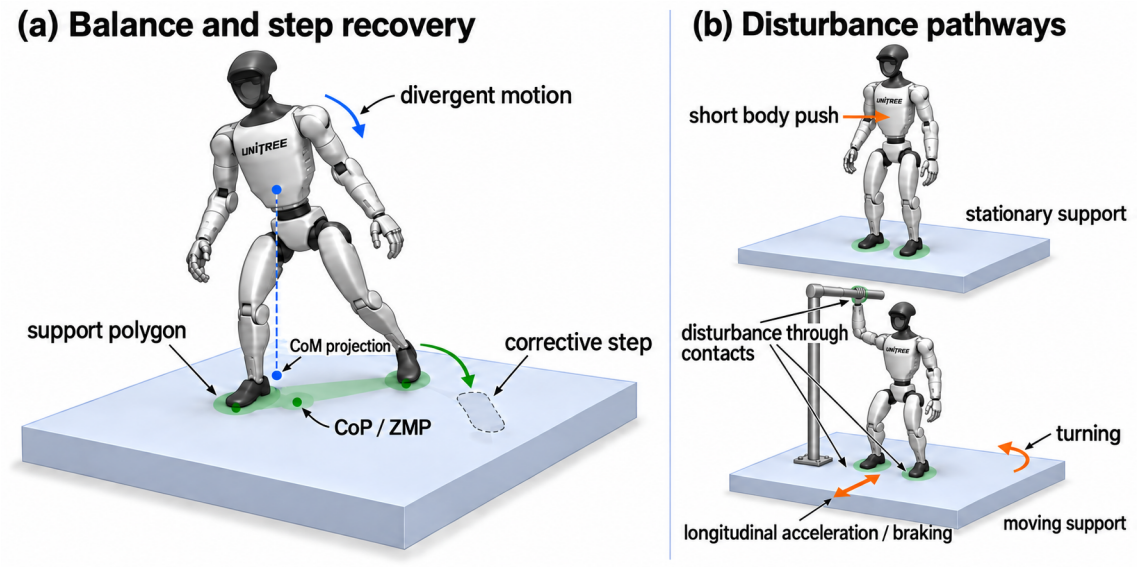


Figure 2.2: Conceptual summary of balance recovery and moving-support disturbances. (a) The CoM projection, CoP/ZMP, and support polygon describe balance under the current contacts, while divergent motion may require a corrective step. (b) A short body push acts directly on the torso while the support remains stationary, whereas longitudinal acceleration, braking, and turning of a moving support act through the active foot and hand contacts.

2.4.1 Free Standing Balance

We first consider balance without a handhold, in which the feet are the only support, and then turn to hand-supported balance. Several studies address parts of the free-standing problem. Nava et al. model and control the iCub humanoid balancing on a seesaw whose motion is coupled to the robot [12]; the support moves about a single axis. Xi et al. learn balance for a small biped on a rotating platform [13], which shows that RL can compensate support rotation but does not address a full-size humanoid or translational acceleration. Gao et al. derive a time-varying angular-momentum model for planar walking on a horizontally swaying surface [14], and Misenti et al. report degraded stability and tracking of commercial quadrupeds under shipboard motion [20]. None of these studies combines a full-size humanoid, several vehicle-motion classes, a finite floor, and hand support, and none evaluates a learned policy in a moving road vehicle.

2.4.2 Hand-Supported Balance

A rigid hand contact enlarges the set of wrenches that the robot can apply to its environment but closes a kinematic chain through the arm, torso, legs, and platform. Excessive reliance on the hand can unload a foot, create large internal forces, or drive the arm toward a joint limit. Multi-contact controllers therefore regulate posture and wrench distribution together [21], and an explicit model of joint flexibility improves force regulation on position-controlled humanoids [22]. These results motivate a separate objective for rigid-handle balance rather than attaching a hand to the no-

handle controller.
A handle suspended from a strap is more restrictive, because the strap is approximately tension-only,

$$\boldsymbol{f}_h = f_h \boldsymbol{e}_r, \qquad f_h \geq 0, \tag{2.6}$$

where f_h is the strap tension and $\boldsymbol{e}_r$ is the unit vector along the strap toward its anchor. A slack strap supplies no force, the load direction changes with the geometry, and the support cannot impose an arbitrary hand pose or moment. The suspended-handle model used in this thesis captures the changing support direction through a rigid link on a ceiling ball joint, but not slack or tension-only loading.

2.5 Humanoid Hardware

Humanoid platforms differ in size and in actuation, and both change the balance problem studied here.
Size sets the time scale and the required contact forces. In the frame of a floor that accelerates horizontally at a_P, Eq. (2.2) gains the inertial term $-a_P$, and holding the CoM at rest relative to the floor requires

$$x - p = \frac{z_c}{g}\, a_P. \tag{2.7}$$

The ZMP offset that a given floor acceleration demands therefore grows with the CoM height, whereas the foot length bounds the offset that the feet can provide. A taller robot also has a lower ω_0, so its divergent motion develops more slowly, but it falls from a greater height. Contact and hand forces scale with the robot mass for a given floor acceleration. Full-size research humanoids such as TALOS are about 1.75 m tall and weigh about 95 kg [29], whereas the Unitree G1 used in this thesis is 1.32 m tall [30].
Three actuation concepts are common in legged robots. Electric motors with harmonic drives provide high reduction ratios in a compact joint and precise position control, but their reflected inertia and friction make the joints stiff and hardly backdrivable; torque control then requires a torque sensor in each joint, as in TALOS [29]. Series elastic actuators place a spring between the gearbox and the load, so that its deflection measures the joint torque and buffers impacts, at the cost of a bandwidth limited by the spring stiffness [31]. Quasi-direct-drive (QDD), or proprioceptive, actuators combine a high-torque motor with a low-ratio gearbox. Their low reflected inertia keeps the joints backdrivable, the joint torque can be estimated from the motor current, and impacts are transmitted to the gearbox less severely [32]; the MIT Humanoid uses actuators of this type [33].
The actuation affects balance on a moving floor in two ways. Stepping and sudden floor motion cause foot impacts, which stiff, highly geared joints pass on to the gearbox and the structure. A rigid handle closes a kinematic chain through the arm, torso, legs, and floor, and small position errors in stiff joints then cause large internal forces, whereas backdrivable joints under moderate proportional–derivative gains yield to them.
The G1 is a short humanoid with QDD-type joints: low-inertia permanent-magnet synchronous motors [30] drive the joints through low-ratio gearboxes, and the joints

have no torque sensors. We chose it for three reasons. Its CoM height lowers the ZMP offset required by Eq. (2.7), and its small size and mass suit a cargo area and reduce the consequences of a fall during tests in a moving vehicle. Its backdrivable joints tolerate the impacts of corrective steps and the closed chain through a rigid handle. Finally, its joint-level position interface and the available simulation model allow a policy trained in simulation to be deployed without a separate torque controller.

2.6 Research Gap

The work reviewed above leaves three gaps that this thesis addresses. First, learned balance of humanoids has been shown on stationary ground and under pushes [10], and moving-support studies consider single motion types or small robots [12, 13, 14]; to our knowledge, a full-size humanoid under combined vehicle-like motion on a finite floor has not been studied. Second, multi-contact control regulates hand and foot wrenches with an explicit model [21, 22], but the reviewed work does not address how an RL objective should prevent a policy from hanging on a rigid handle while still exploiting it. Third, to our knowledge, no learned balance policy of this kind has been evaluated in a moving road vehicle against the robot's own standing controller.

3

System Description and Problem Formulation

This chapter defines the robot, the moving platform, the support conditions, and the balance and recovery task. These definitions are independent of the observations and rewards of the learned policies.

3.1 System Overview

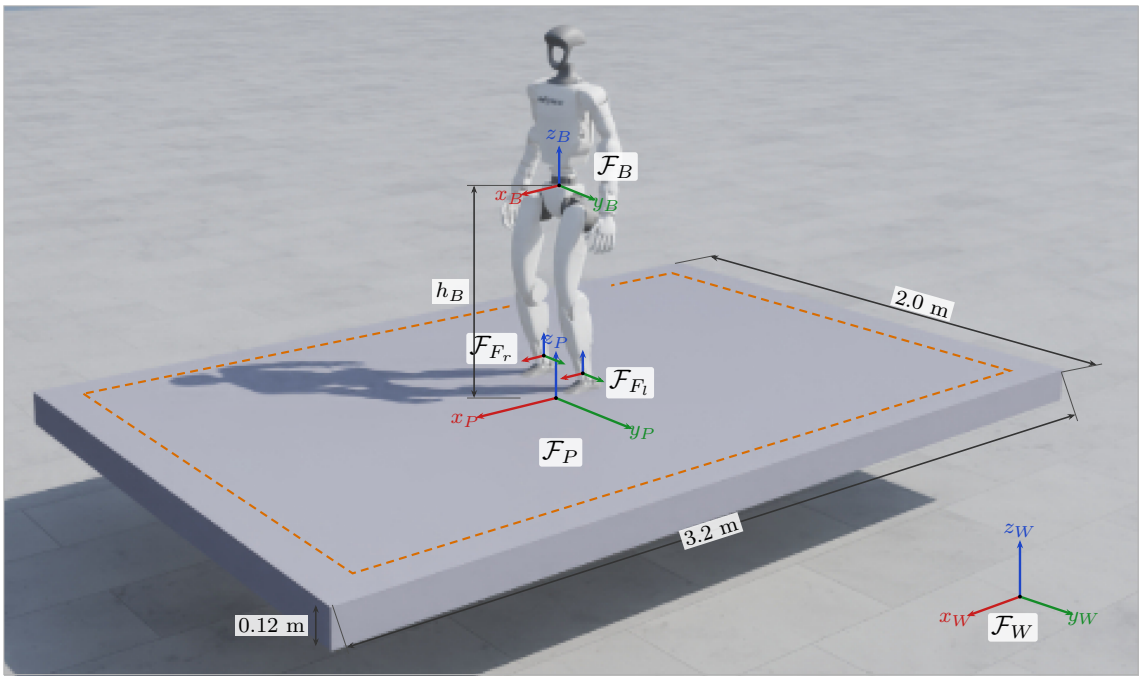
The system consists of a 29-DoF Unitree G1 standing on a finite rigid platform that represents a vehicle floor. Prescribed platform trajectories generate longitudinal and turning motion, which reaches the robot through its contacts. The no-handle training set additionally contains grade motion, and the rigid-handle training set superimposes road-induced heave, roll, and pitch. Because a deployed robot cannot measure the future motion of the vehicle, we give the actor no information about the platform trajectory; it must infer the disturbance from its own inertial and joint measurements.

The robot used throughout this work is a Unitree G1 that we named William. Its state is measured by an inertial measurement unit (IMU) in the torso, which provides angular velocity, linear acceleration, and orientation, and by the encoders of the 29 joint motors, which provide joint positions and velocities. The head also carries a depth camera and a lidar, which we do not use. The deployed controller therefore relies on inertial and joint measurements only; the contact forces that the simulation provides for the reward, the termination checks, and the analysis have no counterpart on the robot.

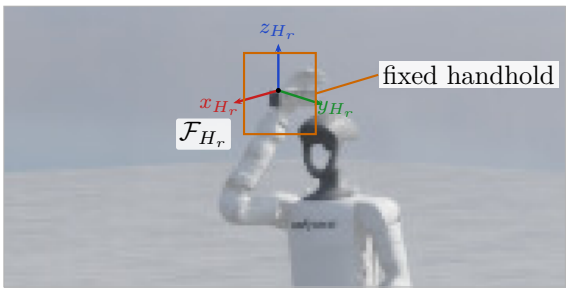
We consider three support conditions, illustrated in Figure 3.1. In the *no-handle* (NH) condition, the feet are the only intended contacts and recovery may include a step. In the *rigid-handle* (RH) condition, the right palm is additionally fixed to a frame attached to the platform. These two conditions form the main study and are evaluated both in simulation and on hardware. In the *suspended-handle* (SH) condition, which we study in simulation, the platform-fixed frame is replaced by a triangular handle that swings about a ceiling pivot.

3.2 Robot and Control Interface

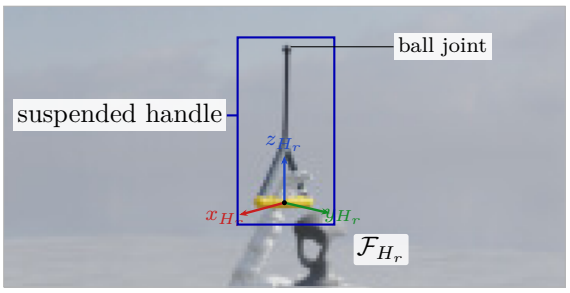
The simulated robot uses the manufacturer's model of the 29-DoF G1, which has 29 revolute joints. Twelve joints actuate the legs, three the waist, and fourteen the



(a) No handle



(b) Rigid handle



(c) Suspended handle

Figure 3.1: Isaac Lab realizations of the finite moving-platform system under (a) no-handle, (b) rigid-handle, and (c) suspended-handle support, with the coordinate frames used in this chapter (x red, y green, z blue). The dimensions refer to the common rigid platform, and h_B is the height of the base above the platform origin. The dashed rectangle in (a) marks the usable floor area of Eq. (3.4). The suspended-handle condition is evaluated only in simulation.

arms:

$$\boldsymbol{q}_j = \begin{bmatrix} \boldsymbol{q}_{\text{legs}}^\top & \boldsymbol{q}_{\text{waist}}^\top & \boldsymbol{q}_{\text{arms}}^\top \end{bmatrix}^\top \in \mathbb{R}^{29}. \tag{3.1}$$

The floating base is unactuated. The nominal stance is slightly crouched (Table 3.1). In the hand-supported conditions, each episode begins with the right arm already in a support-ready posture and the hand attached to the support. The rigid- and suspended-handle policies share this posture, so their nominal postures $\boldsymbol{q}_j^0$ are identical.

Table 3.1 lists the nominal joint positions and the proportional–derivative gains of the training configuration, which are exported with the policy and used unchanged on the robot. Left and right joints share the same values unless a sign is given. The

suspended-handle policy uses the same gains except at the right wrist, whose roll, pitch, and yaw joints use $K_p = 280$, 260, and 260 N m/rad and $K_d = 24$, 22, and 22 N m s/rad.

Table 3.1: Nominal joint positions and joint gains. The right-arm entries in the last column give the support-ready posture of the hand-supported policies.

Joint	K_p (N m/rad)	K_d (N m s/rad)	q^0 (no handle) (rad)	q^0 right arm (handle) (rad)
Hip pitch	100	2	−0.10	
Hip roll	100	2	0	
Hip yaw	100	2	0	
Knee	150	4	0.30	
Ankle pitch	40	2	−0.20	
Ankle roll	40	2	0	
Waist yaw	200	5	0	
Waist roll, pitch	40	5	0	
Shoulder pitch	40	1	0.30	−1.75
Shoulder roll	40	1	±0.25	−0.16
Shoulder yaw	40	1	0	0.05
Elbow	40	1	0.97	−0.10
Wrist roll	40	1	±0.15	0
Wrist pitch, yaw	40	1	0	0

For support condition c, the normalized policy output $\boldsymbol{u}_t$ is converted to desired joint positions as

$$\boldsymbol{q}_{j,t}^{\mathrm{des}} = \boldsymbol{q}_j^0 + \boldsymbol{S}_a^{(c)} \boldsymbol{u}_t, \tag{3.2}$$

where $\boldsymbol{q}_j^0 \in \mathbb{R}^{29}$ is the nominal posture and $\boldsymbol{S}_a^{(c)} \in \mathbb{R}^{29\times 29}$ is a diagonal matrix of joint-wise action scales. A proportional–derivative actuator model tracks these targets. The policy thus acts directly at the joint level, and any footstep, arm motion, or posture change emerges from the learned mapping rather than from a separate planner or whole-body optimizer. Table 3.2 summarizes the robot and platform model.

3.3 Coordinate Frames

We use the inertial world frame ($\mathcal{F}_W$), the platform frame ($\mathcal{F}_P$), the robot-base frame ($\mathcal{F}_B$) attached to the pelvis, the left- and right-foot frames ($\mathcal{F}_{F_l}$, $\mathcal{F}_{F_r}$) attached to the ankle-roll links, and the right-hand frame ($\mathcal{F}_{H_r}$) attached to the right palm, all shown in Figure 3.1. The platform origin lies at the geometric center of the cuboid, so the top surface is 0.06 m above the origin. The axes x_P, y_P, and z_P point in the longitudinal, lateral, and normal directions. A world point X is expressed in platform coordinates by

$$\boldsymbol{p}_X^P = \boldsymbol{R}_{WP}^\top (\boldsymbol{p}_X^W - \boldsymbol{p}_P^W), \tag{3.3}$$

where $\boldsymbol{R}_{WP} \in SO(3)$ maps platform-frame vectors to the world frame and $\boldsymbol{p}_P^W$ is the platform origin. We compute floor boundaries, centering, foot slip, and support

Table 3.2: Common robot and platform model.

Property	Implemented value
Robot	Unitree G1, 29 actuated joints
Joint groups	12 leg, 3 waist, 14 arm joints
URDF link mass	33.34 kg
Nominal stance height	base 0.74 m above the platform surface
Platform top surface	3.2×2.0 m
Platform contact	static friction 1.1, dynamic friction 1.0, zero restitution; friction randomized at every reset to $[0.75, 1.15]$ and $[0.65, 1.05]$
Actor sensing	torso IMU and joint states
Control output	29 desired joint-position offsets

geometry in platform coordinates, so that the translation of the vehicle through the world is not mistaken for robot drift on the floor.

3.4 Moving-Platform Environment

The platform is a $3.2 \times 2.0 \times 0.12$ m rigid cuboid whose pose follows the prescribed trajectory exactly, while the robot remains fully dynamic. Reaction forces from the robot therefore do not alter the vehicle motion, which matches a vehicle whose mass is much larger than that of the robot. With platform-frame base position $[x_B^P, y_B^P, z_B^P]^\top$, the robot remains on the usable floor area while

$$|x_B^P| \leq 1.6 - 0.14, \qquad |y_B^P| \leq 1.0 - 0.12. \tag{3.4}$$

This area is shown as the dashed rectangle in Figure 3.1(a). The boundary is checked on the base position, which is defined at every instant, whereas the feet alternate between support and swing during stepping. The margins of 0.14 m and 0.12 m are of the order of the base-to-foot distance in the nominal stance (0.09 m to the toe and 0.12 m to the ankle laterally), so that a robot standing at the limit still has both feet on the platform. During a step, a foot can come closer to the edge; the condition therefore approximates the usable floor area rather than constraining foot placement exactly.

Longitudinal trajectories contain acceleration, holding, braking, and possible re-acceleration phases. Turning is represented by continuous variable-curvature paths and by constant-radius arcs. For curvature κ and speed v, the yaw rate and normal acceleration are

$$\dot{\psi}_P = \kappa v, \qquad a_\mathrm{n} = \kappa v^2. \tag{3.5}$$

Grade motion, used in no-handle training, changes platform height and pitch consistently. For rigid-handle training, an independently sampled road overlay adds heave, roll, and pitch, composed of band-limited stochastic vibration and an occa-

Table 3.3: Scope of the support conditions.

Condition	Intended contacts	Hand mechanics	Physical test
No handle	feet	none	yes
Rigid handle	feet and right hand	platform–palm fixed joint	yes
Suspended handle	feet and right hand	ball-jointed suspended rigid handle	no

sional transient bump. The trajectories represent disturbance classes that occur in public transport rather than one vehicle suspension or route.

3.5 Support Conditions

Each support condition defines its intended contacts and the conditions that terminate a training episode; Table 3.3 summarizes the three conditions. Reward and termination are implemented separately. All three conditions detect a fall from the base height

$$h_B = \boldsymbol{e}_z^\top \left(\boldsymbol{p}_B^W - \boldsymbol{p}_P^W \right), \tag{3.6}$$

the vertical distance of the base above the platform origin (Figure 3.1), and from the base tilt

$$\theta_B = \cos^{-1}\!\left(\boldsymbol{e}_z^\top \boldsymbol{R}_{WB}\, \boldsymbol{e}_z \right), \tag{3.7}$$

the angle between the base z-axis and the world vertical, where $\boldsymbol{e}_z = [0, 0, 1]^\top$ and $\boldsymbol{R}_{WB}$ is the orientation of $\mathcal{F}_B$ in $\mathcal{F}_W$. The nominal base height is 0.80 m, that is, the 0.74 m stance height above the floor surface plus half of the 0.12 m platform thickness. All three conditions also require the base to remain inside the usable floor area of Eq. (3.4).

3.5.1 No-Handle Support

The intended contacts are the two feet,

$$\mathcal{C}_{\mathrm{NH}} = \{F_l, F_r\}. \tag{3.8}$$

Either foot may leave the floor during recovery, and the policy chooses step timing and foot placement itself. An episode therefore terminates only when the robot falls, detected through h_B or θ_B, or when the base leaves the usable floor area.

3.5.2 Rigid-Handle Support

The rigid condition adds the right hand,

$$\mathcal{C}_{\mathrm{RH}} = \{F_l, F_r, H_r\}. \tag{3.9}$$

A six-DoF fixed joint connects the palm to a platform-fixed frame and remains active throughout the episode. The rigid joint models a firm grasp; finger compliance and grasp loss are outside the model.

At every reset, the right-arm joints are perturbed around the support-ready posture of Table 3.1, and the grasp anchor is placed at the resulting palm pose, so that every episode starts without constraint violation. The perturbation ranges are $[-0.18, 0.135]$ rad for shoulder pitch, $[-0.09, 0.1125]$ rad for shoulder roll, ± 0.1125 rad for shoulder yaw, ± 0.18 rad for the elbow, and ± 0.0525 rad for each wrist joint. Each episode then starts with an idle period of 1.5 s before the platform moves.
An ideal constraint allows a policy to stay upright by hanging from the handle, which a physical grasp would not sustain. We therefore require the rigid-handle policy to keep the feet as the primary support, to bound the handle load, and to return toward the support-ready stance after a disturbance. Besides the fall checks, the rigid-handle condition therefore bounds the grasp wrench, forbids sustained contact of non-foot links, and limits the departure from the support-ready posture. The grasp wrench consists of the force $\boldsymbol{f}_H$ and moment $\boldsymbol{\tau}_H$ that the fixed joint transmits to the palm, expressed in $\mathcal{F}_P$; the force is normalized by the robot weight $B = mg = 327$ N, where m is the total link mass of Table 3.2. A non-foot contact is a contact force on any link other than the ankle links. The departure from the support-ready posture is measured by the root-mean-square joint deviation

$$e_q = \sqrt{\tfrac{1}{29} \left\| \boldsymbol{q}_j - \boldsymbol{q}_j^0 \right\|_2^2} \tag{3.10}$$

and by a stance score that compares the foot placement with that of the support-ready posture. Let $\Delta \boldsymbol{p}_i \in \mathbb{R}^2$ be the error of the horizontal position of foot $i \in \{l, r\}$ relative to the base, expressed in $\mathcal{F}_P$, with respect to its value in the support-ready posture, and let ρ be the lateral distance between the left and right foot divided by its support-ready value. The stance score is

$$s^{\mathrm{st}} = \exp\!\left(-\frac{(e^{\mathrm{st}})^2}{\sigma_{\mathrm{st}}^2} \right) \mathcal{H}(\rho; \rho_0, \rho_1), \qquad e^{\mathrm{st}} = \sqrt{\tfrac{1}{4} \textstyle\sum_i \|\Delta \boldsymbol{p}_i\|_2^2}, \tag{3.11}$$

where

$$\mathcal{H}(x; a, b) = \chi^2(3 - 2\chi), \qquad \chi = \mathrm{clip}\!\left(\frac{x - a}{b - a}, 0, 1 \right) \tag{3.12}$$

is a smooth ramp from zero at a to one at b. With the width $\sigma_{\mathrm{st}} = 0.10$ m and the separation limits $\rho_0 = 0.2$ and $\rho_1 = 0.8$, the score equals one in the support-ready stance, approaches zero as the feet move away from it, and vanishes when the feet come closer than 20% of their nominal separation or cross. Nominal standing and recovery are distinguished by a recovery memory $M_t \in [0, 1]$, a decaying measure of recent disturbance that the rigid-handle reward also uses; nominal standing corresponds to $M_t \leq 0.18$. The posture check tolerates a large departure for a short time during nominal standing and for a longer time during recovery, so that the posture can leave the support-ready configuration temporarily but cannot remain in a strongly deformed stance.

3.5.3 Suspended-Handle Support

The suspended handle consists of a 0.25 m rigid suspension link above a rigid triangular handle of height 0.15 m and width 0.20 m. The moving handle body has a

mass of 0.35 kg, and the grip center lies 0.40 m below the ceiling pivot. A ball joint at the pivot gives the assembly three rotational degrees of freedom: two rotations produce pendulum-like swing of the grip center, and rotation about the suspension axis changes the handle orientation. The anchor body is fixed to the moving platform, and a six-DoF fixed joint connects the center of the horizontal grip to the right palm.
The model therefore represents a support whose direction changes as the handle swings. Because the link is rigid, it can transmit compression and has no slack state; it does not reproduce the tension-only behavior of a strap in Eq. (2.6). We use it to compare fixed with pivoting support geometry, and we do not draw conclusions about strap compliance from it.
Besides the fall checks, the suspended-handle condition terminates an episode on sustained torso tilt θ_T, defined as in Eq. (3.7) for the torso link; on sustained heading drift, that is, a deviation of the base yaw or of either foot yaw from its nominal value relative to the platform; on sustained single support, in which exactly one foot carries a contact force above 8 N; and on excessive hand force. The hand force is estimated from the whole-body translational dynamics,

$$\widehat{F}_h = \left\| m\left(\dot{\boldsymbol{v}}_{\mathrm{CoM}} - \boldsymbol{g}\right) - \sum_{i\in\{l,r\}} \boldsymbol{f}_{F_i} \right\|_2 , \tag{3.13}$$

where m is the robot mass, $\boldsymbol{v}_{\mathrm{CoM}}$ is the center-of-mass velocity, $\boldsymbol{g}$ is the gravity vector, and $\boldsymbol{f}_{F_i}$ is the contact force at foot i. The acceleration is obtained by finite differences at 50 Hz, and the estimate is low-pass filtered with a time constant of 0.08 s. The sustained-force limit depends on a disturbance gate $g_t^{\mathrm{dist}} \in [0, 1]$, which equals one while the platform moves. While the platform is idle or nearly steady (speed below 0.08 m/s and angular speed below 0.08 rad/s), g_t^{dist} rises linearly from zero to one as the magnitude of the measured torso acceleration increases from 0.9 to 1.8 m/s^2.
Table 3.4 lists the thresholds of all termination checks. A check with a grace period acts only after that time has elapsed since the start of the episode, and a check with a duration requires the condition to hold continuously for that time. The thresholds are training safeguards, not certified physical limits of a handle.

3.6 Balance and Recovery Task

For condition $c \in \{\mathrm{NH}, \mathrm{RH}, \mathrm{SH}\}$, the controller is a causal mapping

$$\boldsymbol{u}_t = \pi_c(\boldsymbol{o}_{0:t}), \tag{3.14}$$

where $\boldsymbol{o}_{0:t}$ denotes the history of onboard measurements up to step t. An episode is successful when it reaches its time limit without a termination while the robot retains the contacts required by its condition and remains on the usable floor area. Temporary tilt, high angular velocity, and single support are admissible when the robot subsequently returns to a sustainable stance.
No-handle step recovery follows the qualitative sequence

$$\text{instability} \rightarrow \text{single support} \rightarrow \text{touchdown} \rightarrow \text{stable double support}. \tag{3.15}$$

Table 3.4: Termination thresholds of the three support conditions. Grace periods t_g are measured from the start of the episode; durations T are the times for which a condition must hold continuously.

Quantity	No handle	Rigid handle	Suspended handle
Base height h_B	< 0.64 m	< 0.56 m; $t_g = 0.8$ s	< 0.56 m; $t_g = 0.8$ s
Base tilt θ_B	> 1.0 rad	> 1.1 rad; $t_g = 0.6$ s	> 0.80 rad; $t_g = 0.6$ s
Floor area	Eq. (3.4)	Eq. (3.4)	Eq. (3.4)
Non-foot contact force	–	> 50 N; $T = 0.15$ s	–
Hand force	–	$\|\boldsymbol{f}_H\|_2 > 0.23\,B$; $T = 0.20$ s; $t_g = 0.5$ s	$\widehat{F}_h > 125 + 75 g_t^{\text{dist}}$ N, $T = 0.50 + 0.30 g_t^{\text{dist}}$ s; or $\widehat{F}_h > 280$ N, $T = 0.15$ s; $t_g = 0.5$ s
Hand moment	–	$\|\boldsymbol{\tau}_H\|_2 > 10$ N m; $T = 0.20$ s; $t_g = 0.5$ s	–
Posture	–	$e_q > 0.75$ rad or $s^{\text{st}} < 0.04$; $T = 0.5$ s if $M_t \leq 0.18$, else $T = 2.0$ s	–
Torso tilt θ_T	–	–	> 0.70 rad; $T = 0.75$ s; $t_g = 0.8$ s
Heading drift	–	–	base > 0.70 rad, foot > 0.60 rad; $T = 0.75$ s; $t_g = 0.8$ s
Single support	–	–	$T = 0.75$ s; $t_g = 0.8$ s

This sequence motivates event-gated rewards: lifting a foot is useful only when instability creates a need to step, and touchdown is useful only when it restores stable support. Rigid-handle recovery instead coordinates hand loading with continued foot support.

3.7 Assumptions and Proposed Workflow

We make the following modeling assumptions. The platform motion is prescribed and unaffected by the robot. The scene contains no seats, passengers, or doors. Each episode starts with the robot standing and, in the hand-supported conditions, with the grasp already established. The rigid handle is stiffer than a physical arm–hand–rail connection, and contact and actuator models are approximations of the hardware. We test the consequences of these assumptions by evaluating the policies in a second simulator and, for the no-handle and rigid-handle conditions, in a moving vehicle.

Figure 3.2 summarizes the workflow from the prescribed vehicle motion and finite-platform task to policy learning and evaluation. The suspended-handle policy is evaluated only in Isaac Lab, whereas the MuJoCo and moving-vehicle evaluations use the no-handle and rigid-handle policies. The manufacturer's standing controller appears only as the physical reference.

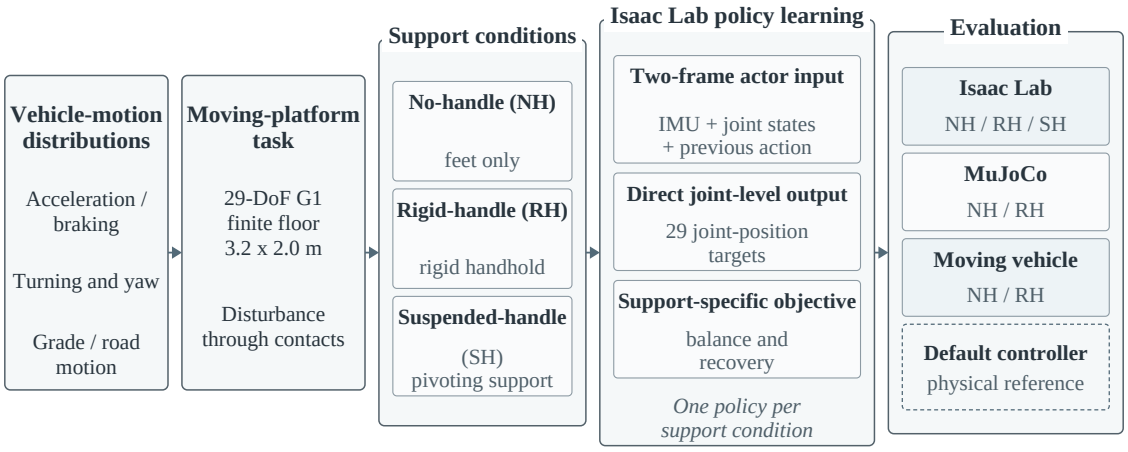


Figure 3.2: Workflow from the moving-platform task to simulation and physical evaluation.

4

Reinforcement Learning Framework

Chapter 3 defined the robot, the moving platform, the support conditions, and the balance and recovery task. This chapter describes how we turn that task into learned feedback policies. We specify the training motions, the observation and action interface, and the support-specific rewards for the no-handle, rigid-handle, and suspended-handle conditions. General RL concepts were introduced in Chapter 2.

4.1 Framework Overview

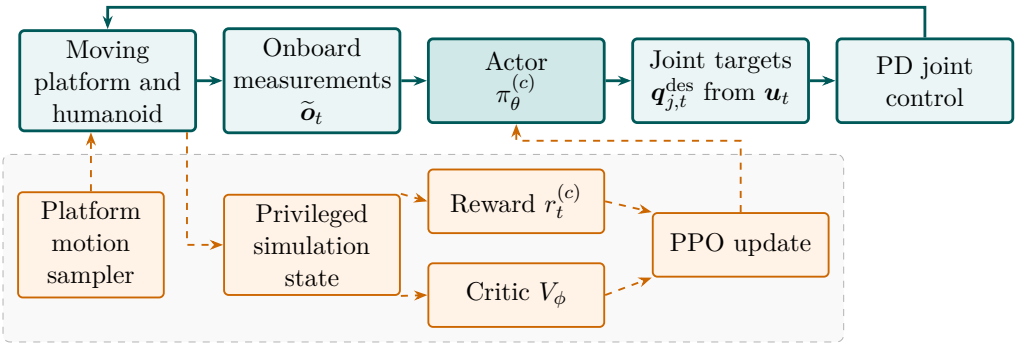


Figure 4.1: Overview of the learning framework. The loop in the upper row is the deployed controller, which runs identically in simulation and on the robot. The components in the shaded box are used only during training in simulation: the critic receives the actor observations together with privileged simulation state, and PPO updates both the actor and the critic.

