BODYREADY: GROUNDING HUMAN-CENTERED PLANNING IN VISUAL PRECONDITIONS AND ACTION EFFECTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Human-centered planning and assistance require more than predicting a plausible next action: they also require knowing whether the observed person can actually begin that action from the current bodily state. Existing benchmarks largely evaluate task progression, symbolic action preconditions, or agent and object feasibility, but do not isolate whether a specific person's visible bodily state permits an action or how preceding actions change what becomes possible next. Instead of treating action plausibility as sufficient, we explicitly ground action applicability in the person's hand use, posture, and target reachability, and evaluate how action-induced changes in bodily state alter subsequent readiness. We introduce BODYREADY, a benchmark for visual action applicability of a specific human, comprising 1,190 questions grounded in 280 exocentric images. Six complementary tasks evaluate whether an action can begin, what prevents it, which preparation can enable it, and where a short action sequence first becomes invalid. Reference answers are derived from human-reviewed visual states, explicit action prerequisites, and bounded completion effects. Across 18 proprietary and open VLMs, the best model achieves 74.96% strict overall accuracy, but only 63.85% on selecting currently applicable actions. Even when it correctly identifies an action as unready, it fails to diagnose the correct set of bodily blocker categories in 28.88% of those cases. Models also choose preparations that cannot themselves begin and fail to identify the first invalid step when the same actions are reordered. These results reveal a gap between predicting plausible actions and reasoning about whether a particular person can execute them from the current bodily state, establishing visually grounded human action applicability as an important capability for planning and assistance.

1 INTRODUCTION

Human-centered planning and assistance require more than predicting a plausible next action: they require knowing whether the observed person can begin that action from the bodily state. An action may be relevant to the task yet immediately infeasible because the person's hands are occupied, the posture is incompatible, or the target is out of reach (Fig. 1). These bodily conditions are dynamic: completing one action can establish or remove the conditions required by another. Reasoning about human action therefore involves *action applicability*, whether the person's current state permits an action, and *state transition*, how an applicable action changes that state (Xie et al., 2025).

Existing benchmarks provide important evaluations of task progression (Chen et al., 2026; Qiu et al., 2026a), symbolic action preconditions and effects, and scene or agent functionality (Maillard et al., 2026). However, these settings do not isolate a significant question: whether the *specific observed person's* bodily state permits a proposed action, and how that applicability changes after preceding actions change the conditions required to begin that action. Instead of treating action plausibility as sufficient, we explicitly ground action applicability in the person's hand use, posture, and target reachability, and evaluate how actions change whether the required bodily conditions hold.

We introduce BODYREADY, a benchmark for *human-specific visual action applicability*. We define *human readiness* as whether the bodily conditions required to directly begin a specified action currently hold. BODYREADY evaluates this capability at three levels. First, *Action Readiness*, *Action*

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

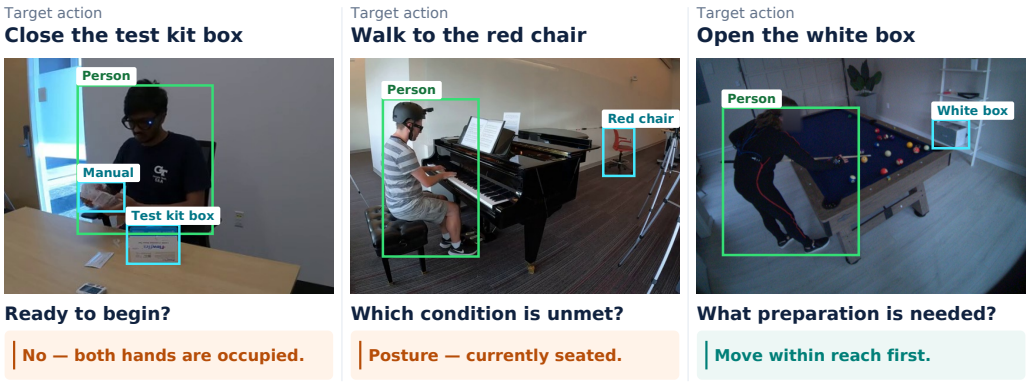

Figure 1: Human readiness constrains the next action. These examples illustrate why human-centered action decisions must account for the person's current bodily conditions. Occupied hands prevent immediate box closing (left); a seated person needs to stand before walking (middle); and an out-of-reach box requires repositioning before opening (right).

