ACTIVEBRAIN: BENCHMARKING AND LEARNING ACTIVE COGNITION IN EMBODIED AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Can embodied agents actively determine what remains unresolved throughout an interaction and act accordingly to complete the original instruction? This requires seeking information beyond the current view, distinguishing the instruction-specified target from similar instances, and revising ineffective behavior. We call this **Active Cognition (AC)**: actively updating decisions based on new observations and action consequences while preserving the instructed goal. To evaluate AC, we introduce **ActiveBrain-Bench**, comprising 1,035 tasks across six tracks with a common policy interface. Its Adapt track separately measures local progress and instruction completion. Zero-shot evaluations of 18 open-weight models and GPT-5.5 reveal fragmented strengths across the six tracks. In Adapt, models may make local progress yet leave the instruction unfinished. To study whether this capability can be learned, **ActiveBrain** constructs 11.5K replay-verified continuations from generated start states and rollout failure states. Each continuation retains RGB–action history and ends in simulator-verified completion. Our post-training pipeline combines supervised fine-tuning, on-policy distillation, and multi-turn reinforcement learning with environment-derived rewards. ActiveBrain-4B raises Qwen3.5-4B's unweighted six-track mean success rate from 8.1% to 44.0%. ActiveBrain-2B reaches 35.0%, above all evaluated zero-shot models. Across seven general multimodal benchmarks, ActiveBrain-4B records an unweighted mean of 72.6, versus 71.1 for its Qwen3.5-4B initialization.

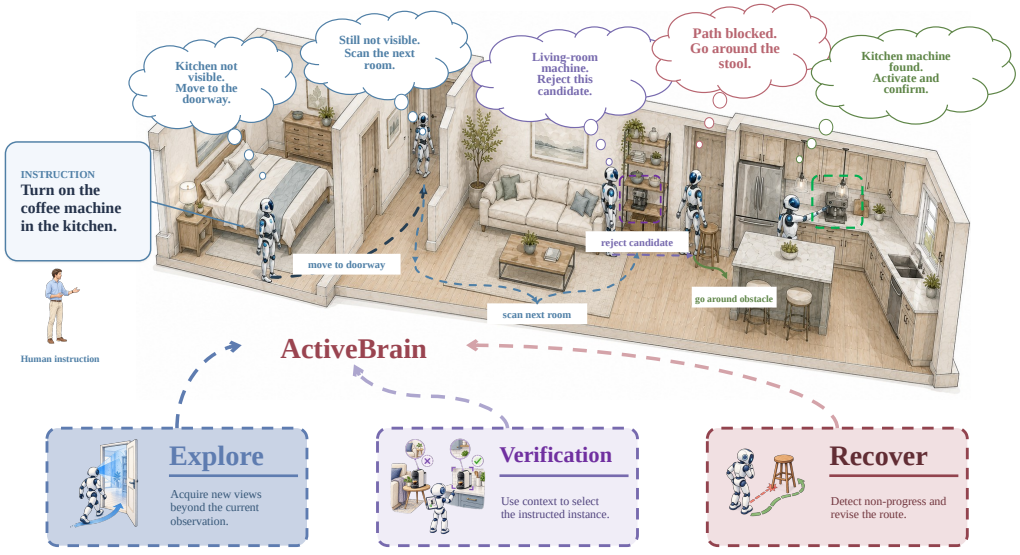


Figure 1: **Active Cognition in interactive embodied task completion.** An agent seeks evidence beyond the current view, verifies the instruction-specified target, and revises behavior after non-progress while continuing to pursue the instructed goal until simulator-verified completion.

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

Embodied agents often begin acting before they have all the information needed to complete an instruction such as "turn on the coffee machine in the kitchen." In Figure 1, the kitchen may lie outside the initial view, a visually similar coffee machine may appear elsewhere, and a route or interaction may fail. Active search helps determine which instance to approach, while action consequences reveal whether the current strategy works. Completing an instruction therefore requires seeking missing information, identifying the instruction-specified target, revising ineffective behavior, and preserving the goal throughout an evolving interaction. We call this capacity **Active Cognition (AC)**: using new observations and action consequences to actively identify what remains unresolved and decide how to act next while preserving the instructed goal.

Prior tasks and methods address individual elements through object navigation and active evidence acquisition (Chaplot et al., 2020; Hong et al., 2026), embodied referring and household task execution (Qi et al., 2020; Shridhar et al., 2020), and behavioral correction after execution failures (Raman et al., 2024; Farag et al., 2025). However, their different policy inputs, starting states, and success criteria prevent a direct test under a common interface: can a policy start from a controlled incomplete state, adjust its behavior from new observations and action consequences, and continue pursuing the instructed goal until completion? An agent may change an ineffective action, approach a relevant object, or clear an obstruction yet still select the wrong instance or stop before completing the instruction. Thus, local behavioral improvement does not imply instruction completion.

To evaluate this capacity, we introduce **ActiveBrain-Bench**. It contains 1,035 fixed tasks, each with an instruction, an incomplete executable simulator state, and a corresponding environment goal. Through a common policy interface, all policies receive the instruction, current RGB observation, and bounded RGB–action history, then continue interacting from the given state. Three core tracks test information seeking (**Explore**), target verification among same-category distractors (**Verification**), and continuation from failure states reached during policy rollouts (**Recover**); three extended settings test adaptation after an ineffective action (**Adapt**), cross-room search (**Cross-Room**), and longer action sequences (**Long-Horizon**). All six use simulator-verified instruction completion; Adapt also reports local progress separately from completion.

Zero-shot evaluations of 18 open-weight models and GPT-5.5 reveal fragmented strengths. The best model reaches only 33.7% mean success across the six tracks. Adapt reveals a further gap: on fixed states, Qwen3.5-4B records a 58.5% local-progress rate but only a 1.0% instruction-completion rate.

To investigate whether AC can be learned, we introduce **ActiveBrain**, which constructs 11.5K replay-verified continuations from generated starts and rollout failure states. Each retains available RGB–action history; replay verifies successful action execution, interaction with the instruction-specified target, and final goal satisfaction. These continuations provide supervised fine-tuning (SFT) targets. We then apply on-policy distillation (OPD) to a student initialized from the base model using student-sampled response prefixes, followed by agentic reinforcement learning (ARL) through multi-turn interaction with environment feedback. ActiveBrain-4B raises Qwen3.5-4B's mean success across the six tracks from 8.1% to 44.0%; ActiveBrain-2B reaches an unweighted six-track mean success rate of 35.0%, above every evaluated zero-shot model on this aggregation. Across seven general multimodal benchmarks, ActiveBrain-4B records an unweighted mean of 72.6 versus 71.1 for its Qwen3.5-4B initialization. In a controlled history intervention on ActiveBrain-4B, shuffling historical action labels relative to the Raw history condition reduces local progress from 80.5% to 76.0% and instruction completion from 39.0% to 24.0%. Under this intervention, completion is more sensitive than local progress to action–observation alignment on Adapt tasks. Our contributions are threefold:

- We formulate **Active Cognition** and introduce **ActiveBrain-Bench** to test whether agents can start from incomplete states, actively identify missing information, unverified targets, or ineffective actions, and adapt their behavior until completing the original instruction.
- We evaluate 18 open-weight models and GPT-5.5, revealing fragmented strengths across tracks and a substantial gap between local progress and instruction completion in Adapt.
- We introduce **ActiveBrain** to inject the ability of Active Cognition by training an agent on newly-constructed 11.5K replay-verified continuations with sequential SFT–OPD–ARL

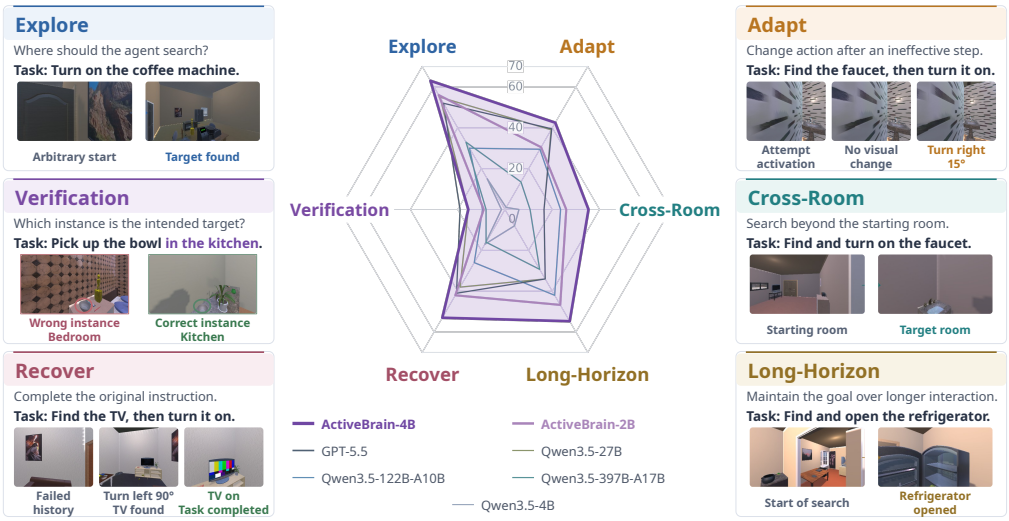


Figure 2: **ActiveBrain-Bench tracks and representative performance.** Side panels show three core AC tracks (left) and three extended settings (right). Adapt follows an ineffective interaction. Recover begins at a failed handoff. The radar reports raw per-track success rates (%). All spokes show instruction completion, with Adapt using Goal@20.

learning. ActiveBrain-4B raises Qwen3.5-4B's unweighted six-track mean success from 8.1% to 44.0% and its seven-benchmark general multimodal mean from 71.1 to 72.6.

2 Related Work

Embodied interaction and diagnostic evaluation. Object navigation, embodied referring, and household interaction benchmarks evaluate target search, language-conditioned localization, and multi-step execution (Chaplot et al., 2020; Qi et al., 2020; Shridhar et al., 2020). EmbodiedBench evaluates vision-driven agents across diverse environments (Yang et al., 2025). ESI-Bench studies active information acquisition (Hong et al., 2026), while Planning with the Views evaluates viewpoint planning (Wang et al., 2026). EmbodiedEval and REAL-Bench broaden interactive task coverage (Cheng et al., 2025; Mi et al., 2026), while Embodied Agent Interface and PRISM support finer-grained diagnosis of embodied decision-making (Li et al., 2024; Lim et al., 2026). These benchmarks vary in policy inputs, starting conditions, and success criteria. ActiveBrain-Bench evaluates instruction completion from controlled incomplete states under a common policy interface, with paired measurements of local progress and instruction completion on the same Adapt tasks.

Learning from failures and interaction feedback. CAPE uses action-precondition errors to propose corrective actions (Raman et al., 2024), while other approaches introduce conditional recovery procedures or combine recovery with learned embodied skills (Farag et al., 2025; Bang et al., 2026). These approaches study how execution problems inform subsequent behavior. DAgger uses supervision on learner-visited states to address distribution shift between training and policy execution (Ross et al., 2011). These works study behavioral improvement during interaction through error feedback, recovery strategies, and the states covered by supervision. ActiveBrain organizes supervision around continuations that retain interaction history and end in verified instruction completion.

3 ActiveBrain-Bench

3.1 Task Definition and Evaluation Protocol

Each task specifies an instruction, an executable simulator start state, and an initially unsatisfied environment goal. At each turn, the policy receives the instruction, current RGB observation, up

to six prior RGB–action pairs, and bounded deterministic summaries of observable history, then outputs one navigation, camera, or object-interaction action. Adapt allows 20 continuation calls and the other tracks 35; invalid outputs consume a call, and the remaining budget is hidden. Each policy is evaluated three times per task. A track's score is the median of its three run-level success rates, and overall success rate (SR) is the unweighted mean of the six track scores. All tracks require simulator-verified instruction completion; Adapt also reports first-action executability, local progress, and action adjustment, with only Goal@20 contributing to SR.

3.2 Track Design and Task Construction

The six tracks use distinct constraints: **Explore** places the target outside the initial view; **Verification** adds same-category distractors distinguishable by the instruction; **Recover** uses solvable failure handoffs from policy rollouts; **Adapt** starts after an ineffective action with a failure-specific local-progress condition; **Cross-Room** places the target outside the starting room; and **Long-Horizon** requires longer within-room action sequences. Tasks are generated in AI2-THOR from held-out ProcTHOR-10K test scenes (Kolve et al., 2017; Deitke et al., 2022). Each candidate is reset and executed; automated checks require an initially unsatisfied goal, a reachable and interactable target, a trajectory completing the instruction, and track-specific constraints. Recover and Adapt handoffs come from four prespecified source models and remain fixed across evaluated policies. Training and evaluation share no scenes or episodes. Target-instance IDs, planner paths, simulator errors, and outcome labels used for construction and scoring are hidden from the policy.

Three human reviewers independently inspect each candidate's step-by-step rendered observations and construction metadata for *correctness* (valid start and trajectory), *solvability* (unambiguous target and feasible continuation), and *nontriviality* (initial-view cues or prior knowledge cannot bypass the track condition). Each criterion is decided by majority vote; any failure excludes the candidate.

3.3 Benchmark Composition and Handoff States

ActiveBrain-Bench has 1,035 fixed tasks: 200 each in Explore, Verification, Recover, and Adapt; 85 in Cross-Room; and 150 in Long-Horizon. Recover and Adapt handoffs are mainly blocked navigation (45.0% and 48.5%, respectively) or premature interaction (42.0% and 48.0%, respectively). Although 76.0% and 75.5% occur at source-rollout steps 31–35, the continuing policy sees at most six prior RGB–action pairs; in half of both sets, those prior RGB frames show only one distinct observation. Full distributions appear in Appendix A.1.3.