We train a separate policy $\pi_\theta^{(c)}$ for each support condition $c \in \{\text{NH}, \text{RH}, \text{SH}\}$. All policies share one deployable interface: the actor maps a short history of onboard inertial and joint measurements to 29 desired joint positions at 50 Hz, and a critic with additional simulator information is used only during training. What differs between the conditions is the contact mechanics and, consequently, the reward. Figure 4.1 summarizes the framework, which has three components:

1. a distribution of platform motions that exposes the policy to the longitudinal, turning, grade, and road-induced disturbances of public transport;
2. a policy interface restricted to measurements available on the physical robot; and

3. support-specific rewards that encode when the robot should hold its stance, when it may step, and how it may use a handle.

4.2 Training Disturbance Distributions

We sample complete platform trajectories at episode reset. The families expose the policies to longitudinal acceleration and braking, turning, grades, and vertical road inputs. We use reported transport motions to motivate the lower part of the disturbance ranges and extend selected ranges to examine recovery under stronger inputs. Table 4.1 lists the families and their parameter ranges.

The no-handle policy is trained on four families sampled with equal probability: straight motion with variable acceleration and braking, variable-curvature (Lissajous) turning, straight motion over a grade, and Lissajous turning over a grade. The two grade families additionally carry a small vertical micro-bump component. The rigid-handle policy is trained on straight, Lissajous, and constant-radius arc motion with probabilities 0.5, 0.4, and 0.1, respectively, and an independently sampled road overlay superimposes heave, roll, and pitch on every family. The overlay combines band-limited continuous vibration with an occasional transient bump. The suspended-handle policy is trained on straight and Lissajous motion with equal probability on a flat floor; its turning is considerably milder than that of the rigid-handle policy, and it sees neither constant-radius arcs nor the road overlay. Figure 4.2 sketches the families and their sampling probabilities. The rigid support permits the training curriculum to explore stronger longitudinal inputs than the no-handle condition. Within each straight-line range, the peak acceleration is not sampled uniformly: most no-handle trajectories lie between 0.4 and 0.8 m/s^2, and 90% of the rigid-handle trajectories between 1.5 and 3.0 m/s^2, with the remainder in the upper tail up to 4.0 m/s^2.

The measured bus values give context for the longitudinal training ranges: the first row reports route maxima, and the second summarizes peaks across braking events in a different dataset. Keller et al. also constructed representative acceleration and braking pulses from city-bus measurements [5]. We chose the no-handle straight-line range of 0.30–1.20 m/s^2 to cover lower-intensity motions in this context. For rigid-handle training, we progressively raised the upper bound to 4.0 m/s^2 to examine recovery under stronger longitudinal inputs.

The turning, grade, and road families broaden the direction and timing of the disturbance. De Graaf and van Weperen reported lateral acceleration up to about 4 m/s^2 during bus cornering and movements into or out of stops [35]. We set the rigid-handle turning cap at 4.5 m/s^2 to include somewhat stronger lateral disturbances. Light-rail track guidance permits sustained main-line grades of 4% without a length limit and 6% over a limited length [36]. The no-handle 3–6% grade range covers these inclines. The rigid-handle road overlay adds vertical, roll, and pitch disturbances; published road-hump work motivates the transient vertical component [37]. The actor is not given the sampled family or its trajectory parameters; it must respond to their consequences in the onboard measurements.

Table 4.1: Moving-platform disturbance parameters used for training. a_x and a_y denote longitudinal and lateral platform acceleration, v path speed, $v_{\max}$ the speed cap of a straight trajectory, R the turning radius, and A_z and a_z the vertical displacement and acceleration of the road overlay.

Condition	Motion family	Parameter range or literature context
No-handle	Straight line	$\lvert a_x \rvert = 0.30$–1.20 m/s^2; $v_{\max} = 1.0$–3.6 m/s
	Lissajous turning	$v = 1.2$–2.5 m/s; $R = 4.5$–14 m
	Straight line with grade	$v = 0.8$–1.8 m/s; grade 3–6%
	Lissajous with grade	$v = 1.1$–2.2 m/s; $R = 4.5$–14 m; grade 3–6%
Rigid-handle	Straight line	$\lvert a_x \rvert = 1.50$–4.00 m/s^2; $v_{\max} = 8$–14 m/s
	Lissajous turning	$v = 3.6$–7.0 m/s; $R = 4.5$–14 m; $\lvert a_y \rvert \leq 4.5$ m/s^2
	Constant-radius arc	$v = 10$–30 km/h; $R = 6$–20 m; $\lvert a_y \rvert \leq 4.5$ m/s^2
	Road overlay (all families)	$A_z \leq 60$ mm; $\lvert a_z \rvert = 0.5$–4.5 m/s^2; roll 0.2–1.8°; pitch 0.2–2.5°
Suspended-handle	Straight line	$\lvert a_x \rvert = 0.75$–3.00 m/s^2; $v_{\max} = 9$–15 m/s
	Lissajous turning	$v = 1.35$–2.10 m/s; $R = 8$–20 m; $\lvert a_y \rvert \leq 0.75$ m/s^2
Measured context	Bus acceleration and braking	Route maxima of 1.33 and 1.82 m/s^2, respectively [34]
	Bus braking events	99% had peak deceleration below 3 m/s^2 in the reported dataset [6]
	Bus cornering	Lateral acceleration reached about 4 m/s^2 in the sampled bus journeys [35]
Design guidance	Light-rail main-line grade	Recommended maximum sustained grades: 4% without a length limit; 6% for up to 750 m [36]

4.3 Policy Interface

4.3.1 Actor Observations

Because a deployed robot cannot measure the future motion of the vehicle, we restrict the actor to quantities that the physical robot measures itself. One observation frame is

$$\boldsymbol{o}_t = \begin{bmatrix} \boldsymbol{\omega}_{B,t}^\top & \boldsymbol{\alpha}_{\mathrm{imu},t}^\top & \boldsymbol{g}_{B,t}^\top & \Delta\boldsymbol{q}_{j,t}^\top & \dot{\boldsymbol{q}}_{j,t}^\top & \boldsymbol{u}_{t-1}^\top \end{bmatrix}^\top \in \mathbb{R}^{96}, \tag{4.1}$$

where $\boldsymbol{\omega}_{B,t} \in \mathbb{R}^3$ is the base angular velocity, $\boldsymbol{\alpha}_{\mathrm{imu},t} \in \mathbb{R}^3$ is the torso-IMU linear acceleration, $\boldsymbol{g}_{B,t} \in \mathbb{R}^3$ is the gravity direction projected into the base frame, $\Delta\boldsymbol{q}_{j,t} = \boldsymbol{q}_{j,t} - \boldsymbol{q}_j^0 \in \mathbb{R}^{29}$ and $\dot{\boldsymbol{q}}_{j,t} \in \mathbb{R}^{29}$ are the joint offsets and velocities, and $\boldsymbol{u}_{t-1} \in \mathbb{R}^{29}$ is the previous action. The previous action tells the policy which target the actuators are currently tracking. The IMU acceleration is the most direct onboard trace of platform motion: together with the projected gravity, it lets the policy infer the inertial load imposed by the vehicle from the robot's own response. Table 4.2 gives

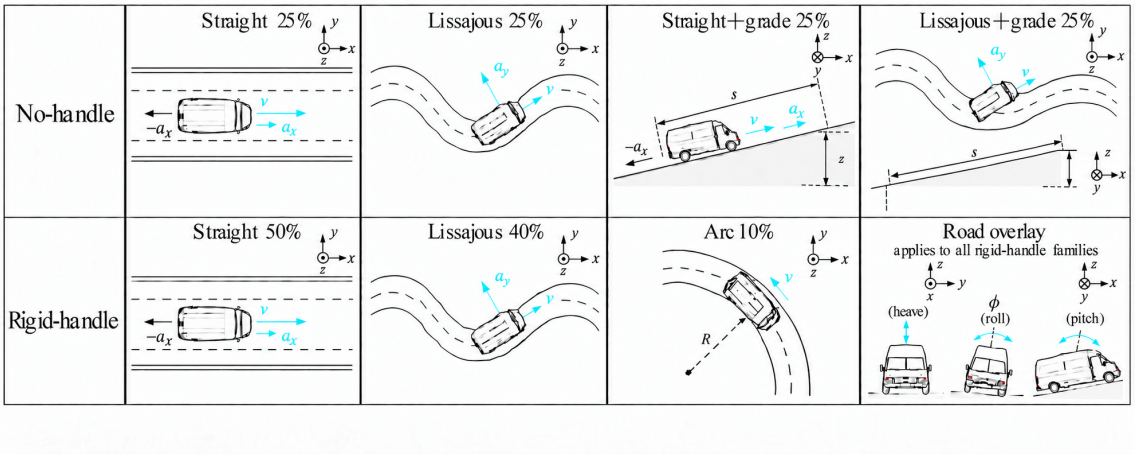


Figure 4.2: Platform-motion families used for training the no-handle and rigid-handle policies. Each sketch shows the qualitative speed, path, or height profile of a family, and the percentage gives the probability with which it is sampled at an episode reset. The road overlay is sampled independently and superimposed on every rigid-handle family. The sketches use platform-fixed axes (x, y, z) (longitudinal, lateral, vertical), path speed v, acceleration components (a_x, a_y), path distance s, turn radius R, roll ϕ, and pitch θ.

Table 4.2: Single-frame actor observation shared by all policies. Noise is uniform and applied during training only.

Quantity	Dimension	Scale	Training noise
Base angular velocity	3	0.2	$[-0.05, 0.05]$
IMU linear acceleration	3	0.1	$[-0.15, 0.15]$
Projected gravity	3	1.0	$[-0.02, 0.02]$
Relative joint position	29	1.0	$[-0.01, 0.01]$
Joint velocity	29	0.05	$[-0.2, 0.2]$
Previous action	29	1.0	none

the scaling and the training noise of each signal.
A single sample does not reveal how the disturbance is evolving, so the actor receives the current and the preceding frame,

$$\tilde{\boldsymbol{o}}_t = \begin{bmatrix} \boldsymbol{o}_{t-1}^\top & \boldsymbol{o}_t^\top \end{bmatrix}^\top \in \mathbb{R}^{192}. \tag{4.2}$$

The two frames are 20 ms apart and give the actor finite-difference information on the measured signals.

4.3.2 Asymmetric Critic

The critic receives the uncorrupted actor quantities together with signals that are available only in simulation. For all conditions, these include the base linear velocity, which gives 198 critic inputs over two frames. The rigid-handle critic additionally receives the six-axis grasp wrench in the platform frame and two foot-support quantities, giving 214 inputs. These signals let the critic distinguish postures that look

alike but distribute the load differently between the feet and the hand, which is central to the rigid-handle objective. Table 4.3 summarizes the interfaces of the three policies.

Table 4.3: Policy observation interfaces.

Condition	Actor	Critic	Training-only critic information
No-handle	192	198	Base linear velocity
Rigid-handle	192	214	Base linear velocity, six-axis grasp wrench, and two foot-support quantities
Suspended-handle	192	198	Base linear velocity

4.3.3 Joint-Position Actions

The actor outputs $\boldsymbol{u}_t \in \mathbb{R}^{29}$, which is converted to joint-position targets using the interface defined in Eq. (3.2). The diagonal action-scale matrix $\boldsymbol{S}_a^{(c)}$ differs by support condition. With the joints grouped as legs and waist (15), left arm (7), and right arm (7), its structure is

$$\begin{aligned} \boldsymbol{S}_a^{(\mathrm{NH})} &= s_{\mathrm{NH}} \boldsymbol{I}_{29}, \\ \boldsymbol{S}_a^{(\mathrm{RH})} &= \mathrm{diag}(s_{\mathrm{RH},L} \boldsymbol{I}_{15}, s_{\mathrm{RH},A} \boldsymbol{I}_7, s_{\mathrm{RH},G} \boldsymbol{I}_7), \\ \boldsymbol{S}_a^{(\mathrm{SH})} &= \mathrm{diag}(s_{\mathrm{SH},L} \boldsymbol{I}_{15}, s_{\mathrm{SH},A} \boldsymbol{I}_7, s_{\mathrm{SH},G} \boldsymbol{I}_7), \end{aligned} \tag{4.3}$$

with the values in Table 4.4. The smaller right-arm scales in the hand-supported conditions reduce commanded motion at the grasp while still allowing the policy to adjust body posture, foot loading, and stepping through the same action interface.

4.3.4 Network Architecture

We train the three policies with PPO. The actor and critic are separate multilayer perceptrons, each with hidden layers of 512, 256, and 128 exponential linear units. As shown in Figure 4.3, the actor receives the 192-component two-frame observation history and outputs the mean of a 29-dimensional diagonal Gaussian action distribution. An action is sampled from this distribution during training and converted to joint-position targets by Eq. (3.2). The critic receives the corresponding training-only observation: 198 components for no-handle and suspended-handle, or 214 for rigid-handle (Table 4.3). It outputs one estimate of the expected discounted return. The extra critic signals do not enter the actor.

4.4 Reward Structure

The three objectives share survival, base-stability, contact, and motion regularization terms, but differ where the available support changes the desired response. No-handle recovery must permit a necessary step without rewarding arbitrary foot

Table 4.4: Joint-position action scales in radians per unit policy output. Each value sets the joint-target change produced by one unit of action.

Condition	Legs and waist	Left arm	Right arm
No-handle	$s_{\mathrm{NH}} = 0.18$	$s_{\mathrm{NH}} = 0.18$	$s_{\mathrm{NH}} = 0.18$
Rigid-handle	$s_{\mathrm{RH},L} = 0.07$	$s_{\mathrm{RH},A} = 0.04$	$s_{\mathrm{RH},G} = 0.015$
Suspended-handle	$s_{\mathrm{SH},L} = 0.07$	$s_{\mathrm{SH},A} = 0.04$	$s_{\mathrm{SH},G} = 0.0225$

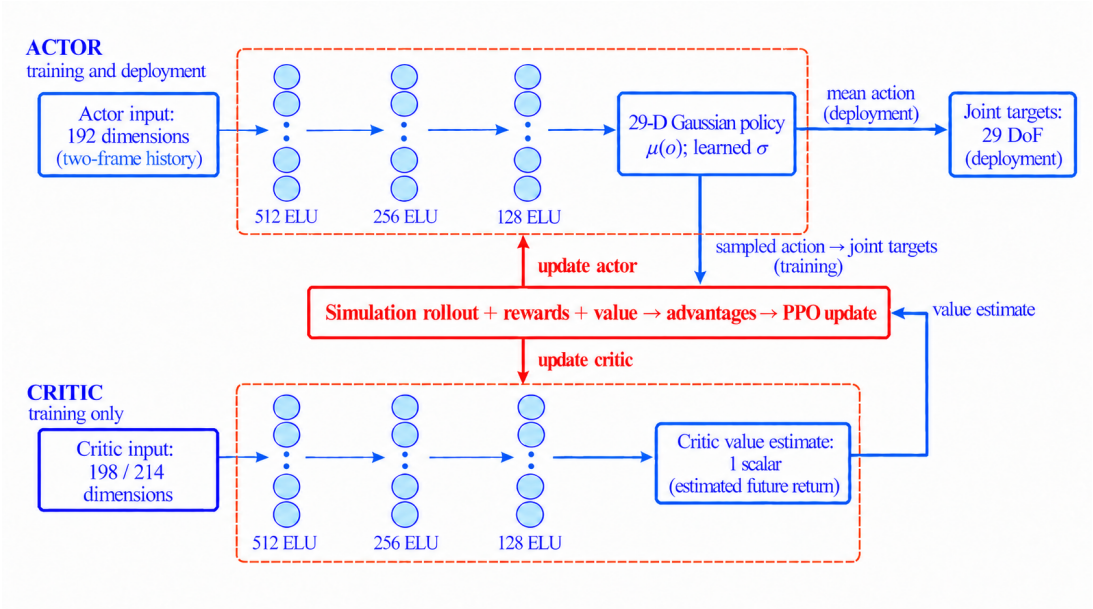


Figure 4.3: Actor–critic structure used for PPO training. The red dashed outlines group the trainable networks, while the red arrows show their parameter updates. Blue arrows show the training data flow and the mean-action path used after deployment. The critic and PPO update are used only during training.

lifting. Rigid-handle recovery can use the grasp while keeping the feet as the primary support. The suspended-handle objective must also account for motion of the grasp point about its pivot.

For condition c, the per-step reward is a weighted sum of scalar terms,

$$r_t^{(c)} = \Delta t \sum_{i \in \mathcal{R}^{(c)}} w_i^{(c)} r_{i,t}^{(c)}, \tag{4.4}$$

where $\mathcal{R}^{(c)}$ is the set of reward terms, $r_{i,t}^{(c)}$ is the term value, $w_i^{(c)}$ its signed weight, and $\Delta t = 0.02$ s is the control interval. Positive weights reward survival, posture or support quality, and recovery; negative weights penalize errors. We normalize force terms by B where indicated. The sum is not clipped. Several terms use the smooth ramp $\mathcal{H}(x; a, b)$ defined in Eq. (3.12). For $a < b$, it increases from zero to one over $[a, b]$, allowing a penalty to set in gradually. We also use $a > b$ for decreasing gates: the same function then rises from zero to one as the measured quantity falls.

Several recovery terms depend on recent events. Their memories have the form $M_t = \max(\lambda M_{t-1}, x_t)$, with $0 < \lambda < 1$. For a memory that only decays, the half-life is $\Delta t \ln(0.5)/\ln(\lambda)$. These memories are internal to reward calculation. The

Table 4.5: Functional comparison of the three policy objectives. Values in parentheses are signed weights $w_i^{(c)}$ in Eq. (4.4).

Function	No-handle (NH)	Rigid-handle (RH)	Suspended-handle (SH)
Posture and stability	Base orientation (−0.25) and group-wise joint-pose costs	Recovery-gated READY joint, stance, and orientation (4, 2, 1.5)	READY joint, stance, and orientation (6, 6, 8), without recovery relaxation
Foot support	Recovery-aware airborne-foot cost (−0.008) and slip (−0.12)	Feet-primary reward (3) and foot-yaw alignment (−0.50)	Upright double support (4), feet-primary support (6), and foot yaw (−0.40)
Recovery	Stable recovery (0.6) and completed recovery step (0.9)	Directional step (1.2), foot-pose repair (0.8), and READY return (1.0)	Through posture and support terms; no event-gated step reward
Hand support	None	Vertical and horizontal force and moment (−4, −1.5, −1), with near-limit costs (−2, −2)	Excess estimated hand loading (−2.5), hand torque (−1), and pivot alignment (−2.5)
Floor use	Platform centering (−1.2) and low base height (−1.2)	Low base height (−1.2)	Low base height (−1.2)
Regularization	Joint velocity and acceleration, action increment, power proxy, joint limits, undesired contact	Joint velocity and acceleration, action increment, power proxy, joint limits, undesired contact, and left-arm quietness	Joint velocity and acceleration, action increment, power proxy, joint limits, undesired contact

deployed actor receives no mode command or privileged signal from them. We adjusted the reward weights and gates after observing the policy at each stage of the disturbance curriculum. We checked whether it stood quietly, stepped when necessary, kept the feet loaded during handle-supported recovery, and returned to a stable stance. Since the terms differ in scale and in when they become active, we judged their balance from these behaviors rather than from the weight values alone. Table 4.5 summarizes the objectives of the three conditions. We organize the comparison by function because the terms use different quantities, normalizations, and activation gates.

Episodes end through the termination checks of Section 3.5 and Table 3.4, or at the time limit. We do not add a terminal penalty to the reward.

4.5 No-Handle Balance and Recovery

Without a handle, one action space must serve two regimes with opposing requirements. During quiet standing, both feet should stay loaded and the body should move little. Once the current contacts can no longer arrest the motion, insisting on double support would prevent the step that restores balance. We therefore combine continuous stabilization terms, which are always active, with event-gated recovery terms, which reward a step only when it follows detected instability and ends in stable support. As the training disturbances increased, we adjusted the standing costs and recovery rewards together. The policy needed to remain quiet under mild motion without being discouraged from stepping when the current support could no longer arrest a stronger disturbance.

4.5.1 Stabilization and Floor Use

Base stabilization is expressed by three penalties: the tilt cost $\|\boldsymbol{g}_{B,xy}\|_2^2$, where $\boldsymbol{g}_{B,xy}$ contains the two horizontal components of projected gravity; the vertical base velocity $v_{B,z}^2$; and the roll–pitch rate $\|\boldsymbol{\omega}_{B,xy}\|_2^2$. Deviation from the nominal posture is penalized joint group by joint group,

$$c_{\mathcal{J}}^{q}(t) = \sum_{j\in\mathcal{J}} \left| q_{j,t} - q_j^0 \right|, \tag{4.5}$$

with the largest absolute weight on the ankles (Table 4.7). We denote the bilateral arm joints (shoulder pitch, roll, and yaw; elbow; and wrist pitch, roll, and yaw) by $\mathcal{J}_{\text{arm}}$. Its shoulder and elbow subsets are $\mathcal{J}_{\text{sh}}$ and $\mathcal{J}_{\text{el}}$. The other groups are the three waist joints $\mathcal{J}_{\text{waist}}$, bilateral hip roll and yaw $\mathcal{J}_{\text{hip,ry}}$, bilateral hip pitch $\mathcal{J}_{\text{hip,p}}$, bilateral knees $\mathcal{J}_{\text{knee}}$, and bilateral ankle pitch and roll $\mathcal{J}_{\text{ankle}}$. The shoulder and elbow costs supplement the whole-arm cost; their overlap is intentional.
The common motion costs are defined on the 29 actuated joints $\mathcal{J}_{\text{act}}$ and on the policy action vector:

$$\begin{aligned} c_t^{\dot{q}} &= \sum_{j\in\mathcal{J}_{\text{act}}} \dot{q}_{j,t}^2, & c_t^{\ddot{q}} &= \sum_{j\in\mathcal{J}_{\text{act}}} \ddot{q}_{j,t}^2, \\ c_t^{\Delta u} &= \|\boldsymbol{u}_t - \boldsymbol{u}_{t-1}\|_2^2, & c_t^{\text{lim}} &= \sum_{j\in\mathcal{J}_{\text{act}}} \left[(q_{j,-}^{\text{soft}} - q_{j,t})_+ + (q_{j,t} - q_{j,+}^{\text{soft}})_+ \right], \end{aligned} \tag{4.6}$$

where $(x)_+ = \max(x, 0)$ and $q_{j,-}^{\text{soft}}$ and $q_{j,+}^{\text{soft}}$ are the soft joint-position bounds. The limit cost increases linearly once a joint passes either bound. The mechanical-power proxy is

$$c_t^{E} = \sum_{j\in\mathcal{J}_{\text{act}}} |\dot{q}_{j,t}|\,|\tau_{j,t}|, \tag{4.7}$$

where $\tau_{j,t}$ is the applied joint torque. The product increases when a joint moves rapidly under a large torque. The condition-specific weights of these motion costs appear in the three reward inventories.
Foot slip is penalized using ankle-link motion relative to the moving floor. The monitored set $\mathcal{K}_{\text{ankle}}$ consists of the left and right ankle-pitch and ankle-roll links. For each $k \in \mathcal{K}_{\text{ankle}}$, we evaluate the platform rigid-body velocity field at the link position, $\boldsymbol{v}_{P,k}^W = \boldsymbol{v}_P^W + \boldsymbol{\omega}_P^W \times (\boldsymbol{p}_k^W - \boldsymbol{p}_P^W)$, and

$$c^{\text{slip}}(t) = \sum_{k\in\mathcal{K}_{\text{ankle}}} \mathbb{1}\left[\|\boldsymbol{f}_k^W\|_2 > F_{\text{load}}\right] \left\|\boldsymbol{v}_{k,xy}^W - \boldsymbol{v}_{P,k,xy}^W\right\|_2^2, \tag{4.8}$$

where $\boldsymbol{p}_k^W$ and $\boldsymbol{v}_k^W$ are the link position and velocity in the world frame, and $\boldsymbol{p}_P^W$, $\boldsymbol{v}_P^W$, and $\boldsymbol{\omega}_P^W$ are the platform position, linear velocity, and angular velocity. Subscript xy selects the world-frame horizontal components. The force $\boldsymbol{f}_k^W$ is the contact force on link k, F_{load} is the loading threshold of Table 4.6, and $\mathbb{1}[\cdot]$ is the indicator function. Using ankle-link velocity penalizes motion of loaded distal links relative to the platform. Contacts on other monitored links contribute to the undesired-contact penalty.

Table 4.6: No-handle event and floor-use parameters. The projected-gravity thresholds are dimensionless; angular-rate thresholds are in rad/s.

Parameter	Value	Role and basis
Instability $(g_{\mathrm{u}}, \omega_{\mathrm{u}})$	$(0.18, 0.35$ rad/s$)$	Tilt gate is approximately 10°; starts recovery memory
Recovered state $(g_{\mathrm{s}}, \omega_{\mathrm{s}})$	$(0.10, 0.18$ rad/s$)$	Lower gates require the motion to settle before reward
Recovery and step decay $(\lambda_{\mathrm{rec}}, \lambda_{\mathrm{step}})$	$(0.96, 0.94)$ per step	Preserve instability and step events for about 0.34 and 0.22 s half-life
Step-memory gate $m_{\min}$; loaded-foot force F_{load}	0.20; 5 N	Require a recent prior event and contact above a small fraction of body weight
Centering safe half-widths $(x_{\mathrm{safe}}, y_{\mathrm{safe}})$	$(0.40, 0.25)$ m	One quarter of platform half-dimensions
Centering scales (L_x, L_y)	$(1.20, 0.75)$ m	Distance from safe-zone edge to platform edge
Soft base-height threshold h_{soft}	0.74 m	Begins the cost 0.06 m below the nominal $h_B = 0.80$ m

Two terms keep the robot usefully placed on the finite floor. With the platform-frame base position (x_t^P, y_t^P), the centering penalty

$$c^{\mathrm{ctr}}(t) = \left(\frac{\max(|x_t^P| - x_{\mathrm{safe}}, 0)}{L_x}\right)^2 + \left(\frac{\max(|y_t^P| - y_{\mathrm{safe}}, 0)}{L_y}\right)^2 \tag{4.9}$$

is zero inside the central region. Table 4.6 gives its safe-zone half-widths and normalization lengths. This term leaves small step-induced displacements unpenalized but discourages using up the remaining floor. The height cost is $c^{\mathrm{h}}(t) = [\max(0, h_{\mathrm{soft}} - h_B)]^2$, with the base height h_B of Eq. (3.6). Its soft threshold lies below the nominal base height, so a small crouch is permitted while a larger drop is penalized. The safe-zone widths follow from the platform dimensions in Table 3.2: they are 25% of its half-length and half-width, while L_x, L_y are the remaining distances to the edges. The foot-loading threshold is small compared with the robot weight B and separates loaded from unloaded monitored links.

4.5.2 Event-Gated Step Recovery

Recovery is detected from the base tilt $g_t = \|\boldsymbol{g}_{B,xy,t}\|_2$ and the roll–pitch rate $\omega_t = \|\boldsymbol{\omega}_{B,xy,t}\|_2$. We define instability and recovered stability by

$$I_t^{\mathrm{u}} = \mathbb{1}\left[g_t > g_{\mathrm{u}} \ \vee \ \omega_t > \omega_{\mathrm{u}}\right], \qquad I_t^{\mathrm{s}} = \mathbb{1}\left[g_t < g_{\mathrm{s}} \ \wedge \ \omega_t < \omega_{\mathrm{s}}\right], \tag{4.10}$$

with angular rates in rad/s. The distinct onset and recovery thresholds prevent a state close to the instability threshold from immediately counting as recovered. Their values are listed in Table 4.6.

Two positive terms reward recovery. The first rewards returning to a stable state shortly after an instability,

$$r_t^{\mathrm{rec}} = I_t^{\mathrm{s}}\,\bar{U}_t, \qquad \bar{U}_t = \max(\lambda_{\mathrm{rec}}\bar{U}_{t-1}, I_t^{\mathrm{u}}), \tag{4.11}$$

where the memory $\bar{U}_t$ decays at the rate listed in Table 4.6. This term remains positive on successive stable control steps while the memory decays. The second term rewards a completed recovery step, that is, the sequence instability → single support → touchdown → stable double support of Eq. (3.15). With $n_t \in \{0, 1, 2\}$ the number of feet carrying more than F_{load}, and with the events $e_t^- = \mathbb{1}[n_{t-1} = 2 \wedge n_t = 1]$ (liftoff) and $e_t^+ = \mathbb{1}[n_{t-1} = 1 \wedge n_t = 2]$ (touchdown), three memories are updated in sequence,

$$\begin{aligned} U_t &= \max\left(\lambda_{\mathrm{step}}U_{t-1},\, I_t^{\mathrm{u}}\right), \\ L_t &= \max\left(\lambda_{\mathrm{step}}L_{t-1},\, e_t^-\,\mathbb{1}[U_t > m_{\min}]\right), \\ T_t &= \max\left(\lambda_{\mathrm{step}}T_{t-1},\, e_t^+\,\mathbb{1}[L_t > m_{\min}]\right), \end{aligned} \tag{4.12}$$

and the step reward is

$$r_t^{\mathrm{step}} = I_t^{\mathrm{s}}\,\mathbb{1}[n_t = 2]\,U_t\,T_t. \tag{4.13}$$

The step memory decays at the rate listed in Table 4.6. Its gate $m_{\min}$ (Table 4.6) admits a liftoff or touchdown only while the preceding event remains recent. A foot lifted without prior instability therefore earns no step reward. Once stable double support is restored, the product in Eq. (4.13) can be positive for more than one control step as the memories decay. All memories start at zero, with $n_{-1} = 2$.
A third term shapes foot contact directly. The airborne penalty equals 2 when neither foot is loaded, 1 when one foot is loaded while $I_t^{\mathrm{u}} = 0$, and zero otherwise. Single support thus costs nothing while the robot is unstable, whereas unnecessary foot lifting during quiet standing and loss of both contacts remain penalized. The resulting no-handle objective is

$$\bar{r}_t^{\mathrm{NH}} = r_{\mathrm{aux},t}^{\mathrm{NH}} + w_{\mathrm{rec}}^{\mathrm{NH}} r_t^{\mathrm{rec}} + w_{\mathrm{step}}^{\mathrm{NH}} r_t^{\mathrm{step}} + w_{\mathrm{air}}^{\mathrm{NH}} c_t^{\mathrm{air}} + w_{\mathrm{ctr}}^{\mathrm{NH}} c^{\mathrm{ctr}}(t) + w_{\mathrm{h}}^{\mathrm{NH}} c^{\mathrm{h}}(t), \tag{4.14}$$

where $r_{\mathrm{aux},t}^{\mathrm{NH}}$ collects the remaining stabilization and regularization terms in Table 4.7, and $r_t^{(\mathrm{NH})} = \Delta t\,\bar{r}_t^{\mathrm{NH}}$. Figure 4.4 illustrates the event logic. The weights of these and the remaining terms are listed in Table 4.7.

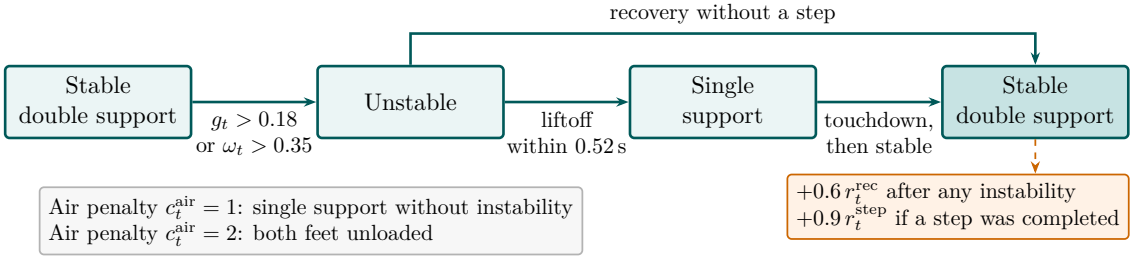


Figure 4.4: Event-gated no-handle recovery terms. Recent instability gates a stable-state term and, following liftoff and touchdown, a step-recovery term. The decay of the memories allows either term to contribute on successive eligible control steps.

Table 4.7: Complete no-handle reward inventory. The w_i values are signed weights. Shoulder and elbow pose costs supplement the whole-arm cost; the waist row combines two equal contributions.

