MimicX: Policy-in-the-Loop Supervision Refinement for Video-Driven Humanoid Motion Tracking

Anonymous authors | Anonymous Website: https://anonymous-for-sub.github.io/MimicX
Paper under double-blind review

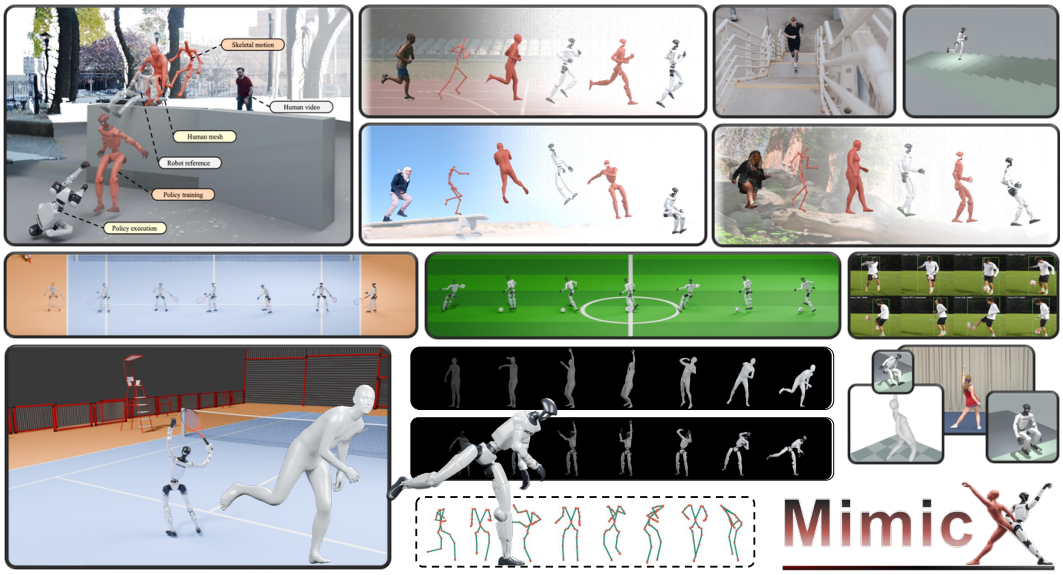


Figure 1: From human video to humanoid motion. Input videos, human reconstruction, robot references, and learned execution span athletic skills and collision-scene tasks.

Abstract

Human videos provide rich motion targets for humanoid learning, yet visually plausible references can still produce persistent failures under physics-based execution. These failures reveal where training supervision should change. We present MimicX, a policy-in-the-loop framework that uses execution feedback to refine video-driven humanoid motion tracking. Starting from reconstructed and retargeted motion, MimicX localizes difficult transitions and affected body regions, then jointly adapts tracking objectives and the reset curriculum for policy continuation. Repeated rollout verification selects execution-priority improvements subject to tracking guards. Across four core video tasks, MimicX consistently improves tracking accuracy and Robust Execution Horizon relative to the Fixed Reference baseline. Task-averaged results show a 25.7% reduction in body-tracking error and a 255.6% increase in execution horizon. Additional video, motion-reference, and collision-scene studies evaluate the method beyond the core tasks, while MimicX-HLoop accelerates feedback through heterogeneous execution. Overall, MimicX turns policy failure into actionable supervision for deciding what to refine and which refinement to retain.

1 Introduction

Human video provides a scalable specification of whole-body behavior. World-grounded reconstruction, retargeting, and physics-based tracking turn that specification into humanoid skills (Shen et al., 2024; Araujo et al., 2025; Liao et al., 2025; Luo et al., 2026). Yet a robot must realize the motion with different geometry and dynamics: plausible contact timing, root motion, or limb coordination

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

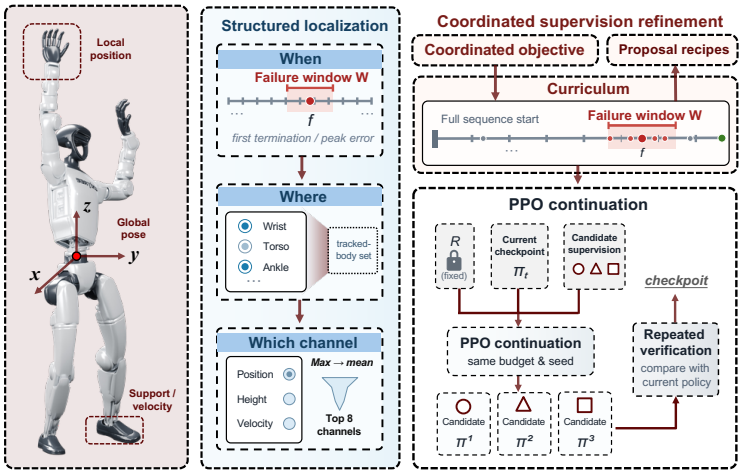


Figure 2: Execution feedback coordinates supervision. Rollout errors localize failure windows, affected bodies, and dominant local, global, or dynamic channels. These signals condition three proposals that coordinate tracking rewards and reset curricula. Each candidate continues PPO from the same checkpoint with a matched budget and seed. Repeated verification compares candidate execution with the current policy to select the retained update using execution-first criteria over repeated rollouts.

can repeatedly fail at the same transition. Policy execution should therefore inform the supervision used to learn it.

Our central insight is that rollout failures reveal *when* execution becomes unreliable, *which* body regions deviate, and *how* local, global, and dynamic errors develop together. MimicX uses this structure in a diagnosis–refinement–verification loop (Figure 2). Reconstruction and retargeting supply a reference; diagnosis conditions objective and curriculum updates, followed by policy continuation. A lower-body failure can jointly strengthen foot position, height, and velocity supervision around the difficult transition, while a root failure emphasizes anchor and torso consistency.

Reliable selection matters because a higher reward need not imply more reliable execution: a candidate can improve reward yet shorten the worst-case execution horizon. MimicX prioritizes completion, horizon, and tracking fidelity across repeated rollouts, retaining the current policy when a proposal fails verification. MimicX-HLoop accelerates feedback through heterogeneous execution. Across four core video skills, MimicX improves both tracking error and worst-case execution horizon relative to Fixed Reference. Tennis achieves full-horizon success in nine of nine evaluations versus one of nine for the baseline. Additional video, supplied-motion, direct-policy, and efficiency studies evaluate broader motion coverage and execution behavior. Figure 22 presents recorded motions in task scenes. Our contributions are **(1)** a policy-in-the-loop refinement framework that converts localized, body-specific execution errors into coordinated objective and curriculum updates; **(2)** repeated execution verification that selects reliable improvements while protecting stronger current policies; and **(3)** a heterogeneous backend that accelerates feedback while preserving selection outcomes.

2 Related Work

Video reconstruction and motion tracking. SMPL-X, GVHMR and GMR provide human geometry, world-grounded motion and robot references (Pavlakos et al., 2019; Shen et al., 2024; Araujo et al., 2025). Physics-based imitation and motion priors support tracking, recovery and partial-motion control (Peng et al., 2018; 2021; Luo et al., 2023; Tessler et al., 2024). BeyondMimic emphasizes difficult segments and guided diffusion; SONIC scales generalist tracking (Liao et al., 2025; Luo et al., 2026). MeshMimic couples reconstructed scene geometry with contact-aware motion retargeting (Zhang et al., 2026). MimicX instead makes measured execution failure an explicit input to supervision revision and validates the resulting policy through repeated rollouts.

Adaptive whole-body execution. Contact-aware tracking, timing adaptation and dynamics alignment motivate coordinated local, global and dynamic supervision (Xie et al., 2025; Huang et al., 2025; He et al., 2025). Recent methods separate motor learning from physical refinement, combine references with task goals, and compose long-horizon or end-effector skills (Wang et al., 2026b;a; Wu et al., 2026; Cao et al., 2026). MimicX uses a diagnosis–proposal–verification loop to select supervision updates through execution. Appendix H provides the detailed method-by-method comparison.

3 POLICY-IN-THE-LOOP SUPERVISION REFINEMENT

The outer loop uses execution errors to revise the supervision of a tracking policy. Its input is a motion reference and a current policy; its output is the policy retained after repeated execution verification. Given a monocular video, GVHMR reconstructs world-grounded SMPL-X motion and GMR retargets it to the humanoid (Shen et al., 2024; Pavlakos et al., 2019; Araujo et al., 2025). The resulting robot reference

$$R = \{\bar{q}_t, \dot{\bar{q}}_t, (\bar{p}_{t,b}, \bar{Q}_{t,b}, \bar{v}_{t,b}, \bar{\omega}_{t,b})_{b\in\mathcal{B}}\}_{t=0}^{T_R-1} \tag{1}$$

contains joint angles q, positions p, orientations Q, and linear/angular velocities v, ω for tracked bodies $\mathcal{B}$; bars denote reference values. A PPO policy $\pi_\theta(a_t \mid o_t)$ (Schulman et al., 2017) drives joint-position commands in MuJoCo/MjLab (Todorov et al., 2012; Zakka et al., 2026). The outer loop maintains an accepted checkpoint θ_k and proposes supervision bundles $U_{k,j} = (W_{k,j}, C_{k,j})$, where W parameterizes the tracking objective and C the reset curriculum. Figure 4 shows the reconstruction and execution stages on two video tasks.

3.1 LOCALIZING EXECUTION-CRITICAL TRANSITIONS

Each strict rollout records termination flags, reward, aligned body errors, absolute anchor errors, height discrepancies, and linear/angular velocity errors. Diagnosis uses the first scheduled evaluation repeat. Let f be its first flagged termination; when execution completes, we use the step with maximum tracked-body error. For evaluation horizon T, diagnosis focuses on

$$\mathcal{W}_k = [\max(1, f - 40), \min(T, f + 40)]. \tag{2}$$

Within this interval, channels are ranked by their maximum and mean values. Their body identities map them to local upper/lower-limb precision, global pelvis/torso stabilization, or dynamics and contact-related supervision. This interval supplies initialization states for practicing the transition.

3.2 COORDINATING TRACKING OBJECTIVES AND PRACTICE

We align reference articulation to robot anchor translation and heading, obtaining $\widetilde{p}_{t,b}$ and $\widetilde{Q}_{t,b}$, while supervising the absolute anchor separately. For body group G, a representative positional term is

$$\phi_p(G, \sigma) = \exp\left[-\frac{1}{|G|\sigma^2}\sum_{b\in G}\|p_{t,b} - \widetilde{p}_{t,b}\|_2^2\right]. \tag{3}$$

The full tracking objective combines selected local limb terms, global anchor/core terms, and dynamic velocity/contact terms, with action smoothing and joint-limit and self-collision penalties $\mathcal{P}_t$:

$$r_t(W) = \sum_{\ell\in\mathcal{L}} w_\ell\phi_\ell + \sum_{g\in\mathcal{G}} w_g\phi_g + \sum_{d\in\mathcal{D}} w_d\phi_d - \lambda_a\|a_t - a_{t-1}\|_2^2 - \mathcal{P}_t. \tag{4}$$

Thus a lower-body failure can strengthen position, height, velocity, and stance terms together; a root failure jointly emphasizes anchor and core consistency. The three groups couple task-critical precision to whole-body motion and physical consistency.

Diagnosis conditions three bounded supervision candidates: failure-window practice, stronger window practice with dynamics emphasis, and full-start consolidation. They adjust reward weights and scales, action smoothing, and the initial-state distribution. For practice interval $[s, e)$,

$$t_0 \sim (1-\rho)\delta_0 + \rho\,\mathrm{Unif}\{s, \ldots, e-1\}, \tag{5}$$

where δ_0 starts the motion from its beginning. The window candidates use $\rho = 0.45$ and 0.65, respectively; consolidation uses $\rho = 0$. This mixture pairs whole-sequence practice with repeated exposure to the difficult transition. Every candidate starts from the same accepted policy and receives the same continuation budget. A common task evaluates the resulting policies; Appendix B.3 derives the roles of objective and curriculum refinement.

3.3 SELECTING UPDATES THROUGH REPEATED EXECUTION

For candidate j and evaluation seed s, let $n_{j,s}$ count terminations and $h_{j,s}$ be the first terminated step, using $T+1$ when no failure occurs. Over S repeats,

$$Z_j = \sum_s \mathbf{1}[n_{j,s} = 0], \qquad H_j = \min_s h_{j,s}, \qquad N_j = \sum_s n_{j,s}. \tag{6}$$

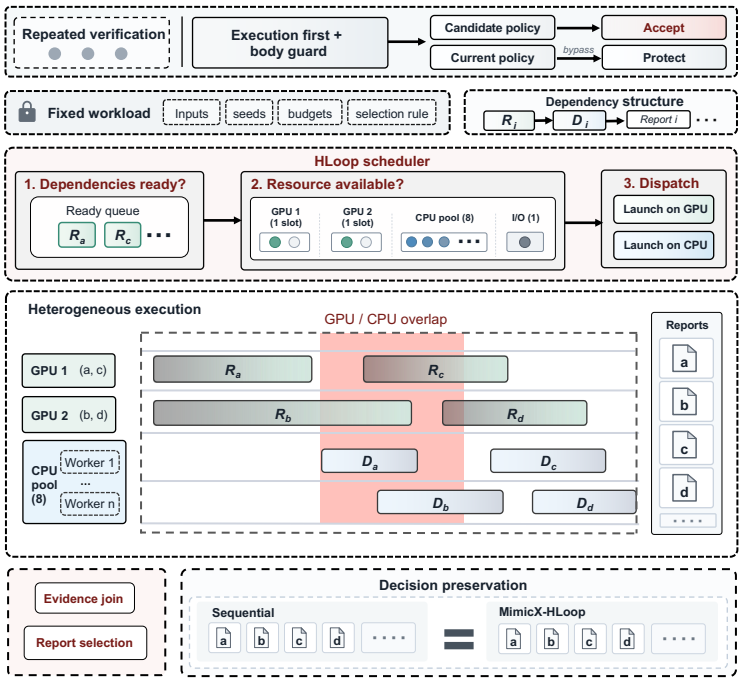


Figure 3: Verified updates with heterogeneous execution. Repeated rollouts rank candidates by completion, reliable motion length, and tracking quality. A body-error guard protects the current policy when no candidate qualifies. The same verification workload can run sequentially or through MimicX-HLoop. Its scheduler dispatches ready rollouts to GPUs and completed trajectories to CPU diagnosis workers, overlapping independent work while respecting dependencies. Reports are joined before selection; inputs, seeds, budgets, and the selection rule remain fixed. Execution order changes feedback latency while preserving the complete evidence supplied to the gate and its policy decision across all repeated evaluation seeds.

The gate ranks complete candidates by

$$K_j = (Z_j, H_j, -N_j, -b_j, -z_j, \tilde{r}_j), \tag{7}$$

where b_j and z_j are medians across repeats of the last recorded maximum body-position and ankle-height errors, and $\tilde{r}_j$ is median rollout reward. A candidate must strictly improve K_j and satisfy $b_j \le (1+\epsilon)b_k$ with $\epsilon = 0.20$. The first passing candidate becomes θ_{k+1}; if none passes, the accepted policy is retained. Completion and reliable motion length take priority; tracking fidelity and reward break ties. Algorithm 1 summarizes refinement; Appendix B.2 proves its selection properties.

Heterogeneous feedback execution. MimicX-HLoop represents rollout, diagnosis, and selection as a dependency graph, allowing completed rollouts to be diagnosed while other simulations continue. Selection begins once all required evaluations are available, preserving the same inputs as sequential execution (Figure 3). With fixed node outputs and deterministic report ordering, dependency-respecting schedules select the same policy. For resource r, let W_r denote total work and m_r its parallel slots. With critical-path duration L_{crit}, the scheduling bound is

$$T_{\mathrm{sched}} \ge \max\left\{ L_{\mathrm{crit}}, \max_r W_r/m_r \right\}. \tag{8}$$

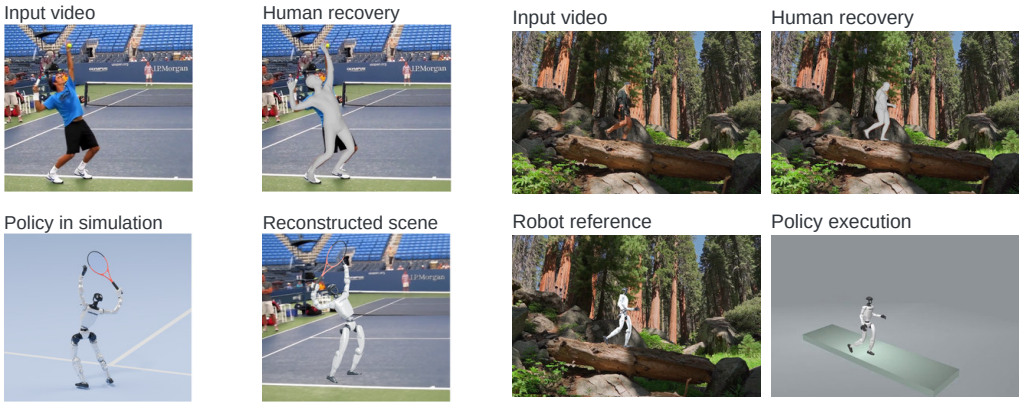
Appendix B.4 derives this bound and establishes scheduling invariance.

4 Experimental Setup

We evaluate whether execution feedback improves tracking, how the supervision components contribute, and whether verification and heterogeneous execution make refinement reliable and efficient.

Tasks and learner. The core tasks are Tennis Swing, Football Juggling, Dance Sequence, and Kung Fu Sequence on a 29-DoF Unitree G1 in MuJoCo/MjLab. Simulation runs at 200 Hz and the PPO policy at 50 Hz. Each task supplies one shared warm-start checkpoint. Four methods and three continuation seeds yield 48 task–method–seed trials; three evaluation repeats per final policy yield 144 rollouts. Architecture and optimization settings appear in Appendix C.

