

Whole-Body Dexterity

Conor Mc Gartoll^{*1}, An Dang^{*1,2}, Pranay Thangeda^{*1}, Arjun Gupta¹, Carolina Higuera¹,
Nima Fazeli^{1,2}, Mustafa Mukadam^{†1}, Manikantan Nambi^{†1}

¹Amazon, ²University of Michigan

^{*}Equal contribution, [†]Equal advising

<https://wholebodydexterity.com>

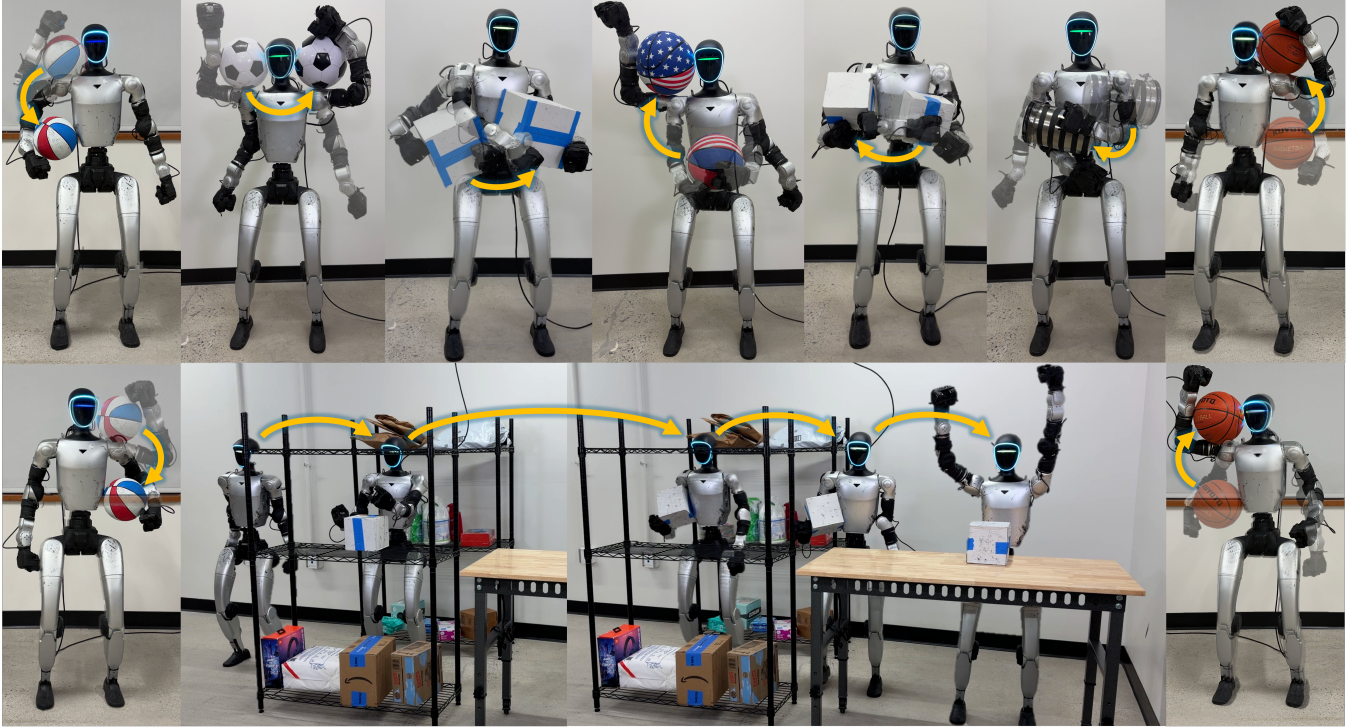


Fig. 1: **Whole-Body Dexterity (WBD)** extends humanoid dexterous manipulation beyond the hands, allowing humanoids to emulate humans’ ability to use their arms and torso as contact surfaces for holding and repositioning objects. This enables tasks such as lifting a box, securing it under one arm, carrying it while walking, and setting it down.

Abstract—Current humanoid dexterous manipulation remains largely confined to end-effectors and simple grasps, falling short of humans’ ability to engage the entire body by bracing and sliding bulky objects against the arms and torso. We introduce Whole-Body Dexterity (WBD), a general framework that uses the robot’s body and appendages as contact surfaces to grasp and manipulate objects. Our sim-to-real pipeline first generates a cache of feasible whole-body grasps, then trains reinforcement-learning experts on a general any-to-any grasp objective by sampling initial and goal grasps from the cache, and finally distills these experts into a single deployable policy. We show that the distilled policy transfers zero-shot to a real Unitree G1, where it repositions a basketball and a cardboard box to regions around the torso and completes a long-horizon task that stows a box under one arm, carries it between shelves, and places it. The controller operates interactively, steered by coarse motion references from a human teleoperator. In simulation, a similar controller refines corrupted expert commands into stable manipulation. Across real-world whole-body manipulation tasks, the distilled controller completes 90% of trials, compared with 18% for direct retargeting and 4% for a generalist whole-body controller.

I. INTRODUCTION

People rarely manipulate with their hands alone. We brace a box against the chest, cradle a basket across both arms, or tuck an object under one arm to free the other. These actions use the limbs and torso as a single highly dexterous manipulator. Humanoids need the same flexibility to carry an object while interacting with the environment or handling another object. We call this ability to reposition objects by continually changing supporting body contacts *whole-body dexterous manipulation* (Fig. 1). Prior humanoid work has used the term for specific settings, such as coordinated whole-body control with dexterous hands [1], [2], while related loco-manipulation confines dexterity to the end-effectors [3], [4], [5], [6], [7] or to a sustained arm grasp on a bulky object, whether carried during locomotion or reoriented on a table [8], [9], [10], [11]. None treats it as a general capability in which contact is established and released supported by any link to move an object between diverse regions of the body.

Under this view, whole-body dexterous manipulation has

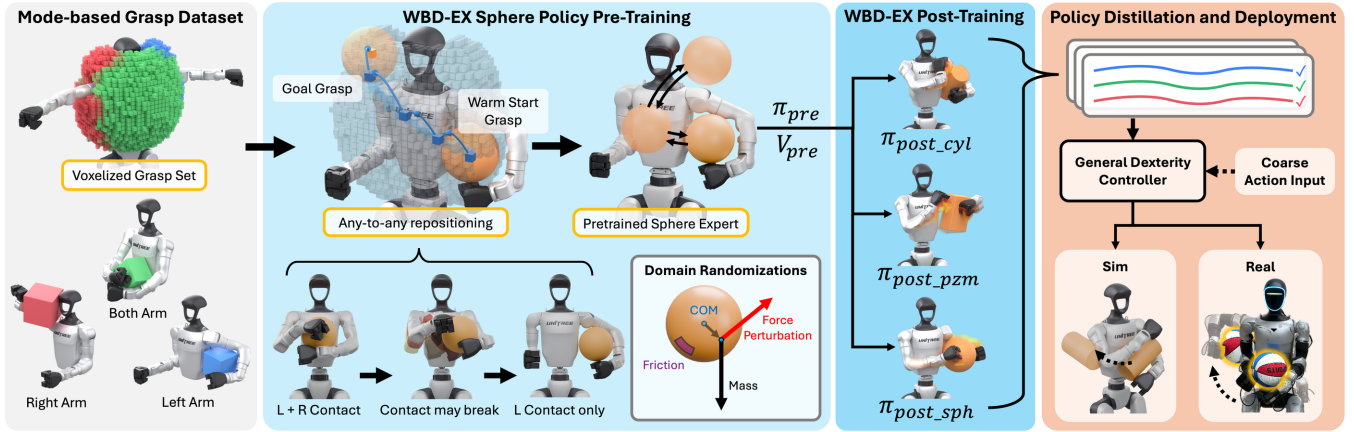


Fig. 2: **Whole-Body Dexterity Pipeline.** (1) Generate a voxelized dataset of stable whole-body grasps. (2) Pre-train a sphere WBD-EX expert on any-to-any repositioning using the cached grasps. (3) Warm-start WBD-EX experts for other object primitives from the sphere expert. (4) Collect WBD-EX trajectories and distill them into a general dexterity controller for simulation and real-world tasks.

a useful parallel in dexterous hands, with the limbs and torso taking the roles of fingers and palm. Recent foundation controllers for dexterous hands cast complex manipulation as sequences of learned grasp-to-grasp transitions [12]. Extending this principle to humanoids introduces a challenge: a dexterous hand with many fingers can often reposition some fingers while the others support the object [13], [14], whereas a humanoid using two arms has fewer independent supports. Releasing one arm may require the other to brace the object against the torso, and changing grasps can require rapid reconfiguration and rolling or sliding contact. These contact requirements also complicate teleoperation: retargeting human motion does not ensure that the robot establishes the contacts needed to hold the object [15], [16]. This additional complexity compels a different approach to repositioning across regions of the body than previous grasp-to-grasp methods.

We introduce **Whole-Body Dexterity (WBD)**, a framework for learning whole-body dexterous manipulation through *mode-constrained re-grasping* (Fig. 2). We first generate and validate diverse whole-body grasps from specified contact configurations, then use them to initialize reinforcement-learning experts, **WBD-EX**. Each goal pairs an object position with a target grasp mode: a left-arm, both-arm, or right-arm hold in our humanoid implementation. Transitions allow brief contact loss during movement but require a stable hold in the target mode at goals, so the experts learn both to move the object and to hand it off between arms, enabling repositioning across difficult regions such as behind the back, on top of the shoulders, and under the arms. A shared reward formulation trains experts across object primitives under randomized dynamics. We then distill their trajectories into a general controller that uses proprioceptive history and an object geometry descriptor to refine coarse motion commands into stable whole-body object repositioning, without privileged object state.

We evaluate WBD in simulation and on a real Unitree G1. Simulation experiments assess command refinement from

intentionally noised object repositioning experts and transfer to object sizes withheld from training. Real-world experiments first demonstrate whole-body dexterous capabilities such as transferring a basketball or cardboard box under an arm or raising it to a shoulder and also showcase a multi-stage shelf pick, stow, carry, and place task. Across these three tasks WBD completes 45 of 50 trials, against 9 of 50 for a retargeting teleoperation baseline and 2 of 50 for the evaluated whole-body controller. Within WBD, our main contributions are:

- 1) A whole-body grasp generation method defined over mode-typed contact configurations, formulated for arbitrary open-link articulated rigid-body systems and demonstrated on a humanoid robot.
- 2) **WBD-EX**, per-object repositioning experts built on a mode-constrained re-grasping formulation that uses the diverse grasp set.
- 3) A distillation procedure that turns these experts into a controller deployable without privileged object state, refining a coarse operator command stream into whole-body dexterous manipulation on a real humanoid.

II. RELATED WORK

Whole-Body Manipulation. In this paper, we advance whole-body dexterous manipulation, in which objects are held and repositioned through contact anywhere on the robot’s body rather than only at designated end-effectors or links. The closely related term, whole-body dexterity, was introduced for a snake-like robot arm that manipulates with its entire body surface [20]. Related in spirit, a humanoid can manipulate from onboard proprioception alone, since joint-encoder readings under compliant contact act as a coarse whole-body tactile channel [21]. A second group manipulates bulky objects that exceed the hand’s size and payload by enveloping them against the arms and torso and distributing contact across the body [8], [22], [23], [24]. Other work uses whole-body tactile sensing to stay balanced while making contact and, in simulation, tilts a box 90° on a table through

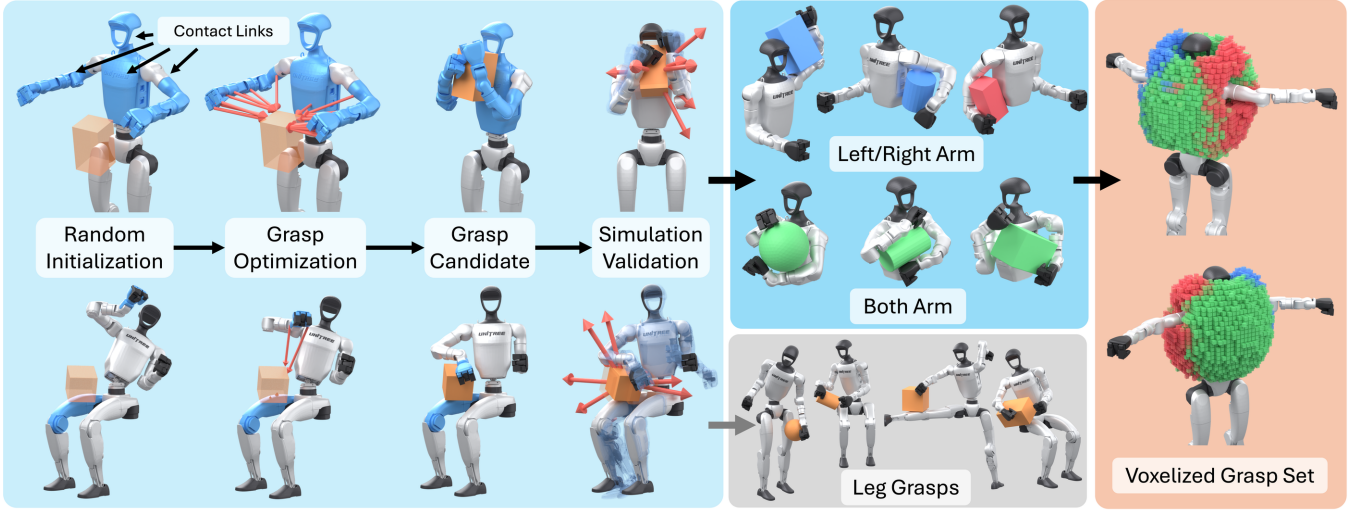


Fig. 3: **Whole-Body Grasp Generation.** Energy-based optimization [17], [18], [19] generates grasp candidates from specified arm, leg, and torso contact links (blue), followed by simulation validation under perturbations. The voxelized training set contains left-arm, right-arm, and both-arm grasps; leg grasps are excluded.

changing contacts [9]. Closest to our setting, reinforcement learning on a soft upper-body humanoid lifts cuboids of varying sizes and masses and performs recovery re-grasping in front of the torso [25], but without standing or walking and without moving the object to other regions of the body.

Classical model-based work formulates this as optimization over contacts placed anywhere on the body surface [26], [27] and plans holding manipulation as a transition graph over object motion and contact states [28], though typically under planar, quasi-static, or box-shaped assumptions; recent learning-based methods instead expose whole-body contact as an explicit, commandable interface [29] or imitate whole-body human-object interactions in simulation from motion-capture demonstrations [30], [31]. Evidence that the hand and the humanoid form a structural continuum comes from re-configurable hardware [32] and from commercial humanoids that already cradle baskets, drape cloth over a shoulder, and lift appliances during everyday tasks [33], [34]. Humanoid loco-manipulation couples locomotion with manipulation, enabling dynamic behaviors such as climbing and wall flips alongside picking, placing, and pushing objects [5], [4], [35], [36], [37], [38]. We instead focus on general object manipulation across body regions.