Term	Weight w_i	Quantity and activation
Alive	1.5	Constant while the episode remains active.
Base vertical velocity	-0.6	Squared world-frame vertical base velocity.
Base roll–pitch rate	-0.12	Squared horizontal components of base angular velocity.
Joint velocity	-3×10^{-4}	Squared norm over actuated joints.
Joint acceleration	-5×10^{-8}	Squared norm over actuated joints.
Action increment	-0.005	Squared difference between consecutive actions.
Joint-limit violation	-1.0	Soft joint-limit cost of Eq. (4.6).
Mechanical-power proxy	-2×10^{-5}	$\sum_j \lvert\dot{q}_j\rvert\lvert\tau_j\rvert$.
Arm pose ($\mathcal{J}_{\text{arm}}$)	-0.010	L1 nominal-pose error over both arms.
Shoulder pose ($\mathcal{J}_{\text{sh}}$)	-0.020	Additional L1 nominal-pose error over both shoulders.
Elbow pose ($\mathcal{J}_{\text{el}}$)	-0.012	Additional L1 nominal-pose error over both elbows.
Waist pose ($\mathcal{J}_{\text{waist}}$)	-0.040	Combined weight on the L1 nominal-pose error of the three waist joints.
Hip roll–yaw pose ($\mathcal{J}_{\text{hip,ry}}$)	-0.018	L1 nominal-pose error.
Hip pitch pose ($\mathcal{J}_{\text{hip,p}}$)	-0.030	L1 nominal-pose error.
Knee pose ($\mathcal{J}_{\text{knee}}$)	-0.025	L1 nominal-pose error.
Ankle pose ($\mathcal{J}_{\text{ankle}}$)	-0.045	L1 nominal-pose error.
Base orientation	-0.25	Squared horizontal projected-gravity components.
Foot slip	-0.12	Platform-relative ankle-link slip while contact force exceeds F_{load}.
Undesired contact	-0.8	Indicator for contact on monitored links other than the ankles.
Platform centering ($w^{\text{NH}}_{\text{ctr}}$)	-1.2	Squared excess of platform-frame base position outside the safe region of Table 4.6.
Low base height (w^{NH}_{h})	-1.2	One-sided quadratic cost below h_{soft}.
Recovery-aware feet airborne ($w^{\text{NH}}_{\text{air}}$)	-0.008	Cost 2 for no loaded feet; cost 1 for one loaded foot only outside instability.
Recovery from instability ($w^{\text{NH}}_{\text{rec}}$)	0.6	Active on stable steps while recent-instability memory remains nonzero.
Successful recovery step ($w^{\text{NH}}_{\text{step}}$)	0.9	Requires instability followed by double–single–double support and stable touchdown; may persist while memories decay.

4.6 Rigid-Handle Balance and Recovery

The rigid grasp provides an additional reaction force while the feet remain the primary support. During nominal motion, the robot should hold a repeatable support-ready stance with most weight on the feet. When a disturbance requires correction, the objective relaxes posture tracking and permits foot unloading, a step, and a bounded hand reaction. Once the disturbance subsides, it rewards a return to the support-ready stance. Figure 4.5 shows this progression before we define its terms.

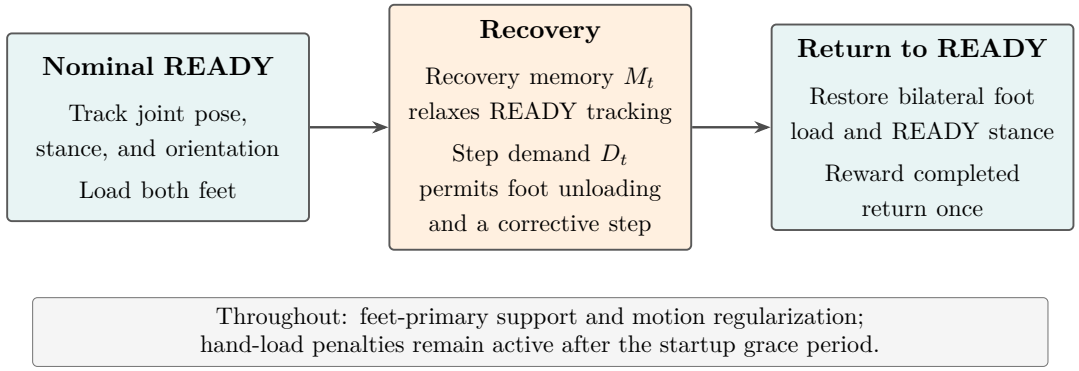


Figure 4.5: Rigid-handle reward hierarchy. Nominal READY tracking and bilateral foot loading give way to relaxed tracking and an optional corrective step when recovery demand rises. The policy then restores foot support and READY posture. Feet-primary support remains rewarded throughout, while hand-load penalties remain active after their startup grace period.

4.6.1 Support-Ready Reference

We call the nominal support-ready configuration READY. It comprises the nominal whole-body joint pose, the foot positions relative to the base, and the base orientation relative to the platform, and it serves only as a reference for the reward; the actor receives no mode or command. We score joint pose using the RMS deviation $e_{q,t}$ defined in Eq. (3.10):

$$s_{q,t} = \exp\left[-\left(\frac{e_{q,t}}{\sigma_q}\right)^2\right]. \tag{4.15}$$

For foot placement, we use the stance score s_t^{st} defined in Eq. (3.11). It compares the foot positions with their support-ready references and suppresses a narrow or crossed stance. The orientation score is $s_{R,t} = \exp[-(\theta_{BP,t}/\sigma_R)^2]$, where $\theta_{BP,t}$ is the angle of the base orientation relative to the platform, including yaw. Table 4.8 gives the joint and orientation score widths and the recovery parameters. We selected the widths while checking that small balance corrections retained posture guidance without suppressing a needed step.

4.6.2 Recovery Demand and Relaxation

To relax posture tracking only when it is needed, we measure how strongly the current state calls for recovery. The instantaneous demand $d_t \in [0, 1]$ is the maximum of smooth gates on the nine signals of Table 4.9, which cover hand load, foot unloading, base motion, stance deformation, slip, and foot yaw. Taking the maximum lets one sufficiently severe signal open the recovery window on its own. Writing $z_{k,t}$ for a listed signal and (a_k, b_k) for its deadband and full-demand values, we define

$$d_t = \max_{k=1,\dots,9} \mathcal{H}(z_{k,t}; a_k, b_k). \tag{4.16}$$

For lighter-foot load and stance score, $a_k > b_k$, so a falling value increases demand. The slip signal is the maximum platform-relative horizontal speed of the two loaded

Table 4.8: Rigid-handle READY, recovery-memory, and foot-support parameters.

Parameter	Value	Role and basis
Joint and orientation widths (σ_q, σ_R)	$(0.30\ \text{rad}, 0.28\ \text{rad})$	Errors at these widths reduce the respective exponential scores to e^{-1}
Recovery-memory decay λ_M	0.99 per step	Gives a 1.4 s half-life so tracking remains relaxed after a transient
READY score floors (f_q, f_f, f_R)	$(0.15, 0.20, 0.15)$	Keeps posture shaping active during recovery
Joint $(D_{q,0}, D_{q,1})$ and stance $(D_{f,0}, D_{f,1})$ step-gate intervals	$(0.20, 0.65)$; $(0.25, 0.70)$	Relaxes posture earlier when a step is needed
Total-foot-load ramp $(F_{\Sigma,0}, F_{\Sigma,1})$	$(0.55B, 0.95B)$	Reaches full score when the feet carry nearly all body weight
Lighter-foot-load ramp $(F_{\min,0}, F_{\min,1})$	$(0.08B, 0.45B)$	Rewards load on both feet during nominal standing
Memory reduction κ_M	0.85	Weakens the lighter-foot requirement as recovery memory rises
Step-gate interval $(D_{Q,0}, D_{Q,1})$	$(0.15, 0.65)$	Opens foot unloading as early-step demand rises
Lighter-foot multiplier floor and increment $(\mu_{\min}, \Delta\mu)$	$(0.05, 0.65)$	Gives $\mu_t = 0.70$ in nominal standing and a nonzero floor during a step

ankle-roll links; a link contributes only above its 10 N contact threshold. It is distinct from the four-link squared slip penalty of Eq. (4.8). The demand is retained in the recovery memory

$$M_t = \max(\lambda_M M_{t-1}, d_t), \tag{4.17}$$

whose decay keeps posture tracking relaxed after the demand has subsided. During the initial 1.5 s stationary preparation phase, the memory is held at zero and READY tracking remains fully active.

A second quantity, the early-step demand $D_t \in [0, 1]$, measures whether a step is required now. It combines the platform-frame offset of the base from the center of support, the stance score, the lateral foot separation, the base tilt, and loaded-foot slip (Table 4.10). Let G_t be the larger of the support-center-offset and decreasing stance score gates, Θ_t^{early} the early-tilt gate, Σ_t^{slip} the loaded-slip gate, Θ_t the ordinary tilt gate, and Θ_t^{sev} the severe-tilt gate. The resulting demand is

$$\begin{aligned} D_t = \max\Big\{ &\Theta_t^{\text{sev}},\ \alpha_G G_t,\ \sqrt{G_t \Theta_t^{\text{early}}},\\ &\sqrt{\Sigma_t^{\text{slip}} \Theta_t^{\text{early}}},\ \alpha_T \Theta_t \Big\}, \end{aligned} \tag{4.18}$$

clipped to $[0, 1]$ and held at zero during the same 1.5 s preparation phase. The coefficients α_G and α_T are listed in Table 4.10. Geometry alone can therefore request

Table 4.9: Signals forming the instantaneous rigid-handle recovery demand d_t. Each gate rises from zero at the deadband to one at full demand; for the lighter-foot load and the stance score, demand rises as the quantity falls.

Signal	Deadband	Full demand
Hand moment	4.5 N m	8.0 N m
Total hand force	$0.11B$	$0.18B$
Horizontal hand force	$0.08B$	$0.16B$
Lighter-foot load	$0.28B$	$0.12B$
Projected-gravity tilt	0.10	0.24
Roll–pitch angular rate	0.25 rad/s	0.60 rad/s
READY stance score	0.86	0.55
Loaded-foot slip	0.05 m/s	0.22 m/s
Mean foot-yaw error	10°	22°

a modest step, while tilt combined with poor geometry or slip raises demand more rapidly. Lateral foot separation enters the stance score rather than a separate gate. The three READY scores are relaxed as

$$\begin{aligned} \widetilde{s}_{X,t} &= [f_X + (1 - f_X)(1 - M_t)(1 - \delta_{X,t})]\, s_{X,t}, \quad X \in \{q, R\}, \\ \widetilde{s}^{\mathrm{st}}_t &= [f_f + (1 - f_f)(1 - M_t)(1 - \delta_{f,t})]\, s^{\mathrm{st}}_t, \end{aligned} \tag{4.19}$$

with score floors and step-gate intervals from Table 4.8. In particular, $\delta_{q,t} = \mathcal{H}(D_t; D_{q,0}, D_{q,1})$, $\delta_{f,t} = \mathcal{H}(D_t; D_{f,0}, D_{f,1})$, and $\delta_{R,t} = 0$, where the two pairs of D thresholds are the joint and stance intervals in that table. The memory M_t relaxes tracking for the whole recovery, whereas the step gates relax the joint and stance scores immediately when a step becomes urgent. Because the floors are positive, READY tracking is weakened but never removed.

4.6.3 Foot Support and Handle Load

The feet should remain the primary support. We rotate the four monitored ankle-link contact forces into the platform frame and sum their positive platform-vertical components into left and right foot loads. With $F_{\Sigma,t}$ their sum and $F_{\min,t}$ the lighter-foot load, the support reward is

$$r^{\mathrm{sup}}_t = \mathcal{H}(F_{\Sigma,t}; F_{\Sigma,0}, F_{\Sigma,1})\, [(1 - \mu_t) + \mu_t \mathcal{H}(F_{\min,t}; F_{\min,0}, F_{\min,1})]\,, \tag{4.20}$$

where the force thresholds are given in Table 4.8. With the READY keep factor $K_t = 1 - \kappa_M M_t$ and the step gate $Q_t = \mathcal{H}(D_t; D_{Q,0}, D_{Q,1})$, the lighter-foot multiplier is $\mu_t = (1 - Q_t)(\mu_{\min} + \Delta\mu K_t) + \mu_{\min} Q_t$. Table 4.8 gives these parameters. The multiplier is $\mu_{\min} + \Delta\mu$ in nominal standing. It decreases as recovery memory rises and reaches $\mu_{\min}$ when the step gate fully opens. The first factor requires the feet to carry most of the body weight; the second favors an even load distribution during quiet standing but allows one foot to unload for a step. A narrow or crossed stance is discouraged by the lateral separation factor in the separate READY stance score, not by an additional flatness factor in this support reward.

Hand load is penalized through the grasp wrench, expressed at the grasp point in the platform frame. The vertical force, horizontal force, and moment enter smooth

ramps, and two further terms penalize the approach to the termination limits (Table 4.11). Because each ramp has a nonzero deadband, ordinary stabilizing reactions at the hand cost nothing and only excessive load is suppressed. These penalties remain active during recovery. With the grasp force $\boldsymbol{f}_H$ and moment $\boldsymbol{\tau}_H$ of Section 3.5, the five ramp inputs are, in order, $|f_{H,z}|/B$, $\|\boldsymbol{f}_{H,xy}\|_2/B$, $\|\boldsymbol{\tau}_H\|_2$, $\|\boldsymbol{f}_H\|_2/B$, and $\|\boldsymbol{\tau}_H\|_2$ again. Their intervals, startup grace period, and weights are listed in Table 4.11. The near-limit ramps reach full cost before the termination thresholds in Table 3.4, giving the policy a rising cost before either episode check is reached.
The remaining terms complete the objective. The vertical velocity and the roll–pitch rate of the base are penalized relative to the platform, so that platform yaw during a turn is not treated as robot instability; foot slip and low base height use the no-handle forms. The free left arm is encouraged to stay quiet during nominal standing through smooth costs on its RMS pose error and joint speed. Both costs are relaxed by a disturbance gate driven by the IMU linear acceleration and angular speed, so that the left arm can move freely to regulate momentum during a disturbance.

4.6.4 Recovery Step and Return to READY

When the relaxed posture terms alone cannot restore support, the robot must step. Three terms reward the phases of a step, and Table 4.10 lists their thresholds. The directional step term rewards liftoff when the early-step demand is high, rewards each new maximum of swing-foot progress along a reference direction, and rewards a loaded touchdown in proportion to the progress achieved and the resulting improvement of the support center. The reference direction is computed inside the reward from the current state: under moderate demand it blends a capture-like direction with motion toward the READY foot position, and under severe demand it points toward a location that rebuilds support beneath the base. The actor receives no foothold or direction; the reference only shapes which of its own steps are rewarded. Conversely, keeping both feet loaded while the early-step demand is high is penalized. For an active swing of foot k, let $p_{k,t}$ be its displacement from liftoff projected onto the reference direction, and let $P_{k,t} = \mathcal{H}(p_{k,t}; p_0, p_1)$. The within-swing shaping term is $c_{\text{prog}}[P_{k,t} - \max_{s<t} P_{k,s}]_+$: passing the same position repeatedly earns no further credit. Liftoff receives $c_{\text{lift}} D_t$ only when one foot unloads with $D_t \geq D_{\text{lift}}$ and $M_t \geq M_{\text{lift}}$ after the cooldown. A touchdown counts only after at least $t_{\text{swing,min}}$ of swing and a renewed load of $F_{\text{touch}} B$; its score is $P_{k,t}(c_{\text{td},0} + c_{\text{td,imp}} \eta_t^{\text{imp}})$, where $\eta_t^{\text{imp}} = \text{clip}[(e_{\text{lift}} - e_t)/\Delta e_{\text{imp}}, 0, 1]$ measures improvement in the base-to-support-center offset. Swing tracking expires after $t_{\text{swing,max}}$ and then enters a t_{cool} cooldown. The parameter values are collected in Table 4.10. Separate stance and foot-yaw terms encourage a well-oriented foot after touchdown.
After a slip or a step, the foot-pose repair term rewards progress toward the READY foot pose. With Φ_t the product of the stance score and a platform-relative foot-yaw score,

$$r_t^{\text{repair}} = \mathbb{1}[M_t \geq M_{\text{repair}}]\,\mathbb{1}[\Phi_t < \Phi_{\text{repair}}]\,\zeta_t\,\text{clip}\left(\frac{\Phi_t - \Phi_{t-1}}{\Delta\Phi_{\text{scale}}}, -1, 1\right), \tag{4.21}$$

where the repair parameters are listed in Table 4.10 and $\zeta_t \in [0, 1]$ decreases while the demand for a support-rebuilding step is high, so that pose repair does not compete

with an urgent step. The signed increment rewards improvement and penalizes deterioration. Finally, a one-time bonus rewards a completed return to READY. It is armed once the recovery memory exceeds M_{arm} and paid when the memory has decayed below M_{pay}, the READY scores are high, and both feet carry load. A displaced but momentarily upright stance therefore does not count as a completed recovery.

Table 4.10: Parameters of the rigid-handle step and return terms.

Quantity	Value
Early-step demand D_t (smooth gates, maximum taken)	
Geometry and ordinary-tilt factors (α_G, α_T)	$(0.55, 0.35)$; bound demand from either signal alone
Base offset from support center	0.035–0.10 m; demand rises as the base moves away from support
READY stance score (decreasing)	0.88–0.55
Lateral separation ratio (decreasing)	0.80–0.50
Base tilt	0.085–0.22; early 0.055–0.16; severe 0.16–0.28
Loaded-foot slip	0.05–0.22 m/s
Directional step term	
Liftoff gates $(D_{\text{lift}}, M_{\text{lift}})$	$(0.15, 0.10)$; require demand and recent recovery memory
Liftoff factor c_{lift}	0.45; scales the liftoff reward with demand
Progress factor c_{prog}; interval (p_0, p_1)	0.40; $(0.02, 0.12)$ m; reward new swing maxima only
Minimum swing $t_{\text{swing,min}}$; touchdown load fraction F_{touch}	0.06 s; 0.15; reject an immediate contact change as a completed step
Touchdown factors $(c_{\text{td},0}, c_{\text{td,imp}})$; improvement scale Δe_{imp}	$(0.60, 0.40)$; 0.05 m; combine progress with support improvement
Swing timeout $t_{\text{swing,max}}$; cooldown t_{cool}	0.90 s; 0.35 s; limit each rewarded swing
Double-support penalty	both feet $\geq 0.16B$, active over $D_t = 0.30$–0.75
Foot-pose repair	
Repair gates $(M_{\text{repair}}, \Phi_{\text{repair}})$; increment scale $\Delta\Phi_{\text{scale}}$	$(0.20, 0.92)$; 0.025; reward pose improvement during incomplete recovery
Return to READY	
Armed; paid $(M_{\text{arm}}, M_{\text{pay}})$	$(0.45, 0.12)$; separate activation from completion
Score thresholds	joint 0.82, stance 0.78, foot yaw 0.80, orientation 0.82
Lighter-foot load	$\geq 0.22B$

Collecting the principal terms, the rigid-handle objective is

$$\begin{aligned}\bar{r}_t^{\text{RH}} = r_{\text{aux},t}^{\text{RH}} + w_q^{\text{RH}}\widetilde{s}_{q,t} + w_f^{\text{RH}}\widetilde{s}_t^{\text{st}} + w_R^{\text{RH}}\widetilde{s}_{R,t} + w_{\text{sup}}^{\text{RH}} r_t^{\text{sup}} - c_t^{\text{hand}} \\ + w_{\text{dir}}^{\text{RH}} r_t^{\text{dir}} + w_{\text{repair}}^{\text{RH}} r_t^{\text{repair}} + w_{\text{ready}}^{\text{RH}} r_t^{\text{ready}},\end{aligned} \tag{4.22}$$

where r_t^{sup} is the support reward of Eq. (4.20), c_t^{hand} is the sum of the five hand-load

costs multiplied by the magnitudes of their negative weights, and r_t^{dir} and r_t^{ready} are the directional step and return terms. The remaining survival, posture, foot-yaw, stability, left-arm, and regularization terms are collected in $r_{\text{aux},t}^{\text{RH}}$. When adjusting these weights across the stronger training stages, we checked both the response to a disturbance and the stance after it. The foot-support reward and hand-load costs were balanced so that the robot could use the hand for recovery without shifting its ordinary standing load away from the feet; the return reward addressed recovery that left it displaced from READY. Table 4.11 gives all their weights.

Table 4.11: Complete rigid-handle reward inventory; w_i denotes the signed weight of each term.

Term	Weight w_i	Quantity and activation
Alive	1.0	Constant while the episode remains active.
READY joint score (w_q^{RH})	4.0	Exponential whole-body RMS joint score with recovery-dependent relaxation; recovery memory and early-step demand are held at zero during the first 1.5 s.
READY stance score (w_f^{RH})	2.0	Foot-position and lateral-separation score with recovery-dependent relaxation.
READY base orientation (w_R^{RH})	1.5	Relative quaternion score with recovery-dependent relaxation.
Feet-primary support ($w_{\text{sup}}^{\text{RH}}$)	3.0	Total and lighter-foot effective load; lighter-foot requirement relaxes during recovery.
Handle vertical force	-4.0	Smooth cost from $0.05B$ to $0.25B$ after 0.5 s.
Handle horizontal force	-1.5	Smooth cost from $0.04B$ to $0.18B$ after 0.5 s.
Handle moment	-1.0	Smooth cost from 1.5 to 7.0 N m after 0.5 s.
Near-limit total force	-2.0	Smooth cost from $0.12B$ to $0.21B$.
Near-limit moment	-2.0	Smooth cost from 6.0 to 9.5 N m.
Hip-roll pose	-0.025	L1 nominal-pose error.
Hip-yaw pose	-0.060	L1 nominal-pose error.
Ankle pose	-0.045	L1 nominal-pose error.
Foot-yaw alignment	-0.50	Smooth cost from 5° to 20°.
Directional recovery step ($w_{\text{dir}}^{\text{RH}}$)	1.2	Liftoff, directional swing progress, and touchdown quality.
Double support under step demand	-0.50	Active as early-step demand rises through 0.30–0.75 while both feet carry at least $0.16B$.
Foot-pose repair progress ($w_{\text{repair}}^{\text{RH}}$)	0.8	Potential difference in stance–yaw score during incomplete recovery.
Return to READY ($w_{\text{ready}}^{\text{RH}}$)	1.0	One-shot event after recovery memory and support-quality conditions are satisfied.
Relative vertical velocity	-0.35	Squared platform-relative vertical base velocity.
Relative roll–pitch rate	-0.10	Squared platform-frame relative angular rate.
Low base height	-1.2	One-sided quadratic platform-relative height cost.
Left-arm pose quietness	-0.6	Smooth RMS pose cost (0.02–0.15 rad), scaled by $1-0.9\nu_t$; the gate ν_t rises with IMU horizontal acceleration (0.10–0.45 m/s^2) and angular speed (0.08–0.30 rad/s), with 0.06 s attack and 0.30 s release.
Left-arm velocity quietness	-0.4	Smooth RMS joint-speed cost (0.05–0.60 rad/s) with the same gate.
Joint velocity	-2×10^{-4}	Squared norm over actuated joints.
Joint acceleration	-5×10^{-8}	Squared norm over actuated joints.
Action increment	-0.003	Consecutive-action rate cost.
Joint-limit violation	-1.0	Soft joint-limit cost of Eq. (4.6).
Mechanical-power proxy	-1×10^{-5}	$\sum_j \lvert\dot{q}_j\rvert\lvert\tau_j\rvert$.
Foot slip	-0.12	Platform-relative ankle-link slip while loaded.
Undesired contact	-0.8	Indicator for contact on monitored non-foot links.

4.7 Suspended-Handle Extension

As a simulation extension, we train a policy for the suspended handle of Section 3.5.3, whose grip swings about a ceiling ball joint. It uses the shared actor interface, the

198-component critic interface, and the suspended-handle action scales of Table 4.4. We train it on the straight-line and Lissajous families of Table 4.1.
The objective follows the feet-primary principle of the rigid-handle design, but keeps full READY tracking during recovery and has no event-gated step reward (Table 4.5). Two terms address the moving grasp point. An alignment penalty grows as the grip swings horizontally away from the pivot, and an estimated hand-load penalty discourages hanging from the handle. The latter uses the filtered resultant force estimate $\widehat{F}_h$ defined in Eq. (3.13). Its penalty interval widens during platform motion, allowing a larger transient stabilizing hand reaction than during nominal standing.
Unlike the rigid-handle orientation score, the suspended-handle READY base score compares the robot and platform up directions and therefore measures tilt without yaw; base yaw has its own anchoring penalty. The suspended feet-primary term uses the total and lighter-foot loads in the same form as Eq. (4.20):

$$r_t^{\mathrm{sup,SH}} = \mathcal{H}(F_{\Sigma,t}; a_{\Sigma,0}, a_{\Sigma,1}) \left[(1-\mu_{\mathrm{SH}}) + \mu_{\mathrm{SH}} \mathcal{H}(F_{\min,t}; a_{\min,0}, a_{\min,1})\right]. \tag{4.23}$$

Here $F_{\Sigma,t}$ and $F_{\min,t}$ are formed by summing the positive world-vertical contact forces on the four monitored ankle links into left and right foot loads. This term retains the bilateral-load preference during disturbance rather than relaxing it with a recovery memory. The separate upright double-support term uses foot forces projected onto the platform normal. Its unweighted score is

$$r_t^{\mathrm{upright}} = \Upsilon_t \ \mathrm{clip}\left(\frac{F_{\Sigma,t}^P}{b_\Sigma B}, 0, 1\right) \ \mathrm{clip}\left(\frac{F_{\min,t}^P}{b_{\min} B}, 0, 1\right) \ e^{-\kappa_{\mathrm{col}} c_t^{\mathrm{column}}}, \tag{4.24}$$

where $\Upsilon_t = \mathrm{clip}[(\cos\theta_{T,t} - \cos\theta_{\mathrm{fail}})/(\cos\theta_{\mathrm{upright}} - \cos\theta_{\mathrm{fail}}), 0, 1]$ for torso tilt $\theta_{T,t}$. The column cost c_t^{column} sums squared horizontal excesses of torso–pelvis and pelvis–foot-center displacement beyond their deadbands, after division by their respective scales. Sagittal and lateral components have different weights, and the total cost is capped. Neither term rewards transferring most body weight to the handle.
Let $g_t^{\mathrm{dyn}} = 1 - g_t^{\mathrm{low}}$ denote the dynamic-motion gate, with $g_t^{\mathrm{low}} = 1$ for an idle or nearly steady platform. The hand-load cost uses the filtered $\widehat{F}_h/B$ in $\mathcal{H}(\widehat{F}_h/B; a_t, b_t)$, where $(a_t, b_t) = (a_{\mathrm{idle}}, b_{\mathrm{idle}}) + g_t^{\mathrm{dyn}}(a_{\mathrm{dyn}} - a_{\mathrm{idle}}, b_{\mathrm{dyn}} - b_{\mathrm{idle}})$. Thus its zero-cost interval is wider under platform motion; the separate incoming-hand-moment cost rises over the interval listed in Table 4.13 after the startup grace period. For grip alignment, let (δ_x, δ_y) be the platform-frame horizontal displacement of the grip point from the ceiling pivot. Its unweighted cost is

$$c_t^{\mathrm{align}} = \min\left\{c_{\max}, \left(\frac{(|\delta_x| - d_x)_+}{\ell_x}\right)^2 + \left(\frac{(|\delta_y| - d_y)_+}{\ell_y}\right)^2\right\}. \tag{4.25}$$

These deadbands permit small grip swings around the ceiling pivot, while the normalization lengths make a larger displacement increasingly costly. The hand-load ramp begins at a higher force fraction during platform motion than during idle standing, preserving a range of transient hand reactions. We adjusted the suspended-handle weights while checking foot loading, grip displacement, and posture during

the staged training motions. The READY and bilateral-support terms remained active during recovery, while the hand-load and alignment costs discouraged hanging on the moving grip. Table 4.12 collects the values used in Eqs. (4.23)–(4.25).

Table 4.12: Suspended-handle foot-support, torso-alignment, and hand-load reward parameters.

Parameter	Value	Role
Foot-load ramps $(a_{\Sigma,0}, a_{\Sigma,1})$; $(a_{\min,0}, a_{\min,1})$	$(0.55B, 0.95B)$; $(0.08B, 0.35B)$	Reward most weight on the feet and load on both feet
Lighter-foot multiplier μ_{SH}	0.70	Retains bilateral-load preference
Upright-support loads $(b_{\Sigma}, b_{\min})$	$(0.82, 0.22)$	Target total and lighter-foot load fractions
Torso angles $(\theta_{\mathrm{upright}}, \theta_{\mathrm{fail}})$; column gain κ_{col}	$(0.08, 0.34)$ rad; 0.70	Full upright score below 0.08 rad, falling to zero at 0.34 rad
Column deadbands; scales	$(0.060, 0.085)$ m; $(0.14, 0.18)$ m	Torso–pelvis and pelvis–foot-center excess
Column sagittal/lateral factors; cap	$(1.4, 0.9)$; 6	Shapes and bounds stacked-body cost
Hand-load ramps $(a_{\mathrm{idle}}, b_{\mathrm{idle}})$; $(a_{\mathrm{dyn}}, b_{\mathrm{dyn}})$	$(0.025, 0.10)$; $(0.08, 0.20)$	Allow a larger hand reaction during platform motion
Alignment deadbands (d_x, d_y)	$(0.05, 0.045)$ m	Permits small grip swings
Alignment scales (ℓ_x, ℓ_y); cap $c_{\max}$	$(0.18, 0.14)$ m; 6	Sets increase and cap of alignment cost

Table 4.13 lists all suspended-handle reward terms with their weights, intervals, and activation rules.

Table 4.13: Complete suspended-handle reward inventory; w_i denotes the signed weight of each term.

Term	Weight w_i	Quantity and activation
Alive	1.0	Constant while the episode remains active.
READY joint pose	6.0	Whole-body READY joint score maintained through recovery.
READY stance geometry	6.0	READY-relative foot-position score without recovery relaxation.
READY base orientation	8.0	Relative base-orientation score without recovery relaxation.
Base-yaw anchoring	-0.25	Platform-relative base-yaw deviation.
Foot-yaw anchoring	-0.40	Platform-relative foot-yaw deviation.
Excess torso tilt	-1.6	Cost beyond the torso-upright region.
Upright double support	4.0	Product of torso-upright, bilateral foot-load, and body-stack scores.
Feet-primary support	6.0	Total and lighter-foot normalized-load quality.
Excess estimated hand loading	-2.5	Motion-dependent hand-load cost with intervals in Table 4.12.
Right-hand torque	-1.0	Smooth incoming-joint-torque cost from 1.5 to 8 N m after 0.5 s.
Grip–pivot alignment	-2.5	Squared sagittal/lateral excess with parameters in Table 4.12.
Relative vertical velocity	-0.35	Squared platform-relative vertical base velocity.
Relative roll–pitch rate	-0.15	Squared platform-frame relative roll–pitch rate.
Relative yaw rate	-0.02	Squared platform-frame relative yaw rate.
Soft base-height limit	-1.2	One-sided quadratic platform-relative height cost.
Joint velocity	-2×10^{-4}	Squared norm over actuated joints.
Joint acceleration	-5×10^{-8}	Squared norm over actuated joints.
Action increment	-0.003	Consecutive-action rate cost.
Joint-limit violation	-1.0	Soft joint-limit cost of Eq. (4.6).
Mechanical-power proxy	-1×10^{-5}	$\sum_j \lvert\dot{q}_j\rvert\lvert\tau_j\rvert$.
Foot slip	-0.12	Platform-relative ankle-link slip while loaded.
Undesired contact	-0.8	Indicator for contact on monitored non-foot links.

5

Implementation and Evaluation Methodology

This chapter describes how the policies are trained, how they are evaluated in a second simulator, and how they are deployed and tested in a moving vehicle. The reward design is given in Chapter 4; here the focus is on the training procedure, the consistency of the policy interface across the three stages, and the measures used to compare the conditions.

5.1 Training

We train the policies with PPO in Isaac Lab on the platform of Section 3.4, with the policy running at the 50 Hz rate used on the robot. Rigid-handle episodes on constant-radius arcs may run for up to 50 s so that the longest arcs can be completed. Table 5.1 lists the settings that differ between the policies, and Table 5.2 lists the settings shared by all of them.

5.1.1 Training Curriculum

So that each policy learns to stand before it faces strong platform motion, we trained it through a sequence of stages of increasing difficulty (Table 5.3), initializing each stage from a checkpoint of the previous one. The sequence first establishes upright standing, then introduces longitudinal acceleration and braking, and only then adds the coupled lateral and yaw motion of turning. The final stage uses the distributions of Section 4.2, including the grade families for the no-handle policy and the road overlay for the rigid-handle policy. Within each stage, the motion distribution is fixed.

5.1.2 Policy Selection and Export

We selected the final policy of each condition by its behavior across the disturbance families, with episode completion as the primary criterion and tilt, unintended contacts, stepping, and support-specific failures as secondary criteria. The selected actor is exported in the Open Neural Network Exchange (ONNX) format together with a configuration file that fixes the interface on the robot: the policy period of 0.02 s, the mapping between the simulation and robot joint orders, the nominal posture, the joint gains, the action scales and offsets, and the scales and two-frame

Table 5.1: PPO settings of the final training stage that differ between the policies.

Setting	No handle	Rigid handle	Suspended handle
Parallel environments	1024	1024	128
Episode length	20 s	26 s; up to 50 s (arc)	26 s
Iterations	25,000	8,000	1,000
Learning epochs	5	3	5
Initial learning rate	10^{-3}	3×10^{-4}	3×10^{-4}
Entropy coefficient	0.010	0.003	0.003
Value-loss coefficient	1.0	0.5	1.0

Table 5.2: Shared PPO and network settings.

