TrackerBench: Stress-Testing Released Humanoid Motion-Tracking Policies

Probe findings — working notes, updated 2026-07-05

Shubham Singh

July 5, 2026

Status: exploratory research in progress. These are working notes from a pre-registered two-night probe. All thresholds were frozen *before* data collection (`trackerbench/KILL_CRITERIA.md`); instrument amendments were logged before the affected data existed (`RUNNING_LOG.md`). Results are simulation-only (MuJoCo sim2sim), two policy families, and should be read as an existence proof — not yet a benchmark release.

1. The question

Humanoid whole-body motion-tracking policies (GMT, TWIST, SONIC, BeyondMimic, –) are ranked by **nominal tracking precision** — mean per-joint position error (MPJPE) and success rate under ideal conditions. Nothing in the literature measures how these released policies degrade under the disturbances any deployment actually faces: shoves, added payload, sensor noise, actuation latency.

Tested claim: the nominal ranking does not predict the ranking under stress — there exist significant, protocol-stable *rank crossovers*, so robustness is an independent evaluation axis and a neutral stress benchmark carries real information.

Pre-registered kill conditions: (i) *KILL-NoReordering* — stress rankings equal nominal rankings everywhere; (ii) *KILL-ProtocolArtifact* — crossovers exist but flip between termination rules, vanish under own-nominal normalization, or fail seed test–retest ≥ 0.8.

2. Protocol

- **Robot:** Unitree G1 humanoid in MuJoCo (CPU), each policy on its **authors' own released sim2sim deploy stack** (model XML, PD gains, observation pipeline copied verbatim; policy is a black box).
- **Policies (so far): GMT** (arXiv 2506.14770, general tracker, authors' TorchScript checkpoint) and **TWIST** (CoRL 2025, general teleoperation tracker, authors' in-repo TorchScript checkpoint). BeyondMimic has no public checkpoint (Isaac-only pipeline); OpenTrack is the next re-host candidate.

- **Behaviors:** GMT's 8 released clips, consumed by both policies (23-DoF retargeted mocap): basic_walk, walk_stand, squat, crouchwalk_stand, dance, dance_waltz, kick_walk, airkick_stand.
- **Stress axes × severities (frozen):** horizontal pelvis **pushes** {50–400 N, 0.2 s every 2 s, seeded random directions}; torso **payload** {2–24 kg}; proprioceptive **observation noise** {1×–16× a base sigma vector}; **action latency** {20–100 ms}.
- **Pairing:** perturbation randomness is seeded by `(clip, axis, severity, seed)` only — both policies receive *identical* push directions and noise draws in every cell. 3 seeds per cell; 560 rollouts per policy.
- **Termination rules (both, post-hoc, one rollout): Rule A** — fall (pelvis < 0.35 m); **Rule B** — fall *or* root-frame mean joint error > 0.35 m ("diverged but upright").
- **Metric:** survival fraction $S \in [0,1]$ (time-to-failure / clip duration), degradation measured against each policy's **own** re-hosted nominal (controls the re-host-gap confound).

3. Instrument validation (all pre-registered gates passed)

- **G-Sanity:** nominal re-hosts reproduce the prior recorded runs to **d = 0.000 cm** on all 8 clips (deterministic); TWIST nominal is clean on 8/8 clips, median local MPJPE 2.79 cm.
- **G-AxisAlive:** after one pre-declared severity recalibration (doubling all-pass grids), all 4 axes exhibit graded failure for GMT: cliffs at ~300 N push, 12$\rightarrow$ $16kgpayload, 8\times\$\$ \rightarrow 16\times\$noise, 60\$ \rightarrow \$100mslatency$.
- **G-SeedStable:** split-half Spearman 0.858 over 64 mid-range cells.
- **One amendment (logged pre-data):** world-frame root error is meaningless for heading-free trackers (GMT's observation discards yaw, so it tracks in its own heading frame); Rule B was re-based on root-frame local error before any grid data existed.