4 Learning from Incomplete States to Instruction Completion

4.1 From incomplete states to verified continuations

As shown in Figure 3, ActiveBrain constructs replay-verified task continuations from generated start states and failure states reached by policies, then learns from them through supervision and environment feedback. Generated starts cover varied information-seeking situations. For failure states reached by policy rollouts, we construct continuations to instruction completion and retain the preceding RGB–action history visible to the policy. We then construct an A^* path to an interactable pose near the instructed target and append the actions needed to complete the instruction. Simulator replay verifies the retained history and continuation together, including action execution, interaction with the instruction-specified target, and final goal satisfaction. This yields 11.5K continuations ending in simulator-verified instruction completion.

4.2 Learning from verified continuations

We first use supervised fine-tuning (SFT) to learn *Active Cognition Chain-of-Thought* (AC-CoT), in which the model reasons over the policy-visible RGB–action history before predicting its next action. For each step, the Qwen3.5-397B-A17B teacher uses the instruction, visible interaction history, and replay-verified next action to generate a rationale and structured action target. Filtering malformed, inconsistent, and information-leaking outputs yields 122,109 turn-level SFT responses from the 11.5K continuations for both the 2B and 4B models.

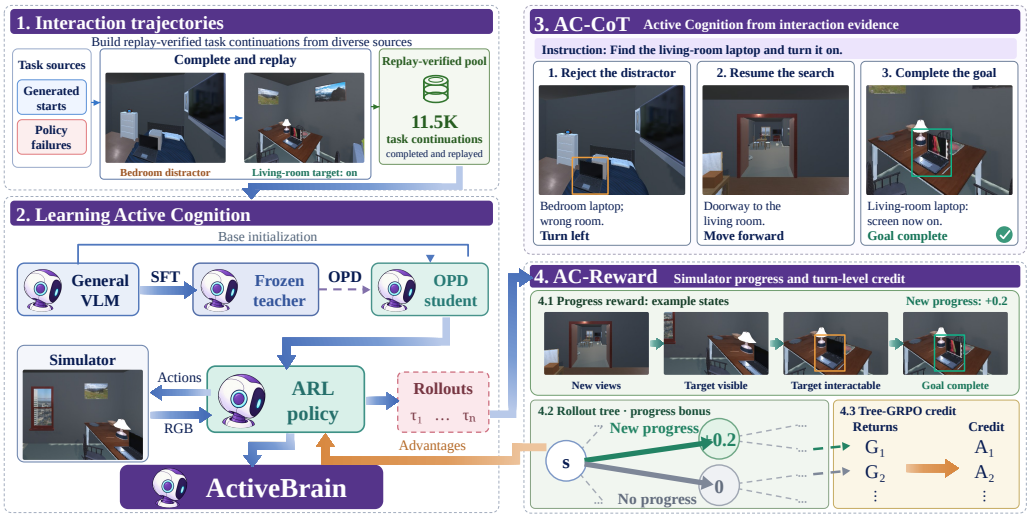


Figure 3: **Learning from incomplete states to instruction completion.** ActiveBrain turns generated start states and failure states reached by policies into 11.5K continuations that retain the policy-visible history and are verified by simulator replay to complete the instruction. The resulting data support SFT, distillation, and multi-turn RL.

To transfer this behavior while preserving the original base-model initialization, we apply on-policy distillation (OPD) (Agarwal et al., 2024). A student initialized from the base model is presented with 21K distinct state prompts twice and samples one response per presentation, while the corresponding SFT model provides token-level supervision on the sampled prefixes. This yields 42K sampled responses. The student is trained to minimize the generalized Jensen–Shannon (GJS) divergence:

$$\mathcal{L}_{\mathrm{OPD}}(\theta) = \mathbb{E}_{x,\, y\sim p_\theta(\cdot|x)} \sum_k D_{\mathrm{GJS}}(p_\theta(\cdot \mid x, y_{<k}), p_T(\cdot \mid x, y_{<k}))\,. \tag{1}$$

The student-sampled responses are not executed in the simulator, and no environment rewards are used during OPD. Closed-loop environment feedback enters only in the subsequent ARL stage.

4.3 Learning from action consequences

Verified continuations teach the policy how to act from known interaction states, but do not expose it to the consequences of its own evolving decisions. We therefore initialize agentic reinforcement learning (ARL) from the OPD model and train through multi-turn interaction on 1,100 interactive simulator tasks in AI2-THOR, with eight rollouts per task. Each predicted action is executed before the next decision. AC-Reward adds a bonus of 0.2 when the simulator verifier detects new task progress, encouraging actions that move the instruction forward rather than repeat ineffective behavior. Tree-GRPO (Ji et al., 2026) compares the downstream returns of branches sharing the same interaction prefix and assigns their relative outcomes back to the corresponding action turn. Verifier signals are used only for training and are never exposed to the policy. The appendix details data filtering, distillation, rewards, and policy optimization. The stages use distinct training conditions. SFT learns rationale–action targets along replay-verified continuations. The SFT model supervises an OPD student initialized independently from the base model on student-sampled prefixes. ARL starts from this student and optimizes multi-turn rollouts. The stages form one sequential path. The reported models are not independent component ablations.

5 Experiments

5.1 Experimental setup

We ask four questions: whether current policies continue acting toward instruction completion, whether local progress predicts completion, whether instruction completion changes across

Table 1: **Model performance on ActiveBrain-Bench.** Per-track success rates (%); Adapt reports Goal@20 and SR is the unweighted six-track mean. ActiveBrain rows are final ARL checkpoints. Gold, silver, and bronze mark the top three values; underlining marks the best zero-shot result. Exp/Ver/Rec/Adp/ CR/LH abbreviate the six tracks. Rank is computed over all policies in Appendix Table A4; omitted rows therefore produce nonconsecutive ranks.

Policy	**Rank↓**	**Core AC tracks**			**Extended task settings**			**SR↑**
		Exp↑	Ver↑	Rec↑	Adp↑	CR↑	LH↑	
Zero-shot policies								
Qwen3.5-4B	13	15.0	1.5	17.0	1.0	5.9	8.0	8.1
Qwen3.5-27B	4	55.0	15.5	38.0	39.0	16.5	34.0	33.0
Qwen3.5-122B-A10B	5	30.0	9.5	26.0	29.5	23.5	42.0	26.8
Qwen3.5-397B-A17B	6	33.0	8.0	16.0	13.5	10.6	29.3	18.4
Qwen3-VL-4B	14	15.5	1.0	3.0	0.5	2.4	6.7	4.9
Qwen3-VL-32B	10	35.5	6.0	0.5	0.0	12.9	22.7	12.9
Qwen3-VL-235B-A22B	7	47.0	9.0	5.0	0.0	11.8	23.3	16.0
RynnBrain1.1-9B	12	19.0	2.5	14.0	7.0	4.7	12.7	10.0
GLM-4.6V	9	29.0	7.0	15.5	0.0	10.6	17.3	13.2
GPT-5.5	3	52.0	**19.0**	41.0	39.5	16.5	34.0	33.7
ActiveBrain								
ActiveBrain-4B	**1**	**63.0**	15.5	**53.0**	**42.5**	**35.3**	**54.7**	**44.0**
ActiveBrain-2B	2	56.0	9.0	42.0	30.5	25.9	46.7	35.0

Table 2: **Local progress, completion, and repeated-action diagnostics (%).** LocalProgress@20 and Goal@20 share 200 Adapt states. Repeat@1 measures immediate repetition on Adapt. On the five non-Adapt tracks, After-failure is the repeat rate after executable failures. Episodes w/ repeat is the share of episodes containing any repeat. ActiveBrain rows are shaded. Gold and silver mark the best two values. Higher is better for Adapt outcomes, lower for repetition.

	Adapt outcome		**Repeated action**		
Policy	LocalProgress@20↑	Goal@20↑	After-failure↓	Episodes w/ repeat ↓	Repeat@1↓
Qwen3.5-4B	58.5	1.0	45.2	88.1	30.0
Qwen3-VL-32B	62.5	0.0	68.9	89.0	21.0
Qwen3.5-27B	74.0	39.0	43.9	56.8	16.5
GPT-5.5	78.0	39.5	4.3	26.5	4.0
ActiveBrain-4B SFT	**85.0**	36.0	2.1	6.6	**0.0**
ActiveBrain-4B OPD	80.0	40.5	**0.3**	**1.0**	**0.0**
ActiveBrain-4B + ARL	81.0	**42.5**	0.5	1.4	**0.0**

continuation-based post-training, and whether general multimodal performance is retained. We compare 18 open-weight zero-shot policies and GPT-5.5 with the Qwen3.5-4B SFT model, its base-initialized OPD student, the subsequent ARL model, and the final ActiveBrain-2B. Throughout, ActiveBrain-2B and ActiveBrain-4B denote the final ARL checkpoints.

5.2 Current policies show fragmented strengths and a gap between local progress and completion

Table 1 and Figure 2 show complementary zero-shot strengths: Qwen3.5-27B leads Explore, GPT-5.5 leads Verification, Recover, and Adapt, and Qwen3.5-122B-A10B leads Cross-Room and Long-Horizon. GPT-5.5 reaches 33.7% SR; medians across the 18 open-weight policies are 15.25% on Explore, 3.0% on Verification, and 8.75% on Recover. On Adapt, Qwen3.5-4B reaches 58.5% LocalProgress@20 but only 1.0% Goal@20; GPT-5.5 reaches 78.0% and 39.5%. Verification peaks at only 19.0%, consistent with the wrong-instance interaction audited in Appendix A.4.1.

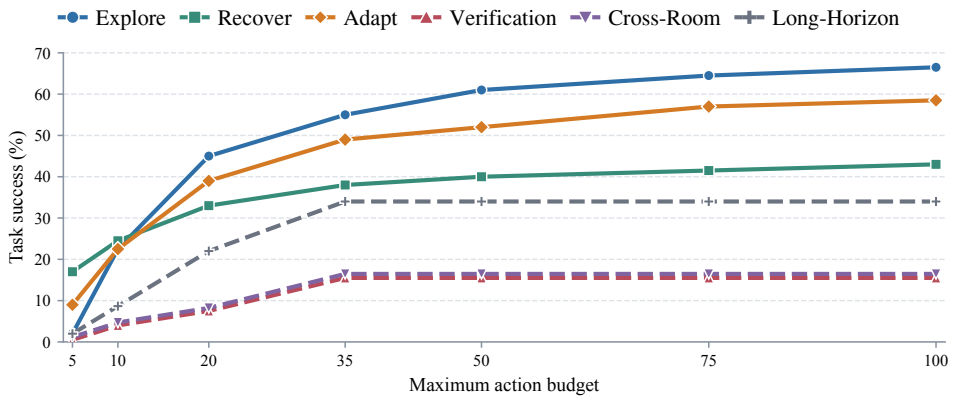


Figure 4: **Task success under increasing action budgets.** Curves report prefix success for the same fixed Qwen3.5-27B tasks at action cutoffs from 5 to 100. The main comparison fixes Adapt at 20 actions and all other tracks at 35 actions for every evaluated policy in Table 1.

Table 3: **ActiveBrain-4B post-training path.** Success rates (%) on fixed sets. Adapt: Goal@20. SR: six-track mean. Lower SFT rows probe rationale supervision, failed-state coverage, and policy-visible history. Gold/silver: column ranks 1/2 in the upper four rows. Ties use the minimum rank.

Model	Exp↑	Ver↑	Rec↑	Adp↑	CR↑	LH↑	SR↑
Qwen3.5-4B	15.0	1.5	17.0	1.0	5.9	8.0	8.1
SFT	55.5	10.0	48.5	36.0	31.8	50.7	38.8
OPD	**70.5**	9.5	43.0	40.5	30.6	**64.7**	43.1
+ ARL	63.0	**15.5**	**53.0**	**42.5**	**35.3**	54.7	**44.0**
SFT w/o rationale	35.5	2.5	31.0	14.0	16.5	27.3	21.1
SFT w/o failed states	41.5	3.5	27.0	19.0	21.2	38.7	25.1
SFT w/o history	31.0	3.0	18.5	12.5	18.8	29.3	18.9

Across the representative zero-shot policies in Table 2, LocalProgress@20 is 58.5–78.0% but Goal@20 is 0.0–39.5%. For Qwen3.5-4B, only 2 of the 117 Adapt episodes meeting LocalProgress@20 also reach Goal@20 (Appendix A.5). Local progress therefore does not imply completion of the instruction on the same fixed tasks under the same 20-step action budget.

Action-budget sensitivity. Figure 4 reports cumulative Qwen3.5-27B success on the same tasks at cutoffs of 5, 10, 20, 35, 50, 75, and 100 actions from rollouts of up to 100 actions per fixed task. The evaluator, not the policy, observes the cutoff, so each shorter-budget result is the corresponding rollout prefix. Explore, Recover, and Adapt improve with more actions, whereas Verification, Cross-Room, and Long-Horizon plateau after 35 actions. More steps do not resolve every failure.

5.3 Instruction completion across the training pipeline

ActiveBrain-4B improves every track over Qwen3.5-4B and raises SR from 8.1% to 44.0%, 10.3 points above the best zero-shot SR. Verification remains below GPT-5.5 (15.5% versus 19.0%). ActiveBrain-2B reaches 35.0%. Table 3 reports the models along the training pipeline.