Compared supervision. *Fixed Reference* continues PPO with the original reference and objective. *Failure Curriculum* concentrates half of its resets around a diagnosed failure window. *Task-Aware Refinement* also adjusts local, global, and dynamic tracking terms. Each uses 250 continuation iterations. Full MimicX evaluates three automatically proposed 250-iteration continuations and selects the final policy through repeated verification. Tennis and Football share the original references across



(a) Tennis: recovered scene context (b) Forest: collision-scene execution

Figure 4: From video observations to humanoid execution. Tennis shows the input, recovered human motion, simulated policy rollout, and the corresponding policy pose in the reconstructed scene at a recovered source phase. Forest shows the input, recovered human motion, retargeted robot reference, and policy execution in the task-equivalent collision scene, sampled at each clip's midpoint.

Table 1: MimicX improves tracking and execution horizon on all four tasks. Means are over three continuation seeds, each evaluated with three fixed rollout seeds. Body mean and p95 summarize aligned worst-body error over time; horizon is the earliest failure across repeats, with $T + 1$ assigned on success. Reduction and gain are relative to Fixed Reference, computed before rounding. Bold marks improvements.

Task	Body mean (m)			Body p95 (m)		Horizon (steps)			Success (%)	
	Fixed	MimicX	*Red. (%)*	Fixed	MimicX	Fixed	MimicX	*Gain (%)*	Fixed	MimicX
Tennis	0.241	**0.157**	**35.0**	0.518	**0.278**	322.0	**801.0**	**+148.8**	11.1	**100.0**
Football	0.286	**0.244**	**14.7**	0.612	**0.504**	53.7	**427.3**	**+696.3**	0.0	0.0
Dance	0.270	**0.244**	**9.9**	0.502	**0.467**	193.3	**341.3**	**+76.6**	0.0	0.0
Kung Fu	0.249	**0.141**	**43.3**	0.634	**0.296**	97.7	**196.0**	**+100.7**	0.0	0.0

methods; Dance and Kung Fu use prepared repaired references for MimicX. Errors are measured against each method's reference. This comparison evaluates successive supervision configurations; Appendix C specifies their profiles, references, and transition budgets.

Execution and fidelity. Evaluation uses deterministic inference, full-start initialization, and fixed termination tests. Full-horizon success requires zero flagged terminations over the evaluation schedule. Robust Execution Horizon is the earliest failure across repeats, with $T + 1$ assigned on success; the reference restarts at motion end. Body error is the temporal mean of $\max_{b \in \mathcal{B}} \|p_{t,b} - \widetilde{p}_{t,b}\|_2$ across recorded rollout steps. We also measure tail body error, anchor error, body velocities, and ankle-height error. Task-macro effects average relative changes after within-task continuation-seed averaging. Appendix C details additional-video, supplied-motion, direct-policy, and timing protocols.

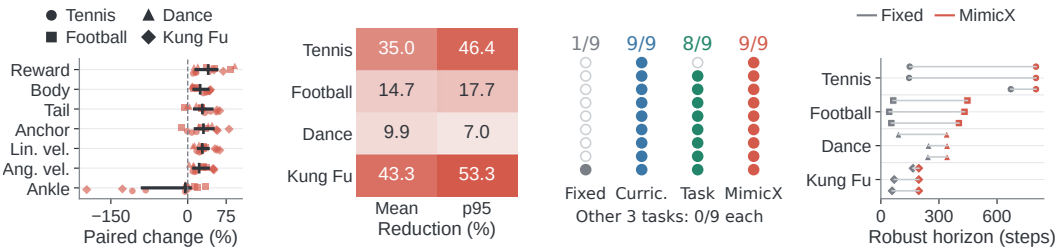
Collision-scene extension. Five additional video-derived motions use task-equivalent fixed collision scenes. Each compares equal 3,000-update continuations from a shared 3,000-update checkpoint, using one training seed and three evaluation seeds (Appendix F.3).

5 Results

5.1 Tracking Quality

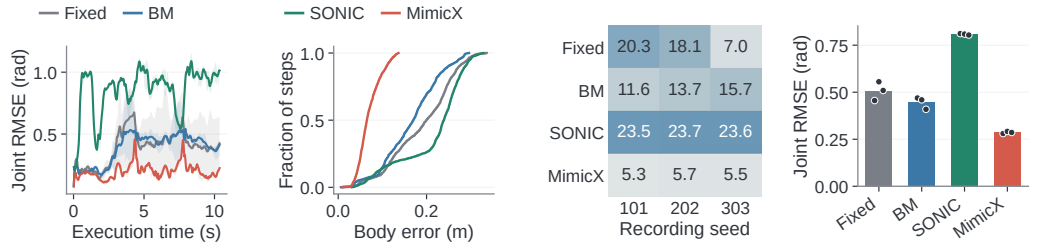
MimicX improves body-tracking error and Robust Execution Horizon in all twelve task-matched comparisons against Fixed Reference (Figure 5, Table 1). Task-level averaging yields a 25.7% reduction in body error and a 255.6% increase in horizon. Tennis reaches full-horizon success in all nine evaluations. These results show that execution-conditioned supervision improves both motion fidelity and progress through demanding transitions.

The paired results show that the gains extend across error channels (Table 11). Both linear- and angular-velocity errors improve in all twelve comparisons, while tail body error and root/anchor error improve in ten. These gains extend from articulation to global motion and dynamics.



(a) Paired effects (b) Mean and tail gains (c) Tennis completion (d) Execution horizon

Figure 5: Controlled evidence across four video-derived skills. (a) All twelve task-matched trials across seven reward and tracking channels. (b) Task-wise reductions in mean and p95 tracking error, in percent. (c) Full-horizon completion across repeated evaluations. (d) Worst-case execution horizon. Black marks in (a) denote median and interquartile range. Fixed, Curriculum, and Task-aware abbreviate Fixed Reference, Failure Curriculum, and Task-Aware Refinement; gray, blue, green, and coral identify the four methods (Curric. and Task in (c)). Tail denotes body-error p95 and Ankle denotes ankle-height error.



(a) Tennis: joint error (b) Football body (c) Tennis body (d) Football joints

Figure 6: More accurate policies for the same demonstrated motion. Common-reference evaluation of Fixed Reference, BeyondMimic's MjLab reproduction (BM), released SONIC, and MimicX. (a) Tennis joint-error mean and range over time; (b) Football control-step body-error distribution; (c) Tennis body error in cm for each recording; (d) mean Football joint errors with every recording. Lower errors are better. All methods are evaluated over the same 518-step Tennis and 454-step Football intervals without failure resets; body error uses root-local G1 forward kinematics. Full protocols and task outcomes appear in Table 16.

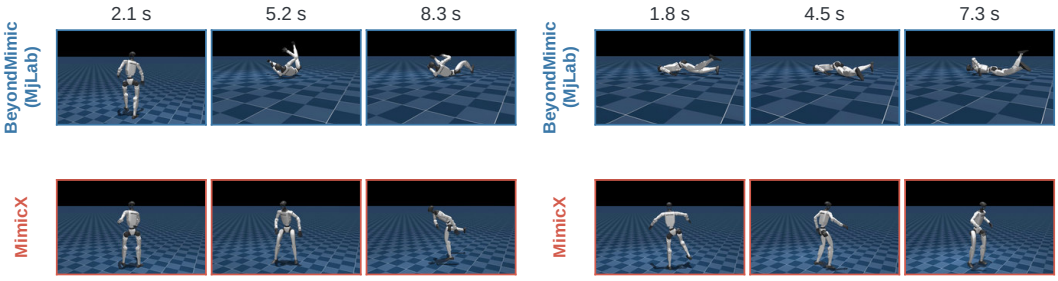
Figure 6 directly compares the standalone MjLab reproduction of BeyondMimic (Liao et al., 2025), released SONIC (Luo et al., 2026), Fixed Reference, and MimicX on identical Tennis and Football references. Relative to BeyondMimic, MimicX reduces joint RMSE and root-local body error on both tasks, with improvements in every paired continuation seed. All recordings cover one complete source interval without failure resets; the common metric and adaptation protocols appear in Appendix F.1; Table 2 reports the Tennis values.

Table 2: Direct policy comparison on the same Tennis Swing reference. MimicX improves both common articulation metrics across all three seeds relative to BeyondMimic's MjLab reproduction. Bold marks minima. Reductions use Fixed Reference as the baseline.

Method	Policy setup	Joint RMSE (rad, ↓)	Reduction (%)	Local body error (m, ↓)	Reduction (%)
Fixed Reference	Fixed objective	0.389 ± 0.166	+0.0	0.151 ± 0.072	+0.0
BeyondMimic (MjLab)	Fixed objective	0.389 ± 0.027	+0.1	0.137 ± 0.021	+9.8
SONIC (released)	Pretrained	0.857 ± 0.018	-120.2	0.236 ± 0.001	-55.8
MimicX	Closed loop	**0.201** ± 0.002	+48.3	**0.055** ± 0.002	+63.4

Mean ± sample SD over three continuation seeds (101/202/303, evaluated with seed 1001) or three SONIC executions. All methods follow the same Tennis reference for 518 steps at 50 Hz without failure resets. Joint error is the temporal mean of per-frame joint RMSE. Local body error averages fourteen-body distances under a common G1 forward-kinematic model with root pose fixed to identity. BeyondMimic uses its standalone MjLab reproduction, continued for 250 PPO updates with 1,024 environments and configured learning rate 10^{-6} from the shared warmstart. SONIC uses its released controller with a five-second settling phase and target-log-synchronized playback; simulation-clock parity is checked and video is rendered offline. MimicX uses its frozen gate-selected policies; candidate-search budgets are specified in Section C.

Figure 7 compares actual BeyondMimic and MimicX recordings at matched source phases: refinement sustains the swing and alternating leg motions after the standalone tracker falls. Figure 4 connects human reconstruction and robot execution across Tennis and Forest Traversal.



(a) Tennis Swing (b) Football Juggling

Figure 7: Execution differences on the same motion timeline. Each case compares BeyondMimic's MjLab reproduction (blue) and MimicX (coral) at 20/50/80% of the source interval. Recorded policies use continuation seed 202 and evaluation seed 1001, with matched camera and crop settings within each task.

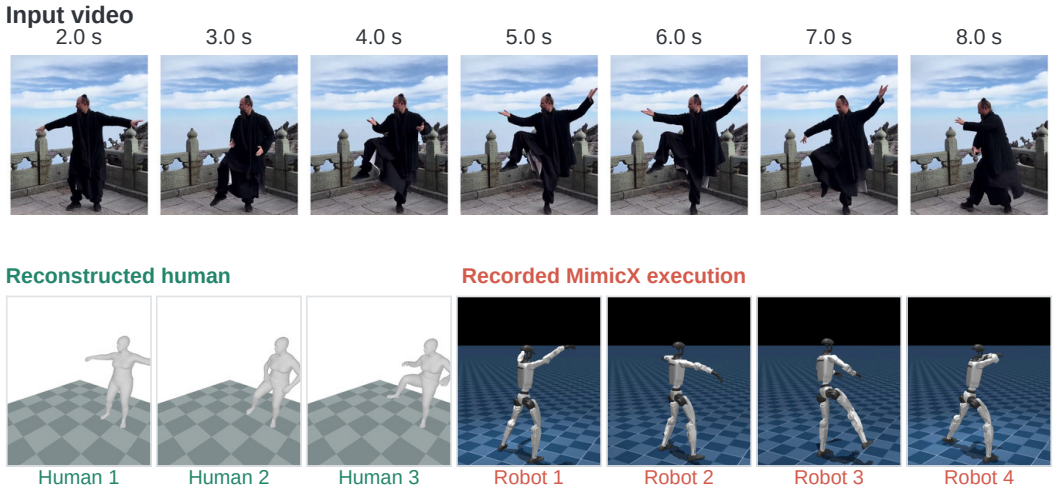


Figure 8: Kung Fu from video to recorded execution. Seven input frames show the selected motion interval. The lower row progresses from recovered human motion to recorded MimicX policy states. Source and rollout timestamps identify the sampled interval; the complete-task evaluation is reported in Table 1.

5.2 Refinement and Verification

Concentrating practice around difficult transitions yields a strong first improvement. Failure Curriculum brings Tennis to full-horizon execution and achieves the lowest task-macro velocity errors (Table 3). Adding task-aware supervision strengthens global consistency: Task-Aware Refinement attains the lowest anchor error and highest normalized horizon. The complete loop attains the lowest mean body error while matching the best task-macro success. The component results thus identify distinct benefits from allocating practice, coordinating tracking objectives, and selecting the retained policy. The task-critical Kung Fu sequence in Figure 8 illustrates whole-body coordination.

Figure 9 shows how reward, body error, and episode length evolve as policies adapt to revised supervision. The component table and learning curves distinguish the effects of practice allocation and tracking objectives; Section 4 specifies continuation and candidate-search budgets.

The gate accepts six updates and protects six current policies across the twelve automated loops. All Tennis and Football trials accept a refinement; Dance and Kung Fu retain the incumbent under the configured execution ordering and body-error guard (Appendix E). The Kung Fu case in Table 5 illustrates the role of execution verification: both candidates improve reward but shorten the worst-case execution horizon, so the current policy is retained. This separates the signal used to train a candidate from the execution criterion used to adopt it. The gate therefore allows objective adaptation while retaining a stronger verified policy when proposals regress. Figure 14 combines execution-first selection, feedback efficiency and time-resolved tracking gains. Appendix Figure 11 resolves orientation, acceleration and angular velocity within a rollout for temporal diagnosis.

Table 3: Supervision components improve complementary tracking channels. Values are unweighted means across four tasks after averaging three continuation seeds within each task. Bold marks column bests.

Method	Supervision			Execution		Tracking		Dynamics	
	C	T	V	Success (%, ↑)	Horizon (%, ↑)	Body (m, ↓)	Anchor (m, ↓)	Linear vel. (m/s, ↓)	Angular vel. (rad/s, ↓)
Fixed Reference	–	–	–	2.8	21.5	0.262	0.219	0.691	2.955
Failure Curriculum	✓	–	–	**25.0**	62.7	0.215	0.135	**0.451**	**1.935**
Task-Aware Refinement	✓	✓	–	22.2	**70.2**	0.222	**0.111**	0.485	2.393
MimicX	✓	✓	✓	**25.0**	64.0	**0.196**	0.147	0.481	2.238

Horizon is normalized by task length. C: failure-window curriculum; T: task-aware tracking; V: repeated verification.

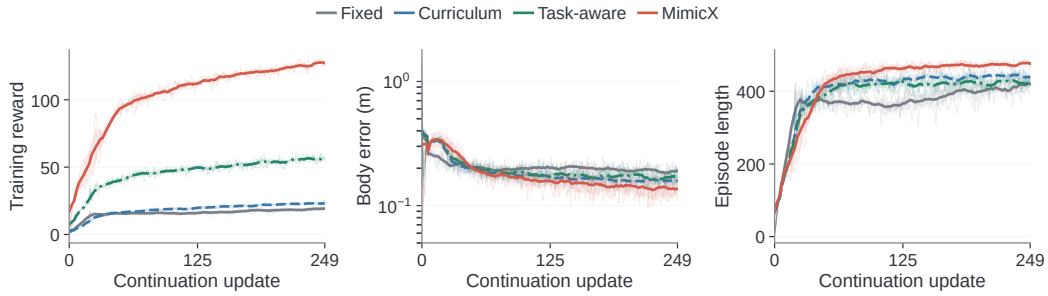


Figure 9: Coordinated supervision changes the learning trajectory. Reward, body error, and episode length across twelve 250-update Tennis continuations. Thin traces retain raw values; emphasized curves use an 11-update moving mean, with sample standard deviation across seeds. A single legend applies to all three panels.

Table 4: Repeated verification. Decisions for four tasks and three continuation seeds yield six accepted refinements and six protected incumbent policies.

Task	Seed 101	Seed 202	Seed 303
Tennis	Accept^{B,K}	Accept^{B,K}	Accept^{B,K}
Football	Accept^{B,K}	Accept^{B,K}	Accept^{B,K}
Dance	Protect^{B}	Protect^{B}	Protect^{B}
Kung Fu	Protect^{B,K}	Protect^{B}	Protect^{B,K}

Accept selects the refined policy; Protect retains the current policy. B: body guard; K: execution ordering. Full protocols appear in Appendix E.

Table 5: Execution overrides reward. Two Kung Fu proposals improve training reward but shorten the repeated-evaluation horizon; both are rejected.

Metric	Current	Proposal 1	Proposal 2
Reward	0.5707	0.6275	0.5838
Δ reward	+0.0000	+0.0568	+0.0131
Horizon	**879**	677	689
Gate	Retain	Reject	Reject

Historical case: horizon is the worst repeated-evaluation failure step. Proposals lose 202/190 steps; reward changes use the current policy.

