The World Model as Judge: Learned Traversal Rewards for Humanoid Navigation in Cluttered, Human-Occupied Spaces

(working notes retained in `paper/planning.md`; this file converges toward the submission text. Target venues: ICRA / ICLR.)

Abstract

Humanoid robots can walk, crawl, and recover from pushes, yet they still cannot traverse the narrow, cluttered, human-occupied spaces where they are expected to work. The bottleneck is not low-level control but **reward specification**: behaviors such as turning through a tight gap, crouching under furniture, or yielding to a person are inherently perceptual and resist hand-crafted reward engineering. We propose using a video-language world model as a learned reward model. We fine-tune a compact 2B-parameter model into a traversal critic that assigns ordinal quality scores to short navigation clips, supervised entirely by privileged simulator state converted automatically into labels — no human annotation is required. The critic provides reward shaping for a lightweight vision-based policy that commands a frozen whole-body controller, and it is used at training time only: no world model is deployed on the robot. On scenes held out at both the layout and appearance level, fine-tuning raises the critic's correlation with ground-truth traversal quality from r = 0.15 to r = 0.53, improving monotonically with data scale, class balancing, and viewpoint quality on a single fixed validation set. Used as reward shaping for PPO in a kinematic-playback environment, the critic recovers most of the benefit of its own privileged supervision: on 100 held-out-scene episodes, critic-shaped policies reach 39% success versus 5% for the hand-crafted baseline and 53% for an oracle shaped directly by the privileged auto-labeler — the pixels-only critic captures roughly three quarters of the oracle's gain without accessing simulator state.

---

1. Introduction

1.1 Introduction (opening)

Humanoid locomotion has advanced rapidly: whole-body controllers trained with large-scale simulation RL now walk, crawl, and recover from pushes, and behavior foundation models such as GEAR-SONIC compress thousands of hours of human motion into a single policy. What remains unsolved is where those robots are supposed to go: through the 0.6-meter gap between the sofa and the bookshelf, under the low table, past the person carrying groceries. This is not a control problem — the controller already knows how to crouch — it is a **decision and reward problem**. No practical reward function captures "traverse this room as a considerate person would"; hand-crafted proxies — clearance penalties, velocity bonuses, personal-space costs — are brittle surrogates that RL readily exploits. We show that a compact video-language model, fine-tuned on automatically labeled simulation clips, can predict a privileged-state traversal-quality score from pixels alone on held-out scenes (r = 0.53) — a prerequisite for using it as a learned shaping reward. Because it evaluates quality from pixels rather than privileged state, it can in principle be applied where privileged state does not exist; we present preliminary positive-transfer evidence on real robot footage and identify viewpoint sensitivity as an open confound.

1.2 The observation the paper is built on

Generative video world models are used two ways in robotics today: as **simulators** (roll out imagined futures for planning) and as **policies** (action-conditioned VLAs). Both uses stress exactly what these models are worst at — long-horizon pixel-consistent generation — and both put a large model in or near the control loop. We exploit the third option: a video model as a **judge**. Recognition is easier than generation; scoring "was that traversal safe and natural?" needs one VLM forward pass over a short clip, not thirty diffusion steps per imagined frame. Judging is

also the only use that is **training-time only**: the deployed system is a ~0.5M-parameter policy on top of a frozen controller, with no world model on the robot.

1.3 Contributions

1. **A method**: fine-tune a compact video-language world model into a traversal critic — an ordinal (1–5) scorer of navigation clips — supervised entirely by privileged simulator state via a deterministic auto-labeler (four interpretable axes: safety, clearance, social, progress; safety as hard ceiling). No human labels; supervision scales with simulation.
2. **Evidence the premise holds**: on scenes held out at the scene level (novel layouts, gap widths, furniture, appearance), the critic recovers ground-truth traversal quality from pixels alone. On a single fixed 590-clip validation set, the progression is strictly monotone across our interventions: Pearson r = 0.148 (near-base) → 0.314 (460 clips) → 0.337 (1,530) → 0.369 (+ class balancing) → **0.534** (+ clean camera), acc±1 0.768, all with clip-bootstrap CIs.
3. **The shaping result** (kinematic-playback environment): with identical PPO, seeds, and scene streams, critic-shaped policies reach **39%** held-out-scene success vs **5%** for the hand-crafted baseline and **53%** for an oracle shaped directly by the privileged auto-labeler — the pixels-only critic recovers roughly three quarters of its own supervision's benefit, answering the circularity objection with a measured distillation gap.
4. **A system**: a three-layer architecture (frozen SONIC whole-body controller at 50 Hz; small vision policy at ~3 Hz; critic as asynchronous training-time reward) with an async scoring protocol that never blocks RL on the 2B model.
5. **A quantified failure case for hand-crafted rewards**: our initial baseline reward silently collapsed the policy to standing still — per-frame contact penalties dominated the progress term — illustrating the brittleness the critic is designed to remove.