Teleoperation is a direct route to whole-body manipulation and is well developed as motion transfer, through real-time human-to-humanoid retargeting and shadowing [15], [16], [39], [40], [41] and tracking controllers that follow whole-body targets [42], [3], [43]. But without force or tactile feedback it is difficult for the operator to tell whether an object is pinned or slipping, leading to mainly hand-only pick-and-place behavior.

Dexterity as Contact Sequencing. In-hand manipulation can be framed as re-grasping: an object is carried through a sequence of grasps by repeatedly breaking and remaking finger contacts, a process known as finger gaitting [14], [13], [44]. Humans perform the body-scale analog when they transfer an object between their own hands to reconfigure the

grasp for better manipulation [45]. A parallel line pretrains a reusable controller once and then prompts it with coarse commands. For a dexterous hand, prior work pretrains a foundational manipulation controller in simulation that high-level policies or teleoperators steer with coarse motion commands [12]. Related goal-conditioned approaches learn grasping and reorientation across procedurally generated tools [46] or pretrain object reposing before adapting to downstream manipulation tasks [47]. Similar in mechanism to our own approach, a constrained sampler over feasible contact-rich states can guide goal-conditioned manipulation RL directly [48]. We lift this re-grasping principle from dexterous hands and arms to the whole humanoid.

III. METHODOLOGY

We present a pipeline for whole-body dexterous manipulation, enabling object repositioning across diverse regions of the body. Beyond end-effector and front-of-torso contact, we define contact surfaces for grasping as any link or link subsection in the robot’s kinematic system. Because the embodiment for whole-body grasping differs greatly from the multifinger hand, much of the regrasping in the whole-body formulation involves sliding and rolling across the torso mesh, which we account for later in this section.

Our method trains and deploys controllers with dexterous whole-body regrasping capabilities in three stages. We first generate a large corpus of whole-body object grasps using energy-based optimization and decompose our torso mesh into subregions to enable diverse whole-body grasping on the robot. Second, we train per-object experts on this grasp corpus to reposition objects around the body, encouraging sliding and rolling contact across the torso. Finally, we distill these experts into deployable controllers without privileged information, which allows real-world whole-body dexterous object repositioning.

A. Whole-Body Grasp Generation

Given a kinematic system, we enumerate *configuration types* (Fig. 3): which bodies or appendages hold the object and which of their links or sub-regions should contact it. A *grasp mode* $m \in \mathcal{M}$ abstracts a configuration type to the appendages contacting the object, and is the level at which goals are commanded in Section III-B. The system model, an object mesh, and the choice of contact links are the only inputs, so the procedure is not specific to a humanoid or to a class of objects.

A large body such as the G1’s torso is contactable from every side, so optimizing against it as one contact target both concentrates contacts wherever they are easiest to find and collapses the distinct directions from which the object can be braced. We therefore subdivide large links into separate contact targets and generate grasps against each independently, which spreads the resulting grasps over the full extent of the body. On the G1 this subdivision yields named torso regions, which the downstream training reuses to name goals, but the mechanism is generic. We showcase a variety of whole-body grasps across arms, legs, and the torso in Fig. 3. Leg grasps would require learning manipulation jointly with balance and locomotion, so we use only arm and torso grasps for WBD-EX training, giving the mode set $\mathcal{M} = \{L, B, R\}$: left-arm, both-arm, and right-arm holds. In a single-arm mode the other arm must not be used for object support; in mode B both arms must contact it.

After defining our desired contact links and body sub-regions, we minimize the differentiable force-closure residual [17] over the object’s signed distance field, drawing contacts onto the object surface and keeping links out of the object and each other. Candidates are initialized from randomized object placements and optimized in GPU batches as in large-scale grasp dataset generation [18], [19], [49].

Due to having only two arms with a large torso to grasp against, G1 grasps depend heavily on gravity and contact link friction. Therefore, we aggressively filter candidates in simulation. Each grasp candidate object is pressed into contact and held under $4\times$ gravity while a sustained wrench disturbance is swept over the six axis directions, and passes only if contact persists; survivors must then drift less than 10 cm and rotate less than 45° under standard gravity without disturbance, giving the settled pose used at episode reset. We index accepted grasps by robot joint position, grasp mode, and settled object position in the torso frame, and drop grasps whose object mesh sits farther than 5 mm from the torso surface or below a size-dependent height threshold (Appendix II-D). This concentrates the set on tightly braced holds covering the body and provides WBD-EX with higher-quality grasps to transition between.

B. WBD-EX Training

We train **WBD-EX** to continually transition between object positions and grasp goals using proximal policy optimization (PPO) [50]. We formulate this problem as a Goal-Conditioned Markov Decision Process (gcMDP) of tuple $(\mathcal{S}, G, \mathcal{A}, \mathcal{R}, \mathcal{P}, \rho_0, \gamma)$ where $\mathcal{S}$ is the state space, G is the

goal set, $\mathcal{A}$ is the action space, $\mathcal{R}$ is the reward function, $\mathcal{P}$ is the transition probability, ρ_0 is the initial state distribution, and γ is the discount factor. The robot seeks a policy $\pi_\theta(\mathbf{a} \mid \mathbf{s}, \mathbf{g})$, parameterized by θ , that maximizes the expected discounted return $\mathcal{J} = \mathbb{E}_{\pi_\theta} [\sum_{t=0}^{\infty} \gamma^t r_t]$ over control steps t , where $\mathbf{s} \in \mathcal{S}$, $\mathbf{g} \in G$, $\mathbf{a} \in \mathcal{A}$, and $r_t = \mathcal{R}(\mathbf{s}_t, \mathbf{a}_t, \mathbf{g})$.

Given the grasp set $\mathcal{D} = \{(q_i, \mathbf{X}_i^O)\}_{i=1}^{N_g}$ for N_g grasps of an object, where q_i is the settled robot joint configuration and $\mathbf{X}_i^O$ is the settled object pose (position and orientation) expressed in the torso frame, we bin grasps into 2 cm voxels by settled object position and group each voxel’s grasps by grasp mode. Each occupied (voxel, mode) pair is a node $n \in \mathcal{V}$ of a goal graph, and $\mathcal{D}(n) \subseteq \mathcal{D}$ denotes the grasps in that voxel with that mode. A transition between nodes is legal only if their modes are compatible: L and R each connect to themselves and to B , but never to each other, so every handoff between arms passes through a both-arm hold.

We initialize each episode at a cached grasp by sampling a mode uniformly, then a voxel carrying that mode uniformly, then $g^{\text{start}} \sim \text{Uniform}(\mathcal{D}(n^{\text{start}}))$. Goals are commanded toward a *far goal* whose mode, then voxel, are drawn uniformly. Each commanded goal advances toward it: we draw a goal distance $d \sim \mathcal{U}(2 \text{ cm}, d_{\text{max}})$ and command the legal-mode node within d of the current node that is closest to the far goal, excluding the current node, with $g^{\text{goal}} \sim \text{Uniform}(\mathcal{D}(n^{\text{goal}}))$. A new far goal is drawn on arrival. With probability 0.2 the command is instead a *hold*, which asks the policy to keep the object at its current position and mode. To encourage human-like repositioning, in which objects roll and slide across body regions between arm handoffs, success depends only on the object’s position, its grasp mode, and the object coming to rest there. This differs from [12] in that our goals and rewards do not include the goal joint configuration or object orientation.

We apply a goal-distance curriculum by ramping the ceiling d_{max} from 2 cm to 8 cm. Because each goal distance is drawn from $[2 \text{ cm}, d_{\text{max}}]$, increasing d_{max} introduces progressively farther goal transitions while retaining closer ones. This exposes the policy to both gradual and more aggressive object repositioning behaviors as training progresses.

A goal is completed only when the object settles at the goal under the correct grasp mode (Appendix III-C). For a single-arm goal the inactive arm must register no object contact at the goal position. Requiring the inactive arm to fully release pushes the policy to learn stable and complete arm-to-arm handoffs around the body. We impose object-contact requirements throughout, but permit brief loss of the required arm contact within a 0.48 s grace window so the policy can execute the fast, dynamic reconfigurations that whole-body re-grasping demands rather than being forced into purely quasi-static holds.

Rewards. A goal is specified only by an object *position* and a target grasp *mode* (L, B, or R), so the policy is free to reach the target however it can. *Retention* $\rho_t \in \{0, 1\}$ indicates that the object has not dropped and keeps the arm contacts required by its current or target grasp mode; e_t is the distance to the commanded position; and the inactive

arm is *clear* when it has no object contact. The reward is a weighted sum grouped as:

$$r = \underbrace{r_{\text{prog}} + r_{\text{settle}} + r_{\text{succ}} + r_{\text{hold}}}_{\text{goal}} + r_{\text{neutral}} - r_{\text{fail}} - r_{\text{reg}}$$

Progress, hold quality and arm neutrality all carry the factor ρ_t , and therefore are nullified whenever retention is lost.

- **Progress** $r_{\text{prog}} = \rho_t \text{clip}(e_{t-1} - e_t, \pm 2 \text{ cm})$. The per-step reduction in distance to the commanded position, rewarded only while the object is retained and clipped so that no single step dominates.
- **Goal-region arrival** r_{settle} . A small per-step reward on movement goals during the dwell period, requiring the object to be settled within a 2 cm goal region in the required grasp mode.
- **Transition completion** r_{succ} . A sparse success bonus for repositioning the object and settling at the goal under the required grasp mode.
- **Hold quality** $r_{\text{hold}} = \rho_t \exp(-e_t/2.5 \text{ cm})$ on hold goals, which command the object to stay where it is. It rewards proximity to the hold position while the object is retained.
- **Arm neutrality** $r_{\text{neutral}} = \rho_t \exp(-\|q_t^{\text{arm}} - q_{\text{def}}^{\text{arm}}\|/0.4)$. Attracts the arms toward their default configuration while the object stays held, so an arm not needed for the current grasp returns to its rest pose.
- **Failure penalty** r_{fail} . A single terminal penalty for a drop, sustained loss of object contact, or another physical or numerical failure (Appendix III-B).
- **Regularizers** r_{reg} , three small penalties for unsupported motion, excessive object speed, and changes between actions (Appendix III-D).

Domain randomization. We train an individual expert per object over a diverse set of primitives (spheres, cubes, rectangular prisms, and cylinders). To make holding and transitions robust to unmodeled dynamics, we randomize object mass, center of mass, friction, and apply random perturbation forces to the object (Appendix III-F). This drives the policy toward grasps and handoffs that stay stable under a range of physical conditions rather than exploiting one exact setting. Evaluation retains object-mass, friction, and center-of-mass randomization, with external perturbation forces disabled.

Expert evaluation. We assess WBD-EX expert quality by *regional route completion*, the fraction of trials that reach every goal of a route in order before failure or timeout, where each route specifies a sequence of grasp goals across body regions. Route definitions, completion requirements and evaluation details are explained in Appendix IV.

Two-stage WBD-EX training. When trained from scratch, experts for smaller and less uniformly shaped objects achieve lower regional route completion than experts for medium-large spheres. To reduce the time to train a new object WBD-EX expert, we pre-train medium-large sphere experts, which provide an effective initialization for the other primitives. We then warm-start additional object primitives from these pre-trained experts with an identical reward formulation. Warm-

started experts achieve comparable or higher regional route completion than from-scratch training at lower amortized GPU cost by sharing the faster sphere pretraining across target objects (Appendix V).

C. WBD Controller Distillation

WBD-EX policies generate continuous trajectories of spatial exploration and transfers between body regions under randomized dynamics. We filter the trajectories to remove failed trajectory tails and stalled intervals, then balance sampling across objects and motion types with cross-body mirroring. We also equally sample three trajectory speed behaviors by sampling goals at goal distances of 2–4, 4–6 and 6–8 cm, unchanging within each trajectory, expressing deliberate and slow as well as faster and dynamic repositioning behavior.

These trajectories are used to train a WBD controller combining a conditional diffusion model with an inverse-dynamics network [12]. Given proprioceptive history and a five-dimensional geometry descriptor comprising log-sorted bounding-box side lengths, log volume, and sphericity, it generates keypoint motions and converts them into joint-target increments.

At inference, coarse commands guide generation through *conditional diffusion guidance* [12]. A tunable guidance weight controls how strongly denoising is steered toward the coarse motion command: larger weights emphasize command tracking, while smaller weights give greater relative influence to the learned motion prior. At each update, the diffusion model samples several candidate motions in parallel; we select the one closest to the command and convert it to refined joint commands with the inverse-dynamics model. This requires no privileged per-step object state and supports human teleoperation with tunable guidance.

IV. EXPERIMENTS AND RESULTS

We evaluate whether **WBD** can be used to generate whole-body dexterous data and train policies that refine imperfect motion commands into stable whole-body manipulation. We first report grasp and **WBD-EX** data generation, then evaluate a distilled WBD controller for on-body repositioning in simulation and on a real Unitree G1. A shelf pick, stow, carry, and place task tests whether these skills compose into a longer manipulation sequence.

A. Experimental Setup

Our object set contains 15 primitives: six spheres, four cubes, three rectangular prisms, and two cylinders. All side lengths and diameters fall between 80 and 240 mm. Of 3.27 million simulation-stable grasps, 2.62 million (80.2%) pass the height and 5 mm torso-distance filters described in Section III-A. Mean per-object coverage of sampled object-center positions around the torso is 90.2% using each object’s minimum radius as the distance tolerance; pooled coverage is 98.4% using the mean minimum radius (Appendix II-D).

