

Goal Revisions and Support Failures in Contact-Rich Whole-Body Skill Composition

Visak Kumar

Independent Researcher

San Francisco, CA, USA · vishak.vk@gmail.com

Project page and video: <https://visakk.github.io/pose-is-not-support>

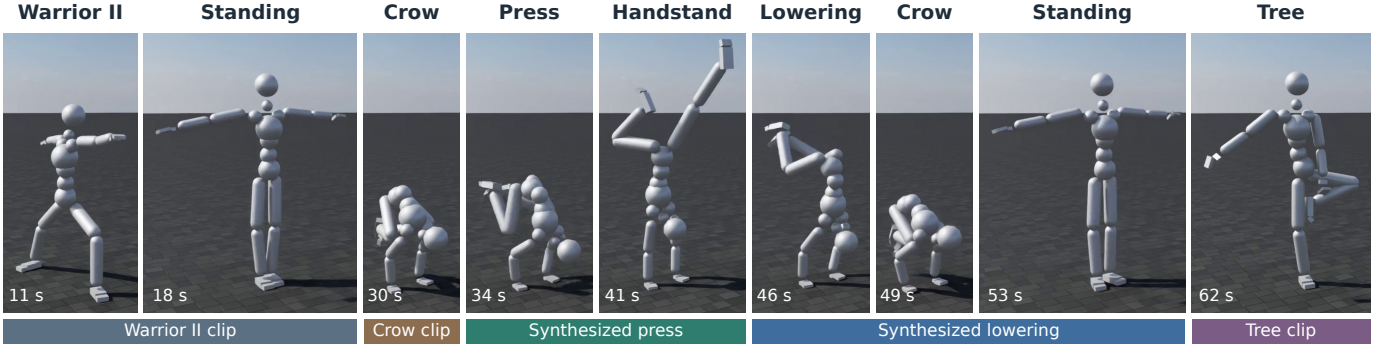


Fig. 1. One rollout of the switch-trained student (seed 0) following a plan that joins five clips without resetting the simulation: Warrior II, crow, a synthesized press to handstand, a synthesized lowering back to crow, and tree. Bars mark the clip that supplies the goals. Consecutive clips meet at standing, crow, or handstand, and switch training held out the plan’s two standing pairings. Every designated hold here passes the pose-and-load rule. Over 409 repeated rollouts, this student completes the whole plan in 258, the other switch-trained seed in 23, and the initial student in none.

Abstract—Contact-rich loco-manipulation requires a whole-body controller to change which limbs carry its weight as a task unfolds, often after the next movement has been revised. A controller that performs two skills individually may still fail when the second is requested during the first. We study this gap in a simulated yoga repertoire of arm balances, inversions, and single-leg supports, distilled online from a privileged teacher into a student that accepts partial pose-and-contact goals. During distillation, we replace a clip’s continuation with a compatible alternative at randomized times, preserving physical state and history, while the teacher, student architecture, and objective stay fixed. Across two fine-tuning seeds per condition, completion of all designated targets in multi-hop chains reaches 65% and 79%, versus below 1% and 18% without switches, and sequences through standing pairings excluded from switch training reach 32% and 21%, versus zero for both controls, while familiar-route success stays near 66%. A reveal-time sweep separates two remaining failures: returning to the original continuation and arriving after the deadline. Requested-support failures persist after switching: pose-only success exceeds pose-and-load success by 14–16 percentage points in the switched students and by 19.5 in the teacher.

I. INTRODUCTION

A humanoid that braces against a wall to push an object, shifts its weight to free a hand, or rises from the floor must change which limbs carry its weight as the task proceeds, sometimes after a planner has revised what comes next [1], [2]. Consider a simulated humanoid holding crow, an arm balance with its feet off the ground. Its familiar motion ends by returning to standing, but a new command asks it to jump back into plank. The controller must preserve the crow hold and then reach plank on schedule, starting from the physical state it

has already reached. Knowing both motions individually does not ensure that it can execute this revised sequence.

Motion coverage and command coverage are different. Labeling the states a student visits [3] or perturbing the expert during demonstration [4] broadens the states a policy learns from, yet a clip-local schedule still pairs each arrival with the same next goal. Supervised learners can generalize poorly to state–goal combinations absent from their data, and relabeling goals across trajectories that meet at nearby states helps [5]; a student that can predict its labels from state and history has little reason to read the goal [6], [7].

We test whether compatible goal switches during collection, an online form of this relabeling, improve composition while the teacher, student architecture, and distillation objective stay fixed. Human yoga capture [8] provides arm balances, inversions, and single-leg balances whose support shifts among feet, hands, forearms, head, and trunk: the non-prehensile, environment-assisted supports of loco-manipulation, without object dynamics.

Switch collection substantially improves multi-hop completion and transfers to standing pairings excluded from collection (Fig. 1), while familiar-route performance stays near its starting level. Asking when a revised command can be followed, whether its deadline is met, and whether requested supports carry load reveals timing-sensitive failures and unloaded supports that remain. The study contributes a controlled measurement of revised-goal experience in online distillation and an evaluation that separates sequence completion, revision timing, and realized loading.

TABLE I

SKILL COMPOSITION AND FAMILIAR-ROUTE PERFORMANCE AT THE SAME FINE-TUNING BUDGET. RATES USE THE POSE-AND-LOAD RULE EXCEPT THE EXPLICIT POSE-ONLY ROW. “ALL TARGETS” REQUIRES EVERY DESIGNATED HOLD TO PASS IN ONE ROLLOUT, NOT EVERY TRANSITION FRAME. STANDING PAIRINGS ARE EXCLUDED FROM SWITCH TRAINING. STUDENT RATES POOL 425 (CHAINS, DESTINATIONS), 2,045 (FLOWS), OR 1,968 (ROUTES) ROLLOUTS. THE TEACHER IS A CAPABILITY REFERENCE WITH DIFFERENT CONDITIONING AND REPLICA COUNTS.

	Teacher	Initial	No switch 0	No switch 1	Switch 0	Switch 1
Standing flows: all targets	.574	.000	.000	.000	.320	.213
Standing flows: second-clip target	.755	.011	.000	.001	.558	.409
Multi-hop chains: all targets	.800	.118	.007	.184	.649	.786
Five revised destinations	1.000	.431	.285	.671	.784	.944
Familiar-route target success	.777	.663	.567	.690	.666	.665
Familiar-route pose-only success	.972	.820	.681	.865	.810	.823
Familiar-route passing plans	32/41	27/41	25/41	28/41	29/41	28/41

II. RELATED WORK

Reusable motion interfaces. Physics-based imitation and adversarial motion priors learn diverse whole-body behavior [9], [10], and privileged trackers are commonly distilled into deployable students by online teacher labeling [11], [12]. MaskedMimic [13] distills a motion-inpainting controller with partial constraints and sequences goals with finite-state machines at deployment; HOVER [14] samples command modes and sparsity masks once per episode; PLAT [15] tracks sparse timed keyframes through a DAgger-trained latent prior refined by residual reinforcement learning. During distillation, these systems draw their commands from the motion being tracked.

Changing commands during training. Controllers that respond to users are often trained under changing commands. DReCon [16] drives its kinematic target with an artificial user that changes direction, speed, and style at random, and Walk These Ways [17] resamples the commanded behavior within training episodes. Switch [18] augments cross-skill connections, starts episodes near transitions, and schedules switching online. SkillMimic composes learned interaction skills with a high-level controller [19], SkillMimic-V2 expands transition coverage by stitching demonstrations into a trajectory graph [20], and motion graphs, sequential composition, and skill chaining connect compatible behaviors more generally [21]–[23]. We measure this principle inside online distillation, where all runs share the teacher, student architecture, objective, and initial weights and optimizer state, under prescribed timed goals rather than an adaptive planner.

Contact-rich whole-body control. Humanoid controllers increasingly use contacts beyond the feet, from contact-staged tasks [24] to getting up from the floor [25], [26]. Omni-Contact [1] chains skills through contact-flow synthesis and replanning, bracing lets a humanoid apply larger manipulation forces [2], and residual learning extends general motion tracking to loco-manipulation [27]. ContactMimic [28] observes that a tracker can reach the correct pose without making meaningful contact with an object, and decouples contact from keypoints with contact-following rewards and trajectory augmentation; we measure the same distinction for ground supports through requested loading.

Adaptation beyond tracking. HIL [29] combines tracking with adversarial imitation and attributes its tracking-only base-

line’s failure on new transitions to conditions sampled from reference sequences; motivated by that diagnosis, we change only how conditions are sampled during distillation.

III. TESTBED AND REVISED-GOAL COLLECTION

A. Repertoire and policy interfaces

We use 56 human clips covering 37 pose families and repair their geometry using the performer’s body, capture, and contact measurements. A graph joins consecutive held configurations, each recorded as a pose, a ground-contact set, and a pelvis-orientation bin; standing connects otherwise separate families. An MPPI-style optimizer [30] synthesizes five missing transitions among crow, handstand, plank, and chaturanga, such as a press from crow to handstand and a jump from crow to plank. Splicing 28 of its trajectories into human lead-ins and exits adds 28 clips, for 84 base clips on 123 nodes and 201 edges, each also with one hold extended by 3 or 7 s. Appendices A–B describe reference construction and its checks.

Both policies control a 24-body, 74 kg avatar through 69 PD targets at 30 Hz in IsaacLab/PhysX [31], [32]. The PPO teacher [33] receives state, contacts, pose history, and two timed full-body goals; dense references supervise its rewards and critics, not its actor. The student receives five masked six-body pose-and-contact goal slots, proprioception, pose history, and contact events. A privileged encoder and a deployable prior supply one of $5^4 = 625$ FSQ codes [34] to a shared decoder with no direct goal input; codes refresh every eight control steps or at contact events. Online teacher labeling [3] supervises student-visited states, and a quarter of environments execute the sampled prior. Appendices C–D specify the interfaces and losses.

B. Training with revised continuations

During initial distillation, goals follow the playing clip’s fixed schedule, so a continuation never changes while its segment is in progress. Switch collection adds randomized events that replace the continuation at standing, at shared holds such as crow and handstand, at the ends of synthesized transitions, or between two recordings of one pose, while physical state, pose history, and contact-event history persist. The playing reference itself changes, so the goals,

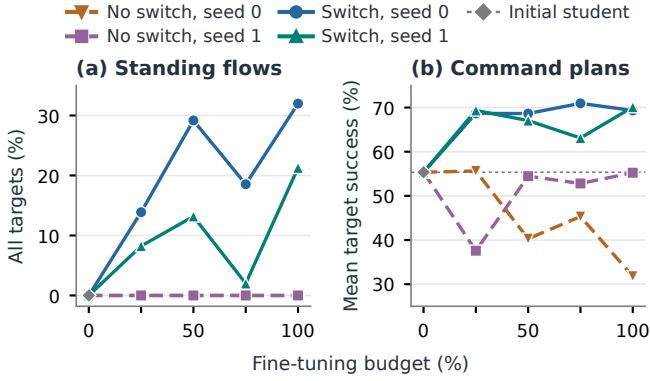


Fig. 2. Composition during fine-tuning. (a) All-target completion of the five standing flows, whose standing pairings are excluded from switch collection; both no-switch runs stay at zero. (b) Mean target success over the 19 other cross-clip plans: revised destinations with their own-route controls, chains, and within-family alternates. Each line is one run from the shared initial student. Replica counts are in Appendix H.

the teacher’s inputs, the privileged encoder’s dense reference, and the tracking-termination reference all follow the revised command. Changing the goals alone would not suffice: the encoder, whose codes the prior learns to predict, and the termination check would still follow the original clip, training the student to ignore revisions and cutting short the episodes that obey them. Up to three switches per episode, at least 2 s apart, form short graph walks; the nominal switch probability is 0.5 over each eligible interval (Appendix E).

Two switch and two no-switch runs start from the same weights and optimizer state and share architecture, loss, and a 1,000-epoch budget; their configurations differ only in the switch settings and the seed. The teacher stays frozen, and the student receives no reinforcement-learning update. The intervention changes continuation exposure, reveal timing, and state visitation together.

IV. EVALUATION

Tasks and comparisons. A cross-clip plan follows one clip to a shared hold and then, without resetting state, commands another clip’s continuation; for example, the human crow clip, which returns to standing, is revised to the synthesized jump to plank. We evaluate 41 familiar routes, which replay a clip’s own schedule; five revised destinations with two own-route controls; five multi-hop chains through the synthesized transitions; five standing flows; and seven within-family alternates between two recordings of one pose. Standing flows join clips through returns to standing; the floor sequence, for example, visits downward dog, plank, side plank, and chaturanga. Their 11 directed standing pairings, including dwell variants, are excluded from switch collection, whereas the revised destinations and chains are trained switch types. These plans are development probes, not an independently reserved test set, and the initial student was selected on them; Appendix G details the plan groups and scoring.

