

Pose Is Not Support: Do Learned Whole-Body Controllers Follow Revised Support Commands?

Visak Kumar
Independent Researcher
San Francisco, CA, USA
vishak.vk@gmail.com

Abstract—Contact-rich loco-manipulation requires whole-body controllers to change which limbs bear weight when a planner or user revises a goal. Evaluating familiar motion sequences by pose accuracy does not establish whether a controller follows such revisions or loads the requested supports. We examine both properties in a simulated yoga testbed with human arm balances, inversions, and single-leg balances, supplemented by five synthesized transitions. A goal-conditioned PPO teacher executes all five transitions after arrivals through human clips. A student distilled to accept partial pose-and-contact goals shows a strong timing dependence: crow-to-plank succeeds on 41 of 42 replicas when the continuation appears as the crow becomes the active goal, but on none of 31 when it appears during the crow hold. Its training schedule contains no such within-hold revision. Separately, pose-only scoring exceeds the teacher’s own-route pose-and-load success by about 20 percentage points, entirely through failures of requested-support loading. Across student checkpoints, own-route scores peak while command success regresses. These results show the value of evaluating command timing and support realization alongside familiar-route performance, and motivate training with revised continuations and explicit support-attainment objectives. Videos and the project page: <https://visakk.github.io/pose-is-not-support/>.

I. INTRODUCTION

A humanoid that kneels to lift a box, braces a hand on a wall, or gets up from the floor changes which limbs carry its weight. A planner may revise these supports during a task, requiring the controller to establish new contacts and release old ones [1]–[3]. Learned whole-body controllers offer goal-conditioned or distilled command interfaces [4]–[6], with longer behaviors formed by sequencing or switching skills [7]–[12]. We examine revisions during balances supported by hands, forearms, or the head, asking whether the controller follows the new goal and whether the requested supports carry load. Both properties matter for manipulation: a braced hand may need to become available for grasping, while a retained brace must transmit the forces assumed by the plan.

Composition requires each switch to place the system where the next skill can succeed [13], [14]. For a learned policy, success may also depend on command timing and observation history. Familiar-sequence pose scores provide limited information about either the response to revised goals or the realized support loads (Fig. 2).

Our testbed uses human yoga capture [15] to study support transitions without object dynamics. Arm balances, inversions, and single-leg balances involve feet, hands, forearms, head support, and self-contact. We repair the references to

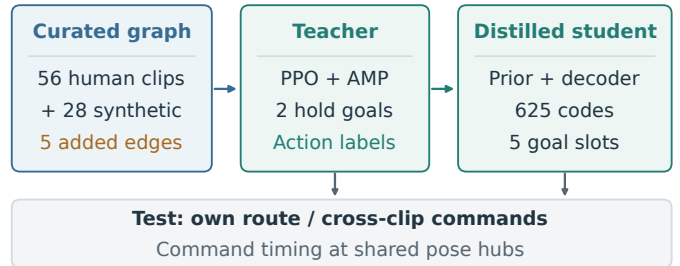


Fig. 1. System and evaluation. Curated goals and synthesized edges train a teacher, whose labels train a student; both are tested on own routes and on commands from other clips at shared poses.

align labeled supports with the simulated body, synthesize five transitions absent from the capture, and train a goal-conditioned teacher. These additions include pressing from crow to handstand and jumping from crow to plank. To reduce the teacher’s requirement for full 24-body goal poses, we follow the masked-conditioning and distillation approach of HOVER and MaskedMimic [4], [5], using a student interface with partial poses and explicit contact requests.

Our contributions are: (i) an evaluation of own-route execution and cross-clip commands that varies command timing and scores ground loading alongside pose (Fig. 1); (ii) a teacher that executes synthesized arm-balance transitions after arrivals through human clips and chains them across clips; and (iii) an empirical analysis of timing sensitivity, missing requested loads, and checkpoint regression.

The policies build on physics-based imitation and adversarial motion priors [16], [17], using established PPO, MPPI, online teacher labeling, and finite scalar quantization [18]–[21]. Command-conditioned locomotion policies commonly resample commands within episodes [22], and goal relabeling decouples goals from trajectories [23]. Our within-clip schedules preserve the association between motion history and demonstrated continuation; the cross-clip tests challenge that association.

II. TESTBED AND POLICIES

Capture. MOYO [15] records one performer in about 200 complex yoga poses with marker-based motion capture, a pressure mat, and SMPL-X fits. Each clip performs one pose as a round trip from standing (two clips start mid-motion),

often through preparatory holds; variants of one pose form a family of separate clips, and no clip moves between two poses. From 60 clips centered on single-leg balances, arm balances, and inversions, we retain 56 (Appendix A): 37 families in 13 single-leg-balance, 13 inversion, 9 arm-balance, and 21 connective clips of 11–55 s.

Contact-configuration graph. We annotate each clip as a sequence of holds (segments), windows in which one contact configuration is held (four per clip at the median). A node is a hold’s pose name, ground-contact set, and pelvis-orientation bin (body–body braces belong to the hold’s contact request, not its node), and a directed edge joins the nodes of consecutive holds in a clip. Because nodes carry pose names, standing is the only node that clips of different families share; half of the 196 captured edges touch it. Mesh geometry, markers, pressure, and rigid-body statics label each hold’s touch, floor loading, and mechanical utility, and a vision-language reviewer supplies provisional brace roles (Appendix A). Support-aware scoring needs references whose labeled supports touch the floor, and the initial references left 434 of 864 labeled supports more than 2 cm up. Rebuilding the body on the performer’s skeleton, using the performer’s own MoSh fit, and re-solving on measured contacts leaves one such support; 270 holds are released (Appendix A).

MPPI in the training simulator synthesizes five edges no clip contains (crow to handstand and back, crow to plank, and crow and handstand to chaturanga), spliced into clips with captured lead-ins and lead-outs. The 56 human clips and 28 synthesized trajectories give 84 base clips and 252 motion instances on 123 nodes and 201 edges. Candidates were selected by hand, with relaxed landing-slip criteria (Appendix B).