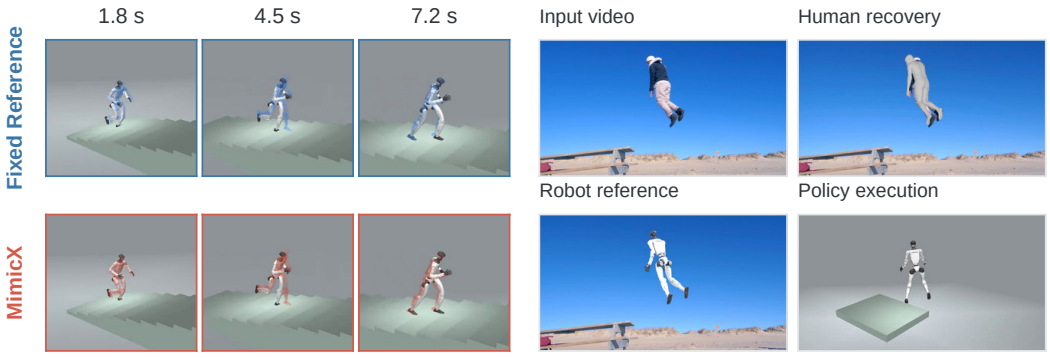
5.3 Task Breadth and Efficiency

Four additional videos cover Basketball, Soccer, Michael Jackson Dance, and Uniandes Dance, with tracking improvements on the latter three and a full-horizon success gain on Soccer. Fourteen supplied-motion cases evaluate the reference-to-policy interface across LaFAN1 and AMASS/ACCAD motions, with reference-averaged body-error reductions of 15.81% and 19.84%, respectively (Table 18). Appendix F reports every task and its protocol. In five collision-scene video tasks, matched-budget refinement raises strict completion from 9/15 to 12/15 rollouts (Table 6). Forest Traversal improves from consistent early failure to 3/3 complete executions. Both body and root tracking improve across all five tasks, extending the gains to scenes (Figure 10). The scene and adapted reference remain fixed within each paired comparison. Diagnosis reallocates resets to 50% true-start initialization, 25% failure-window replay, and 25% full-motion sampling, while revising global, foot, and event-aware tracking objectives. The resulting gains come from changing how the same motion is practiced and tracked in the same environment.

MimicX-HLoop reduces median rollout–diagnosis–selection time from 241.2 s to 124.0 s, a 1.95× speedup over sequential execution (Tables 7 and 8). Selection outcomes remain identical across all paired repetitions. The benchmark holds checkpoints and workload fixed, measuring how overlapping GPU simulation with CPU diagnosis shortens the feedback path. Appendix G details the matched workload and selection-parity analysis. Bulk-synchronous execution has a median of 122.26 s; both parallel modes substantially shorten feedback relative to sequential rollout–diagnosis–selection.

Table 6: Policy-in-the-loop refinement on five collision-scene video tasks. Equal-budget continuations share the same initial policy, scene, and reference. Completion rises from 9/15 to 12/15 full executions; both p95 errors improve on all five tasks. Steps report the minimum valid prefix divided by target length; reductions compare seed-mean errors over matched prefixes before failure.

	Completion		Valid steps / T		p95 reduction	
Task	**Fixed**	**MimicX**	**Fixed**	**MimicX**	**Body**	**Root**
Track Running	3/3	3/3	97/97	97/97	53.80%	24.95%
Stair Ascent	3/3	3/3	449/449	449/449	11.86%	7.88%
Forest Traversal	0/3	**3/3**	39/404	404/404	79.15%	**89.04%**
Platform Jump	3/3	3/3	58/58	58/58	5.78%	50.85%
Parkour	0/3	0/3	220/255	220/255	2.04%	0.59%



(a) Stair Ascent: matched policy states

(b) Platform Jump: source-to-policy stages

Figure 10: Video-driven skills in collision scenes. Stair Ascent compares Fixed Reference and refinement at 20/50/80% of the 449-step schedule, using evaluation seed 2002 and time-specific close crops shared by both methods. Both complete all three evaluations; refinement reduces body and root prefix p95 errors by 11.86% and 7.88%. Platform Jump shows each clip's midpoint: input, human recovery, robot reference, and policy execution.

Dependency-driven execution releases each completed rollout to CPU diagnosis without waiting for the remaining simulations. Reports are joined before selection, so the gate receives the same evidence despite the changed execution order. This decouples the computational schedule from the policy-selection rule. Across five repetitions, MimicX-HLoop takes 119.92–128.88 s, compared with 238.57–242.40 s for sequential execution of the same rollout–diagnosis–selection workload with identical policy checkpoints.

Table 7: Feedback executor comparison. Five matched repetitions; mean, median, and speedup summarize the same rollout-feedback workload.

Statistic	Sequential	Bulk-sync.	HLoop
Mean (s)	240.748	**123.423**	123.896
SD (s)	1.608	2.697	3.596
Median (s)	241.221	**122.258**	124.017
Speedup	1.000×	**1.973**×	1.945×
Selection	Same	Same	Same

Executors preserve selection on a workload with zero policy updates. Speedup is the ratio of medians to sequential time; HLoop is 0.986× as fast as bulk-sync. This measures feedback scheduling, not policy training.

Table 8: Paired feedback timings. Per-run wall time in seconds and ratio to HLoop; bold marks the fastest executor in each repetition.

Run	Seq. (s)	Bulk (s)	HLoop (s)	Seq. / HLoop	Bulk / HLoop
1	242.403	**121.813**	124.017	1.955	0.982
2	239.627	126.440	**121.040**	1.980	1.045
3	238.569	120.474	**119.924**	1.989	1.005
4	241.221	126.131	**125.624**	1.920	1.004
5	241.921	**122.258**	128.876	1.877	0.949

Each ratio divides an executor time by the HLoop time within each run; HLoop / HLoop is 1.000 in all five runs. These repetitions are summarized in Table 7. Headline speedup uses medians, not mean paired ratios.

6 Conclusion

MimicX turns humanoid execution failures into actionable supervision. Temporal and body-specific diagnosis coordinates tracking objectives and practice, while repeated verification selects policy updates. Controlled comparisons demonstrate improved tracking and reliable motion length; same-motion baselines and additional video, motion-reference, and collision-scene tasks establish breadth. MimicX-HLoop accelerates feedback while preserving selection outcomes. Together, these components couple video-derived references with policy learning through execution-guided refinement.

AI USE STATEMENT

For manuscript preparation, AI tools assisted with drafting and language polishing. The authors reviewed and approved the final text.

REFERENCES

Joao Pedro Araujo, Yanjie Ze, Pei Xu, Jiajun Wu, and C. Karen Liu. Retargeting matters: General motion retargeting for humanoid motion tracking. *arXiv preprint arXiv:2510.02252*, 2025.

Zhanxiang Cao, Liyun Yan, Yang Zhang, Sirui Chen, Jianming Ma, Tianyue Zhan, Shengcheng Fu, Yufei Jia, Cewu Lu, and Yue Gao. HiWET: Hierarchical world-frame end-effector tracking for long-horizon humanoid loco-manipulation. In *Proceedings of Robotics: Science and Systems*, Sydney, Australia, July 2026. doi: 10.15607/RSS.2026.XXII.030.

Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1568–1577, 2018.

Félix G. Harvey, Mike Yurick, Derek Nowrouzezahrai, and Christopher Pal. Robust motion in-betweening. *ACM Transactions on Graphics*, 39(4), 2020. doi: 10.1145/3386569.3392480.

Tairan He, Jiawei Gao, Wenli Xiao, Yuanhang Zhang, Zi Wang, Jiashun Wang, Zhengyi Luo, Guanqi He, Nikhil Sobanbabu, Chaoyi Pan, Zeji Yi, Guannan Qu, Kris Kitani, Jessica K. Hodgins, Linxi Fan, Yuke Zhu, Changliu Liu, and Guanya Shi. ASAP: Aligning simulation and real-world physics for learning agile humanoid whole-body skills. In *Proceedings of Robotics: Science and Systems*, 2025. doi: 10.15607/RSS.2025.XXI.066.

Tao Huang, Huayi Wang, Junli Ren, Kangning Yin, Zirui Wang, Xiao Chen, Feiyu Jia, Wentao Zhang, Junfeng Long, Jingbo Wang, and Jiangmiao Pang. Towards adaptable humanoid control via adaptive motion tracking. *arXiv preprint arXiv:2510.14454*, 2025.

Qiayuan Liao, Takara E. Truong, Xiaoyu Huang, Yuman Gao, Guy Tevet, Koushil Sreenath, and C. Karen Liu. BeyondMimic: From motion tracking to versatile humanoid control via guided diffusion. *arXiv preprint arXiv:2508.08241*, 2025.

Zhengyi Luo, Jinkun Cao, Alexander Winkler, Kris Kitani, and Weipeng Xu. Perpetual humanoid control for real-time simulated avatars. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10895–10904, 2023.

Zhengyi Luo, Ye Yuan, Tingwu Wang, Chenran Li, Fernando Castañeda, Sirui Chen, Zi-Ang Cao, Jiefeng Li, David Minor, Qingwei Ben, Jinhyung Park, David Sami, Zi Wang, Xingye Da, Runyu Ding, Cyrus Hogg, Lina Song, Edy Lim, Eugene Jeong, Tairan He, Haoru Xue, Wenli Xiao, Simon Yuen, Jan Kautz, Yan Chang, Umar Iqbal, Linxi "Jim" Fan, and Yuke Zhu. SONIC: Supersizing motion tracking for natural humanoid whole-body control. *Science Robotics*, 11(117):eaed4592, 2026. doi: 10.1126/scirobotics.aed4592.

Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5442–5451, 2019.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10975–10985, 2019.

Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel van de Panne. DeepMimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions on Graphics*, 37(4):143:1–143:14, 2018. doi: 10.1145/3197517.3201311.

Xue Bin Peng, Ze Ma, Pieter Abbeel, Sergey Levine, and Angjoo Kanazawa. AMP: Adversarial motion priors for stylized physics-based character control. *ACM Transactions on Graphics*, 40(4): 144:1–144:20, 2021. doi: 10.1145/3450626.3459670.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. World-grounded human motion recovery via gravity-view coordinates. In *SIGGRAPH Asia 2024 Conference Papers*, 2024. doi: 10.1145/3680528.3687565.

Richard S. Sutton, David A. McAllester, Satinder P. Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, volume 12, 1999.

Chen Tessler, Yunrong Guo, Ofir Nabati, Gal Chechik, and Xue Bin Peng. MaskedMimic: Unified physics-based character control through masked motion inpainting. *ACM Transactions on Graphics*, 43(6), 2024. doi: 10.1145/3687951.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033, 2012. doi: 10.1109/IROS.2012.6386109.

Jiashun Wang, M. Eva Mungai, He Li, Jean Pierre Sleiman, Jessica K. Hodgins, and Farbod Farshidian. Generalizing from references using a multi-task reference and goal-driven RL framework. In *Proceedings of Robotics: Science and Systems*, Sydney, Australia, July 2026a. doi: 10.15607/RSS.2026.XXII.026.

Yunshen Wang, Shaohang Zhu, Peiyuan Zhi, Yuhan Li, Jiaxin Li, Yong-Lu Li, Yuchen Xiao, Xingxing Wang, Baoxiong Jia, and Siyuan Huang. OmniXtreme: Breaking the generality barrier in high-dynamic humanoid control. In *Proceedings of Robotics: Science and Systems*, Sydney, Australia, July 2026b. doi: 10.15607/RSS.2026.XXII.031.

Zhen Wu, Xiaoyu Huang, Lujie Yang, Yuanhang Zhang, Xi Chen, Pieter Abbeel, Rocky Duan, Angjoo Kanazawa, Carmelo Sferrazza, Guanya Shi, and C. Karen Liu. Perceptive humanoid parkour: Chaining dynamic human skills via motion matching. In *Proceedings of Robotics: Science and Systems*, Sydney, Australia, July 2026. doi: 10.15607/RSS.2026.XXII.020.

Weiji Xie, Jinrui Han, Jiakun Zheng, Huanyu Li, Xinzhe Liu, Jiyuan Shi, Weinan Zhang, Chenjia Bai, and Xuelong Li. KungfuBot: Physics-based humanoid whole-body control for learning highly-dynamic skills. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2089.

Kevin Zakka, Qiayuan Liao, Brent Yi, Louis Le Lay, Koushil Sreenath, and Pieter Abbeel. mjlab: A lightweight framework for GPU-accelerated robot learning. *arXiv preprint arXiv:2601.22074*, 2026.

Qiang Zhang, Jiahao Ma, Peiran Liu, Shuai Shi, Zeran Su, Zifan Wang, Jingkai Sun, Wei Cui, Jialin Yu, Gang Han, Wen Zhao, Pihai Sun, Kangning Yin, Jiaxu Wang, Jiahang Cao, Lingfeng Zhang, Hao Cheng, Xiaoshuai Hao, Yiding Ji, Junwei Liang, Jian Tang, Renjing Xu, and Yijie Guo. MeshMimic: Geometry-aware humanoid motion learning through 3D scene reconstruction. *arXiv preprint arXiv:2602.15733*, 2026.

Algorithm 1 MimicX closed-loop refinement

Require: Registered R, accepted θ_0, fixed evaluation seeds and budgets
$\theta \leftarrow \theta_0$
for each refinement iteration **do**
  $E_0 \leftarrow \text{Evaluate}(\theta, R)$; $D \leftarrow \text{Diagnose}(E_0[1])$
  for each validated $U_j \in \text{Propose}(D)$ **do**
    $\theta_j \leftarrow \text{PPOContinue}(\theta, R, U_j)$
    $E_j \leftarrow \text{Evaluate}(\theta_j, R)$ on all fixed seeds
  end for
  Rank complete candidates by Eq. 7 and apply the guard
  Accept the first passing candidate, or protect θ and stop
  Stop if every repeat of the accepted policy completes
end for
return accepted policy θ

A Notation and Algorithmic Details

MimicX uses policy execution to determine how a tracking policy should be supervised. A video-derived robot reference supplies an initial target; policy rollouts identify where that target is difficult to follow; structured proposals alter the training objective and distribution of reference starts; and repeated execution determines which policy to retain. Figure 2 details the failure-conditioned feedback path.

A.1 Reference and Policy Interface

Given a single-person monocular video, we reconstruct world-grounded human motion with GVHMR (Shen et al., 2024), represented using SMPL-X (Pavlakos et al., 2019), and retarget it to the robot with GMR (Araujo et al., 2025). The motion adapter produces

$$R = \{\bar{q}_t, \dot{\bar{q}}_t, (\bar{p}_{t,b}, \bar{Q}_{t,b}, \bar{v}_{t,b}, \bar{\omega}_{t,b})_{b\in\mathcal{B}}\}_{t=0}^{T_R-1}, \tag{9}$$

where q denotes actuated joint angles, p body position, Q orientation, and v, ω linear and angular velocity. Bars indicate reference quantities; the adapter also records the motion frame rate and enforces the robot's joint ordering and quaternion convention. The reference is a kinematic target, while the policy drives the simulated robot through joint-position commands.

The tracking backend follows the whole-body tracking formulation of BeyondMimic (Liao et al., 2025), implemented in MuJoCo/MjLab (Todorov et al., 2012; Zakka et al., 2026). We train $\pi_\theta(a_t \mid o_t)$ with PPO (Schulman et al., 2017), where the observation contains the motion command, reference-anchor pose in the robot frame, base velocities, joint state, and previous action. The outer loop takes a current checkpoint θ_k, reference R, and a fixed verification specification. It creates candidate supervision $U_{k,j} = (W_{k,j}, C_{k,j})$, comprising reward parameters and a reset curriculum, and resumes training from θ_k. Selection links candidate checkpoints, patches, and execution records.

The reference interface also accepts registered repaired motions. In the reported automated loop, the selected motion file is fixed across the current policy and its candidates; reference preparation precedes this search. Consequently, geometric repair supplies supervision to the loop, while policy-conditioned objective and curriculum proposals are generated within it. Section C identifies the tasks that use repaired references.

A.2 Execution-Conditioned Localization

Each verification rollout starts at the beginning of the motion and records termination flags, reward, body-position and height discrepancies, anchor errors, and velocity errors at every control step. For diagnosis, we use the first scheduled evaluation repeat. Let f be its first flagged termination; when no termination occurs, choose the step with the largest tracked-body position error. The diagnostic interval is

$$\mathcal{W}_k = [\max(1, f - 40), \min(T, f + 40)], \tag{10}$$

where T is the evaluation horizon. This interval provides both a local description of the difficult transition and candidate initialization states from which to practice it.

The miner summarizes each recorded diagnostic channel by its maximum and mean inside $\mathcal{W}_k$, sorts these pairs in descending lexicographic order, and retains eight channels. Body-specific channel names identify the affected regions. Wrist, elbow, and shoulder discrepancies suggest upper-body

precision; hip, knee, and ankle discrepancies suggest lower-body position and height; pelvis, torso, and anchor channels suggest global stabilization; velocity channels suggest dynamics and smoothness. These are deterministic mappings from measured discrepancies to reward families. Task-priority body lists and diagnosed windows supply fixed metadata for component comparisons.

A.3 Coordinated Tracking Supervision

Separating articulation from global motion. Let a denote the torso anchor. To measure local articulation without counting horizontal drift twice, align reference bodies with the robot's anchor translation and heading. Define ΔQ_t as the yaw component of $Q_{t,a}\bar{Q}_{t,a}^{-1}$ and $d_t = (p_{t,a}^x, p_{t,a}^y, \bar{p}_{t,a}^z)$. The aligned targets are

$$\widetilde{p}_{t,b} = d_t + \Delta Q_t(\bar{p}_{t,b} - \bar{p}_{t,a}), \qquad \widetilde{Q}_{t,b} = \Delta Q_t \bar{Q}_{t,b}. \tag{11}$$

Absolute anchor tracking is supervised separately, preserving sensitivity to global displacement and orientation even when local articulation is accurate.