Selection, and *Blocker Diagnosis* assess which actions can begin now and which bodily conditions prevent them. Second, *Readiness Reassessment* and *Preparation Selection* test how changes in bodily state alter action applicability and whether an available preparation can make a target action ready. Third, *Sequence Validation* checks whether actions remain applicable as preceding actions change the state. BODYREADY does not attempt unrestricted goal-directed planning; it isolates a prerequisite capability that planning and assistance systems must respect when reasoning about human actions.

The benchmark contains 1,190 questions grounded in 280 exocentric images from Ego-Exo4D and Nymeria (Grauman et al., 2024; Ma et al., 2024). The images span activities including CPR training, musical instrument practice, billiards, and bicycle repair. Reference answers are derived from human-reviewed visual states together with explicit action prerequisites and bounded completion effects. These effects define a controlled reference model for reasoning about short action sequences and their consequences for subsequent readiness, rather than predicting unconstrained physical dynamics. Models receive only the visual observations and natural-language task descriptions; the underlying bodily-state annotations, action prerequisites, and reference derivations are withheld.

We evaluate 18 proprietary and open VLMs, including recently released frontier models Claude Opus 5.5 (Anthropic, 2026) and GPT-6 Astra (OpenAI, 2026b), together with open models from Qwen (Bai et al., 2025), InternVL (Wang et al., 2025b), MiniCPM (Yu et al., 2026), Molmo (Clark et al., 2026), and other families. The highest-scoring model achieves 74.96% strict overall accuracy, but only 63.85% on *Action Selection* and 65.70% on *Blocker Diagnosis*. More importantly, correct binary readiness judgments do not imply complete understanding of the underlying constraints. Even when the highest-scoring model correctly determines that an action cannot begin, it fails to identify the correct set of bodily blocker categories in 28.88% of those cases.

Further errors arise in preparation selection and sequence validation. Models select preparatory actions that cannot themselves begin, including cases in which none of the offered preparations is valid. They also fail to identify the first inapplicable step when the same actions are presented in different orders. Our contributions to evaluating these reasoning capabilities are threefold:

- We formulate *human-specific visual action applicability* through human readiness, relating a person's observed bodily state to action prerequisites and changes in subsequent applicability.
- We introduce BODYREADY, a benchmark of 1,190 questions across six complementary tasks, grounded in human-reviewed visual states and explicit action prerequisites and bounded effects that provide traceable reference answers.
- We evaluate 18 proprietary and open VLMs and identify systematic gaps between binary readiness judgments, complete blocker diagnosis, valid preparation selection, and prerequisite tracking across action sequences.

2 Related Work

Human understanding and interaction feasibility. HumanBench, Human-MME, and HumanVideo-MME evaluate human-centric perception and reasoning (Tang et al., 2023; Liu et al., 2025; Cai et al., 2025), while EgoThink and STAR examine perspective-taking and situated video reasoning (Cheng et al., 2024; Wu et al., 2024). EgoTaskQA evaluates action dependencies, object-state changes, and counterfactual reasoning in egocentric task videos (Jia et al., 2022). Interaction feasibility is studied through interaction localization (Nagarajan et al., 2019; Mo et al., 2021), manipulation prerequisites (Gundawar et al., 2025), and human reachability for assistance (Liu et al., 2026). SceneTeract specifically evaluates scene functionality under agent capability constraints (Maillard et al., 2026). BodyReady makes the observed person the subject of action reasoning: their bodily configuration determines which actions can begin, and their actions change those conditions. Exocentric views expose posture, hand use, and person–object relations.

Visual planning and human assistance. Procedure planning connects visual observations through action sequences (Chang et al., 2020); VLaMP, VideoPlan, and ProAssist study visual planning and proactive human assistance (Patel et al., 2023; Zhang et al., 2026a; 2025b), including evaluation during ongoing human activities (Verghese et al., 2025). EgoPlan-Bench and EgoPlan-Bench2 evaluate next-action planning (Chen et al., 2026; Qiu et al., 2026a), while EgoTV verifies task execution (Hazra et al., 2023). For embodied agents, SayCan connects action selection to skill feasibility, and EmbodiedBench evaluates vision-driven navigation and manipulation (Ahn et al., 2022; Yang et al., 2025). BodyReady evaluates action decisions against the person's bodily conditions. Preparation questions require applicable actions sufficient to establish the target's readiness; sequence questions test whether actions preserve or establish subsequent prerequisites.