Teacher and student. Both control the same 24-body, 74 kg avatar through 69 PD targets at 30 Hz in Isaac Lab/PhysX [24], [25]. The PPO teacher reads proprioception, contact observations, 0.5 s of pose history, and two fully visible hold goals, each with a 24-body pose, ground and brace requests, and timing; dense references shape its rewards and critic but not its actor. An AMP fine-tune precedes training on the expanded graph. The student reads five goal slots, each with a six-body pose (pelvis, head, hands, ankles), the full contact request, pelvis orientation, and timing, plus 2 s of pose history and four measured contact-event tokens; training masks pose and contact visibility. A privileged encoder emits one of 625 intent tokens, a prior predicts it from deployable inputs, and a shared decoder acts; goals reach the decoder only through the token, refreshed every eight control steps or on a contact event. Online distillation uses the frozen teacher’s action labels; after a ramp, 25% of environments execute the sampled prior. No student PPO or perturbation training is included. For both policies, training goals come from the playing clip’s own schedule, not from walks on the graph: the window advances when a segment ends, and a continuation never changes while its segment is in progress (0 of 1,146 segments; Appendix F).

III. EVALUATION

We evaluate the teacher at epoch 3,420, where its run ended, and the student at epoch 5,000 of a run stopped at epoch 7,210, selected using command probes and route retention; these are therefore development-set results. The teacher uses mean actions; the student samples tokens at temperature 1 with top- $p = 0.9$. Replicas measure rollout variation within one training lineage.

Success criterion. For each commanded hold, the temporal mean of the mean position error over six bodies (pelvis, head, hands, ankles), pelvis-relative and best-yaw aligned per frame, must be below 0.15 m; each requested ground zone must carry at least 5 N on at least 90% of hold frames; and no unrequested zone may carry 3% of body weight on 20% of frames. Forces are terrain-filtered. Brace closure, absolute heading, and transition frames are not scored, nor are targets with zero hold duration.

Own routes. Each plan follows one clip’s own goals from its start through its pose hold and one exit hold. We score one target per plan: one pose hold per family, excluding zero-duration Peacock, plus the five edge destinations (41 targets total). A plan passes if at least half its replicas succeed; the teacher and the sampled student use 31 replicas per plan, and a greedy student test uses 15.

Commands from another clip. A plan follows clip A to a node it shares with clip B, then commands B’s continuation without a reset; a per-goal yaw and XY fit places B’s goals in A’s frame. Clips share only standing, nodes within a family, and the ends of the synthesized edges, so the battery has five new-edge commands with two own-route controls, five multi-hop chains, five flows joined at standing, and seven within-family alternates. Under *switch*, B’s continuation appears once the hub’s hold begins, replacing A’s continuation inside a segment, which training never does. Under *approach*, it appears when the hub becomes the active goal, at a window advance, the only moment at which training changes which goals the window holds (Fig. 3). Group scores average target success within plans, then plans equally. Switch uses 31 replicas per plan and approach 42; both reissue goals every 0.25 s.

IV. RESULTS

A. Skill execution transfers incompletely

The student retains 27 of the teacher’s 32 passing own-route targets and adds none (Table I). Crow, handstand, firefly, the other single-leg balances, and the press, lowering, and jump-to-plank targets transfer; the Scorpion and Half Moon targets and the chaturanga landings do not. The landings succeed on at most .35 of student replicas, against 1.00 for the teacher. Conditional on passing the pose criterion, support-success rates are similar: .82 for the student and .80 for the teacher. Hold-level and whole-clip tracking show larger gaps, widest in inversions (Appendix F).

B. The teacher composes across clips

The teacher executes all five new edges after arriving through a human clip, with source and destination both sat-

TABLE I

OWN-ROUTE PERFORMANCE. MEAN SUCCESS INCLUDES POSE AND GROUND LOADS; POSE-ONLY RATES OMIT THE LOAD CONDITIONS.

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

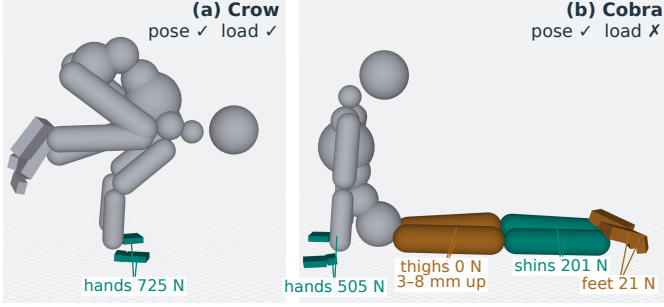


Fig. 2. What a pose-only pass hides. Two teacher own-route targets, each drawn from one scored replica at mid-hold; requested ground zones are teal if that replica loads them under the Section III rule and amber if not; loads are medians over 31 replicas. (a) Crow: the hands carry the body’s weight. (b) Cobra: the pose matches as closely (3 cm), but the thighs carry 0 N on every frame of all 31 replicas and sit 3–8 mm above the floor in the four fully recorded ones, which a 2 cm touch test accepts; the feet touch but load intermittently. MuJoCo render of recorded PhysX states.

ified on every replica (Table II); it also completes the chain crow–handstand–crow–chaturanga on every replica and, under switch, chains clips at standing (crow, firefly, and side crow on every replica). At the new-edge hubs the two clips share the exemplar pose by construction (junction residual 0 m), so these tests probe arrival history more than pose generalization; the within-family alternates, joined at poses 0.02–0.24 m apart, are the stronger evidence (.747).

C. Command timing affects student success

With goal poses, deadlines, dwell, and reissue interval fixed, student jump-to-plank success changes from 41/42 replicas under approach to 0/31 under switch (Table II, Fig. 3). The side-crow alternate shows a similar contrast (note a), while the press succeeds on 31/31 when commanded during the same crow hold. A separate 170-replica diagnostic reproduces the jump contrast after the human crow route (Appendix F). The plank score covers a 0.30 s arrival window; sustained plank holding is not measured.

This timing dependence is consistent with the training schedule. Continuations remain fixed within each training segment: an approach reveal occurs at a window advance, whereas a hold-time revision does not. The student’s 2 s pose history and four contact-event tokens could identify the playing clip and its usual continuation, providing a possible source of causal confusion [26], [27]. The teacher, with 0.5 s of history, accepts revisions during holds but fails under approach in three of five standing flows, where the next pose enters its second slot during the return to standing. Such continuations are rare

TABLE II

COMMANDS FROM ANOTHER CLIP. TOP: DESTINATION SUCCESS AFTER A HUMAN ARRIVAL AT THE SHARED POSE. BOTTOM: MEAN TARGET SUCCESS BY PLAN GROUP, INCLUDING ARRIVAL TARGETS. SW.: CONTINUATION SHOWN ONCE THE HUB’S HOLD BEGINS. AP.: SHOWN WHEN THE HUB BECOMES THE ACTIVE GOAL. ALL RUNS REISSUE GOALS EVERY 0.25 s.