We train one WBD-EX policy per object using the procedure in Section III-B and distill a shared WBD controller

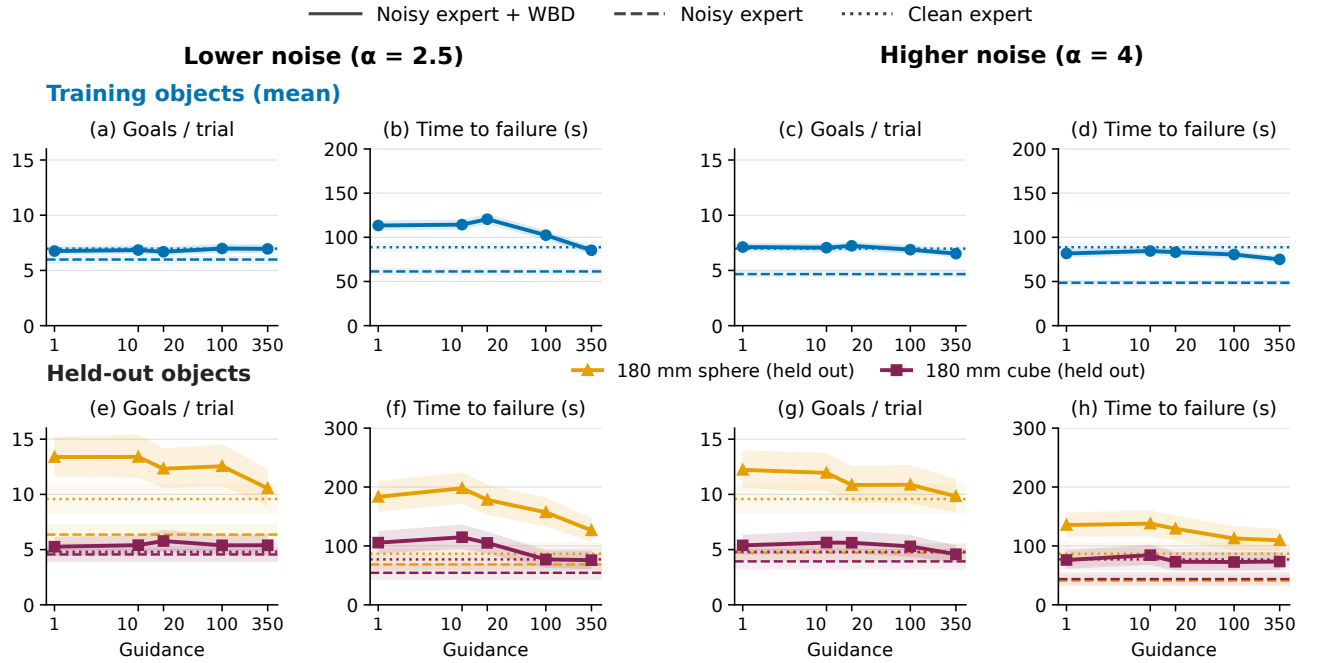


Fig. 4: **Simulated object repositioning under noisy commands.** Trials sample transfers among six body regions (Section IV-B). Training-object mean (top) and held-out sizes (bottom). Solid, dashed, and dotted lines show WBD refinement, noisy experts, and clean experts, respectively. Bands denote pointwise 95% paired bootstrap confidence intervals clustered by source grasp over 256 matched trials per object, with training objects weighted equally. Trials are capped at 600 s.

from their trajectories. Expert proficiency varies with object geometry: sphere experts reposition most effectively across all body regions, while experts for cubes, prisms and cylinders still reposition across all body regions but with lower success rate (Appendix IV-D). Because distillation uses only trajectory segments that pass a quality filter (Appendix VII-B), less proficient experts yield valid demonstrations less efficiently, requiring more collection attempts per retained trajectory. Our training dataset consists of 274,091 trajectories and 873 million recorded frames split equally across all objects. Network architectures, optimization settings, data augmentation, and training costs are detailed in Appendices III, V-C and VII-A.

B. Simulation Experiments

We evaluate on-body repositioning in simulation (Fig. 5) using a WBD controller trained on 13 of the 15 primitives, with a 180 mm sphere and a 180 mm cube withheld from training, validation, and normalization. We degrade each of the 15 object’s WBD-EX policies into coarse command streams and compare the degraded expert alone against the WBD controller refining it. We add independent uniform noise $\mathcal{U}(-\alpha, \alpha)$ to all 14 arm commands. We report results at $\alpha \in \{2.5, 4\}$ and guidance weights $\{1, 10, 20, 100, 350\}$, with clean experts as a reference (Fig. 4). WBD applies bounded corrections to the noisy commands, within ± 2.5 of each noisy command in the experts’ pre-tanh action space (Appendix VII-F). Degrading an expert reproduces the failure we see with real operator commands: the stream still carries the goal direction while individual arm targets move far enough to break supporting contact mid-handoff. In both

settings, the failure is a loss of supporting contact followed by a drop.

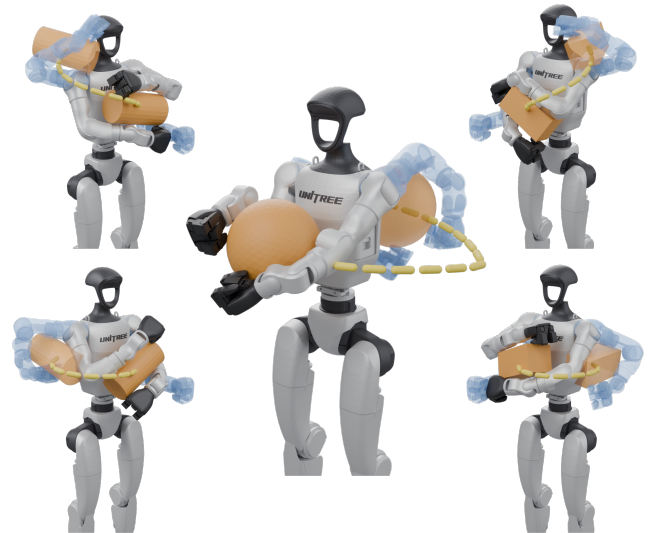


Fig. 5: **Simulated WBD controller experiment.** Examples of regional transfers in which the start grasp is shown on the fully visible g1, while the sampled *far goal* grasp in another region is shown with the blue ghost.

We evaluate transfers among six body regions: front, both underarms, both shoulders, and back. Each trial follows a seeded walk through adjacent regions (Appendix VII-D). The WBD controller is trained across all regions and voxels around the body; see Appendix VII-C. Each underarm and shoulder goal requires the opposite arm to release, with both arms supporting the object at the front and back. Each

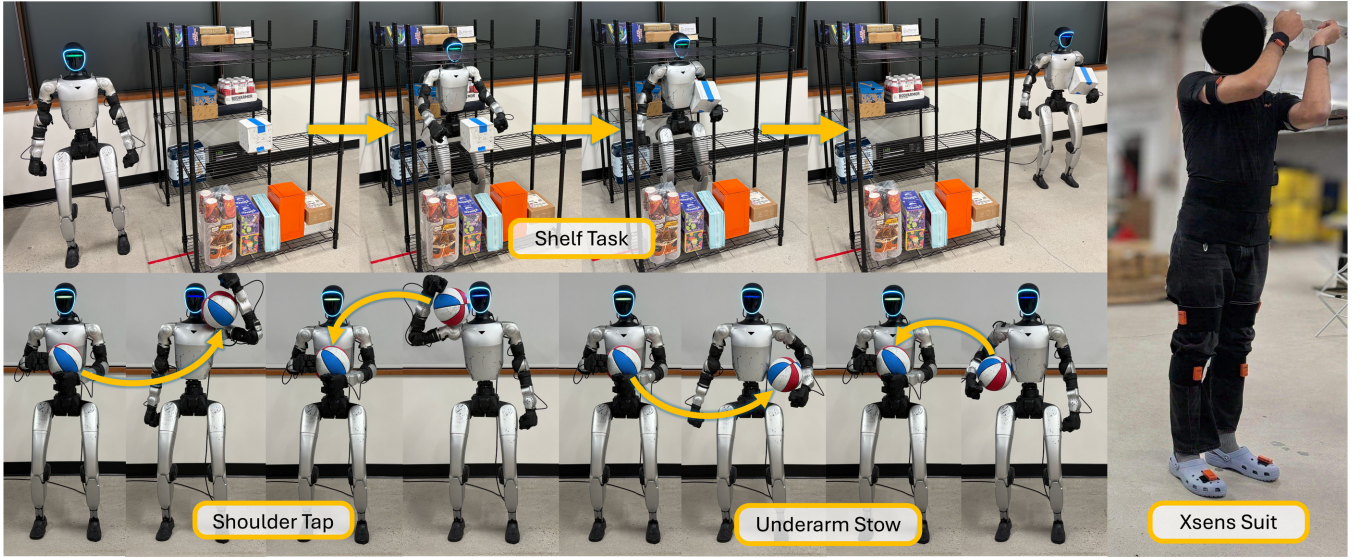


Fig. 6: **Real-world whole-body dexterous manipulation.** On-body repositioning (bottom) and shelf pickup, stowing, and carrying (top) using the WBD controller, guided by the teleoperator (right).

condition uses 256 matched initial grasps per object spanning the six regions; initial states and random seeds are matched across conditions. Trials end at the first failure or 600 s. We report completed goals per trial and time to failure restricted to this horizon (Fig. 4). Each goal is a (voxel, grasp mode) pair sampled from the voxel goal graph along the path towards a randomly drawn *far goal* in the next region of the walk; far goals are included as goals. Reaching a far goal completes a *regional transfer*. Tracking the number of goals achieved is an effective measure of how well a policy can stably transition an object across regions of the body.

Despite noisy commands, WBD largely recovers the clean experts’ mean goal-reaching performance across training objects at both plotted noise levels (Fig. 4). Averaged over these two noise levels and all five guidance values, mean time to failure increases relative to the noisy expert by 71% on training objects and by 168% and 75% on the held-out sphere and cube, respectively. Training objects are weighted equally. At the lower noise level and guidance weights of 1–20, WBD even exceeds the clean experts’ time to failure, by 28–36% for the training-object mean and by 106–128% and 36–49% for the held-out sphere and cube, respectively, showing that the WBD controller maintains stable object holding beyond what the uncorrupted experts achieve. As shown, excessive command tracking can reduce both stability and goals per trial.

C. Real-World Experiments

We evaluate a WBD controller trained on the full 15 object WBD-EX trajectory dataset on two on-body repositioning tasks and a long-horizon shelf task (Fig. 6). For repositioning, a size 3 basketball or a cardboard box starts in a two-handed front grasp. The *shoulder tap* task succeeds when the object contacts the shoulder link; *underarm stow* succeeds when the object center reaches the shoulder link’s depth under a stable single-arm hold. A round trip additionally requires returning to a stable two-arm front grasp. We report

region-reaching, full round-trip success, and mean time to each endpoint over successful trials (Table I). The shelf task requires picking the same cardboard box from a shelf, stowing it under one arm, stably carrying it while walking, un-stowing it, and placing it on a table. Pick and un-stow require a stable two-arm front grasp, and stow uses the underarm criterion, except that the free arm may stabilize the box while walking. We report success counts for each subtask: pick, stow, carry, un-stow, and place (Table II).

All trials are run by a single skilled human teleoperator with 8+ hours of training on these tasks. We use an Xsens motion capture suit to track motion beside and behind the operator, where standard VR headset tracking is limited. The controller can also accept commands from VR headsets and trackers, although limited tracking coverage can restrict dexterity. The operator indicates the desired motion, while WBD refines the arm targets to maintain object support as the arms and torso reconfigure. The operator switches the guidance weight between 10 for on-body repositioning and 100 for pick-and-place phases, balancing stable support with responsive command following.

We compare WBD refinement of motion retargeted by General Motion Retargeting (GMR) [51] against direct GMR control and SONIC [43]. In all three conditions the Xsens stream is converted to an SMPL body motion, which GMR retargets to G1 joint targets for WBD and direct GMR, and which SONIC, using its publicly released checkpoint, consumes directly as the reference for its whole-body tracker. The operator commands the upper body directly, while a separate controller handles the lower body and balance: Adaptive Motion Optimization (AMO) [3] for WBD and direct GMR, and SONIC’s own kinematic planner for SONIC. The real-world results suggest that maintaining object support through contact transitions is a central bottleneck (Tables I and II). GMR let the operator move the arms toward the relevant body regions, but stable contact during reposition-

Task	Controller	Basketball				Cardboard box			
		Reached $\uparrow$	Round trip $\uparrow$	Region (s) $\downarrow$	Total (s) $\downarrow$	Reached $\uparrow$	Round trip $\uparrow$	Region (s) $\downarrow$	Total (s) $\downarrow$
Shoulder tap	SONIC	2/10	0/10	15.2	—	2/10	2/10	15.9	24.9
	GMR	3/10	1/10	26.0	30.1	6/10	4/10	8.0	8.5
	WBD	10/10	9/10	7.7	17.3	10/10	10/10	4.1	6.3
Underarm stow	SONIC	1/10	0/10	28.0	—	2/10	0/10	26.1	—
	GMR	4/10	2/10	31.5	49.8	2/10	1/10	9.3	17.8
	WBD	10/10	9/10	10.6	26.2	10/10	10/10	8.6	13.0

TABLE I: **Real-world on-body repositioning.** Ten trials per controller, object, and task, five per side, with a size 3 basketball and a cardboard box (15 cm cube): region-reaching and full round-trip success. Times are measured from trial start and averaged over trials reaching the respective endpoint.

ing remained difficult; SONIC additionally struggled with reachability in constrained regions. Across the three tasks, counting complete on-body round trips for both objects and full shelf sequences, the WBD controller completed 45 of 50 trials (90%), compared with 9 of 50 (18%) for GMR and 2 of 50 (4%) for SONIC.

In simulation, cube WBD-EX experts have lower mean regional route completion than sphere experts (Table VIII). In the real-world shoulder tap trials, however, GMR completes 4/10 box round trips versus 1/10 with the basketball, with mean completion times of 8.5 versus 30.1 s among successful trials. We hypothesize that the cardboard’s lower friction makes it easier to slide it to the shoulder, whereas the basketball’s higher friction requires rolling along the torso and more deliberate arm manipulation, consistent with the faster successful WBD and GMR box trials. Underarm stowing still remained difficult, requiring a complete single-arm handoff, and yielded low success rates for GMR and SONIC (1/10 and 0/10 complete box round trips, respectively, compared with 10/10 for WBD).

Controller	Pick $\uparrow$	Stow $\uparrow$	Carry $\uparrow$	Un-stow $\uparrow$	Place $\uparrow$
SONIC	10/10	0/10	0/10	0/10	0/10
GMR	10/10	3/10	2/10	1/10	1/10
WBD	10/10	9/10	9/10	9/10	7/10

TABLE II: **Real-world shelf task.** Completed stages over ten trials per controller with the cardboard box.

On the long-horizon task (Table II), all methods completed the initial shelf pickup, but subsequent stowing and transport required support to persist as the robot reconfigured and moved. The WBD controller maintained support through carrying and un-stowing after successful stows, showing that body-supported grasps can serve as reliable intermediate holding states within longer tasks. While performing substantially better than GMR and SONIC, WBD still lost contact on a few trials resulting in dropping the object.

V. CONCLUSION

We presented a pipeline for learning whole-body dexterous manipulation at scale: contact configurations are enumerated into mode-typed whole-body grasps, and those grasps seed large-scale reinforcement learning under common rewards to dexterously reposition objects around the body. We show

how to utilize this large-scale whole-body dexterous manipulation data: training a shared controller for refining a coarse command stream into stable control and teleoperation. Although WBD-EX is trained without a lower-body controller, the distilled policy deploys on a standing and walking G1, holding objects through leaning and walking during the shelf task. The same formulation extends naturally to leg contact: generating combinations of arm and leg grasps, training WBD-EX jointly with a whole-body controller across this expanded grasp dataset, and distilling it to a deployable policy to enable behaviors such as bracing a heavy box on a knee while the arms reposition.