Action prerequisites and state changes. PlanBench and ACPBench evaluate planning and reasoning about change (Valmeekam et al., 2023; Kokel et al., 2025), while proScript studies partially ordered event dependencies (Sakaguchi et al., 2021). Preconditions and effects support state prediction (Xie et al., 2025), embodied planning (Wang et al., 2025a), multi-agent cooperation (Zhang et al., 2025a), and closed-loop evaluation (Zhang et al., 2026b). Embodied Agent Interface evaluates planning modules in VirtualHome and BEHAVIOR, including action sequencing and transition modeling (Li et al., 2024). Its BEHAVIOR implementation automates navigation, whereas BodyReady requires explicit preparation when a target is out of reach. BodyReady evaluates readiness from real human observations, where the domain gap from simulated environments presents grounding challenges.

3 The BodyReady Benchmark

BodyReady evaluates whether a person's observed bodily state permits directly beginning an action and how completing that action changes subsequent readiness. We first define the state representation, action boundaries, prerequisites, and bounded completion effects. We then describe how visual annotations instantiate this representation and support reference answers for the six tasks.

3.1 Human State and Action Boundaries

Given an image I and a designated person p, we represent the person's bodily state as

$$S_p = \langle H_p, P_p, Q_p \rangle. \tag{1}$$

Here, H_p records each hand's availability and its contact or control relations with identified objects. P_p records posture, including standing, bending, sitting, squatting, and kneeling. $Q_p(o)$ records whether target o is reachable from the person's current posture, stance, location, and support configuration. Any component can remain unknown when the visual evidence is insufficient.

Alongside S_p, we record object states and action–role compatibility as F, giving the reference state $W = \langle S_p, F \rangle$. These facts determine object-level eligibility: closing requires an open object, while placement requires a suitable destination. Bodily readiness is evaluated only for eligible actions, so an incompatible object or object state is not labeled as a bodily failure.

Action Boundaries. A precise notion of readiness requires a precise definition of where an action begins. Otherwise, the same situation could receive different labels depending on whether preparatory

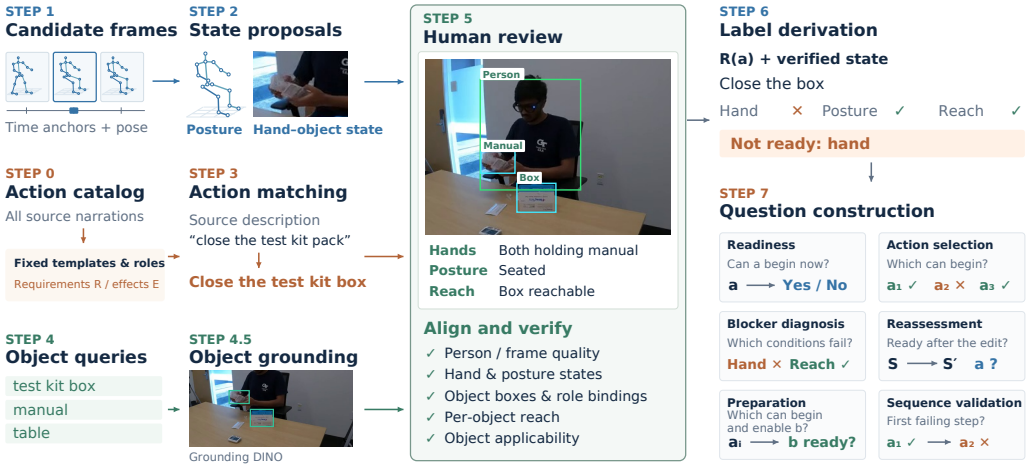

Figure 2: Construction pipeline of BODYREADY. Candidate frames and state proposals (Steps 1–2), catalog-based action matching (Steps 0 and 3), and object queries and grounding (Steps 4–4.5) converge in human review (Step 5). Reviewed states and grounded actions support readiness and blocker derivation (Step 6), followed by six-task question construction using fixed prerequisites and completion effects (Step 7). In the illustrated example, both hands hold a manual, preventing immediate box closing despite the box being reachable.

movements are silently included in the action itself. We therefore distinguish the prerequisites that must already hold from the movements that belong to the operation. This distinction follows temporally extended action models, which separate initiation conditions, internal execution, and termination (Sutton et al., 1999). For example, reaching toward and making initial contact with a reachable drawer can belong to opening the drawer, whereas first putting down an unrelated object to free a hand is treated as a separate preparation.

