HUMANCLAW: CAN VISION-LANGUAGE MODELS ACT THROUGH A BODY?

Anonymous authors
Paper under double-blind review

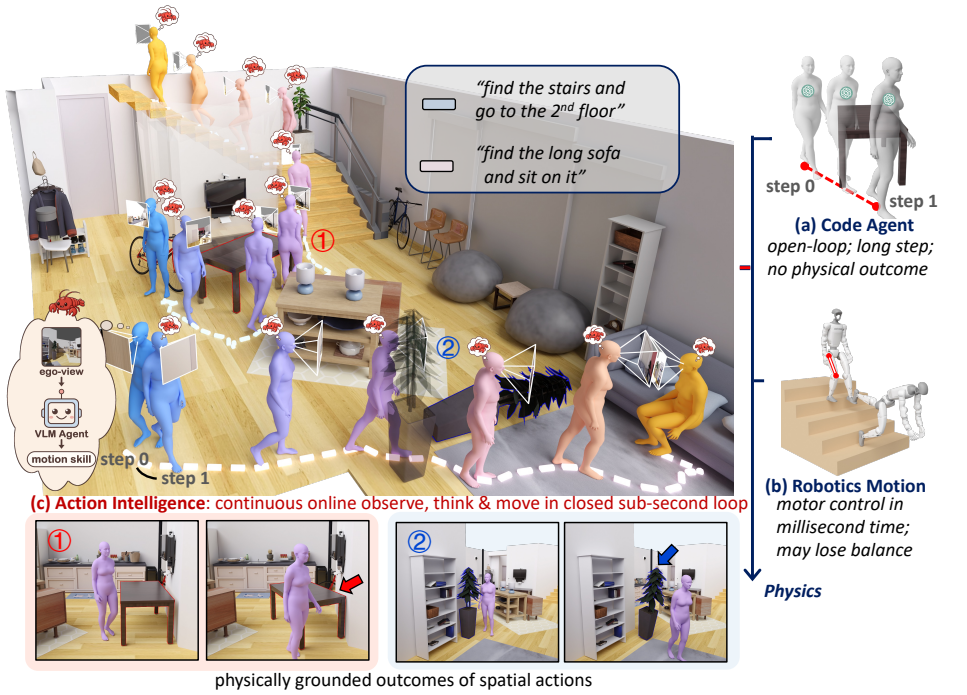

Figure 1: **Where does action intelligence live?** Three regimes of action along increasing physical embodiment: **(a)** action without a body, and **(b)** a body whose action is inseparable from balance and motor control. **(c)** HumanCLAW keeps body and world in a closed sub-second loop but factors out low-level control, so a failed episode reads as a wrong decision, not a lost balance. **Main panel:** one find-navigate-interact episode, each posed body a sub-second decision. **Bottom:** the physical outcomes it produces.

ABSTRACT

Evaluating whether a vision-language model (VLM) can act through a physical body is challenging. The outcome of an action couples the VLM's decision with motor control. When a task fails, it is hard to tell whether the VLM made a bad choice or the motor controller simply failed to execute it, e.g., losing balance and falling. In this work, we introduce **HumanCLAW**, an evaluation framework that decouples action decision-making from low-level execution. At every step, a harnessed, off-the-shelf VLM issues an atomic skill command, and the command is translated into a sub-second chunk of continuous full-body motion with real physical consequences, including gravity and collisions. The body can therefore act freely in the physical world, while execution-side disturbances, balance and motor errors, are factored out. What remains measurable is the model's **action intelligence**: its moment-to-moment choice of what the body should execute next. Based on this framework, we build **HumanCLAW-Bench**: 1,218 long-horizon, egocentric `find-navigate-interact` episodes across 41 indoor scenes. We test nine state-of-the-art VLMs and find that none solves the benchmark; the best model reaches only a 16.8% success rate. Recognizing the target is not the bottleneck. What current VLMs lack is **embodied self-awareness**: they lose track of their own body, failing to tell where it is, whether it has reached the goal, or whether it has hit an obstacle. Video demos, the leaderboard, and analysis are available on the anonymous *supplementary webpage*.

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.

1 INTRODUCTION

Recent vision-language models (VLMs) achieve strong performance on visual grounding, spatial relations, and spatial reasoning (Chen et al., 2024; Yang et al., 2025a). If such a model is given a physical body, an egocentric observation, a history of its actions, and a goal, can it determine an *executable action* appropriate for the current moment? A positive answer would be tantalizing: the decision would come from general-purpose reasoning rather than from a policy fitted to a specific embodiment, so an agent built on it could act in scenarios it was never trained on, without collecting action demonstrations for the decision maker. We call this capacity **action intelligence**: the ability to decide, moment by moment, what a physical body should do next inside the *execution loop*, where each decision is executed and its physical outcome informs the next. We view it as the operational component of *spatial intelligence*, converting spatial understanding into closed-loop embodied decisions by selecting, parameterizing, and sequencing actions as their physical consequences unfold.

In this paper, we explore whether recent VLMs have action intelligence. Turning a VLM's decision into executable motion is not trivial. For a simple embodiment the path is short: a 7-DoF end-effector action can be written out as text (Yang et al., 2025b). A human body offers no shortcut: with well over a hundred pose dimensions changing many times per second, full-body motion is far too much to write out at every control step. Existing evaluations fall on either side of this gap. Agent benchmarks abstract the motion away, executing commands through simulator-defined action abstractions, scripts, or pre-animated routines (Puig et al., 2018; Shridhar et al., 2020; Liu et al., 2025; Chang et al., 2025), so the model never has to realize full-body behavior whose physical consequences emerge from the motion itself. End-to-end VLAs do generate executable motion (Zitkovich et al., 2023; Kim et al., 2025; Black et al., 2024; Lee et al., 2025), but fitting on embodiment-specific trajectories entangles the decision with the learned control policy, making it difficult to isolate how far the generalist's reasoning itself extrapolates into physical action. For a humanoid, motor execution further confounds evaluation: the body may lose balance on a staircase, a failure unrelated to the action decision. Measuring action intelligence therefore requires what neither side provides: a setting where the decision is **decoupled** from execution, yet the body still *acts freely* in the physical world (Figure 1).

HumanCLAW closes this gap without training the decision maker. At each sub-second step, a harnessed off-the-shelf VLM takes its egocentric view and text history, reasons explicitly about what it sees, where its body stands, and what to do next, and outputs one *atomic* whole-body skill with parameters, checked by a skill-specific verifier. Atomic is the operative word: a skill is a minimal, unambiguous unit of bodily action—walk at a given stride, turn by a given angle, sit at a given height—never a compositional instruction like *"sit on the sofa,"* which silently bundles finding, approaching, orienting, and committing into one fuzzy command. Keeping the vocabulary atomic pushes this composition into the VLM's own step-by-step reasoning. A skill-conditioned motion generator supplies what the VLM cannot emit, realizing each accepted skill as continuous, hundred-DoF human motion, and a **half-physics** simulator applies its consequences through contact, collision, gravity, and object displacement while factoring out balance and motor-tracking failures. The resulting egocentric view returns for the next decision, closing the loop: the VLM plays a first-person game (Tan et al., 2025; Zhang et al., 2025a) whose actions are whole-body skills and whose outcomes are determined by physics, so every failure can be attributed at the decision level.

On this framework we build **HumanCLAW-Bench**: $1{,}218$ egocentric whole-body `find-navigate-interact` episodes across 41 indoor houses—find a target object, bring the body to it, and sit on it. Each goal specifies what to achieve but leaves open what the body should do at every moment, so success depends on a sequence of sub-second action decisions; the decision maker stays frozen throughout, and the benchmark measures how far internet-learned reasoning extrapolates when each choice becomes physical motion.

Across nine frontier VLMs, none reliably solves the benchmark: the strongest model completes the full progression in only 16.8% of episodes. The results reveal a consistent gap between *seeing* and *acting through a body*. Once a target is genuinely rendered in the egocentric view, the strongest model's reported target-seeing rate comes within 5 points of the objectively measured visibility rate; the failures concentrate in everything after seeing—the agent cannot reliably tell whether it has arrived, where its body is placed relative to the scene, or when its motion has collided with the world. We call this missing capacity **embodied self-awareness**: an online estimate of one's own body state, its relation to the surrounding scene, and the physical consequences of recent actions. Current VLMs describe the world but do not reliably track the body they now control, and with motor execution factored out by construction, these failures belong to the decision maker.

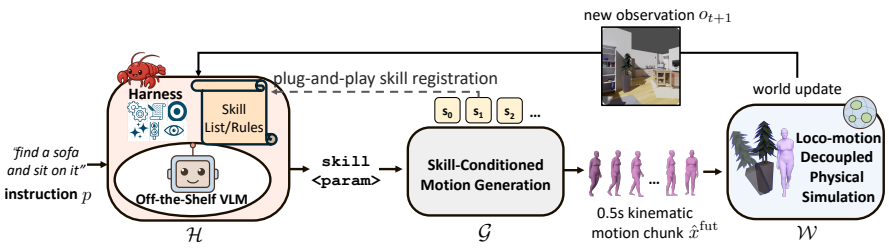

Figure 2: **HumanCLAW closed-loop framework.** A harnessed off-the-shelf VLM maps each egocentric observation to a skill call `skill<param>`; a skill-conditioned generator turns it into a 0.5 s full-body motion chunk, which a half-physics simulator executes under contact and gravity to return the next observation. New skills register plug-and-play, without retraining the VLM.

In summary, our contributions are threefold: (i) **HumanCLAW**, a controlled setting in which an off-the-shelf VLM acts freely in the physical world with the decision decoupled from motor execution, so every failure reads at the decision level; (ii) **HumanCLAW-Bench**, $1{,}218$ egocentric whole-body `find-navigate-interact` episodes across 41 houses with staged, progressive metrics; and (iii) **an extensive evaluation and failure analysis** of nine frontier VLMs, which locates their bottleneck not in perception but in *embodied self-awareness*: knowing where the body is, whether it has arrived, and what its motion has just caused. Related work is discussed in Appendix A.

2 HumanCLAW

Problem setting. HumanCLAW pursues a long-horizon task specified by a single high-level instruction p (e.g., *"find a sofa and sit on it"*) that fixes the goal but leaves every execution detail open, and closes the observe–think–act loop at a sub-second rate. At each step t the agent takes a 2D egocentric RGB observation o_t and a compact textual history h_t of past decisions and feedback, and a skill-decision module built around a harnessed VLM maps them to a single parameterized skill call $\langle s_t, c_t \rangle = H(p, o_t, h_t)$, where s_t is an atomic skill from a fixed set and c_t its continuous parameter. A decoupled motion generator realizes the call as a sub-second, zero-shot-reliable motion chunk $\hat{x}_t^{\text{fut}} = G(x_t, s_t, c_t)$ conditioned on the current body state x_t, and a locomotion-decoupled physical simulator executes the chunk and returns the next observation $o_{t+1} = W(x_t, \hat{x}_t^{\text{fut}})$ (see Figure 2). Rather than committing to a fixed plan, the agent decides what to do next at every step and continually adjusts to the spatial relation between its body and the scene that its own actions reshape.