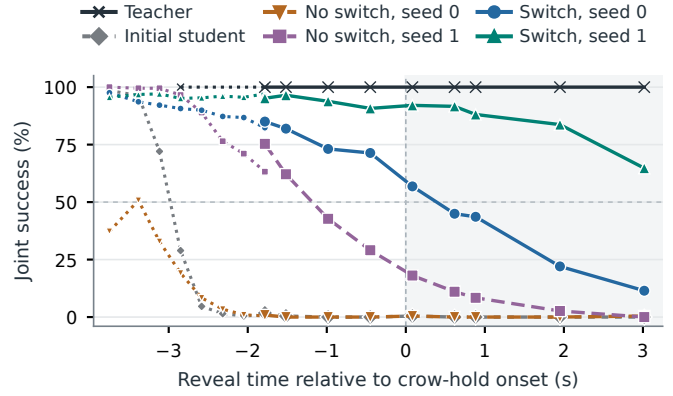


Fig. 3. Timing of crow-to-plank revisions at the fixed fine-tuning endpoints. Joint success requires preserving crow and meeting the scheduled plank target. The extended sweep uses 227 replicas per offset; dotted curves show earlier batches, which differ by up to .12 at -1.78 s, the one shared offset. Shading marks reveals after crow-hold onset; Fig. 8 breaks down the outcomes.

At the common endpoint, student command, flow, and route evaluations use 85, 409, and 48 replicas per plan, rollouts from the same start under the same commands, and sample the deployable prior. The teacher uses 31 replicas, its mean action, and fuller conditioning, so its replicas differ only through simulator nondeterminism. Each controller reaches the revision point along its own trajectory. Replicas estimate rollout variation, not training variability, and repeated batches of one plan can differ by up to .16 (Appendix I).

Pose-and-load success. Each designated hold requires a hold-averaged six-body position error below 0.15 m after pelvis-relative, per-frame best-yaw alignment; every requested ground zone carrying at least 5 N on at least 90% of frames; and no unrequested zone carrying at least 3% of body weight on at least 20% of frames. A plan passes when at least half its replicas succeed. The rule checks local pose and the persistent loading of each requested ground zone; it does not score body-pair braces or transition quality. Passing the load threshold establishes contact loading rather than a particular force distribution; missing it is a requested-support failure even when the controller stays balanced (Appendix G-A).

V. RESULTS

A. Revised-goal experience improves composition

The teacher completes all designated standing-flow targets in 57% of replicas; the initial student completes none (Table I). Switch collection raises completion to 32% and 21%, while both controls stay at zero throughout fine-tuning (Fig. 2a), and second-clip target success rises from 1% to 56% and 41%, against at most 2.1% across either control’s checkpoints. Standing is the most frequent switch site in training (67–68% of switches), and every clip in an excluded pairing also appears in the same role in trained pairings, so this is transfer to new pairings at a familiar hub rather than to a new kind of junction. Figure 1 shows one completed routine.

Multi-hop completion reaches 65% and 79%, versus below 1% and 18% for the controls; revised-destination success

reaches 78% and 94%, versus 28% and 67%. The stronger control acquires useful individual revisions, but averaged over the 19 command plans other than the flows, both switch-trained runs exceed both controls at every saved checkpoint (Fig. 2b). These composition rates change by at most .05 for pose thresholds from 0.10 to 0.20 m. Within-family alternates do not improve consistently, neither the five revision types absent from switch training (initial .21; switched .29 and .20) nor the two present (.54; .50 and .49). Familiar-route success stays near the initial 66% in both switched runs, with individual skills gained and lost; the stronger control reaches 69% while remaining weaker on composition.

B. Following a revision and meeting its deadline

Returning to the opening example, we vary when the crow-to-plank command is revealed while keeping the crow hold and the 0.30 s plank window fixed; joint success requires both. Later reveals leave less preparation time and more evolution under the old standing goal.

Both switched endpoints outperform both controls at all nine offsets (Fig. 3), and every switched destination success also preserves the source hold. At the latest reveal, joint success is 11% and 65%, versus 0–0.4% for the controls; the teacher succeeds at every offset from its own arrivals. A criterion fixed before training, at least .80 on three late crow-to-plank tests at the final checkpoint, is met by neither seed, although each meets it at two of its four saved checkpoints (Appendix H).

The two seeds fail differently. Seed 0 increasingly returns to the original standing continuation: at the latest reveal, 184 of its 201 failures stand through the plank window. Seed 1 follows the request but can arrive late. During the window its failures are already near plank, at a median six-body error of 0.18 m against the 0.15 m threshold and with all four requested supports loaded in 73 of 80, and 79 of the 80 eventually attain plank, 77 only after the window closes. With a 0.20 m threshold, seed 1 would succeed on 217 of 227, while seed 0’s failures are unchanged. These later attainments remain scheduled failures, but they distinguish a missed deadline from a return to the old goal (Appendix I).

C. Accurate poses can leave requested supports unloaded

On familiar routes, the switched students’ pose-only success exceeds their pose-and-load success by 14–16 percentage points, close to the initial student’s 16 and between the controls’ 11 and 17; switch collection does not remove requested-support failures. The teacher shows the same distinction, at 97% by pose alone and 78% with loading conditions, and all 248 replicas in its gap miss requested loading. Unloaded supports can sit within 2 cm of the floor: in Cobra, the teacher reaches 3 cm pose error while its requested thighs hover 3–8 mm above the floor and carry no load (Fig. 4). Pose accuracy and proximity both miss this failure.

VI. DISCUSSION

Why the student needs revised experience. The frozen teacher, trained on the same clip-local schedules, reaches every

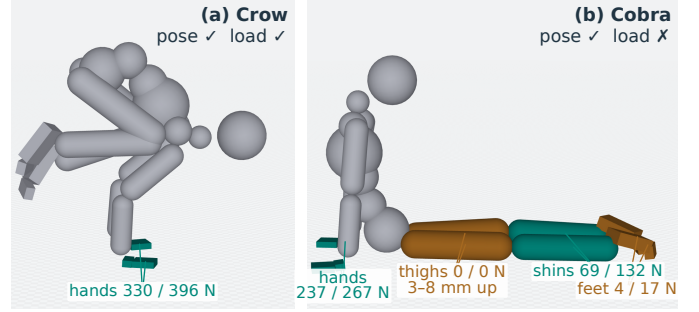


Fig. 4. Pose agreement and requested loading differ. A requested zone passes if it carries at least 5 N on at least 90% of hold frames, checked per zone; teal zones pass and amber zones fail in the drawn replica. Labels give left / right hold-mean loads (medians over 31 replicas). The teacher’s Cobra has 3 cm pose error, yet its requested thighs carry no load in any replica (Appendix J).

revised destination and succeeds at every tested reveal time from its own arrivals, so the gap lies in what the student learned. Its decoder receives the goal only through a held code predicted by the prior. One explanation is a shortcut known from imitation with observation histories [6], [7]: under clip-local collection the requested continuation is fixed by the playing clip and its clock, both of which state and two seconds of history can reveal, so the prior can predict the encoder’s codes without reading the goal. Compatible switches weaken this regularity, because a familiar arrival is now followed by different continuations that only the goal distinguishes, though the experiments do not isolate this mechanism. Resetting the initial student’s history and intent at a revision does not rescue it (Appendix I), so the failure does not lie in an input that can be cleared at test time.

Open problems. Following a revision and meeting its deadline can fail separately, so planners that revise goals online need controllers with predictable response times. Requested supports must carry load, not merely be approached, before a limb can be released or a force transmitted; the teacher’s reward penalizes unwanted loading but never rewards requested support, and its students show the same failures. Pose and contact requests co-vary in human capture, so independent contact control needs data or objectives that decouple them [28]. Objects, external wrenches, real-world contact sensing, and hardware lie outside this study.

Scope. The comparison uses one performer, body, simulator, teacher, and pretrained student, with two seeds per condition; the student’s goal passes through a held discrete code, so other representations need testing. Student task optimization and support-attainment rewards are complementary directions [35], [36] (Appendix L).

VII. CONCLUSION

Compatible goal switches during distillation substantially improve composition, including at standing pairings excluded from collection, while familiar-route success stays near its starting level. Meeting revised deadlines and loading requested supports remain open for these controllers.

REFERENCES

- [1] R. Yu *et al.*, “OmniContact: Chaining meta-skills via contact flow for generalizable humanoid loco-manipulation,” arXiv:2606.26201, 2026.
- [2] Z. Zhang, C. Lehnert, W. N. Browne, and J. M. Roberts, “Brace yourself: Task-conditioned environmental bracing for forceful humanoid manipulation,” arXiv:2609.25486, 2026.
- [3] S. Ross, G. Gordon, and D. Bagnell, “A reduction of imitation learning and structured prediction to no-regret online learning,” in *Proc. Int. Conf. Artif. Intell. Statist. (AISTATS)*, 2011, pp. 627–635.
- [4] M. Laskey, J. Lee, R. Fox, A. Dragan, and K. Goldberg, “DART: Noise injection for robust imitation learning,” in *Proc. Conf. Robot Learn. (CoRL)*, vol. 78, 2017, pp. 143–156.
- [5] R. Ghugare, M. Geist, G. Berseth, and B. Eysenbach, “Closing the gap between TD learning and supervised learning—A generalisation point of view,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2024.
- [6] P. de Haan, D. Jayaraman, and S. Levine, “Causal confusion in imitation learning,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2019.
- [7] C. Wen, J. Lin, T. Darrell, D. Jayaraman, and Y. Gao, “Fighting copycat agents in behavioral cloning from observation histories,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2020.
- [8] S. Tripathi, L. Müller, C.-H. P. Huang, O. Taheri, M. J. Black, and D. Tzionas, “3D human pose estimation via intuitive physics,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023.
- [9] X. B. Peng, P. Abbeel, S. Levine, and M. van de Panne, “DeepMimic: Example-guided deep reinforcement learning of physics-based character skills,” *ACM Trans. Graph.*, vol. 37, no. 4, 2018.
- [10] X. B. Peng, Z. Ma, P. Abbeel, S. Levine, and A. Kanazawa, “AMP: Adversarial motion priors for stylized physics-based character control,” *ACM Trans. Graph.*, vol. 40, no. 4, 2021.
- [11] Z. Luo *et al.*, “Universal humanoid motion representations for physics-based control,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2024.
- [12] T. He *et al.*, “OmniH2O: Universal and dexterous human-to-humanoid whole-body teleoperation and learning,” in *Proc. Conf. Robot Learn. (CoRL)*, vol. 270, 2025, pp. 1516–1540.
- [13] C. Tessler, Y. Guo, O. Nabati, G. Chechik, and X. B. Peng, “Masked-Mimic: Unified physics-based character control through masked motion inpainting,” *ACM Trans. Graph.*, vol. 43, no. 6, 2024.
- [14] T. He *et al.*, “HOVER: Versatile neural whole-body controller for humanoid robots,” in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2025, pp. 9989–9996.
- [15] Z. Wang, J. Wang, C. Ma, X. Shi, and Z. Lu, “PLAT: Sparse timed keyframe motion tracking for humanoid control via privileged latent transition learning,” arXiv:2609.25754, 2026.
- [16] K. Bergamin, S. Clavet, D. Holden, and J. R. Forbes, “DReCon: Data-driven responsive control of physics-based characters,” *ACM Trans. Graph.*, vol. 38, no. 6, 2019.
- [17] G. B. Margolis and P. Agrawal, “Walk these ways: Tuning robot control for generalization with multiplicity of behavior,” in *Proc. Conf. Robot Learn. (CoRL)*, vol. 205, 2023, pp. 22–31.
- [18] Y.-F. Lau *et al.*, “Switch: Learning agile skills switching for humanoid robots,” in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2026, pp. 22476–22483.
- [19] Y. Wang *et al.*, “SkillMimic: Learning basketball interaction skills from demonstrations,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2025.
- [20] R. Yu *et al.*, “SkillMimic-V2: Learning robust and generalizable interaction skills from sparse and noisy demonstrations,” in *Proc. SIGGRAPH Conf. Papers*, 2025.
- [21] L. Kovar, M. Gleicher, and F. Pighin, “Motion graphs,” *ACM Trans. Graph.*, vol. 21, no. 3, pp. 473–482, 2002.
- [22] R. R. Burridge, A. A. Rizzi, and D. E. Koditschek, “Sequential composition of dynamically dexterous robot behaviors,” *Int. J. Robot. Res.*, vol. 18, no. 6, pp. 534–555, 1999.
- [23] Y. Lee, J. J. Lim, A. Anandkumar, and Y. Zhu, “Adversarial skill chaining for long-horizon robot manipulation via terminal state regularization,” in *Proc. Conf. Robot Learn. (CoRL)*, vol. 164, 2022, pp. 406–416.
- [24] C. Zhang, W. Xiao, T. He, and G. Shi, “WoCoCo: Learning whole-body humanoid control with sequential contacts,” in *Proc. Conf. Robot Learn. (CoRL)*, vol. 270, 2025, pp. 455–472.
- [25] X. He, R. Dong, Z. Chen, and S. Gupta, “Learning getting-up policies for real-world humanoid robots,” in *Proc. Robot.: Sci. Syst. (RSS)*, 2025.
- [26] T. Huang *et al.*, “Learning humanoid standing-up control across diverse postures,” in *Proc. Robot.: Sci. Syst. (RSS)*, 2025.
- [27] S. Zhao *et al.*, “ResMimic: From general motion tracking to humanoid whole-body loco-manipulation via residual learning,” arXiv:2510.05070, 2025.
- [28] X. Li, X. He, R. Dong, and S. Gupta, “ContactMimic: Humanoid object interaction via contact control,” arXiv:2607.08742, 2026.
- [29] J. Wang *et al.*, “HIL: Hybrid imitation learning for dynamic athletic control,” *ACM Trans. Graph.*, vol. 46, no. 2, Apr. 2027, published online Oct. 2026, doi: 10.1145/3829364.
- [30] G. Williams *et al.*, “Information theoretic MPC for model-based reinforcement learning,” in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2017, pp. 1714–1721.
- [31] NVIDIA: M. Mittal *et al.*, “Isaac Lab: A GPU-accelerated simulation framework for multi-modal robot learning,” arXiv:2511.04831, 2025.
- [32] C. Tessler *et al.*, “ProtoMotions3: An open-source framework for humanoid simulation and control,” GitHub repository, 2025. [Online]. Available: <https://github.com/NVlabs/ProtoMotions>
- [33] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal policy optimization algorithms,” arXiv:1707.06347, 2017.
- [34] F. Mentzer, D. Minnen, E. Agustsson, and M. Tschannen, “Finite scalar quantization: VQ-VAE made simple,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2024.
- [35] 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,” arXiv:2603.03279, 2026.
- [36] Q. Peng, Y. Lin, Y. Xue, J. Pang, and W. Zhang, “Embodiment-aware generalist specialist distillation for unified humanoid whole-body control,” arXiv:2602.02960, 2026.
- [37] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black, “AMASS: Archive of motion capture as surface shapes,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019.
- [38] G. Pavlakos *et al.*, “Expressive body capture: 3D hands, face, and body from a single image,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019.
- [39] E. Todorov, T. Erez, and Y. Tassa, “MuJoCo: A physics engine for model-based control,” in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2012, pp. 5026–5033.
- [40] L. Pinto, M. Andrychowicz, P. Welinder, W. Zaremba, and P. Abbeel, “Asymmetric actor critic for image-based robot learning,” in *Proc. Robot.: Sci. Syst. (RSS)*, 2018.