REFERENCES

- [1] Y. Ze, S. Zhao, W. Wang, A. Kanazawa, R. Duan, P. Abbeel, G. Shi, J. Wu, and C. K. Liu, “TWIST2: Scalable, portable, and holistic humanoid data collection system,” *arXiv preprint arXiv:2511.02832*, 2025.
- [2] L. Heng, Y. Tang, J. Xu, H. Bao, D. Huang, and Y. Wang, “HumDex: Humanoid dexterous manipulation made easy,” *arXiv preprint arXiv:2603.12260*, 2026.
- [3] J. Li, X. Cheng, T. Huang, S. Yang, R.-Z. Qiu, and X. Wang, “AMO: Adaptive motion optimization for hyper-dexterous humanoid whole-body control,” in *Proceedings of Robotics: Science and Systems (RSS)*, Los Angeles, CA, USA, June 2025.
- [4] Y. Zhang, Y. Yuan, P. Gurunath, I. Gupta, S. Omidshafiei, A.-a. Aghamohammadi, M. Vazquez-Chanlatte, L. Pedersen, T. He, and G. Shi, “FALCON: Learning force-adaptive humanoid loco-manipulation,” in *Proceedings of the 8th Annual Learning for Dynamics and Control Conference (LADC)*, ser. PMLR, vol. 331, 2026, pp. 265–281.
- [5] X. He, S. Xu, X. Li, R. Dong, L. Bian, Y.-X. Wang, and L.-Y. Gui, “ULTRA: Unified multimodal control for autonomous humanoid whole-body loco-manipulation,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2026.
- [6] T. He, Z. Wang, H. Xue, Q. Ben, Z. Luo, W. Xiao, Y. Yuan, X. Da, F. Castañeda, S. Sastry, C. Liu, G. Shi, L. Fan, and Y. Zhu, “VIRAL: Visual sim-to-real at scale for humanoid loco-manipulation,” *arXiv preprint arXiv:2511.15200*, 2025.
- [7] S. Wei, H. Jing, B. Li, Z. Zhao, J. Mao, Z. Ni, S. He, S. Zang, X. Liu, K. Kang, J. Liu, W. Yuan, M. Pavone, D. Huang, and Y. Wang, “ Ψ_0 : An open foundation model towards universal humanoid loco-manipulation,” in *Robotics: Science and Systems (RSS)*, 2026.
- [8] C. Zheng, K. Chen, Z. Bi, Y. Li, L. Pan, J. Zhou, H. Li, and J. Ma, “Embracing bulky objects with humanoid robots: Whole-body manipulation with reinforcement learning,” *arXiv preprint arXiv:2509.13534*, 2025.
- [9] M. Murooka, T. Hoshi, K. Fukumitsu, S. Masuda, M. Hamze, T. Sasaki, M. Morisawa, and E. Yoshida, “TACT: Humanoid whole-body contact manipulation through deep imitation learning with tactile modality,” *IEEE Robotics and Automation Letters*, vol. 10, no. 8, pp. 7819–7826, Aug. 2025.
- [10] J. Dao, H. Duan, and A. Fern, “Sim-to-real learning for humanoid box loco-manipulation,” in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 16 930–16 936.

- [11] A. Rigo, M. Hu, S. K. Gupta, and Q. Nguyen, "Hierarchical optimization-based control for whole-body loco-manipulation of heavy objects," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 15 322–15 328.
- [12] Z.-H. Yin, C. Wang, L. Pineda, F. R. Hogan, C. K. Bodduluri, A. Sharma, P. Lancaster, I. Prasad, M. Kalakrishnan, J. Malik, M. Lambeta, T. Wu, P. Abbeel, and M. Mukadam, "DexterityGen: Foundation controller for unprecedented dexterity," in *Robotics: Science and Systems (RSS)*, 2025.
- [13] H. Qi, A. Kumar, R. Calandra, Y. Ma, and J. Malik, "In-hand object rotation via rapid motor adaptation," in *Conference on Robot Learning*, ser. PMLR, vol. 205. PMLR, 2023, pp. 1722–1732.
- [14] OpenAI, M. Andrychowicz, B. Baker, M. Chociej, R. Józefowicz, B. McGrew, J. Pachocki, A. Petron, M. Plappert, G. Powell, A. Ray, et al., "Learning dexterous in-hand manipulation," *The International Journal of Robotics Research*, vol. 39, no. 1, pp. 3–20, 2020.
- [15] T. He, Z. Luo, W. Xiao, C. Zhang, K. Kitani, C. Liu, and G. Shi, "Learning human-to-humanoid real-time whole-body teleoperation," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 8944–8951.
- [16] T. He, Z. Luo, X. He, W. Xiao, C. Zhang, W. Zhang, K. Kitani, C. Liu, and G. Shi, "OmniH2O: Universal and dexterous human-to-humanoid whole-body teleoperation and learning," in *Conference on Robot Learning (CoRL)*, 2024.
- [17] T. Liu, Z. Liu, Z. Jiao, Y. Zhu, and S.-C. Zhu, "Synthesizing diverse and physically stable grasps with arbitrary hand structures using differentiable force closure estimator," *IEEE Robotics and Automation Letters*, vol. 7, no. 1, pp. 470–477, 2022.
- [18] R. Wang, J. Zhang, J. Chen, Y. Xu, P. Li, T. Liu, and H. Wang, "Dex-GraspNet: A large-scale robotic dexterous grasp dataset for general objects based on simulation," *arXiv preprint arXiv:2210.02697*, 2023.
- [19] J. Chen, Y. Ke, and H. Wang, "BODex: Scalable and efficient robotic dexterous grasp synthesis using bilevel optimization," *arXiv preprint arXiv:2412.16490*, 2025.
- [20] X. Xu, D. Bauer, and S. Song, "RoboPanoptes: The all-seeing robot with whole-body dexterity," in *Proceedings of Robotics: Science and Systems*, Los Angeles, CA, USA, June 2025.
- [21] A. Bhatt, O. Kaidanov, P. Liu, and J. Peters, "Blind dexterity: Whole-body humanoid manipulation via pure proprioception," *arXiv preprint arXiv:2608.29487*, 2026.
- [22] M. Zhang, J. Barreiros, and A. O. Onol, "Plan-guided reinforcement learning for whole-body manipulation," 2023.
- [23] J. A. Barreiros, A. Ö. Onol, M. Zhang, S. Creasey, A. Goncalves, A. Beaulieu, A. Bhat, K. M. Tsui, and A. Alspach, "Learning contact-rich whole-body manipulation with example-guided reinforcement learning," *Science Robotics*, vol. 10, no. 105, p. eads6790, 2025.
- [24] S. Zhao, Y. Ze, Y. Wang, C. K. Liu, P. Abbeel, G. Shi, and R. Duan, "ResMimic: From general motion tracking to humanoid whole-body loco-manipulation via residual learning," 2025.
- [25] C. C. Johnson, C. Alessi, E. Falotico, and M. D. Killpack, "Zero-shot whole-body manipulation with a large-scale soft robotic torso via guided reinforcement learning," *arXiv preprint arXiv:2509.23556*, 2025.
- [26] V. Levé, J. Moura, N. Saito, S. Tonneau, and S. Vijayakumar, "Explicit contact optimization in whole-body contact-rich manipulation," in *2024 IEEE-RAS 23rd International Conference on Humanoid Robots (Humanoids)*. IEEE, Nov. 2024, p. 451–458.
- [27] M. Murooka, K. Okada, and M. Inaba, "Optimization-based posture generation for whole-body contact motion by contact point search on the body surface," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, p. 2905–2912, Apr. 2020.
- [28] M. Murooka, Y. Inagaki, R. Ueda, S. Nozawa, Y. Kakiuchi, K. Okada, and M. Inaba, "Whole-body holding manipulation by humanoid robot based on transition graph of object motion and contact," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2015, pp. 3950–3955.
- [29] X. Li, X. He, R. Dong, and S. Gupta, "ContactMimic: Humanoid object interaction via contact control," *arXiv preprint arXiv:2607.08742*, 2026.
- [30] S. Xu, H. Y. Ling, Y.-X. Wang, and L.-Y. Gui, "Intermimic: Towards universal whole-body control for physics-based human-object interactions," in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2025, pp. 12 266–12 277.
- [31] L. Cao, X. Wu, J. Cui, B. Xu, M. Pei, R. Chen, and M. Xu, "Weave: Learning whole-body dexterous loco-manipulation from human-object interactions," *arXiv preprint arXiv:2609.16683*, 2026.
- [32] R. Li, C. Ma, S. Li, Z. Wei, Y. Yao, H. Shi, C. K. Liu, S. Song, and M. Ding, "Handroid: Bridging dexterous hand and humanoid," *arXiv preprint arXiv:2607.16187*, 2026.
- [33] Boston Dynamics, "Atlas," <https://bostondynamics.com/blog/training-a-humanoid-robot-for-hard-work/>, 2025.
- [34] Figure AI, "Helix 02: Living room tidy," <https://www.figure.ai/news/helix-02-living-room-tidy>, Mar. 2026.
- [35] T. Xie, H. Zhang, J. Park, Z. Wang, B. Wen, J. Li, X. Li, Q. Ben, H. Weng, Y. Ye, D. Minor, T. Wang, C. Jiang, S. Fidler, J. Kautz, L. Fan, Y. Zhu, Z. Luo, U. Iqbal, and Y. Yuan, "GRAIL: Generating humanoid loco-manipulation from 3D assets and video priors," *arXiv preprint arXiv:2606.05160*, 2026.
- [36] I. Taouil, M. Ciebielski, S. Omar, H. Zhao, A. Dai, A. M. Johnson, and M. Khadiv, "MotionDisco: Motion discovery for extreme humanoid loco-manipulation," *arXiv preprint arXiv:2606.06139*, 2026.
- [37] L. Yang, X. Huang, Z. Wu, A. Kanazawa, P. Abbeel, C. Sferazza, C. K. Liu, R. Duan, and G. Shi, "OmniRetarget: Interaction-preserving data generation for humanoid whole-body loco-manipulation and scene interaction," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2026.
- [38] J. Z. Zhang, M. Sorokin, J. Brüdigam, B. Hung, S. Phillips, D. Yershov, F. Niroui, T. Zhao, L. Fermoselle, X. Zhu, C. Cao, D. Ta, T. Pang, J. Wang, P. Culbertson, Z. Manchester, and S. Le Cléc'h, "Sumo: Dynamic and generalizable whole-body loco-manipulation," *arXiv preprint arXiv:2604.08508*, 2026.
- [39] Z. Fu, Q. Zhao, Q. Wu, G. Wetzstein, and C. Finn, "HumanPlus: Humanoid shadowing and imitation from humans," in *Conference on Robot Learning (CoRL)*, 2024.
- [40] X. Cheng, Y. Ji, J. Chen, R. Yang, G. Yang, and X. Wang, "Expressive whole-body control for humanoid robots," in *Robotics: Science and Systems (RSS)*, 2024.
- [41] Y. Ze, Z. Chen, J. P. Araújo, Z.-A. Cao, X. B. Peng, J. Wu, and C. K. Liu, "TWIST: Teleoperated whole-body imitation system," *arXiv preprint arXiv:2505.02833*, 2025.
- [42] T. He, W. Xiao, T. Lin, Z. Luo, Z. Xu, Z. Jiang, J. Kautz, C. Liu, G. Shi, X. Wang, L. Fan, and Y. Zhu, "HOVER: Versatile neural whole-body controller for humanoid robots," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2025.
- [43] Z. Luo, Y. Yuan, T. Wang, C. Li, F. Castañeda, S. Chen, Z.-A. Cao, J. Li, D. Minor, Q. Ben, J. Park, D. Sami, Z. Wang, X. Da, R. Ding, C. Hogg, L. Song, E. Lim, E. Jeong, T. He, H. Xue, W. Xiao, S. Yuen, J. Kautz, Y. Chang, U. Iqbal, L. Fan, and Y. Zhu, "SONIC: Supersizing motion tracking for natural humanoid whole-body control," *Science Robotics*, vol. 11, no. 117, p. ead4592, 2026.
- [44] T. Chen, M. Tippur, S. Wu, V. Kumar, E. Adelson, and P. Agrawal, "Visual dexterity: In-hand reorientation of novel and complex object shapes," *Science Robotics*, vol. 8, no. 84, p. ead9244, 2023.
- [45] N. Wake, A. Kanehira, K. Sasabuchi, J. Takamatsu, and K. Ikeuchi, "A taxonomy of self-handover," *arXiv preprint arXiv:2504.04939*, 2025.
- [46] K. Kedia, T. G. W. Lum, J. Bohg, and C. K. Liu, "SimToolReal: An object-centric policy for zero-shot dexterous tool manipulation," *arXiv preprint arXiv:2602.16863*, 2026.
- [47] J. Lee, J. Yin, A. Rana, N. Blauch, S. Mady, M. Bhardwaj, N. Fazeli, N. Ratliff, K. V. Wyk, and A. Handa, "ADEPT: Accelerating dexterity via pre-training and post-training using reinforcement learning," *arXiv preprint arXiv:2608.19182*, 2026.
- [48] M. Toussaint, C. V. Braun, A. Jordana, S. Audy, E. Cobo-Briesewitz, D. Shcherba, T. Burghoff, and J. Carpentier, "Combined constrained sampling and reinforcement learning for robotic manipulation," *arXiv preprint arXiv:2602.08557*, 2026.
- [49] J. Ye, K. Wang, C. Yuan, R. Yang, Y. Li, J. Zhu, Y. Qin, X. Zou, and X. Wang, "Dex1B: Learning with 1B demonstrations for dexterous manipulation," in *Robotics: Science and Systems (RSS)*, 2025.
- [50] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.
- [51] J. P. Araújo, Y. Ze, P. Xu, J. Wu, and C. K. Liu, "Retargeting matters: General motion retargeting for humanoid motion tracking," *arXiv preprint arXiv:2510.02252*, 2025.
- [52] S. Coumar, G. Chang, N. Kodkani, and Z. Kingston, "Foam: A tool for spherical approximation of robot geometry," 2025. [Online]. Available: <https://arxiv.org/abs/2503.13704>

APPENDIX I WHOLE-BODY GRASP GENERATION DETAILS

We adapt prior energy-based dexterous grasp synthesis [49], [18] to prescribed whole-body contact links and torso regions. We seek joint configurations $q \in \mathcal{C}$ in configuration space $\mathcal{C}$ that minimize the resulting objective. We define K contact points $\mathbf{p}_j$ on the link mesh surfaces throughout the robot that we wish to make contact with the object and, using the signed distance function of the object mesh $\phi : \mathbb{R}^3 \rightarrow \mathbb{R}$, we formulate the following:

$$\min_{q, \mathbf{T}} E_{\text{closure}} + \frac{1}{K} \sum_{j=1}^K [\phi(\mathbf{p}_j) - d^*]_+^2 \quad (1)$$

$$\text{s.t. } \phi(\mathbf{b}_i) \geq r_i, \quad \forall i = 1, \dots, N \quad (2)$$

$$\|\mathbf{b}_i - \mathbf{b}_l\|_2 \geq r_i + r_l, \quad \forall (i, l) \in \mathcal{H} \quad (3)$$

$$\underline{q}_m \leq q_m \leq \bar{q}_m, \quad \forall m = 1, \dots, n_q \quad (4)$$

where E_{closure} is a force-closure energy defined below, $\mathbf{T} \in SE(3)$ is the pose of the robot's kinematic root in the object frame, $\underline{q}_m$ and $\bar{q}_m$ are the limits of joint q_m for the n_q optimized joints, $\mathbf{b}_i$ are the centers of the robot's N collision spheres with radii r_i , and $\mathcal{H}$ is the set of non-adjacent sphere pairs checked for self-collision. We also define $[\cdot]_+ = \max\{\cdot, 0\}$. We constrain the optimization problem above to avoid collisions between the link collision spheres and the objects, avoid self-collision between link collision spheres, and obey joint limits.