2.1 Egocentric Action Decision

As shown in Figure 3, HumanCLAW wraps an open-ended VLM in a contextual reasoning harness that infers the motion skill to execute over the next 0.5 s from the egocentric view. Rather than prompting the VLM to predict motion directly, the harness structures the decision in five stages. **Contextual prompting** specifies the agent's embodiment (a blue humanoid with an egocentric camera), the instruction, the available skills and their usage rules, and a compact *textual history* of recent visual states, reasoning traces, and actions, so the model knows what it has been trying to do and how the environment responded rather than treating each image as an independent captioning problem. **Ego-perception** then requires the VLM to output a formatted visual state—visible objects from left to right with approximate distances, whether the target is visible, and nearby obstacles or affordances—which externalizes its spatial interpretation before any decision is made. **High-to-mid-to-low reasoning** interprets the instruction under this state (search, approach, align, or interact) and derives a mid-level objective that may remain valid for several seconds (e.g., move toward a visible object, turn before sitting); at every step the VLM explicitly decides whether to inherit or revise the previous objective, preserving the stability of hierarchical planning while keeping the policy closed-loop. **Low-level skill choice** then converts the objective into one skill from the predefined pool together with its continuous parameters (e.g., `turn` with a degree, `sit` with a height), emitted as structured JSON for the motion generator. Finally, a **skill-specific verifier** re-examines the proposal from the current view alone with a compact, skill-specific prompt—would walking collide with an obstacle, is the target really reached before stopping, is a stair underfoot before climbing—and accepts it or replaces it with a corrected skill from the same pool. The verifier counters a failure mode we observe in long-context control: as the rollout history grows, the planner's spatial estimates drift and it may hallucinate progress. The per-skill checklist is given in Table 9.

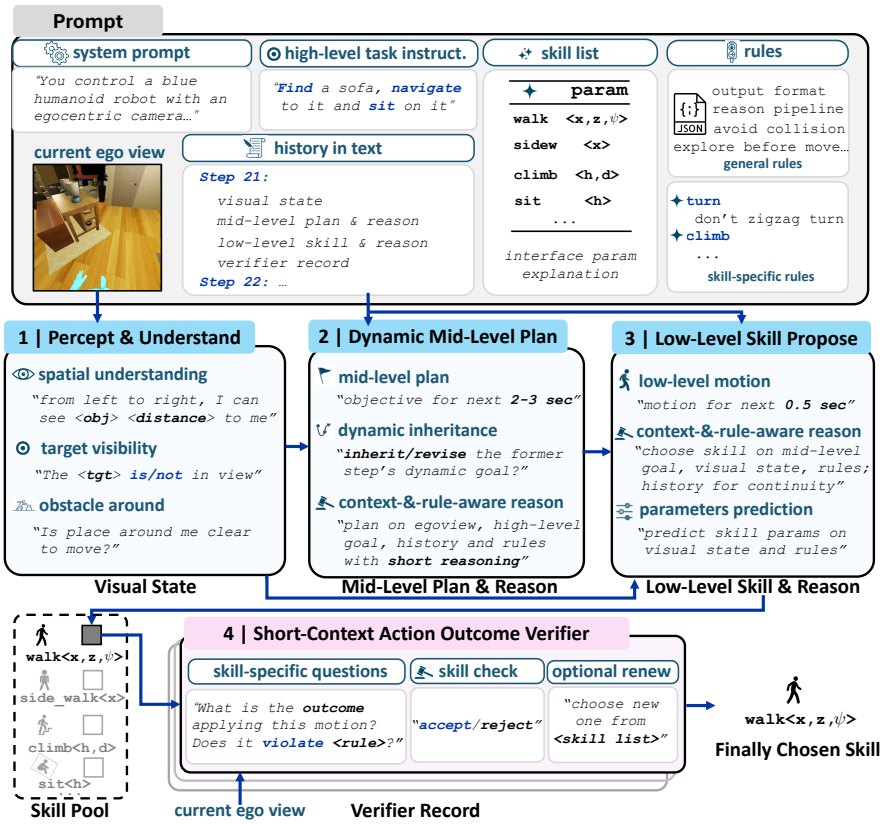

Figure 3: **HumanCLAW harness.** We guide an open-ended VLM from egocentric perception to executable humanoid control through a contextual high-to-mid-to-low reasoning scaffold, followed by a skill-specific verifier that corrects spatially unsafe or premature skill proposals.

2.2 Skill-Conditioned Motion

Real human actions differ from the hard-coded actions of a symbolic agent: a person cannot turn left and immediately right, or reverse direction without decelerating, shifting weight, and recovering balance. These unavoidable transients are part of embodied action intelligence, so decisions should be executed and evaluated under the physical consequences of natural human motion. At the same time, the decoupled design needs a reliable zero-shot interface: if the planner issues `walk`, the generated motion must be a commonsense walk; otherwise failures can no longer be attributed to the decision.

Atomic skills with a deterministic text interface. We therefore define a finite set of atomic, semantically unambiguous skills; the VLM composes them over time, while the generator only realizes the selected primitive. Sitting on a sofa is *not* exposed as `sit(sofa at (x,z))`: the VLM decides whether to walk, turn, step back, or sit from its egocentric state.

$$\texttt{walk}(x,z,\psi),\quad \texttt{side_step}(x),\quad \texttt{step_back}(z),\quad \texttt{turn_in_place}(\theta),$$
$$\texttt{climb_upstairs}(h,d),\quad \texttt{walk_downstairs}(h,d),\quad \texttt{sit_in_place}(h),\quad \texttt{stop}.$$

Here (x,z,ψ) specifies local walking displacement and heading change, θ the signed turning angle, (h,d) stair height and depth, and h the target sitting height. Each skill has a deterministic text/JSON interface, but its execution is generated as continuous human motion rather than retrieved as a fixed animation.

Motion prior and plug-and-play skills. As shown in Figure 4, the generator is a skill-agnostic *Motion Base DiT* trained on AMASS for receding-horizon continuation (Zhang et al., 2025c): given the last five body states it predicts the next fifteen frames at 30 fps (the 0.5 s chunk) with flow matching, conditioning on history by concatenating the clean past frames with the noisy future as motion tokens, which keeps chunk-to-chunk kinematics continuous. Each atomic skill is a ControlNet-style adapter

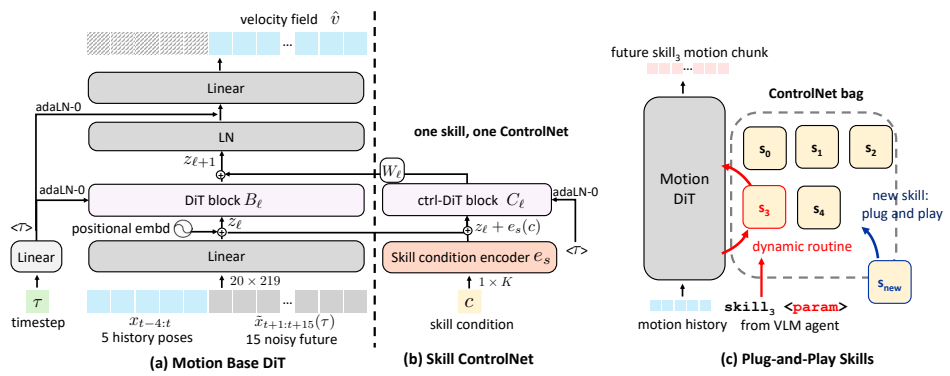

Figure 4: **Skill-conditioned motion generator. (a)** A skill-agnostic Motion Base DiT predicts the velocity field of 15 future frames from 5 history poses via flow matching. **(b)** Each skill is a ControlNet adapter injected through a zero-initialized W_ℓ, leaving the base frozen. **(c)** Adapters form a plug-and-play bag: the VLM-selected skill routes to its own ControlNet, and new skills add one more without touching the rest.

on the frozen base: a trainable copy of the DiT blocks injected through zero-initialized projections and conditioned on the skill parameter through an MLP or Fourier encoder. The parameters are physical quantities readable from the motion chunk itself (final pelvis displacement and yaw for walking, final yaw for turning, step geometry or sitting height for stairs and sitting), so training needs no manual labels beyond curating AMASS clips per skill, which we filter aggressively to keep each skill single-purpose and commonsense-aligned. Adding a skill trains one adapter and touches nothing else. The resulting skills reliably track the commanded magnitudes (Table 4). The architecture, training details, and qualitative controllability results are given in Appendix B.2.

2.3 Locomotion-Decoupled Physical Simulation

Action intelligence must be evaluated through physical outcomes—whether the body passes a narrow gap, knocks over an object, is blocked by a wall, or stops short—yet a fully dynamic humanoid entangles those outcomes with balance and motor failures, while purely kinematic playback removes them entirely (the agent walks *through* walls). HumanCLAW adopts an intermediate ***half-physics*** mechanism (Siyao et al., 2025): the world follows rigid-body physics (collision, friction, gravity, articulated and movable objects), but the humanoid is driven not by joint torques but by equivalent kinematic velocities derived from the generated chunk, computed from consecutive poses and applied as the only actuation signal. The agent thus never fails from balance loss or imperfect motor tracking, yet a wall blocks it, contacts push movable objects, and gravity still constrains unsupported motion. We build on AI Habitat (Savva et al., 2019) with Bullet (Coumans, 2015) at 120 Hz; Appendix B.3 details the actuation and contrasts it with pure kinematics.

3 HumanCLAW-Bench

Task and scenes. We design a `find-navigate-interact` task to probe long-horizon action intelligence: controlling a body, acting decisively at the 0.5 s timescale, and staying aware of how one's own actions reshape the scene. Each episode gives a single instruction, *"find a <obj>, navigate to it with zero distance, and finally sit on it."* The task is *progressive*—each stage can be attempted only after its prerequisite is met—so the agent explores and decides autonomously; interaction is currently instantiated as `sit`. Scenes come from **HSSD** (Khanna et al., 2024), human-designed multi-floor houses with diverse furniture: the 41 validation houses yield 1,218 episodes over six target categories, each specifying a start pose and a target category solved by reaching any instance. `bed`, `couch`, and `toilet` form the 597-episode *sit* subset with the full progression, while `chair`, `potted_plant`, and `tv` form the 621-episode *navigation* subset; find and navigation metrics are computed over all 1,218 episodes and interaction metrics over the 597 sit episodes. To expose the agent to physical consequence, plausibly movable objects are converted from static to *dynamic*, so contact displaces them. Every episode is further scored along three geometric difficulty dimensions—route distance, navigational choice, and clutter—for difficulty-stratified evaluation (Figure 8 in Appendix C.1).

Evaluation metrics. We evaluate along four dimensions. *(1) Progressive success.* Because the task is progressive, staged success requires a stage to be both objectively achieved and subjectively

Table 1: Full validation results on HumanCLAW-Bench (metrics as defined in Section 3; Motion Jerk at a timescale of ≈ 0.27 s, lower is more coherent). best / 2nd per column.

	High-Level Success Rate			Body Misaware & Scene Disturb			Action Quality	Cost		
VLM Backbone	FindSR ↑	NavSR ↑	InteractSR ↑	Coll. ↓	#Dtb/ep ↓	dDtb(m) ↓	Motion Jerk ↓	avg steps	in tok/step ↓	out tok/step ↓
GPT-5.5	55.1%	13.9%	3.4%	43.4%	1.53	1.56	4.2	82.4	4360	354
Gemini-3.1	64.9%	42.4%	16.8%	39.5%	1.50	1.22	5.7	59.1	5890	311
Gemini-2.5	58.5%	21.6%	3.5%	37.7%	1.35	3.03	8.7	71.3	6412	401
Claude-4.8	32.6%	8.6%	1.5%	44.2%	1.73	2.76	5.3	81.0	7047	625
Gemma-4-31B	58.1%	28.7%	11.1%	40.1%	1.66	1.86	4.8	78.5	4632	322
Qwen3.6-27B	51.0%	20.9%	0.2%	43.0%	1.95	2.16	6.5	81.6	4862	482
Qwen3.6-35B-A3B	44.6%	5.8%	0.0%	34.6%	1.45	2.38	7.4	74.0	4766	477
Qwen3.5-27B	37.8%	13.5%	0.0%	34.5%	0.93	2.38	6.8	37.6	4641	473
InternVL3.5-38B	46.8%	0.8%	0.0%	51.2%	1.90	2.56	7.2	93.0	4459	309

acknowledged. **FindSR** requires the target's semantic ID to occupy at least 100 pixels in the 512×512 egocentric render; **NavSR** requires the minimum distance between the body and the target's axis-aligned bounding box to fall below 20 cm; **InteractSR** requires the agent to sit with its pelvis in contact with the target. Subjective completion is read from the agent's stated visual state and its active decision to stop; the headline metrics require both, and GEO-prefixed variants report the objective criterion alone. *(2) Action quality.* Poor skill orchestration produces unstable behavior such as alternating left and right turns; we measure smoothness by the motion jerk, the third time derivative of the agent's kinematic motion. *(3) Body awareness and disturbance.* A body-unaware agent scrapes walls and knocks over objects; we report the fraction of steps with a collision (**Coll.**), the number of movable objects directly or indirectly displaced per episode (**#Dtb**), and their average displacement (**dDtb**). *(4) Cost.* We report input and output tokens per step.