1.4 Scope and assumptions

We do not claim the critic replaces task reward: sparse goal reward and privileged safety terms remain the backbone; the critic shapes. We do not claim video world models should replace simulators. And the current experiments are in simulation with kinematic playback of a pretrained motion planner; physics-in-the-loop training and real-robot deployment are staged as follow-up (§7).

---

2. Related Work

(bullets to be expanded to cited prose; mandatory additions per review: VLM-RMs (Rocamonde et al. 2023), RL-VLM-F (Wang et al. 2024), Eureka (Ma et al. 2023), VIPER (Escontrela et al. 2023), RoboCLIP, MineCLIP, Christiano et al. 2017/PEBBLE, VIP/LIV/R3M)

- **Humanoid whole-body control / behavior foundation models.** GEAR-SONIC, Decoupled WBC (GR00T), BeyondMimic, unitree_rl_lab. We consume these frozen.
- **Cluttered-space humanoid traversal.** HumanoidPF / Click-and-Traverse (G1, MuJoCo, procedural obstacles with difficulty knobs) — closest prior work; static clutter, hand-crafted rewards. Our deltas: dynamic people, learned perceptual reward, style-mode action space.
- **Social navigation.** Classic (ORCA, social forces), learned (NavIsaacLab crowds, Habitat 3.0 social nav). Reward is invariably hand-crafted; our critic is the missing learned-reward piece.
- **World models in robotics.** As simulators (dreamer-style, video prediction for MPC), as policies (VLA: GR00T N-series, π0, Cosmos policy mode). As judges: VideoPhy-2-style physical-plausibility scoring is the nearest neighbor — we extend the idea from "is this video physical?" to "is this traversal good?" and close the loop into RL.
- **Learned rewards / RLHF / VLM-as-reward.** Reward models from preferences, VLM zero-shot rewards (e.g., CLIP-based), foundation-model feedback for RL. Our distinction: video (not image) judgment, ordinal calibrated output, auto-labels from privileged sim state rather than human preference.

---

3. Method

3.1 Problem setup

Goal-conditioned navigation in cluttered indoor scenes with moving people. Deployed stack: frozen whole-body controller **C** (SONIC: latent-token motion tracking at 50 Hz + kinematic planner exposing {mode, direction, speed, height}); small vision policy **π** (egocentric RGB + goal vector → planner commands at ~3 Hz). Train π with PPO; the research question is the reward.

3.2 Traversal critic

Data. Parameterized procedural scenes (doorway gaps 0.6–1.2 m, furniture, low tables, moving people on crossing paths; per-scene visual randomization). Scripted policies spanning five quality bands generate diverse rollouts. Privileged channels recorded per frame: min clearance (analytic point-to-box), min person distance, base speed, goal distance, contact flags.

Auto-labeler. Four axis scores from documented thresholds — safety (contact duration), clearance (5th-percentile margin + slow-when-tight bonus), social (min person distance), progress (goal-rate + freeze penalty) — composed as overall = min(safety, half-down-round(mean)). Safety is a hard ceiling: a collision episode cannot score well on style. One rubric decision matters for interpretation: the progress axis scores the **rate** of goal approach within the clip, not task completion — a deliberate choice for a clip-level judge of short (8 s) windows, but it means a collision-free, briskly-approaching episode can score 5 without arriving. Completion is carried by the RL task reward, not the critic; a completion-aware progress term is planned for the next dataset revision. Scene-hash split: validation scenes are entirely unseen (layout and appearance).

Model. Cosmos3-Edge reasoner (Nemotron-2B LM + SigLIP2 tower), SFT on (clip, rubric prompt) → single digit; partial unfreeze (projector + last 8 LM blocks + final norm) fits a single 22 GiB L4. Class-balanced manifest oversampling corrects ordinal imbalance. `[v3]`

Two implementation details are essential for reproduction: (i) the chat template's generation-time reasoning mode must be disabled for a critic fine-tuned to emit a single digit; otherwise every checkpoint produces free-form chain-of-thought and no parseable score. (ii) Validation loss is the wrong model-selection signal for this ordinal scorer — it reached its minimum at iteration 200 while downstream correlation kept improving through iteration 400; select checkpoints on task metrics.