Following the representation of activities as action–object instructions (Puig et al., 2018), we formalize a grounded action $a = \langle \tau, \rho \rangle$, where template τ defines the operation and its starting boundary, and ρ binds its roles to specific people and objects. Under this boundary, $R(a)$ specifies the bodily conditions that must already hold for a to begin: hand availability or object control in H_p, posture in P_p, and target-specific reachability in Q_p. The prerequisite semantics remain fixed within each template, while the observed state S_p determines whether they are satisfied. This separation between fixed action requirements and instance-specific bodily state makes readiness comparable across visual situations. Appendix C lists the boundaries and bodily prerequisites of all 42 action templates.

3.2 READINESS AND ACTION EFFECTS

Bodily Prerequisites. For each object-eligible grounded action, $R(a)$ specifies the bodily conditions required at its starting boundary. Let r_H, r_P, and r_Q denote satisfaction of the hand, posture, and reach requirements, respectively. Readiness is

$$r(a, W) = r_H(a, W) \wedge r_P(a, W) \wedge r_Q(a, W), \qquad r(a, W) \in \{1, 0, ?\}. \tag{2}$$

A channel with no requirement evaluates to 1. Readiness is 0 whenever any required condition is known to be false, 1 when all required conditions are known to hold, and unknown when no required condition is false but at least one cannot be determined from the available evidence.

The hand state distinguishes availability from object control. Contact with an action target can remain compatible with hand availability, whereas holding an unrelated object occupies that hand. Conversely, controlling a required tool satisfies the corresponding control prerequisite without making the hand available for another role. A one-hand requirement without a specified side accepts either qualifying hand, whereas a two-hand requirement requires two distinct qualifying hands. Posture is treated independently of reachability: failure of Q_p does not imply a posture failure unless the action separately requires a particular posture. Object references are resolved through ρ, while the facts in F continue to determine object-level eligibility separately from bodily readiness.

Completion Effects. To evaluate readiness after an action, we use bounded completion effects that update the reference facts relevant to subsequent judgments, following the state progression formulation used in ACPBench (Kokel et al., 2025). For an object-eligible action a with $r(a, W) = 1$, a fixed effect rule defines its reference state after successful completion:

$$W' = E(a, W) = \langle S'_p, F' \rangle, \qquad S'_p = \langle H'_p, P'_p, Q'_p \rangle. \tag{3}$$

The update refers to the reference state at successful completion specified by the effect rule. Facts unaffected by the action persist, while affected facts whose successor values cannot be determined become unknown. These effects provide a controlled reference transition for benchmark construction rather than a claim that all physical executions of the same action follow an identical trajectory.

For example, consider a person holding unrelated objects in both hands beside a reachable closed drawer. Opening the drawer is initially blocked by hand occupation. If placing one held object on a nearby surface is applicable, completing the placement and releasing the object makes that hand available in H'_p. With posture, the other hand's state, and reachability of the drawer unchanged, opening the drawer becomes ready in W'. The opening prerequisites remain fixed; the preceding placement establishes the hand availability needed to satisfy them.

Applying this check and update process successively allows the benchmark to evaluate whether a preparation enables a later action and whether a sequence remains applicable as the state changes. An action that cannot begin receives no successful effect update. Successor judgments are retained only when the required effects are sufficiently determined. Appendix D details completion rules.

3.3 Data Collection and Annotation

We instantiate the state and action representation using frames from Ego-Exo4D and Nymeria (Grauman et al., 2024; Ma et al., 2024). Figure 2 shows the pipeline from automatic proposals through human review to deterministic reference derivation and question construction. Source sequence and timestamp information are retained throughout.

Candidate frames and state proposals (Steps 1–2). Dataset-specific adapters use source action timing and available person, pose, and hand annotations to select candidate frames and localize the designated person. Source annotations and visual model proposals initialize posture and hand–object relations. These outputs serve only as drafts for review rather than as benchmark labels.

Action catalog and matching (Steps 0 and 3). Source narrations and atomic action descriptions provide operations for constructing the action catalog. Each catalog entry specifies an operation template, its boundary and roles, its bodily prerequisites, and its bounded completion effects according to Sections 3.1 and 3.2. For individual frames, Qwen-assisted proposals map narration-derived candidate actions to these templates using temporal and task context. Scene and object categories provide additional candidate operations, including posture changes. These proposals determine which grounded actions enter review; they do not determine whether those actions are ready.

Object queries and grounding (Steps 4–4.5). Action roles and scene descriptions provide canonical object category queries. Grounding DINO (Liu et al., 2024) proposes candidate bounding boxes, which are consolidated into object instances and verified during human review. The resulting bindings associate each grounded action with the target and required object, tool, or destination.