APPENDIX A REFERENCE CONSTRUCTION

The teacher imitates, and the student is commanded with, references built from human motion capture and extended with synthesized transitions (Appendix B). This appendix describes how the clips were selected and repaired and how their contact labels were checked; every policy in the paper shares these components.

A. Capture, body, and contact evidence

We start from 60 MOYO clips [8] at 60 Hz, with MoSh++ fits [37], a personal SMPL-X template [38], 73 raw markers, and registered pressure measurements. Two lotus clips exceed the modeled knee range; one firefly, about 25° past the modeled spine stop, could not be re-solved inside the joint limits without jerk or new overlap; and one fit is unreadable. These four are dropped, and the 56 retained clips contain 280 annotated hold windows. The simulated character has 24 rigid bodies, 69 PD-actuated coordinates, one collision primitive per body, hard joint limits, and template-volume-based masses totaling 74 kg.

We distinguish geometric touch, external loading, modeled mechanical utility, and semantic brace role. Human mesh floor touch uses a 2.27 cm threshold and body-body touch a 2 cm gap. Registration to MoSh’s fitted markers has per-clip median residual at most 0.7 mm; this is not independent accuracy against raw marker measurements. A Schmitt-triggered raw-marker estimate agrees on 99.0% of frames with decisive extremity states, over the 59 readable clips. Pressure supplies total load and center of pressure; regional loads are attributed geometrically to nearby projected collider patches within 6 cm of the floor. No sensor directly measures internal body-body force.

At each hold exemplar, a friction-pyramid equilibrium program minimizes peak normalized joint effort:

$$\begin{aligned} s^* &= \min_{f, \tau, s} s \quad \text{subject to} \\ g(q) &= S^\top \tau + J(q)^\top f, \quad f \in \mathcal{K}(q), \\ -s\tau_i^{\max} &\leq \tau_i \leq s\tau_i^{\max}. \end{aligned} \quad (1)$$

Here $g(q)$ is gravity load, S selects actuated coordinates, and $\mathcal{K}(q)$ contains unilateral linearized friction constraints. Forces act only at ground contact points within 2 cm of the floor and at body pairs within 1 cm, and are bounded below only by zero. A hold is feasible when every configured ground zone has a point within 2 cm and $s^* \leq 1$; a configured body pair farther apart carries no force. A contact is individually necessary if removing it destroys feasibility; otherwise it is useful if it lowers minimum summed normalized effort by at least 0.02 under a common peak-utilization bound. This is a static calculation at a fixed configuration: it checks equilibrium under the modeled contact and torque constraints, not dynamic realizability of a transition or the contacts established by the learned policy.

Of the final 63 braces across 45 holds, none is individually necessary, 48 are useful, and 15 are redundant under this

TABLE II
REFERENCE QUALITY UNDER A FIXED CONTACT SPECIFICATION,
EVALUATED ON THE SAME 56 RETAINED CLIPS.

	Initial	Own fit	Re-solved
Floating supports (of 864)	434	48	1
Feasible holds (of 280)	35	235	277
Closed braces (of 63)	39	52	63
Mat load $\geq 90\%$ explained	71.2%	99.3%	99.9%
Touch > 0.5 cm in floor	–	43.0%	0.01%
Frames outside joint box ^a	29.1%	32.5%	0
Support slide (cm/s)	4.5	2.96	2.99

Initial: the AMASS-neutral fit’s angles on the initial body. Own fit: the performer’s MoSh fit on the rebuilt body. Floating supports exceed 2 cm at window medians; statics and braces use hold exemplars. Load uses 81,140 loaded frames, and “explained” means the measured load is attributed to colliders within 6 cm of the floor; the initial value uses the grounded conversion before arm-balance edits that the other initial cells include (2,379 frames in 8 clips). Penetration counts human-touch zone-frames. Slide is the median over clips of the 90th-percentile speed. The 63 braces were selected after the re-solve, partly on whether it could close them, so the last column’s brace and feasibility counts are partly by construction. ^aMore than 1° beyond the widened joint box used throughout; against the original box (more than 2°) the initial references exceed it on 83.9% of frames, illustrating why percentages computed with different joint boxes are not directly comparable.

test. Removing one increases peak utilization by more than 0.05 for 22 braces. Braces can still be jointly necessary: in one Peacock hold, removing all four raises the minimum peak utilization from 0.15 to 1.19, beyond the feasible limit. Internal forces cancel from whole-body force and moment balance at a fixed pose, but can redistribute joint loads. Each brace’s feasible force interval begins at zero; its median upper endpoint is 1.764 body weights. These intervals come from an underdetermined model; they estimate neither the performer’s actual squeeze forces nor the semantic role of a brace.

B. Reference repair

The initial body had legs about 5 cm short and a head collider ending about 10 cm below the crown, and the initial references copied the AMASS-neutral SMPL-X fit’s angles onto it, with the wrist rotation applied to the hands. The own-fit condition in Table II uses the performer’s MoSh fit, with finger-based hand rotation and the root at the fitted pelvis, on a body rebuilt on the performer’s skeleton. Its wrist and hand torque limits are 50% higher, so this condition changes both the motion source and the body. We then re-solve root and joint coordinates with Levenberg–Marquardt, anchoring to the fit with a hard joint box and terms for floor contact, flatness, slip, balance, and separation. Human floor states are read frame by frame. Repair uses 164 brace candidates (138 rule-derived, 26 informed-review candidates) to supply proximity-compatible closure requests: 104 close, while 60 remain incompatible at fitted gaps of 2–9.5 cm. The final 63 braces are selected after this step.

The median over clips of the temporal median mean-body edit is 1.37 cm; the clip-median maximum body/frame edit is 10.6 cm, above the 7.8 cm limit set for this step and the only one of its acceptance checks that the re-solve fails. Comparison with the fit detects no added acceleration spikes or deep overlaps, although it does not certify every frame’s dynamics. Of the 280 annotated windows, eight are transitions;

the remaining 272 holds are reviewed in Pass C below, and the curated corpus retains 270 after excluding two scorpion holds beyond the modeled wrist torque limits: 204 pass, 65 are flagged, and one has a masked contact. Excluded windows keep their frames in the imitation references. Synthesized clips later add 112 holds, for 382 in the training graph.

C. Contact review and calibration

Contact labels are checked by an offline reviewer, Claude Opus 5.5, run in a fresh context with read-only access to one packet and a fixed JSON response schema. A packet holds at most six deterministic CPU renders of one moment of a reference (Fig. 5). In the final render version, used for the released labels, each limb segment has its own hue, darker on the left and lighter on the right; the performer’s motion-capture markers appear as small spheres on the skin; and nothing casts a shadow. A thin line drops to the floor from every hand, foot, or the head between 1 and 60 cm above it and from any other part between 1 and 15 cm above it unless another part lies below, so that a hovering part and a touching part look different from every camera. The views include four eye-level cameras around the body, one from high above, one from 2 cm above the floor in which any gap shows floor beneath a part, close-ups under the parts in question, and a short time strip around the moment. Content hashes, structured responses, and call records support replay. A MuJoCo statue test [39] provides an additional witness using stiffened PD gains and equilibrium feed-forward torques; its disabled joint limits make it a different model from the learned controller.

Figure 6 summarizes the evidence hierarchy. Pass A is blind, with a guard rejecting clip names, labels, frame numbers, and timestamps. Pass B supplies candidate labels and measurements to review brace roles, variants, and windows. Pass C compares fit and repaired reference in randomized order against the human mesh. Instrument-checkable classes require empirical precision at least 0.9 over at least 30 claims and answers on half of at least 30 eligible items. Binary classes use the lower per-value precision. Decisive states use stable human touch over ± 0.5 s, avatar floor gaps at most 1 cm or at least 2 cm, and body-pair gaps at most 1 cm or at least 3 cm. Constructed Pass-C controls include identical copies and known pose, height, or limit perturbations. These thresholds and the reported precision values were evaluated on the development set, rather than an independent annotation test set.

Across three gap-rendering versions evaluated on the same 52 holds, re-rendered for each version, blind float precision rises from 0.72 to 0.81 to 0.99. A fourth version also changes the body, packets, and model effort, preventing a controlled comparison. Pose-name precision remains 0.52–0.58 blind and 0.81 informed, below admission. Informed body-contact claims reach 1.00 precision over 162 claims and 0.84 recall, with the measurements supplied. Pass C admits different-pose and closer-to-human judgments on constructed controls, but not general artifact detection; it calls none of 272 repaired

holds a different pose and prefers the fit on 24, which are flagged.

Two fresh-context calls from the same model review each of 93 selected holds for semantic roles. Required-role claims recur in 126/135 ordered comparisons (0.933), but unique-contact agreement is only 63/72 (0.875). Ordered-comparison agreement passes the admission rule, but neither measure establishes semantic validity or independent errors. Agreement from both calls and human mesh contact select 59 braces across 44 holds; four Peacock braces restored by torque feasibility give the final 63 braces across 45 holds. These selected requests should not be read as individually necessary contacts under the static test. Final floor bits are mesh-derived; markers and mat cross-check them. Required support can follow attributed loading of at least 50 N or static necessity.

APPENDIX B

SKILL GRAPH AND TRANSITION SYNTHESIS

A. Holds, nodes, and edges

A hold is an annotated time window in which the performer keeps a configuration: either the pose a clip is named after or an intermediate stage of its entry or exit, such as the crouch before crow. Each hold records a pose label, the set of ground zones touching the floor, its body-pair contacts, and the bin of the pelvis orientation (upright, inverted, prone, supine, or lying on the left or right side). The 15 ground zones are the feet, shins, thighs, upper arms, forearms, and hands on each side, the pelvis, the trunk, and the head.

A node is a distinct combination of pose label, ground-contact set, and orientation bin. Holds that share all three become one node, whichever clip they come from; standing, which ends every human clip, is a single node that all of them share. Body-pair contacts do not define nodes, so two holds that differ only in a brace share a node while each keeps its own brace requests as goals. A directed edge joins the nodes of two consecutive holds in the same clip: it records that a recording moves from the first configuration to the second, and the motion between the two holds is the transition the edge stands for. A human clip is thus a walk through the graph that leaves standing, passes through its holds, and returns to standing; every human clip ends at standing and all but two begin there. The human crow clip, for example, visits standing, a crouch on both hands and the left foot, crow on the hands alone, and standing again, and so contributes three edges. Pose labels separate configurations with the same supports, such as standing and Warrior II, both upright on two feet.