Performance under SFT variants. SFT raises SR from 8.1% to 38.8%, including Recover from 17.0% to 48.5% and Adapt Goal@20 from 1.0% to 36.0%. The SFT w/o rationale variant uses the same verified continuation states, policy-visible history, and action targets as SFT but omits rationale tokens. It reaches 21.1% SR, compared with 38.8% for rationale–action SFT. Removing policy-reached failure states yields 25.1% SR, while removing the accumulated RGB–action context yields 18.9%. These variants probe SFT sensitivity. They do not control data and compute or compare against matched first-action-only supervision. The OPD student reaches 43.1% SR, with its largest gains over SFT on Explore and Long-Horizon.

Table 4: **History intervention for final ActiveBrain-4B on Adapt.** Goal@k measures completion by step k. LocalProgress@20 measures local progress from the ineffective-action state. Shuffled preserves RGB order but misaligns historical action labels. All reported values are percentages. Gold/silver mark ranks 1/2 per metric row. Ties share the minimum rank.

Metric	Full	Raw history	Current only	Shuffled
Goal@5↑	**1.0**	**1.0**	0.5	0.5
Goal@10↑	**18.0**	13.0	5.0	4.0
Goal@15↑	**32.0**	26.5	8.5	15.0
Goal@20↑	**42.5**	**39.0**	**10.0**	**24.0**
LocalProgress@20↑	**81.0**	80.5	38.0	76.0

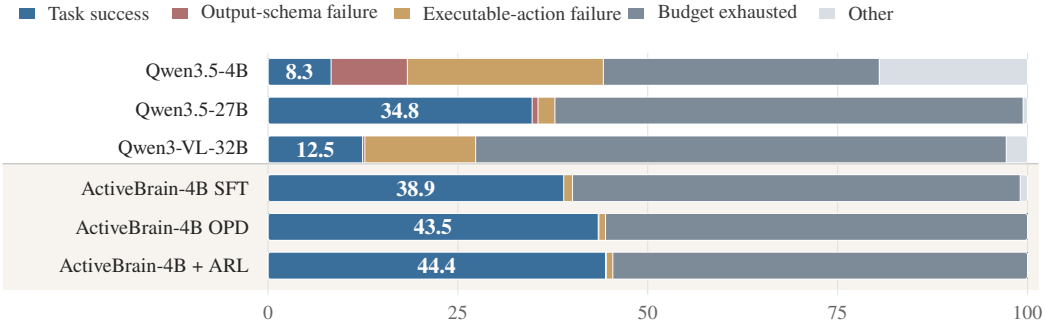


Figure 5: **Failure profiles before and after training.** Each profile pools the median-scoring run from each track (1,035 episodes). Labels show pooled success; SR averages the six tracks equally.

5.4 History use and stage-wise changes in completion

From OPD to ARL, Verification rises from 9.5% to 15.5%, Recover from 43.0% to 53.0%, Adapt from 40.5% to 42.5%, and Cross-Room from 30.6% to 35.3%, while Explore and Long-Horizon decrease. Overall SR rises from 43.1% to 44.0%. Diagnostic definitions are in Appendix A.2.

Aligned history supports instruction completion. Table 4 evaluates the same 200 fixed Adapt handoffs. Raw history removes handcrafted summaries. Current only retains the current RGB frame. Shuffled preserves the chronological RGB sequence but deterministically assigns an incorrect action label to each frame. Removing summaries changes Goal@20 from 42.5% to 39.0%. Shuffling action labels preserves most local progress (80.5% to 76.0%) but lowers Goal@20 to 24.0%. Current RGB alone lowers it to 10.0%. For SFT, the paired Raw–Shuffled differences are 12.5 points in Goal@20 (McNemar $p = 0.0022$, bootstrap 95% CI $[5.0, 20.0]$) and 4.5 points in LocalProgress@20 ($p = 0.262$, CI $[-2.5, 11.5]$, Appendix A.6.1). Under these input interventions, completion changes more than local progress.

The gains are not explained by repetition or output formatting. The after-failure repeat rate falls from 45.2% to 2.1% after SFT and 0.3% after OPD. Episodes with a repeat fall from 88.1% to 6.6% and 1.0%. Yet SFT-to-OPD LocalProgress@20 falls from 85.0% to 80.0% while Goal@20 rises from 36.0% to 40.5%. ARL reaches 81.0% and 42.5%, respectively. Avoiding repetition is therefore insufficient for completion. Output compliance is also insufficient: Qwen3-VL-32B has only three terminal schema failures but completes 129 of the 1,035 diagnostic episodes (Figure 5). SFT removes all 104 schema failures of Qwen3.5-4B while raising completions from 86 to 403. OPD reaches 450 after SFT has already reduced schema failures to zero. The improvements therefore extend beyond suppressing repeated actions or repairing output format.

A non-Adapt behavioral audit also shows a trade-off along this sequential path: from SFT to OPD, object-interaction attempts fall from 2,406 to 1,704 and their executable share rises from 16.4% to 26.6%, while failed Recover episodes with no interaction rise from 36.9% to 60.5% (Table A12). These checkpoint differences describe changed behavior but do not isolate a distillation effect.

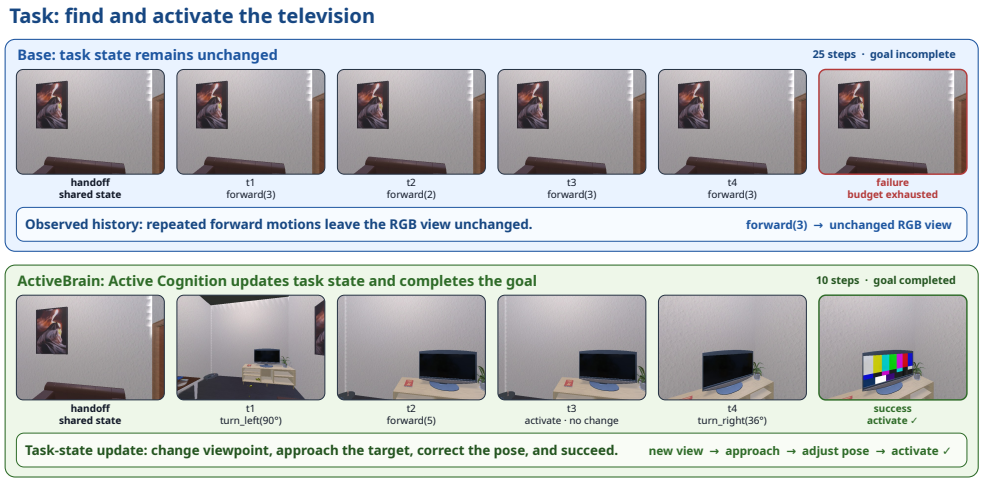


Figure 6: **Matched continuation trajectories.** From the same state and history, Qwen3.5-4B repeats blocked motion. ActiveBrain-4B changes viewpoint, resumes search, and completes the task.

Table 5: **Seven general multimodal benchmarks.** MUIR, Angles, and Spatial abbreviate benchmark names. Avg is the unweighted seven-benchmark mean. Cell arrows show change from Qwen3.5-4B, while gold and silver mark the top two scores in each higher-is-better benchmark column across all compared policies and training stages.

Model	MMBench↑	MMMU↑	MMStar↑	AI2D↑	MUIR↑	Angles↑	Spatial↑	Avg↑
Qwen3.5-4B	91.0	65.0	73.2	86.5	61.6	45.5	74.6	71.1
ActiveBrain-4B								
SFT	82.6↓	50.1↓	58.2↓	74.5↓	45.1↓	18.9↓	67.6↓	56.7↓
OPD	**91.1**↑	**67.0**↑	74.0↑	**87.8**↑	63.3↑	48.5↑	75.7↑	72.5↑
+ ARL	91.0→	66.9↑	**74.2**↑	87.5↑	**63.8**↑	**48.9**↑	**75.9**↑	**72.6**↑

Figure 6 compares continuations from one incomplete state and policy-visible history. The base policy repeats blocked forward motions until its budget expires. ActiveBrain-4B changes viewpoint, approaches the target, adjusts after an ineffective interaction, and completes the instruction.

5.5 Performance on general multimodal benchmarks

Across MMBench (Liu et al., 2023), MMMU (Yue et al., 2024), MMStar (Chen et al., 2024), AI2D (Kembhavi et al., 2016), MUIRBench (Wang et al., 2024), All-Angles (Yeh et al., 2026), and SpatialBench (Xu et al., 2025) (Table 5), SFT raises embodied SR from 8.1% to 38.8% but lowers the seven-benchmark mean from 71.1 to 56.7. The OPD student initialized from the base model restores the mean to 72.5 while further raising embodied SR to 43.1%, matching or exceeding the base model on all seven benchmarks. After ARL, the corresponding values are 72.6 and 44.0%. These point estimates indicate retention on the measured suite after OPD and ARL, but do not separately identify how base initialization and on-policy distillation produce that retention.

6 Conclusion

ActiveBrain studies Active Cognition in embodied agents: actively updating decisions based on new observations and action consequences while pursuing the original instruction to completion. ActiveBrain-Bench reveals fragmented zero-shot strengths across six tracks and a gap between local progress and completion in Adapt. We construct 11.5K replay-verified task continuations retaining RGB–action history and evaluate a sequential SFT–OPD–ARL path. This path raises Qwen3.5-4B's unweighted six-track mean success rate from 8.1% to 44.0%; ActiveBrain-2B reaches 35.0%. Within the tested simulation setting, these results provide evidence that the post-training path can improve Active Cognition, as reflected in higher instruction-completion rates from incomplete states.

AI USE STATEMENT

Generative AI tools were used to support idea discussion, language polishing, and limited code execution and debugging. All research questions, benchmark and method design choices, implementations, experiments, analyses, and conclusions were made or verified by the authors, who take full responsibility for the content of this paper.

REPRODUCIBILITY STATEMENT

The paper and appendix describe the construction and task splits of ActiveBrain-Bench, its evaluation protocols, and the training procedure, implementation details, and main hyperparameters of ActiveBrain. Upon acceptance, we will publicly release the training and evaluation code, the configurations and associated resources for the 1,035 benchmark tasks, and the ActiveBrain-2B and ActiveBrain-4B model weights to support reproducibility and future research.

ETHICS STATEMENT

This work studies the evaluation and post-training of embodied agents in simulated household environments. Our implementation builds upon publicly available simulators, code, pretrained models, and research datasets, subject to their respective licenses and usage terms. The study does not involve human participants or the collection of personal or sensitive data. The evaluated and trained policies may inherit limitations and biases from pretrained models and simulated environments.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b. URL https://arxiv.org/abs/2502.13923.

Jinsik Bang, Jaeyeon Bae, Donggyu Lee, Siyeol Jung, and Taehwan Kim. Environmental understanding vision-language model for embodied agent. *arXiv preprint arXiv:2604.19839*, 2026.

Devendra Singh Chaplot, Dhiraj Prakashchand Gandhi, Abhinav Gupta, and Russ R Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. *Advances in neural information processing systems*, 33:4247–4258, 2020.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2403.20330.

Zhili Cheng, Yuge Tu, Ran Li, Shiqi Dai, Jinyi Hu, Shengding Hu, Jiahao Li, Yang Shi, Tianyu Yu, Weize Chen, Lei Shi, and Maosong Sun. Embodiedeval: Evaluate multimodal LLMs as embodied agents, 2025. URL https://arxiv.org/abs/2501.11858.

Ronghao Dang, Jiayan Guo, Bohan Hou, Sicong Leng, Kehan Li, Xin Li, Jiangpin Liu, Yunxuan Mao, Zhikai Wang, Yuqian Yuan, et al. RynnBrain: Open embodied foundation models. *arXiv preprint arXiv:2602.14979*, 2026. URL https://arxiv.org/abs/2602.14979.

Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Kiana Ehsani, Jordi Salvador, Winson Han, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. Procthor: Large-scale embodied ai using procedural generation. *Advances in neural information processing systems*, 35:5982–5994, 2022.

Youmna Farag, Svetlana Stoyanchev, Mohan Li, Simon Keizer, and Rama Doddipatla. Conditional multi-stage failure recovery for embodied agents. In *Proceedings of the 1st Workshop for Research on Agent Language Models (REALM 2025)*, pp. 200–227, 2025.

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.

Yuxiang Ji, Ziyu Ma, Yong Wang, Guanhua Chen, Xiangxiang Chu, and Liaoni Wu. Tree search for LLM agent reinforcement learning. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ZpQwAFhU13.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European Conference on Computer Vision*, 2016.

Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, et al. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017.

Kehan Li, Bohan Hou, Minghao Zhu, Tianyi Zhang, Zesen Cheng, Zhikai Wang, Sicong Leng, Xin Li, Xiao Lin, Biying Yao, et al. RynnBrain 1.1: Towards more capable and generalizable embodied foundation model. *arXiv preprint arXiv:2607.17977*, 2026. URL https://arxiv.org/abs/2607.17977.

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*, 2024. URL https://arxiv.org/abs/2410.07166.

Yunn Kang Lim, Pengzhan Sun, Ziyi Bai, Xun Xu, Angela Yao, Xulei Yang, and Shijie Li. PRISM: Planning and reasoning with intent in simulated embodied environments, 2026. URL https://arxiv.org/abs/2605.11534.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is your multi-modal model an all-around player?, 2023. URL https://arxiv.org/abs/2307.06281.

Boyu Mi, Mengchen Ma, Yifei Yao, Xing Gao, Junting Chen, Yangzi Li, Zihou Zhu, Guohao Li, Zhenfei Yin, Tai Wang, Yao Mu, Jiangmiao Pang, and Hanqing Wang. Exploratory, communicative, and deployable: Vision-driven embodied agents for open-world mobile manipulation, 2026. URL https://arxiv.org/abs/2607.13653.

OpenAI. GPT-5.5 system card, April 2026. URL https://openai.com/index/gpt-5-5-system-card/.

Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. Reverie: Remote embodied visual referring expression in real indoor environments. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9982–9991, 2020.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Shreyas Sundara Raman, Vanya Cohen, Ifrah Idrees, Eric Rosen, Raymond Mooney, Stefanie Tellex, and David Paulius. Cape: Corrective actions from precondition errors using large language models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 14070–14077. IEEE, 2024.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 627–635. JMLR Workshop and Conference Proceedings, 2011.

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 *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10737–10746. IEEE, 2020.

Huajie Tan, Enshen Zhou, Zhiyu Li, Yijie Xu, Yuheng Ji, Xiansheng Chen, Cheng Chi, Pengwei Wang, Huizhu Jia, Yulong Ao, et al. RoboBrain 2.5: Depth in sight, time in mind. *arXiv preprint arXiv:2601.14352*, 2026. URL https://arxiv.org/abs/2601.14352.

Fei Wang, Xingyu Fu, James Y. Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, Tianyi Lorena Yan, Wenjie Jacky Mo, Hsiang-Hui Liu, Pan Lu, Chunyuan Li, Chaowei Xiao, Kai-Wei Chang, Dan Roth, Sheng Zhang, Hoifung Poon, and Muhao Chen. MUIRBench: A comprehensive benchmark for robust multi-image understanding, 2024. URL https://arxiv.org/abs/2406.09411.

Kangrui Wang, Linjie Li, Zhengyuan Yang, Shiqi Chen, Zihan Wang, Li Fei-Fei, Jiajun Wu, Leonidas Guibas, Lijuan Wang, and Manling Li. Planning with the views. *arXiv preprint arXiv:2605.29563*, 2026.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. URL https://arxiv.org/abs/2508.18265.

Peiran Xu, Sudong Wang, Yao Zhu, Jianing Li, and Yunjian Zhang. SpatialBench: Benchmarking multimodal large language models for spatial cognition, 2025. URL https://arxiv.org/abs/2511.21471.

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 Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 70576–70631. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/yang25f.html.

Chun-Hsiao Yeh, Chenyu Wang, Shengbang Tong, Ta-Ying Cheng, Ruoyu Wang, Tianzhe Chu, Yuexiang Zhai, Yubei Chen, Shenghua Gao, and Yi Ma. Seeing from another perspective: Evaluating multi-view understanding in MLLMs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026. URL https://arxiv.org/abs/2504.15280.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2311.16502.

Z.ai. GLM-4.6V. Hugging Face model card, 2025. URL https://huggingface.co/zai-org/GLM-4.6V.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. URL https://arxiv.org/abs/2504.10479.

Table A1: **Turn-level SFT responses by track.** Each row counts turn-level responses derived from replay-verified task continuations rather than independent trajectories. The full-data 4B and 2B SFT models share the resulting corpus.

Track	# SFT responses
Explore	51,910
Adapt	4,602
Recover	12,989
Verification	31,684
Cross-Room	10,779
Long-Horizon	10,145
Total	**122,109**

A Appendix

A.1 Data and benchmark construction

A.1.1 Training data

Data construction yields 11.5K replay-verified task continuations spanning search, verification, and continuation from incomplete states. Their trajectories supply the retained states and replay-verified actions used to form SFT responses and OPD state prompts. For ARL, we select 1,100 task instances from this continuation pool and collect eight fresh simulator rollouts per task. These counts refer to different units: 11.5K replay-verified task continuations, 122,109 turn-level SFT responses, 21,000 distinct OPD state prompts with 42,000 sampled responses, and 1,100 ARL tasks. Table A1 reports the SFT responses by primary track.

Explore and Verification account for 42.5% and 25.9% of SFT responses. Adapt is the smallest track at 3.8%. We discard examples with invalid coordinates, goals that are already satisfied or whose target instance cannot be resolved, missing images, failed replays, responses containing only a terminal statement, or hidden-information leakage. This leaves 122,109 turn-level SFT responses with no benchmark-scene overlap. For each retained state, Qwen3.5-397B-A17B receives the instruction, RGB–action context, and the next action from an A^* trajectory verified by simulator replay, then generates a rationale and structured action. The simulator validates the action, execution, and terminal goal, while filters validate the response schema and leakage constraints. The rationale itself is not simulator-verified. Policy inputs exclude environment responses, simulator errors, metric distances, target-instance IDs, and outcome labels throughout training and evaluation.

A.1.2 Benchmark construction

The benchmark instantiates scenes from the held-out ProcTHOR-10K test split in AI2-THOR (Kolve et al., 2017; Deitke et al., 2022). It does not import ready-made questions or trajectories from another benchmark. For each candidate task, the generator chooses a scene, a valid start state, a target instance, a language instruction, and a required state change, then applies track-specific conditions: same-category target–distractor pairs for Verification, room and path-length constraints for Cross-Room and Long-Horizon, and incomplete states from failed rollouts for Adapt and Recover. Each candidate is reset and executed in the simulator. Only candidates that pass the checks in Table A2 enter the 1,035-task evaluation set, and the scene metadata used during generation remains hidden from the policy. Each Recover task is defined by the failed state and observable rollout prefix at the handoff. The evaluated policy receives the instruction, current RGB observation, bounded RGB–action history. The evaluator enforces the action budget without exposing it to the policy. A handoff is retained only when the goal remains false and the task remains solvable.

The first five checks ensure that every retained task starts with a false goal and remains solvable under the shared action space. The final check ensures that all policies are ranked on identical track definitions and task sets. Recover and Adapt additionally require an incomplete yet still-solvable handoff, so these tracks evaluate continuation from unresolved states.

Table A2: **Benchmark construction checks.** The checks use planner and simulator state that remains hidden from the evaluated policy throughout every benchmark episode.

Check	Pass criterion
Task specification	Instruction names one target and requires an observable state change
Initial state	Scene resets, target exists, and the goal is initially false
Interaction feasibility	Target can be perceived and manipulated in the action space
Solvability	A planner trajectory reaches the target and satisfies the goal
Failure-state validity	For Recover and Adapt, the goal remains false yet a planner can still finish
Evaluation design	Track definitions and sizes stay fixed across all evaluated policies

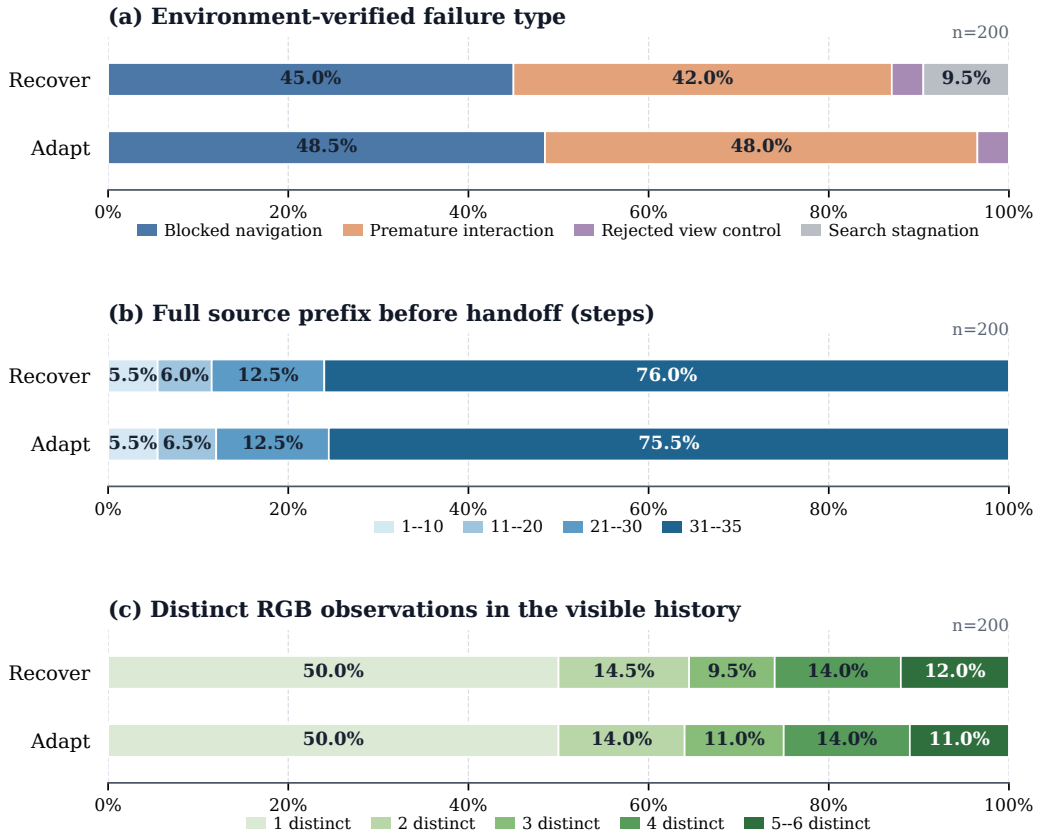


Figure A1: **Recover and Adapt handoffs and visible history.** Each row is normalized over the 200 tasks in its track. (a) Failure type is obtained from the simulator state at the handoff. (b) Prefix length is measured before the handoff. (c) Visual diversity counts distinct RGB observations among the at most six prior history frames shown to the policy at each fixed handoff.

A.1.3 Recover and Adapt handoffs

Figure A1 shows the failure-type, prefix-length, and visible-history distributions for the Recover and Adapt handoffs. The handoffs come from Qwen3.5-27B, Qwen3-VL-32B, InternVL3-38B, and InternVL3.5-38B rollouts, and the handoff set remains fixed across evaluated policies. Recover comprises blocked navigation (45.0%), premature interaction (42.0%), rejected view control (3.5%), and search stagnation (9.5%), while Adapt comprises incomplete states with a local-progress predicate—48.5% blocked navigation, 48.0% premature interaction, and 3.5% rejected view control (camera actions rejected by the simulator, not infrastructure failures). The source rollout prefix and the history shown to the evaluated policy are not the same: 76.0% of Recover and 75.5% of Adapt

Table A3: **ActiveBrain-Bench metrics.** Non-Adapt track metrics use separate fixed task sets. Adapt metrics share one set and mark successive outcomes within it. Every listed metric is higher-is-better under its stated success criterion and corresponding fixed task set.

Metric	Succeeds when	Fails when
Non-Adapt track metrics		
Explore↑	Finds the target and completes the instruction	Search, navigation, or interaction fails
Verification↑	Completes the instructed interaction on the intended instance	Acts on neither candidate, or on the distractor
Recover↑	Completes the instruction after the handoff	Cannot resume from the failed state
Cross-Room↑	Reaches the other room and completes the instruction	Fails to reach the other room or complete the instruction
Long-Horizon↑	Completes a longer action sequence	Exhausts the budget before completion
Adapt metrics		
Valid@1↑	First continuation action is executable	Emits no valid first action
LocalProgress@H↑	Meets the failure-specific progress predicate within H steps	No progress predicate is met
Adapt@H↑	First valid action differs from the ineffective handoff action and LocalProgress@H holds	Repeats that action, emits no valid first action, or makes no local progress
Goal@H↑	Completes the instruction within H steps	Instruction remains incomplete

handoffs fall at steps 31–35 in the prefix, yet the policy sees the current RGB frame plus at most six prior RGB–action pairs (fewer when the prefix is shorter). In exactly half of each track this window contains only one distinct RGB observation, usually because repeated actions did not change the view. The construction therefore keeps accumulated failures while bounding the history available to the policy under the shared interface used by the other tracks.

A.2 Evaluation protocol and metrics

ActiveBrain-Bench uses one policy interface, but scores its metrics on distinct fixed task sets. Each non-Adapt track has its own evaluation set, whereas all Adapt metrics share one set and describe successive outcomes within it. Table A3 summarizes each metric. Rates from different tracks are not subtracted because their fixed task identities differ.

Non-Adapt rows share one endpoint—completing the instruction on that track's fixed tasks—but expose different failures: search, instance choice, handoff continuation, room change, or horizon. Adapt metrics use one task set: Valid@1 scores first-step executability, LocalProgress@H scores local progress, the composite Adapt@H diagnostic additionally requires a changed first action, and Goal@H asks whether the continuation completes the instruction. Adapt Repeat@1 in Table A6 is the first-step repeat rate on the same Adapt evaluation tasks (lower is better). SR in Table 1 averages the five other task-success rates with Adapt Goal@20, not Adapt@H.

Failure-specific LocalProgress@H predicates. LocalProgress@H is true if the original task succeeds within the first H continuation steps, or if any one of those steps meets the predicate associated with the fixed handoff failure. For blocked navigation, the action must execute successfully and change the agent's planar position by at least $0.20\,\mathrm{m}$ in that step. For premature interaction, the step must reduce planar distance to the instructed target instance by at least $0.25\,\mathrm{m}$. Approaching the target is sufficient for local progress and a successful interaction is not required. For rejected view control, a subsequent action must have no protocol failure and execute successfully in the simulator. The action need not itself complete the task. Formally, with planar agent pose p_t, target distance d_t, simulator execution indicator e_t, and protocol-validity indicator v_t, the non-goal predicates are

$$\begin{aligned} \textsc{Blocked}: &\quad e_t = 1 \ \wedge\ \|p_{t+1} - p_t\|_2 \geq 0.20, \\ \textsc{Premature}: &\quad d_t - d_{t+1} \geq 0.25, \\ \textsc{Rejected-view}: &\quad e_t = 1 \ \wedge\ v_t = 1. \end{aligned} \tag{A1}$$

Table A4: **Complete ActiveBrain-Bench results.** Per-track success rates (%); Adapt reports Goal@20. The final column gives their unweighted mean (SR), which determines the global rank (ties share the minimum rank). The Qwen3.5-27B row uses predefined cutoffs from its 100-action evaluation: 20 actions for Adapt and 35 for the other tracks. Exp/Ver/Rec/Adp/CR/LH denote Explore, Verification, Recover, Adapt, Cross-Room, and Long-Horizon. ActiveBrain-2B and ActiveBrain-4B are their final ARL checkpoints; their rows are shaded; gold/silver/bronze mark the top three values; the best zero-shot entry is underlined. ↑: higher is better; ↓: lower is better.