4 EXPERIMENTS

Setup. We evaluate nine off-the-shelf VLMs as the frozen decision maker H: four proprietary (GPT-5.5, Gemini-3.1, Gemini-2.5, Claude-4.8) and five open (Gemma-4-31B, Qwen3.6-27B, Qwen3.6-35B-A3B, Qwen3.5-27B, InternVL3.5-38B), each driving the same harness, motion generator, and simulator for up to 100 steps per episode.

4.1 ZERO-SHOT MOTION FIDELITY

Before attributing failures to decisions, we verify the motor layer. Commanded with a target magnitude, our prior achieves achieved/commanded ratios close to unity with small variance (walk 0.97, turn 0.99, side-step 1.00, step-back 0.99, sit 0.98; stair skills a stable 0.74–0.79 gain), whereas a strong text-to-motion model, MoMask (Guo et al., 2024), prompted with the same commands renders a generic motion per verb and ignores the magnitude, with an order of magnitude larger variance (Table 4 in Appendix B.4). The benchmark therefore largely isolates high-level embodied reasoning, and the failures below are attributable to planning, perception, and body awareness rather than to skill execution.

4.2 BENCHMARK RESULTS

> ***Finding 1:*** No current VLM solves the benchmark.

Table 1 reports the full results. The best model, Gemini-3.1, finds and sits on the target in only 16.8% of episodes, and four of the nine sit on it in at most 0.2%. Behind this low ceiling is a stage-by-stage drop that every model shares: FindSR ranges from 32.6% to 64.9%, NavSR from 0.8% to 42.4%, and InteractSR from 0% to 16.8%, so most of the task is lost twice—between finding a target and reaching it, and again between reaching it and interacting with it. *Seeing a target rarely becomes reaching it*: the landing rate NavSR/FindSR is only 0.65 even for Gemini-3.1, 0.49 for Gemma-4-31B, and 0.02 for InternVL3.5-38B. Recognizing a target is common across models; turning that recognition into a body that stops next to the target is what separates them. *Body awareness and cost must be read together with success*: Qwen3.5-27B records the lowest disturbance (0.93 objects per episode) and the fewest steps (37.6) yet interacts in 0% of episodes, InternVL3.5-38B and Claude-4.8 pair the highest collision fractions (51.2% and 44.2%) with a collapsed NavSR, and Claude-4.8 spends the most tokens per step (7047 in, 625 out) while trailing on every success metric. *Motion Jerk* scores the coherence of a backbone's choreography rather than the shared gait (Appendix D.2) and is decoupled from success: GPT-5.5 is the most coherent (4.2) while Gemini-2.5 is the worst (8.7). Finally, the

Table 2: Harness ablation on the 100-episode mini-val split with Gemma-4-31B, varied from the baseline.

Setting	High-Level Success Rate			Body Misaware & Scene Disturb			Action Quality	Cost		
	FindSR ↑	NavSR ↑	InteractSR ↑	Coll.% ↓	#Dtb ↓	dDtb(m) ↓	Motion Jerk ↓	avg steps	in tok/step ↓	out tok/step ↓
Baseline (hist 10, img 1)	58.0%	27.0%	18.9%	41.4%	1.35	9.38	4.9	78.5	4676	322
Text-history length										
hist 0	65.0%	11.0%	0.0%	38.6%	1.27	6.46	5.1	88.9	2582	292
hist 20	59.0%	27.0%	7.5%	42.9%	1.18	1.56	4.8	80.3	6692	326
hist 50	60.0%	28.0%	7.5%	40.4%	1.24	6.16	4.8	77.6	10538	320
hist 100	56.0%	26.0%	11.3%	37.6%	1.08	7.23	4.7	79.7	12904	318
Visual-history frames										
img 2	57.0%	31.0%	9.4%	43.0%	1.19	2.55	4.7	78.8	5234	321
img 5	55.0%	28.0%	7.5%	41.1%	1.30	8.86	5.1	82.7	6335	324
img 10	53.0%	13.0%	3.8%	34.9%	0.95	3.19	5.2	92.3	7685	303
Design components										
w/o. verifier	51.0%	2.0%	0.0%	33.2%	0.71	10.40	4.8	47.5	4095	247
w/o. mid-level	62.0%	30.0%	0.0%	35.4%	1.22	2.08	5.0	70.3	3769	264

Table 3: Find / Nav / Interact success-rate variants on HumanClawBench full-val. FindSR = render ($\geq$100 px in ego view) *and* claimed (no negation); GeoFindSR = render $\geq$100 px only. NavSR = active stop *and* body–target AABB distance $\leq$ threshold; GeoNavSR = the AABB distance criterion only. InteractSR = a `sit` action lands the pelvis in *mesh* contact with the target at the committed stop; GeoInteractSR = some `sit` action lands the pelvis in target-mesh contact at *any* step (sat on the seat at some point, even if the pose is later lost). best / 2nd per column.

	Find		Nav			Interact	
VLM Backbone	FindSR ↑	GeoFindSR ↑	NavSR@20cm ↑	GeoNavSR@20cm ↑	NavSR@1m ↑	InteractSR ↑	GeoInteractSR ↑
GPT-5.5	55.1%	63.2%	13.9%	27.7%	15.5%	3.4%	9.7%
Gemini-3.1	64.9%	69.9%	42.4%	44.7%	52.5%	16.8%	21.4%
Gemini-2.5	58.5%	66.4%	21.6%	23.6%	31.0%	3.5%	4.5%
Claude-4.8	32.6%	55.7%	8.6%	13.5%	11.4%	1.5%	2.2%
Gemma-4-31B	58.1%	67.7%	28.7%	35.6%	33.4%	11.1%	17.9%
Qwen3.6-27B	51.0%	62.8%	20.9%	34.8%	24.2%	0.2%	0.3%
Qwen3.6-35B-A3B	44.6%	59.0%	5.8%	14.0%	11.4%	0.0%	0.0%
Qwen3.5-27B	37.8%	48.9%	13.5%	17.7%	19.2%	0.0%	0.0%
InternVL3.5-38B	46.8%	59.1%	0.8%	12.6%	1.6%	0.0%	0.0%

strongest open model is close to the frontier: Gemma-4-31B (58.1/28.7/11.1) matches or beats every proprietary model except Gemini-3.1, so progress on HumanCLAW-Bench is open to the community rather than gated behind a single closed model.

4.3 Ablation

We ablate the harness on a 100-episode mini-val split with Gemma-4-31B (Table 2), varying the text-history length, the number of visual-history frames, and the design components around the baseline (hist 10, img 1).

> ***Finding 2:*** Reasoning scaffolds—memory, mid-level objectives, and verification—lift action intelligence; longer text or visual history does not.

The verifier is decisive: removing it collapses NavSR from 27.0% to 2.0% and InteractSR from 18.9% to 0.0% while FindSR barely moves (58.0% to 51.0%)—its effect is on acting, not on seeing—and episodes end far earlier (47.5 vs. 78.5 steps), the signature of an agent that stops or drifts before the loop closes. *Text history is necessary but saturates*: with no history NavSR and InteractSR fall to 11.0% and 0.0%, a short structured memory recovers most of the gap, yet histories of 20–100 steps leave NavSR near 26–28% while input tokens climb from about 4.7K to 12.9K per step. *More image frames can hurt*: at img 10, NavSR drops to 13.0% and InteractSR to 3.8%, so flooding the model with past frames dilutes rather than improves its reasoning over the current view. *Mid-level reasoning carries the long-horizon interaction*: removing it leaves finding and navigating intact (FindSR 62.0%, NavSR 30.0%) but collapses InteractSR to 0.0%, because sitting is a several-second routine—reach the seat, turn to align, then sit—that a purely step-by-step policy cannot hold together.

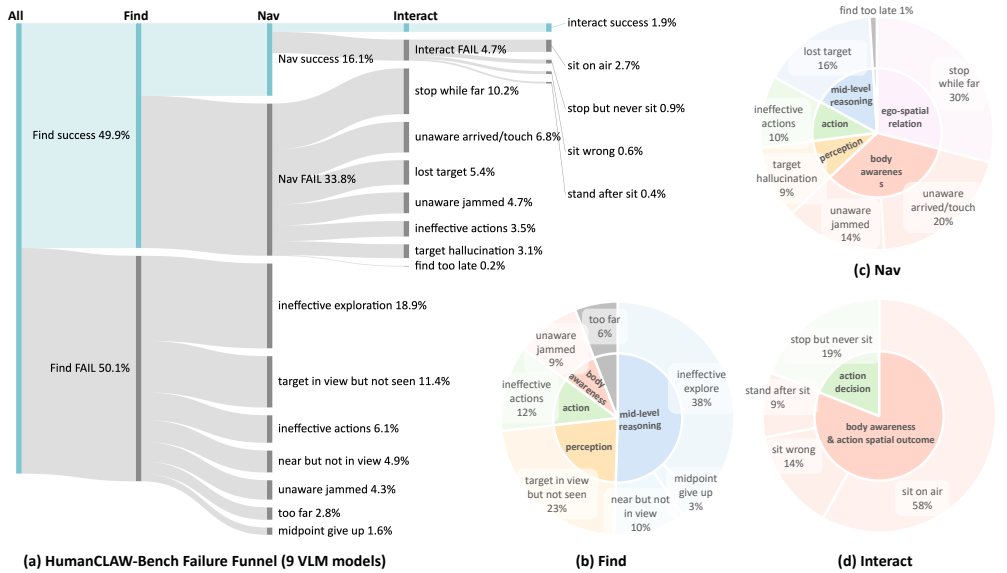

Figure 5: **Where episodes fail across the nine VLMs. (a)** A Find→Nav→Interact funnel over all episodes; **(b–d)** per-stage root-cause breakdowns colored by faculty. Failures concentrate in egocentric body awareness and ego-spatial self-localization—stopping while far, unawareness of arriving or being jammed, and sitting into thin air—not in perception or motor execution.

4.4 Error Analysis and Key Findings

Having seen that no VLM solves the benchmark, we dissect *where* and *why* episodes fail. A capability funnel tracks how episodes drain across the find→navigate→interact stages, and a per-stage breakdown attributes each stage's failures to a single root cause (Figure 5); all labels are assigned automatically by fixed rules over the logged rollouts (Table 10). The appendix adds difficulty-stratified evaluation (Appendix C.1), the objective-versus-acknowledged success variants (Table 3), the construction of Motion Jerk (Appendix D.2), the per-body-part collision breakdown with representative rollouts (Table 5, Figure 9), and the verifier checklist (Table 9), and results for GPT-6 Astra, released after this evaluation (Appendix E). We organize the analysis into four findings, one per stage of the embodied loop.