In a clip’s schedule, each hold with its arrival deadline and duration forms a segment, and the continuation after a hold is the rest of that schedule. Because each human clip performs a single pose family, human clips of different families meet only at standing; holds shared within a family and the ends of the synthesized transitions below provide the other junctions. The 56 human clips give 123 nodes and 196 edges, 15 of them loops where consecutive holds share a node. The five synthesized transitions add edges between existing nodes, for 201.

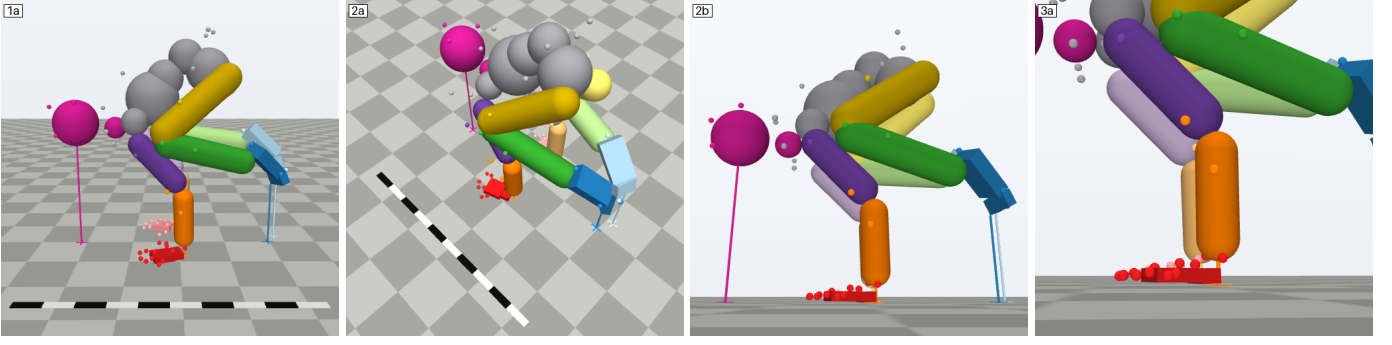
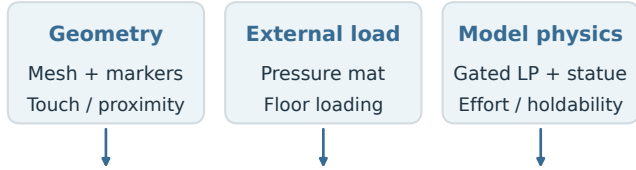


Fig. 5. Four panels of one review packet as the VLM reviewer receives them: a crow hold in the repaired reference on the rebuilt body, with the packet’s own panel tags. 1a: eye level. 2a: from high above. 2b: from a camera 2 cm above the floor, where a gap would show floor beneath a part. 3a: close-up under the left hand. Hues distinguish limb segments and sides (left dark, right light), and small spheres are the performer’s markers. The hands touch the floor and have no drop line; the head and both feet, 26–43 cm up, have one. The bar on the floor is 1 m long in 10 cm segments, matching the floor checks.

1 Measure distinct evidence



2 Review traceable packets

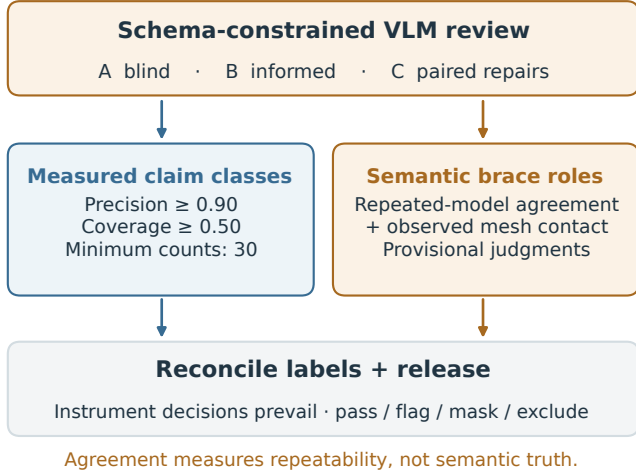


Fig. 6. VLM-assisted contact review. Measured contact states determine instrument-checkable labels. Calibration uses classes with decisive comparison labels; repeated-model agreement measures the repeatability of semantic brace judgments. Advisory variant flags leave contact bits and hold inclusion unchanged.

B. Synthesizing missing transitions

We specified seven missing edges by hand with curated exemplars, phase templates, and intermediate sketches for pressing, lowering, jumping, and landing. Two were dropped after synthesis: a transition from tripod headstand to crow, whose executed head lift was not statically feasible, and one from firefly to crow, which never held crow. The remaining five appear in Table III.

Transition synthesis uses an MPPI-style sampling optimizer [30] in PhysX with the learning body’s PD semantics. It optimizes spline-knot offsets U to nominal targets. For finite rollout costs J_i , the update is

$$\bar{J}_i = \frac{J_i - J_{\min}}{J_{\max} - J_{\min} + 10^{-12}},$$

$$w_i = \frac{\exp(-\bar{J}_i/\lambda_M)}{\sum_j \exp(-\bar{J}_j/\lambda_M)}, \quad U^+ = \sum_i w_i(U + \epsilon_i). \quad (2)$$

Cost-range normalization makes the weights equivalent to raw-cost exponential weights with batch-dependent effective temperature $\lambda_{\text{eff}} = \lambda_M(J_{\max} - J_{\min} + 10^{-12})$. The optimizer therefore adapts its temperature to each sampled batch. Unstable rollouts receive zero weight; the first sample is unperturbed. A short segment is executed before shifting the horizon. Costs combine sketch/destination tracking, velocity, scheduled ground support, planted-hand position and flatness, slip, landing placement and forces, unwanted contact, balance, capture point, torque utilization, speed, smoothness, and limits. Recipes vary these weights. Braces constrain the sketch and destination; there is no separate MPPI body-pair closure cost. Destination brace gaps are checked against 1 cm.

The original endpoint and seam thresholds rejected all candidates. The endpoint test, a final six-body error within 0.05 m, remained mandatory for the press, the lowering, and the jump to plank; for the two transitions into chaturanga it was advisory, and their twelve retained trajectories end 0.078–0.096 m from the exemplar in a lower chaturanga. The retained set uses a relaxed 1 cm planted-hand drift tolerance for the press and the lowering and accepts the exemplars’ own palm offsets of about 12° yaw. Start-seam criteria remain mandatory; landing drift and end-seam tests are advisory: accepted slides were about 7–26 cm to plank, 2–5 cm from crow to chaturanga, and 11–12 cm from handstand to chaturanga. The optimizer’s rollouts run on the training body in PhysX, but two further checks were advisory, and every retained trajectory fails both: some joint’s PD drive demands at least 98% of its torque limit on more than 5% of physics steps (21–100% for the worst joint), and some body’s peak speed or jerk exceeds the

TABLE III
GRAPH ADDITIONS AND TEACHER COMPOSITE EDGE TESTS.

Edge	Trajectories	Passing instances
Crow $\rightarrow$ Handstand	8	24/24
Handstand $\rightarrow$ Crow	3	9/9
Crow $\rightarrow$ Plank	5	15/15
Crow $\rightarrow$ Chaturanga	11	33/33
Handstand $\rightarrow$ Chaturanga	1	3/3

human corpus’s per-body 99th percentile. One of 29 selected candidates was removed for persistent knee contact. Captured lead-ins and lead-outs are spliced with alignment and blending, the settle into the destination and its hold are blended as well, and depenetration moves bodies by up to 2.6 cm. The released references are therefore not certified as dynamically feasible; the teacher’s tracking of them, reported below, shows that they can be followed closely. The synthesized and spliced motions are imitation references, not a separately validated motion-planning output.

Every clip, human or synthesized, appears with 0, 3, and 7 s of extra dwell in one hold, so the 84 base clips (56 human and 28 synthesized) give 252 motion instances on 123 nodes and 201 edges. Because the 84 synthetic instances are dwell variants of 28 trajectories on five edges, they add five edge types rather than 84 independent transitions. The human portion is preserved. The reference package passes 68 identity and consistency checks; motion correctness is assessed separately.

The graph-growth edge test begins at the captured lead-in and requires: 60% of transition frames with every body within 0.5 m of reference; six-body best-yaw destination error below 0.15 m on half the post-deadline hold; and every commanded ground zone within 2 cm of the floor on 90% of its annotated destination window. The per-zone minimum is used rather than simultaneous all-zone satisfaction. This geometric test differs from the loaded-support command rule in Section IV. It excludes neither all unwanted supports nor missing braces. At epoch 1 no instance passes; at epoch 3,420 all 84 pass, every transition frame meets the tracking threshold, and median destination errors are approximately 0.021–0.051 m. Under a 3%-body-weight load threshold, minimum per-zone loaded-frame shares are 1.00 except for the lowering from handstand to chaturanga, at 0.94–0.96.

APPENDIX C TEACHER ARCHITECTURE AND OBJECTIVES

A. Observations and actuation

The asymmetrically trained PPO actor [33], [40] receives simulator kinematics, contact forces/states, proximity, the raw action from two control steps earlier (the history buffer lags it by one step), and pose history at 1, 8, and 15 control steps. A hold goal is

$$g_k = (\mathbf{x}_k, \mathbf{R}_k, \mathbf{c}_k, \mathbf{o}_k, T_k, H_k), \quad (3)$$

with 24 body poses, 104 contact bits (15 ground zones, 89 body pairs), six-bin pelvis orientation, exemplar deadline T_k , and hold end H_k . The actor reads two slots: current hold and successor inside a hold, or upcoming holds during a transition. Each slot includes validity/visibility masks, deadline $\max(T_k - t, 0)$, duration $H_k - T_k$, and remaining time $\max(H_k - t, 0)$; the latter two are clipped to 10 s and divided by 10. Training reveals full poses and contacts. The task critic also reads reference poses 1, 5, 10, and 15 steps ahead; the style critic additionally reads style history. Dense references supervise rewards, resets, and termination, even though the actor does not receive them.

The actor and task critic have six and four hidden ReLU layers of width 2,048. The actor samples 69 pre-transform coordinates with fixed standard deviation $e^{-2.9}$:

$$a_t \sim \mathcal{N}(\mu_\theta(o_t, G_t), e^{-5.8} I), \quad q_t^{\text{des}} = b + s \odot \tanh(a_t). \quad (4)$$

Here $b = 0$, and joint j shares scale $s_j = \min(\pi, 2 \max_i \{|q_{ji}^{\min}|, |q_{ji}^{\max}|\})$ across its three coordinates. The targets are absolute joint positions; the policy does not output residuals on the reference. Torque-limited fixed-gain PD drives execute each action across four physics steps. Evaluation uses the actor mean before the same transform.

B. Task and contact reward

For 24 bodies, E_p is mean squared position error averaged over bodies and three coordinates; E_v, E_ω use the same averaging for linear/angular velocity. E_R is mean squared global-orientation geodesic angle in radians, and E_h is squared root-height error. The implemented task reward is

$$\begin{aligned} r_t^{\text{task}} = & 0.5e^{-25E_p} + 0.3e^{-5E_R} + 0.1e^{-0.5E_v} \\ & + 0.2e^{-0.1E_\omega} + 0.2e^{-100E_h} \\ & - 0.02\|q_t^{\text{des}} - q_{t-1}^{\text{des}}\|_2 \\ & - \min(10^{-5} \sum_j |\tau_{tj} \dot{q}_{tj}|, 0.5) \\ & - 0.3(p_t^{\text{free}} + p_t^{\text{swing}} + p_t^{\text{lean}}). \end{aligned} \quad (5)$$

Let $F_{z,t}$ be positive terrain-filtered vertical load on zone z ; $m_{z,t}$ marks a noncommanded zone annotated as free in the corpus whose reference joint centers exceed 0.25 m. With hold gate I_t ,

$$\begin{aligned} L_t &= I_t \sum_z m_{z,t} F_{z,t}, \\ \bar{L}_t &= (1 - \alpha)\bar{L}_{t-1} + \alpha L_t, \quad \alpha = \frac{\Delta t}{0.25 + \Delta t}, \\ p_t^{\text{free}} &= I_t \text{clip}_{[0,1]}(\bar{L}_t / (0.1mg)). \end{aligned} \quad (6)$$

Averaging load before saturation discourages brief high-force taps: in one arm-balance clip of an earlier run on uncured references, a per-frame clamp let the policy tap a foot at 3–4 Hz, keeping 68–82% of its resting load while paying about a quarter of the penalty. The swing term applies the same construction outside holds with offline swing labels and a 0.1 s time constant. For eligible arm-supported holds, lean penalty is $I_t^{\text{lean}} \text{clip}_{[0,1]}((0.03 - d_t)/0.10)$, where d_t is the signed

distance of the COM inside the commanded support polygon in meters, negative outside it. All three penalties are disabled during reset grace. Required-support/brace attainment, pair load, and slip have zero reward weight. Brace closure is consequently supervised through pose imitation and dynamics rather than an explicit attainment term.