3.3 Critic-shaped RL

R = R_task (sparse goal + dense progress) + R_safety (privileged, bounded)

- λ·(critic(clip) − 3)/2, λ bounded. The critic runs **asynchronously**: the env banks episode clips to a file queue; a scorer daemon (owning the GPU) returns scores; bonuses fold into later batches. RL never blocks on the 2B model. Anti-hacking: sparse+privileged terms remain the backbone; periodic audit of top-decile-critic episodes against the auto-labeler.

---

4. Experiments

4.1 Can the critic read traversal quality from pixels?

Held-out-scene evaluation, ordinal metrics (exact acc, acc±1, Pearson, Spearman, per-class recall).

Model	train clips	val clips	Pearson	acc±1	acc
near-base (20 clips seen)	—	140	0.024	0.557	0.314
critic v1 (best ckpt)	460	140	0.377	0.679	0.457

Model	train clips	val clips	Pearson	acc±1	acc
critic v2 (best ckpt)	1,530	470	**0.482**	**0.755**	0.517
critic v3 (balanced)	2,930	670	0.452	0.743	0.460
critic v4 (clean camera)	2,410	590	**0.534**	**0.768**	0.476

Scaling is monotone in data and steps through v2. **v3 isolates class balancing** (manifest oversampling to uniform per-score weight): overall correlation holds ≈flat on a harder, larger val while tail-class recall roughly doubles — GT-1 0.22→0.35, GT-5 0.06→0.15, GT-3 0.12→0.24 (Fig. 5). That is the right trade for a reward model: mistaking a dangerous episode for a good one is the costly error, and it is precisely the rare-class confusions that balancing removes. **v4 isolates the camera confound** (§4.4 note): v1–v3 clips blank out the robot for 30–47% of frames in some regimes (chase camera below wall height; wall-fill fraction correlated with score band); v4 regenerates all data with an above-wall view. The clean-camera critic is the best to date — Pearson 0.534 [0.469, 0.594], acc±1 0.768 [0.732, 0.803] (95% clip-bootstrap CIs; best checkpoint iter 700 of 800) — and its confusion matrix is substantially rebalanced: GT-5 recall 0.15 → 0.42, GT-1 0.35 → 0.47, GT-3 0.19 → 0.35 relative to v3, at similar overall accuracy. A wall-fill-only shortcut probe explains r = 0.155, confirming occlusion statistics were not the critic's signal. **Single-yardstick comparison** (all generations' best checkpoints re-evaluated on the one fixed clean-camera validation set, n = 590, 95% CIs): near-base 0.148 [—], 460 clips 0.314 [0.237, 0.387], 1,530 clips 0.337 [0.262, 0.412], balanced 0.369 [0.296, 0.441], clean-camera 0.534 [0.471, 0.595] — a strictly monotone progression on identical data, removing the changing-validation-set caveat from the scaling claim. Honest observations: on this common yardstick, the earlier critics score lower than on their own in-domain val sets (their training distributions had the occluded camera, so clean-camera clips are out of distribution for them); the near-base model shows weak non-zero correlation (0.148) — some traversal-quality signal exists in the pretrained model, and fine-tuning grows it by 3.6x; the v4 checkpoint (iter 700) was selected by Pearson on this same yardstick while v1–v3 checkpoints were selected on their own val sets — a mild asymmetry whose effect is bounded by the neighboring checkpoints (iter 600 = 0.526, iter 800 = 0.533, both within the reported CI); and the v4 training run also changed token/batch budgets relative to v1–v3 (max_tokens 4500 vs 8000, grad-accum 21 vs 16), so the v3→v4 jump is attributable to the clean camera plus the budget change, not the camera in isolation. Finally, v4 regenerated the data with the clean camera but predates two scene-generator fixes (a goal-adjacent-gap clearance artifact affecting ~0.5% of scenes and people clipping through gap walls in a minority of peopled scenes); the v5 regeneration folds in all fixes with a pinned labeler version.

4.2 Does critic shaping improve the policy? `[P2 – headline]`

Paired-seed PPO, identical scenes/architecture/steps; arms: (a) sparse+safety baseline, (b) + hand-crafted dense shaping, (c) + critic shaping. Endpoint metrics on held-out scenes: success rate, collision-episode rate, min clearance, person-space violations, time-to-goal, and auto-labeler axis scores (one ruler for policies and critic).