> ***Finding 3:*** Perception matters but is not the bottleneck.

GeoFindSR counts an episode whenever the target is objectively rendered in the ego view ($\geq$ 100 pixels), whether or not the model says so, while **FindSR** additionally requires the model to acknowledge it. For the strongest models the two differ by only 5–10 points (Gemini-3.1 69.9 vs. 64.9, GPT-5.5 63.2 vs. 55.1, Gemma-4-31B 67.7 vs. 58.1; Table 3), so once a target genuinely enters the ego view it is almost always recognized. Finding breaks upstream, in bringing the target into view at all: about half of all episodes never do, and the dominant root cause is *mid-level reasoning*—chiefly **ineffective exploration** (38%), together with approaching without turning to view (10%) and giving up mid-route (3%)—while genuine perception lapses, where the target is rendered but goes unacknowledged, account for a secondary 23% (Figure 5b). Finding is gated by *where and how the agent chooses to explore*, not by whether it can recognize a target once it appears.

> ***Finding 4:*** Egocentric self-localization is the navigation bottleneck.

Among the 5,473 episodes across all nine VLMs in which the agent *actively finds* the target, 3,706 (68%) still fail to navigate to it (reach within 0.2 m and stop). Nearly two-thirds of these failures are errors of *egocentric self-spatial awareness* rather than of perception or low-level control, and they surface as a *termination* decision that fails in both directions. The largest group is **body awareness** (34%): in 20% the agent has arrived within 0.2 m but never signals arrival and keeps navigating, and in 14% it is jammed against an obstacle yet keeps issuing forward commands, unaware it is blocked. A further 30% is an *ego-spatial distance hallucination*: the agent stops and declares arrival while still more than 0.2 m away. The rest is minor: **mid-level reasoning** (17%; turning away and losing

the target), **visual perception** (9%; claiming a target that has left the view), pure **action** (10%), and discovery timing (under 1%). The core weakness is not *whether the agent can see the target* or *whether it can move*, but *knowing where it is relative to, and whether it has reached, the target it already sees*.

> ***Finding 5:*** Reaching is not interacting.

Among the 726 episodes in which the agent *successfully navigates* to the seat, 513 (71%) still fail to sit. Attributing each from the logged pelvis–object mesh contacts and the action stream, the failures are overwhelmingly errors of *body awareness*—the agent commits to sitting but puts its body in the wrong place (81%)—rather than errors of *decision* (19%). The dominant mode is **sitting into thin air** (58%): the agent issues a sit, yet at no step does its pelvis touch any surface. Another 14% sit on the wrong object or the floor, and 9% briefly reach the seat but keep acting and shift off it before the last frame; the remaining 19% reach and stop in front of the seat but never issue a sit. Once the humanoid has walked up to the target, sitting is gated not by *whether* it decides to sit but by *where it puts its body*—the finest form of embodied spatial reasoning, anticipating where the pelvis will land relative to the seat once the sit unfolds.

> ***Finding 6:*** No current VLM perceives its own body.

Underlying failures at every stage is the agent's obliviousness to its own body. Collisions concentrate in the parts it does not watch: across all nine models the legs and feet collide most (28–45% of steps), the arms and hands next (20–35%), while the head, seldom near obstacles, almost never does (under 7%; Table 5). Representative rollouts (Figure 9) show an arm knocking over a chair that was plainly in view a few frames earlier without the agent ever registering the contact, and legs walking repeatedly into an obstacle that stays visible straight ahead. The scene is seen but the body is not: a VLM trained purely by visual question answering behaves like a *ghost*, holding no instinctive model of which pixels are its own limbs. Part of this blindness is inherited from the interface itself—egocentric RGB and text history only, with no proprioceptive body state or contact signal (Section 5)—but it is not a peripheral nuisance: body awareness underlies 34% of navigation failures and 81% of interaction failures (Findings 4–5), a core capability that embodied action has yet to acquire.

5 Discussion

HumanCLAW isolates a middle layer of physical action—continuous whole-body motion with real scene consequences, but without balance or motor-tracking failures—and uses it to ask how well today's VLMs act through a body. Across nine state-of-the-art models the answer is consistent: none solves the benchmark, and the failures do not lie where a vision–language agent is usually judged. Perception is largely intact (Finding 3), yet competence collapses in everything that follows recognition: knowing where the body is and when it has arrived (Finding 4), placing it precisely enough to interact (Finding 5), and sensing when it collides with the world (Finding 6). A VLM trained by visual question answering learns to describe what it sees, not to feel what it does; it treats its own limbs as just more pixels and never forms the proprioceptive, consequence-predicting model that acting through a body requires. Closing this gap will likely need more than sharper recognition or a longer context window: persistent spatial memory, calibrated termination, and an internal model of the body together with the spatial relations each action produces.

Beneath the framework lies a bet that generalizable action intelligence arises from reasoning rather than from fitting action data: the decision maker stays frozen, the motor layer is a fixed reusable prior, and all composition happens zero-shot. Our results leave the bet open but sharpened: today's reasoning does not yet reach the body, yet the failures are decision-level, exactly where better reasoning could be brought to bear. Our study has limits: the interaction vocabulary is small, decision-level attribution is relative to the granularity of the atomic skills, half-physics abstracts away balance so transfer to a physical humanoid remains future work, and the interface carries no tactile channel, so a body-state or contact signal could be the missing input rather than a missing faculty. HumanCLAW-Bench leaves ample headroom—the strongest model completes the full task on only 16.8% of episodes—concentrated exactly where progress in embodied self-awareness will compound.

AI use statement. We used large language models only to aid and polish the writing of this paper, such as improving grammar and wording of author-written text. LLMs were not used for research ideation, experiment design, data generation, analysis, or drafting sections. All LLM-assisted edits were reviewed by the authors, who take full responsibility for the content.

Ethics statement. This work evaluates frozen, off-the-shelf vision-language models in a simulated environment built from publicly available indoor scene assets. No human subjects, personal data, or real robots are involved. The benchmark is intended to measure and improve the reliability of embodied decision-making; we do not foresee direct negative societal impact beyond those generally associated with embodied AI systems.

Reproducibility statement. Sections 2 and 3 describe the evaluation framework, the atomic skill vocabulary, the motion generator, and the construction of HumanCLAW-Bench, including scene selection, episode definitions, and all success criteria and thresholds. Section 4 specifies the evaluated models and their inference settings, and the appendix lists the verifier checks and the automated root-cause rules used in the failure analysis. The benchmark episodes, evaluation code, and motion generator weights will be released.

REFERENCES

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A vision-language-action flow model for general robot control, 2024. URL https://arxiv.org/abs/2410.24164.

Matthew Chang, Gunjan Chhablani, Alexander Clegg, Mikael Dallaire Cote, Ruta Desai, Michal Hlavac, Vladimir Karashchuk, Jacob Krantz, Roozbeh Mottaghi, Priyam Parashar, Siddharth Patki, Ishita Prasad, Xavier Puig, Akshara Rai, Ram Ramrakhya, Daniel Tran, Joanne Truong, John M. Turner, Eric Undersander, and Tsung-Yen Yang. PARTNR: A benchmark for planning and reasoning in embodied multi-agent tasks. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=T5QLRRHyL1.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. SpatialVLM: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14455–14465, 2024.

Jae-Woo Choi, Youngwoo Yoon, Hyobin Ong, Jaehong Kim, and Minsu Jang. LoTa-Bench: Benchmarking language-oriented task planners for embodied agents. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ADSxCpCu9s.

Xiaoyan Cong, Zekun Li, Zhiyang Dou, Hongyu Li, Omid Taheri, Chuan Guo, Abhay Mittal, Sizhe An, Taku Komura, Wojciech Matusik, et al. Umo: Unified in-context learning unlocks motion foundation model priors. *arXiv preprint arXiv:2603.15975*, 2026.

Erwin Coumans. Bullet physics simulation. In *ACM SIGGRAPH 2015 Courses*, pp. 7:1. ACM, 2015. doi: 10.1145/2776880.2792704.

Markos Diomataris, Berat Mert Albaba, Giorgio Becherini, Partha Ghosh, Omid Taheri, and Michael J Black. Moving by looking: Towards vision-driven avatar motion generation. *arXiv preprint arXiv:2509.19259*, 2025.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar,

Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. PaLM-E: An embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 8469–8488. PMLR, 2023. URL https://proceedings.mlr.press/v202/driess23a.html.

Jiafei Duan, Wentao Yuan, Wilbert Pumacay, Yi Ru Wang, Kiana Ehsani, Dieter Fox, and Ranjay Krishna. Manipulate-anything: Automating real-world robots using vision-language models. In *Conference on Robot Learning*, 2024.

Dibya Ghosh, Homer Rich Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, You Liang Tan, Lawrence Yunliang Chen, Quan Vuong, Ted Xiao, Pannag R. Sanketi, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, July 2024. doi: 10.15607/RSS.2024.XX.090.

Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5152–5161, 2022. doi: 10.1109/CVPR52688.2022.00509.

Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. MoMask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1900–1910, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Guo_MoMask_Generative_Masked_Modeling_of_3D_Human_Motions_CVPR_2024_paper.html.

Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael J. Black. Stochastic scene-aware motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11354–11364, 2021. doi: 10.1109/ICCV48922.2021.01118.

Yining Hong, Jiageng Liu, Han Yin, Manling Li, Leonidas Guibas, Li Fei-Fei, Jiajun Wu, and Yejin Choi. ESI-Bench: Towards embodied spatial intelligence that closes the perception-action loop. *arXiv preprint arXiv:2605.18746*, 2026.

Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. VoxPoser: Composable 3d value maps for robotic manipulation with language models. In *Proceedings of the 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 540–562. PMLR, 2023a. URL https://proceedings.mlr.press/v229/huang23b.html.

Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, Pierre Sermanet, Tomas Jackson, Noah Brown, Linda Luu, Sergey Levine, Karol Hausman, and Brian Ichter. Inner monologue: Embodied reasoning through planning with language models. In *Proceedings of the 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pp. 1769–1782. PMLR, 2023b. URL https://proceedings.mlr.press/v205/huang23c.html.

Yiming Huang, Weilin Wan, Yue Yang, Chris Callison-Burch, Mark Yatskar, and Lingjie Liu. CoMo: Controllable motion generation through language guided pose code editing. In *European Conference on Computer Vision*, pp. 180–196. Springer, 2024. doi: 10.1007/978-3-031-73397-0_11.

Brian Ichter, Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, et al. Do as i can, not as i say: Grounding language in robotic affordances. In *Proceedings of the 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pp. 287–318. PMLR, 2023. URL https://proceedings.mlr.press/v205/ichter23a.html.

Wenqi Jia, Zekun Li, Abhay Mittal, Chengcheng Tang, Chuan Guo, Lezi Wang, James Matthew Rehg, Lingling Tao, and Sizhe An. IAM: Identity-aware human motion and shape joint generation. *arXiv preprint arXiv:2604.25164*, 2026.

Mukul Khanna, Yongsen Mao, Hanxiao Jiang, Sanjay Haresh, Brennan Shacklett, Dhruv Batra, Alexander Clegg, Eric Undersander, Angel X. Chang, and Manolis Savva. Habitat synthetic scenes dataset (HSSD-200): An analysis of 3d scene scale and realism tradeoffs for objectgoal navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16384–16393, 2024. doi: 10.1109/CVPR52733.2024.01550.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P. Foster, Pannag R. Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An open-source vision-language-action model. In *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 2679–2713. PMLR, 2025. URL https://proceedings.mlr.press/v270/kim25c.html.