4. Findings

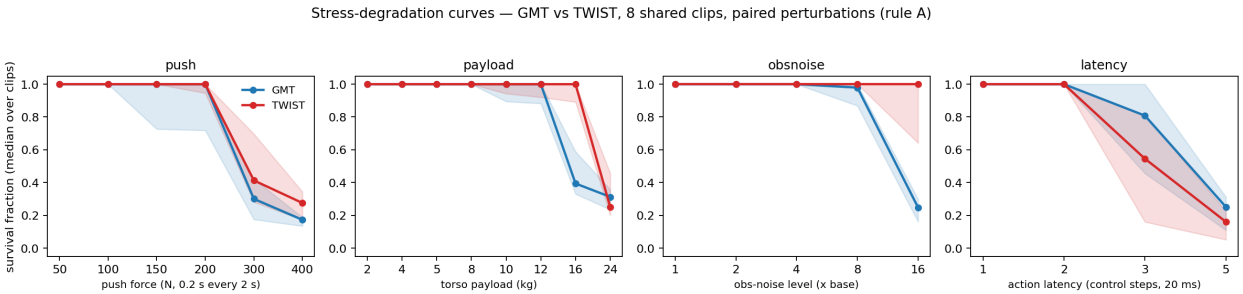

Figure 1: Stress-degradation curves, GMT vs TWIST, median over the 8 shared clips (rule A; bands = IQR). Payload and observation noise show the rank crossover; latency shows the reverse ordering.

F1 — GMT wins the nominal ranking on every clip. Nominal MPJPE 1.39–2.85 cm (GMT) vs 1.67–4.12 cm (TWIST); GMT is the better tracker on 8/8 clips (Fig. 2).

F2 — the ranking inverts under stress, significantly and stably. 33 cells show TWIST — the nominally *worse* policy — surviving significantly better (paired per-seed gap $> 2\sigma$), and the

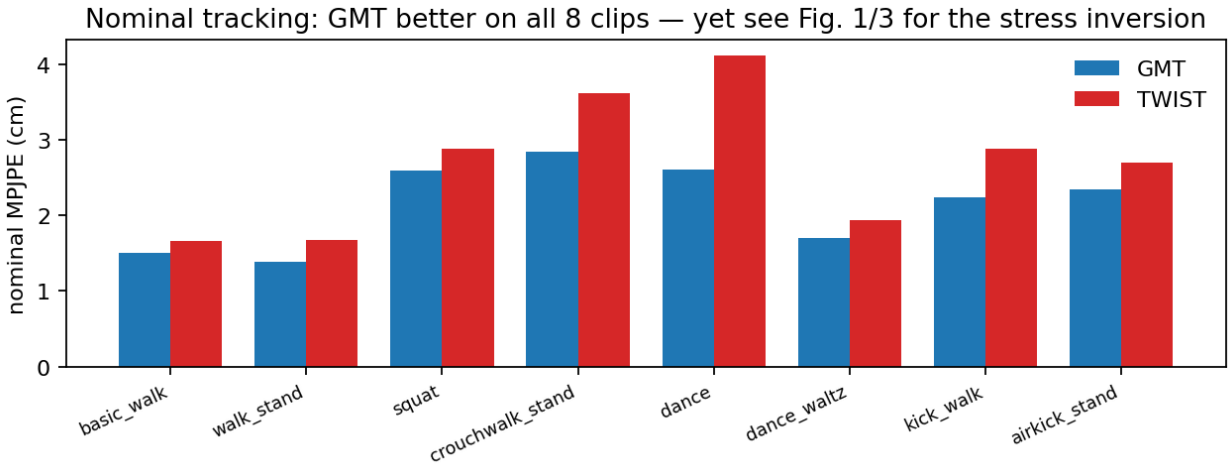

Figure 2: Nominal tracking precision: GMT is the better tracker on all 8 clips.