Since this is a nonconvex problem, and we would like to leverage GPU parallelization capabilities of first-order solvers like Adam to generate many grasps simultaneously, we relax this optimization problem to the unconstrained form:

$$\begin{aligned} \min_{q, \mathbf{T}} \quad & w_{\text{closure}} E_{\text{closure}} + \frac{w_{\text{dist}}}{K} \sum_{j=1}^K [\phi(\mathbf{p}_j) - d^*]_+^2 \\ & + w_{\text{link}} \sum_{i=1}^N [r_i - \phi(\mathbf{b}_i)]_+^2 \\ & + w_{\text{joint}} \sum_{m=1}^{n_q} ([q_m + \delta_q - q_m]_+^2 + [q_m - \bar{q}_m + \delta_q]_+^2) \\ & + w_{\text{self}} \sum_{(i, l) \in \mathcal{H}} [r_i + r_l - \|\mathbf{b}_i - \mathbf{b}_l\|_2]_+^2 \end{aligned}$$

where we define weights w for each optimization term. The contact-distance tolerance $d^* = 2$ mm sets the distance beyond which a contact point is attracted toward the surface. The joint-limit penalty begins $\delta_q = 0.035$ rad inside each limit; joint angles are also clamped to their hard limits after each update. Table III gives the G1 settings.

We express the force-closure constraint as the differentiable energy [17], [18], [49] $E_{\text{closure}} = \frac{1}{K^2} (\|\sum_{k=1}^K \mathbf{n}_k\|_2^2 + \|\sum_{k=1}^K (\mathbf{p}_k - \bar{\mathbf{p}}) \times \mathbf{n}_k\|_2^2)$ where $\mathbf{n}_k = -\nabla\phi(\mathbf{p}_k)/\|\nabla\phi(\mathbf{p}_k)\|_2$ is the inward surface normal of the closest surface point to the query point $\mathbf{p}_k$ and $\bar{\mathbf{p}}$ is the contact centroid.

E_{closure} is the squared net force and moment of unit normal contact forces, so it vanishes when both cancel. Contact magnitudes are fixed at unity and no friction cone enters, making this a frictionless net-wrench residual rather than a force-closure feasibility test, and moments are taken about the contact centroid rather than the object's center of mass.

A. Contact Model and Decision Variables

The humanoid is treated as a dexterous manipulator. The torso is a static “palm” that the object can be pinned against and is simultaneously the kinematic reference frame; each arm or leg is a movable appendage. For the Unitree G1, the fourteen Dex3 finger joints are pinned to constant closed grip values and the twelve leg joints and three waist joints are fixed, leaving the fourteen arm joints as the only controls. Single-arm sweeps additionally freeze the opposite arm, leaving seven.

A candidate is therefore twenty-three scalars optimized jointly: the position of the kinematic root in the object frame, its orientation in a continuous six-dimensional rotation parameterization, and the fourteen joint angles. The signed distance field is defined in the object frame with the object at the origin and the *robot* is moved relative to it, so the stored grasp pose is recovered by inverting the solved transform.

B. Geometry and Contact Sampling

The object is represented by the exact triangle-mesh signed distance field queried through a bounding-volume hierarchy; no voxel volume is used, so all iterations query the mesh directly. Signs are computed by angle-weighted pseudonormal during the search and by ray parity for the final refinement iterations. The robot is approximated by collision spheres rather than by link meshes [49], generated from the robot geometry by spherical approximation [52].

Candidate contact points are a separate, denser set of 626 points, sampled uniformly on the link meshes and then hand-curated, across 34 links, of which 347 over 25 links are active for the bimanual configuration. Rather than using all of them, each iteration samples $K = 5$ of the selected links, always including any required links, and draws one candidate point on each, resampled every iteration [18]. This keeps contacts from collapsing onto a single link and produces the spread of contact patterns the dataset exhibits. Penetration is evaluated over all spheres at every iteration regardless of which contacts are active.

C. Contact Links and Torso Regions

A configuration type names the links allowed to contact the object and, for the torso, which named region of its surface may be used. Our grasp dataset uses three presets: a right-arm sweep over 13 links, a bimanual sweep over 25 links, both requiring torso contact, and a bimanual sweep over 24 arm links with the torso excluded, which produces grasps pinned between the arms alone.

Regions are defined as tags on individual candidate contact points rather than as geometric boxes, and only the torso carries them: the 131 torso points are tagged across six regions shown in Figure 7.



Fig. 7: **Torso contact regions.** One accepted grasp is shown per region, each on a different object.

Restricting to a region has four effects. It filters the candidate contact points, so a torso-region sweep optimizes against roughly a sixth of the available torso points; it biases initialization, since the direction the body is aimed from is set per region; it assigns each region an equal share of the per-object grasp generation quota; and it labels the resulting grasps so that region coverage can be audited afterward. It does not restrict which links may contact.

Every accepted grasp is finally mirrored across the sagittal plane, which doubles the set exactly. The mirror is an exact relabeling of an already-validated grasp, so it is not re-simulated.

D. Pre-Grasp and Post-Grasp Targets

We adapt the contact-normal pressing strategy of [18]. From a solved configuration we derive a retracted pre-grasp and a pressed post-grasp target by backpropagating $\sum_k \nabla \phi(\mathbf{p}_k) \cdot \mathbf{p}_k$ through the kinematics with the SDF gradients held fixed to a joint-space gradient. We normalize this gradient to unit Euclidean norm, multiply it by $+0.5$ for retraction or -0.4 for pressing, and add it to q , clamping the result to the joint limits. The pre-grasp is the approach pose used to start the simulation trial; the post-grasp is the press target held during validation and later reused as the servo setpoint at each WBD-EX episode reset.

E. Solver

All candidates in a batch are optimized in parallel on one GPU: 4096 candidates for 4000 iterations with AdamW at learning rate 10^{-3} , weight decay 10^{-4} , and a cosine schedule annealing to 10^{-4} . Steps after a short warmup are replayed from captured CUDA graphs. Optimization candidates are initialized by sampling a random direction outward from the object centroid and rotating the robot to face the object centroid. Orientation is then jittered by a uniform roll about

the facing axis over $(-\pi, \pi)$ and two orthogonal tilts over $(-\pi/8, \pi/8)$, and the seed posture is jittered per joint by $\pm 25\%$ of that joint’s range.

Parameter	Value
Closure weight w_{closure}	1
Contact-distance weight w_{dist}	300
Object-penetration weight w_{link}	200
Joint-limit weight w_{joint}	1
Self-collision weight w_{self}	10
Contact-distance tolerance d^*	2 mm
Joint-limit margin δ_q	0.035 rad
Pre-/post-grasp gradient multipliers	$+0.5 / -0.4$

TABLE III: **G1 grasp-generation parameters.**

F. Simulation Validation

Because the optimizer minimizes a static surrogate objective, every optimized candidate is validated dynamically in two stages. The torso is pinned to its rest pose every substep and the lower body is held fixed, so the torso frame is exactly static; the object is a free body. Physics runs at 1 ms with four substeps per control frame, an elliptic friction cone, and surface friction of $\mu = 0.6$ on the robot with priority over the object’s, so the contact that matters is governed by the robot-side coefficient.

Each simulation trial dwells at the pre-grasp, ramps to the grasp candidate configuration, presses to the post-grasp target, and then holds. The approach, press, settling and hold phases last 40, 40, 60 and 300 control frames, respectively, with 4 ms per frame. The gravity load is applied during the hold. A link counts as contacting when its normal contact force is at least 0.1 N.

Stage one loads the object at $4\times$ weight and, over the first half of the hold, sweeps a sustained 10 N force and 0.5 N m torque through the six axis directions for 100 ms each, followed by an unperturbed holding phase. A candidate passes if the object is still held in contact with at least two contact links at the end of the holding phase.

Stage two re-runs the survivors with no perturbations, normal gravity, and adds pose gates, requiring the object to drift less than 0.10 m and rotate less than 45° over the hold. The configuration and object pose recorded at the end of stage two are the settled values stored in the cache and used at WBD-EX reset, so the training distribution consists of poses that have settled under contact rather than the optimizer’s solutions.

We refer throughout to a grasp’s *mode* (left, right or both), meaning which arms must contact the object at the goal, as specified by its optimization contact group. Mode *L* requires left-arm contact, *R* right-arm contact, and *B* both; single-arm goals also require the opposite arm to release (Appendix III-C).

APPENDIX II

VOXELIZATION AND THE WBD-EX GRASP CACHE

A. Cache Contents

Each object’s accepted grasps are stored as a single array file that WBD-EX loads at startup. Per grasp we keep the

settled arm configuration q_{sim} (14 joints), the pressed post-grasp target q_{post} used as the servo setpoint at reset, the settled object pose $T_{\text{obj}}^{\text{torso}}$ as a homogeneous transform in the torso frame, and the grasp mode $m \in \{L, R, B\}$.

B. Position Voxelization

We voxelize by the *settled object center* expressed in the torso frame, taking the translation of $T_{\text{obj}}^{\text{torso}}$ and binning it on a 2 cm lattice anchored at the torso origin. A voxel’s representative position is the mean of its member grasp positions. Within a voxel, grasps are grouped by mode, so a *node* of the goal graph is an occupied (voxel, mode) pair. Across the object set this yields $1.0\text{--}5.5 \times 10^3$ occupied voxels and $1.7\text{--}9.1 \times 10^3$ nodes; spheres sit at the low end and the elongated prisms at the high end, because an elongated object reaches a wider spread of settled centers for the same body region. The occupied region spans roughly ± 0.21 m laterally and 0.03–0.45 m vertically.

C. Goal Graph

Mode compatibility and goal sampling follow Section III-B, with $B \leftrightarrow B$ transitions also permitted. A mode change within a single voxel is a legal zero-distance transition. Intermediate goals exclude the current (voxel, mode) node.

Far goals exclude the current voxel. If no legal-mode node lies within the sampled distance, the sampler selects the nearest legal-mode node toward the far goal. If no legal-mode node exists, the episode terminates as a dead end. A curriculum ramps d_{max} linearly from 0.02 m to 0.08 m over the first 5000 iterations. Hold commands are excluded from every transition-credit signal.

D. Cache Construction

Candidate grasps enter the cache only after the two-stage simulation filter of Appendix I. We then apply a safety filter to the *settled* pose with two components. The settled object center must lie at least 9 cm above the torso-frame origin, or 3 cm for objects whose mesh bounding-box height exceeds 15 cm. An exact on-body clearance test then drops any grasp whose object surface sits farther than 5 mm from the torso.

Coverage measures sampled object-center positions around the torso. For each object, we offset exterior torso-face centroids along their normals at six evenly spaced distances between the object’s minimum and maximum bounding-box half-extents. We deduplicate positions on a 2 cm grid and retain torso-frame heights $z \in [0.09, 0.29]$ m. A reference position is covered if a retained grasp center lies within the object’s minimum half-extent. The per-object score is the fraction covered, averaged equally across objects. For pooled coverage, we combine all grasp centers and reference positions, deduplicate the latter on a 1 cm grid, and use the mean minimum half-extent of 7.6 cm as the tolerance.

APPENDIX III

WBD-EX TRAINING DETAILS

A. Observations and Actions

Table IV lists the observation terms. The actor receives 89 dimensions at the current timestep, and the critic receives the

Group	Term	Dim.
Proprioception	Arm joint positions, minus default	14
	Arm joint velocities	14
	Servo error, commanded — measured	14
	Previous action	14
Command	Hold flag, commanded and max. goal distance	3
	Goal position relative to the object	3
Object	Position and 6-D rotation, torso frame	9
	Linear and angular velocity, torso frame	6
Contact	Contacting-link fractions: left arm, right arm, torso	3 e
	Contacting links / total monitored links	1
	Retention flag (ρ_t)	1
	Net contact-force magnitude (divided by 80 N, capped at 2)	1
Grasp mode	One-hot source and target mode	6
Actor total		89
Privileged	Object mass	1
	Object center-of-mass offset	3
	Mean sliding friction, per group	2
Critic total		95

TABLE IV: **WBD-EX observation terms.** Current-timestep inputs, without history. The actor uses the first five groups with training noise; the critic uses their uncorrupted values plus the sampled dynamics parameters.

same terms plus six privileged dimensions giving the realized randomization of Appendix III-F, for 95 dimensions. Actor observations are corrupted during training with additive uniform noise of ± 0.005 rad on joint positions, ± 0.003 m on the goal offset and object position, and ± 0.01 on the rotation entries; the critic receives the same terms uncorrupted.

The action is a 14-dimensional bounded increment on the arm joint targets. A raw action is squashed through tanh and scaled by 0.025 rad per control step, then accumulated into a retained servo setpoint that persists across steps. The setpoint is clamped to lie within 0.35 rad of the measured position and within the joint limits, and is linearly interpolated across the four physics substeps. The lower body and waist are welded, so the torso frame is exactly static and the arms are the only actuated degrees of freedom. Arm stiffness is $k_p = 20$; damping is derived per joint as $k_v \propto \sqrt{k_p}$.

B. Rates, Episodes, Terminations

Physics runs at 200 Hz with a decimation of four, giving a 50 Hz control rate. Training episodes last 40 s (2000 control steps) and rewards are not scaled by the timestep. Contact is read from a sensor over all links against the object collision geometry.

An episode ends on time-out, which is bootstrapped, or on one of six failures: the object falling below 0.5 m in the world frame; contact lost for longer than a grace window of 24 control steps (0.48 s), where contact is maintained if the arm contacts required by either the source or the target grasp mode are present, and torso contact alone does not count; the object leaving a 0.55 m workspace radius or exceeding 6 m/s or 40 rad/s; the object coming to rest on the legs; a non-finite state; or a dead end in the goal graph. The sparse

failure penalty fires on the first five of these; dead ends and time-outs are not penalized.

Each episode is initialized directly from a cached grasp, with the arm configuration and object pose written exactly from the cache grasp and zero velocity. The servo target is set instantly to that grasp’s pressed target with no ramp. Grasps are sampled as in Section III-B. The only continuous randomization at reset is a ± 5 mm jitter on movement goal positions; hold goals are exact.