Jason Lee, Jiafei Duan, Haoquan Fang, Yuquan Deng, Shuo Liu, Boyang Li, Bohan Fang, Jieyu Zhang, et al. MolmoAct: Action reasoning models that can reason in space. *arXiv preprint arXiv:2508.07917*, 2025.

Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Michael Lingelbach, Jiankai Sun, et al. BEHAVIOR-1K: A benchmark for embodied AI with 1,000 everyday activities and realistic simulation. In *Proceedings of the 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pp. 80–93. PMLR, 2023. URL https://proceedings.mlr.press/v205/li23a.html.

Linjie Li, Mahtab Bigverdi, Jiawei Gu, Zixian Ma, Yinuo Yang, Ziang Li, Yejin Choi, and Ranjay Krishna. Unfolding spatial cognition: Evaluating multimodal models on visual simulations. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=fbGmSV6tUw.

Manling Li, Shiyu Zhao, Qineng Wang, Kangrui Wang, Yu Zhou, Sanjana Srivastava, Cem Gokmen, Tony Lee, Li Erran Li, Ruohan Zhang, Weiyu Liu, Percy Liang, Li Fei-Fei, Jiayuan Mao, and Jiajun Wu. Embodied agent interface: Benchmarking LLMs for embodied decision making. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3188. URL https://papers.nips.cc/paper_files/paper/2024/hash/b631da756d1573c24c9ba9c702fde5a9-Abstract-Datasets_and_Benchmarks_Track.html.

Zekun Li, Sizhe An, Chengcheng Tang, Chuan Guo, Ivan Shugurov, Linguang Zhang, Amy Zhao, Srinath Sridhar, Lingling Tao, and Abhay Mittal. Llamo: Scaling pretrained language models for unified motion understanding and generation with continuous autoregressive tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2209–2220, June 2026b.

Zekun Li, Rui Zhou, Rahul Sajnani, Xiaoyan Cong, Daniel Ritchie, and Srinath Sridhar. GenHSI: Controllable generation of human-scene interaction videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 138–149, 2026c. doi: 10.1109/WACV61042.2026.00022.

Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. In *IEEE International Conference on Robotics and Automation*, 2023.

Wenlong Liang, Rui Zhou, Yang Ma, Bing Zhang, Songlin Li, Yijia Liao, and Ping Kuang. Large model empowered embodied AI: A survey on decision-making and embodied learning, 2025. URL https://arxiv.org/abs/2508.10399.

Xiao Liu, Tianjie Zhang, Yu Gu, Iat Long Iong, Yifan Xu, Xixuan Song, Shudan Zhang, Hanyu Lai, Xinyi Liu, Hanlin Zhao, et al. VisualAgentBench: Towards large multimodal models as visual foundation agents. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=2snKOc7TVp.

Yueen Ma, Zixing Song, Yuzheng Zhuang, Jianye Hao, and Irwin King. A survey on vision-language-action models for embodied AI. *IEEE Transactions on Neural Networks and Learning Systems*, 2026. doi: 10.1109/TNNLS.2025.3650584.

NVIDIA, Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, Joel Jang, Zhenyu Jiang, Jan Kautz, Kaushil Kundalia, Lawrence Lao, Zhiqi Li, Zongyu Lin, Kevin Lin, et al. GR00T N1: An open foundation model for generalist humanoid robots, 2025. URL https://arxiv.org/abs/2503.14734.

Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. VirtualHome: Simulating household activities via programs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8494–8502, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Puig_VirtualHome_Simulating_Household_CVPR_2018_paper.html.

Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Tsung-Yen Yang, Ruslan Partsey, Ruta Desai, Alexander Clegg, Michal Hlavac, So Yeon Min, Vladimír Vondruš, Theophile Gervet, Vincent-Pierre Berges, John M. Turner, Oleksandr Maksymets, Zsolt Kira, Mrinal Kalakrishnan, Jitendra Malik, Devendra Singh Chaplot, Unnat Jain, Dhruv Batra, Akshara Rai, and Roozbeh Mottaghi. Habitat 3.0: A co-habitat for humans, avatars, and robots. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=4znwzG92CE.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A platform for embodied AI research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9339–9347, 2019. doi: 10.1109/ICCV.2019.00943.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. ALFRED: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10740–10749, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Shridhar_ALFRED_A_Benchmark_for_Interpreting_Grounded_Instructions_for_Everyday_Tasks_CVPR_2020_paper.html.

Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. Bailando: 3D dance generation by actor-critic GPT with choreographic memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. Bailando++: 3D dance GPT with choreographic memory. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(12):14192–14207, 2023.

Li Siyao, Tianpei Gu, Zhitao Yang, Zhengyu Lin, Ziwei Liu, Henghui Ding, Lei Yang, and Chen Change Loy. Duolando: Follower GPT with off-policy reinforcement learning for dance accompaniment. In *International Conference on Learning Representations (ICLR)*, 2024.

Li Siyao, Yao Feng, Omid Taheri, Chen Change Loy, and Michael J Black. Half-physics: Enabling kinematic 3d human model with physical interactions. *arXiv preprint arXiv:2507.23778*, 2025.

Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M. Sadler, Wei-Lun Chao, and Yu Su. LLM-Planner: Few-shot grounded planning for embodied agents with large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2998–3009, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Song_LLM-Planner_Few-Shot_Grounded_Planning_for_Embodied_Agents_with_Large_Language_ICCV_2023_paper.html.

Andrew Szot, Alexander Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Singh Chaplot, Oleksandr Maksymets, Aaron Gokaslan,

Vladimír Vondruš, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel X. Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra. Habitat 2.0: Training home assistants to rearrange their habitat. In *Advances in Neural Information Processing Systems*, volume 34, pp. 251–266, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/021bbc7ee20b71134d53e20206bd6feb-Abstract.html.

Weihao Tan, Wentao Zhang, Xinrun Xu, Haochong Xia, Ziluo Ding, Boyu Li, Bohan Zhou, Junpeng Yue, Jiechuan Jiang, Yewen Li, et al. Cradle: Empowering foundation agents towards general computer control. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=SJ1kSyO2jwu.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=ehfRiF0R3a.

Zihao Wang, Shaofei Cai, Guanzhou Chen, Anji Liu, Xiaojian Ma, and Yitao Liang. Describe, explain, plan and select: Interactive planning with LLMs enables open-world multi-task agents. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/6b8dfb8c0c12e6fafc6c256cb08a5ca7-Abstract-Conference.html.

Qingxuan Wu, Zhiyang Dou, Chuan Guo, Yiming Huang, Qiao Feng, Bing Zhou, Jian Wang, and Lingjie Liu. Text2Interact: High-fidelity and diverse text-to-two-person interaction generation. *arXiv preprint arXiv:2510.06504*, 2025.

Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. OmniControl: Control any joint at any time for human motion generation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=gd0lAEtWso.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10632–10643, 2025a. doi: 10.1109/CVPR52734.2025.00994.

Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, Marziyeh Movahedi, Manling Li, Heng Ji, Huan Zhang, and Tong Zhang. EmbodiedBench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 70576–70631. PMLR, 2025b. URL https://proceedings.mlr.press/v267/yang25f.html.

Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. PhysDiff: Physics-guided human motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16010–16021, 2023. doi: 10.1109/ICCV51070.2023.01467.

Alex L. Zhang, Thomas L. Griffiths, Karthik R. Narasimhan, and Ofir Press. VideoGameBench: Can vision-language models complete popular video games? *arXiv preprint arXiv:2505.18134*, 2025a.

Libo Zhang, Zekun Li, Tianyu Li, Zeyu Cao, Rui Xu, Xiaoxiao Long, Wenjia Wang, Jingbo Wang, Yuan Liu, Wenping Wang, et al. Egoreact: Egocentric video-driven 3d human reaction generation. *arXiv preprint arXiv:2512.22808*, 2025b.

Yan Zhang and Siyu Tang. The wanderings of odysseus in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20481–20491, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Zhang_The_Wanderings_of_Odysseus_in_3D_Scenes_CVPR_2022_paper.html.

Yan Zhang, Yao Feng, Alpár Cseke, Nitin Saini, Nathan Bajandas, Nicolas Heron, and Michael J. Black. PRIMAL: Physically reactive and interactive motor model for avatar learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12725–12736, 2025c. URL https://openaccess.thecvf.com/content/ICCV2025/html/Zhang_PRIMAL_Physically_Reactive_and_Interactive_Motor_Model_for_Avatar_Learning_ICCV_2025_paper.html.

Kaifeng Zhao, Yan Zhang, Shaofei Wang, Thabo Beeler, and Siyu Tang. Synthesizing diverse human motions in 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14738–14749, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Zhao_Synthesizing_Diverse_Human_Motions_in_3D_Indoor_Scenes_ICCV_2023_paper.html.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Proceedings of the 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 2165–2183. PMLR, 2023. URL https://proceedings.mlr.press/v229/zitkovich23a.html.

A Related Work

A.1 Embodied Tasks and Interactive Environments

Embodied environments differ not only in the tasks they contain, but also in how an agent's actions are grounded and executed in the simulation. VirtualHome (Puig et al., 2018) models household activities as programs over structured scene states and renders atomic instructions with predefined avatar animations, prioritizing activity semantics over closed-loop physical control. ALFRED (Shridhar et al., 2020) introduces egocentric language following, but exposes a discrete set of navigation and interaction actions whose execution is resolved by the simulator. Habitat 2.0 (Szot et al., 2021) and BEHAVIOR-1K (Li et al., 2023) move toward physics-enabled rearrangement and mobile manipulation, supporting articulated objects and persistent physical state under robot-specific control. Habitat 3.0 further introduces humanoid avatars and human-in-the-loop interaction, primarily to study robots acting around or collaborating with human partners (Puig et al., 2024).

Existing embodied environments are often mismatched to general-purpose MLLMs. Program-driven environments resolve actions through predefined transitions, leaving little need to reason about body motion or physical consequence. Fully physics-based simulators, by contrast, require high-frequency continuous control and entangle spatial decisions with motor execution. HumanCLAW strikes a middle ground: a frozen MLLM selects parameterized whole-body motion skills based on egocentric observations, utilizing motion generation model autoregressively generate continuous long horizon behavior. Semi-physics execution preserves collision, contact, and object disturbance while abstracting low-level control, enabling the basic embodied interaction.

Closest to our find-and-navigate setting are object-goal navigation and vision-and-language navigation, where an embodied agent locates and moves to a target from egocentric observations (Savva et al., 2019; Khanna et al., 2024; Anderson et al., 2018). These tasks, however, typically abstract the agent as a moving camera or a velocity-controlled point and score only whether a location is reached, leaving out how a full body arrives, orients, and interacts under physical consequence—exactly the middle layer HumanCLAW targets.

A.2 Multi-modal Foundation as Embodied Decision Makers

Recent surveys distinguish hierarchical agentic systems from end-to-end vision–language–action (VLA) policies (Ma et al., 2026; Liang et al., 2025). Hierarchical agents use multimodal large language models (MLLMs) for task decomposition, spatial grounding, skill selection, feedback-driven replanning (Ichter et al., 2023; Liang et al., 2023; Huang et al., 2023b; Song et al., 2023; Driess et al., 2023; Huang et al., 2023a; Duan et al., 2024; Li et al., 2026c), and open-world memory with reusable skills (Wang et al., 2023; 2024). Frozen VLMs have also been driven in closed loop through first-person games (Tan et al., 2025; Zhang et al., 2025a), where actions remain key presses whose outcomes are resolved by the game script. Although suited to long-horizon reasoning, they often reduce execution to symbolic actions or skill outcomes, obscuring body-level embodied interaction with environments. End-to-end VLA policies instead learn direct motor grounding from embodied data (Zitkovich et al., 2023; Ghosh et al., 2024; Kim et al., 2025; Black et al., 2024; NVIDIA et al., 2025), but couple task reasoning to robot-specific data and low-level control. Action-reasoning models such as MolmoAct (Lee et al., 2025) externalize an editable mid-level spatial trace before emitting actions, yet the decision maker is still trained on action data and evaluated with execution in the loop, rather than queried as a frozen generalist.