C. Adversarial prior and PPO update

AMP [10], [29] uses eight 163-dimensional frames at offsets $\{1, 2, 3, 4, 6, 9, 14, 20\}$. Features are root height, gravity and root velocities in the root frame, parent-relative rotations, and root-relative head/hand/toe positions, avoiding a projected-heading singularity in inverted poses. The unconditioned discriminator has hidden widths 1,024 and 512 and logit D_ψ . With demonstrations $\mathcal{D}$ and equal current-policy/replay negative mixtures,

$$\begin{aligned} \mathcal{L}_D = & \frac{1}{2} \mathbb{E}_{\mathcal{D}} \text{softplus}(-D_\psi) + \frac{1}{4} \mathbb{E}_{\pi} \text{softplus}(D_\psi) \\ & + \frac{1}{4} \mathbb{E}_{\text{replay}} \text{softplus}(D_\psi) + 10 \mathbb{E}_{\mathcal{D}} \|\nabla_{\phi} D_\psi\|_2^2 \\ & + 10^{-4} \|W_D\|_F^2 + 10^{-2} \|w_{\text{out}}\|_2^2, \\ r_t^{\text{style}} = & -\rho(m_t) \log \max\{1 - \sigma(D_\psi(\phi_t)), 10^{-4}\}. \end{aligned} \quad (7)$$

The gradient penalty uses normalized demonstration features. The positive pool has 55 repaired human base clips, excluding one hands-only Scorpion recording, all dwell extensions, and all synthetic clips. Sampling is uniform over usable reference times after the first 20 control steps. During graph fine-tuning, $\rho = 0.5$ for synthetic lineage and 1 for human clips. The score serves as a training signal and should not be read as a calibrated measure of naturalness.

Task and style critics estimate separate returns. GAE uses original reward/value units with $\gamma = 0.99$, $\lambda = 0.95$, true-termination bootstrap masking, and episode-boundary recursion masking. Before PPO, advantages are combined and normalized:

$$\tilde{A}_t = \text{clip}_{[-4,4]} \frac{A_t^{\text{task}} + w_A A_t^{\text{style}} - \mu_A}{\max(\sigma_A, 0.02)}. \quad (8)$$

Running statistics update at factor 0.05. PPO clips the probability ratio to $[0.8, 1.2]$; critics separately regress normalized GAE returns with clipped value losses. The style critic has three hidden layers of width 1,024. The discriminator batch is 4,096 and replay capacity 200,000 histories. Fixed action variance makes the configured entropy bonus inactive.

APPENDIX D

STUDENT ARCHITECTURE AND DISTILLATION

The custom student combines masked conditioning inspired by MaskedMimic [13] with finite scalar quantization (FSQ) [34] (Fig. 7). The following equations define its discrete encoder–prior interface and shared action decoder. Each of its five goal slots carries a six-body pose (pelvis, head, hands, ankles), the full 104-bit contact request, pelvis orientation, and timing. The pose and contact halves are revealed with probabilities 0.75 and 0.85; a revealed pose shows all six bodies with probability 0.75 and otherwise a random subset

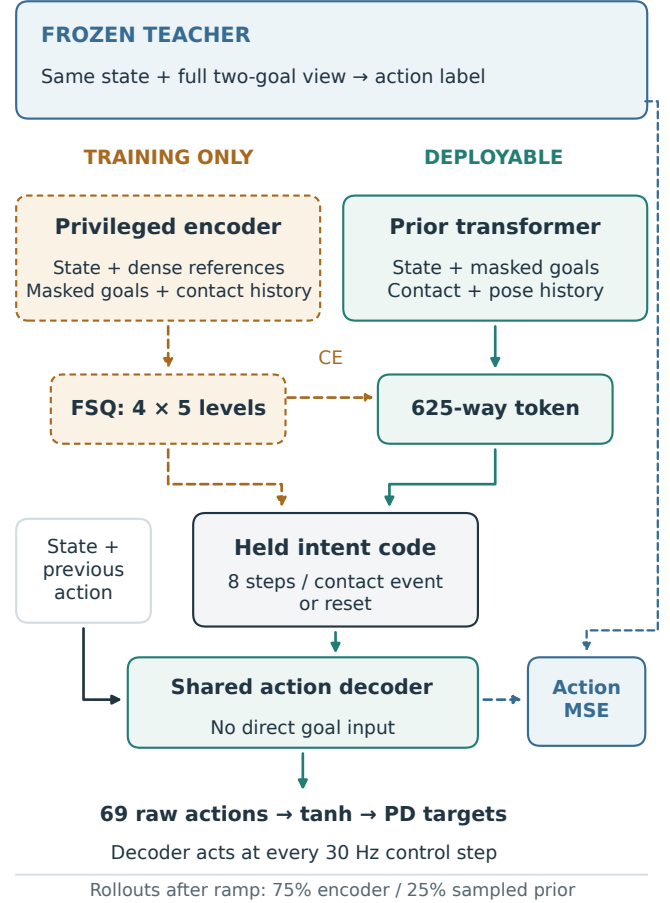


Fig. 7. Teacher/student interfaces. Dense references inform the teacher’s critics, reward and resets, and the student encoder. The deployable prior sees state, masked goals, and history, then supplies a discrete code to the decoder. The frozen teacher labels every visited state. The decoder’s action input is the raw action from two control steps earlier.

of bodies, each revealing position, rotation, or both; the first slot never hides both halves, so its contact half is revealed whenever its pose is hidden; and masks are drawn when a goal enters the window and travel with it. Target root XY is aligned with the current root. The frozen teacher retains its fully visible two-slot, 24-body goal view at the same physical state. We verified all twelve teacher actor inputs against the original environment in 1,497 batched comparisons across starts, training resets, and manual goals. Teacher labels are mean pre-tanh actions from (4).

Let x_t contain proprioception, measured contacts, and proximity; $\tilde{g}_t$ denote masked goals; and h_t denote history: poses at 1, 8, 15, 30, and 60 control steps (2 s) and four measured contact-event tokens. The privileged encoder also reads dense reference poses r_t at offsets $[1, 5, 10, 15]$ and contact-event history h_t^c . At refresh steps,

$$z_t^E = Q_{\text{FSQ}}(E_\eta(x_t, r_t, \tilde{g}_t, h_t^c)) \in \{-2, -1, 0, 1, 2\}^4. \quad (9)$$

The bounded-and-rounded operator uses straight-through gradients. The four scalar indices are packed into a single 625-way token rather than emitted as four sequential tokens. The

prior’s context transformer reads $(x_t, \tilde{g}_t, h_t)$ and its categorical head predicts that token. Both streams use decoder $D_\theta(x_t, a_{t-2}, z_t)$, with no direct goal input; a_{t-2} is the raw action from two control steps earlier, the same buffer lag as the teacher’s. Codes refresh every eight control steps, at a committed contact event, or after reset; the decoder acts each step. During command evaluation, goal advances also flush held intent, whereas periodic command reveals can precede the next normal refresh (Appendix I). Goal information has no direct path into the decoder apart from the held code; state feedback remains available at every action step. Between refreshes, stored encoder codes are detached during replay. Deployment samples at temperature 1 and top- p 0.9; greedy evaluation takes argmax.

The encoder has five hidden ReLU layers of width 1,024, a 512-dimensional output, and a 256–128–4 projection. The prior context transformer has four layers, four heads, width 512, and feedforward width 1,024; its categorical head is a two-layer transformer of the same width followed by a linear 625-way output. Context includes one state token, five goal tokens, five pose-history samples at [1, 8, 15, 30, 60] steps, and four contact-event tokens. The decoder has three hidden layers of width 1,024 and 69 outputs. Quantized scalar codes enter it without division by two.

A. Online collection and supervised loss

Both collection streams execute student actions and obtain frozen-teacher labels. Initially every environment uses the privileged encoder stream. A fixed block switches to sampled prior actions from epoch 1,000, ramping to 25% by epoch 1,500. This is online teacher labeling in the spirit of DAgger [3]. During initial distillation, goals follow the playing clip’s schedule: no continuation changes within a segment or is replaced at a held hub. Revised-goal collection (Appendix E) changes that command distribution while retaining the following losses.

Let $\mathcal{B}_P \subseteq \mathcal{B}$ contain prior-action rows, $\ell(a, b) = \|a - b\|_2^2/69$, and sg denote stop-gradient. Stored prior codes are re-decoded as $\bar{a}_t^P = D_\theta(x_t, a_{t-2}, \text{sg}(z_t^P))$. With encoder-stream action a_t^E and teacher label a_t^T ,

$$\begin{aligned}\mathcal{L}_E &= \frac{1}{|\mathcal{B}|} \sum_{t \in \mathcal{B}} \ell(a_t^E, \text{sg}(a_t^T)), \\ \mathcal{L}_P &= \frac{1}{|\mathcal{B}_P|} \sum_{t \in \mathcal{B}_P} \ell(\bar{a}_t^P, \text{sg}(a_t^T)), \\ \mathcal{L}_{\text{CE}} &= \frac{1}{|\mathcal{B}|} \sum_{t \in \mathcal{B}} \text{CE}_{0.01}(p_\phi(\cdot | c_t), \text{tok}(\text{sg}(z_t^E))), \\ \mathcal{L} &= \mathcal{L}_E + 0.1\mathcal{L}_P + \lambda(e)\mathcal{L}_{\text{CE}}, \\ \lambda(e) &= \text{clip}_{[0,1]} \frac{e - 100}{500}.\end{aligned}\tag{10}$$

Here $c_t = (x_t, \tilde{g}_t, h_t)$; $\mathcal{L}_P = 0$ for an empty subset. Cross-entropy has label smoothing 0.01 and includes refresh and held-code rows; on a held-code row its target is the code stored at the last refresh. Encoder imitation gradients pass through the quantizer only on refresh rows. Because sampled tokens are stop-gradient, the prior-action loss updates only the decoder;

the cross-entropy term is what trains the prior. Student training includes neither PPO nor a reinforcement objective for goal attainment. The four fine-tuning runs restore the weights and Adam state of the selected checkpoint but number their epochs from zero, so both schedules are overridden: $\lambda = 1$ and the 25% prior-action block apply from their first epoch.

APPENDIX E REVISED-GOAL COLLECTION

Revised-goal collection changes only the command distribution under which the student is distilled. Without it, goals always come from the playing clip’s own schedule, so an arrival at a given hold is always followed by the same continuation. With it, a random switch replaces the rest of that schedule with another clip’s continuation from a compatible hold while the simulation carries on.

What a switch changes. The playing reference moves to the new clip at the matching time. The goal window, the teacher’s inputs and hence its action labels, the privileged encoder’s dense reference, and the reference used by the tracking termination all follow the new clip. Physical state, pose history, and contact-event history persist, and the held intent code is kept until its next refresh, so for up to seven steps after a switch the cross-entropy target is still the code chosen before it.

Where switches can occur. A precomputed table lists, for every motion instance, the windows in which a switch is allowed and the continuations it may lead to. Standing rows connect one clip’s final standing to another clip’s opening standing. Crow and handstand rows connect clips that share a span of reference frames exactly, with the clip clock mapped across. Chain rows connect a clip that has arrived at crow or handstand, usually through a synthesized transition, to a synthesized transition leaving from that hold. Within-family rows connect two recordings of the same pose. Standing, chain, and within-family rows are admitted by geometric and heading compatibility, and their continuations are translated horizontally to the current pelvis position, without rotation. A quarter of each crow and handstand row’s weight goes to the same clip at another dwell, so a switch need not introduce a different clip.

When. Inside an eligible window, a per-step hazard gives a nominal total switch probability of 0.5. An episode allows at most three switches, at least 2 s apart, so one episode can follow a short walk through the graph.

What is held out. The 11 directed standing pairings used by the five evaluation flows, including their dwell variants, are removed from the table. Each clip in a held-out pairing still appears in the same role in 6 to 64 other standing pairings, so the flows test new pairings at a familiar junction rather than a new kind of junction.

Resulting exposure. About a third of switch-trained episodes (32%) contain a switch, with 1.24 switches on average among them, and 5.2% of environment steps fall within 2 s after a switch. Two-thirds of switches occur at standing (67–68%),

TABLE IV
TRAINING SEQUENCE. TRANSITIONS ARE NOMINAL TRAINING
COLLECTIONS, EXCLUDING EVALUATION. EPOCHS ARE LOCAL TO EACH
STAGE.

Stage	Envs.	Epochs	Steps (10^6)	Hours
Initial teacher	4,096	15,500	2,031.6	41.30
AMP fine-tuning	4,096	5,000	655.4	15.42
Graph fine-tuning	4,096	3,420	448.3	10.35
Selected student	1,024	5,000	163.8	30.81
Stopped student	1,024	7,210	236.3	44.33

The last two rows are checkpoints of the same run and are not additive. Hours include scheduled evaluation and visualization but exclude curation, planning, standalone analyses, exploratory runs, and the failed launch.

followed by crow (10%), chain (10%), within-family (7%), and handstand (6%) switches.