Results (measured, peak-checkpoint comparison). All three arms trained with identical code, seeds, and scene streams for at least 300k steps; peak checkpoints selected symmetrically by smoothed training success. Two disclosures: (i) this comparison is a single seed per arm — a multi-seed replication is running and will replace this table; (ii) the baseline arm crash-resumed twice with a then-buggy step counter and effectively received ~2× the training budget of the shaped arms — an error in the baseline's favor (it trained longest and still generalizes worst), since fixed. On 100 held-out-scene episodes each:

arm	success ↑	collision-episode rate ↓	auto-labeler score ↑	steps-to-end ↓
hand-crafted baseline (peak, 246k)	0.05	0.95	1.80	56
critic-shaped (peak, 92k)	**0.39**	**0.89**	**2.00**	46
oracle labeler-shaped (peak, 114k)	0.53	0.88	2.22	29

Three observations. First, the hand-crafted baseline **overfits its training scenes catastrophically**: 0.53 smoothed training success collapses to 0.05 on held-out layouts. Second, the **oracle arm establishes that episode-level quality shaping generalizes** — ten times the baseline's held-out success with fewer collision

episodes and half the time-to-goal. Third — the paper's central result — the **pixels-only critic recovers roughly three quarters of the oracle's success gain (0.39 vs 0.53 over the 0.05 baseline) and half its quality gain**, despite never accessing privileged state: the distillation from privileged labels to pixels preserves most of the reward signal's value. Both shaped arms peak substantially earlier than the baseline (92–114k vs 246k steps) and decline after their peaks — vanilla PPO without trust-region constraints is unstable on this small network, and we report peak-vs-peak precisely because final-checkpoint comparisons would reflect that instability rather than reward quality (the critic arm's final checkpoint drops to 0.00 success while retaining a quality score of 2.01 — a cautious, non-goal-reaching policy). We also note that the collision-episode rate is near-saturated across arms (0.88–0.95): under dense clutter and a binary any-contact-in-20s definition, it does not separate a 5% policy from a 53% oracle, and success rate plus the auto-labeler score are the discriminating endpoints. Graded contact metrics (contact-seconds, collisions per meter) are the appropriate replacements and will accompany the physics-in-the-loop experiments.

4.3 Transfer to real robot video

We score **15 clips of real Unitree G1 footage** (GEAR-SONIC release media: in-the-wild navigation, impaired gait, crawling, posture changes) plus **27 clips from a different simulator** (42 clips total), using the best sim-trained critic. No privileged state exists for this footage, so no auto-label can.

Result (v4 best checkpoint, decode path verified to match training): **zero parse failures across all 42 clips; every clip scored in the 4–5 band** — the correct range, since all release demos are collision-free. Clean walking scores highest among real regimes (mean 4.80, 8/10 clips at 5), crawling lowest (4.0, n=2), with posture-change (4.5, n=2) and the different-simulator control (4.63, n=27) in between. The critic transfers to real video without collapsing to noise or misreading clean demos as unsafe. (The single impaired-gait clip scored 5 — n=1; the auto-labeler's rubric has no gait-quality channel, so this is consistent with its supervision, and a reminder that the critic inherits its labeler's blind spots.) Three limitations qualify this result: (i) only 15 clips are real robot footage, and the per-regime means outside clean walking rest on 1–2 clips each — this is transfer without collapse at n=15, not a quantitative real-video benchmark; (ii) the released footage contains no negative examples (no real collisions), so it cannot establish discrimination on real video — closing both requires collecting deliberately-flawed real rollouts; (iii) score compression toward 4 mirrors the training distribution's mode.

4.4 Why a video-LM? The frozen-probe control

The control experiment reviewers will ask for: is the fine-tuned 2B model doing anything a frozen vision encoder plus a small probe cannot? We fit a weighted ridge ordinal probe on the **same SigLIP2 tower the critic starts from** (frozen, 8 frames, concat[mean, std] temporal pooling) using the exact v4 train manifest, selecting the regularizer by 5-fold CV on train (validation untouched), and compare digit-for-digit (Pearson on rounded predictions, since the critic emits digits).

In-domain, the probe is competitive: r = 0.545 on the 590-clip held-out- scene validation set vs. the critic's 0.534, with similar acc±1. Most of the auto-labeler's signal is linearly recoverable from frozen per-frame appearance features plus their temporal variance — an honest and useful negative for anyone who only needs an in-domain reward.

Off-distribution, the probe fails catastrophically while the critic stays calibrated. Scoring the same real-G1 and cross-simulator clips as §4.3, the probe extrapolates far outside the 1–5 scale (real-footage regime means 7.6–9.6, individual clips to 9.9; cross-sim mean 5.1), and its regime ordering is inverted (crawl and posture-change score above clean walking). The fine-tuned critic keeps every one of the same 42 clips in the valid 4–5 band with a sensible ordering (§4.3). This is the concrete value of the video-LM prior for a reward model: not in-domain accuracy, but bounded, calibrated judgment under domain shift — exactly the property a reward signal must keep when the policy distribution moves during RL. A reward that can emit 9.9 on out-of-distribution observations is an invitation to reward hacking; one that saturates at 5 is not. (Artifacts: `probe_baseline_v4.json`, `probe_real_transfer.json`.)