Setting	Value
Physics time step	0.005 s
Physics steps per action	4 (policy rate 50 Hz)
Policy steps per rollout	24
Mini-batches	4
Discount factor γ	0.99
Generalized-advantage parameter λ	0.95
Clipping parameter	0.2
Clipped value loss	yes
Learning-rate schedule	adaptive, desired KL divergence 0.01
Maximum gradient norm	1.0
Initial action standard deviation	1.0
Observation normalization	none

Table 5.3: Training stages of increasing platform-motion difficulty.

Stage	Platform motion	Purpose
1	Stationary	Upright standing and nominal posture control
2	Straight, low speed	Longitudinal acceleration and braking at low severity
3	Straight, higher speed	Wider longitudinal speed and inertial-load range
4	Curved, low speed	Lateral acceleration and yaw at limited severity
5	Curved, higher speed	Higher turning speeds before the final distribution

history of every observation term (Table 4.2). The robot-side controller reads this file, so that the same values are used in simulation, in MuJoCo, and on the robot. The critic and all training-only quantities remain in simulation.

5.1.3 Software and Model Versions

Table 5.4 lists the software and hardware used for training, evaluation, and deployment.

Table 5.4: Software and hardware.

Component	Version or model
Isaac Sim	5.1.0
Isaac Lab	2.3.2
RSL-RL	3.1.2
PyTorch	2.7.0 (CUDA 12.8)
Python	3.11.13
Training CPU	AMD Ryzen Threadripper PRO 9995WX (96 cores)
Training GPU	NVIDIA RTX PRO 6000 Blackwell Workstation Edition (96 GB)
Training memory	1 TB RAM
MuJoCo	3.3.6
Unitree SDK (unitree_sdk2)	2.0.2
ONNX Runtime	1.22.0

5.2 Evaluation in Simulation

We evaluate the selected policies in Isaac Lab and, after export, in MuJoCo. In both simulators the actor runs with deterministic actions and without reward access, and the policy parameters and interface remain fixed. Table 5.5 compares the platform motions of these evaluations with the training ranges and with the nominal vehicle tests.

5.2.1 Isaac Lab

The Isaac Lab evaluation disables observation noise and consists of three parts. First, each motion family of Table 4.1 is evaluated with 200 episodes drawn from its training range. Second, a severity sweep raises the peak longitudinal acceleration of the straight-line family in steps of 0.5 m/s^2 from 1.0 to 3.5 m/s^2 for the no-handle policy and in steps of 1.0 m/s^2 from 2.0 to 6.0 m/s^2 for the rigid-handle policy, with 200 episodes per level; the upper levels lie outside the respective training ranges. Third, representative episodes are recorded at the policy rate together with rendered frames to show the recovery motion in detail. The suspended-handle policy is evaluated on its two training families and on one fixed straight-line trajectory.

5.2.2 MuJoCo

MuJoCo implements the robot and contact models independently of Isaac Lab. The exported actors run with the same configuration file as on the robot (Section 5.1.2), through the same deployment controller and state machine. The evaluation has

Table 5.5: Platform motions in training, in the simulation evaluations, and in the vehicle tests. Straight-line motion is characterized by the peak longitudinal acceleration $|a_x|$ and the top speed, turning by the speed v, the radius R, and the lateral acceleration $|a_y|$. Physical values are the nominal targets given to the driver.

Policy	Setting	Straight line	Turning
No handle	Training	$\lvert a_x\rvert = 0.30$–1.20 m/s^2; up to 3.6 m/s	$v = 1.2$–2.5 m/s; $R = 4.5$–14 m
	Isaac Lab and MuJoCo, sweep	$\lvert a_x\rvert = 1.0$–3.5 m/s^2	–
Rigid handle	Training	$\lvert a_x\rvert = 1.5$–4.0 m/s^2; up to 14 m/s	v up to 30 km/h; $R = 4.5$–20 m; $\lvert a_y\rvert \leq 4.5$ m/s^2
	Isaac Lab and MuJoCo, sweep	$\lvert a_x\rvert = 2.0$–6.0 m/s^2	–
Suspended handle	Training	$\lvert a_x\rvert = 0.75$–3.0 m/s^2; up to 15 m/s	$v = 1.35$–2.1 m/s; $R = 8$–20 m
	Isaac Lab, fixed episode	$\lvert a_x\rvert = 3.0$ m/s^2; 10 m/s	–
No and rigid handle	MuJoCo, fixed episodes	$\lvert a_x\rvert = 1.0$ m/s^2; 6.5 m/s	$v = 4.17$ m/s; $R = 20$ m; $\lvert a_y\rvert = 0.87$ m/s^2
Default, no and rigid handle	Vehicle tests	10–45 km/h	$v = 10$–20 km/h; $R = 8$–20 m

two parts. First, the no-handle and rigid-handle policies are each run on two fixed platform trajectories: a straight-line acceleration of 1.0 m/s^2 up to 6.5 m/s, and a turn that reaches 4.17 m/s with a peak yaw rate of 0.208 rad/s. This is a functional check: it tests whether the policies keep their behavior under a different contact model, and whether the interface later used on the robot is correct.
Second, we repeat the straight-line severity sweep of the Isaac Lab evaluation with the same acceleration levels and 30 episodes per level. The platform trajectories come from a port of the same straight-line generator, with the peak acceleration fixed and the other parameters sampled for each episode. Each episode starts with the robot suspended from a virtual elastic band; the state machine moves it to its ready posture and starts the policy, and the band is released before the platform moves. For the rigid-handle policy, the joints are first set to the nominal pose $\boldsymbol{q}_j^0$ and the grasp is attached at the resulting palm position, as at an Isaac Lab reset. The hand-load checks of Table 3.4 are evaluated after each episode instead of ending it, so that every episode also shows whether the robot stays upright.

5.2.3 Metrics

All measures are computed at the policy rate of 50 Hz over the samples $t = 1, \ldots, K$ of an episode, and forces are normalized by the robot weight B of Section 3.5.
Completion. An episode is completed when it reaches its time limit without any of

the terminations of Section 3.5. Of n episodes of one setting, n_c are completed, and we report the completion rate $\hat{p} = n_c/n$ with the 95% Wilson score interval

$$\frac{1}{1+z^2/n}\left(\hat{p}+\frac{z^2}{2n} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}+\frac{z^2}{4n^2}}\right), \qquad z = 1.96, \tag{5.1}$$

which, unlike the normal approximation, stays inside $[0, 1]$ when $\hat{p}$ is close to zero or one.

Attitude. Body attitude is measured by the platform-relative base tilt

$$\alpha_t = \cos^{-1}\left(\boldsymbol{e}_{z,P}^{\top}\boldsymbol{e}_{z,B}\right), \tag{5.2}$$

where $\boldsymbol{e}_{z,P}$ and $\boldsymbol{e}_{z,B}$ are the z-axes of $\mathcal{F}_P$ and $\mathcal{F}_B$ in world coordinates, and we report its episode maximum $\alpha_{\max} = \max_t \alpha_t$.

Floor use. The boundary margin

$$d_t^{\mathrm{b}} = \min\left(\bar{x} - |x_{B,t}^P|,\ \bar{y} - |y_{B,t}^P|\right) \tag{5.3}$$

is the distance between the base and the edge of the usable floor area, where $\bar{x} = 1.6 - 0.14$ m and $\bar{y} = 1.0 - 0.12$ m are the limits of Eq. (3.4), and we report its episode minimum $d_{\min}^{\mathrm{b}} = \min_t d_t^{\mathrm{b}}$.

Corrective steps. Foot $i \in \{l, r\}$ is in contact when the vertical contact force on it exceeds 5 N, $c_{i,t} = \mathbb{1}[f_{z,i,t} > 5\ \mathrm{N}]$. A liftoff of foot i is a sample with $c_{i,t-1} = 1$ and $c_{i,t} = 0$, and a touchdown a sample with $c_{i,t-1} = 0$ and $c_{i,t} = 1$. A touchdown counts as a corrective step when the horizontal position of the foot in $\mathcal{F}_P$ has moved by more than 0.05 m since the preceding liftoff of the same foot, so that brief unloading without displacement is not counted. The number of corrective steps n_s of an episode is the step count of the recovery. We also report the step rate $10\, n_s/T_{\mathrm{ep}}$ in steps per 10 s, where T_{ep} is the episode duration, so that episodes that end early can be compared with completed ones.

Support and hand load. The foot-support fraction

$$\eta = \frac{1}{K}\sum_{t=1}^{K} \mathbb{1}[f_{z,l,t} + f_{z,r,t} \geq 0.5B] \tag{5.4}$$

is the share of time in which the feet together carry at least half of the body weight. For the rigid handle, the hand force $\|\boldsymbol{f}_{H,t}\|_2/B$ and the hand moment $\|\boldsymbol{\tau}_{H,t}\|_2$ of Section 3.5 are summarized by their 95th percentiles over the episode, $F_{H,95}$ and $\tau_{H,95}$; a percentile is used instead of the maximum so that a single sample does not dominate. For the suspended handle, we report the maximum handle swing angle, that is, the angle between the suspension link and $\boldsymbol{e}_{z,P}$, the 95th percentile of the estimated hand force $\widehat{F}_h/B$ of Eq. (3.13), and the 95th percentile of the torque magnitude transmitted through the right wrist joints.

5.3 Evaluation in the Real World

5.3.1 Physical Deployment

The hardware controller is implemented in C++ on the Unitree low-state and low-command channels. ONNX Runtime executes the actor in a dedicated 50 Hz thread,

and a 1 kHz finite-state-machine loop publishes the latest desired joint positions with the exported stiffness and damping. The IMU angular velocity, linear acceleration, and orientation, together with the 29 joint positions and velocities, form the deployed observation. Figure 5.1 summarizes the controller.

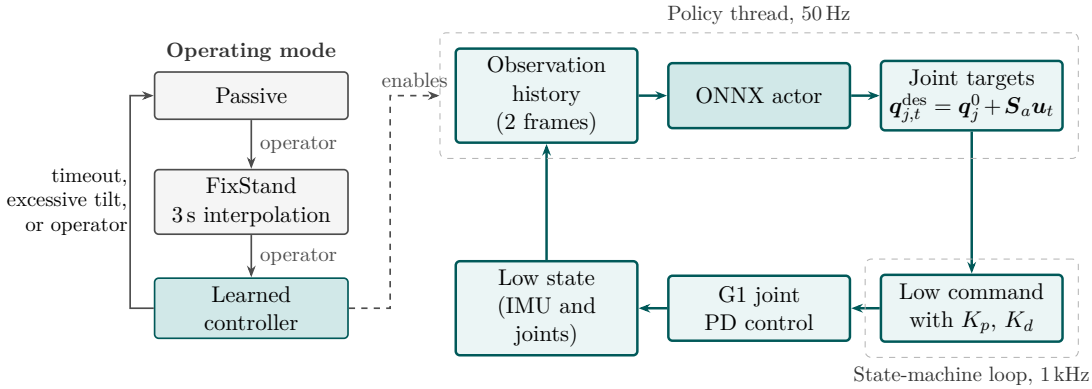


Figure 5.1: Onboard controller on the G1. Left: operating modes of the finite-state machine. Right: the policy thread builds the observation history from the low state, evaluates the actor, and converts its output into joint targets at 50 Hz; the state-machine loop sends the latest targets with the joint gains to the robot at 1 kHz.

The finite-state machine starts in Passive. An operator first enters FixStand, which interpolates to a known posture over 3 s, and then enables the learned moving-platform controller. A low-state communication timeout or excessive orientation returns the controller to Passive, and the operator can request this transition at any time.

5.3.2 Test Site and Vehicle

We conducted the experiments on the skidpad of the Volvo Cars Hällered Proving Ground in south-western Sweden, a broad, level paved area within a closed test facility. Its open layout provided space for approximately constant-radius turns and for successive acceleration and braking maneuvers, and restricted access separated the experiment from public traffic. The vehicle could therefore return to a common start region between runs, and the driver and site operator could stop a maneuver if the robot approached the boundaries of the cargo compartment.
The moving platform was a Chalmers REVERE research van, a first-generation Volkswagen Crafter. The G1 stood inside the cargo compartment on a flat floor installed between the vehicle sides (Figure 5.2). The usable floor between the wheel-arch covers is about 3.0 m long and 1.35 m wide. The central floor area allowed foot repositioning, while the covered wheel arches and the side walls bounded the available lateral space, so the compartment is a finite moving support surface like the simulated platform. An overhead metal frame carried a passive safety tether attached behind the robot. The tether stayed slack during standing and corrective motion and became taut only during a large loss of balance, when it limited the

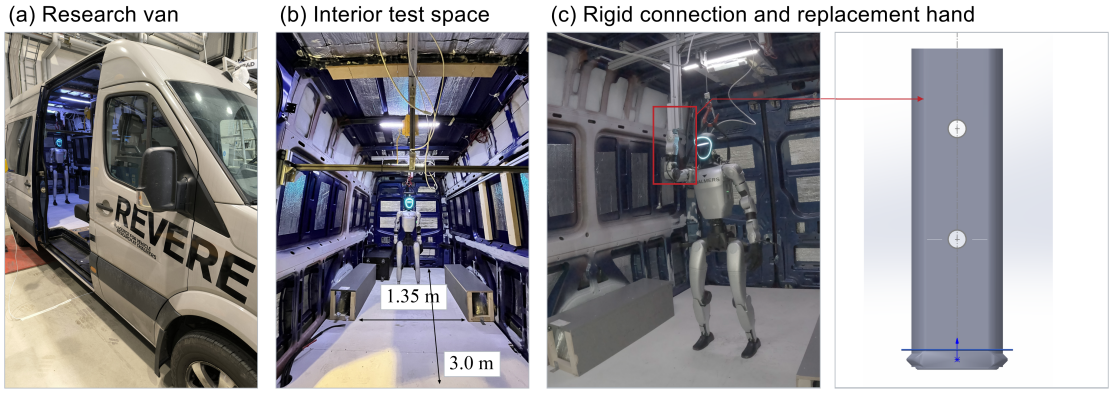


Figure 5.2: Physical test platform. (a) The Chalmers REVERE research van. (b) The Unitree G1 in the prepared cargo compartment, with the approximate length and width of the usable floor. The flat central floor, covered wheel-arch structures, overhead support frame, and rear safety tether define the robot's physical test environment. (c) The rigid-handle installation and a CAD view of the replacement hand component; the red box marks the connection to the vehicle-fixed support.

downward motion and inclination of the robot. It is safety equipment rather than part of the controller, and we treat every observed tether engagement as an external intervention.

5.3.3 Controller–Support Conditions

We evaluated three controller–support conditions on the 29-DoF G1. In the *default* condition, the manufacturer's standing controller balanced the robot without hand contact; it serves as the physical reference. In the *no-handle* condition, the learned no-handle policy balanced the robot on its feet alone, with the space to step and to move its arms freely. In the *rigid-handle* condition, the learned rigid-handle policy balanced the robot while its right hand was fixed to a support attached to the vehicle, aligned with the support-ready posture used in training. Because the rigid-handle condition changes both the controller and the contact set, its comparison with the other two conditions characterizes the complete support strategy rather than the controller alone. The suspended-handle policy was evaluated in simulation only. The two learned policies run with the joint gains of Table 3.1, which are identical for the no-handle and rigid-handle policies. The default condition uses the standing controller supplied with the robot, which runs with its own gains; these are part of the manufacturer's firmware and not accessible to us. Every run began in Passive, proceeded through FixStand, and enabled the selected controller before the vehicle moved.

5.3.4 Instrumentation and Synchronization

Table 5.6 summarizes the measurement system. The robot's onboard stream recorded time-stamped base inertial signals, the orientation quaternion, and the position, ve-

Table 5.6: Instrumentation used in the physical experiments.

Source	Recorded quantities	Role in the analysis
Unitree G1 onboard stream	Base IMU and joint position and velocity signals	Robot attitude and whole-body response
PEAK PCAN GNSS+IMU module	GNSS-derived speed and position, linear acceleration, and angular velocity	Realized maneuver input and trajectory context
Forward-facing GoPro	Road view and vehicle execution	Maneuver and timing verification
Robot-facing GoPro	Body motion, foot and hand contacts, cabin proximity, and visible tether engagement	Recovery behavior and safety-event review

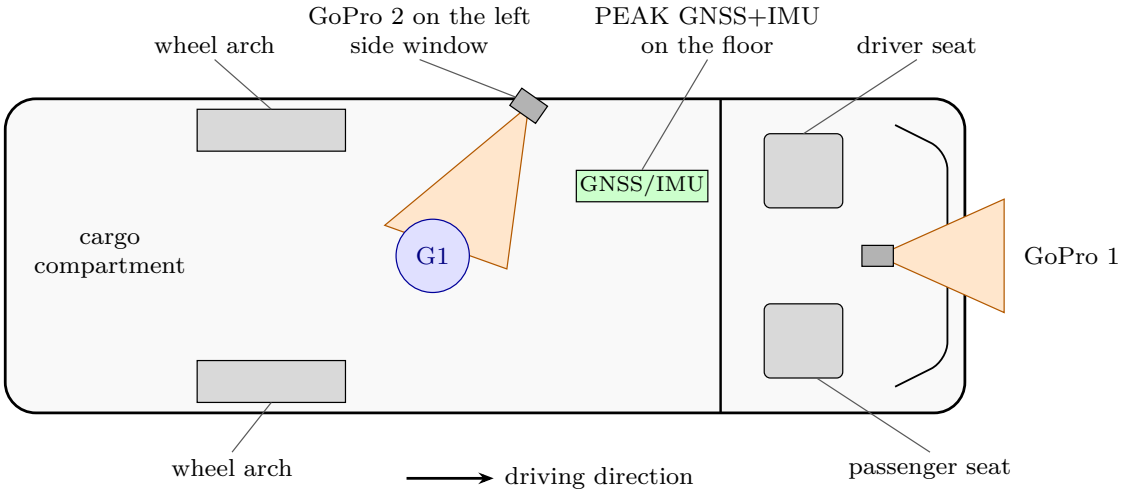


Figure 5.3: Top view of the test vehicle with the measurement equipment (schematic, not to scale). The shaded areas indicate the fields of view of the two cameras.

locity, and estimated torque of all joints. A PEAK PCAN GNSS+IMU module, placed on the floor at the front of the cargo compartment, measured the vehicle motion: GNSS-derived speed and position together with linear acceleration and angular velocity. Its CAN messages were logged on a laptop and decoded with the corresponding CAN database (DBC) file. Two GoPro cameras recorded the test: one between the driver and passenger seats facing forward through the windshield, and one on the left side window of the cargo compartment facing the robot. Figure 5.3 shows where the equipment was mounted. The vehicle channels describe the disturbance transmitted through the floor, the robot stream records the resulting body and joint motion, and the camera views document maneuver execution, stepping, hand contact, cabin proximity, and tether engagement.

Table 5.7: Nominal physical-test matrix. The target values guided the driver; the measured vehicle motion was retained for the quantitative analysis.

Straight-road acceleration and braking					
Pedal instruction / target speed	10 km/h	18 km/h	27 km/h	36 km/h	45 km/h
Light input	S1	S2	S3	S4	S5
Slightly harder input	S6	S7	S8	S9	S10

Constant-radius turning			
Radius / target speed	10 km/h	15 km/h	20 km/h
8 m	T1	T2	T3
10 m	T4	T5	T6
12 m	T7	T8	T9
20 m	T10	T11	T12

5.3.5 Maneuver Matrix and Execution

The test matrix comprises straight-road acceleration–braking tests and constant-radius turning tests (Table 5.7). The straight-road tests expose the robot to longitudinal disturbances of opposite sign during acceleration and braking; the turning tests impose a sustained lateral disturbance whose magnitude follows from the realized speed and radius. For S1–S5, the driver was instructed to apply a light accelerator input, and for S6–S10 a slightly stronger one; each run accelerated toward its target speed and then braked. The turning tests combined target speeds of 10, 15, and 20 km/h with radii of 8, 10, 12, and 20 m and were driven as approximately constant-radius, constant-speed arcs. In addition, T13 probed the rigid-handle condition at a higher turning speed on the 20 m radius, to approach the practical speed limit of that configuration.

Because a human driver executed each run, the target speeds, radii, and pedal instructions define nominal maneuvers rather than reproduced inputs. The S/T identifiers provide the pairing structure across conditions, while the measured vehicle signals characterize the realized acceleration, speed, and turn rate of every run and are shown alongside each robot response. For each controller–support condition, the runs progressed from lower to higher disturbance levels. The driver and site operator stopped a run whenever the robot approached the floor or vehicle boundaries.

Figure 5.4 shows the measured speed of every straight-road run, and Table 5.8 compares the realized turning speed and radius with the targets. The speed is the GNSS speed of the PEAK module, which updates at 1 Hz. The peak speeds of the straight runs lie within 3 km/h of their targets, except for the default S5 run, which stayed 4.4 km/h below. In the turns, the realized radius lies within 1.3 m of the nominal value except in the default T10 and T12 runs, which were 3.8 and 1.8 m wider. The turning speed was mostly 1–3 km/h below the target, by up to 5.3 km/h in the no-handle T3 run, and 3.2 km/h above it in the rigid-handle T10 run.

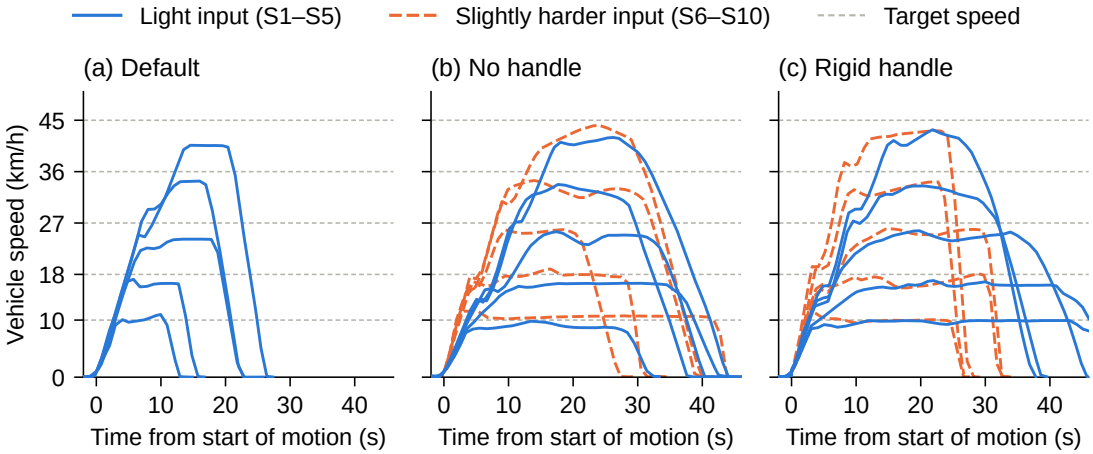


Figure 5.4: Measured vehicle speed of the straight-road runs, aligned at the start of motion. The dashed horizontal lines mark the target speeds of Table 5.7.

Table 5.8: Target and realized turning maneuvers. The realized radius $\hat{R}$ is the median of $v/|\dot{\psi}|$ and the realized speed $\bar{v}$ the mean GNSS speed over the steady part of each run, in which the yaw rate exceeds 80% of its 95th percentile. A dash marks a maneuver that was not driven in that condition.

	Target		Default		No handle		Rigid handle	
Trial	R (m)	v (km/h)	$\hat{R}$ (m)	$\bar{v}$ (km/h)	$\hat{R}$ (m)	$\bar{v}$ (km/h)	$\hat{R}$ (m)	$\bar{v}$ (km/h)
T1	8	10	8.5	11.3	7.1	7.0	7.3	9.2
T2	8	15	–	–	–	–	7.0	12.9
T3	8	20	–	–	6.8	14.7	6.9	17.0
T4	10	10	10.7	10.6	9.5	8.4	9.3	9.4
T5	10	15	–	–	–	–	9.2	13.0
T6	10	20	10.6	17.0	9.3	16.1	9.1	17.4
T7	12	10	13.3	10.7	11.9	9.3	11.3	9.1
T8	12	15	–	–	–	–	10.9	12.9
T9	12	20	12.8	19.0	11.2	17.3	11.2	17.6
T10	20	10	23.8	10.1	20.4	9.6	19.4	13.2
T11	20	15	–	–	–	–	19.7	14.1
T12	20	20	21.8	19.1	19.6	18.0	19.6	18.2
T13	20	–	–	–	–	–	18.4	31.4

5.3.6 Data Processing and Metrics

For every run, we convert the robot and vehicle logs to a common time base, crop them to the documented maneuver interval, and align them with the trimmed video. Each vehicle acceleration record contains a constant offset, which we remove by subtracting its median over the samples in which the vehicle is at rest, that is, with a speed below 0.5 km/h. Roll and pitch are computed from the robot quaternion

Table 5.9: Quantities used in the physical-data analysis.

Quantity	Source	Interpretation
Vehicle speed and a_x or a_y	Vehicle CAN data	Realized disturbance
Roll and pitch	Robot quaternion	Trunk excursion and variation
J_x and its percentiles	Robot quaternion and joint state	Whole-body deviation and activity
σ_{att}	Median-centered roll and pitch	Attitude variation
Input-timed frames	Synchronized video	Visible posture and contact response

and expressed relative to their median over the first 2 s of the trial. All signals are lightly smoothed and resampled to a common time grid, with the same parameters for all conditions.
We quantify the robot response by the state-deviation index

$$J_x = \phi^2 + \theta^2 + \sum_{j=1}^{29} (q_j - q_{j,\text{ref}})^2 + 0.05 \sum_{j=1}^{29} \dot{q}_j^2, \tag{5.5}$$

where ϕ and θ are roll and pitch in radians, q_j and $\dot{q}_j$ are joint position and velocity, and $q_{j,\text{ref}}$ is the first joint position of the cropped trial. The index combines trunk attitude, deviation from the initial whole-body pose, and joint activity. A large value can therefore reflect either a large body excursion or an active multi-joint correction such as a step, so we read J_x together with the attitude and the video rather than as a failure criterion. The attitude variation

$$\sigma_{\text{att}} = \sqrt{\text{RMS}(\Delta\phi)^2 + \text{RMS}(\Delta\theta)^2} \tag{5.6}$$

uses roll and pitch after subtracting their medians over the evaluation window. It measures how much the torso moves about its typical orientation within a trial, so that the static lean during a turn does not count as oscillation. Low σ_{att} thus indicates a steady torso, whereas J_x also captures joint displacement and corrective activity.
For the cross-condition comparison, we interpolate the vehicle and robot streams to a common 0.05 s grid and evaluate each trial over the central 80% of its overlapping interval, omitting the first and last 10% to reduce sensitivity to the interval boundaries. For each trial we report the 95th percentile J_x^{95} and σ_{att}; a percentile is used instead of the maximum so that a single sample does not dominate. The same rule applies to every trial and condition. Table 5.9 summarizes the measures.
The vehicle setup did not include a handle force sensor, and we did not annotate step onsets or touchdowns with a common criterion. The physical comparison therefore rests on the kinematic measures above, complemented by the synchronized video, while handle load and step statistics are reported for the simulation evaluation.
Table 5.10 lists the fixed processing parameters, which are the same for all three controller–support conditions. The Gaussian smoothing of the vehicle speed sup-

presses the quantization of the GNSS-derived speed, and the 60 Hz playback table serves the synchronized inspection with the video.

Table 5.10: Fixed parameters used to process the physical vehicle and robot records.

Processing operation	Parameter
Acceleration zero-offset estimate	Median while vehicle speed is below 0.5 $\mathrm{km\,h^{-1}}$
Vehicle-acceleration smoothing	Centered 0.18 s moving mean
Main comparison time grid	0.05 s
Initial vehicle-speed smoothing	0.10 s window
GNSS-speed quantization smoothing	Gaussian kernel with $\sigma = 0.50$ s
Roll and pitch preprocessing	Centered 0.10 s mean and 0.10 s resampling
Aggregate evaluation interval	Central 80% of the overlapping vehicle–robot interval
Aggregate state-deviation statistic	95th percentile, J_x^{95}
Video-inspection playback table	60 Hz

5.3.7 Trial Inclusion

When an identifier was driven more than once, typically because an earlier attempt was judged unsatisfactory, we use the last complete run. An identifier enters the three-condition comparison only when vehicle and robot records are available for all three conditions. We examine S4 and T12 in detail because they contain clearly identifiable straight-road and steady-turn phases. With one run per identifier and different realized inputs, the physical results are a descriptive comparison of these runs rather than estimates of failure probability or repeatability.

6
Simulation Results

This chapter reports the simulation results of the learned policies. We first summarize the final training stage, then evaluate the no-handle and rigid-handle policies in Isaac Lab, describe the recovery behavior that they produce, and check the exported actors in MuJoCo. The suspended-handle policy is reported at the end of the chapter as a simulation-only extension. The evaluation protocol and metrics are defined in Section 5.2.

6.1 Training

Figure 6.1 shows the mean episode reward and the share of episodes ending by each condition over the final training stage of the two principal policies. Both stages continue from the preceding stage of the curriculum in Table 5.3, so the curves show refinement on the final disturbance distribution rather than learning from scratch. After a short transient at the change of distribution, the no-handle reward increases over the second half of the stage, and the share of episodes reaching the time limit rises from about 70% to about 80% before falling back towards 72% near the end. The rigid-handle reward and time-limit share settle within the first thousand iterations and then fluctuate around a constant level.
Table 6.1 lists the termination shares over the last 100 iterations. The no-handle policy rarely falls: low base height and excessive tilt together end 0.6% of episodes. Leaving the platform instead ends 27.8%, so that in the training distribution the finite floor, rather than loss of upright balance, is the dominant limitation. This

Table 6.1: Share of episodes ending by each condition, averaged over the last 100 iterations of the final no-handle (25,000 iterations) and rigid-handle (8,000 iterations) training stages. An episode can meet more than one condition at the same step, so a column can sum to slightly more than 100%.

End of episode	No handle	Rigid handle
Time limit reached	71.7%	78.5%
Platform boundary	27.8%	0.0%
Low base height	0.5%	0.0%
Excessive tilt	0.1%	0.0%
Handle torque overload	–	16.4%
Handle force overload	–	4.9%
Severe non-foot contact	–	1.0%
Gross posture deviation	–	0.4%

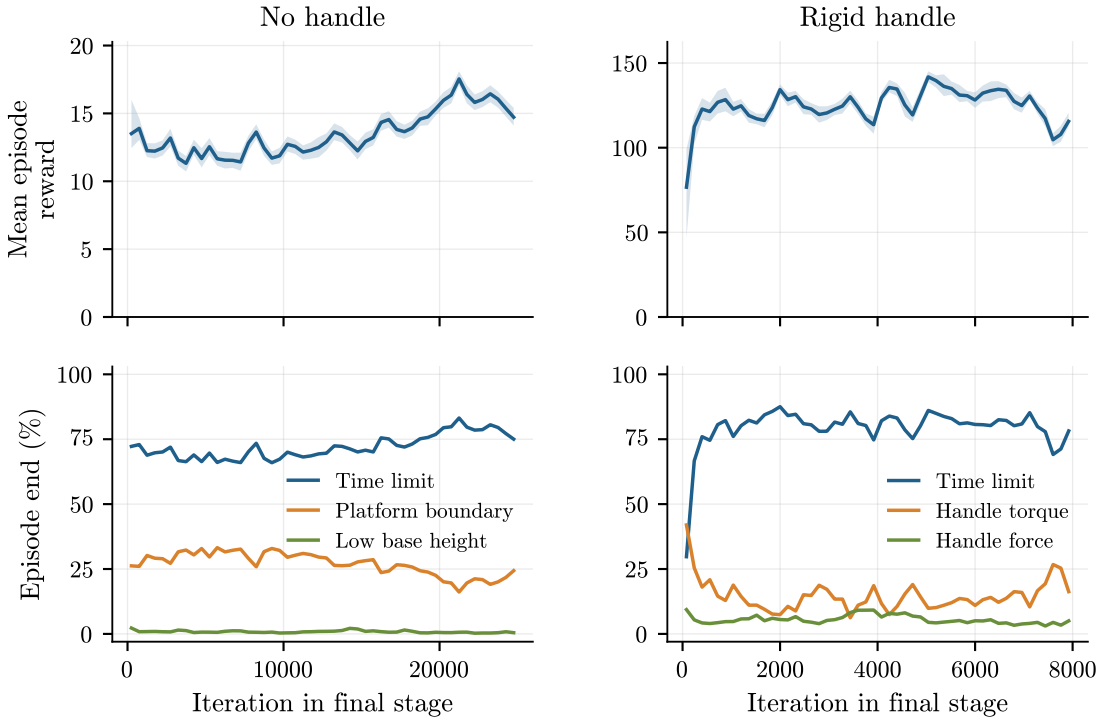


Figure 6.1: Final training stage of the no-handle and rigid-handle policies. Top: mean episode reward. Bottom: share of episodes ending by the time limit and by the most frequent termination conditions. Lines show the mean over consecutive blocks of 500 (no handle) and 160 (rigid handle) iterations; the shaded band in the top row is the interquartile range within each block.

is consistent with the stepping recovery of Section 3.6: successful recovery steps restore balance but also translate the robot across the floor, and repeated steps can eventually reach the boundary. The rigid-handle policy neither falls nor leaves the platform. Its early terminations are dominated by handle torque overload (16.4%) and handle force overload (4.9%), so the handle-load limits, rather than loss of foot-supported balance, bound its training performance. Torque overload ends more than three times as many episodes as force overload, so the grasp is limited by the moment about the wrist rather than by the pulling force. Because the training distribution deliberately includes severe cases, these shares are not evaluation success rates.