Runs. All four fine-tuning runs start from the same checkpoint of the original distillation run, selected with command probes and familiar-route retention (Appendix K), and restore its weights and Adam state; the teacher stays frozen. Each run trains for 1,000 epochs, or 32.8 million nominal environment steps (about 5.7 h on one GPU for each timed run), and saves checkpoints after 250, 500, 750, and 1,000 epochs. Switch and no-switch runs differ only in the switch settings and the seed. The seed affects resets, motion sampling, prior tokens, minibatch order, and switch draws, so equal seed numbers do not pair a switch run with a no-switch run.

APPENDIX F IMPLEMENTATION, SCHEDULES, AND COMPUTE

All stages use GPU PhysX in IsaacSim 5.1.0 via IsaacLab [31] and ProtoMotions [32]. The software versions are ProtoMotions 3.1, IsaacLab Python package 0.50.5, PyTorch 2.7.0, and Python 3.11.14. Each run uses one 24 GB NVIDIA RTX A5000, an Intel Core i9-12900KS, and 64 GB host memory. Temporal Gauss–Seidel uses four position and four velocity iterations. Physics/control rates are 120/30 Hz. Self-collision is enabled; external pushes and domain randomization are absent. Training episodes of both policies terminate when any body is more than 0.5 m from its reference position.

All optimizers are Adam, with $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$, zero optimizer weight decay, and norm clipping at 50. Teacher and student learning rates are 2×10^{-5} ; both teacher critics and the discriminator use 10^{-4} . One epoch collects 32 steps per environment (Table IV). Teachers use minibatches of 16,384 and one pass; students use 8,192 and six passes. Two 4,096-row microbatches with size-weighted accumulation and pooled normalization moments reduce peak student GPU memory from 24,027 to 17,547 MiB. Reported student results use this microbatch configuration; a full-batch launch that failed with a PhysX allocation error is excluded. These pretraining stages use seed 0. The four additional student fine-tuning runs use two seeds per condition, as specified in Appendix E.

The initial curated-reference teacher trains for 15,500 epochs. Its AMP fine-tuning runs 5,000: $w_A = 0$ through epoch 200, is calibrated there to target 0.5, and reaches

that weight by epoch 500. Graph fine-tuning starts from this checkpoint, keeps $w_A = 0.5$, freezes actor/task-critic observation statistics, and stops at epoch 3,420. Teacher fine-tuning preserves optimizer state. Initial teacher, AMP fine-tuning, and student initialize near arrivals with probability 0.6 and up to 0.5 s pre-roll. Graph fine-tuning allocates 0.4 to arrival, 0.2 to departure, 0.08 to beginnings, and 0.32 uniformly in time. Teacher stages draw 80% of motions from their base sampling weights and 20% in proportion to one minus each clip’s smoothed hold-and-tracking score; in graph fine-tuning the base weights are the package’s, where each new edge weighs as much as three human clips, independently of its variant count. Student normalization is separate and its motion-sampling distribution is fixed.

Teacher training evaluation spans 2,250 control steps; student evaluation spans 1,890, covering the longest 63 s reference motion. Both policies terminate training episodes at the same 0.5 m body deviation, but their scheduled evaluators use different tracking criteria, so those success values are not directly comparable. The teacher stages total 67.1 GPU-hours of training and 3.14 billion nominal transitions. Summed training time through the selected student is 97.9 hours, or 111.4 including the rest of its stopped run.

APPENDIX G EVALUATION PROTOCOL

Plans, targets, and replicas. Every command evaluation runs a plan: a timed sequence of hold goals, each with a pose, requested ground zones, an arrival deadline, and a hold duration, issued to the controller without resetting the simulation. Designated holds, or targets, are scored; the other goals guide the motion and are not. A replica is one rollout of a plan; all replicas of a plan start from the same state and receive the same commands. Student replicas differ through sampled intent codes and simulator nondeterminism, whereas the teacher acts with its mean action, so its replicas differ only through the simulator. Replicas therefore measure variation between rollouts, not between training runs.

Plan groups. Familiar routes (41 plans) replay a clip’s own schedule, as in training, and test whether a controller still performs the corpus; each has one target. The cross-clip suite holds 24 further plans. Revised destinations (five plans) follow a human clip to a shared hold and then command another clip’s continuation, such as the jump from crow to plank in place of the crow clip’s return to standing; two own-route controls follow the crow and handstand clips through the same holds without a revision. Multi-hop chains (five) revise several times in a row through the synthesized transitions, for example crow, press to handstand, and lowering back to crow. Standing flows (five) join clips of different pose families through returns to standing, as in Fig. 1; their standing pairings are held out of switch collection. Within-family alternates (seven) switch between two recordings of the same pose at a shared hold. The 19 cross-clip plans other than the flows are the command plans averaged in Fig. 2b.

Revealing a revision. Until a revision is revealed, the controller sees only the goals of the clip it is following. Under the switch protocol, used unless stated otherwise, the new continuation appears once the hold before it begins, as after a switch in training. Under the approach protocol it appears earlier, when that hold becomes the active goal. Each continuation is placed with a rigid yaw and horizontal offset, fitted between the two clips’ reference holds at the junction, so that it continues from the previous clip’s reference; the residual mismatch is zero at synthesized junctions, 1–3 cm at standing, and up to 0.24 m within families. Goals are reissued every 0.25 s, which rounds to eight control steps, except in the checkpoint sweep of Appendix K.

Scores. A target succeeds in a replica when it meets the pose-and-load rule of Section IV over its hold window. Teacher and student are scored on the same six bodies (pelvis, head, hands, and ankles), recomputed from saved errors and load flags. The pose error subtracts each side’s pelvis position, rotates the result about the vertical axis by the least-squares yaw of each frame, and averages the six body distances over the hold window; the pelvis term is therefore zero, and the 0.15 m threshold corresponds to 0.18 m averaged over the head, hands, and ankles. Hold windows are half-open on the 30 Hz control grid, so the 0.30 s plank window holds nine frames, and the 90% loading requirement means all nine. Target success is the fraction of replicas in which a target succeeds. Group means first average target success within each plan, including arrival targets such as the crow before a revision, and then weight plans equally. All-target completion requires every target of a plan in one replica; for a flow, second-clip success scores the first target in the flow’s second clip. A plan passes when at least half its replicas succeed. Three designated targets have zero hold duration and are not scored: the continuation of one Warrior III alternate, the standing exit of the crow control, and the only target of a Peacock route, which is therefore not among the 41. The first two plans contribute only their arrival targets, so that Warrior III alternate tests no revision.

Replica counts. At the fixed fine-tuning endpoint, the students run 85 replicas of each cross-clip plan, 409 of each standing flow, and 48 of each familiar route, sampling the deployable prior; the teacher runs 31 replicas of each plan. The initial student’s evaluations before fine-tuning come from earlier batches (Appendix K).

A. Scope of the pose-and-load rule

The 5 N threshold is about 0.7% of the avatar’s body weight, so passing it shows persistent contact loading, not the force needed to release another limb or to transmit an external wrench. Each zone is checked on its own: with four requested zones, 90% occupancy per zone allows, in the worst case, only 60% of frames on which all four are loaded together. A measure of simultaneous loading would complement the rule but was not computed for these results. Pose is compared after pelvis-relative, best-yaw alignment, so the rule checks where the head, hands, and ankles lie relative to the pelvis, not world-

frame placement, pelvis height, heading, or the full articulated pose. It does not score the 89 body-pair requests, the quality of transitions, or contacts between designated holds, and the 0.30 s plank window after the jump from crow tests attainment rather than a sustained plank.

APPENDIX H FINE-TUNING RESULTS

This appendix follows the four fine-tuning runs of Table I and Fig. 2 across their saved checkpoints and lists the criteria fixed during development.

A. Trajectories

Table V gives each measure at the four checkpoints. Both switch runs exceed both no-switch runs on revised destinations with their own-route controls, chains, standing flows, and the command-plan mean at every checkpoint; within-family alternates are mixed. These measures overlap, since the command-plan mean includes the first two, and they are repeated measurements of the same four runs rather than independent tests. Neither no-switch run completes a flow at any checkpoint, although clip-local training alone recovers some individual revisions.

The two no-switch runs differ markedly. Seed 1 keeps far more command performance than seed 0: its endpoint command-plan mean is .552 against .318, and the press from crow to handstand succeeds on 81 of 85 replicas against 4. Their final imitation losses (7.30×10^{-5} and 7.40×10^{-5}) and code cross-entropies (1.454 and 1.453) are nearly equal, so training losses do not reveal the difference.

Individual checkpoints trade composition against familiar routes. At 750 epochs, switch seed 1 reaches 978 of 2,045 second-clip targets but completes only 41 of 2,045 flows, while no-switch seed 0 has its best familiar-route mean (.722, 31 passing routes) with a command-plan mean of .453. Aggregate route success also hides gains and losses of individual skills: the switched endpoints keep 26 and 24 of the 27 routes the initial student passes and gain three and four others. Over pose thresholds from 0.10 to 0.20 m, the composition rates in Table I change by at most .05. The four runs share pretraining, teacher, corpus, and development plans, so their checkpoint ranges describe variation within this setup, and replica counts do not estimate uncertainty across independent training runs.

B. Criteria fixed during development

A criterion set before training asked for at least .8 success at the final checkpoint on each of three late crow-to-plank tests: the standard revised plan and two diagnostics with the base and the extended crow dwell. Neither switch run meets it at its final checkpoint, although each meets it at two of its four checkpoints (Table VI).

A broader criterion, fixed after switch seed 0 and before switch seed 1, asked that revised-destination, chain, flow, and command-plan means exceed those of the initial student and no-switch seed 0 at matched checkpoints, that second-clip flow success exceed .1, and that at least 25 of 41 familiar

TABLE V

FINE-TUNING TRAJECTORIES. EACH CELL LISTS THE CHECKPOINTS AFTER 250 / 500 / 750 / 1,000 ADDITIONAL EPOCHS. THE FIRST FIVE ROWS ARE MEAN TARGET SUCCESS IN THE 85-REPLICA CROSS-CLIP SUITE: PER PLAN GROUP, AND OVER THE 19 COMMAND PLANS OF THE FIRST, SECOND, AND FOURTH GROUPS. THE FLOW ROWS BELOW USE THE 409-REPLICA FLOW EVALUATIONS, EXCEPT AT 250 AND 500 EPOCHS FOR BOTH SWITCH RUNS AND NO-SWITCH SEED 1, WHICH COME FROM THE 85-REPLICA SUITE.

Measure	Initial	No switch 0	No switch 1	Switch 0	Switch 1
Revised destinations and own-route controls	.780	.800/.672/.718/.530	.516/.796/.753/.824	.971/.943/.974/.912	.953/.912/.871/.971
Multi-hop chains	.586	.656/.188/.297/.241	.306/.675/.539/.646	.852/.845/.892/.868	.880/.917/.767/.909
Standing flows	.302	.271/.276/.279/.221	.257/.300/.203/.295	.541/.637/.589/.662	.535/.532/.504/.577
Within-family alternates	.304	.241/.287/.300/.161	.286/.200/.296/.214	.285/.317/.315/.350	.301/.253/.294/.282
Command-plan mean (19 plans)	.554	.556/.403/.453/.318	.376/.545/.528/.552	.687/.686/.710/.693	.694/.670/.631/.701
Flows: second-clip target	.011	.000/.021/.002/.000	.016/.002/.003/.001	.478/.539/.558/.558	.332/.391/.478/.409
Flows: all targets	.000	.000/.000/.000/.000	.000/.000/.000/.000	.139/.292/.185/.320	.082/.132/.020/.213
Familiar routes passing (/41)	27	28/30/31/25	31/30/22/28	26/29/30/29	26/27/28/28
Familiar-route target success	.663	.665/.688/.722/.567	.704/.703/.556/.690	.617/.664/.676/.666	.634/.619/.635/.665

TABLE VI

LATE CROW-TO-PLANK SUCCESS BY ADDITIONAL EPOCH. CELLS GIVE STANDARD-PLAN / BASE-DWELL / EXTENDED-DWELL COUNTS OUT OF 85 / 170 / 170. A STAR MARKS AT LEAST .8 SUCCESS IN ALL THREE.

Epoch	Switch seed 0	Switch seed 1
250	83/136/167*	83/100/162
500	72/124/147	85/164/170*
750	75/157/152*	81/166/166*
1,000	35/39/94	79/127/155

TABLE VII

ALL-TARGET COMPLETION OF EACH STANDING FLOW AT THE FIXED ENDPOINTS, OUT OF 409 REPLICAS. BOTH NO-SWITCH RUNS COMPLETE NONE.