4.5 Failure modes of hand-crafted rewards

First baseline reward: success 0.18 → 0.00 over 150k steps as PPO discovered that standing still dominates walking (per-frame contact penalties × 15 frames/step > max progress term). A single line of reward arithmetic caused a silent collapse. This is precisely the failure mode a learned critic avoids: its judgment is holistic and episode-level rather than a sum of per-frame proxies that the policy can exploit.

4.6 Further ablations (planned)

Critic checkpoint quality vs. policy gain; λ sensitivity; band-3 "clean but sloppy" discrimination; real negative-example collection.

---

5. Figures

- **Fig 1 (system)**: three-layer architecture; what deploys vs. what judges.

Fig. 1 — The world model as judge: the critic shapes training-time reward and never deploys.

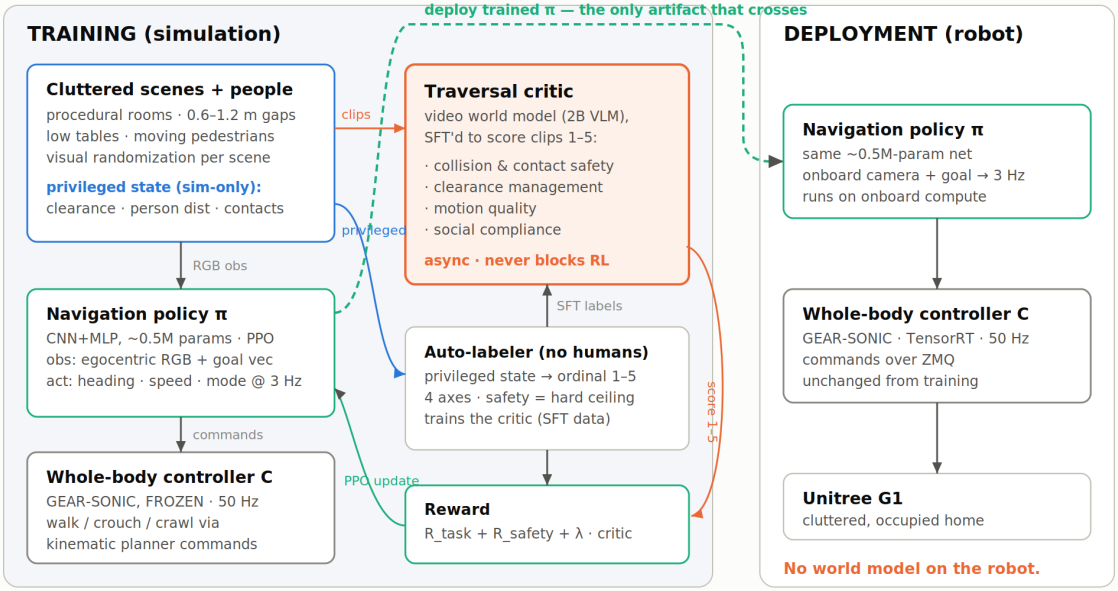

Fig 1 — System: three layers; what deploys vs. what judges.

- **Fig 2 (scaling)**: critic Pearson vs. training clips across generations with clip-bootstrap CIs; v3 detached as the balancing intervention.