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>				
New-edge hubs + 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 (0/42).

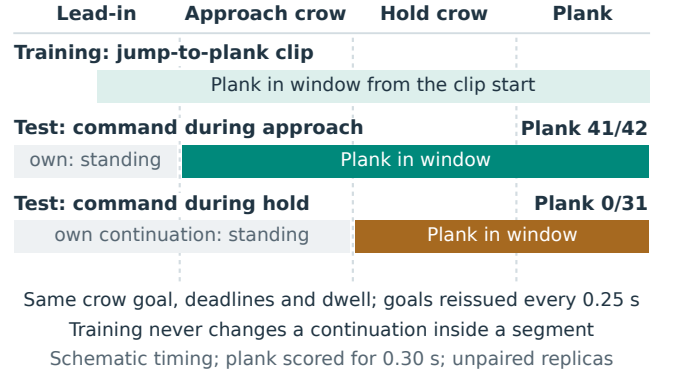


Fig. 3. When the plank enters the student’s goal window after the human crow route. Training exposes the jump clip’s plank goal from the clip start. The approach protocol reveals it at an active-goal change; switch replaces the continuation during the crow hold. Goal poses, deadlines, and dwell are held fixed; phase lengths are schematic.

in training: 228 of the 252 motion instances end standing, and mid-clip standing with a subsequent goal occupies only 1.4% of standing-segment time. This scarcity may also contribute to the student’s failure to leave standing for another clip under every tested protocol. Greedy decoding reproduces the plank and flow failures. Full lookahead also fails to resolve them: the source crow then succeeds on only 3/42 replicas.

D. Pose-only scores miss requested-load failures

Pose-only scoring passes targets with missing or insufficiently persistent requested loading. On own routes the teacher passes the pose criterion on .972 of target replicas and the full rule on .777 (Table I, Fig. 4a). All 248 replicas in the gap fail a requested-zone load condition, across eight families. Examples include the thighs in Cobra (Fig. 2b), the hands in the forearm stand, the trunk in Plow and Shoulderstand, and the head in Koundinya. None fails the unrequested-support condition; the student shows the same pattern in 185 of 188 cases. In each

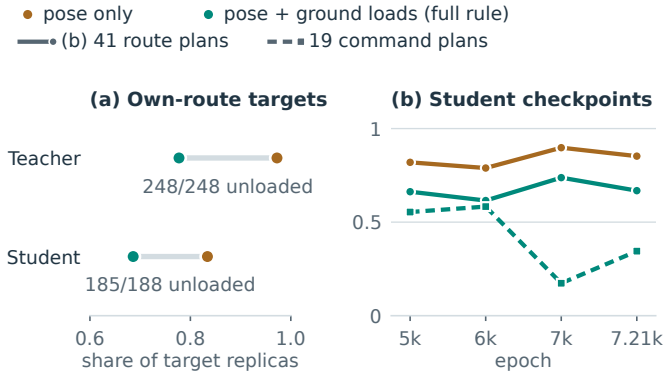


Fig. 4. Pose and support scores across targets and checkpoints. (a) Fraction of own-route target replicas passing pose alone and the full rule; all teacher cases in the gap fail a requested-zone load condition. (b) Student checkpoints: mean success over the 41 own-route plans, scored by pose alone (brown) and by the full rule (solid teal) on the same rollouts, and mean target success over 19 command plans (dashed teal). Appendix F gives the counts and protocols.

family’s fully recorded replicas, the missing supports remain within 2 cm of the floor throughout the hold. The forearm stand’s hands even touch the floor while the forearms carry 97% of body weight. These cases pass a 2 cm proximity test despite failing the load criterion. The teacher’s reward has a corresponding asymmetry: it penalizes unwanted support but assigns zero weight to requested-support attainment (Appendix C). We have not tested an added attainment term. Its design must account for commands that specify support zones without force targets and for intermittent contact, such as the 3–4 Hz tapping observed with an earlier unwanted-support penalty. Some failures may also reflect attribution between adjacent zones, such as upper arms and trunk.

E. Own-route scores miss a command regression

Own-route and command scores diverge across student checkpoints, whether routes are scored by pose alone or with support checks on the same rollouts (Fig. 4b): from epoch 6,000 to 7,000, passing plans rise from 33 to 39 by pose and from 25 to 32 with support checks, while mean success on 19 command plans falls from .583 to .173. More than half of the decline occurs at hub holds reached through long-dwell

human clips. There the crow retains its pose but loads the left foot on 49–52 of 53 replicas, compared with 9 of 48 on its own-route plan. The subsequent press fails on every replica although its own route succeeds (Appendix F). Thus own-route scores miss a regression that includes source-hold stability as well as continuation performance.

V. DISCUSSION

A. Scope of the evidence

The study uses one performer, one body, one training seed per stage, and development-selected checkpoints. The six-body criterion covers ground loads and a subset of pose coordinates; brace closure and partial-goal control are untested, and destination windows can be short. Pose and contact requests also co-vary, so the tests do not isolate use of the contact-command channel. The benefits of curation, AMP, synthesis, and quantization have not been separated, and object interaction and hardware transfer remain untested.

B. Implications and next steps

The results motivate command-revision and load measurements in contact-rich benchmarks, including controllers distilled from demonstration goals such as HOVER and Masked-Mimic [4], [5]. Randomized continuation switches during DAGger collection would directly test the training-schedule hypothesis; the teacher can label the held-hub switches on which it succeeds. History interventions at matched arrivals could help distinguish dependence on past observations from differences in physical state. Requested-support rewards should address both loading and persistence. Support-scored command probes can also complement own-route evaluation when selecting checkpoints. Appendix F develops these comparisons.

VI. CONCLUSION

The teacher and student expose two gaps in familiar-route pose evaluation: requested supports can remain insufficiently loaded, and a continuation that succeeds under one command schedule can fail under another. Own-route scores can also improve while command-battery performance regresses. Evaluating revised goals and realized support loads makes these failures visible and provides a basis for testing changes to training schedules and support rewards.