Flow	Switch seed 0	Switch seed 1
Arm balances	92	117
Floor sequence	174	263
Inversions	1	0
Mixed routine	258	23
Standing balances	129	32

routes pass. Both switch runs meet it at all four checkpoints. The .1 threshold marks early progress rather than reliable success: switch seed 1's second-clip success stays below .5 at every checkpoint. No-switch seed 1, trained last, kept more command performance than predicted before its run, but its second-clip flow success stayed at most .016, below the predicted ceiling of .05.

C. Completion by flow

Table VII splits all-target completion by flow. The seeds are strong on different flows: seed 0 completes the routine of Fig. 1 on 258 of 409 replicas and the standing balances on 129, against 23 and 32 for seed 1, while seed 1 is stronger on the floor sequence and the arm balances. Both almost never complete the inversion flow, whose Feathered Peacock target fails requested hand loading even for the teacher (Appendix J). The teacher is not a per-flow ceiling either: it completes the routine on only 3 of 31 replicas and the inversion flow on none.

APPENDIX I REVISION TIMING

A. Protocol

The timing experiment holds one plan fixed and varies only when its revision is revealed. The plan follows the human crow clip, whose own continuation returns to standing, and revises it to the synthesized jump to plank. The crow hold is scored over [10.85, 14.98] s and the plank over [18.6, 18.9] s; goals, dwell, and deadline are identical in every condition. The reveal offset is the time at which the plank continuation replaces the standing goal, measured from crow-hold onset: negative offsets reveal it while the controller still approaches crow, positive ones during the hold. Joint success requires the crow and the plank target in the same replica. The extended sweep covers nine offsets from -1.78 to $+3.02$ s with 227 replicas each (Fig. 8, Table VIII). An earlier, finer sweep covers offsets from -3.75 to -1.78 s with 204 or 408 replicas per point; it was run for the initial student, at every checkpoint of the switch runs, and at the endpoints of the no-switch runs. A reveal takes effect at a goal reissue, every 0.267 s, or at a goal advance, which sets the exact offsets. A reveal does not force an immediate intent refresh: the held code changes at its next scheduled refresh, whereas a goal advance flushes it.

B. Two kinds of failure

Both switched endpoints keep the crow: seed 0 preserves it on 224–227 of 227 replicas per offset and seed 1 on all of them. Neither shows an early departure, that is, nine consecutive frames after crow-hold onset with crow pose error above 0.30 m, beginning more than one reissue period (0.267 s) before the crow's commanded end, and their destination and joint success counts are equal at every offset, so their failures are failures at the destination. To tell these apart, attainment is detected as the first nine consecutive frames with plank pose error below 0.15 m and all four requested zones carrying at least 5 N, regardless of the scheduled window and of unwanted loads, and is timed from the first of those frames.

The two seeds fail differently. At the latest offset, $+3.02$ s, seed 0 has 201 destination failures; none attains plank during the rollout, and 184 stand through the plank window on

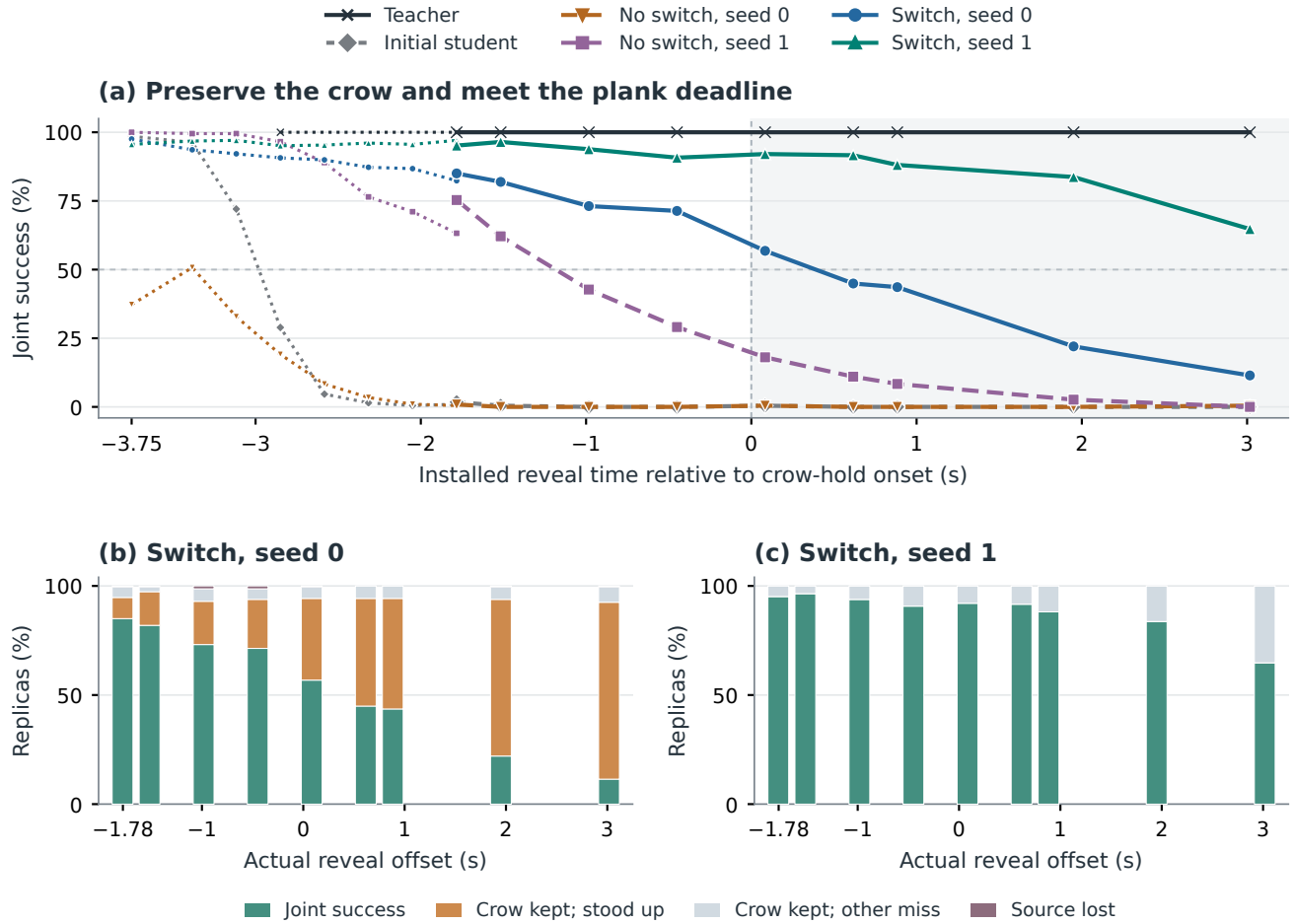


Fig. 8. Revision timing at the fixed fine-tuning endpoints. (a) Joint success preserves the source hold and meets the destination’s scheduled pose-and-load window. The extended sweep uses 227 replicas per offset; dotted curves pool earlier batches with identical installation times (204 or 408 student replicas and 58 teacher replicas per point). Shading marks reveals after source-hold onset; the hold ends at +4.133 s. (b–c) Outcomes of each switch run: joint success, crow kept and then standing, other destination failures with crow kept, and source failures.

their feet with an upright pelvis, a median 0.83 m from the plank pose. These replicas return to the crow clip’s original continuation. Seed 1’s 80 failures are mostly late arrivals: 79 eventually attain plank, at a median of 0.27 s after the window closes, 77 of them for the first time after closure; the other two attain it inside the window but load their feet on only four and five of its nine frames. During the window they are already near plank, at a median six-body error of 0.18 m, with all four requested zones loaded in 73. At +0.88 and +1.95 s, every one of seed 1’s failures eventually attains plank. Late attainment still counts as failure: the plank window closes at 18.9 s, when the plan advances to standing and flushes the held intent, so a longer deadline would change everything that follows. The teacher attains plank at about 15.6 s from its own arrivals and succeeds at every offset.

Later reveals leave less notice: the time between the reveal and the required plank onset falls from 9.5 to 4.7 s across the sweep, while the controller spends longer under the old goal. Preparation time, the quality of the arrival state, and the

controller’s response are therefore coupled in this experiment.

C. Sensitivity and repeatability

Raising the pose threshold from 0.15 to 0.20 m lifts switch seed 1 to at least 223 of 227 joint successes at every earlier offset and 217 at the latest, leaves seed 0 unchanged, and changes no-switch seed 1 by at most one replica. Linear interpolation places the 50% crossing at about -1.18 s for no-switch seed 1 and $+0.39$ s for switch seed 0; switch seed 1 stays above 50% at every tested offset, and the initial student crosses at about -2.98 s in the earlier sweep. Repeated batches of one condition differ by up to .16: at -1.78 s, no-switch seed 1 moves from 129/204 to 171/227, and at the standard $+0.08$ s reveal, switch seed 0 moves from 35/85 to 129/227.

D. Clearing history

Clearing the student’s memory does not rescue a late revision. At the initial student’s late reveal, resetting pose history, contact-event history, and the held intent, in all eight combinations, leaves the failing plank plans with at most

TABLE VIII

JOINT CROW-AND-PLANK SUCCESSES AT THE FIXED ENDPOINTS, OUT OF 227 REPLICAS PER REVEAL OFFSET FROM CROW-HOLD ONSET. GOALS, DWELL, AND DEADLINE ARE FIXED; EACH CONTROLLER, INCLUDING THE TEACHER, REACHES CROW ALONG ITS OWN TRAJECTORY.

Offset (s)	-1.783	-1.517	-0.983	-0.450	+0.083	+0.617	+0.883	+1.950	+3.017
Teacher	227	227	227	227	227	227	227	227	227
Initial student	4	1	0	0	1	0	0	0	0
No switch, seed 0	2	0	0	0	1	0	0	0	1
No switch, seed 1	171	141	97	66	41	25	19	6	0
Switch, seed 0	193	186	166	162	129	102	99	50	26
Switch, seed 1	216	219	213	206	209	208	200	190	147

one success among 126–129 replicas per condition, while accepted press revisions and own-route controls still succeed. At standing, no such reset meaningfully recovers the first target of the new clip.

APPENDIX J

REQUESTED-SUPPORT FAILURES

A replica can pass the pose test and still fail the pose-and-load rule. This appendix describes what such replicas have in common.

Teacher. On the 41 familiar routes, the teacher’s 1,271 target replicas include 1,236 that pass pose and 988 that pass the full rule. Every one of the 248 replicas in that gap misses a requested load: the forearm in Bridge, the thighs and feet in Cobra, the forearms in Dolphin Plank, the hands in the forearm stand, the head, trunk, and forearm in Legs Up the Wall, the trunk in Plow and Supported Shoulderstand, and the head in Koundinya.

The unloaded supports sit close to the floor. The evaluation records full body states for four replicas per plan, 32 of them in this gap. In all 32, every missing support stays within 2 cm of the floor on every hold frame, at most 1.89 cm for Koundinya’s head; the Cobra’s thighs hover 2.8–7.8 mm up (Fig. 4). A 2 cm touch test would pass all of them; only load separates them. In the forearm stand, the hand colliders touch the floor, within 2 mm, while the forearms carry 706 N (median over 31 replicas). In the four recorded replicas, the left hand averages under 4 N and reaches 5 N on only about a fifth of hold frames, and the right averages about 16 N but reaches 5 N on only 80–83% of them, short of the 90% the rule requires.

Students. The initial student shows the same pattern in its 31-replica route evaluation: with sampled codes, 1,060 pose passes and 872 full passes, with 187 of the 188 differences missing a requested load; with greedy decoding, 516 and 413 of 615, with all 103 differences missing one. At the fixed endpoints, 284 of 285 and 310 of 311 gap replicas of the switched students miss requested loading, against 298 of 309 for the initial student in its 48-replica batch and 344 of 344 for no-switch seed 1. On switch seed 1’s Dolphin route, all 48 replicas pass pose and miss right-forearm loading. In the inversion flow, the Feathered Peacock target fails requested hand loading even for the teacher, which passes pose on 31 of 31 replicas and the full rule on none.

Relation to the reward. The teacher’s reward penalizes load on zones that should be free (Appendix C) but gives requested-

TABLE IX

INITIAL FAMILIAR-ROUTE PERFORMANCE. MEAN SUCCESS INCLUDES POSE AND GROUND LOADS; POSE-ONLY RATES OMIT THE LOAD CONDITIONS. THE TEACHER AND SAMPLED STUDENT USE 31 REPLICAS PER ROUTE AND THE GREEDY STUDENT 15.

Policy	Passing plans	Mean success	Pose only
Teacher	32/41	.777	.972
Student, sampled	27/41	.686	.834
Student, greedy	27/41	.672	.839

support attainment zero weight, so no reward term requires a requested zone to carry load. The students, distilled from this teacher, show the same failures; whether an attainment term would remove them is untested (Appendix L).

APPENDIX K

TEACHER AND INITIAL-STUDENT EVALUATION

The teacher and the initial student were evaluated before fine-tuning with the plans and scoring of Appendix G: 31 replicas per plan for the teacher and the sampled student, 15 for the greedy student, and 42 under the approach protocol. The teacher’s results are those of Table I; the initial student’s come from earlier batches than that table’s and differ slightly from them.