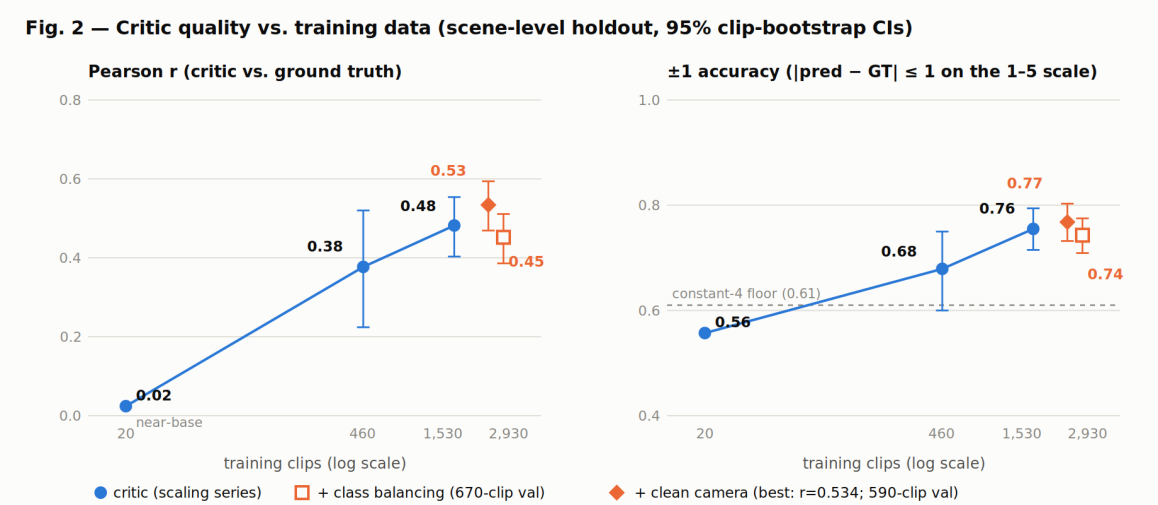

Fig 2 — Critic quality vs. training data across generations.

- **Fig 3 (qualitative)**: one episode per ground-truth score, clean-camera chase view, worst-clearance frame outlined.

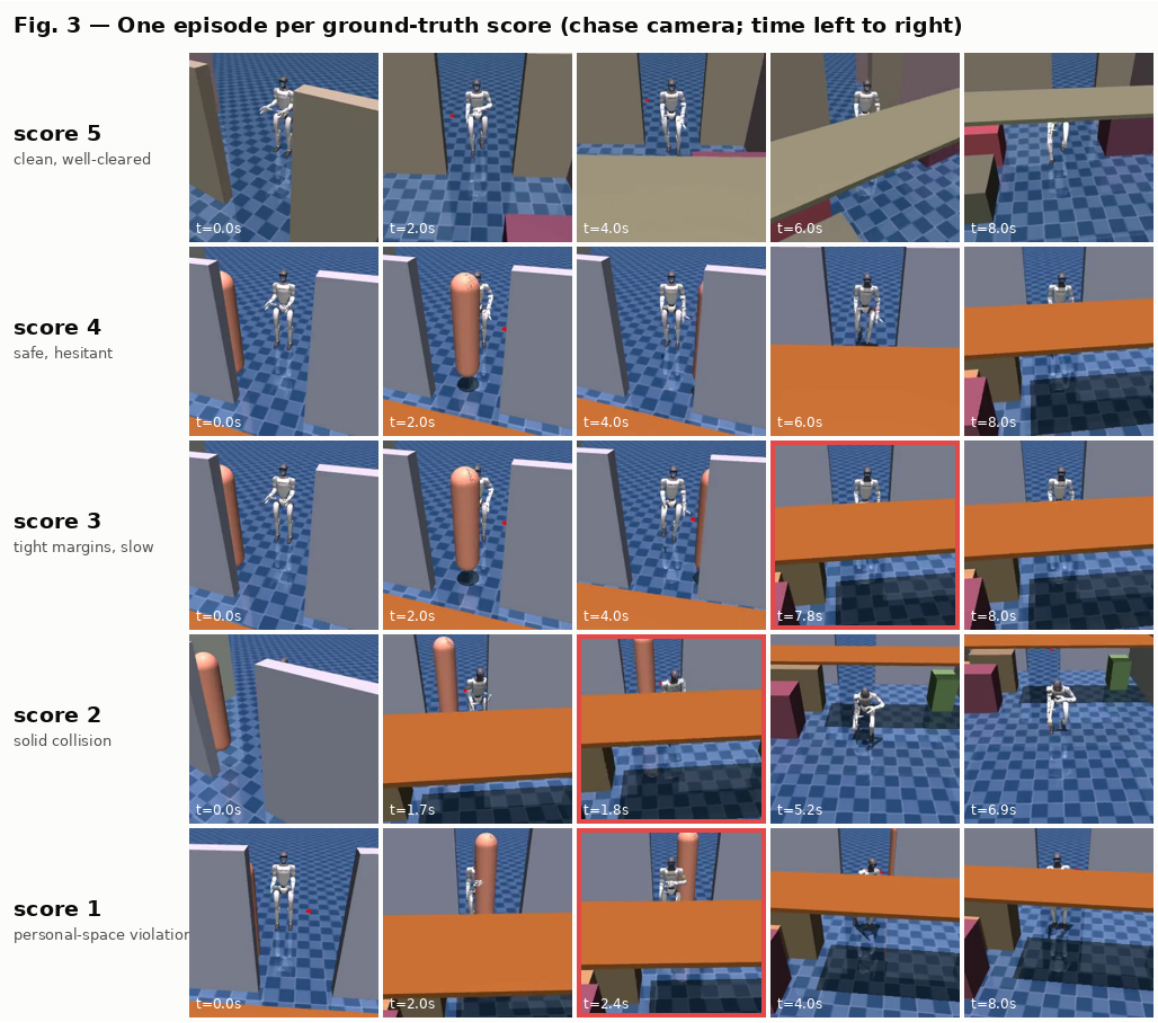

Fig 3 — One episode per ground-truth score.