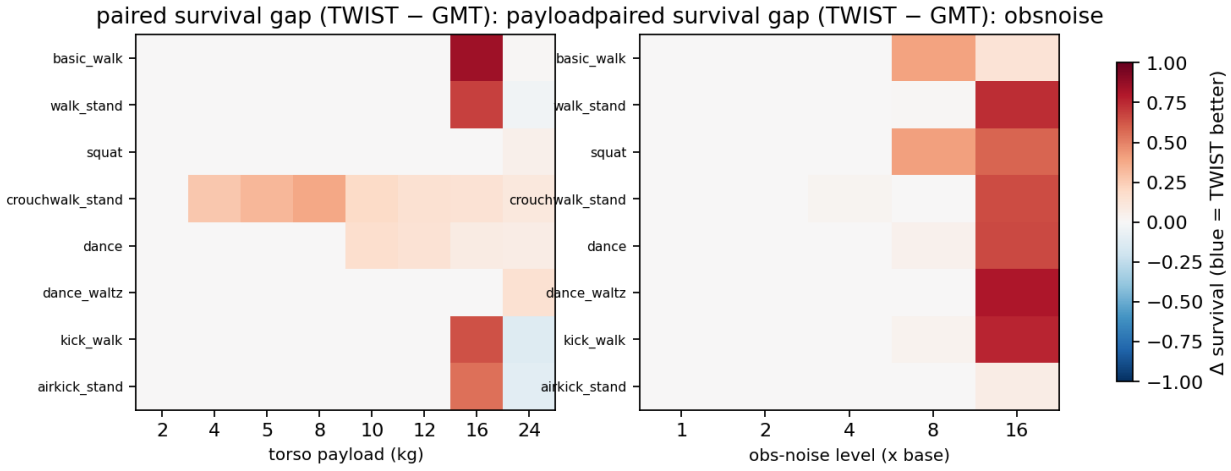

Figure 3: Paired per-cell survival gaps (TWIST - GMT, median over 3 paired seeds). Blue = the nominally worse policy (TWIST) survives longer.

same 33 cells cross under **both** termination rules (zero rule-flips). Seed test–retest sign-agreement 0.883 (gate: ≥ 0.8). Largest gaps: basic_walk @ 16 kg payload — GMT S = 0.15, TWIST S = 1.00, seed $\sigma = 0$; walk_stand @ 16× noise — gap +0.74; dance_waltz @ 300 N push — gap +0.58 (Figs. 1, 3).

F3 — vulnerability fingerprints are axis-specific, not scalar. 14 *reverse* cells exist where GMT out-survives TWIST (notably latency at 60 ms and several push cells) — so neither policy is uniformly tougher. A single "robustness score" would hide exactly the structure that matters; per-axis degradation curves are the right release artifact.

F4 — released trackers are far more robust than their nominal numbers suggest. GMT absorbs 200 N shoves, 12 kg payload, and 8× sensor noise without a single failure across all clips and seeds. The interesting regime — and the discriminative one — starts beyond severities most papers never test.

Verdict per the frozen decision table: SURVIVE. Both kill conditions failed to fire; the benchmark premise stands on its first two families.

5. Threats to validity (open, honest)

1. **Two families.** The crossover exists for GMT-vs-TWIST; the benchmark's value scales with families (target: 4–6). OpenTrack is next; PBHC joins as a per-motion specialist family.
2. **Sim2sim only.** No hardware anchor yet; the claim is about evaluation protocol, not transfer.
3. **Author-shipped clips.** Both policies are evaluated on GMT's released motions — favorable terrain for GMT (which makes the crossover *conservative*, but a neutral clip set is still needed).
4. **Two of five planned axes remain:** terrain irregularity and reference corruption are designed but not yet frozen/run.

6. Roadmap

Re-host OpenTrack (+ PBHC specialist track) → freeze terrain + reference-corruption axes → neutral clip set → 4–6-family release with per-axis degradation curves + vulnerability fingerprints as a living leaderboard (mjlab integration) → hardware spot-check on a real G1.

Method note: experiments run under a pre-registered killshot protocol (frozen kill criteria, paired seeds, post-hoc dual termination rules), executed with an automated harness; ~1,120 rollouts, CPU-only, one night.