C. Goal Completion Criteria

We define the object as *settled* when its linear and angular speeds are below 0.12 m/s and 1.5 rad/s, respectively. A commanded goal is completed when all of the following hold at once: the object is within 0.02 m of the commanded position, the object is settled, the arms required by the goal grasp mode are in contact, and these conditions persist for 10 control steps for movement goals or 25 for hold goals. A single-arm goal additionally requires the other arm to register no object contact for five consecutive steps. Training and evaluation both apply this requirement.

D. Reward Details

Table V lists the signed coefficients for the reward terms and penalty expressions defined in Section III-B and below. Reward weights are fixed for the whole of training; the two curricula modify the environment rather than the objective.

The object’s linear and angular speeds are denoted by v_t and ω_t . The three regularization penalties are:

- *Unsupported motion*, $(1 - \rho_t) \max(v_t - 5 \text{ cm/s}, 0)^2$: penalizes linear object speeds above the threshold when $\rho_t = 0$, discouraging unsupported motion.
- *Speed cap*, $\max(v_t - 0.35 \text{ m/s}, 0)^2 + 0.02 \max(\omega_t - 3 \text{ rad/s}, 0)^2$: penalizes linear and angular speeds above their respective thresholds.
- *Action rate*, $\|a_t - a_{t-1}\|^2$: penalizes changes between consecutive actions.

Term	Coefficient
Progress	100
Goal-region arrival	0.15
Transition completion	20
Hold quality	0.2
Arm neutrality	0.5
Failure penalty	−20
Unsupported motion	−1
Speed cap	−0.5
Action rate	−10 ^{−3}

TABLE V: **Reward coefficients.** Negative values denote penalties.

E. Optimization

Experts are trained with PPO [50] and the asymmetric actor–critic inputs specified in Appendix III-A. We use 64 steps per environment, five epochs and 16 minibatches per update, discount 0.99, GAE $\lambda = 0.95$, clip 0.2, value-loss coefficient 1, gradient-norm clip 1, and advantage normalization over the collected batch. The learning rate starts at 10^{−3}

under a KL-targeted adaptive schedule with target 0.015, clamped to $[10^{-5}, 10^{-2}]$; the second stage inherits the pre-trained sphere policy’s converged rate. Entropy coefficient is 0.005 in the first stage and 0 in the second. The $\varnothing 180$, 200 and 220 mm sphere experts are not warm-started from another object; they are trained to 50,000 iterations with the first-stage settings. Actor and critic are separate 512×256 ELU networks with empirical observation normalization, and the policy is Gaussian with a learned standard deviation initialized at 0.3 and hard-clamped to $[0.05, 0.6]$ in the first stage and $[0.05, 0.4]$ in the second. Production runs use 8192 environments on each of eight NVIDIA H100 GPUs, giving 65,536 environments and roughly 4.2×10^6 transitions per iteration.

F. Domain Randomization

Randomization ranges and schedules are listed in Table VI.

Parameter	Range and sampling schedule
Robot friction	Per-geometry multiplier $\mathcal{U}(0.75, 1.25)$, sampled on the first reset and fixed per environment.
Object mass and inertia	Mass 0.4–1.2 kg for every object, with inertia co-scaled. Sampled at startup and fixed per environment.
Object center of mass	Offset ± 5 mm independently per axis. Sampled at startup and fixed per environment.
Perturbation force	Magnitude $\mathcal{U}(0, F_{\max})$ and isotropic direction, applied at the object center. No applied torque.
Force duration	0.05–0.20 s per perturbation.
Cooldown	0.75–2.0 s between perturbations.
Force cap $F_{\max}$	0 for the first 2000 iterations, then a linear ramp to 20 N over the next 8000; 20 N thereafter.

TABLE VI: **WBD-EX domain randomization.**

APPENDIX IV

WBD-EX REGIONAL ROUTE COMPLETION

WBD-EX policies are evaluated on a standardized set of routes traversing between six regions across the entire body as described in Table VII.

A. Body Regions

Evaluation is defined over six named regions of the torso frame, applied to the settled object center (Fig. 8). The region bands are fixed across all object sets so that scores are comparable; the voxels a given region admits differ across objects only because the reachable grasp set does. Front and back occupy the three innermost 2 cm columns on either side of the body within ± 3 cm laterally and 0.090–0.19 m vertically, and both require a both-arm grasp. The under-arms lie beyond ± 0.11 m laterally, within -0.04 – 0.06 m front-to-back and 0.090–0.22 m vertically, and require the corresponding single-arm grasp. The shoulders lie beyond ± 0.05 m laterally, within ± 0.06 m front-to-back and 0.29–0.49 m vertically, also requiring the corresponding single arm. All expert evaluations use the completion and inactive-arm release requirements in Appendix III-C.

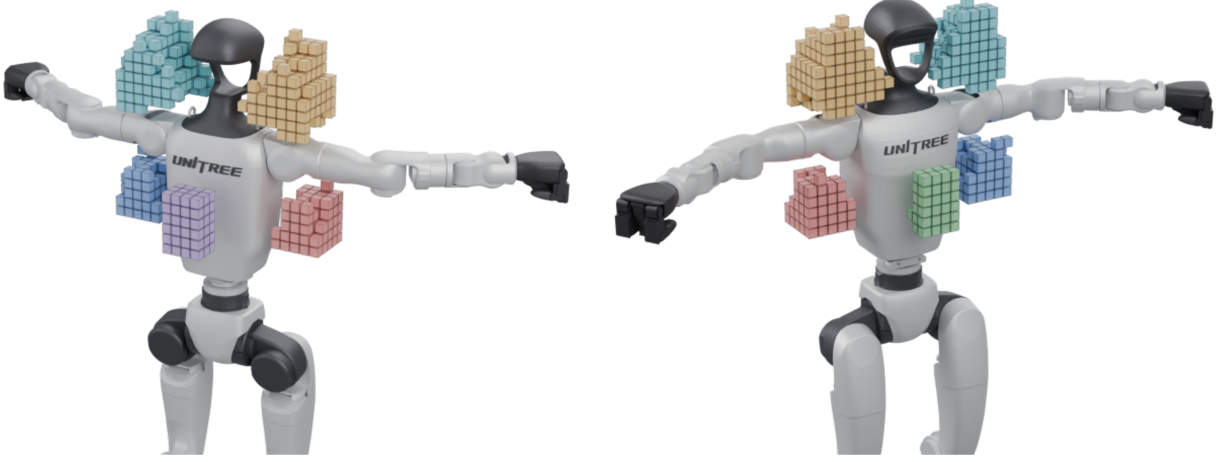


Fig. 8: **Evaluation regions.** Six torso-frame goal regions shown from two viewpoints. Required grasp modes are both arms for front and back, and the corresponding single arm for the underarms and shoulders.

B. Routes

A route is a fixed sequence of regions. The environment is initialized inside the first region, and each subsequent region becomes a commanded goal that must be reached in order. We name a route by its region sequence, abbreviating front as F, back as B, the left and right underarms as LA and RA, and the left and right shoulders as LS and RS. A route’s name therefore states its itinerary: route F_LA_B starts at the front, passes through the left underarm, and finishes at the back. Table VII lists the fourteen.

Route	Region sequence
<i>Underarm</i>	
F_LA_F	front → left underarm → front
F_RA_F	front → right underarm → front
LA_F_LA	left underarm → front → left underarm
RA_F_RA	right underarm → front → right underarm
<i>Shoulder</i>	
F_LS_F	front → left shoulder → front
F_RS_F	front → right shoulder → front
LS_F_LS	left shoulder → front → left shoulder
RS_F_RS	right shoulder → front → right shoulder
<i>Back</i>	
F_LA_B	front → left underarm → back
B_RA_F	back → right underarm → front
B_LA_B	back → left underarm → back
B_RA_B	back → right underarm → back
<i>Loop</i>	
F_LA_B_RA_F	front → left underarm → back → right underarm → front
B_LA_F_RA_B	back → left underarm → front → right underarm → back

TABLE VII: **Evaluation routes.** Fourteen routes grouped for region. Each sequence includes the initial region; subsequent regions are goals.

Regional route completion is the fraction of trials that reach every goal of a prescribed route in order before failure or timeout; partial progress along a route scores zero. The studies in Appendices IV-D, V and VI specify their own checkpoint selection, route subsets and initial-state filtering.

C. Trials and Budget

Every policy is scored against the same completion requirements regardless of how it was trained, and the dwell, position threshold and contact grace match training. Each route is evaluated with 2,048 trials at a fixed seed and a 1,500-step budget per trial, with trials retired at the first termination, giving 28,672 trials per checkpoint over the fourteen routes. Evaluation retains the object-mass, friction and center-of-mass randomization of Appendix III, freezes the goal-distance curriculum at its final value, and disables the perturbation impulses.

Object (mm)	Completion by region group (%)				All 14
	Underarm	Shoulder	Back	Loop	Mean (%)
Sphere $\varnothing 120$	75.9	39.8	69.0	61.5	61.57
Sphere $\varnothing 140$	80.7	61.4	72.8	71.6	71.63
Sphere $\varnothing 160$	85.6	61.3	80.9	80.2	76.56
Sphere $\varnothing 180$	86.1	76.0	85.4	83.2	82.59
Sphere $\varnothing 200$	88.5	77.8	85.6	86.6	84.33
Sphere $\varnothing 220$	87.9	83.2	92.2	90.4	88.14
Cube 140	72.2	57.4	45.4	19.7	52.81
Cube 160	75.6	62.6	44.1	19.8	54.91
Cube 180	65.5	59.3	41.4	11.5	49.10
Cube 200	61.3	69.7	12.5	0.0	40.98
Prism $100 \times 100 \times 200$	70.7	47.1	48.5	18.2	50.12
Prism $80 \times 160 \times 220$	47.0	29.7	10.9	0.1	25.05
Prism $140 \times 160 \times 240$	48.6	33.6	11.2	0.7	26.77
Cylinder $\varnothing 120 \times 220$	66.9	33.2	29.2	6.2	37.82
Cylinder $\varnothing 140 \times 160$	69.2	45.5	49.1	21.3	49.83

TABLE VIII: **WBD-EX expert per object.** Region-group entries are the equal-weight mean over that group’s routes; the final column is the equal-weight mean over all fourteen. Evaluations use batches of 2,048 trials per route.

D. Expert Evaluation Across Object Primitives

One WBD-EX expert is trained per object and evaluated on all fourteen routes of Table VII under the protocol of

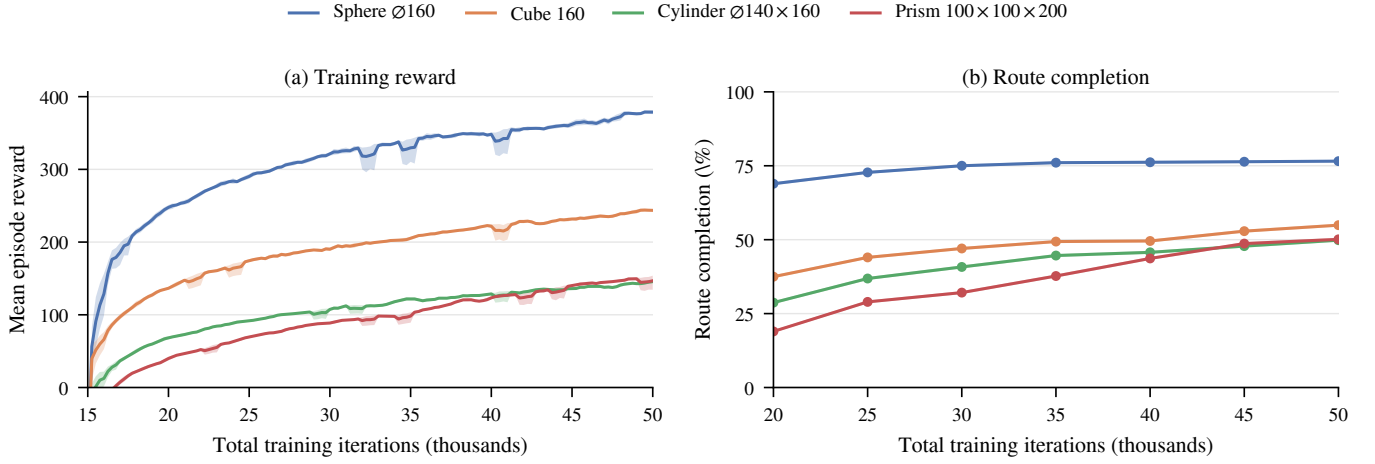


Fig. 9: **Training progress for one expert per geometry class.** (a) Mean episode reward during training, averaged over up to 1,000 logged updates; bands are approximate 95% confidence intervals (CIs) from within-window block resampling. (b) Regional route completion, averaged equally over fourteen routes at 2,048 trials per route; bands are pointwise 95% evaluation CIs. One training seed per object.

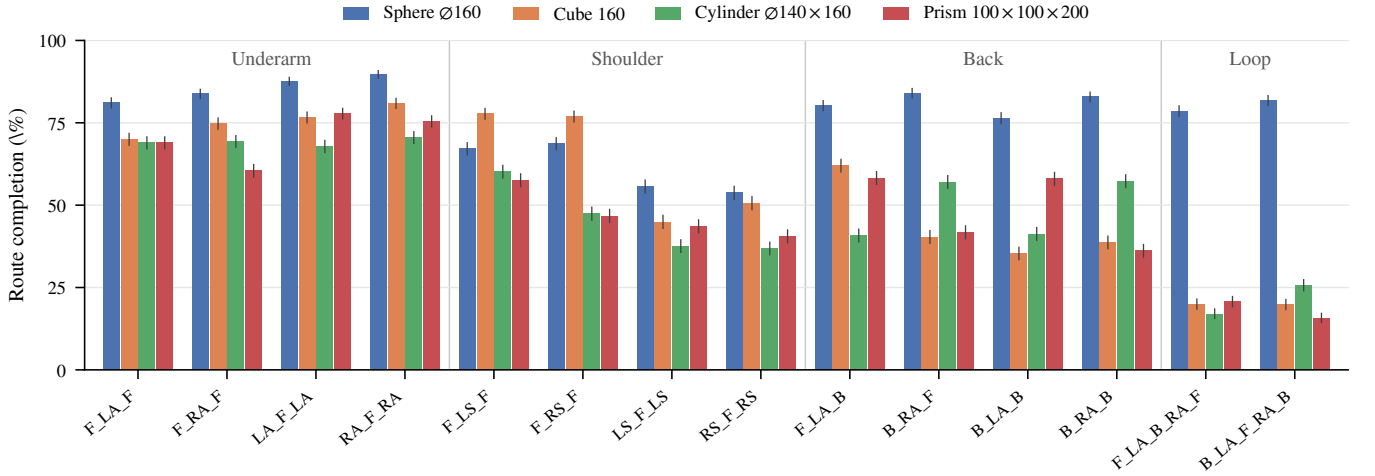


Fig. 10: **Regional route completion at 50,000 iterations** for one expert per geometry class, grouped by the body region the route visits. Error bars are pointwise 95% CIs from 2,048 trials per route. One training seed per object.