REFERENCES

- [1] 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.
- [2] S. Zhao *et al.*, “ResMimic: From general motion tracking to humanoid whole-body loco-manipulation via residual learning,” arXiv:2510.05070, 2025.
- [3] R. Yu *et al.*, “OmniContact: Chaining meta-skills via contact flow for generalizable humanoid loco-manipulation,” arXiv:2606.26201, 2026.
- [4] 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.
- [5] 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.
- [6] C. Zhang, W. Xiao, T. He, and G. Shi, “WoCoCo: Learning whole-body humanoid control with sequential contacts,” in *Proc. Conf. Robot Learn. (CoRL)*, PMLR vol. 270, 2025, pp. 455–472.
- [7] L. Kovar, M. Gleicher, and F. Pighin, “Motion graphs,” *ACM Trans. Graph.*, vol. 21, no. 3, pp. 473–482, 2002.
- [8] R. Yu *et al.*, “SkillMimic-V2: Learning robust and generalizable interaction skills from sparse and noisy demonstrations,” in *Proc. SIGGRAPH Conf. Papers*, 2025.
- [9] Z. Wu *et al.*, “Perceptive humanoid parkour: Chaining dynamic human skills via motion matching,” in *Proc. Robot.: Sci. Syst. (RSS)*, 2026.
- [10] P. Xu *et al.*, “Learning to ball: Composing policies for long-horizon basketball moves,” *ACM Trans. Graph.*, vol. 44, no. 6, 2025.
- [11] 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.
- [12] Q. Liao *et al.*, “BeyondMimic: From motion tracking to versatile humanoid control via guided diffusion,” *Sci. Robot.*, vol. 11, no. 117, Art. no. eadx8924, 2026.
- [13] 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.
- [14] 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)*, PMLR vol. 164, 2022, pp. 406–416.
- [15] 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.
- [16] 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.
- [17] 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.
- [18] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal policy optimization algorithms,” arXiv:1707.06347, 2017.
- [19] G. Williams *et al.*, “Information theoretic MPC for model-based reinforcement learning,” in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2017, pp. 1714–1721.
- [20] 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.
- [21] F. Mentzer, D. Minnen, E. Agustsson, and M. Tschannen, “Finite scalar quantization: VQ-VAE made simple,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2024.
- [22] N. Rudin, D. Hoeller, P. Reist, and M. Hutter, “Learning to walk in minutes using massively parallel deep reinforcement learning,” in *Proc. Conf. Robot Learn. (CoRL)*, PMLR vol. 164, 2022, pp. 91–100.
- [23] M. Andrychowicz *et al.*, “Hindsight experience replay,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017.
- [24] NVIDIA: M. Mittal *et al.*, “Isaac Lab: A GPU-accelerated simulation framework for multi-modal robot learning,” arXiv:2511.04831, 2025.
- [25] C. Tessler *et al.*, “ProtoMotions3: An open-source framework for humanoid simulation and control,” GitHub repository, 2025. [Online]. Available: <https://github.com/NVlabs/ProtoMotions>
- [26] P. de Haan, D. Jayaraman, and S. Levine, “Causal confusion in imitation learning,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2019.
- [27] 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.
- [28] 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.
- [29] 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.
- [30] 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.
- [31] 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.
- [32] J. Wang *et al.*, “HIL: Hybrid imitation learning for dynamic athletic control,” *ACM Trans. Graph.*, 2026, doi: 10.1145/3829364.

APPENDIX A CAPTURE CURATION

The reference corpus combines motion capture, contact measurements, and model-based checks. The evaluations below characterize reference quality and reviewer performance; they do not isolate the effects of AMP, curation, contact annotations, graph expansion, or quantization on policy learning.

A. Capture, body, and contact evidence

We start from 60 MOYO clips [15] at 60 Hz, with MoSh++ fits [28], a personal SMPL-X template [29], 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 performer 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. 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. Ground contacts must be within 2 cm and body-pair contacts within 1 cm. Feasibility requires every configured support and $s^* \leq 1$. 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.

Of the final 63 braces across 45 holds, none is individually necessary, 48 are useful, and 15 are redundant under this test. Removing one increases peak utilization by more than 0.05 for 22 braces. Collective necessity differs across clips. 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.

TABLE III
FIXED-CONTACT REFERENCE AUDIT 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; penetration counts human-touch zone-frames. Slide is the median over clips of the 90th-percentile speed. ^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, so the earlier 84.5% versus 32.5% contrast mostly reflects the box.

B. Reference repair

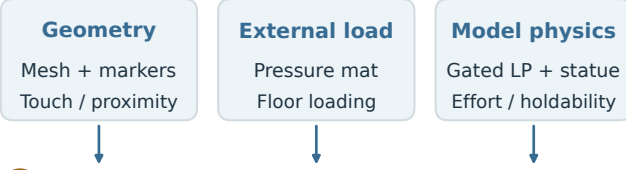
The initial body had legs about 5 cm short and a head collider ending about 10 cm below the crown, and the initial references copied the AMASS-neutral SMPL-X fit’s angles onto it, with the wrist rotation applied to the hands. The own-fit condition in Table III uses the performer’s MoSh fit, with finger-based 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. Comparison with the fit detects no added acceleration spikes or deep overlaps, although it does not certify every frame’s dynamics. The release retains 270 holds: 204 pass, 65 are flagged, and one has a masked contact. Eight transition windows and two scorpion holds beyond the modeled wrists are excluded from the goal graph and hold evaluation; their frames remain in the imitation references.

C. Contact review and calibration

Figure 5 summarizes the evidence hierarchy. Packets contain at most six deterministic CPU-rendered images, with distinct limb hues, shadowless marker glyphs, and colored drop lines for parts 1–15 cm above the floor (hands, feet, and head up to 60 cm). Content hashes, structured responses, and call records support replay. The reviewer uses Claude Opus 5.5 in a fresh context, with read-only packet access and a fixed JSON response schema. A MuJoCo statue test [30] 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.