Local, global, and dynamics terms. For a body group G, define the positional term

$$\phi_p(G, \sigma) = \exp\left[-\frac{1}{|G|\sigma^2}\sum_{b\in G}\|p_{t,b} - \widetilde{p}_{t,b}\|_2^2\right]. \tag{12}$$

Height tracking replaces the squared norm by its vertical component; orientation tracking uses squared quaternion angular distance. Velocity terms compare robot and reference velocities in world coordinates. The objective combines these terms as

$$r_t(W) = \sum_{\ell\in\mathcal{L}} w_\ell \phi_\ell + \sum_{g\in\mathcal{G}} w_g \phi_g + \sum_{d\in\mathcal{D}} w_d \phi_d - \lambda_a\|a_t - a_{t-1}\|_2^2 - \mathcal{P}_t, \tag{13}$$

where $\mathcal{L}$ contains selected limb position and height terms, $\mathcal{G}$ contains absolute anchor and core/body pose terms, and $\mathcal{D}$ contains linear/angular velocity and reference-conditioned stance terms. $\mathcal{P}_t$ collects joint-limit and self-collision penalties. The full candidate profile includes a stance-slip reward: when a reference ankle lies below a prescribed height, it penalizes the robot foot's squared horizontal velocity plus a weighted vertical component. Thus contact-related supervision is grounded in reference ankle height and physical foot motion.

Proposals coordinate these groups instead of substituting a single local reward for the tracking objective. Every generated proposal strengthens the absolute anchor and core-position terms, retains action smoothing, and adds position, height, or velocity rewards according to the diagnostic families. For example, lower-body emphasis adds position and height precision together with linear-velocity tracking; a dynamics proposal additionally supervises core angular velocity. The full profile retains its existing wrist, core, and stance terms when these diagnostic additions are applied.

A.4 Bounded Proposals and Policy Continuation

The generator emits a finite set of typed patches over recognized reward terms and tracked bodies. Nonnegative reward weights lie in $[0, 10]$, exponential scales in $[0.01, 5]$, and the action-rate coefficient in $[-1, 0]$. A proposed replay interval must lie within the horizon and overlap the diagnosed window; its probability lies in $[0.2, 0.8]$. Proposal generation leaves evaluation seeds, termination criteria, and acceptance rules fixed, keeping training adaptation separate from verification.

The heuristic generator produces three candidates: window replay with probability 0.45, stronger replay with probability 0.65 and added dynamics emphasis, and full-start consolidation. For a replay probability ρ, the initial reference frame follows

$$t_0 \sim (1-\rho)\delta_0 + \rho\ \mathrm{Unif}\{s, s+1, \ldots, e-1\}, \tag{14}$$

with $[s, e)$ taken from the diagnostic interval and clipped to the available motion. Consolidation uses $\rho = 0$. Replay therefore means reference-state initialization followed by new on-policy simulation, not reuse of an offline transition buffer. The mixture preserves whole-sequence practice while increasing exposure to the difficult transition.

All candidates in an iteration start from the same accepted checkpoint and use the same training seed and continuation budget. PPO updates the policy under each candidate's supervision; its optimization rule is unchanged. Checkpoint resumption carries the learned policy and optimizer state forward under the configured continuation learning rate. The resulting policy is evaluated under the common strict tracking task, without its candidate-specific reward patch. This makes evaluation reward comparable within a fixed-reference comparison despite differences in training rewards.

Algorithm 2 MimicX policy-in-the-loop supervision refinement

Require: Registered reference R, current checkpoint θ_0, fixed evaluation seeds, iteration and candidate budgets

$\theta \leftarrow \theta_0$
for each permitted refinement iteration **do**
  $E_0 \leftarrow \text{Evaluate}(\theta, R)$ on all fixed seeds
  $D \leftarrow \text{Diagnose}(E_0[1])$
  $\mathcal{U} \leftarrow \text{Validate}(\text{Propose}(D))$
  for each budgeted supervision patch $U_j \in \mathcal{U}$ **do**
    $\theta_j \leftarrow \text{PPOContinue}(\theta, R, U_j)$
    $E_j \leftarrow \text{Evaluate}(\theta_j, R)$ on all fixed seeds
  end for
  Rank candidates using Equation 16
  $j^* \leftarrow$ first fully evaluated candidate with $K_j > K_0$ satisfying Equation 17
  if such a candidate exists **then**
    $\theta \leftarrow \theta_{j^*}$; record acceptance and checkpoint hash
  else
    Record current-policy protection; **break**
  end if
  if all selected-policy repeats complete without termination **then**
    break
  end if
end for
return θ and the proposal/evaluation/selection record

A.5 Repeated Execution and Protected Selection

For candidate j and evaluation seed s, let $n_{j,s}$ count termination events and $h_{j,s}$ be the first terminated step, with $h_{j,s} = T + 1$ when no termination is flagged within the evaluation budget. Over S repeats, define

$$Z_j = \sum_{s=1}^{S} \mathbf{1}[n_{j,s} = 0], \qquad H_j = \min_s h_{j,s}, \qquad N_j = \sum_s n_{j,s}. \tag{15}$$

Z_j/S is full-horizon execution success and H_j is the worst-case Robust Execution Horizon. Let b_j and z_j be the medians across repeats of the *last recorded* maximum body-position and ankle-height errors, and $\widetilde{r}_j$ the median rollout-average evaluation reward. The recorded protocol maximizes the lexicographic key

$$K_j = (Z_j, H_j, -N_j, -b_j, -z_j, \widetilde{r}_j). \tag{16}$$

Success is considered first, then the earliest failure across repeats, then the number of resets. Tracking and reward resolve subsequent ties.

A candidate must have all S repeats, strictly exceed the current policy's key, and satisfy the configured body-error guard,

$$b_j \leq (1 + \epsilon) b_k, \qquad \epsilon = 0.20. \tag{17}$$

Missing guard measurements reject a candidate. Candidates are tested in descending rank order, and the first passing candidate replaces the current policy; otherwise, the current checkpoint is retained. Reward can therefore break an execution-and-tracking tie, but cannot reverse an earlier comparison. Figure 3 illustrates this accept-or-protect decision.

A.6 Heterogeneous Feedback Execution

Each iteration retains its manifest, input hashes, candidate patches, checkpoints, repeated metrics, and selection reasons. Accepted checkpoint state is persisted using atomic file replacement; resumption checks the manifest and recorded outputs. The loop stops on verified full-horizon execution, a protected current policy, or budget exhaustion (Algorithm 2).

MimicX-HLoop expresses independent work as a dependency graph with GPU, CPU, and I/O jobs. Ready work starts once dependencies and resources are available, allowing CPU diagnosis to overlap other GPU rollouts. Selection waits for all required evaluations. Execution modes preserve inputs, seeds, and budgets, and are compared through their outputs and selection outcomes. We measure this executor on the fixed-checkpoint rollout–diagnosis–selection workload specified next.

B Analysis of Policy-in-the-Loop Refinement

This section characterizes how execution feedback governs policy continuation. We first express refinement as a finite-budget two-level problem, then derive the protected selection property, the complementary roles of tracking rewards and practice distributions, and the scheduling invariance of heterogeneous feedback execution. The analysis uses the reference, verification key, and curriculum defined in Section A.

B.1 Finite-Budget Two-Level Optimization

Let $\mathcal{A}_B$ denote B PPO updates, including the checkpoint's optimizer state, and let ξ_k collect the prescribed training randomness. At iteration k, diagnosis D_k generates a finite proposal set $\mathcal{U}(D_k)$. Candidate j is obtained by

$$\theta_{k,j} = \mathcal{A}_B(\theta_k; R, W_{k,j}, C_{k,j}, \xi_k), \qquad (W_{k,j}, C_{k,j}) \in \mathcal{U}(D_k). \tag{18}$$

Here R remains fixed within the comparison. The inner procedure optimizes the candidate's training objective; the outer decision evaluates its resulting policy under the common verification specification $\mathcal{V}$. Writing the actual finite optimization trajectory makes the training budget part of the problem, as in optimization-based bilevel formulations (Franceschi et al., 2018). Equation 18 uses the PPO continuation operator itself, rather than replacing it by an exact inner optimum.

Let $\widehat{K}_{\mathcal{V}}(\theta)$ be the key in Equation 16, computed from the prescribed repeated rollouts. Define $\mathcal{F}_k$ to contain the incumbent and all fully evaluated candidates with available guard measurements satisfying Equation 17. The outer decision is

$$\theta_{k+1} \in \arg\max_{\theta \in \mathcal{F}_k}{}_{\text{lex}} \widehat{K}_{\mathcal{V}}(\theta), \tag{19}$$

with incumbent preference on an equal key and a fixed proposal order for remaining ties. This is equivalent to testing candidates in descending key order and selecting the first guarded strict improvement. Thus the outer criterion acts on executed behavior, while objective and curriculum changes remain instruments for producing better candidates. Execution defines the outer objective.

B.2 Protected Execution-Priority Selection

Fix the reference, horizon, termination rules, evaluation seeds, and metric aggregation. For the following statements, a checkpoint has a reproducible verification record under this specification. This defines a common empirical ordering across refinement iterations.

Proposition B.1 (Monotone verified selection)**.** *Every completed refinement iteration satisfies*

$$\widehat{K}_{\mathcal{V}}(\theta_{k+1}) \succeq_{\text{lex}} \widehat{K}_{\mathcal{V}}(\theta_k). \tag{20}$$

The inequality is strict when a candidate is accepted. For any accepted candidate, the guarded body statistic satisfies $b_{k+1} \le (1+\epsilon)b_k$; otherwise $\theta_{k+1} = \theta_k$.

Proof. The incumbent belongs to $\mathcal{F}_k$, so the maximum in Equation 19 cannot have a smaller key. Incumbent preference excludes replacement by an equal key, giving strict improvement whenever replacement occurs. Membership of an accepted candidate in $\mathcal{F}_k$ gives the body bound. If no admissible candidate strictly improves the key, the incumbent is retained. Applying the same argument successively yields Equation 20 over all completed iterations. □

The ordering protects the specified execution priorities: Z cannot decrease; when Z is unchanged, H cannot decrease; when both are unchanged, N cannot increase. Lower-priority coordinates resolve the remaining ties. The body guard additionally restricts b, the median of the last-record maximum body errors, even when a higher-priority execution coordinate improves. These are properties of the common verification records; the mean and tail tracking statistics reported in Tables 10 and 11 provide separate empirical measurements of motion fidelity.

Corollary B.2 (Reward cannot overturn an execution rejection)**.** *If a candidate's execution prefix $(Z_j, H_j, -N_j)$ is lexicographically smaller than the incumbent's, no increase in its evaluation reward can make it acceptable. Likewise, a violated body guard rejects a candidate independently of reward.*

Proof. Lexicographic comparison is determined by the first differing coordinate. The execution prefix precedes all tracking coordinates and reward, so changing the final reward coordinate leaves an inferior execution prefix inferior. The body guard is a separate membership condition for $\mathcal{F}_k$ and does not depend on reward. □

This explains the Kung Fu case in Table 5: both proposals increase reward but shorten the worst execution horizon, so the stronger execution record is retained.

B.3 Complementarity of Objective and Curriculum

Practice changes which states contribute to learning. Let μ_0 be the reset-state distribution for a full-start rollout and $\mu_{\mathcal{W}}$ the distribution induced by reference-state initialization inside the practice window. Both include the prescribed reset randomization. For fixed W and window, define

$$J_{\mu,W}(\theta) = \mathbb{E}_{x_0\sim\mu,\,\tau\sim\pi_\theta}\left[\sum_{t=0}^{L-1}\gamma^t r_t(W)\right], \qquad \mu_\rho = (1-\rho)\mu_0 + \rho\mu_{\mathcal{W}}, \tag{21}$$

where L is the rollout budget, rewards after termination are zero, and $\gamma \in [0,1]$. The episode clock t is distinct from its reference phase.

Proposition B.3 (Mixture objective and window exposure)**.** *For a fixed policy and supervision, the expected return satisfies*

$$J_{\mu_\rho,W} = (1-\rho)J_{\mu_0,W} + \rho J_{\mu_{\mathcal{W}},W}. \tag{22}$$

Where these objectives are differentiable and ρ, W, and the reset distributions are held fixed, their gradients satisfy the same mixture identity. If $p_{\mathcal{W}}$ is the probability that a full-start rollout visits the practice window, counting initialization inside it as a visit, then

$$p_\rho = \rho + (1-\rho)p_{\mathcal{W}}, \qquad p_\rho - p_{\mathcal{W}} = \rho(1-p_{\mathcal{W}}) \geq 0. \tag{23}$$

Proof. Condition on the Bernoulli choice of reset distribution. The law of total expectation gives Equation 22; differentiation of this finite weighted sum gives the gradient identity. The visit probability is one under window initialization; under a full start, it is $p_{\mathcal{W}}$. Marginalization yields Equation 23. □

The exposure gain is largest when the current policy seldom reaches the diagnosed transition. For example, with $p_{\mathcal{W}} = 0.1$, the two proposal probabilities $\rho = 0.45$ and 0.65 give visit probabilities 0.505 and 0.685, respectively. This calculation concerns opportunities to practice; successful execution through the window is measured by the subsequent full-start verification. The gradient identity characterizes the objective sampled by continuation, with PPO supplying its finite-budget update rule.

Reward parameters change sensitivity within visited states. For one positional group, write $E_G = |G|^{-1}\sum_{b\in G}\|e_b\|^2$ and $r_G = w_G\exp(-E_G/\sigma_G^2)$, where $e_b = p_b - \tilde{p}_b$. Holding the error coordinates independent gives

$$\frac{\partial r_G}{\partial e_b} = -\frac{2w_G}{|G|\sigma_G^2}\exp(-E_G/\sigma_G^2)e_b. \tag{24}$$

This is the reward's sensitivity to tracking error. Policy optimization uses sampled returns and policy gradients (Sutton et al., 1999), rather than differentiating this expression through the simulator. Increasing w_G scales the sensitivity, while changing σ_G changes which error magnitudes receive it. For fixed $E_G > 0$, the scale-dependent factor $u^{-1}\exp(-E_G/u)$, with $u = \sigma_G^2$, has derivative $\exp(-E_G/u)(E_G - u)/u^3$ and is maximized at $u = E_G$. Thus reducing a scale indefinitely does not indefinitely strengthen supervision at a nonzero error. This motivates the bounded joint search over weights and scales in Section A.

For the nonnegative tracking terms indexed by $\mathcal{I} = \mathcal{L}\cup\mathcal{G}\cup\mathcal{D}$, Taylor's theorem yields the explicit quadratic approximation

$$\sum_{i\in\mathcal{I}} w_i e^{-E_i/\sigma_i^2} = \sum_i w_i - \sum_i \frac{w_i}{\sigma_i^2}E_i + \mathcal{R}, \qquad 0 \leq \mathcal{R} \leq \frac{1}{2}\sum_i w_i\left(\frac{E_i}{\sigma_i^2}\right)^2. \tag{25}$$

Indeed, $e^{-x} = 1 - x + e^{-\zeta}x^2/2$ for some $\zeta \in [0,x]$ when $x \geq 0$; summing with $w_i \geq 0$ gives the bound. Locally, the tracking objective therefore assigns effective coefficients w_i/σ_i^2 to local, global, and dynamic squared discrepancies. Action smoothing and physical penalties remain additional terms as in Equation 13. Curriculum modifies exposure to these discrepancies, while objective refinement modifies their relative sensitivity. The two operations act on different parts of the learning problem, providing a rationale for their combination in the component comparison of Table 12.

B.4 Decision-Preserving Heterogeneous Execution

Represent the fixed-checkpoint feedback workload by a directed acyclic graph $\mathcal{G} = (V, E)$. Each node v consumes immutable external inputs a_v, outputs from its predecessors, and a node-local random stream ξ_v:

$$y_v = F_v(a_v, \{y_u : u \in \mathrm{pred}(v)\}, \xi_v)\,. \tag{26}$$

Assume these functions produce schedule-independent semantic outputs for the same inputs and random streams, with no shared mutable state between independent nodes. Timing and log-order metadata are not part of y_v. The selector consumes all required reports in a fixed identity order and uses a deterministic tie rule. These conditions specify the reproducible workload whose execution is rescheduled; no policy update is inserted into it.

Proposition B.4 (Scheduling invariance)**.** *Two dependency-respecting schedules completing the same graph under these conditions produce identical semantic outputs and decisions.*

Proof. Proceed by topological induction. Source nodes receive identical external inputs and random streams, so their outputs agree. If every predecessor of a node has matching outputs, Equation 26 supplies the same arguments to its function in both schedules, and its output also agrees. The graph is finite and acyclic, so this covers all nodes. In particular, the selector receives the same complete ordered report collection; deterministic selection then gives the same decision. □

The assumption concerns reproducible node outputs, not merely equal seed integers. It can be checked independently of elapsed time by comparing semantic outputs and decisions. Table 7 reports identical selection outcomes for the measured feedback workload, alongside a median $1.945\times$ speedup over sequential execution.

Where the time savings arise. For fixed node service times d_v, suppose each node occupies one slot of its assigned resource class $r \in \{\mathrm{GPU}, \mathrm{CPU}, \mathrm{I/O}\}$, with m_r available slots. Write $W_r = \sum_{v:r(v)=r} d_v$, and let $L_{\mathrm{crit}} = \max_P \sum_{v\in P} d_v$ be the longest dependency-path duration. Every valid schedule has makespan

$$T_{\mathrm{sched}} \geq \max\left\{L_{\mathrm{crit}}, \max_r \frac{W_r}{m_r}\right\}. \tag{27}$$

Nodes on one path must run in sequence, giving the first bound; a class with m_r slots can supply at most $m_r T_{\mathrm{sched}}$ slot-time, giving the second. Sequential execution pays $\sum_v d_v$ before orchestration overhead, whereas heterogeneous execution can overlap independent rollout and diagnosis work. The bound identifies the critical-path and resource-load constraints; the measured wall times include actual scheduling overhead. HLoop thus accelerates feedback by changing when independent work executes while preserving what information reaches selection.