Policy	**Rank↓**	**Core AC tracks**			**Extended task settings**			**SR↑**
		Exp↑	Ver↑	Rec↑	Adp↑	CR↑	LH↑	
Zero-shot policies								
Qwen3.5-4B	13	15.0	1.5	17.0	1.0	5.9	8.0	8.1
Qwen3.5-9B	16	7.5	1.5	4.0	0.0	4.7	4.0	3.6
Qwen3.5-27B	4	55.0	15.5	38.0	39.0	16.5	34.0	33.0
Qwen3.5-35B-A3B	11	14.0	4.0	10.5	5.0	10.6	17.3	10.2
Qwen3.5-122B-A10B	5	30.0	9.5	26.0	29.5	23.5	42.0	26.8
Qwen3.5-397B-A17B	6	33.0	8.0	16.0	13.5	10.6	29.3	18.4
Qwen3-VL-4B	14	15.5	1.0	3.0	0.5	2.4	6.7	4.9
Qwen3-VL-8B	19	2.0	1.0	1.5	1.5	0.0	1.3	1.2
Qwen3-VL-30B-A3B	8	27.5	7.5	14.0	0.0	12.9	23.3	14.2
Qwen3-VL-32B	10	35.5	6.0	0.5	0.0	12.9	22.7	12.9
Qwen3-VL-235B-A22B	7	47.0	9.0	5.0	0.0	11.8	23.3	16.0
RynnBrain-30B-A3B	15	7.5	3.5	6.0	0.0	4.7	7.3	4.8
RynnBrain1.1-9B	12	19.0	2.5	14.0	7.0	4.7	12.7	10.0
GLM-4.6V	9	29.0	7.0	15.5	0.0	10.6	17.3	13.2
Qwen2.5-VL-7B	18	1.5	0.5	2.5	3.5	1.2	2.0	1.9
RoboBrain2.5-4B	17	0.5	1.0	11.0	0.0	0.0	1.3	2.3
RoboBrain2.5-8B-NV	21	0.0	0.0	3.5	0.0	0.0	0.0	0.6
RynnBrain-8B	19	0.0	0.0	7.0	0.0	0.0	0.0	1.2
GPT-5.5	3	52.0	**19.0**	41.0	39.5	16.5	34.0	33.7
ActiveBrain								
ActiveBrain-4B	**1**	**63.0**	15.5	**53.0**	**42.5**	**35.3**	**54.7**	**44.0**
ActiveBrain-2B	2	56.0	9.0	42.0	30.5	25.9	46.7	35.0

These are local transition criteria: they record displacement from a blocked pose, progress toward the target after a premature interaction, or successful execution after a rejected command. They neither require completion nor establish that the original cause of failure has been removed. Goal@H reports task completion on the same episodes and at the same horizon.

Table A4 reports results for all evaluated policies.

A.3 Training and evaluation implementation

This section specifies the shared policy interface and the three-stage training setup used for all reported ActiveBrain models. The remaining appendix sections analyze zero-shot and trained behavior under the same interface and fixed task definitions.

A.3.1 Policy input and bounded memory

The prompt below is the policy-input renderer used for both training and evaluation. Angle brackets mark episode-specific values. They are not extra annotations.

ActiveBrain Policy Prompt

Role and observable evidence. Act as an embodied agent in a partially observable 3D environment. At each step, use the instruction, the current egocentric RGB frame and up to six prior frames in chronological order, the action preceding each frame, and bounded summaries computed from those action calls.

Allowed actions.

- `navigate`: forward/backward motion or left/right rotation, with a positive magnitude.
- `look`: up/down camera tilt, with a positive magnitude.
- `interact_pixel`: `activate`, `deactivate`, `open_access`, `close_access`, or `pick` at normalized coordinates $(x, y) \in [0, 1000]^2$.

Visual and history rules. Use the current RGB frame to determine visibility and interaction coordinates. Use earlier RGB–action pairs to track motion, repeated behavior, search coverage, and previous interaction attempts. Treat same-category objects as candidates unless visible room, support, or state cues identify the requested instance.

Rendered user message. Frame $t-k$ (role=history, previous_action=a_{t-k-1}) <image>
...
Frame t (role=current, previous_action=a_{t-1}) <image>
Task: <instruction>
Action Replay: <recent actions; repetition and alternation>
Spatial Belief: <frame count; signed heading change; recent search actions>
Object Evidence: <three most recent interaction intents>

Constraints and output. No action limit, remaining-step count, simulator result flags, metric distances, target-instance identifiers, oracle actions, or verifier state are provided. Return an optional <think>...</think> block followed by exactly one JSON object containing only `skill_name` and `arguments`; do not return markdown, result claims, or additional actions.

A.3.2 Action parser and protocol failures

The policy returns an optional <think> block followed by one JSON invocation with keys `skill_name` and `arguments`. The parser removes the reasoning tags, extracts one JSON payload, normalizes Unicode quotation marks and trailing commas, and accepts the official primitive aliases for motion, rotation, and camera tilt, but does not fill in a missing skill or required argument. The invocation is then checked against three executable skills: `navigate` with one of {`forward`, `backward`, `turn_left`, `turn_right`} and a positive magnitude; `look` with direction {`up`, `down`} and a positive magnitude; and `interact_pixel` with one of the five task intents and finite image coordinates. Model-specific chat templates, image preprocessing, and coordinate normalization follow the official configuration for each model before the invocation is mapped to the common simulator action. Empty, malformed, multi-action, unsupported, and out-of-range outputs consume one action step but do not change the environment: we record them as protocol failures, count them toward the fixed invalid-action budget, and never repair an action on the policy's behalf. For the analysis in Figure A6, any invocation outside this shared three-action interface is counted as an output-schema failure rather than treated as an additional action.

A.3.3 SFT, OPD, and ARL training

The Qwen3.5-397B-A17B data-generation teacher produces the rationale–action targets. After quality filtering, the ActiveBrain-SFT corpus contains 122,109 turn-level responses shared by the 4B and 2B SFT models. For Qwen3.5-4B, SFT freezes the vision encoder and fully fine-tunes the remaining parameters for one epoch with AdamW, learning rate 10^{-5}, cosine decay, 10% warmup, weight decay 0.1, and maximum sequence length of 20,480 tokens.

For each backbone, the SFT model is frozen as the OPD teacher while the OPD student starts independently from the original base model; teacher and student share architecture and capacity. OPD transfers the teacher's token distributions to contexts produced by the student initialized from the base model. Teacher and student have the same capacity, so OPD does not compress the model. Instead, it trains the student on its own response prefixes rather than only on fixed teacher-forced contexts. Here, *on-policy* refers only to response prefixes sampled from the student; those responses

Table A5: **Training configuration for ActiveBrain.** SFT uses the same corpus and schedule for both model sizes; + ARL denotes agentic reinforcement learning.

Setting	SFT (4B / 2B)	OPD	+ ARL
Initialization	Qwen3.5-4B / Qwen3.5-2B	Student initialized from base model; frozen same-size SFT teacher	OPD student
Training data	122,109 turn-level SFT responses	42,000 sampled responses from 21,000 distinct state prompts	1,100 interactive simulator tasks
Training passes	1 epoch	1 epoch	1 pass over the task set
Batch / rollouts	Global batch size 64	Batch size 32; one student sample per prompt presentation	Eight sibling rollouts per task; turn-level updates
Optimizer / schedule	AdamW; lr 10^{-5}; weight decay 0.1; cosine; 10% warmup	Adam; lr 2×10^{-6}; 10 warmup updates	Adam; lr 3×10^{-7}; grad clip 0.5; cosine; 3% warmup
Context / sampling	20,480-token context; teacher-forced targets	Prompt 18,432 / response 2,048; temperature 0.8; top-p 0.95; top-k 50	Context 20,480 / response 512; $\leq$20 turns; temperature 0.8; top-p 0.95; top-k 50
Learning signal	Next-token likelihood	Generalized JS (α=0.5) over student top-100 tokens + tail bucket	AC-Reward ($+0.2$ for newly detected progress); intra-/inter-tree relative advantage; PPO clip 0.1; KL 0.05

are not executed in the simulator during OPD. OPD presents 21,000 distinct state prompts twice. It samples one student response per presentation, yielding 42,000 sampled responses with maximum response length 2,048. At every student-generated prefix, it computes teacher and student token distributions from the same observable context, retaining the student's top-100 tokens plus a tail-probability bucket and minimizing their generalized Jensen–Shannon divergence with $\alpha = 0.5$. OPD uses no task reward, reward model, or policy-gradient loss at this stage.

ARL then starts from the OPD student and trains on 1,100 interactive simulator tasks with the progress-reward Tree-GRPO objective described in Section 4. Each task produces eight sibling rollouts, organized into trees whose branches share the interaction prefix before their actions diverge. Neither OPD nor ARL exposes simulator-side audit signals to the policy. Table A5 summarizes the three-stage configuration and its optimization settings.

Only ARL uses environment returns; OPD receives no task reward.

A.3.4 Training objectives

ActiveBrain-SFT minimizes the standard next-token objective over an observable policy input x_t and a response target $u_t^* = (r_t, a_t^*)$, where r_t is the filtered Qwen3.5-397B-A17B rationale written after the teacher receives the next action a_t^* from a simulator-verified trajectory. The simulator checks a_t^*, not the rationale; we do not treat the rationale as a verified account of the policy's internal reasoning:

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = -\mathbb{E}_{(x_t,u_t^*)\sim\mathcal{D}_{\mathrm{ActiveBrain}}} \log p_\theta(u_t^* \mid x_t).$$

For a response $y \sim p_\theta(\cdot \mid x)$ sampled from the student policy, let $c_t = (x, y_{<t})$ denote the student-generated prefix. We write $\widetilde{p}_\theta(\cdot \mid c_t)$ and $\widetilde{p}_T(\cdot \mid c_t)$ for the student and frozen-teacher distributions over the student's top-100 tokens plus a shared tail bucket, and define their mixture as

$$m_t = \tfrac{1}{2}\widetilde{p}_\theta(\cdot \mid c_t) + \tfrac{1}{2}\widetilde{p}_T(\cdot \mid c_t).$$

The OPD objective is

$$\mathcal{L}_{\mathrm{OPD}}(\theta) = \mathbb{E}_{x,\, y\sim p_\theta(\cdot|x)} \sum_t \frac{1}{2}\left[D_{\mathrm{KL}}(\widetilde{p}_\theta(\cdot \mid c_t) \,\|\, m_t) + D_{\mathrm{KL}}(\widetilde{p}_T(\cdot \mid c_t) \,\|\, m_t)\right].$$

AC-Reward progress signal. The simulator verifier tracks the progress events illustrated in Figure 3: observing a previously unseen view, making the instructed target instance visible, making that instance interactable, and satisfying the task goal. A target-visible or target-interactable event requires an exact target-instance match; the latter additionally requires support for the instructed interaction, distance at most 1.5, and a detection box in the current frame. A same-category distractor therefore does not count as target progress. A turn receives a $+0.2$ progress reward when it triggers a progress event not previously credited in that rollout, and receives zero otherwise. All predicates are computed from simulator-side metadata by the verifier. They are used only to form the training reward and are never included in the policy input during training or evaluation.

Tree-GRPO return and relative advantage. Let H_i denote one of the eight rollouts for task q. We first compute the discounted return-to-go from the progress rewards:

$$G_{i,t} = \sum_{u=t}^{T_i-1} \gamma_{\text{tree}}^{u-t} r_{i,u}^{\text{prog}}, \qquad \gamma_{\text{tree}} = 0.97. \tag{A2}$$

The scalar return for a completed rollout is $R(H_i) = G_{i,0}$. Following Tree-GRPO (Ji et al., 2026), let $\mathcal{G}_{\text{intra}}(T_i)$ contain the completed rollouts in the same rollout tree as H_i, and let $\mathcal{G}_{\text{inter}}(q)$ contain the rollouts across all trees for task q. For either non-degenerate group $g \in \{\text{intra}, \text{inter}\}$, the group-relative advantage is

$$\widehat{A}_g(H_i) = \frac{R(H_i) - \text{mean}_{H_j \in \mathcal{G}_g} R(H_j)}{\text{std}_{H_j \in \mathcal{G}_g} R(H_j)}. \tag{A3}$$

The two estimates are added as in Tree-GRPO:

$$A^{\text{tree}}(H_i) = \widehat{A}_{\text{intra}}(H_i) + \widehat{A}_{\text{inter}}(H_i). \tag{A4}$$

If a group has no return variance, it contributes zero relative advantage. Because branches share the policy calls before a divergence, descendant returns provide relative credit to the action chosen at that branch. No separate depth-relative advantage is added.

Proximal policy optimization update. All action tokens generated along rollout H_i use its tree-relative advantage. With token-level policy ratio $\rho_{i,k}(\theta)$, the clipped Tree-GRPO objective is

$$\mathcal{L}_{\text{PPO}} = -\mathbb{E}\left[\min\left(\rho_{i,k}(\theta) A^{\text{tree}}(H_i), \text{clip}(\rho_{i,k}(\theta), 0.9, 1.1) A^{\text{tree}}(H_i)\right)\right]. \tag{A5}$$

We add a Kullback–Leibler (KL) penalty of 0.05 to the frozen SFT reference policy. Optimization uses Adam with learning rate 3×10^{-7}, cosine decay, 3% warmup, and gradient clipping at 0.5.