1 Measure distinct evidence



2 Review traceable packets

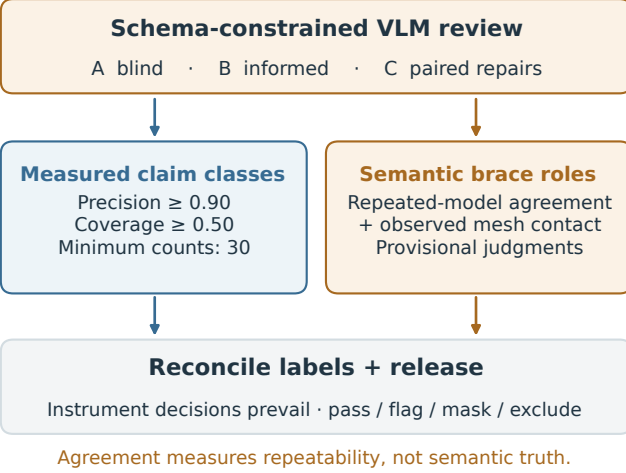


Fig. 5. Curation reviewer. Instrument-decided states remain authoritative. Claim-class calibration applies where decisive comparison labels exist; semantic brace roles use repeated-model agreement and remain provisional. Advisory variant flags do not change contact bits or exclude holds.

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 admission thresholds were set on the development set and carry no statistical guarantee.

Across three gap-rendering versions evaluated on the same 52 packets, 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. Final critical braces require both calls’ agreement and human mesh contact: 59 across 44 holds, plus four Peacock braces restored by torque feasibility, giving the final 63 braces across 45 holds. 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

GRAPH EXPANSION AND MOTION SYNTHESIS

Nodes combine pose name, floor-contact set, and pelvis-orientation bin; brace sets remain occurrence-specific. The initial graph has 123 nodes and 196 directed edges. We selected five missing edges by hand (Table IV) and specified each with curated exemplars, phase templates, and intermediate sketches for pressing, lowering, jumping, and landing; the procedure does not discover useful edges on its own.

The release regenerates the edges by MPPI in PhysX with the learning body’s PD semantics. MPPI [19] optimizes spline-knot offsets U to nominal targets. For finite rollout costs J_i , the update is

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

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

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.

No candidate passed the original endpoint and seam criteria. We subsequently relaxed the planted-hand drift tolerance to 1 cm for the press/lowering edges and accepted exemplar palm offsets of up to about 12° yaw. Landing drift and end-seam tests remained advisory, while the start-seam criterion remained mandatory: accepted slides were about 7–26 cm to plank, 2–5 cm from crow to chaturanga, and 11–12 cm from handstand to chaturanga. One of 29 selected candidates was removed for persistent knee contact. Captured lead-ins and lead-outs are spliced with alignment, blending, and depenetration; seam frames need not be unmodified dynamical rollouts.

Each of 28 trajectories receives 0, 3, and 7 s extra source dwell, yielding 84 synthetic instances. The final release contains 84 clips, 252 motion instances, 123 nodes, and 201 edges. Because the synthetic instances are dwell variants of 28 trajectories on five edges, they add five edge types rather than

TABLE IV
GRAPH ADDITIONS AND TEACHER COMPOSITE EDGE TESTS.

ID	Edge	Trajectories	Passing instances
E1	Crow $\rightarrow$ Handstand	8	24/24
E3	Handstand $\rightarrow$ Crow	3	9/9
B1	Crow $\rightarrow$ Plank	5	15/15
E2	Crow $\rightarrow$ Chaturanga	11	33/33
E5	Handstand $\rightarrow$ Chaturanga	1	3/3

84 independent transitions. The human portion is preserved. The release 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 III. 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 float-down, at 0.94–0.96.

APPENDIX C TEACHER ARCHITECTURE AND OBJECTIVES

A. Observations and actuation

The asymmetrically trained PPO actor [18], [31] 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 known free in the release 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 an earlier run on uncurated references, a per-frame clamp of such a load was beaten by a foot tapping at 3–4 Hz that kept 68–82% of its resting load for a quarter to a third 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 COM distance inside the commanded support polygon in meters. All three penalties are disabled during reset grace. Required-support/brace attainment, pair load, and slip have zero reward weight. Thus brace closure can arise through pose imitation and dynamics; direct brace-command control is not established.

C. Adversarial prior and PPO update

AMP [17], [32] 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 Scorpion-b, 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 student uses masked conditioning [4] and finite scalar quantization (FSQ) [21]. 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; the first slot always reveals at least its contact half; 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,

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

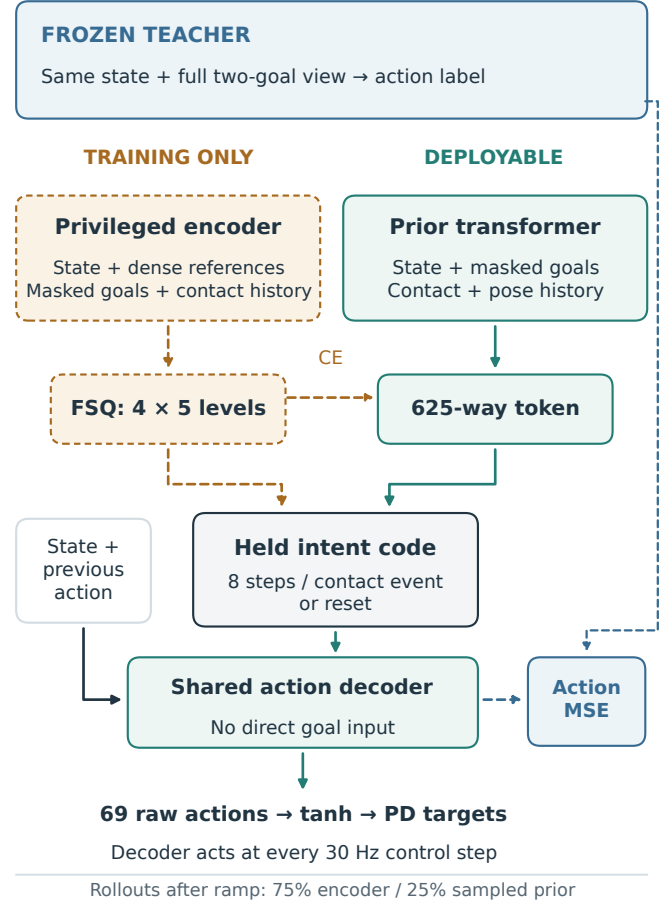


Fig. 6. Teacher/student interfaces. Dense references inform the teacher’s critics, reward and resets, and the student encoder. The deployable prior sees state, masked goals, and history, then supplies a discrete code to the decoder. The frozen teacher labels every visited state.