C Complete Experimental Protocol

C.1 Tasks, Robot, and Learning Backend

The core study comprises Tennis Swing, Football Juggling, Dance Sequence, and Kung Fu Sequence, with evaluation horizons of 800, 455, 784, and 1,062 control steps, respectively. The corresponding reference lengths are 519, 455, 784, and 1,062 frames at 50 Hz. All use a 29-DoF Unitree G1 on flat terrain in MuJoCo/MjLab (Todorov et al., 2012). Simulation advances at 200 Hz and the policy at 50 Hz. Fourteen bodies are tracked, with the torso as the anchor. The actor has 160 input dimensions and 29 joint-position outputs; actor and critic use multilayer perceptrons with hidden widths $(512, 256, 128)$, ELU activations, and observation normalization.

PPO (Schulman et al., 2017) collects 24 steps from each of 1,024 environments per iteration, followed by five optimization epochs and four minibatches. Its clipping parameter is 0.2, discount 0.99, advantage decay 0.95, entropy coefficient 0.005, and target KL 0.01. Task-ordered continuation learning-rate settings are $(10^{-6}, 7\times10^{-7}, 9\times10^{-7}, 5\times10^{-7})$, with an adaptive schedule. The tracking configuration disables observation corruption, push perturbations, joint initialization noise, and the configured friction, encoder-bias, and center-of-mass randomizations.

C.2 Controlled Continuation Protocol

Each method uses continuation seeds 101, 202, and 303 on every core task. Each task has one registered warmstart checkpoint shared by all methods and continuation seeds.

Fixed Reference. We resume PPO for 250 iterations using the original retargeted motion and fixed tracking objective. The backend's existing adaptive motion-start sampling remains active.

Failure Curriculum. We use the same reference and reward profile, replacing start sampling by a 50% mixture of full-start and window initialization. Previously diagnosed windows are $[340, 430)$ for Tennis Swing, $[380, 430)$ for Football Juggling, $[500, 545)$ for Dance Sequence, and $[165, 225)$ for Kung Fu Sequence.

Task-Aware Refinement. We retain this curriculum and add task-priority position, height, and linear-velocity rewards, pelvis/torso position and angular-velocity rewards, stronger anchor tracking, and adjusted action smoothing. Priority groups are the arms for tennis, both legs for football, pelvis/torso and ankles for dance, and pelvis/torso and both legs for kung fu.

MimicX. We run one automated refinement iteration with three heuristic candidates, each receiving 250 PPO iterations from the shared warmstart. Its full supervision profile includes inherited wrist, core, and stance-slip terms before diagnostic patches are added. Thus the first three methods each use $250 \times 24 \times 1{,}024 = 6{,}144{,}000$ new transitions; MimicX uses three such candidate budgets, or 18,432,000 transitions, plus verification. Dance Sequence and Kung Fu Sequence use registered repaired motion files for MimicX, including its current-policy evaluation; the other methods use the original references. Tennis Swing and Football Juggling share the original reference across methods. Body errors are evaluated against each method's selected reference, not re-evaluated against a common original target.

C.3 Verification and Metric Reduction

All reported core checkpoints are evaluated with seeds 1001, 2002, and 3003, yielding nine rollouts per task–method cell. Each rollout uses one environment, deterministic policy inference, full-start initialization, and active termination tests. These tests flag anchor-height error above 0.35 m, either ankle's reference-relative height error above 0.25 m, or an absolute difference above 0.8 between the vertical components of reference and robot anchor-frame gravity. The orientation threshold is a projected-gravity discrepancy. MimicX reports the selected policy's verification repeats; protected trials report the shared warmstart again.

Success and Robust Execution Horizon follow Equation 15: tennis success uses the sentinel $800+1 = 801$, not the reference length. At motion end, the backend restarts the reference and reinitializes robot state. Success thus means no strict termination over the evaluation schedule. The evaluator also continues after termination resets. The instantaneous whole-body scalar is

$$e_t^{\text{body}} = \max_{b \in \mathcal{B}} \|p_{t,b} - \widetilde{p}_{t,b}\|_2. \tag{28}$$

Mean, nearest-rank 95th percentile, and peak pool all steps across three repeats; the mean is a temporal mean of worst-body error. Anchor error uses world-position distance, velocity errors average bodywise Euclidean discrepancies, and the end-effector height scalar takes the maximum over the two ankles. These summaries differ from the final-record medians used by the guard. Task summaries average training-seed values, while success counts all nine repeats.

C.4 Breadth and External Evaluation

Four additional videos cover Basketball, Soccer, Michael Jackson dance, and Uniandes Dance. Their Fixed Reference/MimicX comparisons use three training seeds, 300 iterations per continuation, and three evaluation repeats; MimicX allows two iterations with one candidate each. Reference-input breadth covers 11 LaFAN1 references (Harvey et al., 2020) and three AMASS/ACCAD references (Mahmood et al., 2019), including two retargetings of the Form 1 source motion, using training seed 101, three repeats, and the 250-iteration candidate budget.

The direct tracker study compares the standalone MjLab reproduction of BeyondMimic (Liao et al., 2025), released SONIC (Luo et al., 2026), Fixed Reference, and MimicX on Tennis and Football. Dance and Kung Fu add common-reference comparisons of Fixed Reference, SONIC, and MimicX. Each task uses identical target coordinates across its compared policies and one full interval at 50 Hz without failure resets (Table 16). BeyondMimic is continued from the shared warmstart for 250 PPO updates with 1,024 environments, continuation seeds 101/202/303, and configured learning rate 10^{-6}. Evaluation uses seed 1001. SONIC is evaluated in three independent executions, following a five-second controller settling phase; logged target frames synchronize the motion, and offline rendering preserves real-time simulation. Joint RMSE and fourteen-body root-local forward-kinematic error measure articulation under common definitions (Table 2).

C.5 Heterogeneous Execution Benchmark

The MimicX-HLoop benchmark freezes the four Fixed Reference checkpoints trained with seed 101. Two evaluation seeds per checkpoint produce eight GPU rollout jobs, eight dependent CPU diagnosis jobs, and one selection job, with zero policy updates. The selector retains diagnoses whose frontier reaches the requested horizon and uses a deterministic ordering in every execution mode. Sequential, bulk-synchronous, and dependency-driven execution share this 17-job graph; parallel modes use two GPU slots and at most eight CPU workers. After one warmup per mode, five repetitions rotate execution order. We report median wall time, its ratio to sequential execution, and identical selection outcomes for this timed workload. Table 9 summarizes the sampling units across the core, breadth, direct-policy, and timing studies.

Table 9: Experimental coverage and independent sampling units. Training seeds, repeated evaluations and task aggregates are distinguished explicitly.

Study	Recorded evidence	Tasks	Methods	Train seeds	Eval. repeats
Core tracking	48 policies; 144 rollouts	4	4	3	3
Repeated verification	12 decisions; 6 accept, 6 protect	4	–	3	3
Same-motion trackers	42 eligible recordings	4	3–4	3[a]	1
Collision scenes	30 final rollouts	5	2	1	3
Additional videos	4 paired task aggregates	4	2	–	–
Supplied motions	11 LaFAN1; 3 ACCAD references	14	2	–	–
Learning trajectories	3,000 updates; 250 per curve	1	4	3	–
Milestone evaluation	48 curves; 240 checkpoints	4	4	3	1
Heterogeneous execution	15 wall times; report parity	–	3	–	5

[a] Three continuation seeds or released-controller repetitions. Dashes denote an inapplicable axis or an aggregate-only dataset, not zero trials. Milestones evaluate one fixed seed; final core policies use three repeated evaluations.

D Full Core Results

D.1 Closing the loop improves execution and fidelity

Table 10 and Figure 5 evaluate the four component configurations on every core task. Relative to Fixed Reference, MimicX improves both reported body-tracking error and worst-case execution horizon for all twelve task-matched continuation-trial pairs. The task-macro effects are a 25.72% reduction in body error and a 255.57% increase in execution horizon, using the reference and metric definitions in Appendix C. Table 10 reports task-level values and Table 11 reports paired effects across metrics.

MimicX achieves full-horizon success in all nine Tennis evaluations; on the other three tasks, where every method records 0/9 full-horizon completions, it improves both execution horizon and mean body error over Fixed Reference.

Table 10 combines robust execution horizon with mean and 95th-percentile tracking errors, capturing both sustained execution and the magnitude of large tracking deviations. Improvements extend beyond mean body error. Both velocity metrics improve in all twelve pairs, while tail body and root-and-anchor errors improve in ten (Table 11).

D.2 Cumulative components address different execution errors

The component comparison in Table 12 identifies complementary effects of the supervision components. Failure-conditioned sampling makes difficult segments more influential in learning. It already brings Tennis to full-horizon execution and substantially extends the other tasks' execution horizons. Task-aware supervision further changes the allocation of tracking effort: on Football, for example, it further extends execution horizon while reducing root-and-anchor error.

The full loop selects a different operating point by combining proposals with repeated execution guards. It provides the lowest mean body error among the four methods on Tennis and Kung Fu, and the longest Football execution horizon. Intermediate variants lead on some other channels, including Dance execution horizon and several root-tracking values. These results show how coordinated refinement and explicit acceptance criteria balance complementary execution and tracking objectives. Table 13 keeps local height, global tracking, and velocity channels visible together.

Body error is the temporal mean of the maximum aligned tracked-body error, pooled over rollout samples. Anchor is world-coordinate torso-anchor error; velocities average world-coordinate errors

Table 10: Four-task continuation comparison. Entries show mean ± sample SD over three continuation-seed trials, each aggregating three evaluation rollouts. Bold marks each task–metric best mean, including ties.

Method	Horizon (steps, ↑)	Reward (↑)	Body mean (m, ↓)	Body P95 (m, ↓)	Anchor error (m, ↓)	Linear vel. (m/s, ↓)	Angular vel. (rad/s, ↓)
Tennis Swing ($H = 800$ steps)							
Fixed Reference	322.0 ± 302.2	0.0657 ± 0.0023	0.241 ± 0.020	0.518 ± 0.125	0.301 ± 0.023	0.566 ± 0.053	2.287 ± 0.031
Failure Curriculum	**801.0** ± 0.0	**0.0816** ± 0.0008	0.184 ± 0.002	0.282 ± 0.007	0.208 ± 0.014	**0.426** ± 0.004	**1.798** ± 0.012
Task-Aware Refinement	768.0 ± 57.2	0.0788 ± 0.0014	0.178 ± 0.011	0.293 ± 0.012	**0.177** ± 0.023	0.454 ± 0.003	2.074 ± 0.027
MimicX	**801.0** ± 0.0	0.0750 ± 0.0012	**0.157** ± 0.008	**0.278** ± 0.009	0.276 ± 0.006	0.439 ± 0.017	1.979 ± 0.064
Football Juggling ($H = 455$ steps)							
Fixed Reference	53.7 ± 10.5	0.0437 ± 0.0043	0.286 ± 0.017	0.612 ± 0.096	0.130 ± 0.009	0.840 ± 0.056	3.669 ± 0.241
Failure Curriculum	264.3 ± 2.1	**0.0785** ± 0.0007	**0.237** ± 0.007	**0.395** ± 0.005	0.093 ± 0.006	**0.535** ± 0.006	**2.051** ± 0.025
Task-Aware Refinement	360.0 ± 39.8	0.0672 ± 0.0011	0.246 ± 0.007	0.448 ± 0.002	**0.083** ± 0.014	0.599 ± 0.019	2.765 ± 0.076
MimicX	**427.3** ± 22.4	0.0683 ± 0.0041	0.244 ± 0.017	0.504 ± 0.049	0.108 ± 0.026	0.568 ± 0.021	2.489 ± 0.032
Dance Sequence ($H = 784$ steps)							
Fixed Reference	193.3 ± 88.6	0.0435 ± 0.0115	0.270 ± 0.008	0.502 ± 0.051	0.215 ± 0.042	0.811 ± 0.056	3.779 ± 0.216
Failure Curriculum	506.3 ± 14.6	**0.0679** ± 0.0007	**0.243** ± 0.006	0.488 ± 0.035	0.100 ± 0.010	**0.622** ± 0.007	**2.892** ± 0.025
Task-Aware Refinement	**522.3** ± 0.6	0.0614 ± 0.0004	0.256 ± 0.008	0.482 ± 0.016	**0.077** ± 0.003	0.631 ± 0.003	3.484 ± 0.032
MimicX	341.3 ± 1.2	0.0581 ± 0.0001	0.244 ± 0.001	**0.467** ± 0.000	0.133 ± 0.000	0.686 ± 0.004	3.447 ± 0.010
Kung Fu Sequence ($H = 1062$ steps)							
Fixed Reference	97.7 ± 59.4	0.0565 ± 0.0046	0.249 ± 0.005	0.634 ± 0.189	0.229 ± 0.115	0.548 ± 0.062	2.084 ± 0.038
Failure Curriculum	298.7 ± 144.1	0.0832 ± 0.0070	0.194 ± 0.044	0.362 ± 0.084	0.140 ± 0.037	**0.220** ± 0.041	**0.998** ± 0.155
Task-Aware Refinement	**415.0** ± 94.0	0.0780 ± 0.0020	0.209 ± 0.011	0.358 ± 0.029	0.108 ± 0.007	0.255 ± 0.006	1.249 ± 0.093
MimicX	196.0 ± 0.0	**0.0871** ± 0.0000	**0.141** ± 0.000	**0.296** ± 0.000	**0.070** ± 0.000	0.230 ± 0.000	1.038 ± 0.000

Body error is the temporal mean of the maximum aligned tracked-body error, pooled over rollout samples. P95 is the temporal 95th percentile of the same aligned tracked-body maximum, pooled across three rollouts per continuation seed. Anchor denotes world-coordinate torso-anchor error. Horizon is the earliest first-failure step across three rollouts ($H + 1$ if none). Reward is dimensionless. Continuation seeds share a task-specific warmstart; protected Dance/Kung Fu trials retain the same checkpoint (Section C).

across bodies and time. Normalized horizon is $100 \min(h, H)/H$ per trial. Continuation, reference, and retained-checkpoint protocols are in Section C.

Scores are dimensionless (↑): each channel averages $\mathrm{clip}((b-a)/(|b|+10^{-6}), -2, 2)$ across its error terms. Local: wrists, ankles, and maximum ankle-height error. Global: world anchor position plus aligned bodywise mean and maximum errors. Dynamics: anchor/body linear and angular velocity. Progress is event step divided by task horizon.

D.3 Learning dynamics under selected supervision

Figure 9 and Table 14 summarize twelve 250-update Tennis continuations. These traces resolve the initial body-error transient, subsequent tracking improvement, reward development, and episode-length evolution across the full continuation.

For MimicX, the curves follow the continuation selected by the execution gate. All curves share a 250-update continuation axis; the total candidate budgets are specified in Appendix C. Jointly reporting reward, body error, and episode length reveals how fidelity and sustained execution develop under each supervision configuration.

Table 11: Paired improvements of MimicX over Fixed Reference. All eight recorded metrics are retained. Relative changes are percentages, with positive values denoting reward/horizon increases or error reductions; wins count strictly positive task–continuation-seed pairs.

Metric	Task macro (%, ↑)	Paired mean (%, ↑)	Wins	Paired median (%, ↑)	Minimum (%)	Maximum (%)
Reward	+39.44	+41.96	12/12	+39.97	+10.58	+91.90
Robust execution horizon	+255.57	+320.70	12/12	+255.78	+18.07	+904.65
Mean aligned worst-body error	+25.72	+25.65	12/12	+24.26	+7.35	+44.58
95th-percentile aligned worst-body error	+31.13	+29.25	10/12	+28.54	-5.51	+62.52
Root-and-anchor error	+33.28	+31.38	10/12	+30.66	-11.79	+80.65
Body linear-velocity error	+32.07	+31.81	12/12	+28.49	+10.17	+62.60
Body angular-velocity error	+26.14	+26.03	12/12	+22.34	+2.72	+51.20
End-effector height error (ankles)	-24.47	-38.27	5/12	-4.66	-197.04	+34.99

For baseline b and refined value a, let $r(b, a) = 100s(a - b)/b$, with $s = +1$ for reward/horizon and $s = -1$ for errors. Task macro is $\frac{1}{4}\sum_t r(\bar{b}_t, \bar{a}_t)$; paired mean is $\frac{1}{12}\sum_{t,k} r(b_{tk}, a_{tk})$. Neither is the ratio of pooled means. Body error is the temporal mean of the maximum aligned tracked-body error, pooled over rollout samples. Body p95 is the temporal 95th percentile of the same spatial maximum. Shared-warmstart and protected-policy interpretation: Section C.

Table 12: Complementary components. Task-balanced means over four motions.

Metric	Fixed Ref.	FC	TA	MimicX
Success (%)	2.8	**25.0**	22.2	**25.0**
Horizon (%)	21.5	62.7	**70.2**	64.0
Body (m)	0.262	0.215	0.222	**0.196**
Anchor (m)	0.219	0.135	**0.111**	0.147
Linear (m/s)	0.691	**0.451**	0.485	0.481
Angular (rad/s)	2.955	**1.935**	2.393	2.238

FC: Failure Curriculum; TA: Task-Aware Refinement. Higher success/horizon and lower errors are better. Bold marks metric bests. Horizon is $100\,\min(h, H)/H$.

Table 13: Task-critical channels. Strongest selected event for training seed 101, evaluation seed 1001.

Channel	Tennis	Football	Dance	Kung Fu
Phase (%)	25.1	100.0	87.6	37.4
Local	**0.432**	**0.740**	0.678	0.706
Global	0.248	0.630	**0.765**	**0.770**
Dynamics	0.225	0.193	0.527	0.671
Dominant	Local	Local	Global	Global
Gate	Accept	Accept	Protect	Protect

Bold identifies the dominant channel. Tennis/Football accept refinement; Dance/Kung Fu protect the current policy. Scores are dimensionless improvements; phase locates the selected event.