Existing benchmarks assess goal interpretation, action sequencing, and long-horizon planning (Li et al., 2024; Choi et al., 2024; Chang et al., 2025; Yang et al., 2025b; Hong et al., 2026), yet primarily score plans, simulator actions, or robot trajectories. HumanCLAW instead evaluates how frozen off-the-shelf VLMs sustain egocentric spatial action over long-horizon tasks. Their skill-level decisions are realized as continuous whole-body motion in a physical world, which preserves collision, contact, gravity, and object dynamics while largely abstracting away balance and motor-tracking failures. This setup evaluates progressive task success together with action coherence, body awareness, and environmental disturbance.

A growing line of work probes the spatial intelligence of VLMs directly (Chen et al., 2024; Yang et al., 2025a;b; Li et al., 2026a), reporting that current models struggle with metric distance, orientation, and multi-step spatial simulation. These evaluations, however, mostly pose spatial understanding as question answering over images, videos, or cross-view pairs, in open-loop settings without

environmental feedback. Action intelligence, as we use the term, is the operational component of spatial intelligence that such settings leave unmeasured: converting spatial understanding into closed-loop embodied decisions—selecting, parameterizing, and sequencing actions as their physical consequences unfold—with the model's own body state part of the problem. It is likewise narrower than a planner in the generic sense (given an observation and a goal, output what the agent should do next): a planner is typically scored open-loop on the plausibility of its proposals, whereas action intelligence is measured through execution, where every decision is realized as motion and its consequence returns in the next egocentric view. Our error analysis is consistent with the spatial-reasoning findings: even when perception succeeds, the dominant failures are egocentric-spatial—self-localization, reachability, and knowing when the body has arrived—so HumanCLAW serves as a physically grounded, closed-loop stress test of exactly this bottleneck.

A.3 Controllable Human Motion Generation as Agentic Tools

Text-to-motion models learn strong priors for plausible full-body movement from large motion-language corpora (Guo et al., 2022; Tevet et al., 2023; Guo et al., 2024; Huang et al., 2024). Beyond just using text, following controllable motion generation methods further condition motion on spatial constraints, character identity, human interaction, ego perception, physics, and music (Hassan et al., 2021; Zhang & Tang, 2022; Zhao et al., 2023; Xie et al., 2024; Yuan et al., 2023; Siyao et al., 2025; Wu et al., 2025; Jia et al., 2026; Cong et al., 2026; Zhang et al., 2025b; Li et al., 2026b; Zhang et al., 2025b; Diomataris et al., 2025; Zhang et al., 2025c; Siyao et al., 2022; 2023; 2024). Although additional control modalities improve motion fidelity, their input protocols beyond the output format of existing MLLMs agent, i.e., text-only output. An agentic executable tool set requires a compact and predictable interface: each tool-calling should follow its semantic intent and continuous parameters while producing a reliable local outcome. Therefore, HumanCLAW proposes a shared receding-horizon motion prior and a plug-and-play adapter for each atomic skill based on text templates with parameters, leaving long-horizon skill selection, parameterization, and composition to VLMs.

B Framework Details

B.1 Closed-Loop Pipeline

Figure 2 shows the three modules of the loop described in Section 2: the harnessed VLM, the skill-conditioned motion generator, and the half-physics simulator.

B.2 Motion Generator

Figure 4 in the main text shows the architecture; this section gives the full formulation, training details, and qualitative results (Figure 6).

Motion Base DiT. Unlike most motion generation models that synthesize a complete sequence, our backbone follows Zhang et al. (2025c) and is trained for receding-horizon continuation. At step t, it observes the last five clean body states and generates the next fifteen frames at 30 fps, corresponding to the 0.5s motion chunk executed by the simulator. The base model is intentionally skill-agnostic: it inherits the current motion dynamics, but not the skill name or task intent. It therefore serves as a general prior over locally natural human motion, while semantic control is added later through skill ControlNet.

A key design choice is how the past motion state conditions generation. Instead of compressing history into a single condition vector (Zhang et al., 2025c), we concatenate it directly with the noisy future as motion tokens:

$$u_\tau = [x_{t-4:t}, \tilde{x}_{t+1:t+15}(\tau)] \in \mathbb{R}^{20\times d}, \quad d = 219. \tag{1}$$

Each frame state is linearly projected to a 38M-parameter, 10-layer DiT with 8-head self-attention processes the full 20-frame sequence. The flow time τ is embedded by a sinusoidal MLP and injected through adaLN-Zero in every DiT block; the base model uses no history encoder and no semantic condition. The network predicts a velocity field for all tokens, and we keep only the future-token outputs:

$$\hat{v}_\theta = G_\theta([x_{t-4:t}, \tilde{x}_{t+1:t+15}(\tau)], \tau)_{t+1:t+15}\,. \tag{2}$$

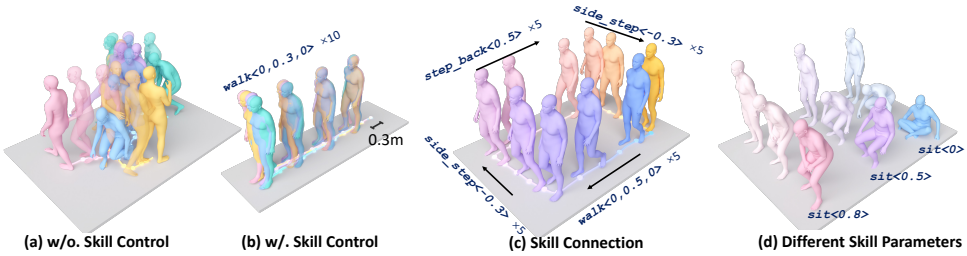

Figure 6: **Skill controllability.** **(a)** Without skill control the prior collapses into colliding poses; **(b)** with control, `walk` tracks the requested 0.3 m stride. **(c)** Skills compose with natural transitions (`walk`/`side_step`/`step_back` into a loop); **(d)** `sit` at heights 0/0.5/0.8 realizes the commanded parameter.

This history-as-token conditioning makes kinematic continuity explicit: future tokens can directly attend to the exact recent pose and velocity states, rather than relying on a compressed history representation that may lose boundary information.

We train the backbone on AMASS motions preprocessed into 20-frame chunks at 30Hz. Before training, each segment is calibrated into a chunk-local frame whose origin and heading are defined by the initial body state. This removes global translation and yaw, so the base DiT models local human dynamics rather than absolute world motion. The first five frames are clean history, and the last fifteen frames form the clean future chunk $x^{\text{fut}} = x_{t+1:t+15}$. Flow matching samples Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$ and a time $\tau \in [0, 1]$, constructs

$$\tilde{x}_{t+1:t+15}(\tau) = (1 - (1 - \sigma_{\min})\tau)\,\epsilon + \tau x^{\text{fut}}, \quad v^\star = x^{\text{fut}} - (1 - \sigma_{\min})\epsilon, \tag{3}$$

and minimizes $\|\hat{v}_\theta - v^\star\|_2^2$ on the future frames. At inference, we start from noise and solve the learned ODE with a 30-step midpoint solver. The resulting base DiT provides a skill-agnostic motion prior: it produces motion that is kinematically continuous with the past, while the next stage aligns this prior with the text-level skill name and its continuous parameters.

Plug-and-Play Skills as ControlNet. Once the base DiT provides a kinematically continuous motion prior, the next problem is to align this prior with the zero-shot skill interface exposed to the VLM. As shown in Figure 4, we do this with one ControlNet-style adapter per atomic skill, while keeping the base DiT frozen. Thus, adding a new skill only involves training a new ControlNet and does not change any previously trained skill. At rollout time, the VLM-selected skill name simply chooses which adapter to attach, and the skill parameters provide the adapter condition.

For a skill s with continuous parameter c_t, we copy the DiT blocks of the base model into a trainable control branch and add a skill-condition encoder $e_s(c_t)$. The encoder is chosen according to the parameter type: scalar conditions such as side-step distance or sitting height use an MLP, while spatial conditions such as walking displacement and yaw use Fourier features before projection to the DiT hidden dimension. At layer ℓ, the frozen base branch and trainable control branch are combined as

$$z_{\ell+1} = B_\ell(z_\ell, \tau) + W_\ell C_\ell(z_\ell + e_s(c_t), \tau), \tag{4}$$

where B_ℓ is the frozen base block, C_ℓ is the trainable control block, e_s is the skill-condition encoder, and W_ℓ is initialized to zero. The final projection layer remains the one from the base DiT; the adapter only modifies the hidden token stream.

The skill parameterization is deliberately tied to physical outcomes that can be read from the motion chunk itself. For walking, the condition is the chunk-local final pelvis displacement and yaw change. For side stepping and stepping backward, it is the corresponding lateral or backward displacement. For turning, it is the final body yaw. For stairs and sitting, the parameters are heuristic geometric quantities such as step height/depth or target sitting height. Therefore, training does not require manually labeled action parameters: after we curate or filter AMASS clips for a skill, the continuous condition is extracted from the calibrated chunk geometry. This does not mean that the training data is used indiscriminately. We manually review and aggressively filter the clips for each ControlNet to keep the skill distribution single-purpose and commonsense-aligned. Clips with abnormal style or mixed intent, such as unstable walking, walking while performing unrelated upper-body actions, or stair motions with excessive body twisting, are removed. This curation is important for zero-shot

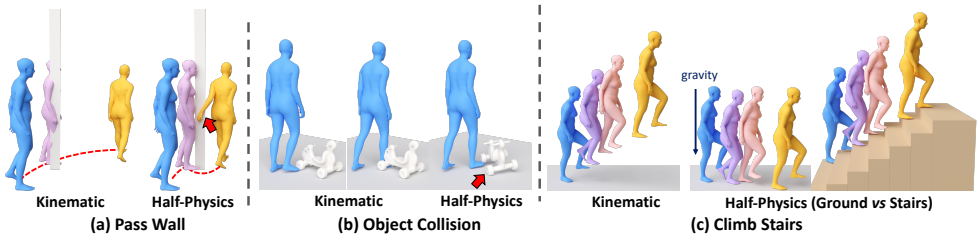

Figure 7: **Half-physics versus pure kinematics.** Pure kinematic playback walks *through* a wall and climbs stairs in mid-air; half-physics subjects the same motion to contact and gravity—the wall blocks the body and the stairs support it—while still factoring out balance and motor-tracking failures.

alignment: when the VLM calls `walk`, `turn`, or `climb_upstairs`, the corresponding adapter should realize the ordinary meaning of that skill rather than an idiosyncratic AMASS motion.

Each adapter is trained with the same flow-matching objective as the base DiT. The base network is frozen, and only the condition encoder, copied control blocks, and zero-initialized residual projections are optimized. Given a clean history, a noisy future, and the extracted skill condition, the adapter predicts the future-frame velocity field and is supervised with $\|\hat{v}_{\theta,s} - v^\star\|_2^2$. We train skill adapters with AdamW using learning rate 3×10^{-4}, batch size 2048, and 5×10^5 to 1.5×10^6 optimization steps depending on the skill. We do not use classifier guidance or classifier-free guidance; skill control is provided directly by the selected adapter and its deterministic parameter interface.