Appendix IV. Table VIII reports the measured fourteen-route mean for each object expert, grouped by the route groups of Table VII, because objects do not rank from easy to hard along a single axis: they fail in region-specific ways.

To characterize what a trained WBD-EX expert looks like, we take one object from each geometry class: the $\varnothing 160$ mm sphere, the 160 mm cube, the $\varnothing 140 \times 160$ mm cylinder and the $100 \times 100 \times 200$ mm prism. We report their training reward and regional route completion (Fig. 9) and their final per-route profiles (Fig. 10). All four are warm-started, so the horizontal axes of Fig. 9 start after the pre-trained sphere policy’s 15,000 iterations.

Completion is highest for spheres and lowest for the slender prisms and cylinder. The gap between geometry classes is largest on the most difficult trajectories: the four-goal loops, which chain the most transfers. Loop completion is near zero for the 200 mm cube and the two larger prisms ($80 \times 160 \times 220$ mm and $140 \times 160 \times 240$ mm; Table VIII). Their size may contribute to this difficulty on the G1. Although route

completion differs across object shapes and sizes, all experts supply trajectories for distillation. Continued improvement of expert success across all object shapes and sizes is left to future work.

APPENDIX V

WARM-START VERSUS TRAINING FROM SCRATCH

Appendix III describes the two-stage procedure: a sphere WBD-EX expert is trained first, and every other object’s expert is initialized from its complete training state. Here we compare that initialization against training the same object from scratch.

A. Setup and Matched Budgets

We compare warm-started and from-scratch experts for six objects: a $100 \times 100 \times 200$ mm prism, cubes of side 140 and 200 mm, a $\varnothing 140 \times 160$ mm cylinder, and spheres of diameter 140 and 160 mm. We warm-start these experts

Object (mm)	Underarm		Shoulder		Back		Loop		All 14 routes	
	Scr.	Warm	Scr.	Warm	Scr.	Warm	Scr.	Warm	Scr.	Warm
Prism 100×100×200	35.3	70.7	27.0	47.1	18.9	48.5	0.8	18.2	23.30	50.12
Cube 140	67.5	72.2	55.2	57.4	46.2	45.4	15.6	19.7	50.48	52.81
Cube 200	55.1	61.3	68.9	69.7	0.3	12.5	0.0	0.0	35.53	40.98
Cylinder Ø140×160	68.2	69.2	29.6	45.5	56.7	49.1	25.2	21.3	47.74	49.83
Sphere Ø140	81.1	80.8	39.1	60.6	72.7	73.1	71.8	71.5	65.37	71.50
Sphere Ø160	89.4	85.6	42.1	61.3	83.8	80.9	83.9	80.2	73.51	76.56

TABLE IX: **Warm-start versus from-scratch experts by body region**, at 50,000 matched total iterations with 2,048 trials per route. Entries are mean regional route completion (%), equally weighted over each group’s routes. Bold marks the higher score when the 95% evaluation CI for the difference excludes zero. One training seed per condition.

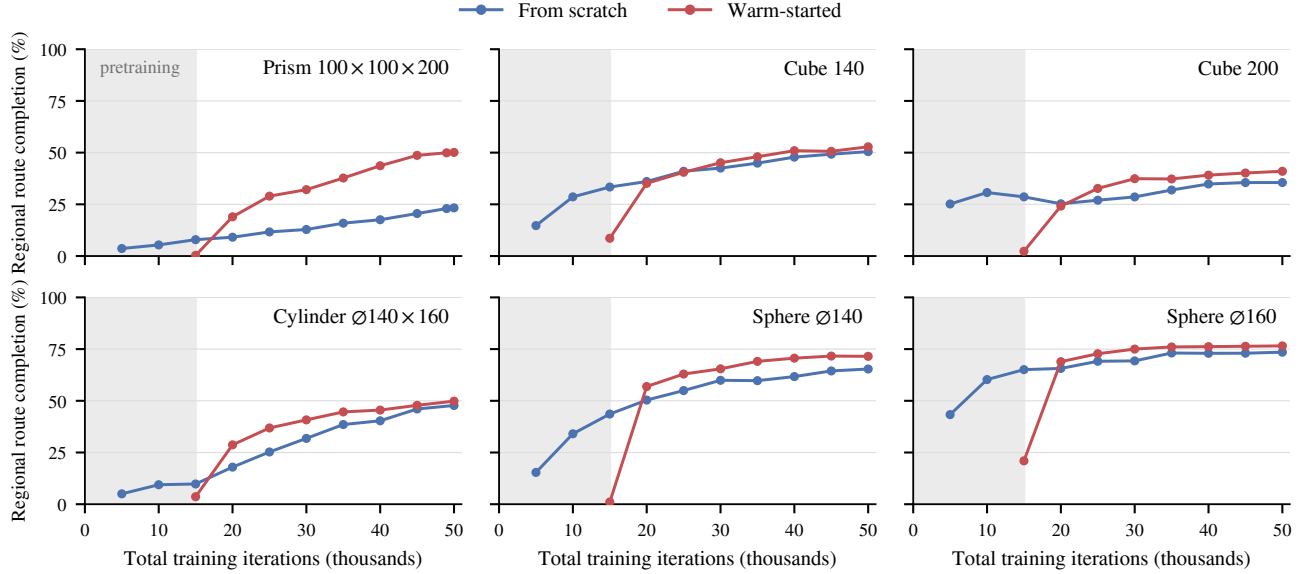


Fig. 11: **Warm-started and from-scratch learning curves**. Warm-started curves include 15,000 sphere-pretraining iterations (shaded); their first point precedes target-object training. Markers show mean regional route completion, weighting all fourteen routes equally. Bars show pointwise 95% evaluation CIs (binomial trial uncertainty or evaluation-batch bootstrap), often smaller than the markers. One training seed per condition.

from policies pre-trained on spheres with diameters of 200–220 mm. All share one training recipe and the randomization of Appendix III-F.

Budgets are matched at 50,000 total iterations. The warm-started expert is charged its pre-trained sphere policy’s 15,000 iterations plus 35,000 on its own object, so it reaches the comparison point having seen its target object for 15,000 fewer iterations. All six objects are compared on all fourteen routes.

B. Results by Body Region

Table IX breaks the comparison down by body region and Fig. 11 shows their training progress. All six warm-started experts have higher mean regional route completion, although some regional scores are lower.

Shoulder completion is higher for every warm-started expert. The prism has the largest aggregate difference, from 23.30% to 50.12%, including loop completion of 0.8% versus 18.2%. For cube 200, mean completion across the four back routes is 0.3% from scratch and 12.5% with warm-starting. Both conditions have 0.0% loop completion.

Other regions show smaller differences or regressions. The warm-started cylinder has lower back and loop completion,

and the Ø160 mm sphere has lower underarm, back and loop completion despite its higher overall mean.

C. Training Cost

The 15,000-iteration sphere pretraining cost is shared by all experts initialized from that policy. For N target experts, each post-trained for 35,000 iterations, the average budget is $35,000 + 15,000/N$ iterations per expert, compared with 50,000 iterations per expert trained from scratch. Counting the six objects here with their two pre-trained sphere policies, the total is 240,000 iterations against 300,000, a 20% reduction in aggregate iterations with all six final means higher.

Representative runs on eight NVIDIA H100 GPUs use 15,000 sphere-pretraining iterations at approximately 270 H100 GPU-hours, followed by 35,000 post-training iterations per object at 660–815 H100 GPU-hours. Training an expert from scratch for 50,000 iterations costs approximately 1,030 H100 GPU-hours. The shared pretraining contribution is therefore approximately $270/N$ H100 GPU-hours per target. In these runs, the sphere also has the lowest cost per iteration, complementing its shorter pretraining budget. These are GPU costs for the runs that produced these experts.

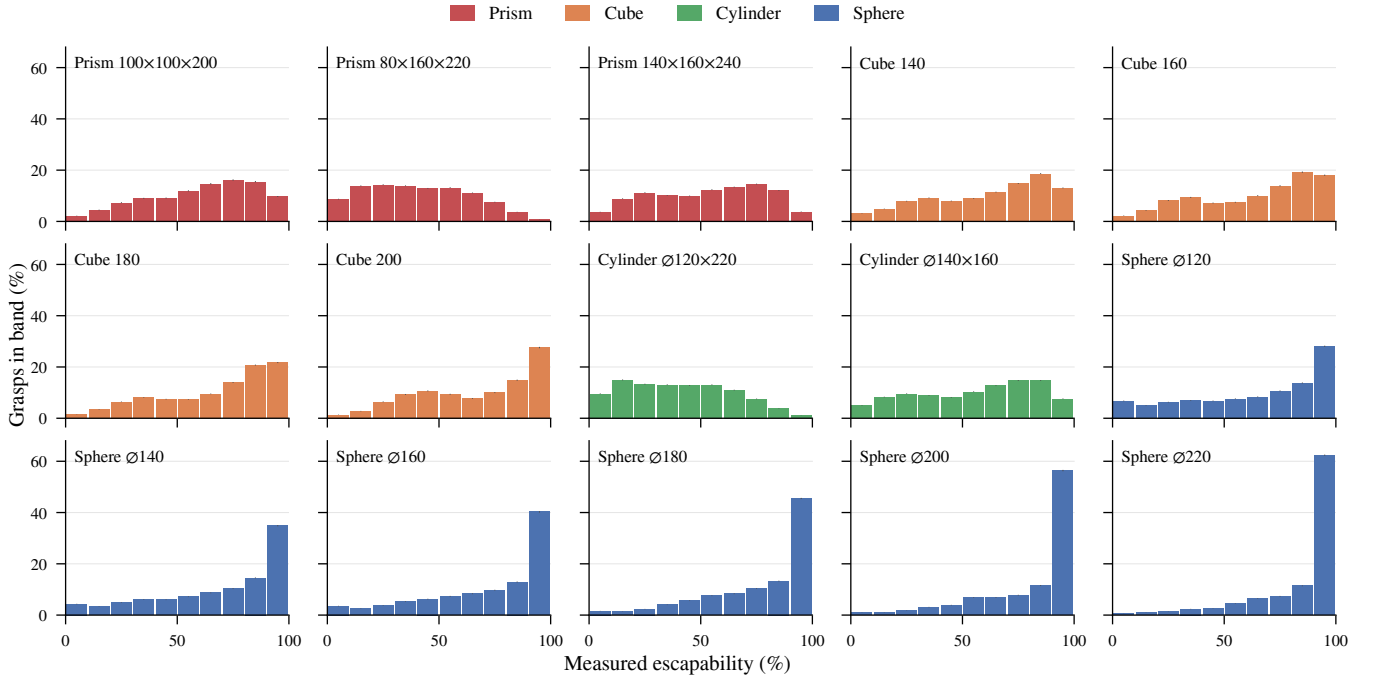


Fig. 12: **Grasp escapability distributions.** Bars give the percentage of that object’s grasps falling in each ten-point escapability band, colored by geometry class. Error bars are pointwise 95% bootstrap CIs over grasps, with measured scores held fixed; counts are exact for the scored pools.

Figure 11 shows several warm-started experts crossing the from-scratch endpoint well before their own budget is spent (the prism after 10,000 target iterations and both spheres within 15,000), suggesting additional savings if post-training stops at matched regional route completion.

The initialization transfers the actor, critic, optimizer state, observation normalizers and curriculum progress together, and their separate contributions are not isolated here.

APPENDIX VI ESCAPABILITY-BASED SELECTION OF EXPERT INITIALIZATIONS

Grasp validation (Appendix I) certifies that a grasp holds the object, not that a trained expert can reposition the object from it, and experts leave some validated grasps far more readily than others. Down-weighting low-quality grasps during WBD-EX training underperformed training on the full grasp set, so we instead score each grasp with its trained expert and filter after training, when selecting initial states for trajectory collection (Appendix VII-B). Within each voxel we remove the lowest-scoring grasps while preserving occupied voxel–grasp-mode pairs.

Policy weights, the goal pool and intermediate routing remain unchanged. We report regional route completion among retained starts to characterize the effect of this selection using the same completion requirements as the expert evaluations.

A. Escapability Scoring

For a grasp g , *escapability* is the fraction of transitions attempted from g , each to a single commanded goal, that reach the commanded position and grasp mode under the requirements of Appendix IV. Targets are drawn over the

eligible mode-compatible position voxels 6–8 cm from the source voxel’s representative position, with a budget of 3 s (150 control steps) per attempt. The score is specific to the frozen WBD-EX expert being scored. We evaluated all 2,510,864 grasps across the fifteen objects with 128 sampled targets per grasp, for 3.2×10^8 single-goal transition trials in total.

Figure 12 shows that the distribution’s shape, not just its mean, varies systematically with geometry, and that the variation is large. The two most slender objects (the $\varnothing 120 \times 220$ mm cylinder and the $80 \times 160 \times 220$ mm prism) have nearly flat distributions with mean escapability near 40% and only about 1% of their grasps in the top band. Sphere grasps are increasingly concentrated at high escapability with diameter, from 28% of grasps in the top band at $\varnothing 120$ to 62% at $\varnothing 220$, with means of 64% and 86%. Thus, grasps that pass the grasp generation simulation filtering can still have low transition success under their fully-trained expert, and the distribution of this success varies across the object-specific libraries.