The 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. 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 has two layers at the same width. 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 [20]; goals still follow the played clip’s own schedule, so a continuation never changes during its segment and no foreign continuation appears at a held hub.

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 ,

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

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. 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.

APPENDIX E

IMPLEMENTATION, SCHEDULES, AND COMPUTE

All stages use GPU PhysX in Isaac Sim 5.1.0 via Isaac Lab [24] and ProtoMotions [25]. The software versions are ProtoMotions 3.1, Isaac Lab 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.

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 V). 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. Each stage has one seed (0).

TABLE V
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 continuation	4,096	5,000	655.4	15.42
Graph continuation	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.

The initial curated-reference teacher trains for 15,500 epochs. Its AMP continuation runs 5,000: $w_A = 0$ through epoch 200, is calibrated there to target 0.5, and reaches that weight by epoch 500. Graph continuation 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 continuation, and student initialize near arrivals with probability 0.6 and up to 0.5 s pre-roll. Graph continuation allocates 0.4 to arrival, 0.2 to departure, 0.08 to beginnings, and 0.32 uniformly in time. Each new edge has the package sampling weight of 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 release motion. Different training tracking gates prevent interpreting their scheduled success values as directly comparable. The teacher stages total 67.1 elapsed GPU-hours and 3.14 billion nominal transitions. Total wall-clock time through the selected student is 97.9 hours, or 111.4 including the rest of its stopped run.

APPENDIX F

EXTENDED EVALUATION

A. Common scoring and alignment

All command comparisons use a common six-body pose criterion: pelvis, head, both hands, and both ankles. Pose is pelvis-relative with per-frame best-yaw alignment, then averaged over the hold, so it discounts heading and omits uncommanded body details. We recompute the success conjunction from saved six-body errors and ground-load flags because the original logs used each actor’s conditionable subset: 24 bodies for the teacher and six for the student. Checking cases near rounded thresholds leaves the reported success unchanged. A matched 24-body command rescore across all replicas remains future work. Targets with zero hold duration are not scored: the commanded continuation of the Warrior III a→b alternate and the standing exit of the crow control, so those two plans contribute only their arrival targets.

Each foreign continuation is placed with a rigid yaw/XY map fitted on the two 24-body hub exemplars and com-

posed along the chain, separately for each goal slot. Junction residuals are 0 at synthetic hubs, 0.014–0.028 m at standing, and 0.022–0.237 m within families. No state reset occurs at a junction. All composition batteries and timing diagnostics reissue goals every 0.25 s, including the teacher’s matched switch run at 0.25 s. A teacher switch run at 0.10 s gives identical destination counts and group means within 0.016 (family alternates .763 against .747, from one Koundinya plan). The checkpoint sweep’s command runs use 0.10 s throughout. Hold durations are clip-specific. In particular, the off-route B1 plank target has only a 0.30 s scoring window, so its success does not certify a long sustained plank.

A route passes if at least half its replicas satisfy the joint pose/ground-load criterion of Section III. Under this six-body scoring, mean success is 0.777 full/0.972 pose-only for the teacher (0.779/0.974 in the 0.10 s run), 0.686/0.834 for the sampled student, and 0.672/0.839 for the greedy student. Passing plans are respectively 32/41, 27/41, and 27/41; greedy and sampled students pass different sets. The zero-duration Peacock target is excluded.

B. Checkpoint sweep

Table VI compares identical sets of 19 command plans (new-edge hubs and controls, chains, and family alternates; no standing flows). From epoch 6,000 to 7,000 the command mean falls by 0.41, of which 0.22 comes from each plan’s first target, the hub hold reached along a human route, and 0.19 from later targets. At epoch 7,000 the crow in the crow-start plans passes pose on 50–52 of 53 replicas while the left foot carries load on 49–52, and the press from that crow fails (0/53) although the press’s own route succeeds on 48/48 under the full rule. Scored with the full rule on the same 41 route plans (48 replicas), the student passes 27, 25, 32, and 28 plans at the four checkpoints (mean success .663, .615, .738, and .668), against 34, 33, 39, and 36 by pose alone on the same rollouts. The crow’s own route plan, whose hold lasts 1.1 s rather than the command plans’ 4.1 s, degrades too (48, 48, 39, and 21 of 48 replicas, with the left foot loaded on 0, 0, 9, and 25), but other route plans improve more. Epoch 5,000 is selected using these development probes and route retention. Its crow–handstand command succeeds on 53/53 replicas in the sweep, compared with 37/53 at epoch 6,000. The evaluation is therefore not an independent test set, and rollout replicas capture simulator and action-sampling variation rather than seed-to-seed training variation.

C. Matched clip tracking and conditional hold quality

A separate evaluation starts all 84 base clips at time zero for 12 replicas each without termination. A tracking frame passes if every body is within 0.5 m of dense reference; a whole clip passes only if every frame does. This dense-tracking measure is distinct from sparse-goal success and uses the same evaluation settings for both policies.

Per-frame pelvis-XY alignment changes the sampled student’s human tracked share only from 0.800 to 0.804, indicating little explanation by global drift. Synthetic failures cluster

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

Epoch	Pose panel	Full routes	Crow ctl.	Cmd. mean	All tgts.
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 panel: separate pose-only runner, 48 replicas, passing threshold 0.5, goals reissued every 0.5 s. Full 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 own-route crow plan in the command battery (4.1 s crow hold).

TABLE VII
MATCHED CLIP TRACKING. ENTRIES ARE HUMAN / SYNTHETIC; METRICS WEIGHT CLIPS EQUALLY.

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

near fast presses, jumps, and kick-ups, followed by recovery. In a separate evaluation of replica zero across 382 holds (270 human, 112 synthetic), 375 teacher holds and 309 student holds track at least 90% of the annotated window. Conditional geometric support rates are 357/375 versus 289/309; six-body exemplar-pose rates are 327/375 versus 265/309. Support requires each zone within 2 cm on 90% of its window, and pose uses temporal-median error below 0.15 m after the exemplar deadline. Both are conditional geometric quantities and differ from the joint loaded-support command score. Inversions show 67/69 versus 44/69 tracked holds. The teacher and student each show nine substitutions, seven of them shared; the student’s other two are one knee-supported landing and one foot down in Side Crow -c. These single-replica counts vary across reruns.