Figure 6 shows these properties qualitatively. On its own the base DiT produces diverse but unconstrained trajectories **(a)**; attaching the skill ControlNet converges them to the zero-shot-reliable behavior the skill name denotes **(b)**. Because all skills share the same motion backbone, transitions between consecutive skills stay smooth, with inertia consistent with human dynamics **(c)**. Moreover, as the continuous condition is read directly from motion geometry, this unsupervised parameterization affords fine-grained control over the skill outcome, e.g., sitting to different target heights **(d)**.

This design makes the skill set extensible. Adding a new primitive only requires defining its text/JSON schema, collecting or filtering motion clips with the relevant outcome parameter, and training a new adapter on top of the frozen base. Conversely, improving one skill adapter does not affect the VLM prompt format, the simulator, the base prior, or the other skills.

B.3 Half-Physics Simulation

Figure 7 contrasts half-physics with pure kinematic playback. Concretely, we build our environment on AI Habitat (Savva et al., 2019) with the Bullet (Coumans, 2015) physics engine. Given a generated 0.5s pose sequence from Sec. 2, we convert consecutive poses into equivalent root and joint velocities. For two adjacent poses q_t and q_{t+1} with simulation step size Δt, the linear joint velocity is computed as

$$\dot{q}_t = \frac{q_{t+1} - q_t}{\Delta t},$$

with rotational components computed from relative rotations in the corresponding tangent space. These equivalent velocities are applied as the only actuation signal for the humanoid. Following prior passive-body simulation practice (Siyao et al., 2025), we run the simulator at 120 Hz and use passive joint stiffness $\lambda = 1.0$ to preserve compliant contact response.

B.4 Zero-Shot Skill Fidelity

Low-level skills are zero-shot reliable. Our controllable motion prior—a shared base diffusion transformer with a per-skill ControlNet, trained once on AMASS/BABEL and queried *zero-shot* at deployment—faithfully realizes the commanded direction and magnitude of every skill without any per-task or per-scene tuning. In Table 4, commanding a skill with a target magnitude yields an achievement ratio close to unity with small variance: walking tracks the commanded step distance at 0.97, turning the commanded angle at 0.99, side-stepping at 1.00, stepping back at 0.99, and sitting reaches the commanded pelvis height at 0.98, while stair climbing realizes 0.74–0.79 of the commanded riser as a stable, calibratable gain. The per-command standard deviation is tiny (≤ 0.12), so the controller executes the same command the same way every time—the near-constant ratios reflect a fixed, predictable gain rather than noise. We contrast this with MoMask (Guo et al., 2024), a

Table 4: **Skill functional fidelity.** Achievement ratio = achieved / commanded (mean$_{\pm \text{std}}$) over rollouts, evaluated under identical conditions for both methods. Our controllable prior tracks every commanded magnitude at ≈ 1 with small variance, whereas the text-to-motion baseline renders a generic motion per verb and ignores the requested magnitude, giving ratios that deviate from 1 with an order of magnitude larger variance.

Method	walk	side_step	step_back	turn_in_place	climb_upstairs	walk_downstairs	sit_in_place
Ours	$0.966_{\pm 0.022}$	$1.002_{\pm 0.123}$	$0.986_{\pm 0.056}$	$0.994_{\pm 0.038}$	$0.794_{\pm 0.019}$	$0.738_{\pm 0.052}$	$0.977_{\pm 0.078}$
MoMask (Guo et al., 2024)	$1.921_{\pm 1.031}$	$1.717_{\pm 1.234}$	$2.072_{\pm 1.067}$	$0.608_{\pm 0.284}$	$0.820_{\pm 0.826}$	$1.540_{\pm 0.421}$	$1.817_{\pm 1.279}$

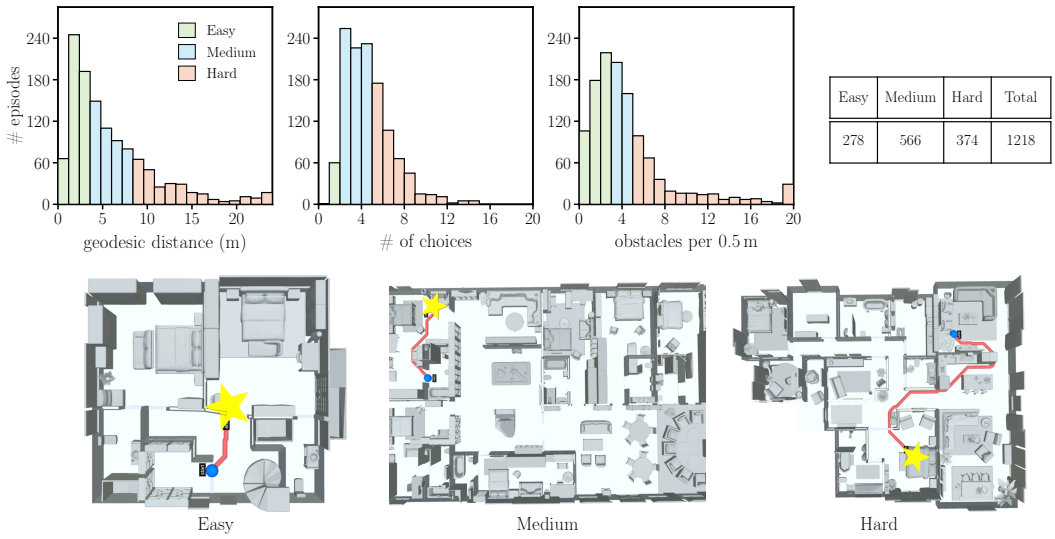

Figure 8: **Difficulty distributions over the** 1,218 **benchmark episodes.** Each episode is scored along three geometric dimensions—distance (geodesic length to the nearest target), choice (turns plus rooms traversed), and obstacle (average obstacles within 1 m of the route)—and split into easy/medium/hard by fixed thresholds; bar color encodes the tier (green/blue/pink) and the inset table lists the overall counts. Distance is right-skewed, while choice and obstacle spread smoothly, giving well-separated difficulty tiers.

strong text-to-motion model, prompted with the identical commands rendered in natural language with the magnitude spelled out (e.g. "*walks forward at* 0.8 *meters per second*"). Even with the number in the prompt, MoMask renders a *generic* motion for each verb instead of the requested magnitude—it walks at a habitual speed and sits to a habitual chair height regardless of the command—so its fidelity ratios depart from unity and scatter with an order of magnitude larger variance (standard deviations of 0.3–1.3 versus 0.02–0.12 for ours). Because our motor layer is this reliable out of the box, the benchmark largely isolates *high-level embodied reasoning*—where to go, what to perceive, and how to sequence actions—rather than low-level motor competence, and the failures we analyze are attributable to planning, perception, and body-awareness rather than to execution of the underlying skills.

C Benchmark Details

C.1 Difficulty Analysis

Difficulty Analysis. To enable difficulty-stratified evaluation, we score every episode along three interpretable geometric dimensions of the route to the target: **distance**, the dataset geodesic distance to the nearest same-category goal; **choice**, the navigational branchiness, given by the number of $> 30^\circ$ turns on the Douglas–Peucker-simplified path plus the number of rooms it traverses; and **obstacle**, the clutter, given by the average number of obstacles within 1 m of the route. Since HSSD provides only a scalar geodesic distance and no path geometry, we first reconstruct the shortest navigable route by running A* on a per-floor 5 cm 2D occupancy grid of walls and static furniture; its length matches the ground-truth geodesic to within a couple percent (median ratio 0.99), confirming the routes are faithful. Each dimension is split into *easy*/*medium*/*hard* by fixed physical thresholds (distance $\leq 3.5/\leq 8$ m; choice $\leq 2/\leq 5$; obstacle $\leq 2.5/\leq 5$) that spread the episodes across tiers (roughly 23/46/31%) rather than forcing equal thirds, and the **overall** difficulty averages the three

Table 5: Per-body-group collision breakdown on HumanClawBench. **Coll.%** = any non-ground body part; the four group columns are per-group step fractions and overlap. best / 2nd per column.

VLM Backbone	Coll.% ↓	Arm/Hand ↓	Torso ↓	Leg/Foot ↓	Head ↓
GPT-5.5	43.37	23.24	13.78	38.83	2.99
Gemini-3.1	39.48	20.97	11.72	33.29	2.62
Gemini-2.5	37.69	22.70	13.63	30.87	2.94
Claude-4.8	44.18	28.97	16.90	37.94	6.93
Gemma-4-31B	40.09	21.21	11.38	34.72	2.66
Qwen3.6-27B	42.97	25.48	12.70	36.86	3.54
Qwen3.6-35B-A3B	34.58	21.20	11.40	27.99	3.71
Qwen3.5-27B	34.53	20.59	10.93	29.58	2.87
InternVL3.5-38B	51.17	34.68	23.12	44.75	6.10

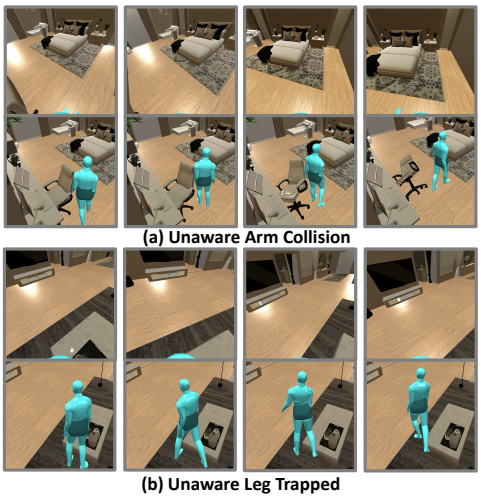

Figure 9: **Body-unawareness failures.** Two episodes, each as egocentric (top) and third-person (bottom) views over consecutive steps: **(a)** the arm scrapes furniture unnoticed; **(b)** the legs jam against a low object while forward commands continue—the costs captured by **Coll.** and **#Dtb**.

levels. Fig. 8 reports the resulting distributions with one worked example per tier, letting us analyze performance and each geometric factor separately across easy, medium, and hard episodes.

D Additional Results

D.1 Objective versus Acknowledged Success

Table 3 separates the objective (geometry-only) success criteria from the model-acknowledged ones at every stage; Finding 3 in the main text is read from its Find columns.

D.2 Motion Jerk as Action Quality

Action quality reflects the coherence of a model's choreography. Task success says nothing about *how* an agent moves, so we report a *Motion Jerk* column—the denoised third time-derivative of the pelvis trajectory at the decision timescale—measuring the coherence of a backbone's choreography rather than the fixed gait of the shared generator. By construction it penalizes incoherent *sequencing*, not the amount of motion: a single skill scores 4–6 (forward walk 5.2, turn 4.3) and a purposeful walk–turn–walk trajectory 5.8, whereas an aimlessly sequenced policy reaches 9.7; fixing the skill content and varying only order confirms this—five forward-walk chunks then five lateral steps score 4.7, but interleaving the same ten chunks yields 7.4 (1.6× higher) purely from ordering. Under this measure (lower is smoother), GPT-5.5 is the most coherent (4.2), followed by Gemma-4-31B (4.8), and quality is decoupled from success: Gemini-2.5, despite a competitive success rate, is the *worst* (8.7), oscillating side to side with the highest turn-reversal rate (about one step in five), while Qwen3.6-35B-A3B is second worst (7.4), adding bouts of in-place spinning.