6.2 Isaac Lab Evaluation

Table 6.2 summarizes the evaluation within the training distribution. The no-handle policy completes almost all straight-line episodes, with and without grade, and takes two to five corrective steps per episode. Turning is more challenging: it completes 38.5% of the Lissajous and 57.0% of the Lissajous-with-grade episodes, and nearly all failures occur because the robot reaches the platform boundary, rather than because the acceleration disturbance directly causes a loss of balance. The rigid-handle policy completes about 90% of the straight-line and Lissajous episodes and

half of the constant-radius arcs, and handle-torque overload is the dominant failure mode. In all three families, its feet carry at least half of the body weight for more than 99% of the time. Its peak tilt includes the forward lean of about 6.5° of the support-ready posture.

Table 6.2: Isaac Lab evaluation within the training distribution, 200 episodes per motion family. Peak tilt is the median of the per-episode maximum; the fourth column gives the median number of corrective steps per episode (no handle) or the median 95th-percentile hand force as a fraction of body weight (rigid handle).

Policy	Motion family	Completed	$\alpha_{\max}$	Steps / $F_{H,95}$	Main failure
No handle	Straight line	99.5%	2.9°	2	tilt (1)
	Lissajous turning	38.5%	15.8°	6	boundary (111)
	Straight with grade	98.5%	6.6°	5	base height (3)
	Lissajous with grade	57.0%	9.3°	6	boundary (80)
Rigid handle	Straight line	89.5%	12.9°	$0.14B$	hand torque (21)
	Lissajous turning	90.0%	12.8°	$0.13B$	hand torque (20)
	Constant-radius arc	49.5%	12.9°	$0.13B$	hand torque (74)

Figure 6.2 shows how the two policies degrade as the peak acceleration and braking of the straight-line family is raised. The no-handle policy completes 99% of the episodes at 1.0 m/s^2 and 70% at 1.5 m/s^2, just above its training range, but only 14% at 2.0 m/s^2 and none beyond. Almost all failures end at the platform boundary: the policy keeps the torso upright by stepping backwards, and at 2.0 m/s^2 and above it leaves the usable floor area within about 3 s. The step rate rises from about 4 to more than 30 steps per 10 s and keeps rising at the levels where no episode is completed. At these levels the robot takes a rapid sequence of seven to ten backward steps (median) before the base crosses the boundary after about 2 s, so the number of steps per episode stays similar while the episode becomes much shorter. The step rate therefore reflects how fast the robot steps to keep upright, not a longer recovery. The rigid-handle policy completes all episodes at 2 m/s^2 and 78.5% at 3 m/s^2. From 4 m/s^2, which lies in the rarely sampled upper tail of its training range, nearly all episodes end by handle-torque overload, as the 95th-percentile hand torque reaches the 10 N m limit. In both conditions, the tolerable acceleration is therefore limited by the available support, the floor area or the handle load, rather than by a loss of upright balance.

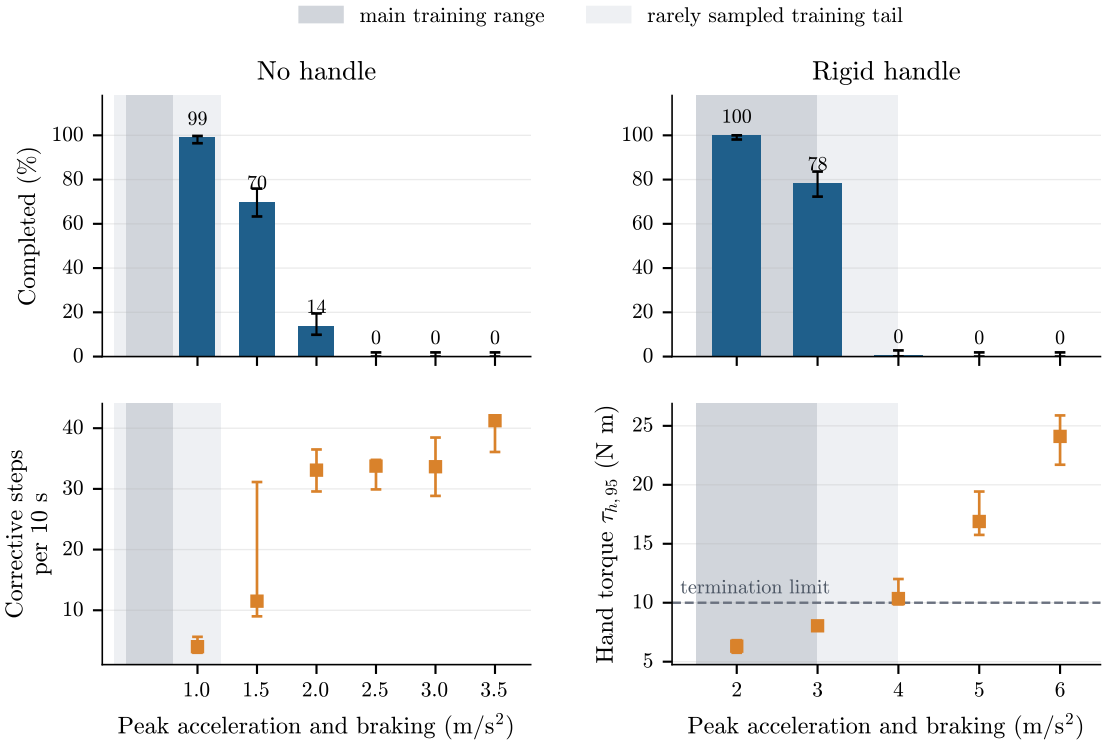


Figure 6.2: Straight-line severity sweep in Isaac Lab, 200 episodes per level. Top: completed episodes with 95% Wilson intervals. Bottom: corrective steps per 10 s (no handle) and 95th-percentile hand torque (rigid handle), as median and interquartile range. Dark shading marks the main training range of the peak acceleration and light shading the rarely sampled tails.

6.3 Recovery Behavior

6.3.1 No-Handle Step Recovery

Figure 6.3 shows a completed no-handle episode at a peak acceleration and braking of 1.5 m/s^2. During the acceleration phase the robot does not hold its position: it steps backwards in alternating single support and moves its base about 1.3 m towards the rear edge of the floor, while the tilt stays below 5.5° (frame (b)). During braking it steps forward again and comes to rest about 0.3 m behind the platform center, where it stands on both feet for the rest of the episode (frames (c) and (d)). The policy thus recovers by stepping with the disturbance. This keeps the torso upright but consumes floor area, which is the mechanism behind the boundary failures at higher accelerations.

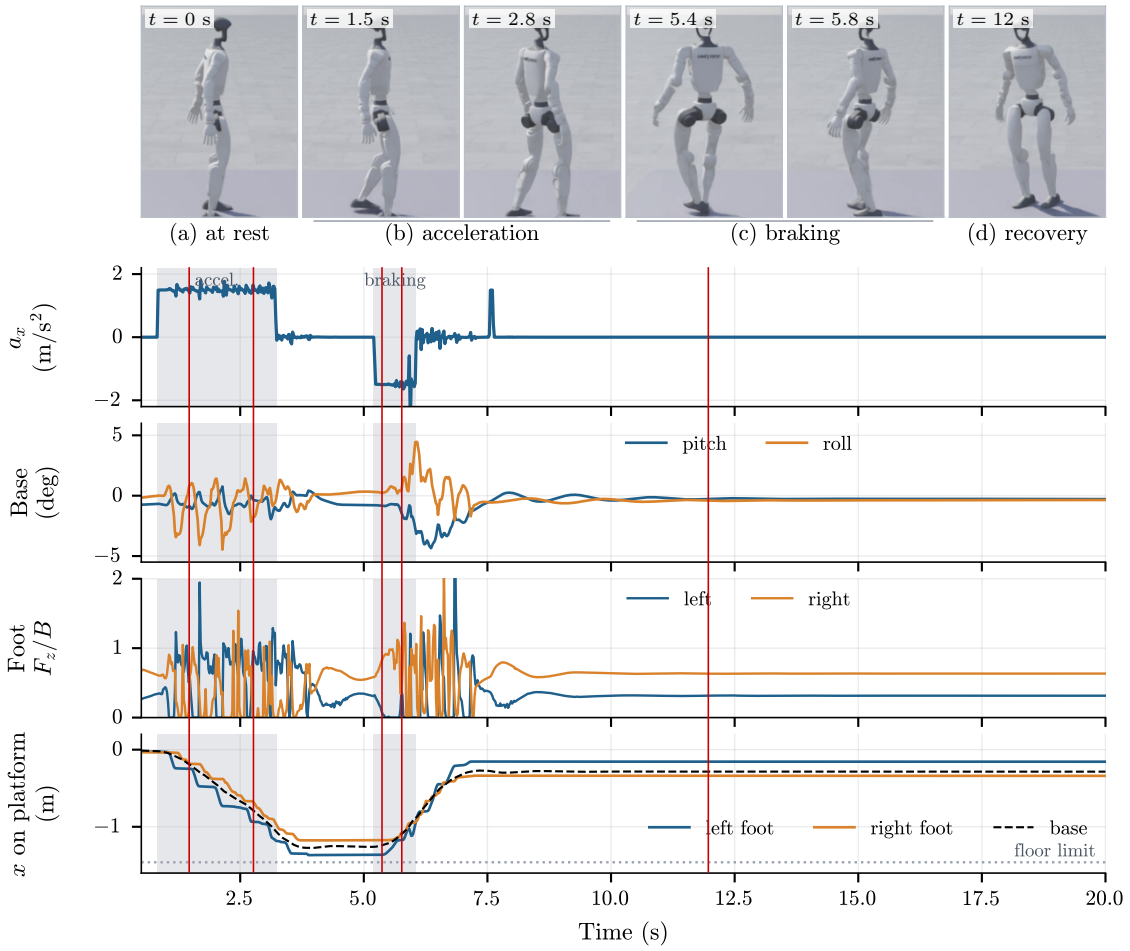


Figure 6.3: No-handle episode at a peak acceleration and braking of 1.5 m/s^2 in Isaac Lab. Top: frames (a) at rest, (b) during acceleration, (c) during braking, and (d) after recovery. Bottom: platform acceleration, platform-relative base pitch and roll, vertical foot forces normalized by the body weight B, and longitudinal position of the feet and the base on the platform; the dotted line marks the edge of the usable floor area. Shaded intervals mark acceleration and braking, and the thin red lines mark the approximate times t of the frames.

6.3.2 Rigid-Handle Support

Figure 6.4 shows a rigid-handle episode with two acceleration and braking cycles at 3 m/s^2. The robot does not step, its base moves by less than 0.1 m, the free arm swings out during acceleration (frame (b)), and neither foot carries less than about a quarter of the body weight. The base pitches from the 6.5° lean of the support-ready posture to about 15° during acceleration and returns at constant speed. The hand force stays below $0.22B$, whereas the hand torque briefly reaches the 10 N m limit during the acceleration phases. The hand torque is therefore the quantity that bounds the tolerable acceleration, consistent with Figure 6.2.

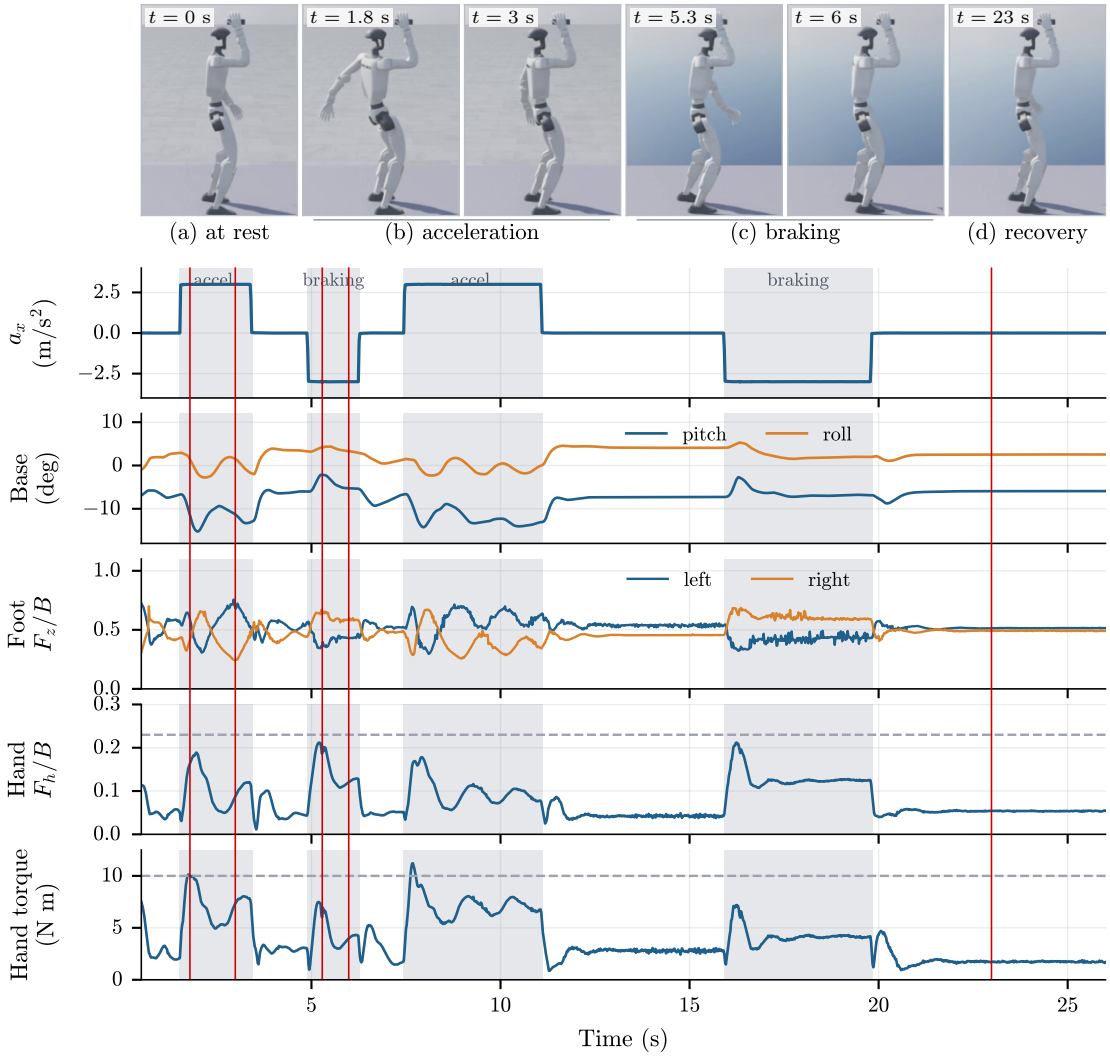


Figure 6.4: Rigid-handle episode at a peak acceleration and braking of 3 m/s^2 in Isaac Lab. Top: frames (a) at rest, (b) during acceleration, (c) during braking, and (d) after recovery. Bottom: platform acceleration, platform-relative base pitch and roll, vertical foot forces, hand force, and hand torque; dashed lines mark the termination limits and shaded intervals mark acceleration and braking; the thin red lines mark the approximate times t of the frames.

6.4 MuJoCo Sim-to-Sim Evaluation

Figure 6.5 shows both policies in MuJoCo on the two trajectories of Section 5.2. In the straight-line case the platform accelerates at 1.0 m/s^2 until it reaches 6.5 m/s after 6.5 s and then holds its speed. In the turning case it accelerates to 4.17 m/s while the yaw rate rises to about 0.21 rad/s, so that the horizontal acceleration settles near 0.9 m/s^2. Both policies remain standing throughout both trajectories. The no-handle policy regulates the disturbance through foot contact and whole-body posture adjustment, whereas the rigid-handle policy keeps its right-hand grasp. The

test thus also confirms the deployment interface.
Figure 6.6 repeats the severity sweep of Figure 6.2 in MuJoCo. The no-handle policy completes 97% of the episodes at 1.0 m/s^2, 53% at 1.5 m/s^2, and 13% at 2.0 m/s^2, and none beyond; as in Isaac Lab, most failures end at the platform boundary. The rigid-handle policy completes all episodes at 2 m/s^2 and 67% at 3 m/s^2, and from 4 m/s^2 all episodes end by handle-torque overload. Apart from one tilt failure at 6 m/s^2, the robot stays upright in every rigid-handle episode. Both simulators therefore show the same limits: the no-handle policy leaves the usable floor area from about 2 m/s^2, and the rigid-handle policy overloads the handle from 4 m/s^2.

6.5 Suspended-Handle Extension

Table 6.3 summarizes the suspended-handle policy on its two training families. It completes all Lissajous episodes and 74.5% of the straight-line episodes. Most straight-line failures end by foot-yaw drift (44 of 51 episodes), where the feet rotate on the floor beyond the limit of Table 3.4; hand-force overload ends five episodes. In both families the feet together carry at least half of the body weight for more than 99% of the time, so the policy balances on its feet rather than hanging from the handle. The handle swings by about 15° in straight-line motion and 11° in turning, and the estimated hand force reaches about $0.35B$ and $0.19B$, respectively.
The suspended handle thus transmits a larger hand force than the rigid handle, whose median $F_{H,95}$ is $0.14B$ in straight-line motion and $0.13B$ in turning (Table 6.2). This is not a direct comparison of the support types: the suspended-handle force is an estimate, and this policy was trained on milder motion, with straight-line accelerations of 0.75–3.0 m/s^2 instead of 1.5–4.0 m/s^2 and a lateral acceleration of at most 0.75 m/s^2 in turning. It does show how the policy uses the swinging support: the weight stays on the feet, and the hand force rises with the swing angle as the handle is deflected from the vertical (Figure 6.7).

Table 6.3: Suspended-handle policy in Isaac Lab, 200 episodes per training family. Tilt, swing angle, hand force, and wrist torque are medians over episodes of the per-episode maximum or 95th percentile.

	Straight line	Lissajous turning
Completed episodes	74.5%	100%
Peak tilt $\alpha_{\max}$	9.3°	5.5°
Peak handle swing angle	15.4°	11.2°
Estimated hand force $\widehat{F}_{h,95}/B$	0.35	0.19
Right-wrist torque, 95th percentile	9.0 N m	8.6 N m
Time with feet carrying $\geq 0.5B$	99.8%	99.8%

Figure 6.7 shows one episode on a fixed straight-line trajectory: the platform accelerates at 2.0 m/s^2 to 10 m/s, holds this speed for 4 s, and brakes at 2.0 m/s^2 to rest. The policy completes the episode. During acceleration the handle swings by up to 15°, the base pitches by about 6°, and the load shifts towards one foot, whose share drops to about $0.15B$; the estimated hand force peaks at $0.33B$. At constant speed the handle returns close to the vertical and the hand force falls below $0.1B$. During

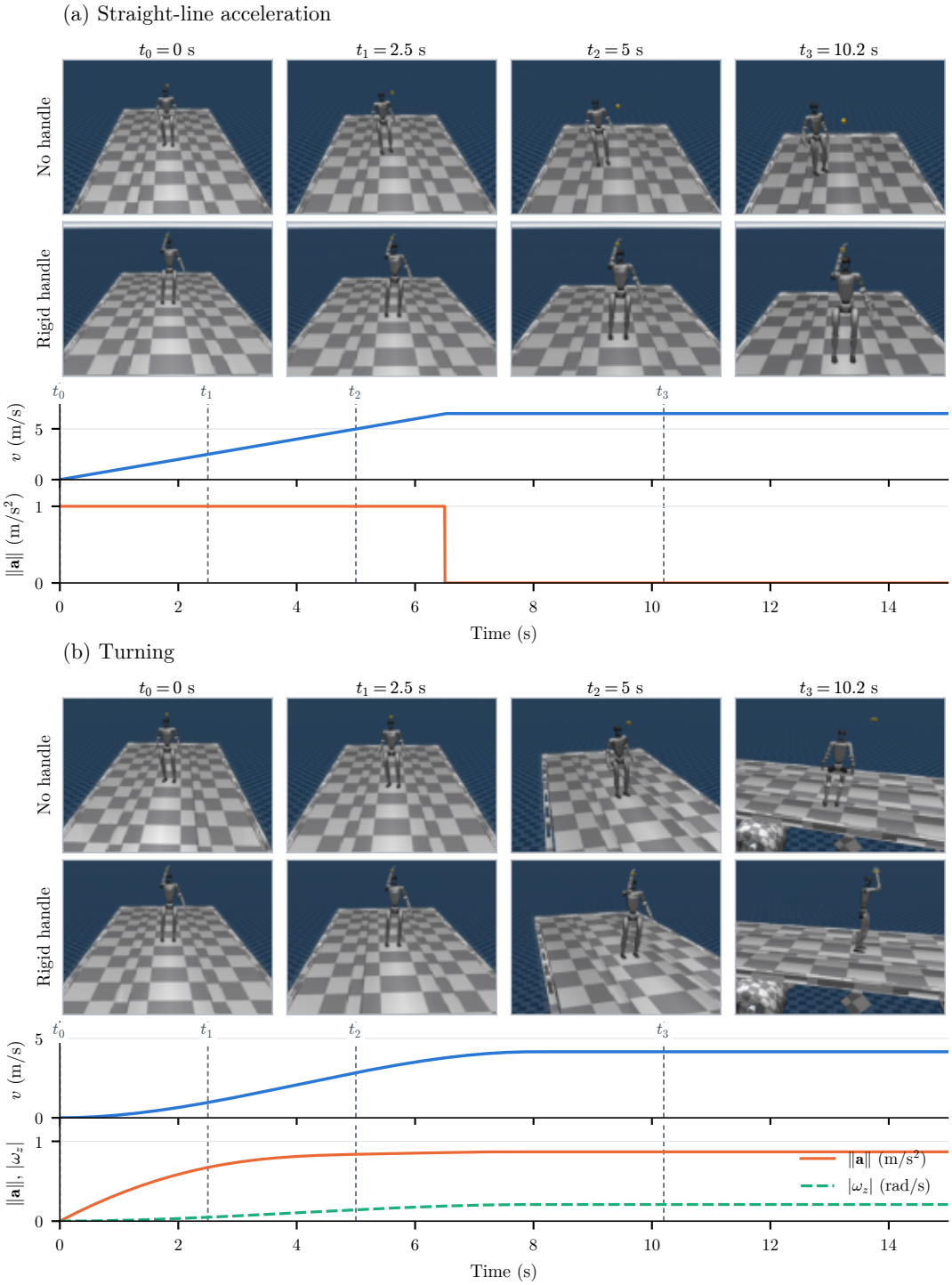


Figure 6.5: MuJoCo sim-to-sim evaluation of the no-handle and rigid-handle policies under (a) straight-line acceleration and (b) turning. The frames show both policies at the instants t_0–t_3 marked below them, starting from rest at $t_0 = 0$. The lower panels show the platform input: speed v, magnitude of the horizontal platform acceleration $\|\boldsymbol{a}\|$, and, for the turn, the yaw rate $|\omega_z|$.

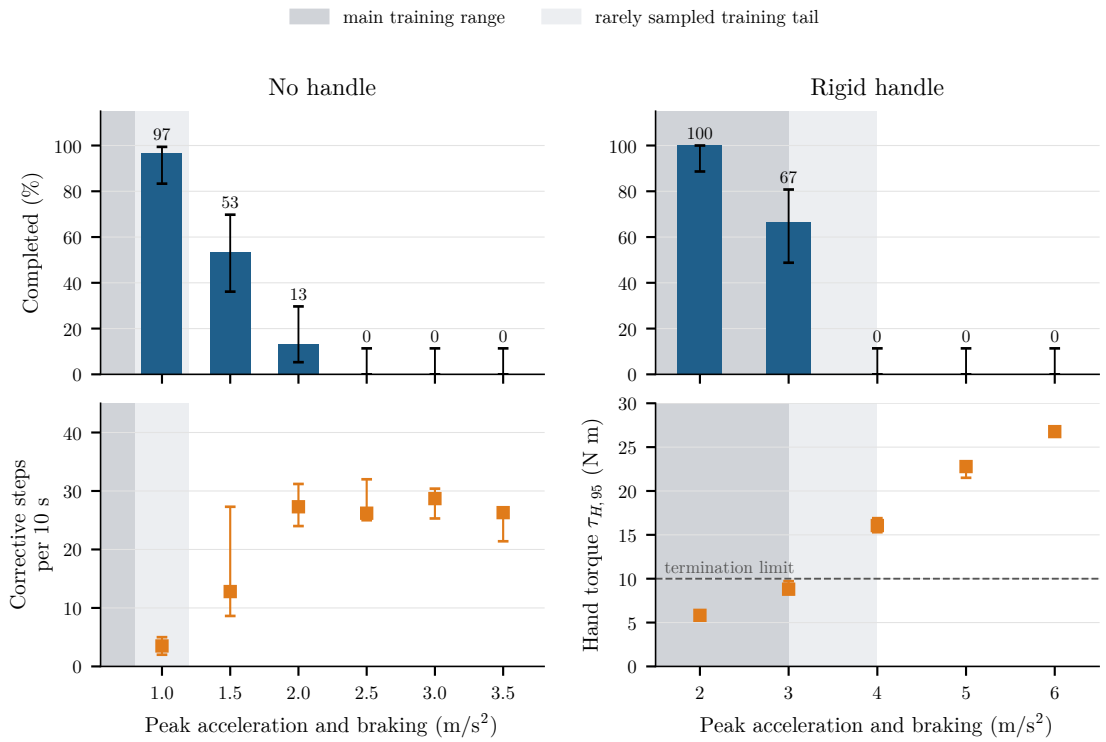


Figure 6.6: Straight-line severity sweep in MuJoCo, 30 episodes per level. Top: completed episodes with 95% Wilson intervals. Bottom: corrective steps per 10 s (no handle) and 95th-percentile hand torque (rigid handle), as median and interquartile range. Shading as in Figure 6.2.

braking the response is oscillatory: the hand force repeatedly reaches about $0.35B$ while the two feet share the load almost equally and the tilt stays below 5°. The largest tilt, 8.4°, occurs when the platform stops abruptly at the end of braking, and it falls below 3° about 0.8 s later. Averaged over each phase, the feet carry about the full body weight.

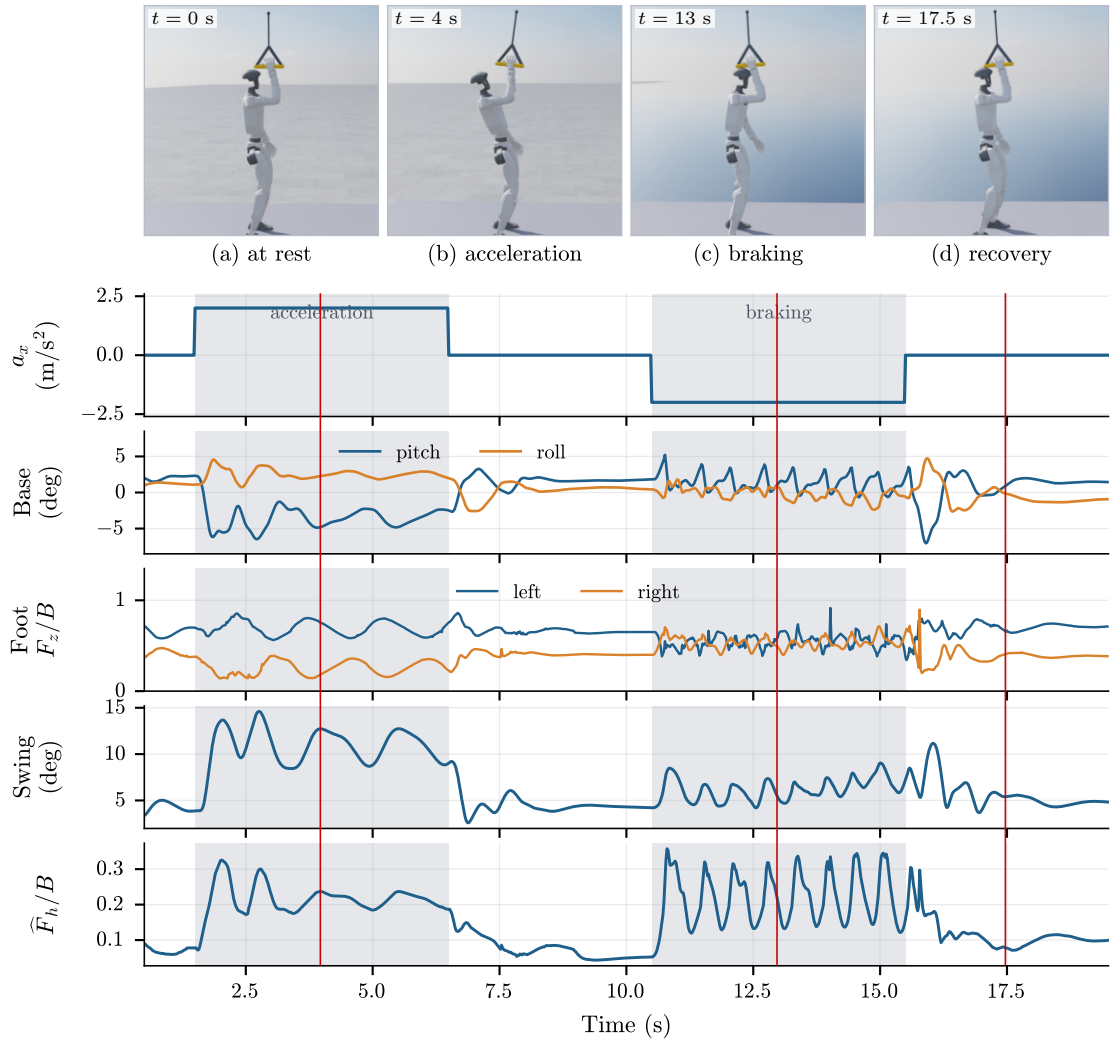


Figure 6.7: Suspended-handle episode on a fixed straight-line trajectory in Isaac Lab. Top: frames (a) at rest, (b) during acceleration, (c) during braking, and (d) recovery after the platform has stopped. Bottom: platform acceleration, platform-relative base pitch and roll, vertical foot forces, handle swing angle, and estimated hand force; forces are normalized by the body weight B. Shaded intervals mark acceleration and braking, and the thin red lines mark the approximate times t of the frames.

Table 6.4 lists how the suspended-handle evaluation episodes of Table 6.3 ended.

6.6 Summary

Neither principal policy falls in a noticeable share of episodes; their performance is bounded by the available support. The no-handle policy completes almost all straight-line episodes of its training distribution but fails more often in turning. It recovers by stepping with the disturbance, which keeps the torso upright but

Table 6.4: End of the suspended-handle evaluation episodes, 200 per family.

End of episode	Straight line	Lissajous turning
Time limit reached	149	200
Foot-yaw drift	44	0
Hand-force overload	5	0
Torso orientation	1	0
Base-yaw drift	1	0

consumes floor area, so that stronger accelerations drive it off the platform. The rigid-handle policy keeps both feet loaded and does not step; its limit is set by the hand torque. In MuJoCo, both exported policies remain standing under straight-line acceleration and turning and keep their support behavior, and the severity sweep shows the same limits as in Isaac Lab. The suspended-handle policy completes all turning episodes and three quarters of the straight-line episodes of its training distribution while keeping its weight on the feet; its main failure mode is foot-yaw drift in straight-line motion, and the swinging handle makes its response to braking oscillatory.

7

Physical Experiments and Results

This chapter reports the physical vehicle experiments with the manufacturer's standing controller, the learned no-handle policy, and the learned rigid-handle policy, using the setup and protocol of Section 5.3.2 and the measures of Section 5.3.6. In this chapter, these three controller–support conditions are labeled *default*, *no-handle RL*, and *rigid-handle RL*; tables and figures use the short labels Default, No-handle, and Rigid-handle. The rigid-handle configuration refers to the combined learned controller and rigid support. The chapter first compares the three conditions over their common test identifiers, then examines one straight-road and one turning case in detail, compares the two learned policies over a stronger straight-road sequence, and finally uses the synchronized robot-view recordings to distinguish active corrective motion from loss-of-balance events and engagement of the rear safety tether.

7.1 Overview of Completed Trials

Table 7.1 summarizes the trials retained for analysis, each with synchronized vehicle data, robot data, and video.

Table 7.1: Coverage of the analysis-ready physical dataset. The rigid-handle turning count includes the exploratory T13 run; the three-way subset contains only target identifiers available in every condition.

Condition	Straight	Turning	Three-way subset
Default	5	7	12
No-handle	10	8	12
Rigid-handle	10	13	12

The common comparison set comprises S1–S5 and T1, T4, T6, T7, T9, T10, and T12. Each point in the aggregate comparison represents one trial under each of the three conditions. After the S5 loss-of-balance event, the default-controller straight-road sequence was stopped for safety. Both learned configurations continued through S6–S10, which are therefore compared only between the two learned configurations. Appendix A shows the time histories of all retained trials other than S4 and T12, which are analyzed in Sections 7.3 and 7.4.

7.2 Paired Cross-Condition Comparison

Figure 7.1 compares the common test identifiers across the three controller–support conditions. Thin lines connect equal identifiers and diamonds show the condition medians. Table 7.2 reports the corresponding median and interquartile range.