D. Timing diagnostics and alternate decoding

The same-plan crow-to-plank comparison in Section IV holds goals, deadlines, dwell, and reissue interval fixed and changes visibility protocol, improving student plank success from 0/31 to 41/42. Counts differ and simulator states are not paired. The corresponding teacher succeeds on 31/31 under switch and 42/42 under approach, at the same 0.25 s interval. The larger diagnostic in Table VIII also varies source dwell: early variants inherit B1’s 0.30 s dwell, while late variants inherit the human crow clip’s 1.13 s dwell. This shifts the destination deadline, so these results support the same-plan comparison but do not isolate reveal timing. One protocol difference remains in the same-plan comparison: under switch the reveal lands at the next periodic reissue without forcing an intent refresh, so the code refreshes within eight steps, whereas under approach the reveal coincides with an active-goal change and a refresh. Codes then keep refreshing every eight steps, more than a dozen times before the plank’s reach, and the +3 s variant in Table VIII also fails (0/170), so this delay is unlikely

TABLE VIII
STUDENT JUMP-TO-PLANK DIAGNOSTIC, 170 REPLICAS PER CELL. EVERY CROW HUB SUCCEEDS; THE TEACHER HAS 170/170 DESTINATION SUCCESSES THROUGHOUT.

Lead-in	Plank first visible	Base dwell	+3 s
B1 own route	Episode start	170/170	170/170
Human crow	First crow step	166/170	169/170
Human crow	Crow hold	1/170	0/170

TABLE IX
ADDITIONAL COMMAND CONTROLS: MEAN TARGET / ALL-TARGET SUCCESS. GREEDY STUDENT USES SWITCH AND 15 REPLICAS; LOOKAHEAD USES 42.

Group	Student greedy	Student early ^a	Teacher early ^a
New hubs + controls	.786 / .571	.561 / .497	1.000 / 1.000
Multi-hop chains	.632 / .173	.338 / .100	.949 / .848
Standing flows	.304 / .000	.211 / .000	.515 / .133
Family alternates	.329 / .162	.293 / .143	.571 / .429

^a“Early” is full lookahead from plan start. It differs from approach for the student and is identical to approach for the teacher.

to explain the 0/31 result; an immediate-refresh control would confirm it.

Showing every future goal from the start does not resolve the student gap: with plank several holds ahead, only 3/42 replicas satisfy source-crow pose and support. Table IX gives additional aggregate controls. The teacher has two goal slots and the student five, so lookahead exposes more distant future goals to the student. For the teacher’s two slots, approach and lookahead give identical inputs at every step, so the teacher’s early results in Table IX and approach results in Table II are independent runs of one protocol; their group means differ by at most 0.003. In three of five standing flows the teacher falls on all 42 early-command replicas; student standing flows have zero all-target success under every tested timing.

E. Training goal schedule and own-route support failures

Both policies’ training windows come from the playing clip’s own schedule, with the current segment included: slot 0 is the segment containing the current time, or the next segment in a gap, and the other slots hold the clip’s following segments. The window advances when a segment ends. Over the 1,146 segments of the released references, none changes its continuation while in progress, and a continuation sits in slot 1 for a median of 5.1 s (10th–90th percentile 1.9–13.8 s) before it becomes current. Of 396 standing segments, 228 end a clip, 162 begin one, and 6 lie mid-clip; by clip time, 70.3% of standing-segment time is a final standing with no goal ahead, 28.3% an opening standing, and 1.4% a mid-clip standing with a goal ahead. The student’s prior reads poses 1, 8, 15, 30, and 60 control steps back and four contact-event tokens; the teacher’s actor reads 1, 8, and 15 steps and no event history.

On the 41 own-route targets the teacher’s 1,271 target replicas include 1,236 that pass pose and 988 that pass the full rule (1,238 and 990 in the 0.10 s run). All 248 cases in the gap fail a requested-zone load condition: Bridge (forearm), Cobra

(thigh and foot), Dolphin Plank (forearms), the forearm stand (hands), Legs Up the Wall (head, trunk, and forearm), Plow and Supported Shoulderstand (trunk), and Koundinya (head). The battery 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, Koundinya’s head; the Cobra’s thighs 2.8–7.8 mm), and the forearm stand’s hand colliders touch the floor (within 2 mm) while the hands carry 20 N (2.8% of body weight) and the forearms 706 N (medians over its 31 replicas). Fig. 2 shows replica 0 of the crow and Cobra targets at mid-hold, with colors indicating that replica’s support outcomes. The sampled student has 1,060 pose passes and 872 full-rule passes, with 185 of 188 differences failing a requested-zone load condition; the greedy student has 516 and 413 of 615, with 102 of 103 differences failing that condition. Under lookahead, the student’s destination counts in Table II read 34/42, 5/42, 0/42, 21/42, 6/42, and 41/42; greedy decoding under switch reads 15/15, 0/15, 0/15, 4/15, 14/15, and 2/15.

F. Additional teacher diagnostics

AMP continuation reduces joint discrepancy from about 12.6° to 7.3–7.4° and a held-out-clip policy/reference classifier’s AUROC from 0.991 to about 0.79–0.81. It also changes support curriculum and adds 5,000 training epochs; a matched no-AMP continuation was not run. Its last-three-evaluation drag is 51.9J, above the 39.6J target. After graph continuation, human group scores averaged over the final three scheduled evaluations are 0.998/0.948/0.997/0.994 for single-leg/inversion/arm-balance/connective groups, but unwanted-support substitutions average 7.7 against a threshold of six. The selected epoch-3,420 checkpoint itself has eight substitutions and drag 49.9J, versus final-three drag 36.7J. These running averages describe several evaluations and differ from the selected checkpoint’s results.

Additional held-exemplar command panels achieve 163/163 destination poses for each new edge, while a training-derived standing–crow–handstand route reaches 94%. These panels start from held exemplars; the main command tests instead use human-clip arrivals. Action differences under changed goals are insufficient by themselves: the pre-growth teacher also shows them while failing transitions. A broad recovery region over perturbed, policy-reached sources remains unmeasured. New edges reach body-speed peaks 1.37–2.74 times the human corpus’s per-body 99th percentile; naturalness and general robustness are not established. An extended-hold test of Warrior III variant a falls on all 128 replicas.