A.3.5 Evaluation interface

All closed-loop evaluations use AI2-THOR (Kolve et al., 2017), identical per-track tasks and budgets, and the shared policy interface. Explore, Cross-Room, and Long-Horizon contain 200, 85, and 150 disjoint cases; Verification, Recover, and Adapt contain 200 each. The policy receives the instruction, current RGB frame, at most six prior RGB–action pairs, and bounded summaries derived only from observable history. Action-execution flags, distances, errors, target IDs or coordinates, planner actions, and planner state remain hidden.

Model sources. The evaluated policies include Qwen3.5 (Qwen Team, 2026), Qwen3-VL (Bai et al., 2025a), Qwen2.5-VL (Bai et al., 2025b), RynnBrain (Dang et al., 2026), RynnBrain1.1 (Li et al., 2026), RoboBrain2.5 (Tan et al., 2026), GLM-4.6V (Z.ai, 2025), and GPT-5.5 (OpenAI, 2026). Recover and Adapt handoff collection additionally uses InternVL3 (Zhu et al., 2025) and InternVL3.5 (Wang et al., 2025).

A.4 Zero-shot policy analysis

This section examines zero-shot behavior through illustrative rollouts, an instance-level Verification failure, terminal outcomes, repeated actions after failure, and whether local progress on Adapt or continuation on Recover leads to task success. Figures A2 and A3 use the same evaluation runs as

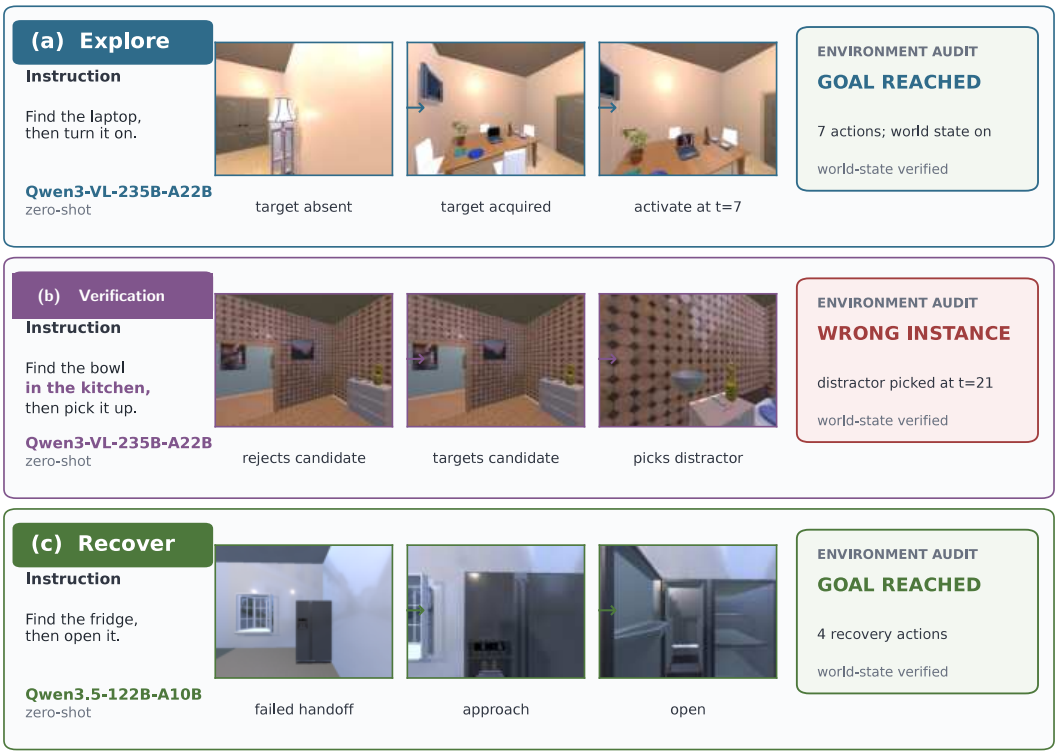


Figure A2: **Zero-shot behavior on Explore, Verification, and Recover. Explore:** Qwen3-VL-235B-A22B finds a previously unseen laptop and completes the interaction. **Verification:** the same policy rejects a visible bowl in language, then selects that distractor. **Recover:** Qwen3.5-122B-A10B resumes from a failed handoff and opens the refrigerator. Right-hand cards are simulator audits that remain unavailable to the evaluated policy throughout interaction.

the quantitative tables; RGB frames and action labels are policy-visible, while outcome cards are environment audits shown only to the reader. The examples are illustrative and do not affect ranking.

The core cases already separate the three demands: Qwen3-VL-235B-A22B succeeds on Explore but fails Verification even after reaching the relevant object category, whereas Qwen3.5-122B-A10B resumes a failed handoff on Recover. Figure A3 extends this contrast: Qwen3.5-4B clears the immediate Adapt obstruction without completing the instruction, while other policies search across rooms or complete a long trajectory. Together, these examples mirror the quantitative separation among search, instance verification, local progress, and completion of the instruction within and across policies under the shared interface, fixed task sets, and common action budgets.

A.4.1 Instance verification audit

The audit compares the object ID resolved by each executed `interact_pixel` action with the target–distractor pair, separating failure to select either candidate from selection of the wrong instance after a resolved interaction. Figure A4 shows the Qwen3-VL-235B-A22B case from Figure A2; the same check is applied to SFT and OPD in Section A.6.

The executed interaction in Figure A4 is syntactically valid and reaches an object of the correct category, but the resolved object ID is the distractor. This documents an instance-verification error rather than an output-format or category-search failure: verbal rejection of a candidate is not sufficient unless the eventual pixel interaction is grounded to the instruction-specified target.

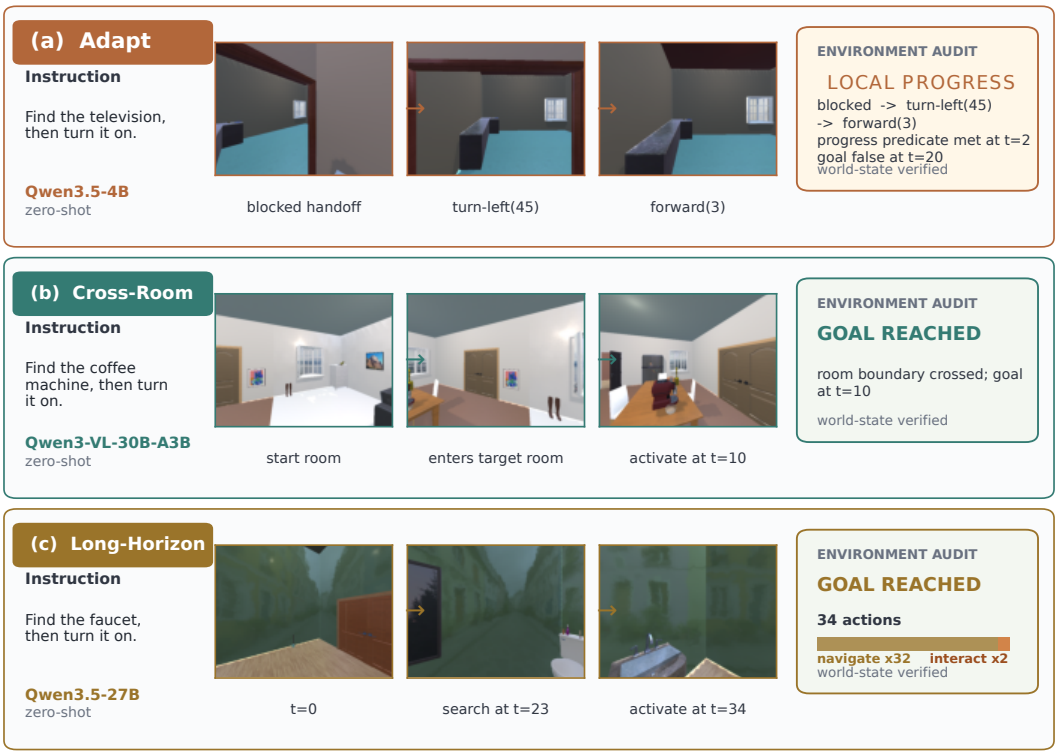


Figure A3: **Zero-shot behavior on Adapt, Cross-Room, and Long-Horizon. Adapt:** Qwen3.5-4B clears a local blockage but leaves the goal incomplete. **Cross-Room:** Qwen3-VL-30B-A3B reaches the target room and finishes. **Long-Horizon:** Qwen3.5-27B follows a long navigation sequence and completes the faucet objective within the fixed interaction budget.



Figure A4: **Audited instance-verification failure.** The instruction asks for the kitchen bowl; the policy later interacts with the paired distractor. Object and room IDs appear only in the audit.

A.4.2 Terminal outcomes and action repetition

Figure A5 reports stopping labels on the five non-Adapt tracks; Figure A6 aggregates each policy's diagnostic profile and includes ActiveBrain-4B after SFT and OPD. Both use the criteria and cutoffs in Table 1 for every evaluated policy and checkpoint; Figure A6 uses Goal@20 on Adapt.

Budget exhaustion dominates Verification, Recover, and Cross-Room (50.4%, 52.7%, and 56.5%). For the 18 open-weight zero-shot policies on ActiveBrain-Bench, pooled terminal outcomes are 51.6% budget exhaustion, 18.3% output-schema failure, 13.1% executable-action failure, 10.3% task success, and 6.6% other; 49.8% of episodes contain at least one protocol violation. Protocol

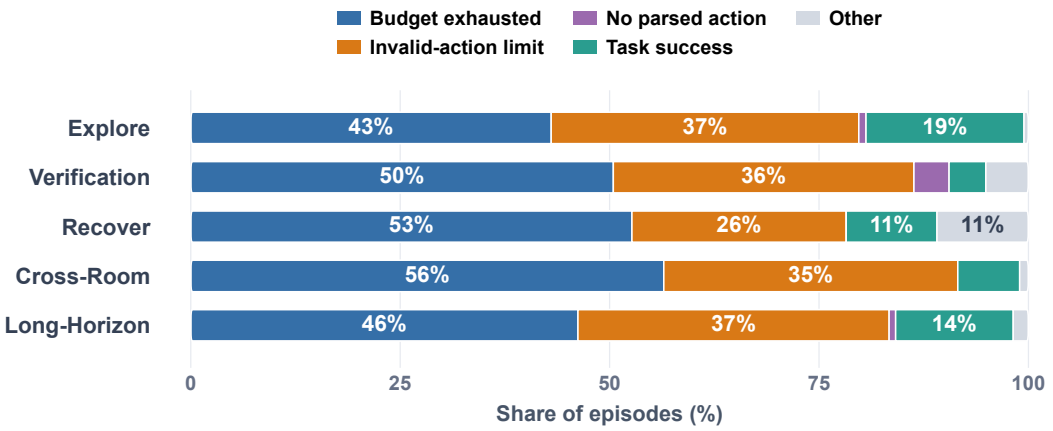


Figure A5: **Task-level zero-shot failure profiles.** Mutually exclusive terminal outcomes for open-weight policies on the five tracks excluding Adapt. Task success uses the criterion and cutoff in Table 1; the other segments use the terminal outcome category.

errors dominate several weak policies, but compliance is insufficient: Qwen3-VL-32B and GLM-4.6V have violation rates of only 0.3% and 2.1%, yet complete just 12.5% and 13.3% of episodes; 69.9% and 74.0% instead exhaust the action budget on their fixed tasks under the shared evaluation protocol.

For Qwen3.5-4B, successes in the corresponding diagnostic profiles increase from 86 zero-shot to 403 after SFT and 450 after OPD. Terminal schema failures change from 104 to 0 and 1, executable-action failures from 267 to 12 and 9, and episodes with any protocol violation from 664 to 8 and 1. Budget exhaustion accounts for 58.9% and 55.6% of episodes in the SFT and OPD diagnostic profiles, respectively. Because OPD adds 47 successes after SFT has already eliminated terminal schema errors, its gain cannot be attributed to format repair alone. Conversely, the same zero score may reflect invalid actions, budget exhaustion, or local Adapt progress without Goal@20.

Terminal labels omit repeated behavior. Across open-weight policies, 80,201 of 152,777 executable follow-ups after a failed action exactly repeat the previous action (52.5%). GPT-5.5 repeats 454 of 10,626 such follow-ups (4.3%). Table A6 reports adjacent, after-failure, episode-level, and Adapt Repeat@1 rates computed from the evaluation trajectories.

After-failure repeat rates are 92.1% for RoboBrain2.5-8B-NV, 89.0% for RynnBrain-8B, and 74.5% for RynnBrain-30B-A3B, versus 4.3% for GPT-5.5. Adapt Repeat@1 is often lower because a policy may initially change course but repeat later. Local progress can likewise diverge from completion: Table A7 splits RynnBrain1.1-9B's Adapt results by failure type.

Overall LocalProgress@20 is 60.5%, but Goal@20 is 7.0%. After premature interaction the rates are 43.8% and 1.0%; after blocked navigation, 74.2% of episodes make local progress but only 12.4% finish. Figures A7 and A8 compare two policies that differ substantially on Recover task completion under the shared evaluation interface.

On premature-interaction Adapt handoffs, Qwen3-VL-32B always changes the first action and makes local progress in 52.1% of episodes, yet never reaches Goal@20; Qwen3.5-122B-A10B changes in 75.0%, makes local progress in 66.7%, and completes 38.5%. On the matched Recover handoffs, Qwen3-VL-32B immediately repeats the failed interaction in 82.1% of states and never finishes, whereas Qwen3.5-122B-A10B repeats in 44.0% and later completes 47.6% of those states. Changing course and clearing a local obstruction therefore do not by themselves decide whether the instruction is completed within the evaluation budget.