Table 14: Dense Tennis learning-log summaries from all 3,000 records: four methods, three continuation seeds, and 250 recorded updates per trajectory. Entries show mean ± sample SD, computed across seeds after obtaining each trajectory's endpoint or integral. Training reward retains its logged scale and differs from evaluation reward.

Method	Training reward			Body error (m)			Anchor error (m)		
	Initial	Final	AUC	Initial	Final	AUC	Initial	Final	AUC
Fixed Reference	0.29 ± 0.03	19.30 ± 1.29	15.96 ± 0.55	1.032 ± 0.021	0.189 ± 0.008	0.205 ± 0.002	0.080 ± 0.015	0.351 ± 0.015	0.388 ± 0.004
Failure Curriculum	0.23 ± 0.01	23.70 ± 1.91	18.45 ± 0.18	1.159 ± 0.355	0.152 ± 0.013	0.188 ± 0.001	0.095 ± 0.004	0.261 ± 0.023	0.377 ± 0.002
Task-Aware Refinement	1.12 ± 0.03	56.33 ± 1.07	45.89 ± 0.58	1.159 ± 0.355	0.163 ± 0.007	0.197 ± 0.005	0.095 ± 0.004	0.332 ± 0.036	0.436 ± 0.011
MimicX	4.36 ± 0.35	126.25 ± 1.59	103.74 ± 0.88	0.800 ± 0.741	0.143 ± 0.018	0.181 ± 0.019	0.056 ± 0.009	0.359 ± 0.035	0.457 ± 0.023

The logged aligned bodywise mean is distinct from the evaluation maximum: it averages aligned position errors across tracked bodies. Initial/final refer to updates 0/249. Normalized AUC uses trapezoidal integration divided by 249, preserving units. Shared-warmstart continuation: Section C.

D.4 Evaluation milestones across all core tasks

Table 15 follows 48 selected policy trajectories through 240 milestone evaluations. Task-Aware Refinement provides the highest macro horizon AUC, while MimicX provides the lowest final body error. This complements the dense training logs with checkpoint-level evidence on actual execution.

D.5 Temporal diagnostics connect fidelity to dynamics

Figure 11 aligns root orientation, acceleration, and angular velocity on a common execution clock. These time-resolved signals complement the episode-level gate statistics by locating rapid changes in orientation and motion dynamics; restart-crossing derivatives are excluded.

Table 15: Execution quality throughout policy continuation. Every cell averages three continuation seeds. The normalized horizon AUC integrates five checkpoints from 20% to 100% of the recorded continuation; the checkpoint columns report horizon fractions, followed by final body error. The four-task, four-method, three-seed design gives 48 curves and 240 evaluations.

Task	Method	AUC	20%	40%	60%	80%	100%	Body (m)
Tennis	Fixed Reference	0.580	0.463	0.674	0.725	0.462	0.460	0.239
	Failure Curriculum	**0.963**	0.705	1.000	1.000	1.000	**1.000**	0.183
	Task-Aware Refinement	0.941	0.845	0.861	1.000	1.000	0.959	0.183
	MimicX	0.921	1.000	1.000	0.842	0.842	**1.000**	**0.156**
Football	Fixed Reference	0.170	0.566	0.113	0.097	0.130	0.118	0.286
	Failure Curriculum	0.574	0.566	0.570	0.576	0.578	0.581	**0.237**
	Task-Aware Refinement	0.630	0.568	0.572	0.577	0.692	0.791	0.246
	MimicX	**0.744**	0.569	0.599	0.739	0.886	**0.939**	0.244
Dance	Fixed Reference	0.329	0.307	0.398	0.318	0.323	0.247	0.268
	Failure Curriculum	0.550	0.479	0.478	0.545	0.613	0.647	0.243
	Task-Aware Refinement	**0.622**	0.471	0.585	0.667	0.666	**0.666**	0.256
	MimicX	0.435	0.435	0.435	0.435	0.435	0.435	**0.241**
Kung Fu	Fixed Reference	0.114	0.176	0.167	0.065	0.090	0.092	0.249
	Failure Curriculum	0.257	0.187	0.214	0.270	0.308	0.288	0.194
	Task-Aware Refinement	**0.344**	0.238	0.297	0.425	0.340	**0.392**	0.211
	MimicX	0.185	0.185	0.185	0.185	0.185	0.185	**0.141**

Milestones use evaluation seed 1001. The three retained Kung Fu policies remain constant across their milestones. The continuation axis follows the selected trajectory; total candidate-search budgets are specified in Section C. Final core results separately aggregate three evaluation seeds.

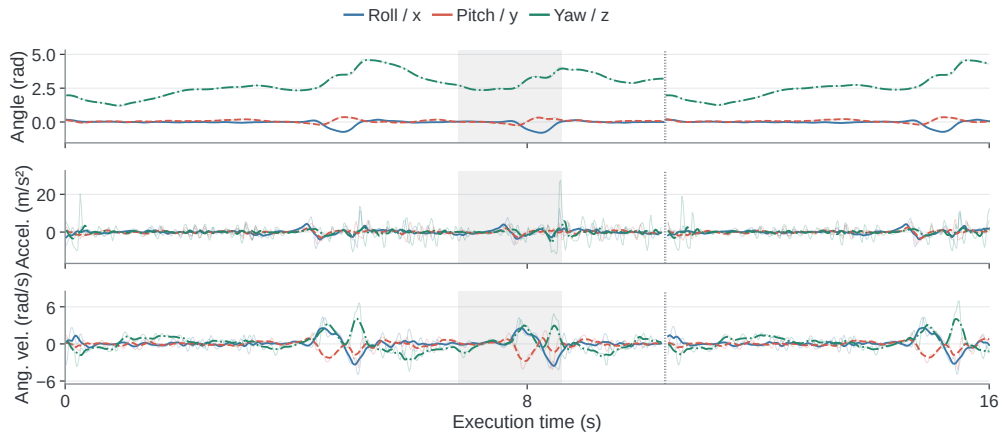


Figure 11: Aligned motion-state traces. Orientation, acceleration and angular velocity share a clock across 800 recorded Tennis states. Acceleration uses position differences; angular velocity uses quaternion increments. Derivative stencils crossing resets are excluded. Raw values and 11-sample means distinguish short transients from the motion trend.

E Gate and Rejection Case Studies

Across twelve full-loop trials, six proposals are accepted and six current policies are protected (Table 4). Every Tennis and Football trial accepts a proposal. Tennis's final policies complete all repeated evaluations; Football improves its worst-case execution horizon. On Dance and Kung Fu, the body-error guard and execution ordering retain the current policy. Candidate exploration therefore preserves the stronger verified execution whenever a proposed continuation regresses.

B: body-error guard; K: execution ordering. Superscripts identify the criteria determining each decision. The body guard uses the last recorded aligned worst-body error, not its temporal mean. Continuation seeds share a task-specific warmstart; protected Dance/Kung Fu trials retain the same checkpoint (Section C).

Reward is dimensionless; changes are relative to the current policy. Bold identifies the longest horizon, the execution criterion relevant to this decision.

The Kung Fu case in Table 5 isolates the reason for this separation. The current policy has reward 0.5707 and worst-case horizon 879. Two proposals increase reward to 0.6275 and 0.5838 but reduce that horizon to 677 and 689. Both are rejected. This case study uses its own repeated-verification

protocol. The increased reward accompanies earlier failure for both candidates, directly explaining why the execution ordering preserves the incumbent.

F Breadth and Direct Policy Evaluation

F.1 Direct same-motion tracker comparison

Table 2 compares policies on the identical Tennis reference. The standalone BeyondMimic tracker uses the MjLab reproduction linked by the original implementation (Liao et al., 2025). MimicX improves joint RMSE and root-local body error in every seed pair. Figure 6(a,c) shows the common articulation comparison, including Fixed Reference and the released SONIC controller (Luo et al., 2026). The same metric definitions apply to SONIC. All twelve recordings cover the same 518-step motion interval without failure resets.

The two articulation metrics use a common joint order and a common G1 kinematic model. Fixing the root pose for the body-position calculation isolates local configuration, complementing the global anchor and execution statistics in Appendix D. The comparison uses the shared warmstart and per-candidate continuation budget; Table 2 specifies the adaptation and evaluation settings.

Football Juggling provides a second matched-reference comparison. MimicX improves both articulation metrics over BeyondMimic in every seed pair (Table 16). Figures 6(a) and 14(d) resolve Tennis joint error and Football body-error gains across their complete intervals. Figure 12(a–c) adds recording-level means, body-error distributions, and joint-error phases.

Table 16 reports all common-reference task outcomes. On Dance, SONIC attains the lowest local articulation errors. On Kung Fu, the best joint and body means belong to Fixed Reference and SONIC, respectively. This non-reset, full-interval evaluation measures articulation through the complete clip; the core experiment measures reliable execution under its termination schedule and supervision.

Table 16: Same-reference policy evaluation across motions. Mean ± sample SD; lower is better. Bold marks the lowest task-wise mean.

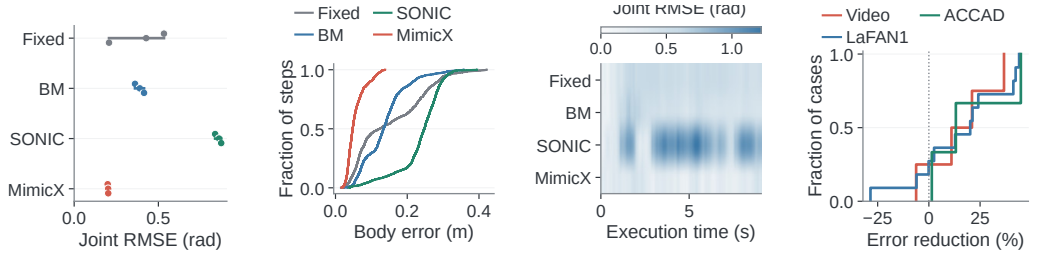
Task	Method	Runs	Joint RMSE (rad)	Local body error (m)
Tennis	Fixed Reference	3	0.389 ± 0.166	0.151 ± 0.072
	BeyondMimic (MjLab)	3	0.389 ± 0.027	0.137 ± 0.021
	SONIC (released)	3	0.857 ± 0.018	0.236 ± 0.001
	MimicX	3	**0.201** ± 0.002	**0.055** ± 0.002
Football Juggling	Fixed Reference	3	0.507 ± 0.050	0.190 ± 0.023
	BeyondMimic (MjLab)	3	0.446 ± 0.032	0.166 ± 0.007
	SONIC (released)	3	0.808 ± 0.004	0.216 ± 0.001
	MimicX	3	**0.286** ± 0.006	**0.070** ± 0.003
Dance	Fixed Reference	3	0.517 ± 0.047	0.167 ± 0.026
	SONIC (released)	3	**0.489** ± 0.004	**0.134** ± 0.002
	MimicX	3	0.530 ± 0.001	0.1695 ± 0.0004
Kung Fu	Fixed Reference	3	**0.603** ± 0.027	0.205 ± 0.021
	SONIC (released)	3	0.725 ± 0.003	**0.189** ± 0.003
	MimicX	3	0.624 ± 0.015	0.224 ± 0.005

All rows within a task share one reference hash and a full, non-reset playback interval: 518/454/783/1,061 steps for Tennis/Football/Dance/Kung Fu. Joint error is the temporal mean of per-frame joint RMSE. Local body error uses the same fourteen-body G1 forward kinematics with root pose fixed to identity. BeyondMimic denotes the MjLab reproduction; its original-reference Dance and Kung Fu continuations are not pooled with the repaired-reference replays. Seeds and continuation budgets follow Table 2.

Table 17: Additional video-task outcomes. Body error is reported in meters (↓); bold marks the lower error within each task. Values are recorded task aggregates. Success change is in percentage points (pp).

Video task	Fixed Reference	MimicX	Error reduction (%, ↑)	Δ success (pp, ↑)
Basketball	**0.204**	0.217	-6.22	+0.0
Soccer	0.279	**0.248**	+11.03	+33.3
Michael Jackson Dance	0.413	**0.326**	+21.04	+0.0
Uniandes Dance	0.289	**0.183**	+36.68	+0.0

Body error is the temporal mean of the maximum aligned tracked-body error, pooled over rollout samples.



(a) Tennis: run means (b) Tennis: body CDF (c) Football: joint phases (d) Cross-task gains

Figure 12: Complementary tracking diagnostics across recordings and tasks. (a) Tennis joint-error means for all three recordings and their range. (b) Distribution of all 3×518 Tennis body-error measurements per method. (c) Football joint-error means at each of 454 control steps. (d) Body-error reductions on 18 additional video and supplied-motion cases, grouped by input source; the full distribution includes negative changes. BM denotes the standalone BeyondMimic MjLab tracker; ACCAD includes the Form 1 sequence. Panels (a–c) use the common-reference comparison, while (d) compares MimicX with Fixed Reference on additional tasks.

F.2 Additional videos and motion-reference inputs

The additional video study extends the framework to Basketball, Soccer, Michael Jackson Dance, and Uniandes Dance (Table 17). Soccer improves both body tracking and full-horizon success, while Michael Jackson Dance and Uniandes Dance also reduce body error. Basketball retains the same success rate with higher body error; Table 17 reports the task-level values.

Table 18: Motion-reference breadth: fourteen evaluated references grouped by source dataset. Body error is reported in meters (↓); bold compares methods within a row. Cohort reductions average per-reference relative reductions rather than taking the ratio of cohort-mean errors.

LaFAN1: 11 evaluated references							
Motion reference	**Fixed**	**MimicX**	**Red. (%)**	**Motion reference**	**Fixed**	**MimicX**	**Red. (%)**
Dance 1	0.2866	**0.2499**	+12.81	Jumps 1	0.2871	**0.1606**	+44.06
Dance 2	0.2961	**0.2885**	+2.55	Run 1	0.2523	**0.1985**	+21.30
Fall and Get Up 1	**0.3746**	0.3972	-6.03	Sprint 1	0.2861	**0.2171**	+24.12
Fall and Get Up 2	**0.3540**	0.3544	-0.13	Walk 1	0.2462	**0.1419**	+42.36
Fight 1	0.2988	**0.2385**	+20.18	Walk 4	**0.2341**	0.3008	-28.51
Fight and Sports 1	0.3099	**0.1821**	+41.24	Reference-weighted mean	0.2932	**0.2481**	+15.81
AMASS/ACCAD: 3 evaluated references							
Motion reference	**Fixed**	**MimicX**	**Red. (%)**	**Motion reference**	**Fixed**	**MimicX**	**Red. (%)**
Form 1 (provided retargeting)	0.2140	**0.1858**	+13.17	Run, Change Direction	0.4376	**0.4313**	+1.43
Form 1 (new retargeting)	0.3728	**0.2054**	+44.90	Reference-weighted mean	0.3415	**0.2742**	+19.84

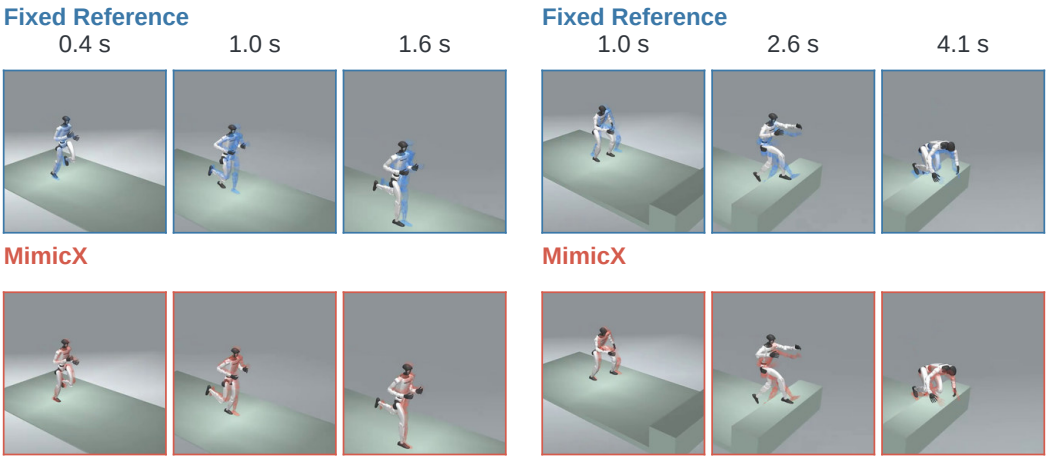
Body error is the temporal mean of the maximum aligned tracked-body error, pooled over rollout samples. Form 1 originates from AMASS/ACCAD. AMASS/ACCAD contains three references from two source motions, including two Form 1 retargetings. References are equally weighted.

Fourteen motion-reference cases provide a complementary test after the reference-construction interface (Table 18): eleven LaFAN1 references and three AMASS/ACCAD references, including two retargetings of the same Form 1 source motion. They include locomotion, dance, fighting and recovery motions. The results show sizeable body-error reductions for several skills, including jumps, walking, fight-and-sports and the ACCAD form sequence. Figure 12(d) retains every additional video and supplied-motion case in the gain distributions. The motion-input cases test the policy-side loop with an existing reference; the eight video tasks exercise the complete reconstruction-to-policy path.

F.3 Collision-scene video extension

Track Running, Stair Ascent, Forest Traversal, Platform Jump, and Parkour pair video-derived motions with fixed task-equivalent collision scenes. Each task uses a shared 3,000-update PPO checkpoint followed by equal 3,000-update Fixed Reference and Policy-in-the-Loop Refinement continuations, with training seed 101 and 1,024 parallel environments. Evaluation seeds 1001, 2002, and 3003 yield 30 final rollouts, with terminations enabled. The reset mixture and coordinated objective updates are described in Section 5. Execution and tracking guards govern acceptance independently of reward.