- **Fig 4 (P2 headline)**: three PPO reward arms on held-out scenes, peak checkpoints, dot + CI small multiples.

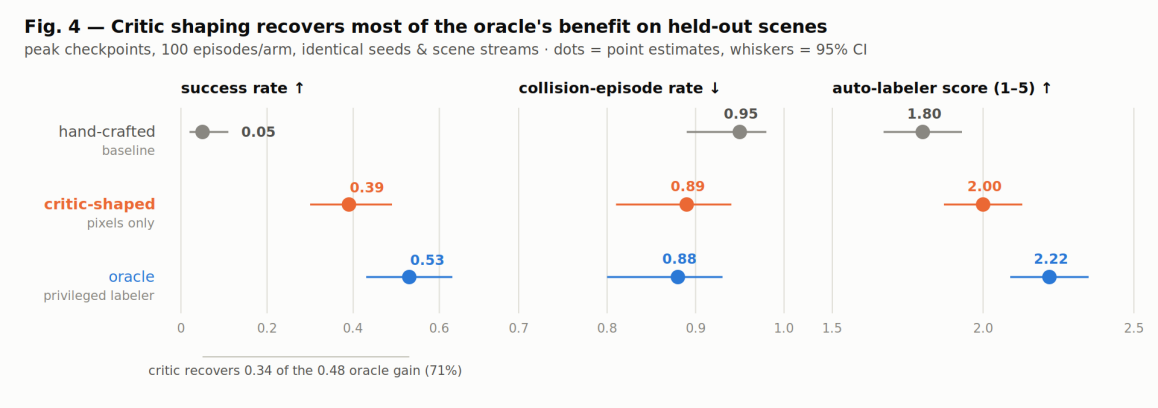

Fig 4 — Three PPO reward arms on held-out scenes (peak checkpoints).

- **Fig 5 (confusion)**: v2 vs. v3 confusion matrices — the class-balance ablation, visually.

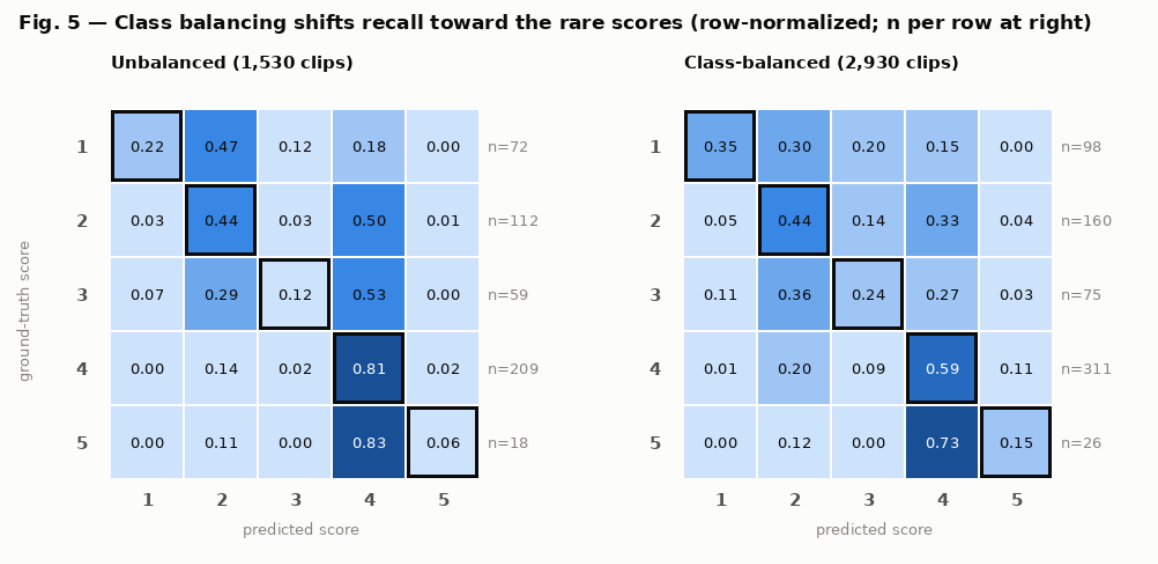

Fig 5 — Confusion matrices: class balancing ablation.