G. Future controlled comparisons

The following proposed comparisons would test the mechanisms discussed above. They would match optimization steps, reset curricula, and checkpoint-selection rules across multiple training seeds. Final testing would use reserved command combinations and switch times, reporting both route retention and support-aware command success.

A direct test of the training-schedule explanation would compare DAgger collection with held-hub clip switches against own-schedule collection from the same checkpoint, using matched motion weights and training budgets. Randomizing when continuations are revealed and switched during collection, then sweeping test switches from the source’s first contact step through its hold, would measure sensitivity to command timing. Teacher labels should initially cover verified successful held-hub switches, since arbitrary early-switch labels require validation. Late-command success would be measured alongside retention of the current 27/41 route passes. Chained-clip collection and greater exposure to difficult inversion and landing segments could address the observed skill gaps.

The roles of history and code selection can be examined separately. Clearing pose and contact history at standing junctions while keeping the command plan fixed would test history dependence. Rolling out candidate intent codes would reveal whether useful departures exist but are missed by the prior. Enumerating all 625 codes would test single-code reachability, although it could miss a successful longer code sequence. A later student token-PPO stage with reference-independent graph-walk goals could improve code selection; with a frozen decoder, it would not supply missing motor behavior.

Teacher fine-tuning could randomize switch times and source dwell, correct support errors, and test perturbed arrivals around policy-reached hubs. Paired arrivals would reduce simulator variation in command comparisons. Longer destination holds, explicit transition scoring, a common 24-body command evaluation, and separate partial-goal probes would broaden the present evidence.

The effect of AMP requires matched continuations with and without the adversarial prior from the same pre-AMP teacher, controlling data, support curriculum, environment steps, and selection criteria. Posture and motion smoothness would be measured alongside route success, support substitutions, drag, and command composition. The present improvement also includes a changed support curriculum and additional training, so it cannot be assigned to the prior alone.

Reviewer-assisted curation can be compared with deterministic geometry, pressure, and statics while preserving instrument-decided floor labels. Holding the repaired corpus fixed would isolate changes to semantic brace labels, while a separate end-to-end comparison would measure coverage changes caused by admission decisions. Independent knowledgeable annotators would assess held-out brace roles and legitimate pose variants, since repeated-model agreement establishes repeatability rather than semantic validity.

Contact-specific control requires separating the conditioning channel from its supervision. A pose-only versus pose-and-contact comparison would hold measured contact observations, body, data, rewards, and training budget fixed; annotation quality and contact-reward terms would then be varied separately. Feasible fixed-pose alternatives with different requested contacts, scored by ground loads and body-pair closure, would test whether the controller responds to contact

requests. In current training, pose and contact requests co-vary and direct brace attainment has zero reward weight. Adding a requested-support reward with load averaged before saturation would test whether the loading failures in Section IV follow from the reward; sustained contact would still need to be assessed by the occupancy criterion.

Finally, the benefit of graph expansion can be measured against an equally long continuation without new-edge data. Continuous-latent and direct-action students with matched conditioning, rollout collection, training budget, and selection criteria would provide baselines for FSQ. The current single-teacher experiment does not establish an advantage of quantization or evaluate multi-expert consolidation.

APPENDIX G QUALITATIVE EXAMPLES

Figures 7 and 8 show selected teacher and student arm balances and inversions. The pictured hands-only Scorpion variant b lies outside labeled hold windows; the Scorpion route evaluation uses variant a with hand and forearm support. The frames illustrate the repertoire, while Section IV reports pose-and-support success over the scored holds.

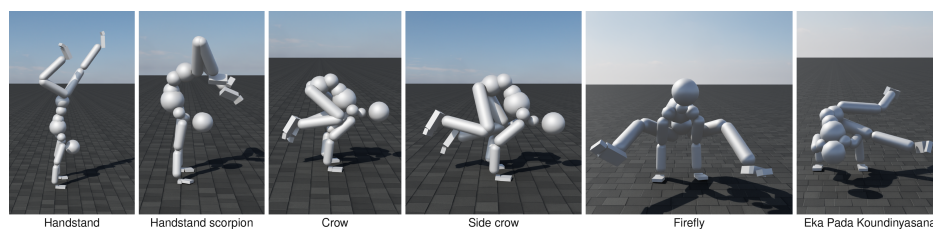


Fig. 7. Selected teacher poses at epoch 3,420: handstand, handstand scorpion, crow, side crow, firefly, and Eka Pada Koundinyasana.

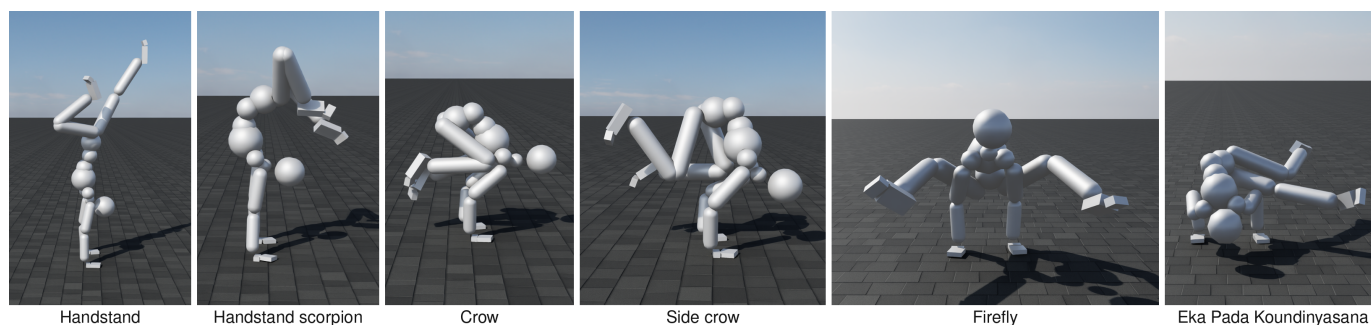


Fig. 8. Selected student poses at epoch 5,000, using the teacher gallery's plans and cameras: handstand, handstand scorpion, crow, side crow, firefly, and Eka Pada Koundinyasana.