Strict completion requires the full scheduled interval without a flagged termination. Table 6 reports the minimum number of valid pre-failure steps across evaluation repeats, divided by the target length T; completed executions are recorded as T/T. Tracking reductions use the common pre-first-failure



(a) Track Running (b) Parkour transitions

Figure 13: Execution on additional task-equivalent collision scenes. Fixed Reference and MimicX at matched 20/50/80% source phases, evaluation seed 2002. Crops are identical between methods at each time. Track Running completes all repeated evaluations with lower body and root errors after refinement. Parkour panels depict the shared pre-failure interval; its full-horizon outcome is reported in Table 6.

Table 19: Execution-guided refinement in developmental case studies. Source-verified before/after transitions show complementary improvements in stability, task-critical articulation and motion completion. These are selected single-rollout mechanism cases, separate from the controlled multi-seed results.

Case	Measurement	Before	After	Change
Tennis initialization	Terminations in 800 steps	7	0	Stable rollout
Tennis articulation	Mean max-body error (m)	0.1411	0.1202	14.81% lower
Tennis articulation	Right-wrist error (m)	0.0977	0.0497	49.13% lower
Tennis articulation	Left-wrist error (m)	0.0932	0.0453	51.39% lower
Football execution	Terminations in 455 steps	1	0	Complete interval
Dance execution	Terminations in 784 steps	1	0	Complete interval

The articulation pair preserves zero terminations. Dance combines local reference repair with continued policy learning; the rows characterize their respective recorded supervision configurations.

prefix of each paired evaluation. We compute the temporal p95 of maximum body-position error and world-root position error, average each over the three repeats, and report the relative reduction. Thus completion measures the full execution, while prefix errors compare tracking at the same motion phases before either policy resets.

Refinement increases aggregate strict completion from 60% to 80%. Forest Traversal changes from 0/3 to 3/3 completions while reducing body and root prefix errors by 79.15% and 89.04%, respectively. Track Running, Stair Ascent, and Platform Jump preserve 3/3 completion with improved tracking. Parkour improves both prefix errors and reaches 220 of 255 steps in both conditions. Figure 10 complements these results with close views of Stair Ascent and the source-to-policy stages of Platform Jump. Figure 13 adds matched Track Running and Parkour transitions from policy evaluations.

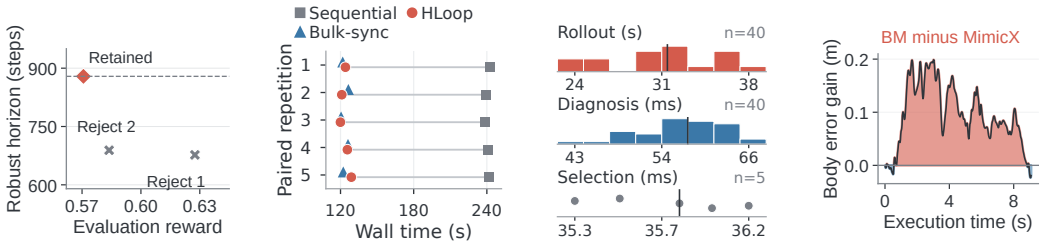
F.4 Developmental Refinement Cases

The developmental runs illustrate how execution feedback targets different stages of refinement (Table 19). Initialization and root-focused practice remove the observed Tennis terminations; subsequent articulation refinement reduces wrist errors while preserving zero terminations. Separate Football and Dance cases illustrate task-specific continuation and local reference repair, respectively.

G HLoop Execution Analysis

MimicX-HLoop reduces median wall time from 241.22 s to 124.02 s, a 1.945× speedup over sequential execution (Table 7), while preserving the same selection outcome across repetitions. Figure 14(b,c) shows paired timings for all three schedulers and their stage-duration distributions.

Table 8 gives each paired repetition. All execution modes use the same fixed-policy rollout–diagnosis–selection workload and selection criterion. Bulk-synchronous execution has a median time of 122.26 s.



(a) Execution-first gate (b) Paired wall time (c) Stage distributions (d) Football: body gain

Figure 14: Reliable refinement and faster feedback across motion inputs. (a) Execution verification rejects reward-improving regressions. (b) Five paired wall-time measurements: connectors link sequential execution and HLoop; triangles show bulk-synchronous execution. (c) Per-job durations: rollout and diagnosis histograms, five selection measurements, and median markers; each stage has its own linear axis and units. (d) BeyondMimic minus MimicX body error over the complete 454-step Football interval, averaged over three recordings; positive values indicate improved tracking.

Both parallel modes substantially shorten the feedback workload relative to sequential processing. Dependency-driven execution exposes each ready diagnosis as its rollout finishes, providing a scheduling interface for the policy feedback stage.

H Extended Related Work

Video-derived humanoid supervision. SMPL-X provides an expressive human representation (Pavlakos et al., 2019), GVHMR recovers world-grounded motion from monocular video (Shen et al., 2024), and GMR shows that retargeting quality materially affects downstream humanoid tracking (Araujo et al., 2025). These methods construct the motion consumed by a controller. MimicX adds an execution-feedback layer: rollout failures determine where the reference-conditioned objective and training distribution should change.

Physics-based motion tracking. DeepMimic, AMP, PHC, and MaskedMimic establish powerful mechanisms for reference imitation, reusable motion priors, recovery, and partial-motion conditioning (Peng et al., 2018; 2021; Luo et al., 2023; Tessler et al., 2024). BeyondMimic prioritizes empirically difficult motion segments and extends a tracker with guided diffusion (Liao et al., 2025); SONIC scales model, data, and compute for generalist whole-body tracking (Luo et al., 2026). In 2026, OmniXtreme separates general motor learning from actuation-aware physical refinement (Wang et al., 2026b), and reference-and-goal multi-task RL uses human motion as a prior for behavior that generalizes beyond the reference distribution (Wang et al., 2026a). MimicX uses measured policy failure to revise supervision and verifies continuation through repeated execution.

Adaptive execution and long-horizon skills. KungfuBot adapts tracking tolerances and processes contact-aware motions (Xie et al., 2025); AdaMimic adapts timing and actions from sparse keyframes (Huang et al., 2025); ASAP uses real trajectories to align simulation and robot dynamics (He et al., 2025). Perceptive Humanoid Parkour composes retargeted human skills into long-horizon trajectories before tracking and distillation (Wu et al., 2026), while HiWET couples world-frame end-effector accuracy to whole-body stability (Cao et al., 2026). These works reinforce the importance of coordinating local goals, global motion, and dynamics. Our method uses this structure inside a diagnosis–proposal–verification loop and applies an execution-first gate whenever the supervision bundle changes.

Scene-aware video imitation. MeshMimic reconstructs human trajectories and scene geometry, then combines kinematic consistency optimization with contact-invariant retargeting to learn motion–terrain interactions from monocular video (Zhang et al., 2026). MimicX complements geometric grounding by using policy execution feedback to revise tracking objectives and practice distributions, with repeated verification determining which continued policy is retained.

I Reproducibility and Artifact Organization

The refinement state links each motion reference, supervision proposal, policy continuation and repeated evaluation to a versioned source record. The accepted policy is updated only after complete candidate verification. Interrupted work resumes from the last accepted state and candidate set.

Table 9 summarizes the evidence units. The controlled analysis retains all tasks, training seeds and repeated evaluations; the broader video and motion-input studies have their own denominators. The systems comparison uses a fixed workload and checks selection outcomes across execution modes. Each study reports its own sampling units and evaluation protocol.
Figure provenance distinguishes input RGB, reconstructed human geometry, robot reference, dynamic policy execution, diagnostic views and scene presentation. The paper uses vector numerical plots and experimental images derived from preserved masters. Quantitative results come from numerical tables and logged trajectories; scene renderings provide spatial context for the recorded motion. The accompanying video collection retains input, intermediate and final execution views so that temporal behavior can be inspected beyond the selected stills.

J Qualitative Evidence and Video Correspondence

J.1 Human reconstruction and learned execution

Figure 4 shows distinct pipeline stages for two tasks. Figures 18 and 19 add Fixed Reference comparisons at task-critical phases. The sequences below connect original video frames to recovered human motion and recorded policy states. Each upper row contains seven input frames; the lower row progresses from three human states to four robot states. Tennis and Football use the recovered source-phase mapping. Dance uses nominal source and rollout timestamps over a selected interval. Figure 8 gives the Kung Fu example in the main text; Figure 21 adds Track Running and Stair Ascent.

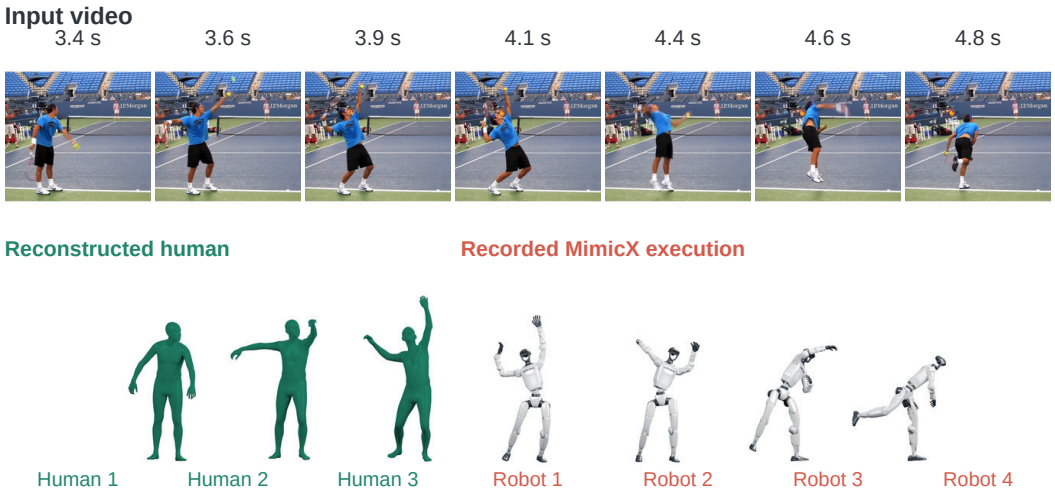


Figure 15: Tennis across time and embodiment. Seven original input frames accompany three reconstructed human states and four policy states. Source phases associate the two rows.

J.2 Task-critical events reveal execution differences

The task-event plates show Fixed Reference and MimicX at selected execution moments for Tennis, Football Juggling, Dance and Kung Fu (Figures 18–20b). In these plates the solid robot is the policy execution and the translucent method-colored robot is its reference. Fixed Reference uses gray and MimicX uses coral. The comparison focuses on visible pose and reference alignment; the accompanying numerical tables summarize every declared trial.

J.3 A continuous scene makes motion progression readable

Figure 22 shows two seven-state sequences in task scenes. The Tennis view includes the racket and registered ball; Football ball placement follows observations at the recovered source phases. Spatial offsets separate the recorded states and increasing opacity encodes chronology. Here translucent robots denote earlier policy states, not reference overlays.

J.4 Scene reconstruction connects the rendered policy to the input

Figure 4 links the observed Tennis frame, camera-space human reconstruction, policy in simulation, and the same policy presented in a reconstructed scene. The stage panels use a common layout. The final presentation registers the recovered camera to a Gaussian background while preserving the

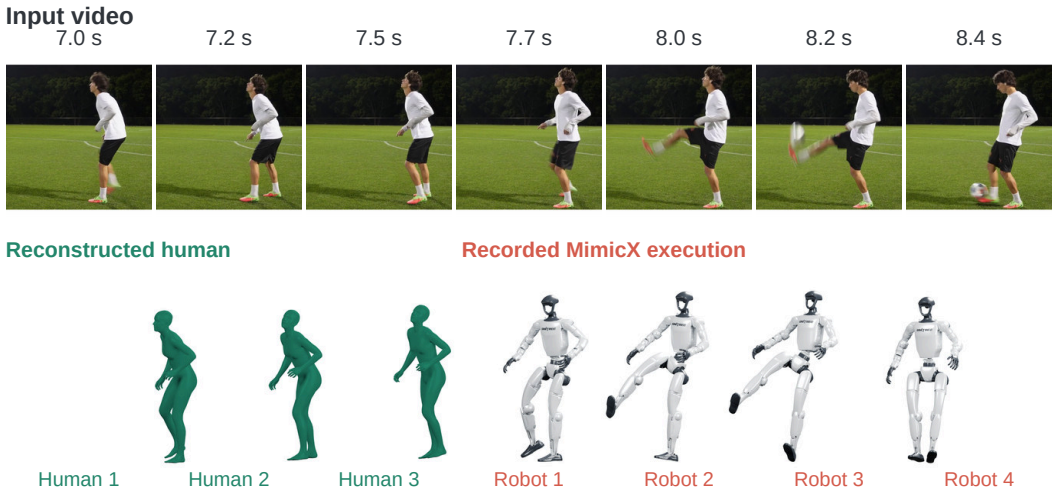


Figure 16: Football Juggling across time and embodiment. Original-video frames provide the scene and action context for the reconstructed human motion and policy sequence below.

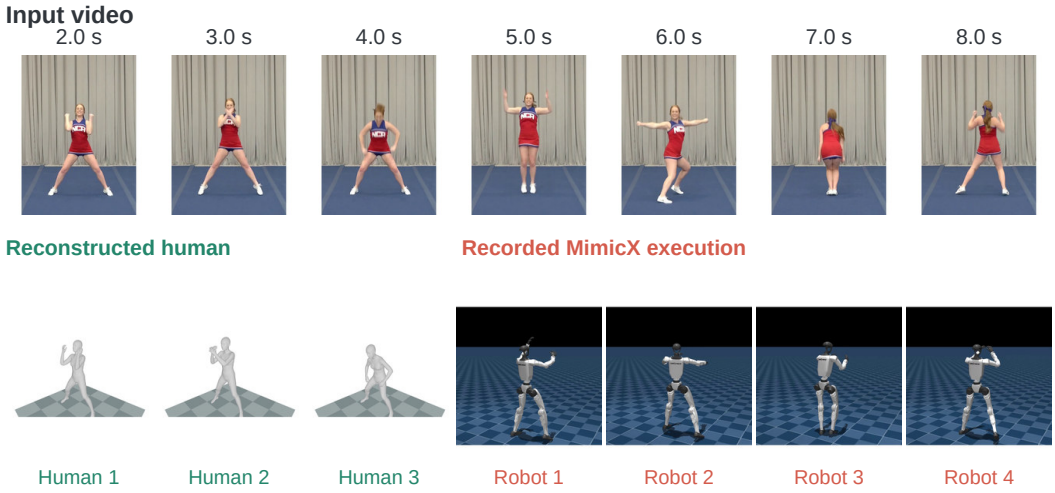


Figure 17: Dance across time and embodiment. Seven original-video frames show the selected interval; the lower row follows human reconstruction and policy execution.

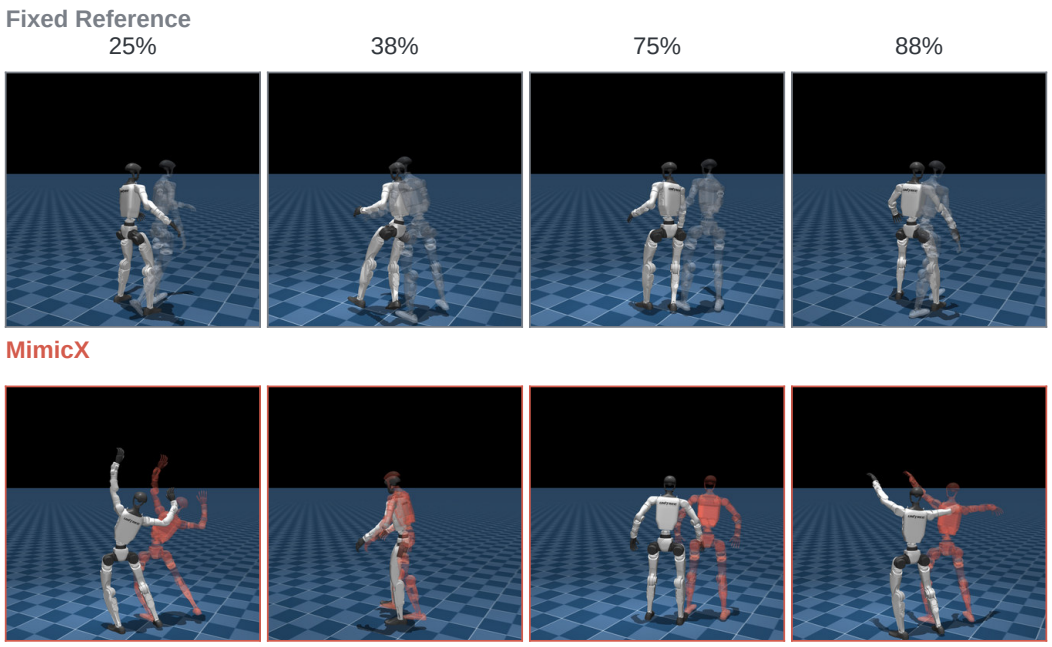


Figure 18: Tennis task-critical execution. Fixed Reference and MimicX at four recorded event times. The neutral robot is the policy and the translucent method-colored robot is the reference overlay. Gray identifies Fixed Reference and coral identifies MimicX.