A.4.3 Step-by-step executed-action trace

Figure A9 shows a complementary failure pattern: valid actions that still hit the wrong pixels.

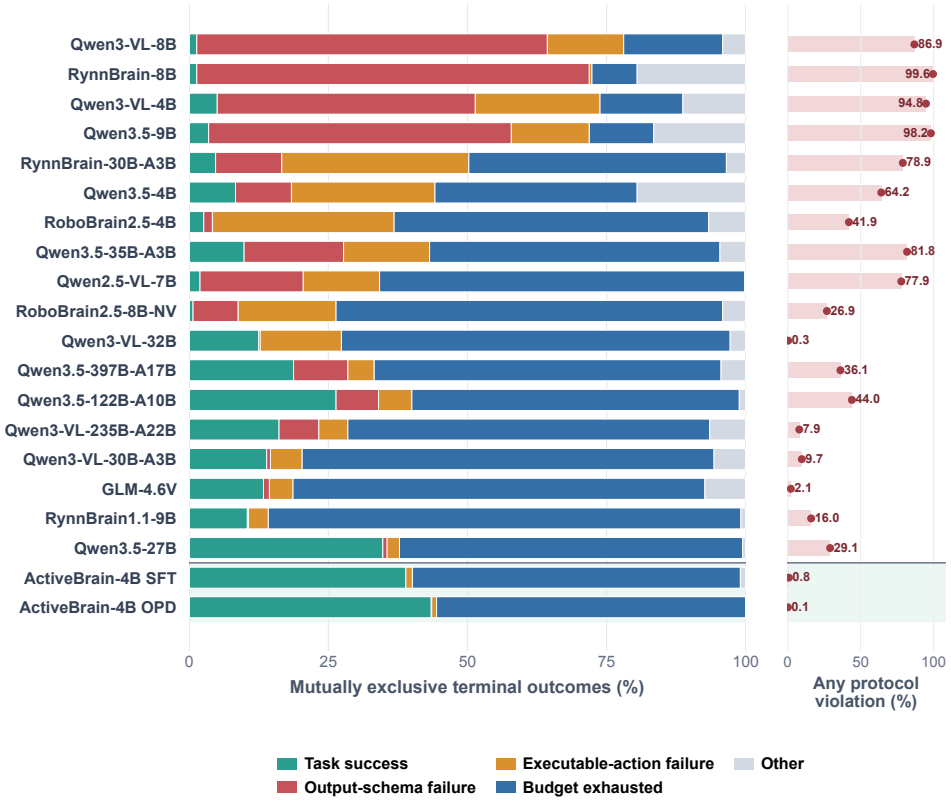


Figure A6: **Model-level failure profiles before and after training.** Five mutually exclusive terminal outcomes in each policy's diagnostic profile, including zero-shot policies and ActiveBrain-4B after SFT and OPD. Success is track-specific and uses Goal@20 on Adapt; local progress alone is not success. Output-schema failure denotes a terminal attempt rejected by the shared `navigate`/`look`/`interact_pixel` interface, which contains no separate reporting action. The right panel separately reports episodes with any protocol violation and can therefore overlap all terminal outcomes. "Other" includes Adapt episodes that stop before the 20-step horizon without Goal@20. Trained rows appear below the separator; these episode-pooled profiles are diagnostic rather than the reported equal-weight benchmark score.

In a Qwen3.5-397B-A17B Adapt episode, two attempted faucet interactions resolve to the wall. Both interaction commands pass the parser, and the policy changes action after the first failure; protocol validity and action switching therefore do not reveal this geometric grounding error, which only appears after resolving the selected pixels in the environment rather than from output syntax or action switching alone.

A.5 Interaction-length and Adapt-horizon analyses

We next examine how interaction length and local Adapt progress relate to completion. Tables A8 and A9 report interaction length and the full Adapt-horizon breakdown. Model size alone is not analyzed separately because Table 1 is already non-monotonic within the Qwen3.5 family. Step counts summarize trajectory length rather than efficiency-normalized returns. The interaction-length table includes SFT and OPD, while the Adapt-horizon table also includes ARL. All trained models use the same fixed tasks and action budgets as the zero-shot policies used throughout the main comparison.

Among open-weight policies, mean length over all episodes ranges from 13.7 (RynnBrain-8B) to 30.4 (RynnBrain1.1-9B) steps. The successful-only means are consistently lower because completed tasks stop as soon as the goal is reached. The SFT and OPD models have similar mean interaction

Table A6: **Repeated-action rates across policies (%).** *Adjacent* is an exact repeat of the previous executed action; *After-failure* restricts this rate to simulator failures; and *Episodes w/ repeat* is the fraction of rollout episodes with at least one such repeat. *Repeat@1* (Adapt only) records whether the first continuation repeats the failed handoff action. Action-level rates use model-dependent denominators; all four reported repetition metrics are lower-is-better.

	Five tracks excl. Adapt			Adapt
Policy	Adjacent↓	After-failure↓	Episodes w/ repeat ↓	Repeat@1↓
Zero-shot policies				
Qwen3.5-4B	39.4	45.2	88.1	30.0
Qwen3.5-9B	27.8	37.1	41.3	16.0
Qwen3.5-27B	30.6	43.9	56.8	16.5
Qwen3.5-35B-A3B	20.7	22.9	65.7	30.0
Qwen3.5-122B-A10B	21.5	28.2	67.4	23.0
Qwen3.5-397B-A17B	20.9	11.9	27.9	10.0
Qwen3-VL-4B	19.2	23.3	52.3	4.0
Qwen3-VL-8B	26.4	27.8	29.8	16.5
Qwen3-VL-30B-A3B	40.4	52.7	93.9	25.0
Qwen3-VL-32B	54.1	68.9	89.0	21.0
Qwen3-VL-235B-A22B	34.1	52.2	70.7	8.5
RynnBrain-30B-A3B	61.3	74.5	79.0	21.0
RynnBrain1.1-9B	44.3	62.3	48.6	9.0
GLM-4.6V	29.0	36.2	82.2	21.0
Qwen2.5-VL-7B	43.9	55.3	63.8	9.0
RoboBrain2.5-4B	70.9	79.4	78.3	6.0
RoboBrain2.5-8B-NV	86.1	92.1	86.2	46.5
RynnBrain-8B	85.7	89.0	18.6	0.5
GPT-5.5	7.4	4.3	26.5	4.0
ActiveBrain-4B				
SFT	3.1	2.1	6.6	0.0
OPD	1.8	0.3	1.0	0.0

Table A7: **RynnBrain1.1-9B by Adapt failure type.** All tasks use a 20-action budget; columns split the handoffs by failure type (n in parentheses). LocalProgress@20 and Goal@20 follow Table A3, and higher is better for both metrics on the fixed 200-task Adapt set.

Metric	Blocked navigation (n=97)	Premature interaction (n=96)	Rejected view control (n=7)	All (n=200)
LocalProgress@20 (%)↑	74.2	43.8	100.0	60.5
Goal@20 (%)↑	12.4	1.0	14.3	7.0

lengths (25.6 and 24.7 steps over all episodes), so their performance difference is not explained by systematically longer trajectories or a different evaluation budget.

LocalProgress@20 often substantially exceeds Goal@20: Qwen3-VL policies reach 58–66.5% local progress but at most 1.5% completion, and GLM-4.6V reaches 66.5% versus 0%. Training raises Qwen3.5-4B from 58.5%/1.0% (LocalProgress/Goal) to 85.0%/36.0% after SFT, 80.0%/40.5% after OPD, and 81.0%/42.5% after ARL. The OPD and ARL models have higher Goal@20 than the SFT model without a corresponding increase in LocalProgress@20. For the zero-shot model, only 2 of 117 episodes meeting LocalProgress@20 complete the task, motivating the use of Goal@20 in SR.

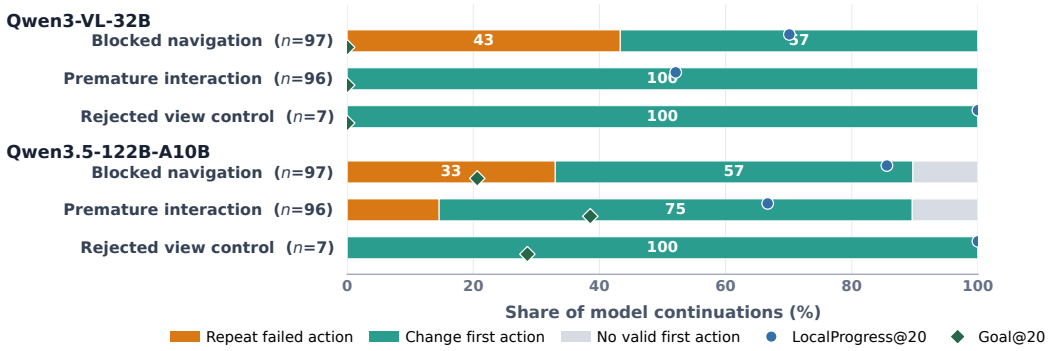


Figure A7: **Adapt behavior for two representative zero-shot policies.** Bars show the first continuation (repeat, change, or no valid action); circles show LocalProgress@20 and diamonds show Goal@20 on the same fixed Adapt handoffs and 20-step action budget.

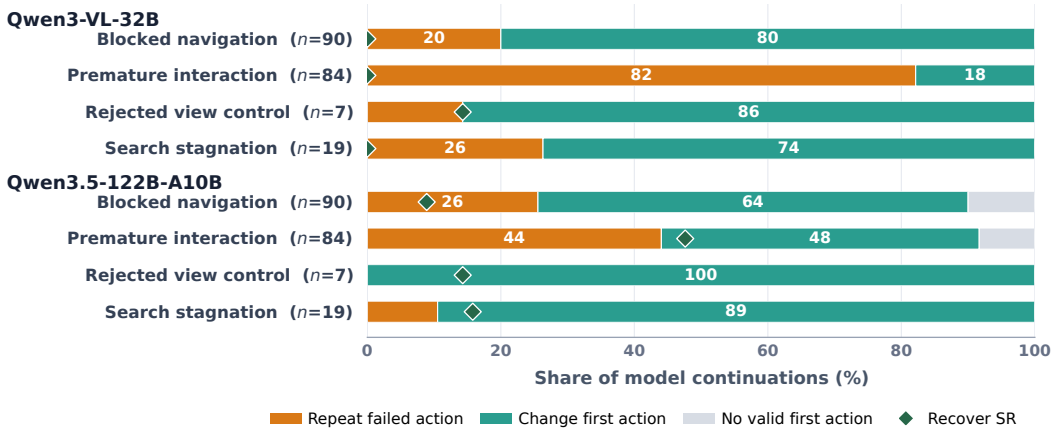


Figure A8: **Recover behavior for the same two policies.** Bars use the categories in Figure A7; diamonds mark Recover task success. Rates should be compared only within this figure.

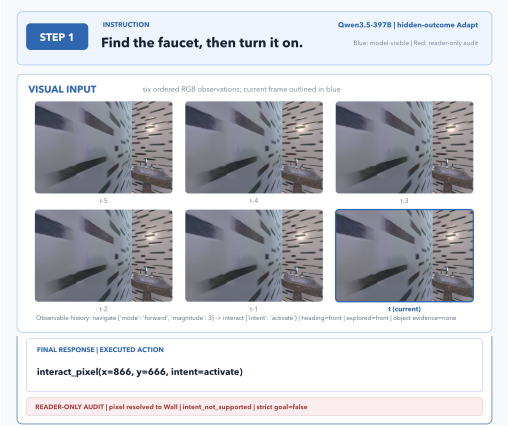
A.6 From zero-shot to trained policies

Table A10 places zero-shot Qwen3.5-4B and the ActiveBrain-4B models from the training pipeline on the same fixed track sets. OPD starts from the base model with a frozen SFT teacher; ARL starts from the OPD model. The columns follow this order and are not single-factor ablations.

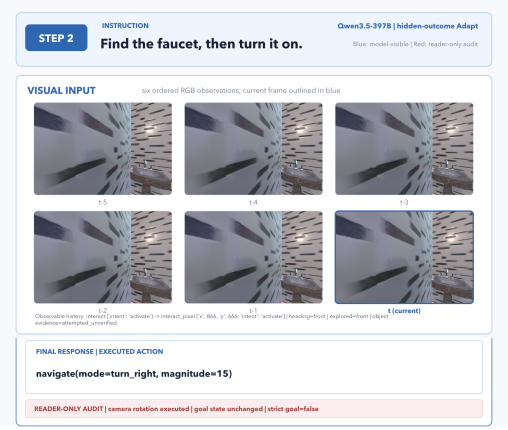
The reported SR values differ by +30.7, +4.3, and +0.9 points between adjacent columns. The largest zero-shot-to-SFT gains occur on Long-Horizon (+42.7), Explore (+40.5), and Adapt Goal@20 (+35.0). Later changes are non-uniform: relative to SFT, the OPD model is higher on Explore and Long-Horizon but lower on Recover; relative to OPD, the ARL model is higher on Recover and Verification but lower on Explore and Long-Horizon. On the same 200 Adapt handoffs, Goal@20 increases from 1.0% to 36.0%, 40.5%, and 42.5%, directly measuring completion of the instruction after failure.