D.3 Collisions by Body Part

Table 5 reports, for each model, the fraction of motion steps on which any non-ground body part is in contact with fixed geometry (Coll.%) and the same fraction restricted to four body groups; a step can involve several groups, so the group columns overlap and do not sum to Coll.%. The ordering is the same for every model: legs and feet collide most (28–45% of steps), arms and hands next (21–35%), the torso less (11–23%), and the head almost never (3–7%). The spread across models is modest next to the spread in success: InternVL3.5-38B collides most (51.2%) and Qwen3.5-27B least (34.5%), while Gemini-3.1, the strongest model, sits in the middle (39.5%). A low collision rate on its own does not mean that a model moves more carefully: an agent that mostly turns in place, as Qwen3.6-35B-A3B does, also avoids contact. Figure 9 shows two representative rollouts as egocentric (top) and third-person (bottom) frames over consecutive steps: in (a) the arm sweeps

through furniture that was in view a few frames earlier without the agent registering the contact; in (b) the legs are caught on a low object while the agent keeps issuing forward commands. In both cases the visible-state text describes the scene ahead and never mentions the contact, the body unawareness that Finding 6 summarizes.

E GPT-6 Astra: Results and Analysis

Recently, GPT-6 is released. We evaluated it on all 1,218 full-validation episodes with the same harness and metric definitions (Section 3). Tables 6 to 8 place GPT-6 next to Gemini-3.1, the strongest of the nine models, and GPT-5.5, its predecessor.

Table 6: GPT-6 Astra on HumanCLAW-Bench full-val, in the format of Table 1. The Gemini-3.1 and GPT-5.5 rows are repeated from the main table. best / 2nd per column.

	High-Level Success Rate			Body Misaware & Scene Disturb			Action Quality	Cost		
VLM Backbone	FindSR ↑	NavSR ↑	InteractSR ↑	Coll. ↓	#Dtb/ep ↓	dDtb(m) ↓	Motion Jerk ↓	avg steps	in tok/step	out tok/step
GPT-6 Astra	75.5%	57.1%	46.6%	41.5%	1.70	0.65	5.3	60.5	4847	308
Gemini-3.1	64.9%	42.4%	16.8%	39.5%	1.50	1.22	5.7	59.1	5890	311
GPT-5.5	55.1%	13.9%	3.4%	43.4%	1.53	1.56	4.2	82.4	4360	354

Table 7: Success-rate variants of GPT-6 Astra, Gemini-3.1 and GPT-5.5, in the format of Table 3. best / 2nd per column.

	Find		Nav			Interact	
VLM Backbone	FindSR ↑	GeoFindSR ↑	NavSR@20cm ↑	GeoNavSR@20cm ↑	NavSR@1m ↑	InteractSR ↑	GeoInteractSR ↑
GPT-6 Astra	75.5%	78.0%	57.1%	59.3%	58.9%	46.6%	47.9%
Gemini-3.1	64.9%	69.9%	42.4%	44.7%	52.5%	16.8%	21.4%
GPT-5.5	55.1%	63.2%	13.9%	27.7%	15.5%	3.4%	9.7%

Table 8: Per-body-group collision breakdown of GPT-6 Astra, Gemini-3.1 and GPT-5.5, in the format of Table 5 (per-group step fractions in %, groups overlap). best / 2nd per column.

VLM Backbone	Coll.% ↓	Arm/Hand ↓	Torso ↓	Leg/Foot ↓	Head ↓
GPT-6 Astra	41.47	20.62	9.59	35.00	2.08
Gemini-3.1	39.48	20.97	11.72	33.29	2.62
GPT-5.5	43.37	23.24	13.78	38.83	2.99

Headline results. GPT-6 sharply improves every stage over the best of the nine models: FindSR 64.9% → 75.5%, NavSR 42.4% → 57.1%, and InteractSR 16.8% → 46.6% (Table 6). Three changes stand behind the numbers. *Perception.* Failures caused by not recognizing a target that is in view fall from 11.4% for the nine-model pool to 2.5%. *Body self-localization.* For the nine models the dominant navigation problem was locating the body relative to the target: stopping while far, not stopping on arrival, and being jammed unawares made up two thirds of navigation failures (64%, or 21.7% of all episodes). For GPT-6 they fall to 8.1% of all episodes and to 44% of the remaining navigation failures. It still collides as often (Coll. 41.5%), but the objects it touches move far less (dDtb 1.22 m → 0.65 m): its contacts are brief brushes rather than pushes. *Decisiveness.* In this agentic loop GPT-6 acts continuously and commits: it never stops in front of the seat without sitting (0% of its interaction failures, against 18.7% for the nine models), never sits and stands up again (Figure 10d), and completes far more tasks within the same budget of 60.5 decision steps per episode, the level of the most frugal of the nine (Gemini-3.1, 59.1).

Where GPT-6 still fails. Figure 10 repeats the analysis of Figure 5 for GPT-6. The main challenge is still concentrated in ***body awareness***. Of the navigation failures (18.3% of all episodes), 44% are still self-localization errors: the agent stops while far from the target (23%), stands against it without stopping (12%), or is jammed without noticing (9%). The interaction failures are all misplaced sits: the agent commits to the seat and issues the sit, but the pelvis comes down without touching it (69%) or onto the wrong surface (31%). At the same time, a sizeable share of failures shows that the spatial outcome of an action is still not clear to the agent: GPT-6 still fails through ineffective pacing or spinning in 35% of the find failures and 17% of the navigation failures (Figure 10). Videos of representative success and failure cases are provided on the supplementary webpage. In sum, HumanCLAW-Bench remains unsolved.

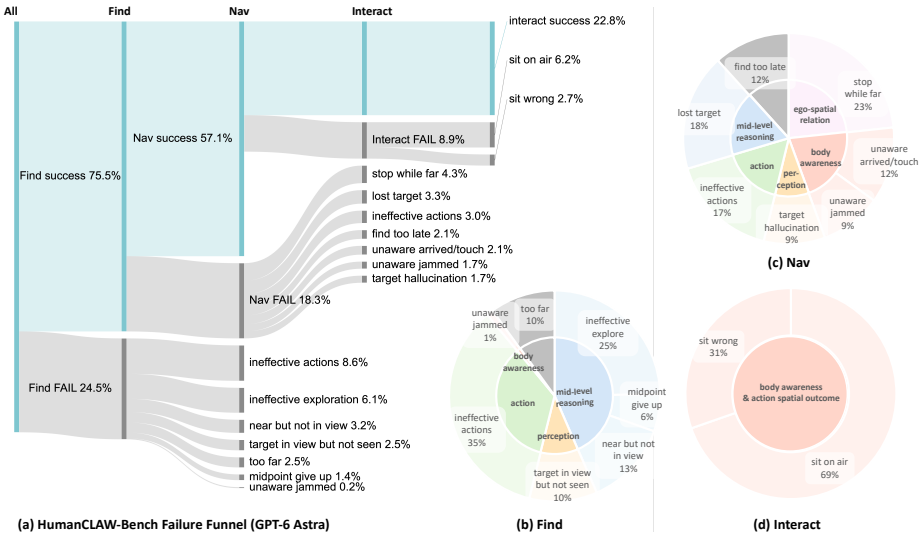

Figure 10: **Where GPT-6 Astra fails. (a)** A Find→Nav→Interact funnel over all 1,218 episodes; **(b–d)** per-stage root-cause breakdowns colored by faculty, with the same taxonomy and rules as Figure 5.

F Skill-Specific Verifier Checks

The verifier applies a fixed, skill-triggered checklist. Its purpose is to counter long-context degradation: as the rollout history grows, the planner's estimates of spatial distance and relation become unreliable, so each check is re-answered from the current egocentric view alone, through a compact, skill-specific prompt. The verifier runs only when one of the skills in Table 9 is proposed; if a check fails, the proposal is replaced with a corrected skill from the same action pool. Skills whose precondition cannot be assessed from the egocentric view—e.g. stepping back, where the region behind the body is unobservable—carry no check.

Table 9: **Skill-specific verifier checks.** Each rule fires only when its skill is proposed; a failed check replaces the proposal with a corrected skill from the same action pool.

Trigger skill	Checks
walk forward	Is the path ahead clear; would the motion collide with a wall or furniture; does moving forward advance the current goal.
stop	Is the high-level goal actually complete; has the body arrived at, or made contact with, the target.
climb up	Is a stair step directly underfoot; is the body facing the staircase.
climb down	Same checks as climb up: a step directly underfoot and the body facing the staircase.
turn (before sitting)	Is the body already flush against the seat (zero distance); after the turn, would the seat lie directly behind the body.

G Automated Root-Cause Attribution

All root-cause labels in Figure 5(b–d) are assigned automatically—no human annotation is involved. A deterministic classifier replays each failed episode's logs and evaluates, per decision step, the geodesic distance from the pelvis to the nearest target instance on the scene's navigation mesh, pelvis–object mesh contacts, the target's rendered semantic pixel count in the egocentric view, the action stream, and the model's own stated visible state. Within each stage, the rules in Table 10 are evaluated top to bottom and the first matching rule assigns the label, so every failed episode receives exactly one error type and the labeling is exactly reproducible from the logs.

Table 10: **Automated root-cause rules** behind the per-stage breakdowns in Figure 5(b–d). All quantities are read from the logged rollouts (navmesh geodesic distances, pelvis–object mesh contacts, per-step egocentric renders, the action stream, and the model's stated visible state); contact is exact mesh contact, not bounding-box overlap.

Error type	Rule (evaluated top to bottom; first match assigns the label)
Find failures — no step both renders the target with ≥ 100 semantic pixels in the 512×512 egocentric view and has the model acknowledge it	
target in view but not seen	The target renders ≥ 100 pixels at some step, but the model never acknowledges it in its visible-state output.
near but not in view	The pelvis comes within 1.5 m of a target instance, yet the target never renders ≥ 100 pixels: the agent reaches the object without ever bringing it into view.
too far	The geodesic distance from the starting pose to the nearest target instance exceeds 20 m.
unaware jammed	The agent issues locomotion commands for ≥ 8 consecutive steps while per-step body displacement stays below 0.1 m.
ineffective actions	A window of ≥ 25 consecutive steps in which the agent keeps issuing movement or turning commands while the body remains confined within a 2 m box.
midpoint give up	The agent first cuts its geodesic distance to the target to below half the initial value, but then retreats until the episode ends at more than $0.6\times$ the initial distance.
ineffective exploration	All remaining cases in which the agent never brings the target into view.
Nav failures — the target is actively found, but the agent never both actively stops and brings the pelvis within 0.2 m of the target	
unaware arrived/touch	The pelvis comes within 0.2 m of the target at some step, but the agent never actively stops: it keeps issuing navigation actions after arrival.
stop while far	The agent actively stops and declares arrival, but the pelvis has never come within 0.2 m of the target.
find too late	The target is first rendered only after step 70 of the 100-step budget, leaving too few steps to reach it.
unaware jammed	The target is still in view (≥ 100 pixels) at the final frame, and the trajectory contains ≥ 8 consecutive locomotion steps with body displacement below 0.1 m: the agent is lodged against geometry yet keeps issuing movement commands.
ineffective actions	The target is in view at the final frame and the body moves freely (no jammed run), yet the agent's actions never bring it within 0.2 m.
target hallucination	The target renders <100 pixels at the final frame, but the model's final visible-state still claims to see it.
lost target	The target is out of view at the final frame and the model no longer claims to see it: the agent turned away and never re-acquired it.
Interact failures — navigation succeeds (active stop within 0.2 m), but the sit fails	
stand after sit	Some sit step achieves pelvis–seat mesh contact with the target, yet the agent continues acting and its pelvis is off the seat at the final frame.
sit on air	A sit is issued, but the pelvis never makes mesh contact with any surface throughout the episode.
sit wrong	A sit is issued and the pelvis ends resting on a different object, the floor, or a wall rather than the target seat.
stop but never sit	The agent reaches the seat and stops, but never issues a sit.