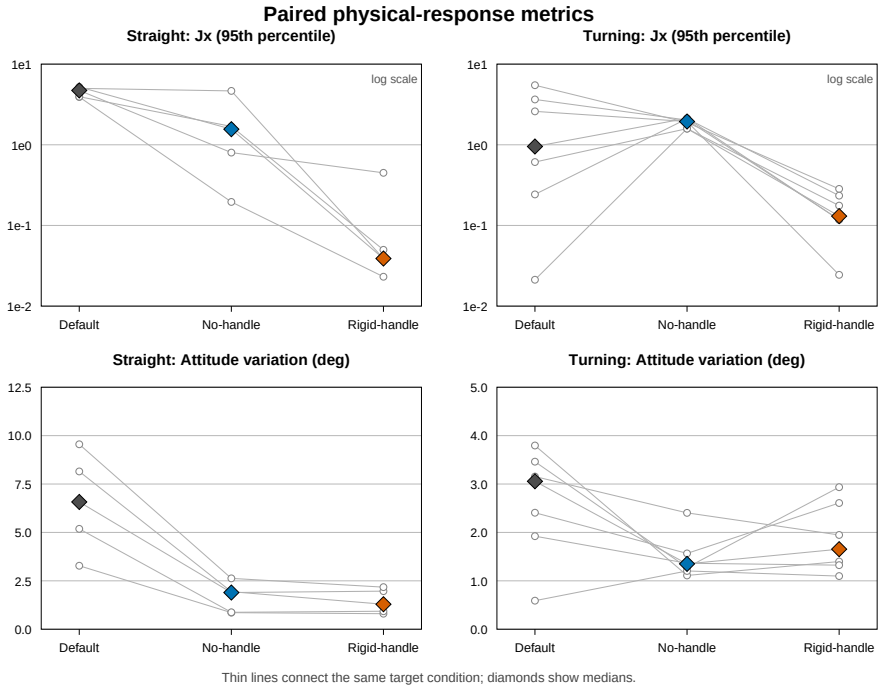


Figure 7.1: Paired physical-response metrics for the common straight and turning identifiers. Each circle is one recorded trial, thin lines connect equal target identifiers, and diamonds denote condition medians across the tested identifiers. J_x^{95} is displayed on a logarithmic scale.

Table 7.2: Median [interquartile range] across the tested identifiers for the two physical-response metrics. Each trial metric is calculated over the central 80% of its overlapping vehicle–robot interval.

Maneuver	Condition	J_x^{95}	σ_{att} (deg)
Straight	Default	4.719 [3.947, 5.006]	6.574 [5.185, 8.150]
	No-handle	1.561 [0.801, 1.690]	1.895 [0.874, 1.939]
	Rigid-handle	0.0389 [0.0388, 0.0499]	1.291 [0.929, 1.967]
Turning	Default	0.952 [0.427, 3.122]	3.057 [2.165, 3.308]
	No-handle	1.942 [1.736, 2.066]	1.350 [1.240, 1.465]
	Rigid-handle	0.131 [0.118, 0.205]	1.654 [1.361, 2.278]

The rigid-handle configuration gave the strongest overall result in the paired comparison. It had the lowest median J_x^{95} in both straight driving and turning, together with low torso attitude variation in both maneuver groups. This shows that combin-

ing the learned policy with rigid hand support was the most effective of the tested configurations for limiting the overall robot response.
Because J_x also responds to joint displacement and velocity (Section 5.3.6), a low σ_{att} can coexist with a larger J_x when the robot steps or reconfigures its body to keep the torso steady.
The straight-road comparison gives a clear ordering. Across S1–S5, the median J_x^{95} decreased from 4.719 with the default controller to 1.561 with no-handle RL and 0.0389 with rigid-handle RL. The median attitude variation decreased in the same order, from 6.574° to 1.895° and then to 1.291°. At the individual-trial level, no-handle RL was below the corresponding default value on both measures in all five straight tests. The straight-road data therefore support the conclusion that no-handle RL improved on the default controller, while the rigid-handle condition produced the lowest overall response values.
This ordering corresponds to three physically different responses. Under the default controller, the longitudinal disturbance produced both larger torso motion and larger deviation of the complete robot state. No-handle RL reduced the torso excursion while retaining the freedom to change foot placement and joint configuration. This behavior is consistent with the earlier event-gated recovery objective, which permits a single-support step when standing with the current contacts is no longer sufficient. The rigid-handle configuration added a rigid reaction path through the right hand and constrained the global motion of the robot, producing a further reduction in J_x^{95}.
The turning results show a difference between torso regulation and whole-body activity. No-handle RL reduced the median attitude variation from 3.057° for the default controller to 1.350°, and its attitude variation was lower than the corresponding default value in six of the seven paired turns. Its median J_x^{95} followed a different pattern, increasing from 0.952 to 1.942; no-handle RL was lower than the default controller on this measure in three turns and higher in four. The combination of steadier torso orientation and larger J_x in several turns indicates more active joint motion, stepping, or repositioning during no-handle recovery. Rigid-handle RL combined a low median attitude variation of 1.654° with the lowest median J_x^{95}, 0.131. Its J_x^{95} was lower than no-handle RL in all seven paired turns and lower than the default controller in six, giving the smallest typical whole-body deviation in the turning comparison.
The no-handle turning pattern is therefore a behavioral signature of active recovery rather than a conflict between the two metrics. The policy generally held the torso closer to a steady orientation while redistributing the body over the feet through joint motion and repositioning. The rigid-handle strategy used a different mechanism: the additional contact supplied an external load path, so lateral disturbance could be resisted with less displacement of the remaining whole-body state. These aggregate patterns are developed further in the S4 and T12 case studies.

7.3 Straight-Driving Results

Across the paired straight-road trials, no-handle RL reduced both response metrics relative to the default controller, and the rigid-handle configuration produced the

lowest median values. S4 illustrates how this ordering developed during acceleration, travel near the target speed, and braking.
S4 is used as a detailed case study near a target speed of 36 $\mathrm{km\,h^{-1}}$. Unlike the aggregate comparison, this case is shown over the complete overlapping vehicle–robot interval, so that acceleration, travel near the target speed, and braking appear in one trace. The input is summarized by the 1st and 99th percentiles of longitudinal acceleration and the 99th-percentile speed, and the response by the 99th percentiles of the largest absolute roll or pitch angle, Tilt_{99}, and of J_x. As shown in Figure 7.2, the measured 99th-percentile speeds were similar: 34.35, 33.73, and 33.49 $\mathrm{km\,h^{-1}}$ for the default, no-handle, and rigid-handle conditions, respectively. Table 7.3 lists the input and response summaries.

Table 7.3: S4 case-study input and response summaries. The acceleration is given as its 1st and 99th percentiles.

Condition	Speed_{99} (km/h)	$a_{x,1}/a_{x,99}$ ($\mathrm{m\,s^{-2}}$)	Tilt_{99} (deg)	$J_{x,99}$
Default	34.35	−2.12/1.46	25.87	15.23
No-handle	33.73	−1.12/1.16	5.57	13.15
Rigid-handle	33.49	−1.29/1.24	9.09	0.150

The similar achieved speeds were reached through different acceleration histories. In the default S4 interval, the 1st and 99th percentiles of the acceleration are −2.12 and 1.46 $\mathrm{m\,s^{-2}}$, compared with −1.12/1.16 $\mathrm{m\,s^{-2}}$ for no-handle RL and −1.29/1.24 $\mathrm{m\,s^{-2}}$ for rigid-handle RL. The default trial thus contains the largest acceleration span, while the two learned-policy trials have comparable input ranges.
Figure 7.2 separates the response into the same three phases. During acceleration and the interval near the target speed, the robot response develops gradually. The largest default-controller excursion appears as the near-target-speed phase ends and the strong braking input begins: torso attitude and J_x rise sharply while the vehicle speed falls. The no-handle record remains comparatively steady through the earlier phases and develops its main J_x transient during the later speed reduction, when the policy changes foot placement and reorganizes the body. The rigid-handle trace remains small across all three phases and shows no comparable whole-body peak.
The default trial, which also contained the strongest braking, reached a 99th-percentile maximum absolute roll-or-pitch angle of 25.87° and $J_{x,99} = 15.23$. The corresponding no-handle values were 5.57° and 13.15, so the smaller attitude excursion coexisted with a large transient in the whole-body state-deviation index. The rigid-handle trial had a larger absolute attitude envelope than the no-handle trial (9.09°), but its $J_{x,99}$ was only 0.150. Unlike σ_{att}, the reported Tilt_{99} is calculated from the recorded roll and pitch without subtracting the trial median and therefore retains posture offset. In this S4 record, the rigid-handle configuration showed the smallest large-excursion whole-body response.
Rigid hand contact permits the robot to maintain a supported equilibrium with a non-zero torso offset. That offset contributes to Tilt_{99}, while the limited joint displacement and corrective activity keep the complete J_x response small. Conversely,

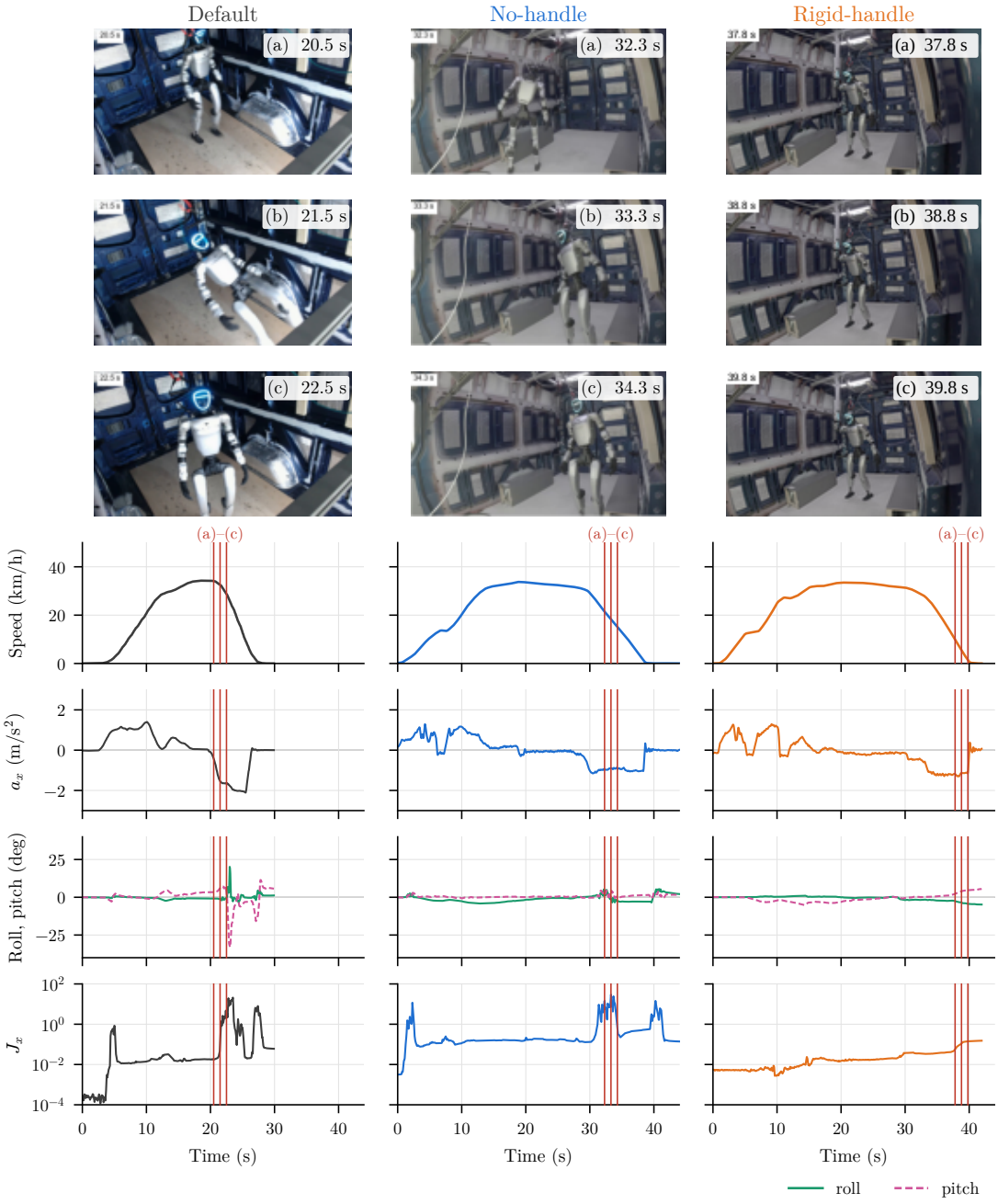


Figure 7.2: S4 straight-road case. Top: robot-view frames before (a), during (b), and after (c) the main response in each condition. Bottom: time histories over the complete overlapping analysis interval, showing vehicle speed, longitudinal acceleration a_x, torso roll and pitch relative to their median over the first 2 s, and the state-deviation index J_x on a logarithmic scale. Vertical lines mark the frame times (a)–(c).

the no-handle policy kept a smaller torso excursion but produced a large J_x through its corrective step and subsequent repositioning.
The frames (a)–(c) in Figure 7.2 show the posture before, during, and after the main S4 response. In the default sequence, the robot loses its upright posture and the rear tether visibly becomes taut. The no-handle sequence shows a distinct corrective step and postural reconfiguration followed by re-centering, whereas the rigid-handle sequence maintains hand contact with comparatively little visible displacement.

7.4 Turning Results

The seven paired turning identifiers cover nominal radii from 8 to 20 m. Table 7.2 shows that both learned configurations reduced median torso variation relative to the default controller, while the rigid-handle configuration produced the lowest median J_x^{95}. The higher no-handle J_x^{95} in several turns reflects a recovery strategy that maintains the torso through more active whole-body correction.
Turning imposes a different physical demand from straight-road acceleration and braking. The longitudinal cases contain two transient inputs of opposite sign, whereas the steady-turn windows maintain lateral acceleration over an extended arc. The robot must therefore sustain a laterally shifted support condition rather than respond only to a short impulse. Foot placement, joint configuration, and hand support can all contribute to maintaining this quasi-steady balance.
T12 provides the detailed 20 m-radius comparison. To separate sustained turning from turn entry and exit, the steady-turn window is selected from the vehicle data: a 1-s moving window counts as steady turning when the speed exceeds 15 km/h and the lateral-acceleration magnitude exceeds 0.8 $\mathrm{m\,s^{-2}}$ for at least 80% of its duration. Within the selected window, the input is summarized by the 95th percentiles of speed and $|a_y|$, and the response by σ_{att} and the 99th percentile of J_x. The selection identified windows of 27.5 s for the default controller, 26.5 s for the no-handle policy, and 35.5 s for the rigid-handle policy. Their measured 95th-percentile speeds were 20.48, 18.90, and 19.07 $\mathrm{km\,h^{-1}}$, respectively. The 95th-percentile lateral-acceleration magnitudes were similar: 1.73, 1.75, and 1.74 $\mathrm{m\,s^{-2}}$, respectively. These measured inputs are shown with the robot response in Figure 7.3, and Table 7.4 lists the summaries.

Table 7.4: T12 steady-turn case-study input and response summaries.

Condition	Speed_{95} (km/h)	$\lvert a_y\rvert_{95}$ ($\mathrm{m\,s^{-2}}$)	σ_{att} (deg)	$J_{x,99}$
Default	20.48	1.73	1.07	0.192
No-handle	18.90	1.75	2.36	4.442
Rigid-handle	19.07	1.74	1.30	0.118

In T12, the default controller had the smallest attitude variation (1.07°), followed by the rigid-handle strategy (1.30°) and the no-handle policy (2.36°). The rigid-handle record also had the smallest $J_{x,99}$, 0.118, compared with 0.192 for the default controller and 4.442 for the no-handle policy. Because the $|a_y|_{95}$ values are nearly

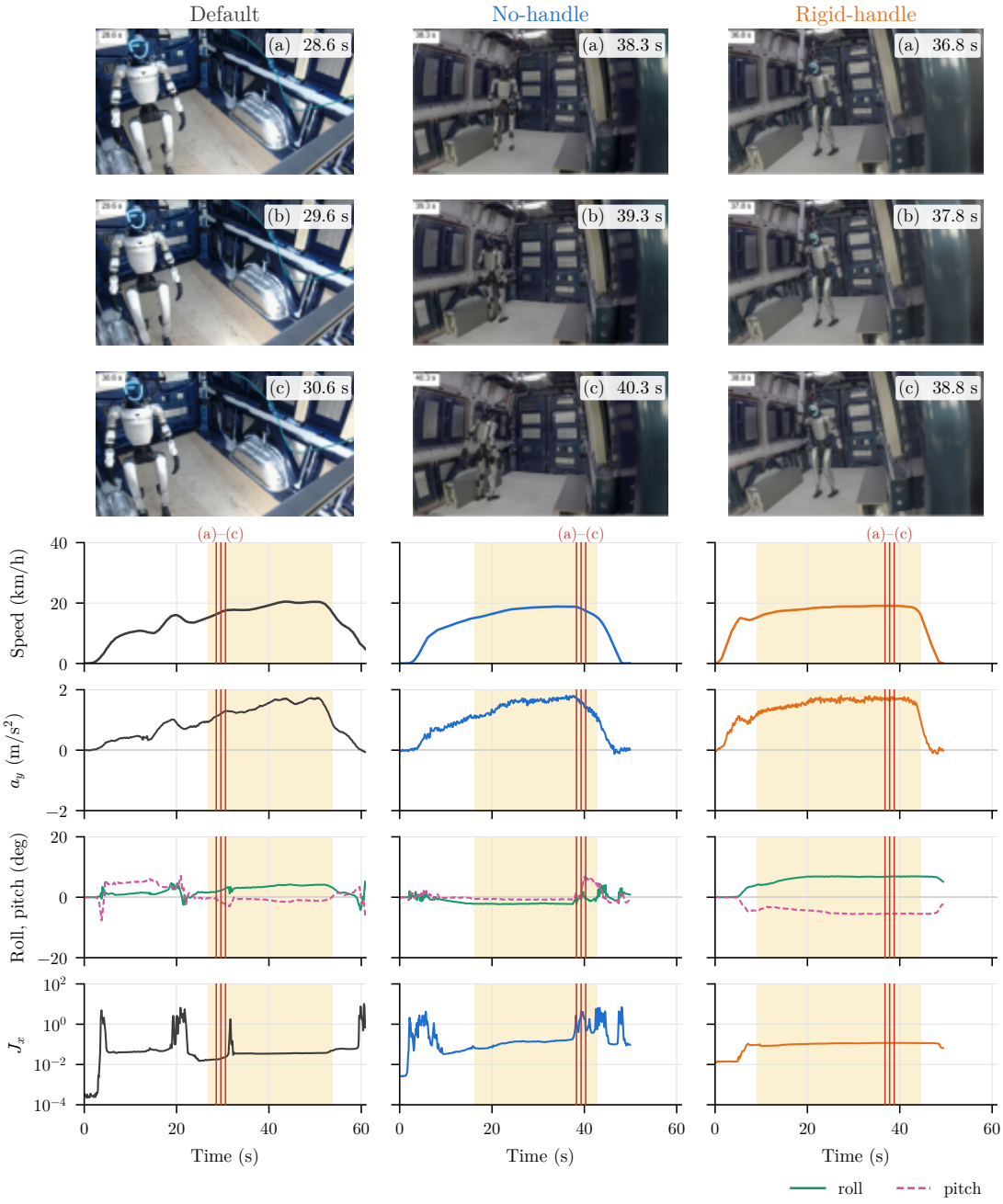


Figure 7.3: T12 turning case. Top: robot-view frames (a)–(c) showing the observed motion within each steady-turn window. Bottom: time histories over the complete trial, showing vehicle speed, lateral acceleration a_y with the sign chosen positive in the turn, torso roll and pitch relative to their median over the first 2 s, and J_x on a logarithmic scale. The shaded band marks the automatically selected steady-turn window, and the vertical lines mark the frame times (a)–(c).

equal, these response differences are not explained by a large difference in the sustained lateral input magnitude. The no-handle J_x trace increases sharply near the end of the selected window. The corresponding postural and foot adjustment is visible in frames (a)–(c) of Figure 7.3, illustrating how active whole-body recovery can produce a larger J_x response.
The frame sequence shows the observed responses within the respective steady-turn windows. The default sequence shows only a modest visible correction, consistent with its small $J_{x,99}$ in this particular turn. In the no-handle sequence, torso orientation remains bounded for most of the window before a postural and foot adjustment produces the late J_x increase. The policy then repositions the body while the sustained lateral load remains present. The rigid-handle sequence maintains the intended right-hand contact and a constrained supported stance; the additional load path limits whole-body displacement and produces the smallest $J_{x,99}$.
T12 is therefore useful as a mechanism-level example. Its individual attitude ordering differs from the aggregate medians, but the time histories and frames show how each strategy generates its measured response. The default controller requires little visible correction in this turn, no-handle RL uses active repositioning, and rigid-handle RL maintains a low-deviation supported stance.
The exploratory T13 run extends the rigid-handle evaluation at a 20 m radius. Over the same central-80% analysis window used for the aggregate metrics, its measured Speed_{95} was 33.17 $\mathrm{km\,h^{-1}}$ and $|a_y|_{95}$ was 5.38 $\mathrm{m\,s^{-2}}$. The corresponding responses were $J_x^{95} = 0.693$ and $\sigma_{\mathrm{att}} = 6.475°$. This lateral acceleration is the largest of all physical trials and exceeds the 4.5 $\mathrm{m/s^2}$ cap of the rigid-handle training. The robot kept its supported stance, while its attitude variation rose to about four times the rigid-handle turning median.

7.5 Extended Straight-Road Sequence

Figure 7.4 and Table 7.5 compare the two learned configurations over S6–S10 using the same central-80% window and response metrics as the main aggregate comparison.

Table 7.5: Extended S6–S10 straight-road results: median [interquartile range].

Condition	J_x^{95}	σ_{att} (deg)
No-handle	1.688 [1.285, 3.501]	2.455 [1.900, 2.584]
Rigid-handle	0.091 [0.076, 0.198]	1.754 [1.436, 2.058]

For all five identifiers, the connecting segment slopes downward from the no-handle to the rigid-handle condition in both metric panels. Across this subset, the median J_x^{95} was 94.6% lower and the median σ_{att} was 28.5% lower in the rigid-handle records. The difference between the two percentage reductions is physically informative. Rigid hand support reduced the whole-body index much more strongly than it reduced torso-attitude variation. The supported advantage was therefore not limited to keeping roll and pitch steady; it was most pronounced in the joint displacement and corrective activity also represented by J_x. This agrees with the main S1–S5

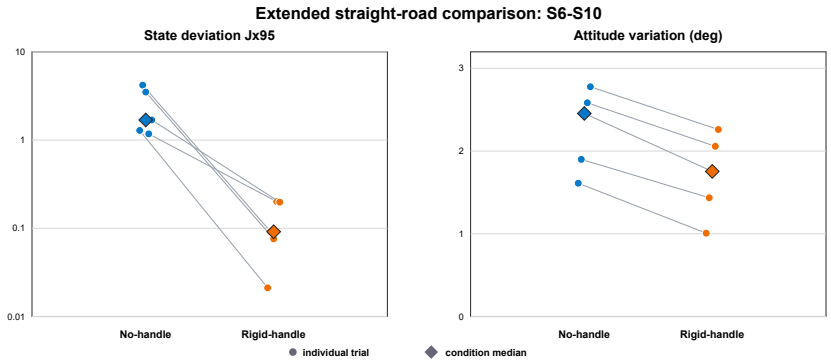


Figure 7.4: Extended straight-road comparison between the no-handle and rigid-handle learned configurations. Circles show the individual S6–S10 results, colored lines connect equal identifiers, and diamonds denote condition medians. The J_x^{95} panel uses a logarithmic scale.

comparison and with the constrained response visible in the rigid-handle case studies.

7.6 Failure Cases and Safety Interventions

The synchronized robot-view recordings show two loss-of-balance events under the default controller, in S4 and S5. In both events, the robot left the safe upright stance, moved toward the wheel arch or cabin interior, and caused the rear safety tether to become taut. The two events are shown in Figure 7.5.
In S4, the robot develops a rapid backward and rightward excursion between 20.5 and 21.5 s. Its torso and hip move into the wheel-arch region, while the safety tether becomes taut and arrests further descent. At 22.5 s the robot remains substantially displaced and supported rather than having independently returned to its original upright stance. The S5 sequence shows a comparable event: between 23.2 and 24.2 s the robot approaches the right-side vehicle interior and the tether again limits the subsequent motion.

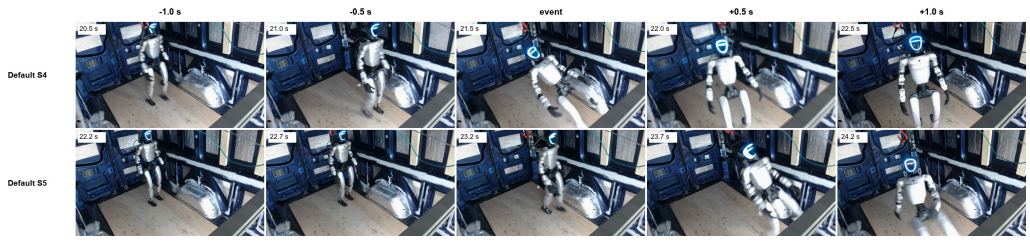


Figure 7.5: Video sequences of two visible loss-of-balance cases under the default controller. In both S4 and S5, the robot moves toward the right-side wheel-arch or cabin interior and the rear safety tether becomes visibly taut, limiting further descent.

These events are physically different from the no-handle S4 response even though

both can produce a large whole-body metric. In the no-handle sequence, the disturbance is followed by a corrective step, postural reconfiguration, and subsequent repositioning toward the usable floor area. In the default S4 and S5 sequences, displacement continues toward the vehicle structure until the external safety system arrests the motion. The distinction is therefore not the existence of a large response alone, but whether the robot converts that response into an active recovery or requires tether intervention.

7.7 Summary and Key Physical Findings

Taken together, the aggregate comparisons, case-study time histories, and synchronized robot-view recordings support the following six findings:

1. **No-handle RL improved torso regulation.** Relative to the default controller, no-handle RL produced lower median σ_{att} in both straight-driving and turning trials. It also reduced the median straight-road J_x^{95} from 4.719 to 1.561.
2. **No-handle RL used active whole-body recovery.** The S4 and T12 recordings show corrective stepping, postural adjustment, and subsequent repositioning. These actions raise the whole-body J_x measure; in S4 they coexist with low torso variation.
3. **Rigid-handle RL produced the smallest measured whole-body response.** It had the lowest median J_x^{95} in both maneuver families and the smallest S4 and T12 $J_{x,99}$ values.
4. **Both learned configurations completed the extended tests.** Both RL configurations were evaluated through S1–S10 without observed loss-of-balance events requiring tether intervention. In the stronger S6–S10 tests, the rigid-handle median J_x^{95} and σ_{att} were 94.6% and 28.5% lower, respectively, than the no-handle medians.
5. **The learned configurations avoided the observed loss-of-balance events.** In the five common straight-road tests (S1–S5), the default controller maintained balance without tether intervention in 3/5 trials (60%), compared with 5/5 (100%) for both RL configurations. The default controller lost balance in S4 and S5, with nominal target speeds of 36 and 45 km/h, respectively.
6. **The learned policies transferred to the physical G1.** The ONNX policies operated on the robot inside a moving road vehicle during acceleration, braking, and constant-radius turning, with synchronized vehicle, robot, and video records documenting their physical response.

8

Discussion

This chapter interprets the simulation and physical results in relation to the three research questions of Chapter 1. The central finding is that a direct joint-position policy can balance a humanoid on a moving floor from onboard measurements alone, and that the recovery strategy it learns depends on the available contacts. Without a handle, the policy recovers by stepping and whole-body motion; with a rigid handle, it keeps both feet loaded and uses the hand as an auxiliary support. In both cases, the policy stays upright, and the tolerable disturbance is set by the support: without a handle, the robot runs out of usable floor area; with a rigid handle, the hand moment reaches the 10 N m limit that the training imposes on the grasp. The latter is a limit on the load at the grasp, not a torque limit of the arm joints.

8.1 Balance without Direct Platform-Motion Observation

The first research question asked what balance and recovery behavior a direct joint-position policy can achieve when the platform motion is not part of its observation, and how this behavior differs between the support-specific objectives. This section addresses the first part. Neither policy receives the vehicle velocity, acceleration, path, or maneuver phase. The actor observes the disturbance only through its effect on the robot: a two-frame history of base angular velocity, IMU acceleration, projected gravity, and joint states. The simulation and vehicle results show that this information suffices to distinguish quiet standing from a developing loss of balance and to respond to it. Within their training distributions, both policies rarely fall: in the final training stage, low base height and excessive tilt together end fewer than 1% of the episodes of either policy (Table 6.1).

The results also show where this capability ends. When the peak acceleration is raised beyond the training range (Figure 6.2), neither policy fails by falling. The no-handle policy keeps its torso upright by stepping backwards and leaves the usable floor area once the acceleration reaches about 2 m/s^2; the rigid-handle policy stays in place but exceeds the hand-torque limit from about 4 m/s^2. The MuJoCo sweep shows the same two limits (Figure 6.6). The answer to the first part of RQ1 is therefore positive, with the qualification that the floor area and the hand-load limit, not upright balance, set the tolerable disturbance. In turning, the no-handle policy completes only 38.5–57% of its training episodes, again because it reaches the platform boundary rather than because it falls.

8.2 Support-Specific Recovery Strategies

The second part of RQ1 concerned the behaviors induced by the support-specific objectives. The two principal policies solve the same problem, remaining upright on a moving finite floor, with different contact strategies.

8.2.1 No-Handle Stepping

The no-handle reward pays a recovery bonus only after the sequence of instability, liftoff, touchdown, and stable double support, and it penalizes single support during quiet standing (Section 4.5). The resulting policy stands still under mild disturbances and steps when its current stance can no longer arrest the body motion. Under sustained acceleration, however, it does not step once and stop: it keeps stepping with the disturbance. At 1.5 m/s^2 in simulation, just above its training range, it retreated about 1.3 m towards the rear edge of the floor during acceleration and walked forward again during braking, ending about 0.3 m behind the center (Figure 6.3). It thereby trades floor area for torso stability, which the centering penalty does not prevent.
The physical S4 and T12 cases show the same pattern on the vehicle: a corrective step, postural reconfiguration, and repositioning. The stepping also explains why J_x, which includes joint displacement and velocity, can be larger for the no-handle policy than for the default controller in some turns even though its torso is steadier.

8.2.2 Rigid-Handle Support

With a rigid handle, an ideal grasp would allow the robot to hang from the handle. The rigid-handle objective prevents this by keeping the feet as the primary support, penalizing hand load, and returning the robot to its support-ready stance. The simulation confirms that the policy uses the handle in this way: it does not step, the feet carry at least half of the body weight for more than 99% of the time, and the 95th-percentile hand force stays around $0.13B$ (Table 6.2). Its failures are hand-torque overloads, not falls.
On the vehicle, the rigid-handle configuration produced the lowest median J_x^{95} in both straight driving and turning, the smallest $J_{x,99}$ in S4 and T12, and a smaller response than the no-handle policy in every S6–S10 pair. Its torso-attitude variation was the lowest in straight driving, whereas in turning it was slightly higher than that of the no-handle policy (1.65° against 1.35°). The added contact thus mainly reduces whole-body motion, while torso regulation is comparable to that of the no-handle policy.

8.2.3 Suspended Handle

The suspended-handle policy completed all turning episodes and three quarters of the straight-line episodes of its training distribution while keeping its weight on the feet. Its training distribution is considerably milder than that of the rigid-handle policy, particularly in turning, so the two handle conditions cannot be compared

directly. The representative episode (Figure 6.7) shows that the swinging handle adds its own dynamics: the response to braking is oscillatory, and the main straight-line failure is foot-yaw drift. The results show that the support-specific formulation extends to a moving grasp point in simulation.

8.3 Transfer from Isaac Lab through MuJoCo to Hardware

The second research question asked to what extent the learned behaviors transfer from Isaac Lab through MuJoCo to the physical G1. In MuJoCo, both exported actors remained standing under straight-line acceleration and turning and kept their support behavior, and in the severity sweep their completion rates followed those in Isaac Lab. Because the actors ran through the deployment interface, this also verified the observation order, history, action scaling, and timing before the vehicle tests. On the vehicle, the same actors ran at 50 Hz inside the 1 kHz state machine and reproduced the behaviors seen in simulation: stepping and repositioning without a handle, and a steady supported stance with the rigid handle.
Two features of the training formulation support this transfer. The actor observes only signals available on the robot, so deployment removes no input that the policy relied on in training, and the motion families expose it to longitudinal, lateral, and yaw disturbances rather than to one repeated waveform. In straight driving, the physical inputs stayed within the no-handle training range: in S4, the 1st and 99th percentiles of the longitudinal acceleration were -1.12 and 1.16 m/s^2. In turning, the lateral acceleration of the physical turns exceeded that of the no-handle turning families (at most about 1.4 m/s^2), and T13 exceeded the 4.5 m/s^2 cap of the rigid-handle training. The vehicle results therefore also show some generalization beyond the trained envelope.

8.4 Comparison with the Manufacturer's Standing Controller

The third research question compared the physical response with that of the manufacturer's standing controller. With the same foot-only contact set, the no-handle policy reduced the median torso-attitude variation in both straight driving and turning. In straight driving, it also reduced the median J_x^{95} from 4.7 to 1.6 and was lower than the default controller on both measures in all five paired trials. In turning, its median J_x^{95} was higher than that of the default controller (1.9 against 0.95), because the policy corrects actively through the feet and joints. Within the T12 steady-turn window, the default controller showed the smallest attitude variation of all three conditions. The advantage of the no-handle policy therefore lies in torso regulation and in recovery from large disturbances, not in every metric.
The clearest difference appears in the largest events. In S4 and S5, the default controller, whose S4 trial also contained the strongest braking, moved towards the wheel arch or cabin interior until the safety tether arrested it, and its sequence was

stopped for safety. The no-handle policy responded to S4 with a corrective step and repositioning, and both learned configurations continued through the stronger S6–S10 sequence. The rigid-handle configuration gave the smallest whole-body response overall; for a vehicle that provides a handhold, it is the most effective of the tested configurations.