Familiar routes and revisions. The teacher passes 32 of 41 familiar routes and the initial student 27, with sampled or greedy codes, although the two decodings pass different sets (Table IX). The teacher reaches all six destinations of Table X under both protocols, although not every target of the plan groups below them. The initial student accepts some revisions only when they are revealed early: the jump from crow to plank succeeds on 0 of 31 replicas when shown at the crow hold and on 41 of 42 when shown during the approach, and the side-crow alternate behaves alike. This matches its training schedule. During distillation, the goal window lists the current segment and the clip’s following ones. A continuation is the next goal, behind the active one, for a median 3.8 s (10th–90th percentile 0.9–11.8 s) before it becomes active, and it never changes while its segment is in progress. Standing is mostly an ending: 70% of the time in standing segments is a clip’s final standing with no goal ahead, so a final standing was never followed by another clip’s goals.

A diagnostic varies when the plank becomes visible (Table XI). Shown from the start of the episode or from the

TABLE X

INITIAL TEACHER/STUDENT CROSS-CLIP COMMANDS. TOP: DESTINATION SUCCESS AFTER A HUMAN ARRIVAL AT THE SHARED HOLD. BOTTOM: MEAN TARGET SUCCESS BY PLAN GROUP, INCLUDING ARRIVAL TARGETS. SW.: CONTINUATION SHOWN ONCE THE SHARED HOLD BEGINS (SWITCH PROTOCOL). AP.: SHOWN WHEN THAT HOLD BECOMES THE ACTIVE GOAL (APPROACH PROTOCOL). ALL RUNS REISSUE GOALS NOMINALLY EVERY 0.25 s (EIGHT CONTROL STEPS).

Destination	Teacher		Student	
	Sw.	Ap.	Sw.	Ap.
Crow → handstand (press)	31/31	42/42	31/31	42/42
Crow → plank (jump)	31/31	42/42	0/31	41/42
Crow → chaturanga	31/31	42/42	0/31	0/42
Handstand → crow	31/31	42/42	10/31	18/42
Handstand → chaturanga	31/31	42/42	27/31	37/42
Side-crow squat → side crow ^a	31/31	42/42	0/31	40/42
<i>Plan group (mean target success)</i>				
Revised destinations + controls	1.000	1.000	.786	.876
Multi-hop chains	.933	.952	.561	.615
Standing flows	.836	.513	.305	.310
Family alternates	.747	.571	.297	.279

^a A within-family alternate. Under approach the student skips the squat's support: the squat passes on 0/42 under the full rule, with the left foot unloaded on every replica, although its pose passes on 32/42.

TABLE XI

INITIAL-STUDENT JUMP-TO-PLANK DIAGNOSTIC: 170 REPLICAS PER CELL, VARYING WHEN THE PLANK FIRST BECOMES VISIBLE. EVERY CROW SUCCEEDS; THE TEACHER REACHES PLANK ON 170 OF 170 THROUGHOUT.

Lead-in	Plank first visible	Base dwell	+3 s
Synthesized clip	Episode start	170/170	170/170
Human crow	First crow step	166/170	169/170
Human crow	Crow hold	1/170	0/170

first crow step, it is reached on almost every replica; shown once the crow hold begins, on 1 of 170, and 3 s of extra crow dwell does not help (0 of 170). The early variants inherit a shorter crow dwell, which shifts the deadline, so this table supports but does not isolate the reveal-time effect measured in Appendix I. Neither decoding nor lookahead closes the gap: greedy decoding gives similar group means, and showing every future goal from the start of the plan lowers them (chains .338 against .561).

Selecting the initial student. The initial student is the checkpoint after 5,000 epochs of the original distillation run, chosen among the four in Table XII with these probes and familiar-route retention; its crow-to-handstand command succeeds on 53 of 53 replicas in the sweep, against 37 of 53 at 6,000 epochs. The two latest checkpoints score higher on familiar routes but much lower on commands: from 6,000 to 7,000 epochs the command-plan mean falls from .583 to .173 while passing routes rise from 25 to 32 of 41 under the full rule. Of that 0.41 drop, 0.22 comes from each plan's first target, the shared hold reached along a human route. At 7,000 epochs, in the six command plans that start from crow, the crow passes pose on 50–52 of 53 replicas while the left foot carries at least 3% of body weight on at least a fifth of the hold on 49–52, and the press from that crow fails on all 53. Aggregate familiar-

TABLE XII

STUDENT CHECKPOINT SWEEP: FAMILIAR-ROUTE SCORES, POSE-ONLY OR FULL-RULE, PEAK AT EPOCH 7,000 WHILE COMMAND SCORES FALL.

Epoch	Pose-only routes	Pose-and-load routes	Crow control	Command mean	All targets
5,000	35/41	27/41	53/53	.554	.311
6,000	34/41	25/41	53/53	.583	.347
7,000	39/41	32/41	32/53	.173	.072
7,210	36/41	28/41	14/42	.345	.209

Pose-only routes: a separate pose-only evaluation, 48 replicas, passing threshold 0.5, goals reissued every 0.5 s; scoring the pose-and-load rollouts by pose alone instead gives 34, 33, 39, and 36. Pose-and-load routes: the same 41 route plans scored with the full rule, 48 replicas, goals reissued every 0.25 s. Commands and crow control: sampled stream, switch protocol, goals reissued every 0.10 s, 53 replicas at epochs 5,000–7,000 and 42 at epoch 7,210; shared six-body pose and ground-load criterion. Crow control: full-rule success of the familiar-route crow plan in the command evaluation suite (4.1 s crow hold).

TABLE XIII

CLIP TRACKING FROM TIME ZERO WITHOUT TERMINATION, 12 REPLICAS PER CLIP. ENTRIES ARE HUMAN / SYNTHETIC, WEIGHTING CLIPS EQUALLY. A FRAME PASSES WHEN EVERY BODY IS WITHIN 0.5 M OF THE REFERENCE; A CLIP PASSES WHEN EVERY FRAME DOES.

Policy	Frame share	Whole clip	Mean error (m)
Teacher	.983 / 1.000	.808 / 1.000	.065 / .042
Student, sampled	.800 / .916	.205 / .110	.165 / .069
Student, greedy	.764 / .895	.226 / .104	.183 / .079

route scores, even support-aware ones, would have missed this regression: the crow's own route, with a 1.1 s rather than a 4.1 s hold, degrades too (48, 48, 39, and 21 of 48 replicas), but other routes improve more. The plans used for this choice are the development probes of the main evaluation, so the evaluation is not an independent test.

Clip tracking. A separate evaluation starts each of the 84 base clips at time zero, 12 replicas each, without termination. A frame passes when every body is within 0.5 m of the reference, and a clip passes only if every frame does (Table XIII). The teacher tracks most clips in full, including every synthesized one; the student tracks most frames but few whole clips, and on synthesized clips its failures cluster near fast presses, jumps, and kick-ups, followed by recovery. Aligning the pelvis's horizontal position per frame changes the student's tracked share of human-clip frames only from 0.800 to 0.804, so global drift explains little.

Teacher diagnostics. AMP fine-tuning reduces the teacher's joint-angle discrepancy from the human references from about 12.6° to 7.3–7.4°, and the AUROC of a classifier separating policy from reference motion on held-out clips from 0.991 to about 0.80; no matched run without AMP isolates its effect. Commanded from held exemplars, the teacher reaches every new edge's destination on 163 of 163 replicas, whereas before graph growth the same tests reached both chaturanga destinations but none of the press, the lowering, or the jump to plank. In the teacher's executions of the synthesized edges, one replica per instance, body speeds peak at 1.3–2.8 times the human corpus's per-body 99th percentile, so their naturalness is not established.

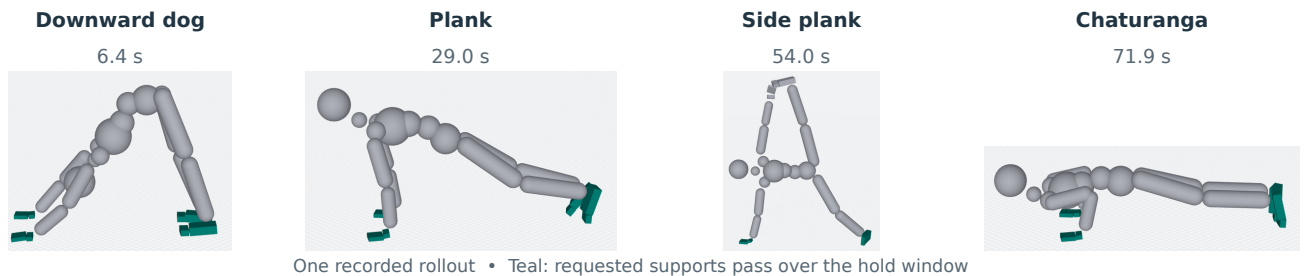


Fig. 9. One successful floor sequence from switch seed 1 at the fixed endpoint: downward dog, plank, side plank, and chaturanga are connected *via standing*. Intervening standing returns are omitted from these snapshots. All four designated holds pass the pose-and-load rule; teal marks requested supports. This flow completes in 263/409 replicas.

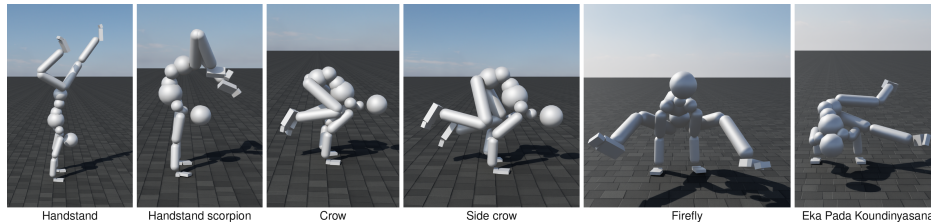


Fig. 10. Selected poses of the teacher used throughout: handstand, handstand scorpion, crow, side crow, firefly, and Eka Pada Koundinyasana.

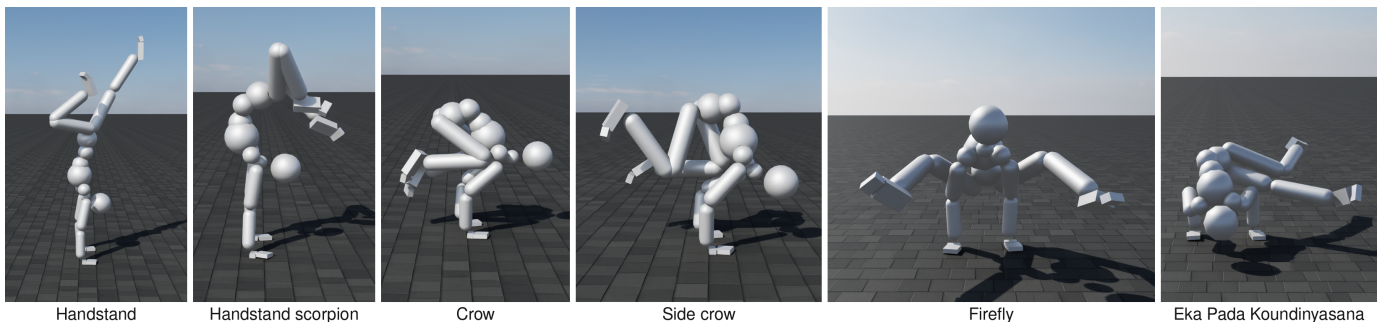


Fig. 11. Selected poses of the initial student, using the teacher gallery’s plans and cameras: handstand, handstand scorpion, crow, side crow, firefly, and Eka Pada Koundinyasana.

APPENDIX L FUTURE COMPARISONS

Several comparisons would separate contributors to the gains reported here. Exposure-matched fixed-time switches, and one switch against several per episode, would separate timing randomization from the number of transitions experienced. Takeovers from common pre-action states and histories would separate arrival-state quality from downstream execution, and independent changes to dwell and deadline would separate preparation time from evolution under the old goal. Matched students with a continuous latent or a direct goal input would show how the held discrete code affects the benefit of switch collection, and crossing clip-local and revised collection with and without student policy optimization would compare a change of collection with a change of objective. Requested-support rewards need their own comparison, because the loading failures already occur in the teacher, and feasible goals that request different contacts for

similar poses would test whether contact commands can be followed independently of pose. AMP, curation, and graph expansion are shared by every run here; attributing benefits to them would need matched controls.

APPENDIX M QUALITATIVE EXAMPLES

Figure 9 shows a successful standing-joined floor sequence from switch seed 1. The intervening returns through standing connect the four familiar skills in one rollout.

Figures 10 and 11 show selected teacher and student arm balances and inversions. The pictured hands-only Scorpion comes from a recording whose holds lie outside the labeled windows; the Scorpion route evaluation uses another recording with hand and forearm support. The frames illustrate the repertoire, while Section V reports pose-and-support success over the scored holds.