A.6.1 History intervention on Adapt

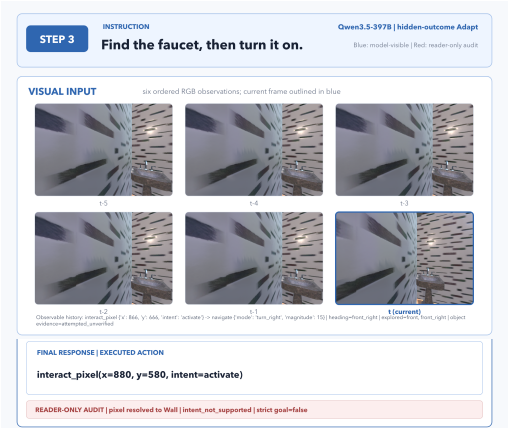
Table A11 reports a paired input intervention for the SFT model on the same 200 fixed Adapt handoffs; Table 4 reports the corresponding intervention for the final ARL model in the main paper. Full uses the complete policy input; Raw history removes the handcrafted Action Replay, Spatial Belief, and Object Evidence summaries but retains the chronological RGB–action history; Current only retains the current RGB observation; and Shuffled preserves the chronological RGB sequence but



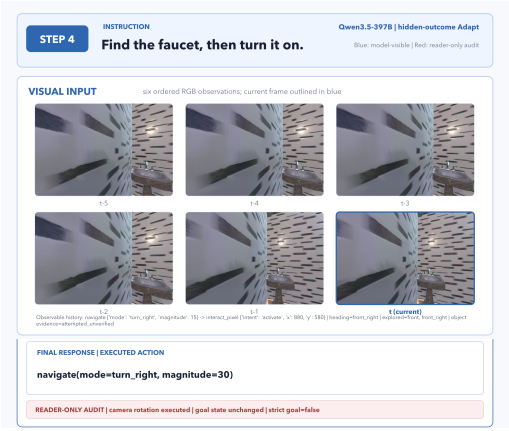
(Step 1 of 4) The policy receives the instruction, current RGB observation, and bounded history. It emits a response and a valid pixel interaction, but the coordinate resolves to the wall. The red audit row is not part of the policy input and is displayed only for post-hoc analysis; it is never returned to the policy after simulator execution completes.



(Step 2 of 4) After the failed interaction, the policy reorients the camera with a valid navigation action, but the task goal remains false.



(Step 3 of 4) The changed viewpoint does not correct the object-to-pixel selection: the second valid interaction again resolves to the wall rather than the faucet, so the intervening camera motion still leaves the task incomplete at this step.



(Step 4 of 4) The policy continues to search and issues another valid camera rotation, but the goal remains unmet after the fourth displayed step.

Figure A9: **Executed-action failure trace (steps 1–4).** Zero-shot Qwen3.5-397B-A17B Adapt episode. Each panel shows the instruction, RGB–action context, executed action, and post-hoc environment audit. The audit is not a policy input at any interaction step.

deterministically misaligns the historical action labels. All four conditions use identical task IDs and starting handoffs, so every reported comparison is paired by task.

Removing handcrafted summaries changes SFT Goal@20 from 36.0% to 33.0% (McNemar $p = 0.504$; bootstrap 95% CI $[-4.5, 10.5]$), whereas removing raw history lowers it to 8.5% (Raw–Current: +24.5 points, $p = 1.36 \times 10^{-10}$, CI $[17.5, 31.5]$). Shuffling action labels while preserving the RGB sequence yields 20.5% (Raw–Shuffled: +12.5 points, $p = 0.0022$, CI $[5.0, 20.0]$). The intervention changes alignment on 3,616 of 3,733 eligible turns across all 200 episodes. By contrast, LocalProgress@20 differs by only 4.5 points between Raw and Shuffled ($p = 0.262$; CI $[-2.5, 11.5]$). Chronological observations therefore help local progress, while aligned RGB–action history is more clearly reflected in completion of the instruction.

Table A8: **Mean interaction length across policies (steps).** Results cover the five tracks excluding Adapt; successful-only means include episodes that satisfy the task-success criterion in Table 1.

Policy	All episodes	Successful only
Zero-shot policies		
Qwen3.5-4B	23.3	13.5
Qwen3.5-9B	19.6	10.4
Qwen3.5-27B	27.6	14.5
Qwen3.5-35B-A3B	25.1	15.2
Qwen3.5-122B-A10B	25.2	14.7
Qwen3.5-397B-A17B	26.1	16.4
Qwen3-VL-4B	19.1	12.3
Qwen3-VL-8B	19.3	10.3
Qwen3-VL-30B-A3B	29.2	21.0
Qwen3-VL-32B	27.4	16.6
Qwen3-VL-235B-A22B	26.5	15.4
RynnBrain-30B-A3B	22.6	13.4
RynnBrain1.1-9B	30.4	17.3
GLM-4.6V	29.1	17.7
Qwen2.5-VL-7B	29.2	18.2
RoboBrain2.5-4B	26.8	7.3
RoboBrain2.5-8B-NV	27.0	1.6
RynnBrain-8B	13.7	3.6
GPT-5.5	25.6	16.2
ActiveBrain-4B		
SFT	25.6	14.6
OPD	24.7	15.0

ARL shows the same ordering: Goal@20 is 42.5% with Full, 39.0% with Raw, 24.0% with Shuffled, and 10.0% with Current only. LocalProgress@20 remains high for Full, Raw, and Shuffled (81.0%, 80.5%, and 76.0%) but falls to 38.0% with Current only. Thus chronological history, and especially aligned RGB–action correspondence, remains important after ARL.

A.6.2 Observed changes from SFT to OPD

Table A12 compares ActiveBrain-4B trajectories after SFT and OPD on the fixed non-Adapt tracks unless a row is explicitly marked, connecting aggregate changes to action-level behavior.

OPD issues 29.2% fewer `interact_pixel` actions, yet more of them execute (16.4%→26.6%), while exact repeats after failure fall from 2.1% to 0.3% and episodes with an action-type repeat after failure from 11.7% to 2.2%. These behavioral counts do not reveal intent, but are consistent with more selective interactions: fewer are attempted and a larger fraction execute. On matched tasks, OPD changes 50 SFT Explore failures to success and reverses 20 successes; Long-Horizon changes are 39 and 18. Verification is nearly unchanged (13 improved, 14 regressed), as is Cross-Room (six improved, seven regressed). Table A13 gives the instance-level audit.

Episodes that never select either candidate increase from 79.5% to 87.5%, while distractor changes among pair completions remain similar (40.6% versus 43.8%). Thus the nearly unchanged success rate masks more failures to interact with either candidate, not improved instance discrimination.

A related trade-off appears on Recover: the OPD model improves 21 tasks and regresses 32, while failed episodes with no interaction rise from 36.9% to 60.5%. Table A14 localizes the change.

OPD is 2.4 points higher after premature interaction but 5.5 lower after blocked navigation and 21.0 lower after search stagnation; the view-control subset has only seven cases. On Adapt, OPD improves 39 SFT failures and regresses 30 successes: LocalProgress@20 decreases from 85.0% to 80.0%, but Goal@20 rises from 36.0% to 40.5%, and completion conditional on local progress

Table A9: **Adapt metrics across action horizons.** All rows use the same 200 Adapt states. Metric definitions follow Table A3; all reported Adapt, LocalProgress, and Goal rates are higher-is-better.

	Adapt@H↑					LocalProgress @20↑	Goal@H↑		
Policy	@1	@5	@10	@15	@20		@5	@10	@20
Zero-shot policies									
Qwen3.5-4B	6.0	17.0	24.0	27.5	32.0	58.5	0.0	0.5	1.0
Qwen3.5-9B	3.5	5.0	10.5	10.5	10.5	26.0	0.0	0.0	0.0
Qwen3.5-27B	5.0	41.5	55.0	59.0	63.0	74.0	9.0	22.5	39.0
Qwen3.5-35B-A3B	8.0	18.5	26.0	31.5	31.5	63.5	0.0	0.5	5.0
Qwen3.5-122B-A10B	7.5	33.0	46.5	49.0	50.0	77.0	1.5	12.5	29.5
Qwen3.5-397B-A17B	6.5	35.0	46.5	50.5	55.0	65.0	1.5	7.0	13.5
Qwen3-VL-4B	8.5	30.5	42.0	47.0	50.0	66.5	0.0	0.5	0.5
Qwen3-VL-8B	9.0	35.0	44.5	49.0	51.5	65.0	0.0	0.0	1.5
Qwen3-VL-30B-A3B	6.0	16.0	25.0	32.5	40.0	58.0	0.0	0.0	0.0
Qwen3-VL-32B	7.0	30.5	38.5	47.0	48.0	62.5	0.0	0.0	0.0
Qwen3-VL-235B-A22B	4.5	26.0	40.0	48.0	55.5	61.0	0.0	0.0	0.0
RynnBrain-30B-A3B	3.0	13.5	16.0	19.0	19.5	27.0	0.0	0.0	0.0
RynnBrain1.1-9B	9.5	34.5	41.5	45.5	50.5	60.5	0.0	3.0	7.0
GLM-4.6V	9.0	30.5	40.0	47.0	49.5	66.5	0.0	0.0	0.0
Qwen2.5-VL-7B	6.5	35.5	41.5	47.0	48.5	54.0	0.0	0.0	3.5
RoboBrain2.5-4B	4.5	13.5	17.5	22.5	25.5	29.5	0.0	0.0	0.0
RoboBrain2.5-8B-NV	7.0	10.5	11.0	11.5	11.5	27.0	0.0	0.0	0.0
RynnBrain-8B	0.5	1.0	1.0	1.0	1.0	17.5	0.0	0.0	0.0
GPT-5.5	11.5	60.5	69.0	72.0	74.0	78.0	0.5	10.0	39.5
ActiveBrain-4B									
SFT	4.0	63.5	75.5	81.0	85.0	85.0	4.0	15.0	36.0
OPD	5.0	62.0	70.0	74.5	80.0	80.0	1.0	17.0	40.5
+ ARL	5.0	63.0	71.0	75.0	81.0	81.0	1.0	18.0	42.5

Table A10: **Performance across the ActiveBrain-4B training pipeline.** Per-track success rates (%) on the fixed evaluation tasks of Table 1; Adapt reports Goal@20. The final row is the unweighted mean (SR), and each Δ is the percentage-point difference from the preceding model. + ARL denotes agentic reinforcement learning; success rates are higher-is-better, while each Δ retains its signed change for every track and aggregate SR.

Track	Zero-shot↑	SFT↑	Δ	OPD↑	Δ	+ ARL↑	Δ
Explore	15.0	55.5	+40.5	70.5	+15.0	63.0	−7.5
Verification	1.5	10.0	+8.5	9.5	−0.5	15.5	+6.0
Recover	17.0	48.5	+31.5	43.0	−5.5	53.0	+10.0
Adapt (Goal@20)	1.0	36.0	+35.0	40.5	+4.5	42.5	+2.0
Cross-Room	5.9	31.8	+25.9	30.6	−1.2	35.3	+4.7
Long-Horizon	8.0	50.7	+42.7	64.7	+14.0	54.7	−10.0
SR	8.1	38.8	+30.7	43.1	+4.3	44.0	+0.9

rises from 42.4% to 50.6%. Across these audits, OPD attempts fewer unsuccessful or repeated interactions and more often converts local progress to completion, while more Verification episodes fail to select either candidate and more failed Recover episodes contain no interaction. Candidate search and post-handoff interaction therefore remain clear targets for further interactive training under the current three-stage path.

Table A11: **History intervention for ActiveBrain-4B SFT on Adapt.** Rates (%) are computed on the same 200 fixed Adapt handoffs. Goal@k measures completion of the instruction by step k; LocalProgress@20 measures local progress after the ineffective-action state. All metrics are higher-is-better and use identical task IDs and starting handoffs across input conditions. Gold/silver mark ranks 1/2 per metric row; ties share the minimum rank.

Metric	Full	Raw history	Current only	Shuffled
Goal@5↑	**4.0**	3.0	2.0	1.0
Goal@10↑	**15.0**	10.5	4.0	3.5
Goal@15↑	**28.0**	23.0	7.5	13.0
Goal@20↑	**36.0**	**33.0**	**8.5**	**20.5**
LocalProgress@20↑	**85.0**	84.5	40.0	80.0

Table A12: **Behavioral audit of ActiveBrain-4B after SFT and OPD.** Rows exclude Adapt unless marked; the Recover row includes only failed Recover episodes. "Executable interactions" means that the action ran, not that the task succeeded. ↑: higher is better; ↓: lower is better.

Statistic	SFT	OPD
Action behavior		
`interact_pixel` actions issued	2,406	1,704
Executable interactions (%)↑	16.4	26.6
Exact repeats after a failed action (%)↓	2.1	0.3
Episodes with action-type repeat after failure (%)↓	11.7	2.2
Failed Recover episodes with no interaction (%)↓	36.9	60.5
Adapt		
LocalProgress@20 (%)↑	85.0	80.0
Goal@20 (%)↑	36.0	40.5
Goal@20 among LocalProgress@20 states (%)↑	42.4	50.6

Table A13: **Instance-level audit of ActiveBrain-4B on Verification.** Conditional rates use each row's subset. The last row matches Verification task success in Tables 1 and A10; other rows are object-ID checks. ↑: higher is better; ↓: lower is better.

Statistic	SFT	OPD
Never selects either pair member (%)↓	79.5	87.5
First resolved selection is distractor (%)↓	36.6	40.0
Pair completions that change distractor (%)↓	40.6	43.8
Task-success rate (%)↑	10.0	9.5

Table A14: **Recover task success by failure type.** SFT and OPD use the same handoffs; higher-is-better rates are conditional on the handoff failure type.

Failure type	n	SFT↑	OPD↑
Blocked navigation	90	33.3	27.8
Premature interaction	84	64.3	66.7
Search stagnation	19	42.1	21.1
Rejected view control	7	71.4	14.3