B. Removal Threshold Selection

Within each original 2 cm position voxel we remove the lowest-scoring grasps up to a nominal percentage cap, retaining at least twenty where available and every original voxel and grasp-mode combination. These retention constraints mean the realized removal is usually smaller than the cap. We then re-score each expert by filtering the saved evaluation trials to the starts that survive the cut, sweeping the cap from 0 to 90%. Figure 13 reports the twelve two-goal routes of Table VII; the two loops are treated separately below because

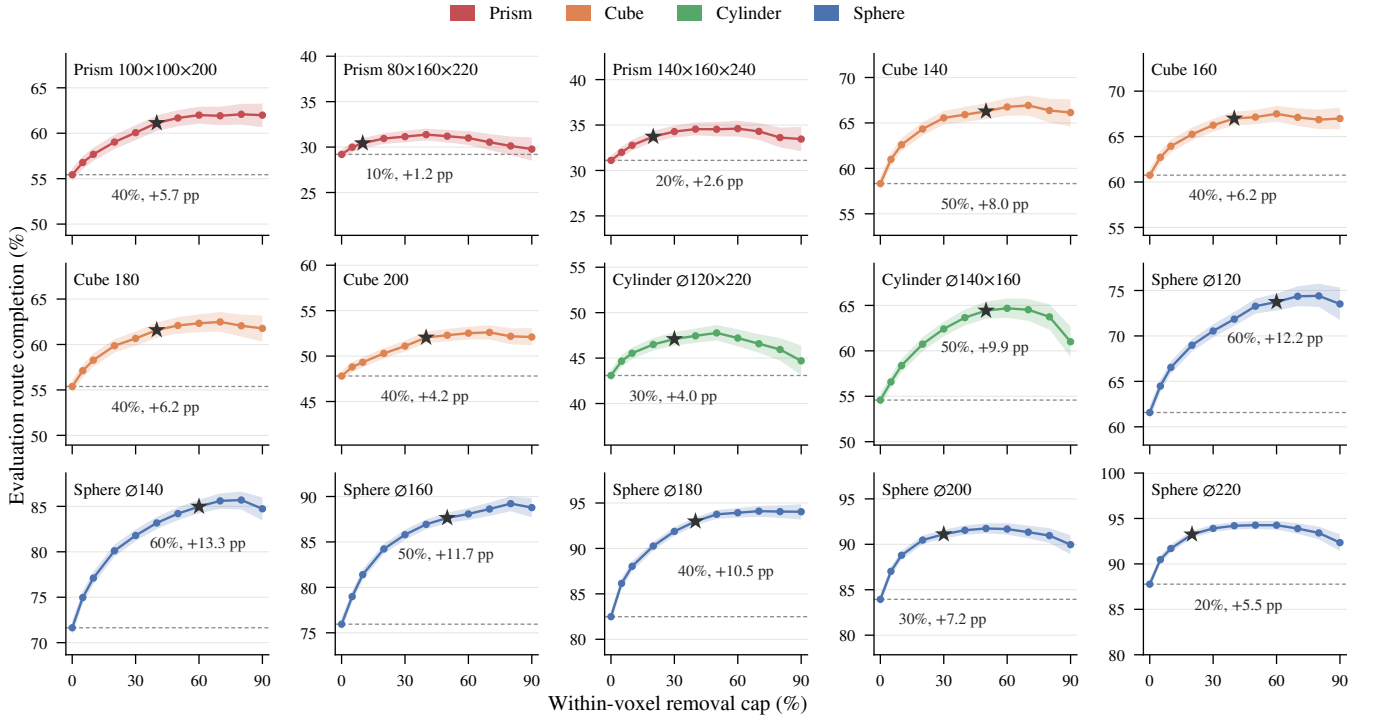


Fig. 13: **Regional route completion among retained initial grasps.** Curves equally weight the twelve two-goal routes. Bands are pointwise 95% paired bootstrap CIs (2,048 resamples over evaluation trials). Draws are shared across routes and removal caps. Dashed lines show full-library baselines; stars mark selected removal caps and gains on the selection trials (percentage points). Each vertical axis spans twenty percentage points with an object-specific offset.

their completion rates are far lower and would compress the curves.

At the adopted cuts, the mean regional route completion over two-goal routes among retained starts exceeds the full grasp dataset baseline for all fifteen objects by 1.2–13.3 percentage points (Fig. 13). The cut preserves every occupied voxel–grasp-mode pair and retains at least 98.8% of each object’s average per-joint span at every cap up to 70%. It preserves spatial and grasp-mode coverage while changing the distribution of grasp configurations within each voxel.

Repeating the sweep on the two four-goal loop routes with the same masks and completion rule, loop completion rises for eleven of the fifteen objects at the adopted cut and falls for none, by a mean of 4.8 percentage points. The largest gains are again on the spheres, from 5 to 15 points, with the Ø140 sphere rising from 71.6 to 86.5% and the Ø160 sphere from 80.2 to 92.3%.

The removal each object tolerates tracks the shape of its escapability distribution. Objects with flat distributions take small cuts for small gains: 10% removal for 1.2 points on the slender prism and 30% for 4.0 points on the slender cylinder, against 20–60% removal for 5 to 13 points on the spheres. Where escapability is uniformly moderate there is little low-escapability tail to remove, and the curve for the slender prism is nearly flat across the whole sweep. At larger removal caps, completion for the slender cylinder and the two largest spheres peaks and then declines.

We adopt, per object, the smallest cap within one percentage point of the best mean available to it, subject to a

70% ceiling and to retaining at least 95% of the original average joint-span coverage, and we keep the full grasp dataset where the interval on the gain is not positive. The adopted caps range from 10 to 60%. The gains above are measured on the trials used to select the caps. The separate trajectory collection comparison below measures usable data yield directly.

C. Trajectory Collection Yield

We compare trajectory collection from the full and escapability-filtered grasp libraries using 13,440 collection attempts per condition across all fifteen objects. Both conditions use the same frozen experts, set of eligible goals, sampling mixture, randomization and quality filter (Appendix VII-B). Each trajectory follows successive voxel-to-voxel goals until failure or a 120 s timeout. A trajectory is usable if its retained prefix passes the quality filter; failed attempts remain in the denominator. Pooling the two samplers by attempt count, usable yield increases from approximately 29.0% to 32.6% (+3.5 percentage points), and usable frames per attempt from approximately 938 to 1,053 (+12%).

The fraction of attempts failing within the first 10 s falls from approximately 26.7% to 22.9%. Retained trajectories contain approximately 3,200 frames under both conditions, suggesting that filtering mainly increases the number of retained trajectories.

The improvement in collection yield is consistent but modest. We use the filtered libraries for dataset generation because they raise the usable data per collection attempt without notably reducing the coverage of the library.

APPENDIX VII

WBD CONTROLLER DISTILLATION DETAILS

Section III-C gives the WBD controller’s architecture and its guidance mechanism. Here we specify the network settings, training data, sampling recipe and inference procedure needed to reproduce the controller.

A. Architecture and Optimization

We follow the diffusion-and-inverse-dynamics formulation and published state-encoder/U-Net configuration of [12], adapting the representation and conditioning to whole-body manipulation. Our U-Net uses one middle block, two linear layers per block and a 256-dimensional timestep embedding.

Each arm contributes four keypoints: one at the elbow and three on the hand, at the thumb, index finger and middle finger. Forward kinematics expresses these points in the torso frame using the fixed closed-fist configuration.

Each history entry contains 24 torso-frame arm-keypoint coordinates, 14 measured joint positions, 14 applied joint targets and 14 joint tracking errors. Four entries at offsets $(-15, -10, -5, 0)$, in 50 Hz control steps, and five static geometry features form the 269-dimensional diffusion input; no task-mode embedding is used. The model predicts 48 keypoint-displacement coordinates at future offsets $(5, 10)$, with targets defined as future measured keypoint positions minus their current measured positions. The inverse model takes the current history entry, predicted motions and geometry features (119 inputs), using width 1536 and four two-layer residual blocks to output 14 joint-target increments.

We train both networks jointly for 130,000 AdamW updates with a global batch of 2,560 training samples and base learning rate 3×10^{-4} , including a final 10,000-update cosine cooldown. The inverse model is supervised on an equal mixture of recorded expert motions and diffusion-generated motions, using the expert’s realized joint-target increment at the same state for both: the applied target after action scaling, tracking-error clamping and joint limits, minus the current target. Frames whose collected action was perturbed are excluded from inverse supervision. Diffusion training uses 100 timesteps with a cosine schedule (offset 0.008, β capped at 0.999).

Training the simulation and real-world controllers uses approximately 22 H100 GPU-hours per model across eight H100 GPUs.

B. Trajectory Collection

Each frozen WBD-EX expert starts from its object’s escapability-filtered grasp library (Appendix VI) and follows successive position-and-grasp-mode goals without resetting between goals. Trajectories are recorded at 50 Hz until failure or a 120 s timeout.

Random voxel-sampled routes sample eligible goal voxels and compatible grasp modes, with initial left/right/both modes in a 1:1:2 ratio. *Random region-graph routes* sample initial and target grasps from voxels in connected body regions. The first transfer follows one of sixteen directed region pairs: front–shoulder, front–underarm, shoulder–underarm

and underarm–back, in both directions and on both sides; subsequent goals follow the voxel goal graph (Appendix II). Voxel-sampled routes with at least four completed goals also supply a longer-trajectory subset. Both samplers balance the 2–4, 4–6 and 6–8 cm goal-distance bands, while keeping the maximum distance used to build the goal graph at 8 cm.

We retain usable prefixes of failed trajectories. We discard the final 100 frames leading up to failure termination, leave clean timeouts untrimmed, and require at least two completed goals and 50 retained frames. Stalled intervals are removed using 50-frame windows in which the goal identifier is unchanged, all joint spans are within 0.02 rad, the object position bounding box is within 5 mm and the maximum pairwise object rotation is within 0.1 rad. Each training sample is drawn from a contiguous 26-frame trajectory segment that supplies the required observation history and future supervision. This segment never crosses a removed interval or an episode boundary.

C. Sampling and Augmentation

Batches mix training samples from random voxel-sampled routes, region-graph routes, and the longer-trajectory subset. Each 320-sample per-GPU batch draws 125, 165 and 30 samples from these three sources and 107, 107 and 106 samples from the 2–4, 4–6 and 6–8 cm goal-distance bands, with objects equally represented, while spatial and grasp-mode sampling promote coverage of body regions and single-arm grasp goals. Training samples are drawn from eligible timesteps throughout each retained trajectory; spatial sampling uses the object’s position at each sample’s current frame.

Each initial-grasp pool entry is assigned to validation with probability 5% by a fixed hash, giving 5.45% of initial-grasp groups in the simulation controller. Trajectories sharing an initial-grasp identity, their derived segments, and augmented views inherit the same assignment.

Half of all training samples are reflected across the sagittal plane, applied as one coherent transform over history, actions and future targets in physical coordinates before normalization, with validation and evaluation left unaugmented. All retained trajectories are collected under randomized dynamics and the corpus contains no nominal-physics fraction. Mass, center-of-mass and friction randomization follow Appendix III-F; the experts’ perturbations remain enabled at their trained 20 N cap. The randomized parameters are not controller inputs.

D. Simulation Evaluation Routes

For the simulation evaluation in Fig. 4, the region graph connects front to each underarm and shoulder, and each underarm to the shoulder on the same side and to the back, in both directions. These define 16 directed transfers across the six regions in Fig. 8; direct shoulder–back and cross-side connections are excluded. After each regional transfer, the next adjacent region is sampled uniformly using a fixed random seed. Dead ends and other failures remain included.

E. Command-Smoothing Comparison

We test whether simple temporal smoothing can reproduce WBD’s gains under command corruption. We compare the noisy expert, exponential moving average (EMA) smoothing of its commands, and WBD refinement. For corrupted raw arm commands u_t , EMA applies $z_t = (1 - \beta)z_{t-1} + \beta u_t$ before action bounding and scaling, initialized with $z_0 = u_0$ and retained across goal changes. We select $\beta = 0.5$ from $\{0.05, 0.1, 0.25, 0.5\}$ by mean completed regional transfers on 64 separate validation starts per training object and noise level, weighting the 13 objects and two noise levels equally. Test starts and their mirrored families are excluded from selection. The same coefficient also ranks first by goal count on these validation starts.

This separate comparison repeats front $\rightarrow$ left underarm $\rightarrow$ front $\rightarrow$ right underarm at $\alpha \in \{2.5, 4\}$, using 256 matched starts per object and terminating at the first failure or 600 s. We use the goals-per-trial metric defined in Section IV-B. All failures remain included. WBD uses guidance 100; the expert checkpoints, action scaling, joint limits and task physics are shared across conditions.

α	Method	Goals / trial	Time to failure (s)
2.5	Noisy expert	8.47	29.1
	Noisy + EMA	7.82	20.5
	Noisy + WBD	24.76	88.7
4	Noisy expert	6.24	24.2
	Noisy + EMA	6.32	17.6
	Noisy + WBD	21.77	63.9

TABLE X: **Command smoothing versus WBD refinement.** Means weight the 13 training objects equally.

The selected EMA filter yields similar goal counts and shorter time to failure than the noisy expert. WBD improves both metrics over this filter at both noise levels (Table X).

F. Real-Time Inference

Each inference update samples 16 motion candidates using 12 DDIM denoising steps at timesteps 88, 80, $\dots$, 0. The guidance request expresses keypoint displacement from the current commanded joint configuration to the requested configuration. Predicted motions are instead measured from the measured configuration; we do not convert between the two references, and the joint tracking error is part of the controller input.

Command chain. Let c be the current applied joint target and q the measured joint configuration. In simulation, the noisy raw expert action $\tilde{a} = a + \epsilon$, $\epsilon \sim \mathcal{U}(-\alpha, \alpha)^{14}$, defines the requested configuration $q^{\text{req}} = \Pi(c + 6 \cdot 0.025 \tanh \tilde{a})$, where Π applies the 0.35 rad tracking-error clamp around q and the joint limits. On hardware, q^{req} is the retargeted operator configuration clipped to the joint limits. The guidance displacement is $\mathbf{d} = \text{FK}(q^{\text{req}}) - \text{FK}(c)$. The inverse model predicts an increment, giving $\hat{c} = c + \Delta \hat{c}$. In simulation, $\hat{c}$ is mapped back to a raw action $\hat{a} = \text{atanh}(\text{clip}((\hat{c} - c)/0.025, \pm 1))$, clipped to $[\tilde{a} - 2.5, \tilde{a} + 2.5]$, and applied through the experts’ action pipeline $\Pi(c + 0.025 \tanh \hat{a})$;

this ± 2.5 raw-action radius is the bounded correction in Section IV-B. On hardware, the increment is clipped to ± 0.025 rad per joint, and the published target is limited to 1.25 rad/s, to within 0.35 rad of q , and to the joint limits.

With $\mathbf{d} \in \mathbb{R}^{24}$ the guidance displacement defined above, for predicted keypoint displacements $\Delta \mathbf{x}_5$ and $\Delta \mathbf{x}_{10}$ at the two future offsets, guidance minimizes

$$D = \|\Delta \mathbf{x}_5 - \mathbf{d}/2\|_2 + \|\Delta \mathbf{x}_{10} - \mathbf{d}\|_2.$$

Each norm spans all keypoint coordinates in physical space. We select the generated candidate with the smallest summed squared error to these targets.

Inference and communication delay can make the command used to compute a prediction outdated by the time that prediction arrives. We therefore separate the applied command in the observed state from the next command predicted by the inverse model. Let c^{obs} denote that observed command and $\hat{c}$ the predicted command. A separate 50 Hz publisher applies their difference once to its latest command c^{pub} :

$$c^{\text{new}} = c^{\text{pub}} + (\hat{c} - c^{\text{obs}}). \quad (5)$$

This preserves the predicted increment when its original command reference has become stale. Between predictions, the publisher continues toward the resulting target without accumulating the same increment again. State histories contain the commands actually applied and timestamped measurements at the offsets in Appendix VII-A, independently of inference timing. On hardware, inference runs on an NVIDIA RTX 4090 GPU on the operator workstation and is capped at 50 Hz. With these sampler settings, one inference update took 11.8 ms median (17.3 ms 95th percentile), updates arrived at 34.9 Hz on average, and the delay from state publication to the robot’s command buffer was 50.1 ms median (73.7 ms 95th percentile), excluding sensing and motor latency.