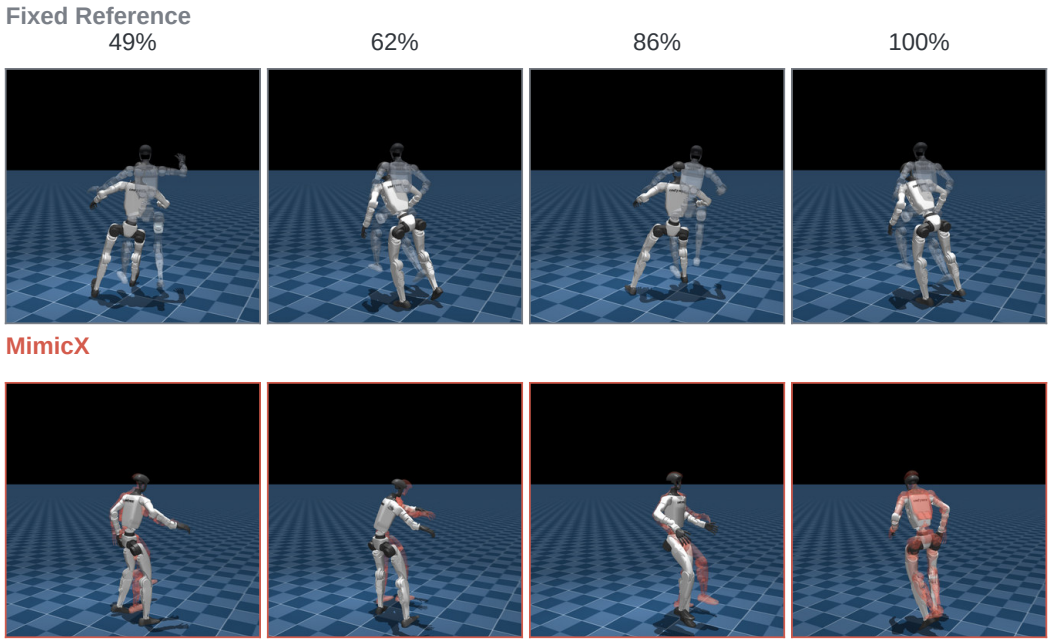
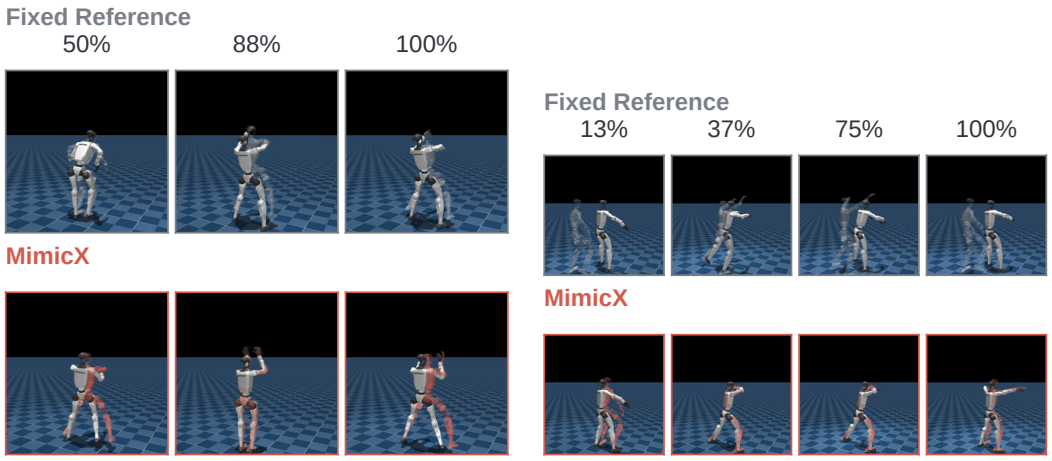
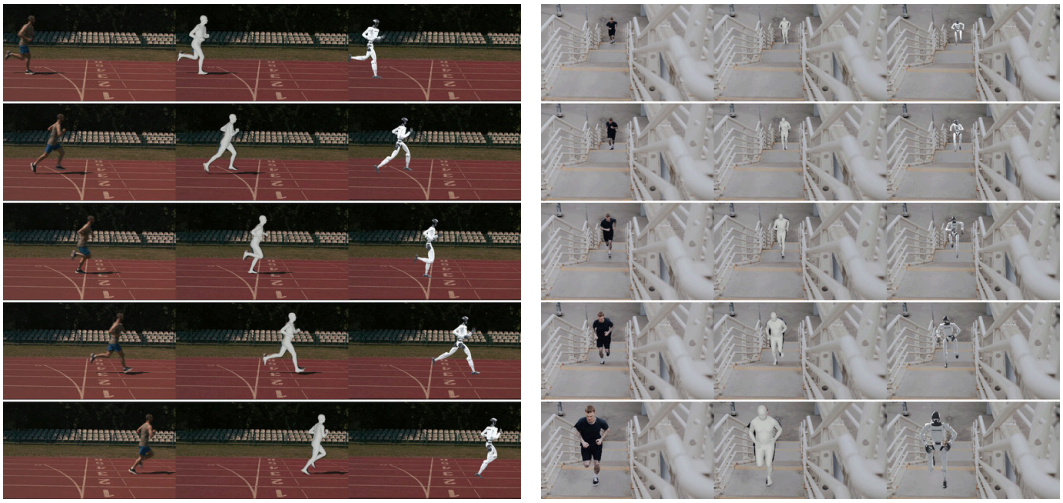


Figure 19: Football Juggling task-critical execution. Four recorded event times compare policy alignment with the reference motion. Alternating support and raised legs reveal whole-body juggling coordination.



(a) Dance: three task-critical events (b) Kung Fu: four task-critical events

Figure 20: Task-critical execution on articulated whole-body motions. Fixed Reference and MimicX are compared at matched events. Solid robots show execution; translucent robots show method-colored references.



(a) Track Running (b) Stair Ascent

Figure 21: Source-view correspondence in Track Running and Stair Ascent. Columns show input video, reconstructed human, and G1 reference; rows follow five synchronized instants.

robot's recorded pose. This gives the reader both an unobstructed simulation view and the original scene context for the recorded tracking experiment.

The Tennis presentation video contains 260 frames at 25 Hz. The foreground is rendered at 1624×1440, and the source-view Gaussian background at 2436×2160 from an 812×720 source image. Camera registration has a median feature reprojection error of 0.647 source pixels, and all foreground contours remain inside the output canvas. These measurements quantify camera-registration accuracy and foreground coverage in the reconstructed-scene presentation.

J.5 Visual breadth beyond the four core tasks

Figures 23 and 24 show two additional video-derived tasks with positive tracking results. The motion-reference plates in Figures 25 and 26 illustrate the reference-input pathway. Together with the numerical breadth table, these examples show how the same diagnosis, proposal and verification interfaces apply after different reference-construction routes.

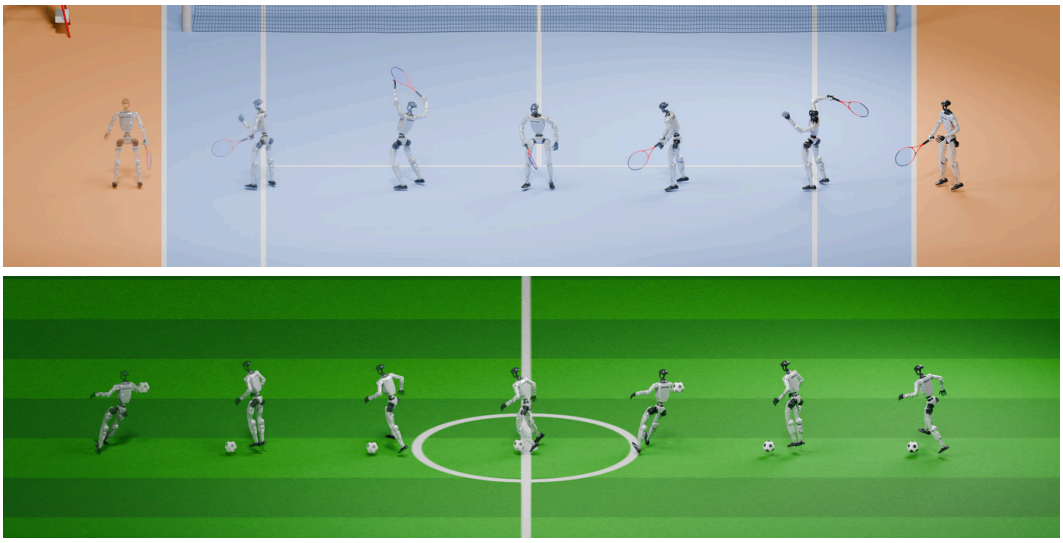


Figure 22: Video-driven motion in task scenes. Tennis (top) and Football Juggling (bottom), each shown as seven recorded robot states. Spatial offsets separate the poses and increasing opacity encodes chronology. The selected Tennis sequence is MimicX; the Football sequence is Fixed Reference. Scene objects provide presentation context for the recorded motion.

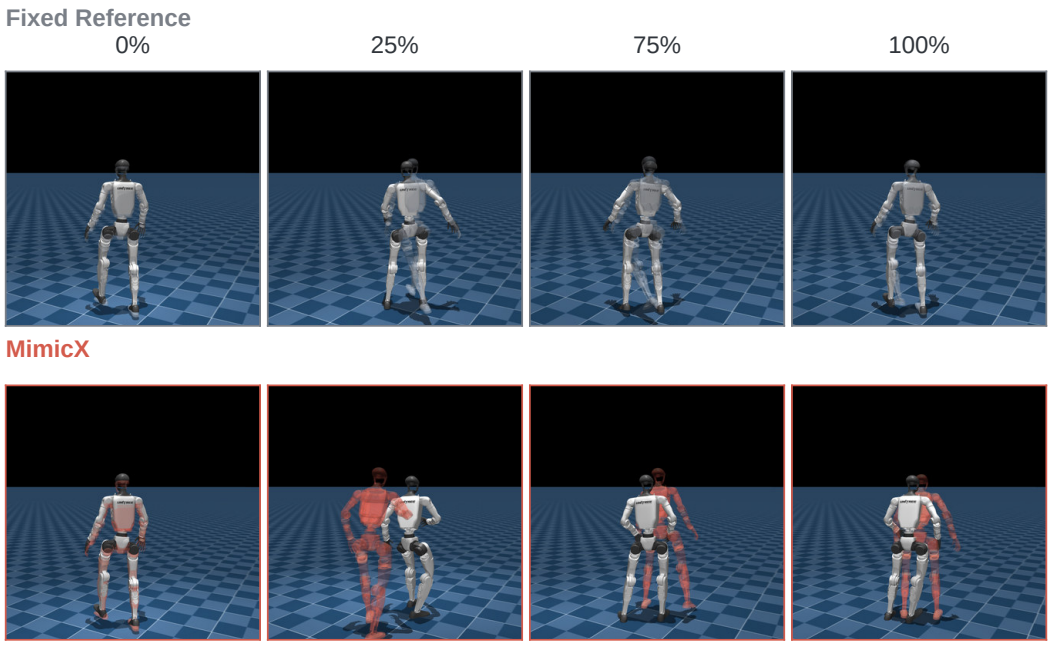


Figure 23: Additional Soccer video task. Fixed Reference and MimicX at four recorded events, with method-colored reference overlays. This task is separate from the core Football Juggling video.

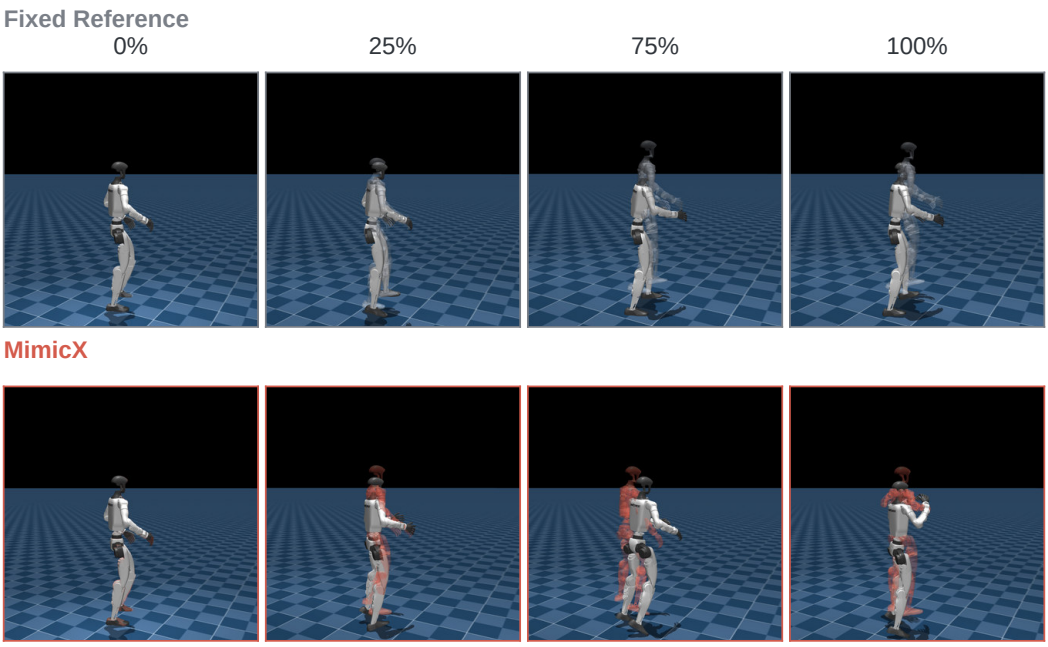


Figure 24: Uniandes Dance video task. Four task-critical events show policy–reference alignment.

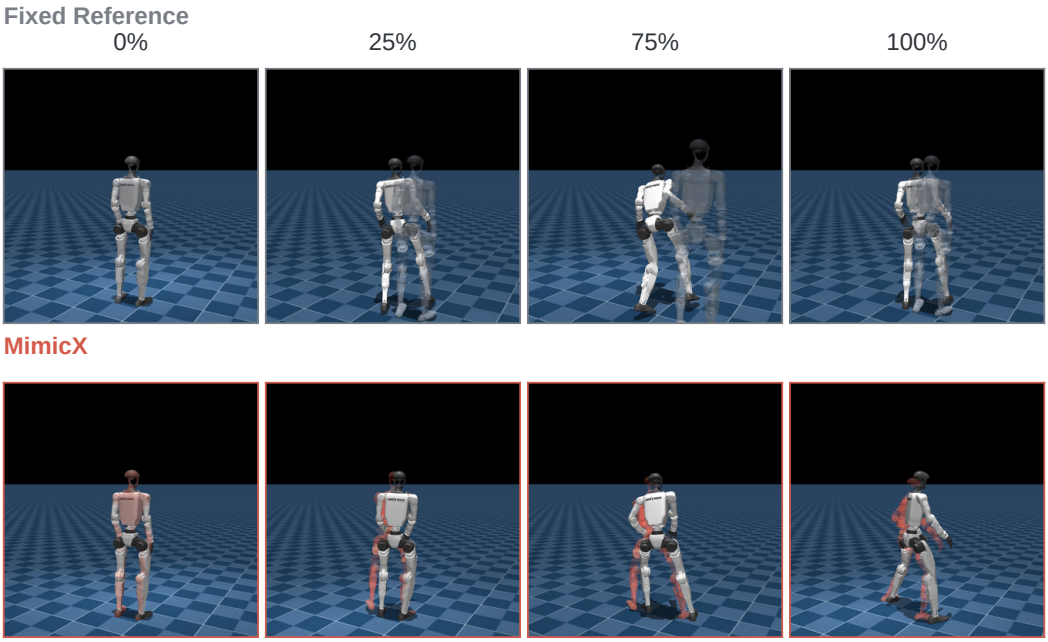


Figure 25: AMASS/ACCAD Form 1: provided retargeting. Fixed Reference and MimicX on the provided robot-motion reference, with recorded policy states and method-colored reference overlays. Source attribution follows the motion audit in Table 18. This pathway starts from an existing motion reference.

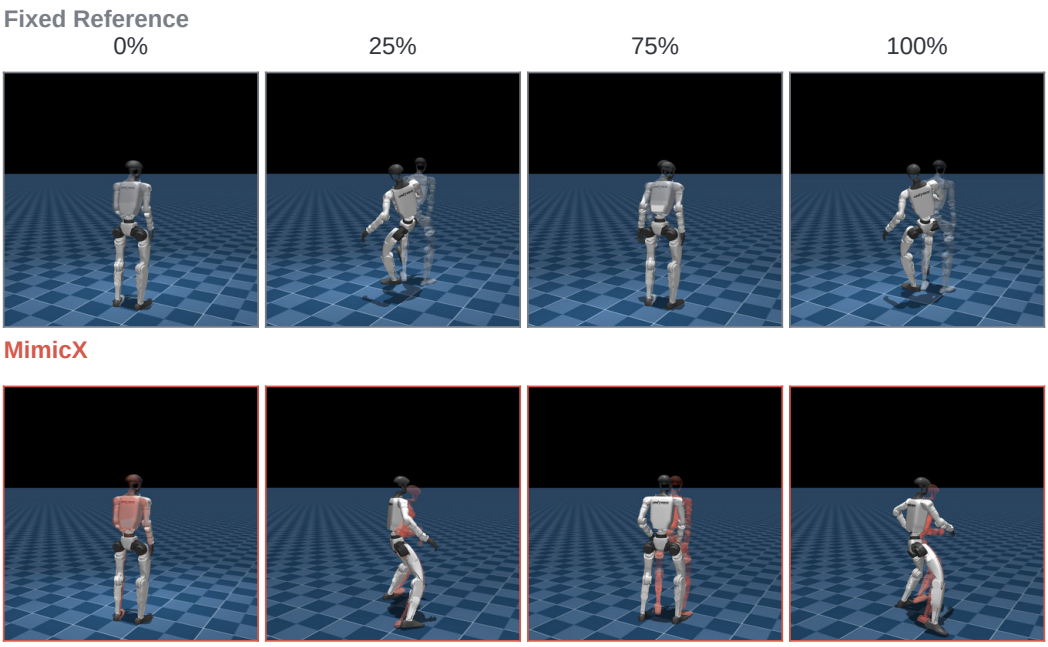


Figure 26: AMASS/ACCAD Form 1: new retargeting. A second retargeting of the same human source motion, evaluated through the shared policy-refinement interface. The two plates compare executions from different robot references derived from this common human motion.