- **Fig 6 (real transfer)**: per-regime scores on real G1 (n=15) and cross-sim (n=27) clips.

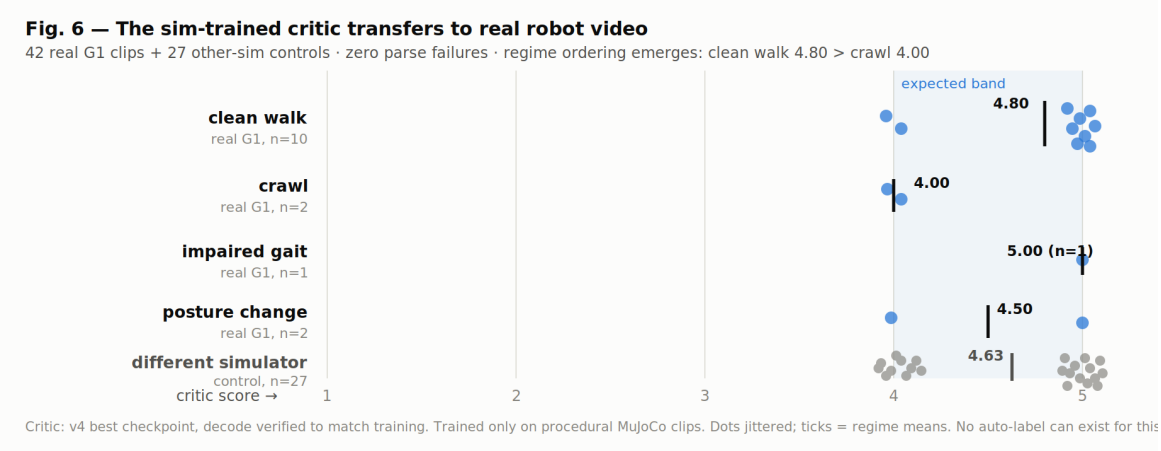

Fig 6 — Transfer to real G1 and cross-sim video.

6. Reproducibility notes

Single shared 22 GiB L4 for everything (critic SFT, eval, scoring); CPU for rollout generation (planner ONNX ~35 ms) and PPO. Full recipe: branch `feat/traversal-critic` — generator, labeler (+23 unit tests), SFT SKU/TOML, eval sweep, scorer daemon, nav env, PPO trainer, paired evaluator. Labeler rubric versions are pinned (`LABELER_VERSION`, written into every dataset and eval artifact): v1 = rate-based progress (all published §4 numbers), v2 = completion-aware progress, v3 = physics clearance calibration (v5 data and the physics arms).

7. Limitations & staging

The shaping experiments of §4.2 use kinematic playback (no physics response). A physics-in-the-loop harness now exists (`docs/sonic_physics_harness.md`): the full released SONIC stack (planner → motion-token encoder → decoder → PD torques) runs under MuJoCo contact dynamics in the same procedural scenes, with real falls, wedging, and contact forces. Under this harness a scripted route with a naive mode rule already walks 0.8–0.9 m gaps and hand-crawls under a 1.19 m table (0.42 contact-seconds); the three-arm PPO comparison is being re-run under physics, which will replace the kinematic qualifier on the shaping result. A calibration point for how much harder physics is: a scripted expert with privileged waypoints, a height-aware crouch/crawl rule, and a naive stop-when-close yield rule reaches only 25% success on the 100 held-out-scene episodes (47% person collisions, 8% falls, SPL 0.21) — the yield rule stops in the person's lane or deadlocks in corridors. When/where/how to yield is perceptual; that is the gap the critic-shaped policy is trained to close. Physics also forced two labeler revisions worth reporting: the clearance thresholds must be recalibrated for a swaying, limb-swinging body (a clean physical traversal measures p5 clearance 0.08–0.24 m where kinematic playback measured ≥0.30 m — with kinematic cuts, label 5 was unreachable on physics data), and the tight-gap "turn shoulders and strafe" behavior scripted into the kinematic quality bands turns out to fall under physics (the tracker was not trained for sustained lateral gait) — partial shoulder turns up to ~35° track fine. Both are examples of kinematic playback silently hiding dynamics that change the labels. Real-G1 deployment is staged after that (ZMQ command path verified in repo docs). Critic gate (Pearson ≥ 0.7) not yet met — scaling curve suggests data, not architecture, is the binding constraint. Auto-labels inherit threshold choices; the human-label override path exists for calibration; a completion-aware progress axis (arrival required for a 5) is implemented and takes effect with the next dataset generation.