8.5 Implications for Humanoid Operation in Public Transport

The results suggest that balance in a moving vehicle can be treated as a reactive whole-body control problem that does not require knowledge of the future vehicle motion. A controller that responds to the disturbance as it enters through the contacts can be deployed without an interface to the vehicle.
The finite floor is part of this problem. A policy that steps with every sustained acceleration survives each event but consumes the space needed for the next one. For unsupported operation, a stronger incentive to return to the center, or to lean into sustained accelerations instead of stepping, would be needed before the policy can handle longer or stronger acceleration phases. A handhold removes this limitation and shifts the limit to the load that the hand and grasp can sustain, which makes hand support an effective, but load-limited, resource.

8.6 Limitations

The grasp is idealized. In simulation, the rigid-handle grasp is a fixed joint between the palm and the platform, and the suspended handle is a rigid body, so neither models grip compliance or slip. For safety, the physical experiments also fixed the right hand to the support instead of grasping it with the dexterous hand of the G1, so grip compliance and slip were not tested.
The vehicle maneuvers were driven by a human driver. Speed, acceleration, and the duration of each phase could therefore be matched only approximately between runs and conditions, which makes the comparison between conditions less direct than in simulation. Straight-line and constant-radius maneuvers on a proving ground also cover only part of the disturbances that occur in regular bus or tram service, such as stop-and-go traffic, combined braking and turning, and uneven road surfaces.
The policies perceive the disturbance only through their own inertial and joint measurements and therefore react to it without anticipation. A passenger standing with closed eyes is in a similar situation: without seeing the road and the traffic, they cannot brace before the vehicle brakes or turns and are more easily knocked off balance. The robot likewise starts to respond only after the disturbance has reached its body.
Residual discrepancies between simulation and the physical robot remain despite the sim-to-sim and real-world validation. The contact and friction models, the robot dynamics, and the actuator response in Isaac Lab and MuJoCo do not reproduce the G1 exactly, and the real vehicle adds high-frequency vibration and floor motion that

the synthetic road overlay of the training represents only approximately. The MuJoCo and vehicle results show that the policies tolerate the differences encountered here, but they do not quantify the remaining margin.

9
Conclusions and Outlook

9.1 Summary of the Work

This thesis investigated whether a full-size humanoid can remain standing and recover on a finite floor that moves as it would in a road vehicle. We formulated the problem as direct joint-level control of the 29-DoF Unitree G1. The actor receives only measurements available on the robot and does not observe the vehicle trajectory, acceleration, or maneuver phase. We trained the policies in Isaac Lab with proximal policy optimization and an asymmetric critic, checked the exported actors and deployment interface in MuJoCo, and evaluated the two principal support conditions on the physical robot in a moving research vehicle.
The main methodological choice was to make the objective depend on the available support. Without a handle, the policy was trained to stand quietly under small disturbances and to complete a recovery sequence when a step was needed. With a rigid handle, the objective instead retained a support-ready posture, kept the feet as the primary load-bearing contacts, and limited the load applied through the hand. A ceiling-pivoted suspended handle was also studied in simulation to examine how the formulation extends to a moving grasp point. The physical evaluation compared the no-handle and rigid-handle configurations with the manufacturer's standing controller during measured straight-road acceleration and braking and during constant-radius turning.

9.2 Main Findings

The main findings are listed below. The first two answer RQ1, and the third and fourth answer RQ2 and RQ3, respectively.

1. **Reactive balance is possible without direct observation of the moving platform.** The policies inferred the developing disturbance from base angular velocity, IMU acceleration, projected gravity, and joint states. Within the tested motion families, this information was sufficient for quiet standing, postural correction, and recovery. The severity sweep also showed that the useful operating range was not ultimately bounded by a loss of upright attitude. The no-handle policy reached the platform boundary after repeated steps, whereas the rigid-handle policy reached its hand-torque limit. The available support therefore determined how the response failed.
2. **The support-specific objectives produced distinct recovery strategies.** The no-handle policy used whole-body motion and corrective stepping.

In simulation it completed almost all straight-line episodes of its training distribution but only about half of the turning episodes, and sustained acceleration caused it to keep stepping with the disturbance and consume the finite floor area. The rigid-handle policy remained in place, its feet together carried at least half of the body weight for more than 99% of the time, and it used the hand as an auxiliary load path rather than hanging from it. The suspended-handle study showed that a foot-loaded strategy could also be learned with a moving grasp point, although the handle swing introduced an additional oscillatory response.

3. **The learned behaviors transferred through the complete deployment chain.** Both exported actors remained standing in the MuJoCo checks while preserving their characteristic support behavior, and the MuJoCo severity sweep showed the same acceleration limits as Isaac Lab. The same actors then ran at 50 Hz within the robot's 1 kHz control state machine and reproduced the expected responses in the moving vehicle: stepping and repositioning without a handle, and a constrained supported stance with the rigid handle. This demonstrates transfer from Isaac Lab through an independent simulator to the physical G1 for disturbances within the tested operating range.
4. **The learned configurations improved the physical response, but the advantage depended on the support and the response measure.** Over the straight-road trials S1–S5, the no-handle policy reduced the median torso-attitude variation from 6.6° to 1.9° and the median state-deviation measure J_x^{95} from 4.7 to 1.6 relative to the manufacturer's controller. In the turns, it again reduced the median attitude variation, from 3.1° to 1.4°, while its median J_x^{95} increased from 0.95 to 1.9 because recovery involved more active joint motion and repositioning. The rigid-handle configuration produced the lowest median J_x^{95} in both maneuver groups and a smaller response than the no-handle policy in every S6–S10 comparison. The manufacturer's controller lost its upright stance in S4 and S5 and was arrested by the safety tether, whereas both learned configurations continued through the stronger straight-road sequence.

Taken together, these findings support the central conclusion of the thesis: a direct joint-position policy trained on a moving finite platform can transfer to a physical humanoid in a road vehicle without receiving the vehicle motion as an input. The controller should, however, be designed for the contacts that are actually available. Foot-only recovery preserves balance by spending floor area, while hand-supported recovery reduces whole-body motion by spending hand-load capacity. A practical balance system must therefore manage the corresponding support resource rather than treat upright posture as its only objective.

9.3 Outlook

The first priority for unsupported balance is to make recovery economical in floor space (Section 8.5). Future training should distinguish transient disturbances from persistent acceleration and reward a controlled lean, an early arrest of the stepping sequence, and a return towards the center once the disturbance has passed. Training

on longer sequences of maneuvers would expose the policy to the floor area that successive events consume together, rather than to one event at a time.
The hand-supported formulation should next use the dexterous hand of the G1. In this work the grasp was a fixed joint in simulation and a fixed attachment on the vehicle, chosen for safety. A grasp with the dexterous hand would turn the hand-load limit into a physical property of the grip, since the handle slips once the grip capacity is exceeded, and would bring the supported strategy closer to how a human holds a handrail. The same hand would allow the controller to acquire and release a handle, and a physical evaluation of the suspended handle with a real strap would test whether the oscillatory response found in simulation persists when the strap, grasp, and arm are all compliant.
The physical evaluation should move from a proving ground into regular service. A robot riding a tram from the first to the last stop would meet the full range of conditions that a closed test cannot reproduce: stops and departures, curves of varying radius, track irregularities, traffic, and combined braking and turning. Such rides would also provide measured vehicle motion, including high-frequency vibration, that can be replayed in simulation. Training and evaluating on these recorded disturbances, together with identifying the actuator response of the G1, would further reduce the residual discrepancies between simulation and hardware.
Finally, the policy could anticipate the disturbance instead of only reacting to it. A passenger who can see out of the window braces before the vehicle brakes or turns. Robot vision could provide the same cue: the motion and blur of the scenery outside the windows indicate the vehicle speed and whether it is accelerating or braking, and the approaching stop or curve indicates what comes next. Such an estimate could enter the policy as an additional input, while the reactive policy developed here would remain responsible for disturbances that are not foreseen. Combined with detecting whether a suitable handhold is available and selecting between unsupported and supported balance, this would move from isolated balance controllers towards a humanoid that manages its contacts and support throughout a complete journey.

Bibliography

[1] S. Hong and S. Park, "Biomechanical optimization and reinforcement learning provide insight into transition from ankle to hip strategy in human postural control," *Scientific Reports*, vol. 15, p. 13640, 2025.

[2] P. Gosine, V. Komisar, and A. C. Novak, "The effect of handrail cross-sectional design and age on the speed and quality of reach-to-grasp reactions to recover balance," *Human Factors*, vol. 66, no. 1, pp. 56–70, 2024.

[3] S. Krašna, A. Keller, A. Linder, A. P. Silvano, J.-C. Xu, R. Thomson, and C. Klug, "Human response to longitudinal perturbations of standing passengers on public transport during regular operation," *Frontiers in Bioengineering and Biotechnology*, vol. 9, p. 680883, 2021.

[4] J.-C. Xu, A. P. Silvano, A. Keller, S. Krašna, R. Thomson, C. Klug, and A. Linder, "Identifying and characterizing types of balance recovery strategies among females and males to prevent injuries in free-standing public transport passengers," *Frontiers in Bioengineering and Biotechnology*, vol. 9, p. 670498, 2021.

[5] A. Keller and S. Krašna, "Accelerations of public transport vehicles: A method to derive representative generic pulses for passenger safety testing," *Frontiers in Future Transportation*, vol. 4, p. 931780, 2023.

[6] M. Dodd, I. Knight, A. Edwards, M. Edwards, P. Ball, J. Appleby, and J. Radcliffe, "Bus safety standard: Bus braking data analysis," TRL Limited, Wokingham, UK, Tech. Rep. PPR1013, 2022.

[7] J. P. Powell and R. Palacín, "Passenger stability within moving railway vehicles: Limits on maximum longitudinal acceleration," *Urban Rail Transit*, vol. 1, no. 2, pp. 95–103, 2015.

[8] S. Kajita, F. Kanehiro, K. Kaneko, K. Fujiwara, K. Harada, K. Yokoi, and H. Hirukawa, "Biped walking pattern generation by using preview control of zero-moment point," in *Proceedings of the 2003 IEEE International Conference on Robotics and Automation*, 2003, pp. 1620–1626.

[9] J. Pratt, J. Carff, S. Drakunov, and A. Goswami, "Capture point: A step toward humanoid push recovery," in *Proceedings of the 2006 IEEE-RAS International Conference on Humanoid Robots*, 2006, pp. 200–207.

[10] A. Duburcq, F. Schramm, G. Boéris, N. Bredeche, and Y. Chevaleyre, "Reactive stepping for humanoid robots using reinforcement learning: Application to standing push recovery on the exoskeleton atalante," in *Proceedings of the 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2022, pp. 6092–6099.

[11] X. Cheng, Y. Ji, J. Chen, R. Yang, G. Yang, and X. Wang, "Expressive whole-body control for humanoid robots," in *Proceedings of Robotics: Science and Systems*, 2024.

[12] G. Nava, D. Pucci, N. Guedelha, S. Traversaro, F. Romano, S. Dafarra, and F. Nori, "Modeling and control of humanoid robots in dynamic environments: iCub balancing on a seesaw," in *Proceedings of the 2017 IEEE-RAS International Conference on Humanoid Robots*, 2017, pp. 263–270.

[13] A. Xi, T. W. Mudiyanselage, D. Tao, and C. Chen, "Balance control of a biped robot on a rotating platform based on efficient reinforcement learning," *IEEE/CAA Journal of Automatica Sinica*, vol. 6, no. 4, pp. 938–951, 2019.

[14] Y. Gao, Y. Gong, V. Paredes, A. Hereid, and Y. Gu, "Time-varying ALIP model and robust foot-placement control for underactuated bipedal robotic walking on a swaying rigid surface," in *Proceedings of the 2023 American Control Conference*, 2023, pp. 3282–3287.

[15] T. Belvedere, N. Scianca, L. Lanari, and G. Oriolo, "Joint-level IS-MPC: A whole-body MPC with centroidal feasibility for humanoid locomotion," in *Proceedings of the 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024, pp. 11 240–11 247.

[16] S. H. Bang, J. Lee, C. Gonzalez, and L. Sentis, "Variable inertia model predictive control for fast bipedal maneuvers," in *Proceedings of the 2024 IEEE 63rd Conference on Decision and Control (CDC)*, 2024, pp. 4334–4341.

[17] I. Bergonzani, M. Popescu, S. Kumar, and F. Kirchner, "Fast dynamic walking with RH5 humanoid robot," in *2023 IEEE-RAS 22nd International Conference on Humanoid Robots (Humanoids)*. Austin, TX, USA: IEEE, 2023.

[18] C. Yang, T. Komura, and Z. Li, "Emergence of human-comparable balancing behaviors by deep reinforcement learning," in *Proceedings of the 2017 IEEE-RAS International Conference on Humanoid Robots*, 2017.

[19] X. Gu, Y.-J. Wang, X. Zhu, C. Shi, Y. Guo, Y. Liu, and J. Chen, "Advancing humanoid locomotion: Mastering challenging terrains with denoising world model learning," in *Proceedings of Robotics: Science and Systems*, 2024.

[20] S. Misenti, B. Hertel, B. Weng, R. Donald, A. Jawaji, M.-T. Kosoko-Thoroddsen, J. G. Trafton, A. Norton, R. Azadeh, and Y. Gu, "Experimental evaluation of commercial quadruped robots: Stability and performance in non-inertial environments," *International Journal of Intelligent Robotics and Applications*, vol. 9, pp. 1109–1134, 2025.

[21] S. Samadi, J. Roux, A. Tanguy, S. Caron, and A. Kheddar, "Humanoid control under interchangeable fixed and sliding unilateral contacts," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 4032–4039, 2021.

[22] Q. Rouxel, S. Ivaldi, and J.-B. Mouret, "Multi-contact whole-body force control for position-controlled robots," *IEEE Robotics and Automation Letters*, vol. 9, 2024.

[23] J. Englsberger, C. Ott, and A. Albu-Schäffer, "Three-dimensional bipedal walking control based on divergent component of motion," *IEEE Transactions on Robotics*, vol. 31, no. 2, pp. 355–368, 2015.

[24] P. M. Wensing, M. Posa, Y. Hu, A. Escande, N. Mansard, and A. Del Prete, "Optimization-based control for dynamic legged robots," *IEEE Transactions on Robotics*, vol. 40, pp. 43–63, 2024.

[25] N. Rudin, D. Hoeller, P. Reist, and M. Hutter, "Learning to walk in minutes using massively parallel deep reinforcement learning," in *Proceedings of the 5th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, vol. 164, 2022, pp. 91–100.

[26] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[27] J. Schulman, P. Moritz, S. Levine, M. I. Jordan, and P. Abbeel, "High-dimensional continuous control using generalized advantage estimation," in *Proceedings of the 4th International Conference on Learning Representations (ICLR)*, 2016, arXiv:1506.02438.

[28] L. Pinto, M. Andrychowicz, P. Welinder, W. Zaremba, and P. Abbeel, "Asymmetric actor critic for image-based robot learning," in *Proceedings of Robotics: Science and Systems (RSS)*, Pittsburgh, PA, USA, 2018.

[29] O. Stasse, T. Flayols, R. Budhiraja, K. Giraud-Esclasse, J. Carpentier *et al.*, "TALOS: A new humanoid research platform targeted for industrial applications," in *2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids)*, Birmingham, UK, 2017.

[30] Unitree Robotics, "Unitree G1 humanoid agent," https://www.unitree.com/g1, 2026, accessed: 2026-10-20.

[31] G. A. Pratt and M. M. Williamson, "Series elastic actuators," in *Proceedings of the 1995 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, vol. 1, 1995, pp. 399–406.

[32] P. M. Wensing, A. Wang, S. Seok, D. Otten, J. Lang, and S. Kim, "Proprioceptive actuator design in the MIT Cheetah: Impact mitigation and high-bandwidth physical interaction for dynamic legged robots," *IEEE Transactions on Robotics*, vol. 33, no. 3, pp. 509–522, 2017.

[33] M. Chignoli, D. Kim, E. Stanger-Jones, and S. Kim, "The MIT humanoid robot: Design, motion planning, and control for acrobatic behaviors," in *2020 IEEE-RAS 20th International Conference on Humanoid Robots (Humanoids)*, 2021, pp. 1–8.

[34] X. Karekla, K. Gkiotsalitis, and N. Tyler, "The impact of a passenger-safety-driven acceleration limit on the operation of a bus service," *Accident Analysis & Prevention*, vol. 148, p. 105790, 2020.

[35] B. De Graaf and W. Van Weperen, "The retention of balance: An exploratory study into the limits of acceleration the human body can withstand without losing equilibrium," *Human Factors*, vol. 39, no. 1, pp. 111–118, 1997.

[36] Parsons Brinckerhoff Quade and Douglas, Inc., "Track design handbook for light rail transit," Transportation Research Board, National Research Council, Washington, DC, Tech. Rep. Transit Cooperative Research Program Report 57, 2000.

[37] J. Kennedy, C. Oakley, S. Sumon, I. Parry, E. Wilkinson, and J. Brown, "Impact of road humps on vehicles and their occupants," Transport Research Laboratory, Crowthorne, UK, Tech. Rep. TRL614, 2004.

A

Physical Trial Records

A.1 Trial Inventory

Table A.1 lists the identifiers of the physical trials of Table 7.1.

Table A.1: Retained physical trials by controller–support condition.

Condition	Straight-road trials	Turning trials
Default	S1–S5	T1, T4, T6, T7, T9, T10, T12
No-handle	S1–S10	T1, T3, T4, T6, T7, T9, T10, T12
Rigid-handle	S1–S10	T1–T13

A.2 Additional Physical Results

We report the per-trial values underlying the aggregate physical results in Chapter 7 in Tables A.2 and A.3. For every record, we report J_x^{95} as the 95th percentile of the whole-body state-deviation index and σ_{att} as the combined roll–pitch attitude variation defined in Section 5.3.6. We evaluate both measures over the central 80% of the overlapping vehicle–robot interval. We use a dash when a maneuver was not executed under a given condition; the dash does not represent a zero-valued result.

Table A.2: Per-trial response metrics for the retained straight-road tests.

	Default		No-handle		Rigid-handle	
Trial	J_x^{95}	σ_{att} (deg)	J_x^{95}	σ_{att} (deg)	J_x^{95}	σ_{att} (deg)
S1	3.9145	3.279	0.1957	0.856	0.0231	0.803
S2	4.7188	5.185	0.7999	0.873	0.4492	0.929
S3	5.2136	8.150	1.5607	1.939	0.0389	1.291
S4	5.0059	6.574	4.6293	1.889	0.0388	1.967
S5	3.9474	9.553	1.6896	2.627	0.0499	2.174
S6	–	–	1.2855	1.611	0.0211	1.008
S7	–	–	4.1924	1.900	0.0915	1.436
S8	–	–	3.5009	2.455	0.0759	1.754
S9	–	–	1.1747	2.584	0.2007	2.058
S10	–	–	1.6885	2.778	0.1979	2.261

Table A.3: Per-trial response metrics for the retained turning tests. The radius is the nominal skidpad radius specified by the test plan.

		Default		No-handle		Rigid-handle	
Trial	Radius (m)	J_x^{95}	σ_{att} (deg)	J_x^{95}	σ_{att} (deg)	J_x^{95}	σ_{att} (deg)
T1	8	3.6481	3.796	2.0252	1.111	0.1179	1.399
T2	8	–	–	–	–	0.3117	2.007
T3	8	–	–	1.1126	0.979	0.4885	3.906
T4	10	2.5964	3.057	1.9419	1.350	0.0244	1.654
T5	10	–	–	–	–	0.1288	2.428
T6	10	5.4777	3.462	1.8855	1.276	0.2839	2.935
T7	12	0.6108	1.923	1.5857	1.364	0.1762	1.324
T8	12	–	–	–	–	0.1735	2.081
T9	12	0.2425	2.407	2.1077	1.566	0.2341	2.608
T10	20	0.0212	0.590	1.5852	1.204	0.1306	1.098
T11	20	–	–	–	–	0.1595	1.865
T12	20	0.9523	3.155	2.1310	2.404	0.1177	1.948
T13	20	–	–	–	–	0.6930	6.475

A.3 Time Histories of All Trials

Figures A.1–A.18 show the vehicle input and the robot response of every retained trial except S4 and T12, which are shown in Figures 7.2 and 7.3. Each column shows one condition, or one identifier where only the rigid-handle configuration was tested. The rows show the vehicle speed, the longitudinal acceleration a_x for straight-road trials or the lateral acceleration a_y for turning trials, torso roll and pitch relative to their median over the first 2 s, and J_x on a logarithmic scale. The lateral acceleration is shown with the sign that makes it positive in the turn. The rigid-handle S1 record contains no stationary samples; its acceleration offset is therefore taken as the median offset of the other rigid-handle trials, which were recorded with the same sensor mounting.

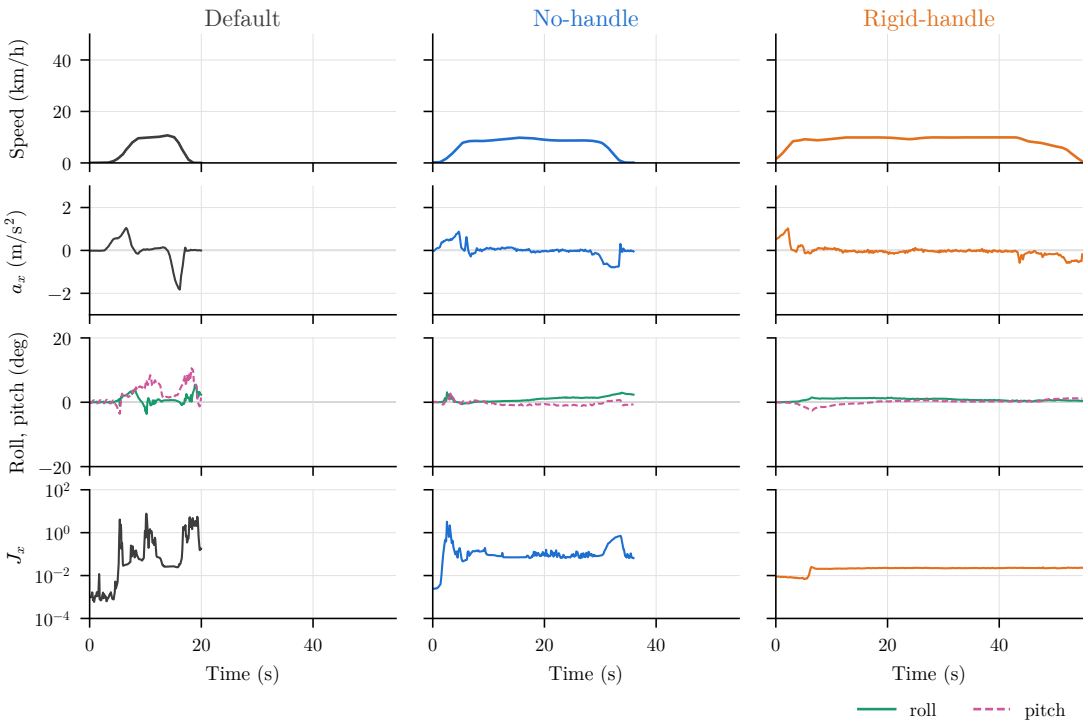


Figure A.1: Time histories of straight-road trial S1.

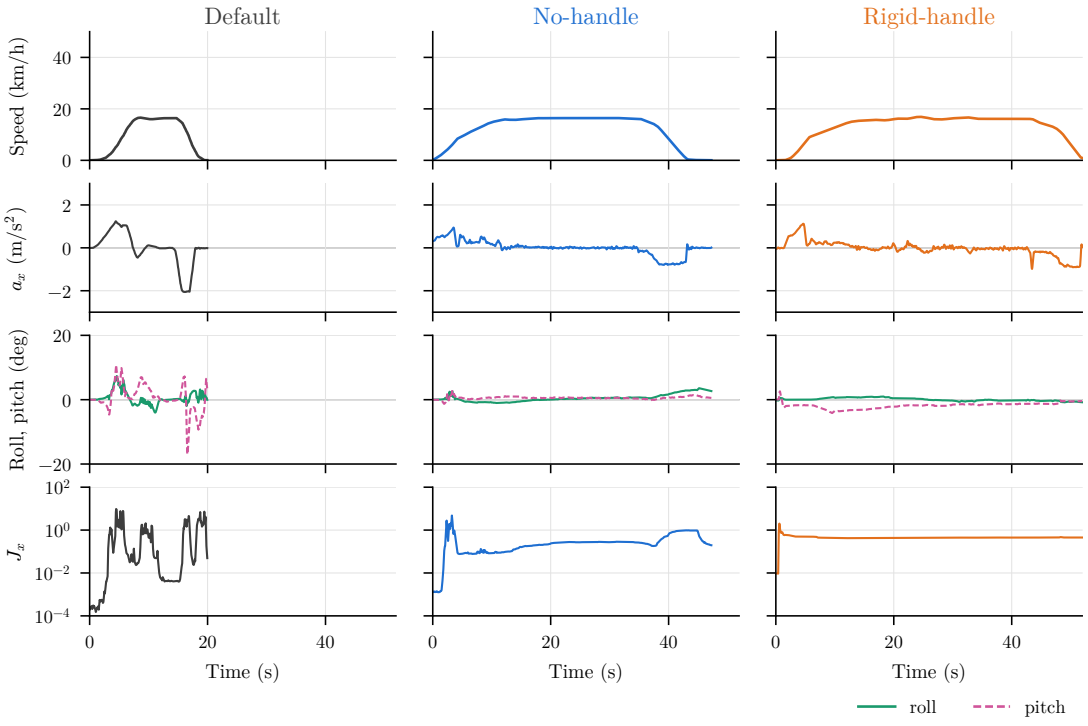


Figure A.2: Time histories of straight-road trial S2.

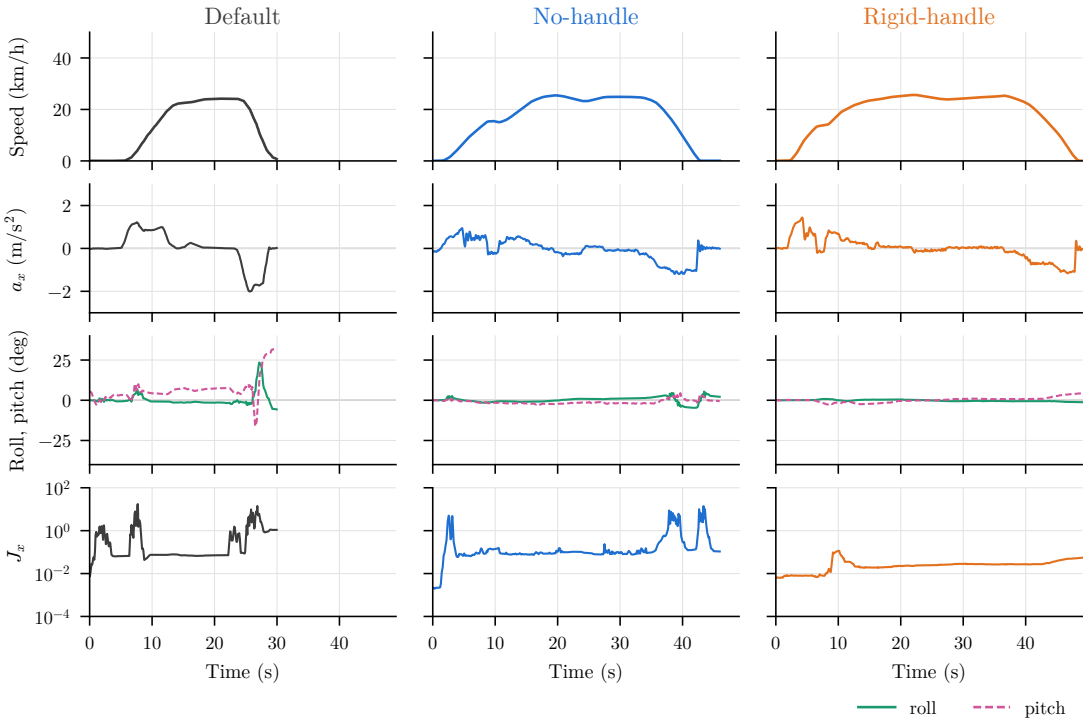


Figure A.3: Time histories of straight-road trial S3.

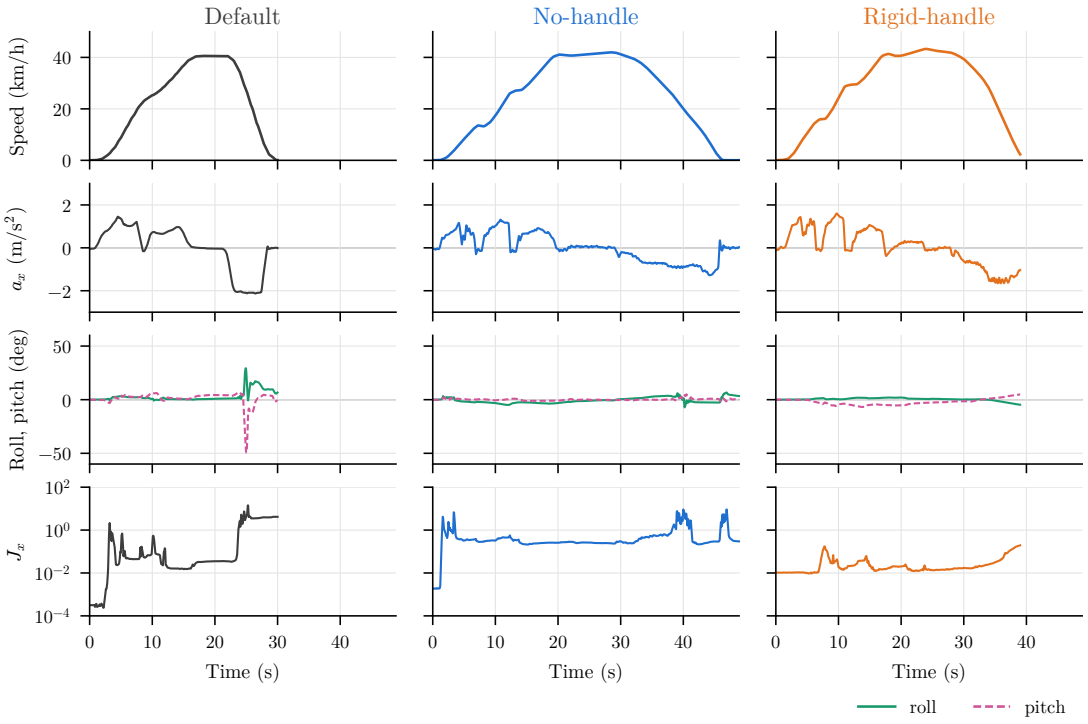


Figure A.4: Time histories of straight-road trial S5.

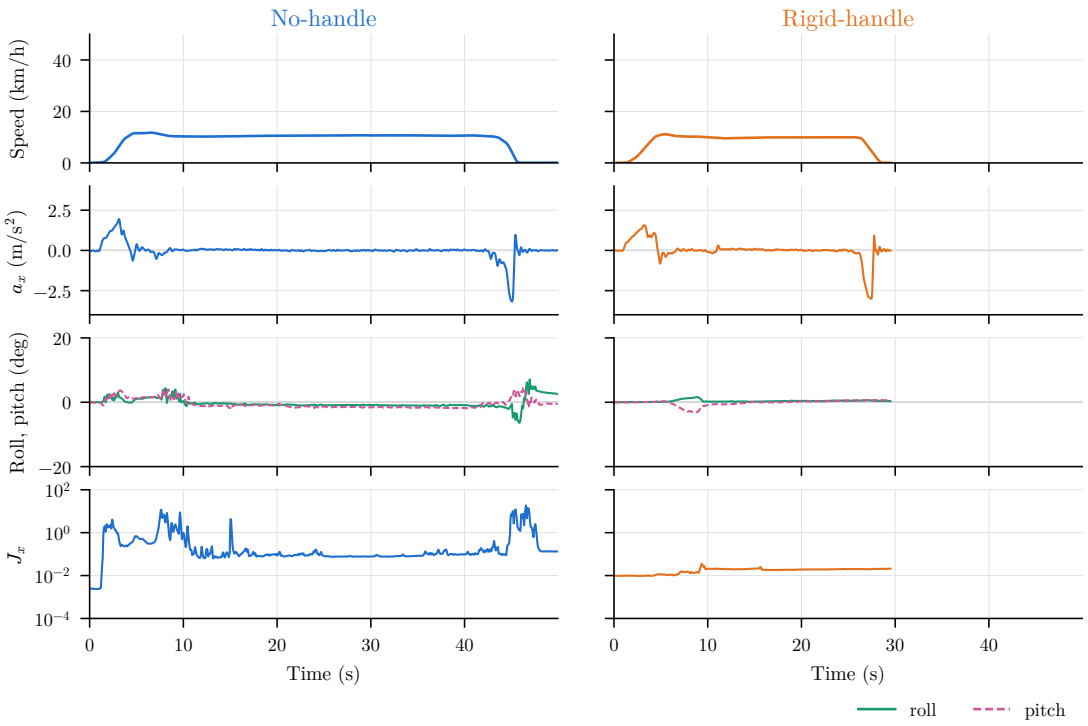


Figure A.5: Time histories of straight-road trial S6.

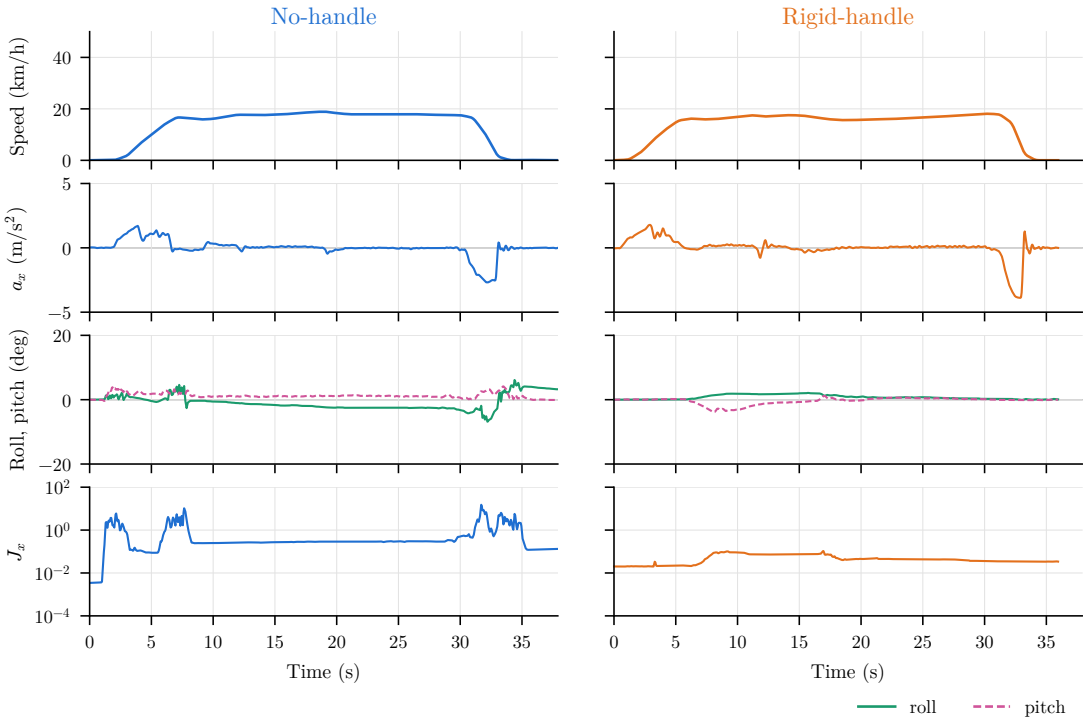


Figure A.6: Time histories of straight-road trial S7.

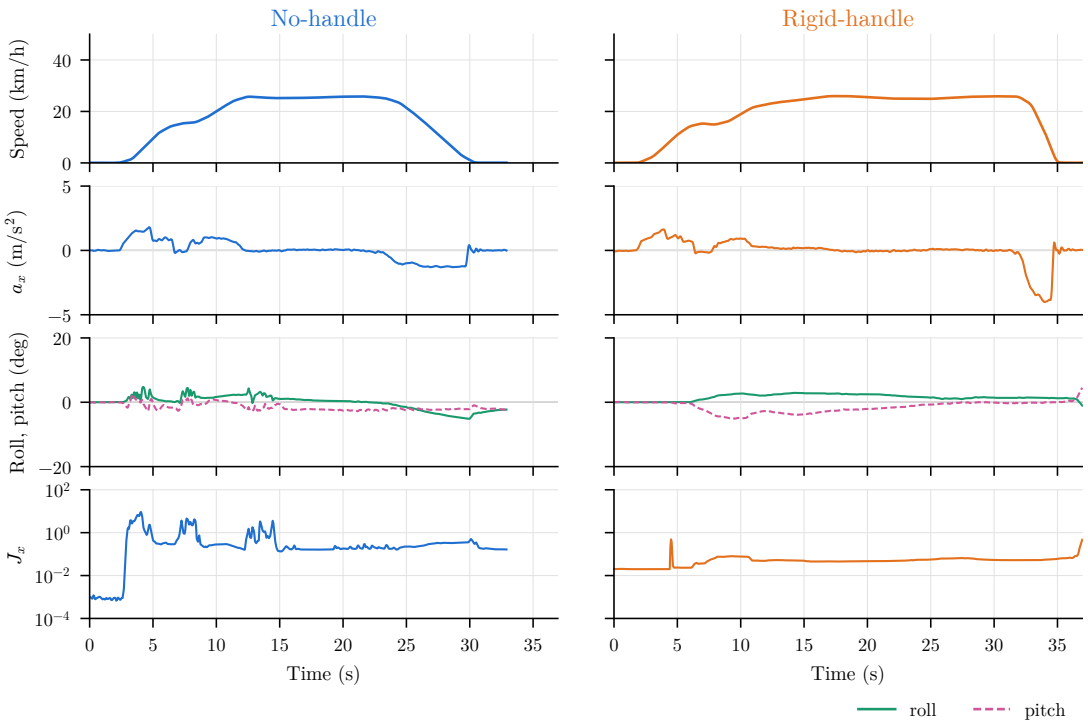


Figure A.7: Time histories of straight-road trial S8.

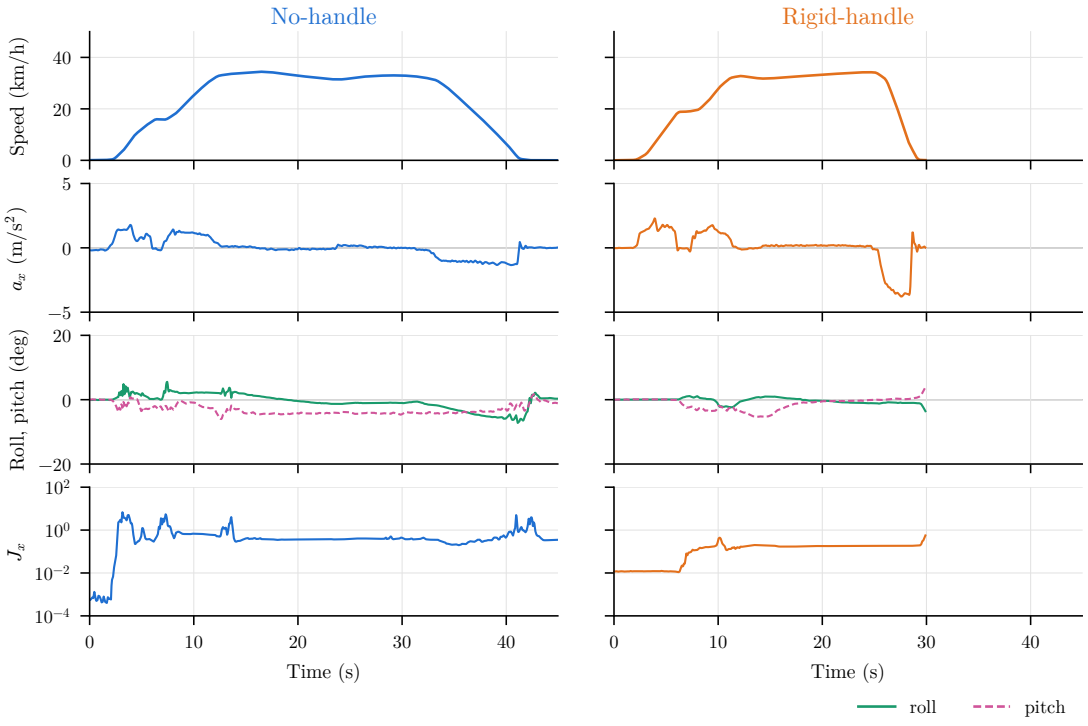


Figure A.8: Time histories of straight-road trial S9.

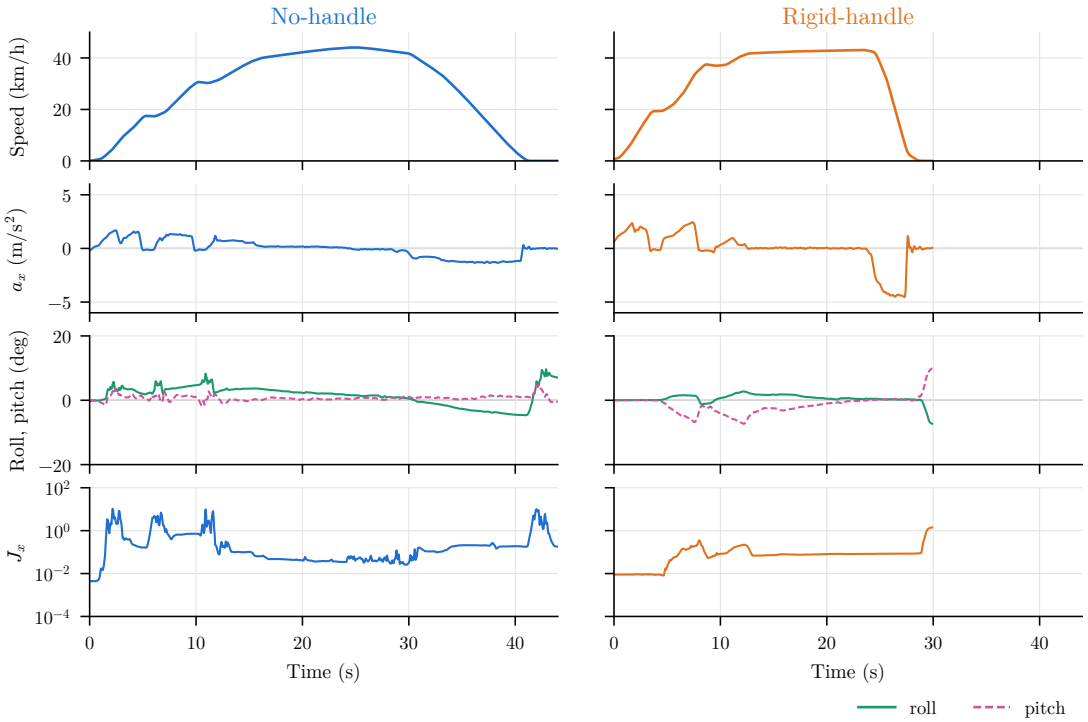


Figure A.9: Time histories of straight-road trial S10.

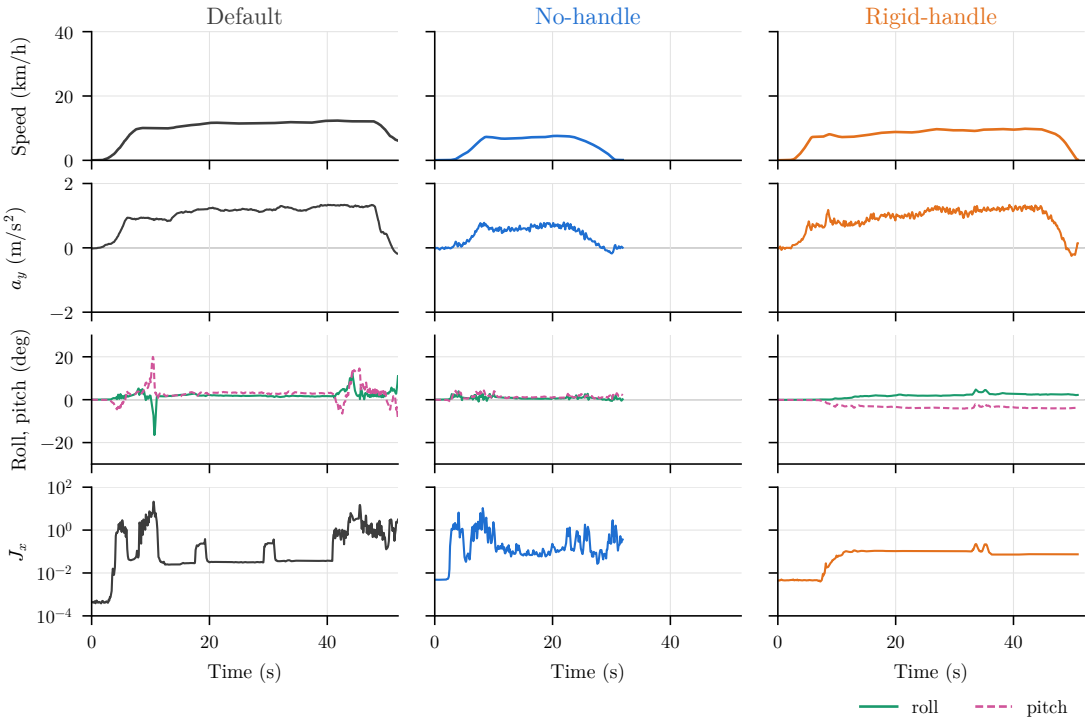


Figure A.10: Time histories of turning trial T1 (nominal radius 8 m).

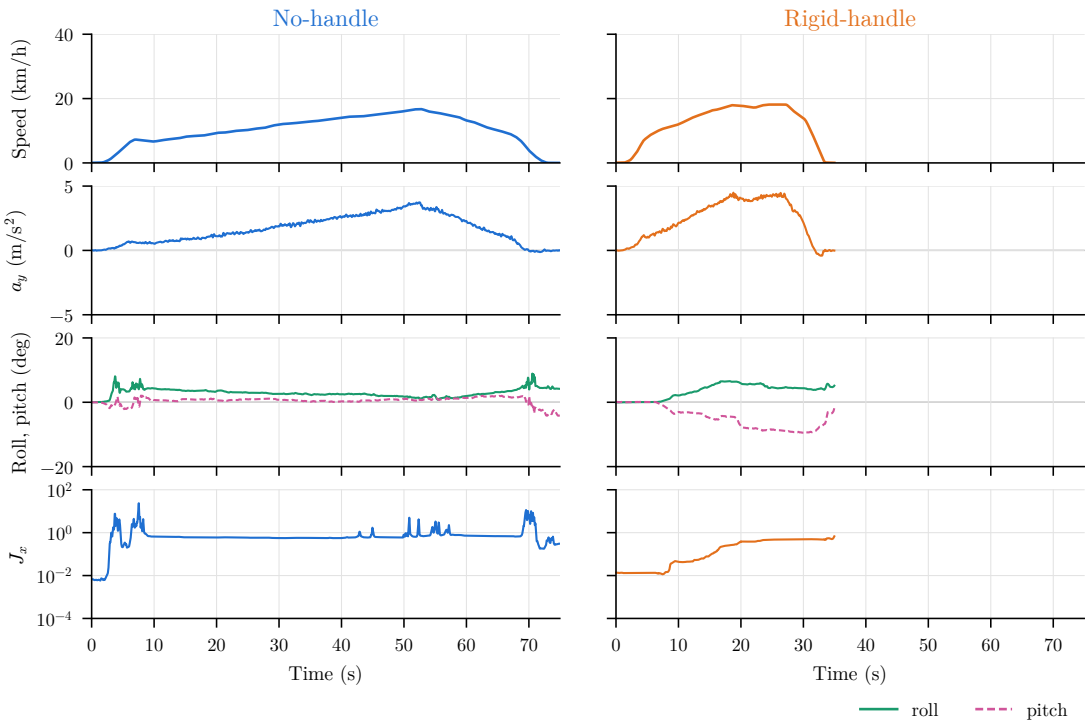


Figure A.11: Time histories of turning trial T3 (nominal radius 8 m).

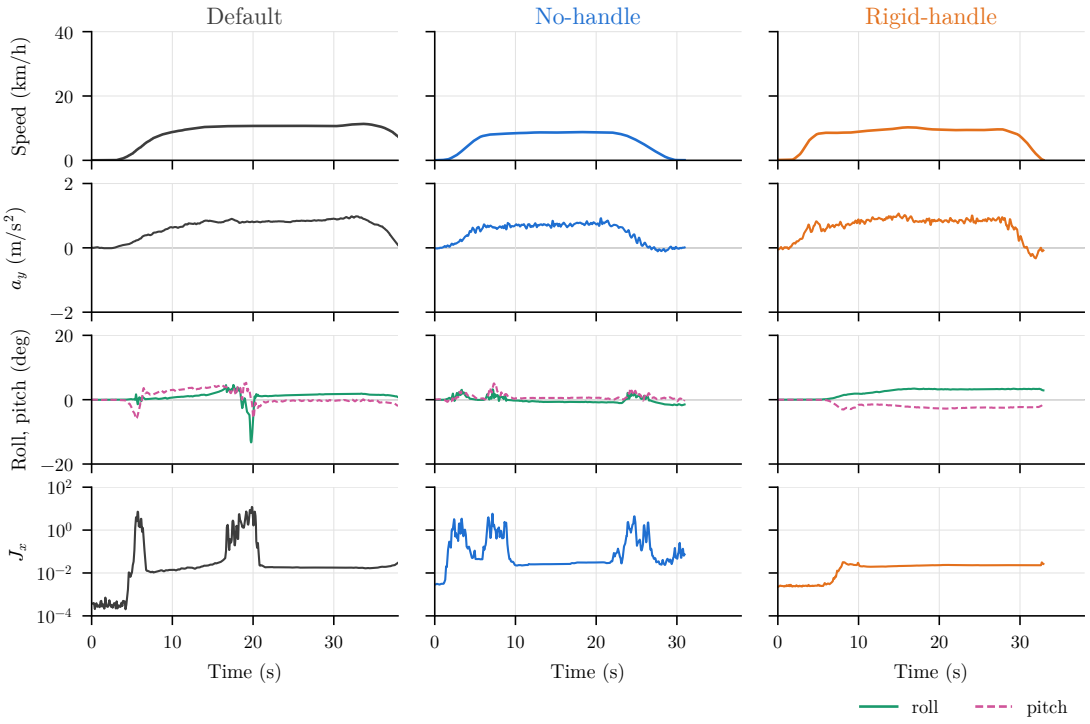


Figure A.12: Time histories of turning trial T4 (nominal radius 10 m).

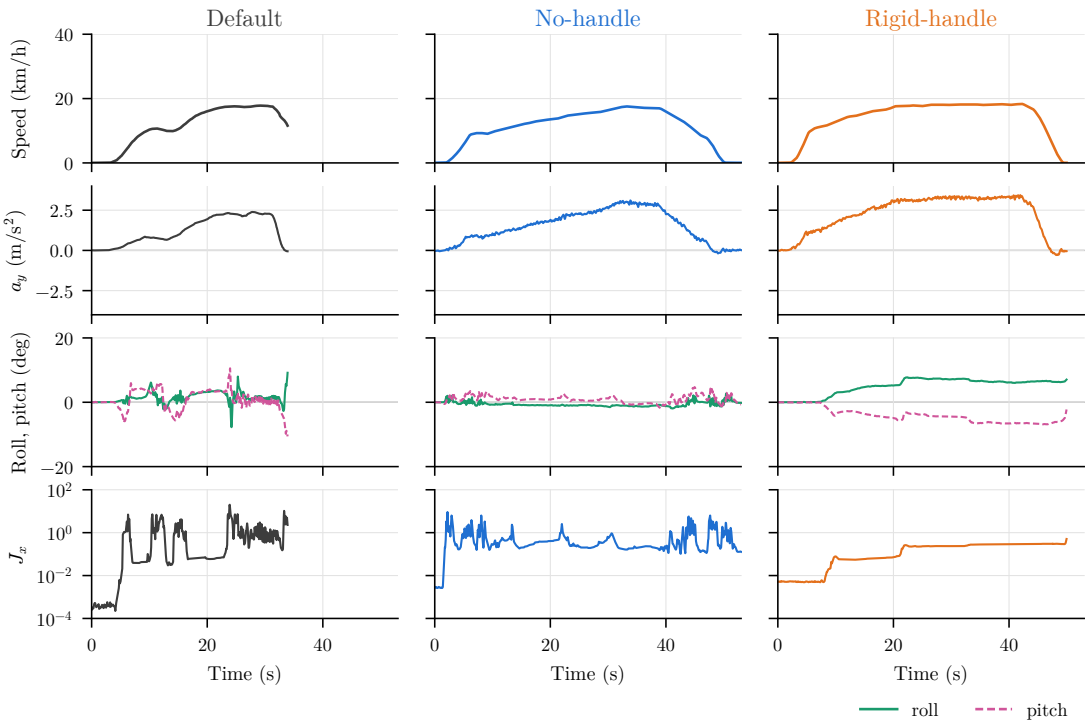


Figure A.13: Time histories of turning trial T6 (nominal radius 10 m).

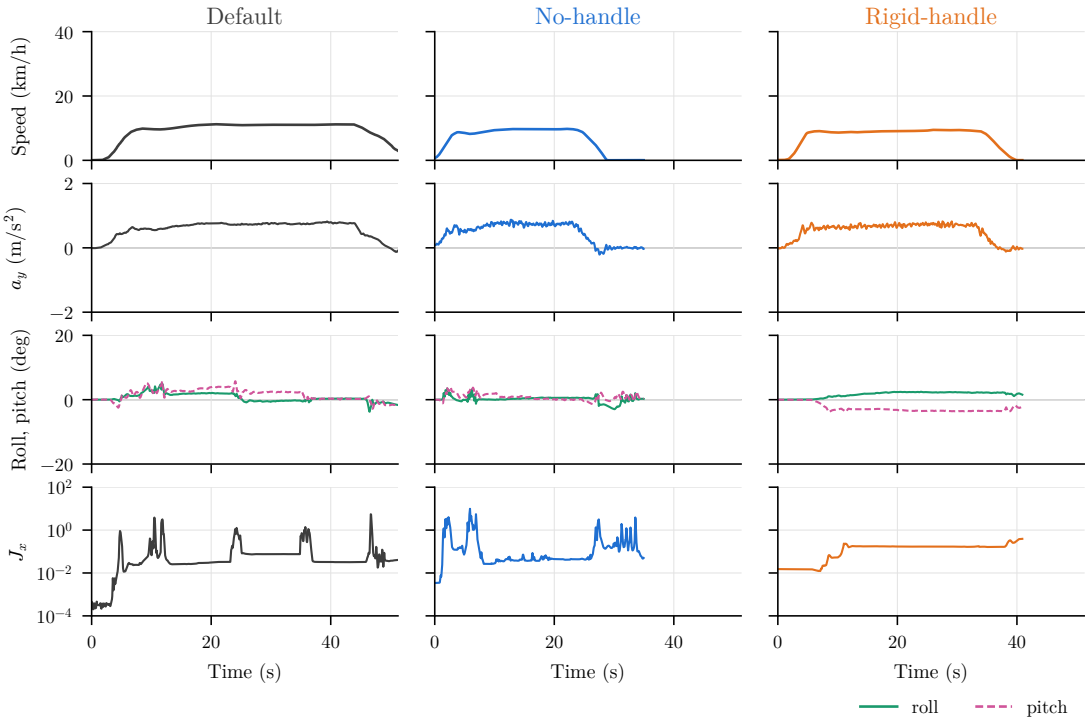


Figure A.14: Time histories of turning trial T7 (nominal radius 12 m).

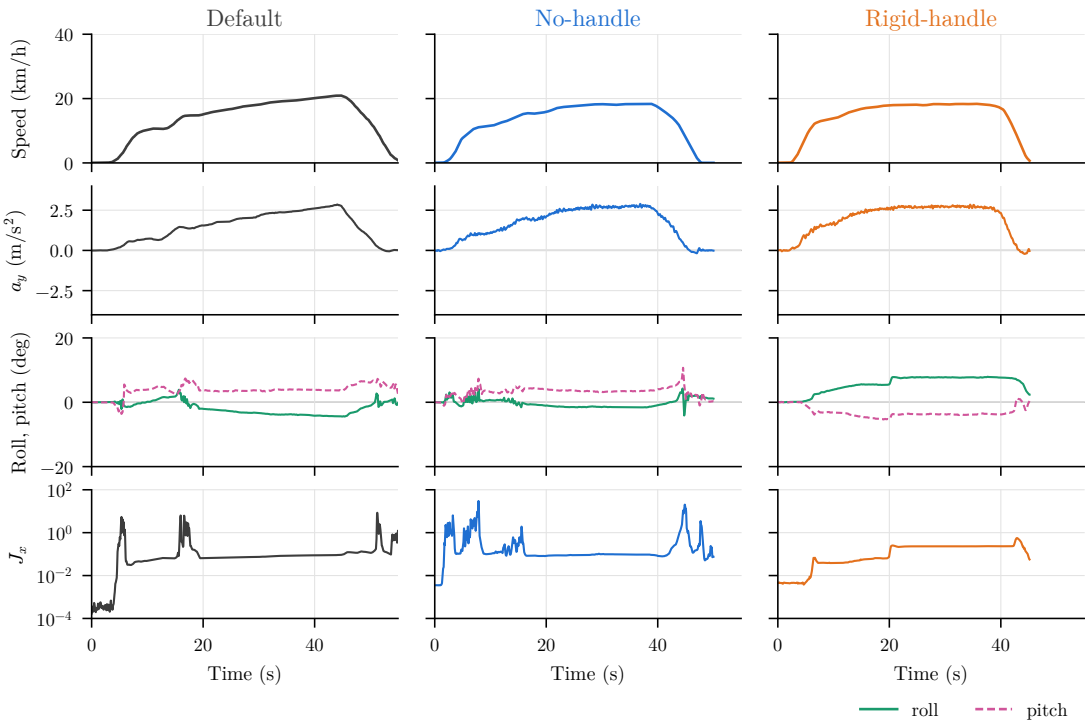


Figure A.15: Time histories of turning trial T9 (nominal radius 12 m).

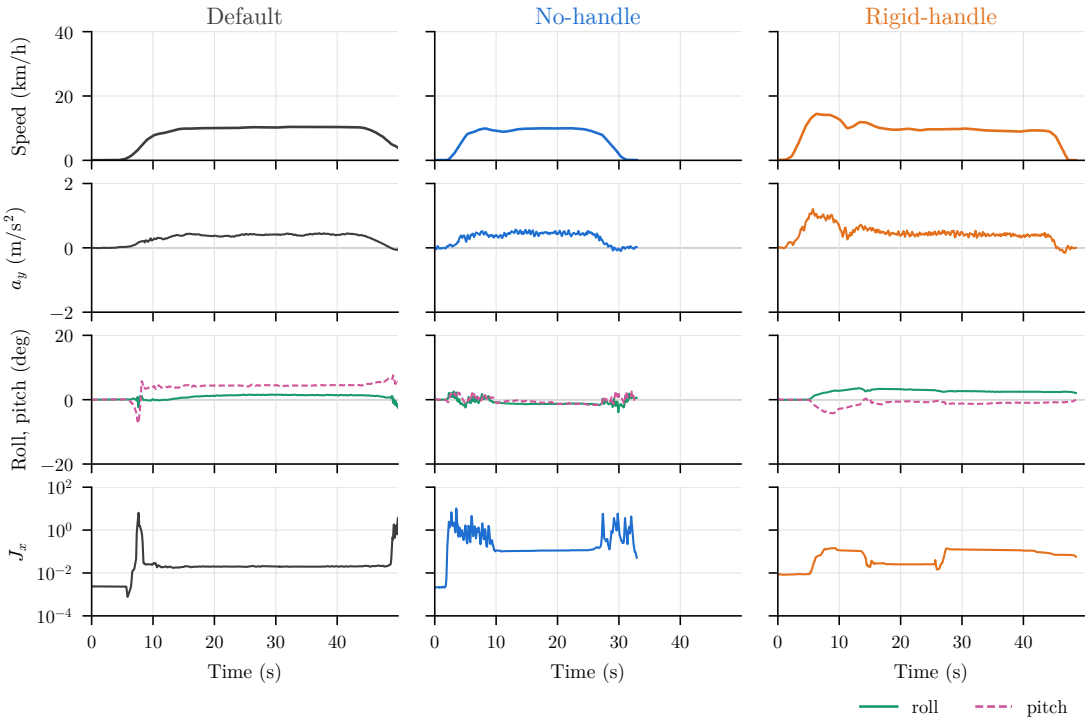


Figure A.16: Time histories of turning trial T10 (nominal radius 20 m).

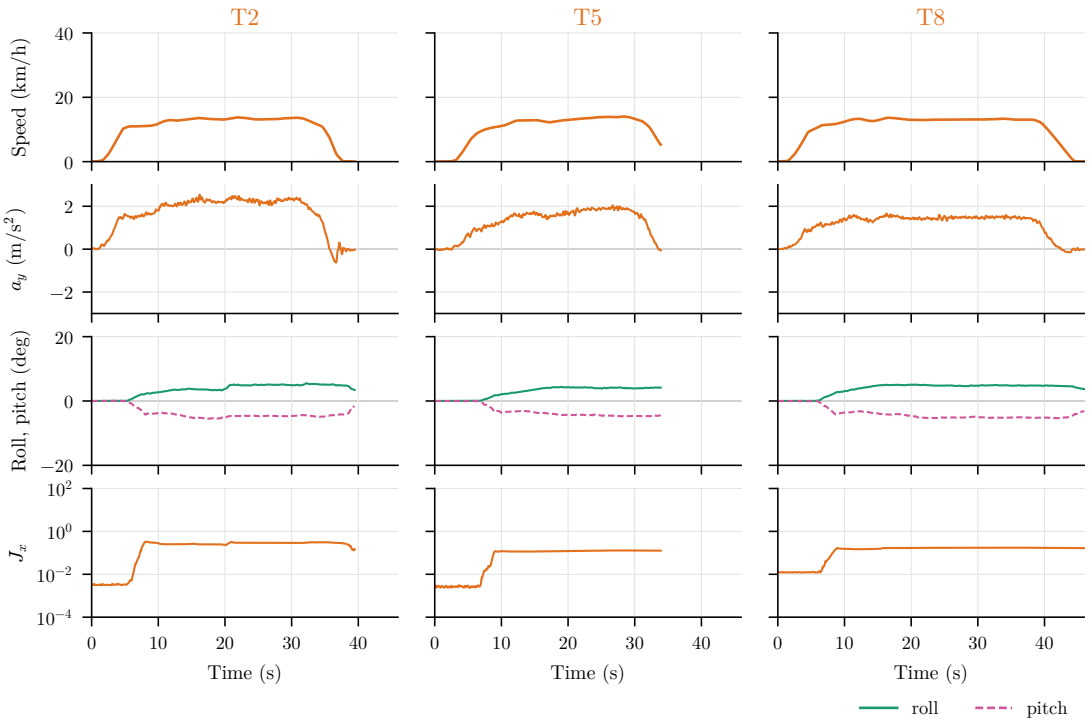


Figure A.17: Time histories of the rigid-handle turning trials T2, T5, and T8 (nominal radii 8, 10, and 12 m).

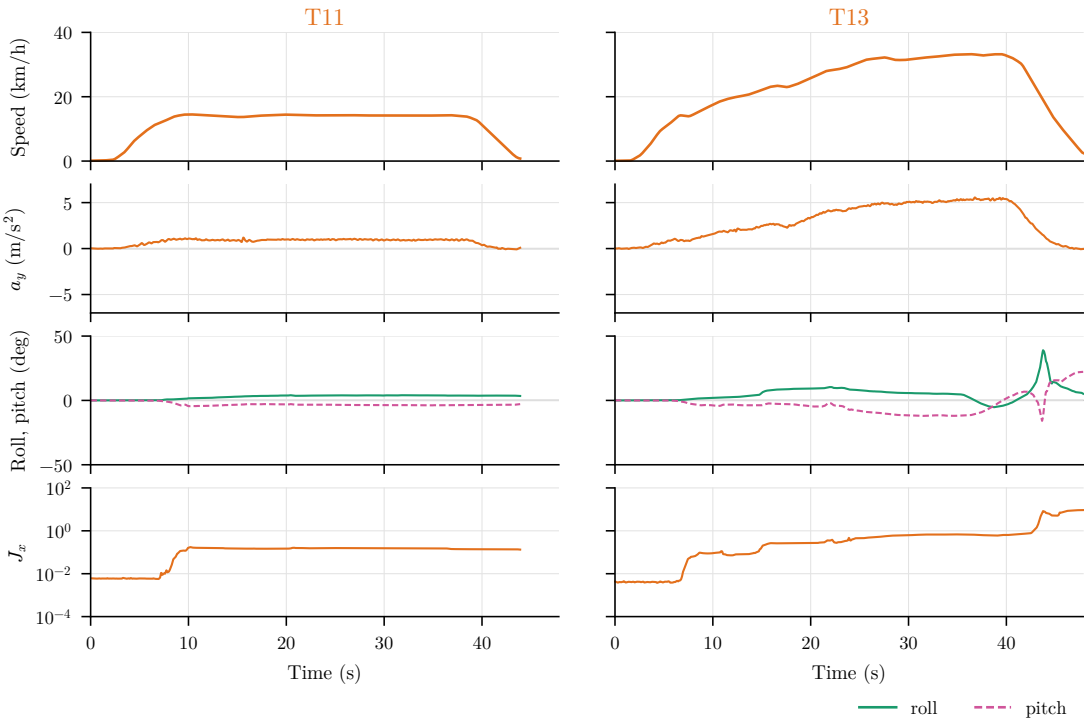


Figure A.18: Time histories of the rigid-handle turning trials T11 and T13 (nominal radius 20 m).

DEPARTMENT OF MECHANICAL ENGINEERING
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2026
www.chalmers.se



CHALMERS
UNIVERSITY OF TECHNOLOGY