Human review (Step 5). Two reviewers divided the annotation pool between them for review, jointly discussing uncertain cases and marking unresolved values as unknown. They verified person and frame quality, hand availability and object control, posture, object boxes, action role bindings, and object eligibility, and annotated target-specific reachability. Of the 287 person–image records supporting the benchmark, 273 have saved automatic drafts for comparison. Reviewers revised left and right hand status in 97 (35.5%) and 100 (36.6%) records, respectively, rejected 1,680 of 5,859 traceable object box proposals (28.7%), and supplied 3,181 reachability judgments absent from the drafts. The reviewed bodily states and object relations support subsequent reference derivation.

Label derivation and question construction (Steps 6–7). Qwen proposes candidates for review; human-reviewed annotations supply W and action-role bindings. For object-eligible actions, fixed $R(a)$ rules derive readiness and unmet bodily requirements. Subsequent-action questions apply E after applicable actions. Section 3.4 defines the six tasks and conditions for determinate reference answers.

Table 1: The six BodyReady tasks. Candidates in AS and PS are considered independently from the same initial state. RR supplies a hypothetical state change; SV supplies an ordered action sequence.

Code	Task	Question answered	n
AR	Action Readiness	Can the specified action directly begin now?	170
AS	Action Selection	Which candidate actions can directly begin now?	130
BD	Blocker Diagnosis	Can the action begin, and which bodily conditions are unmet?	344
RR	Readiness Reassessment	Can the action begin after the specified state change?	188
PS	Preparation Selection	Which candidates can begin and make the target ready?	80
SV	Sequence Validation	Which is the first step that cannot begin, if any?	278
	Total		1,190

3.4 Task Construction and Dataset Composition

Using the reviewed states and fixed action rules, we construct six tasks whose reference answers share a state representation, action prerequisites, and, where needed, bounded completion effects. Action Readiness, Action Selection, and Blocker Diagnosis evaluate these prerequisites in the observed state. Readiness Reassessment evaluates them after an explicit hypothetical state modification. Preparation Selection and Sequence Validation evaluate them in states produced by completing preceding actions. This shared formulation connects current readiness, preparation choices, and sequence validity through the same action model. Table 1 summarizes each task.

Current Applicability and Controlled State Changes. AR evaluates $r(a, W)$ for one grounded action. AS evaluates multiple candidate actions independently from the same reference state W and returns the complete set that can begin. BD additionally requires the complete set of bodily channels that prevent the specified action from beginning,

$$\mathcal{B}(a, W) = \{c \in \{H, P, Q\} : r_c(a, W) = 0\}. \tag{4}$$

Ready actions therefore have an empty blocker set. RR evaluates whether applicability changes after an explicitly supplied state modification U_δ by computing $r(a, U_\delta(W))$. The modification is treated as a premise. The task evaluates its consequence for readiness and does not require the model to infer an action that physically realizes the modification.

Preparation and Sequence Reasoning. PS begins with an unready target action b and asks which candidate preparations both can begin in the current state and, after successful completion, make b ready. The correct candidate set is

$$\mathcal{J}^* = \{j : r(a_j, W) = 1 \ \wedge \ r(b, E(a_j, W)) = 1\}. \tag{5}$$

This definition separates two requirements that can be conflated: a preparation must be currently applicable, and its completion must establish the missing conditions for the target action. Negative candidates therefore include actions that cannot begin as well as actions that can begin but do not make the target ready. An empty answer means that none of the candidates satisfies both requirements.

SV evaluates a supplied action sequence in order. Each applicable action updates the reference state before the next action is checked. Evaluation stops at the first action whose entry conditions are not satisfied. The answer is its one-based position, or zero if all actions can begin in sequence. Because no subsequent decision depends on the effect of the final action, the final step requires only an entry judgment. SV therefore tests whether action order remains consistent with the changing bodily state, rather than asking the model to generate an unrestricted plan.

Question Validity and Final Composition. We retain a question only when the visual facts and, where required, successor facts needed to determine its reference answer are resolved. BD requires a determinate complete blocker diagnosis; RR requires a determinate judgment after the supplied state change; and PS and SV require sufficiently specified completion effects for every successor judgment involved. Object-level eligibility must also be established at each evaluated step. If a preceding action leaves the object conditions required by a subsequent step unknown, the sequence is excluded. Unrelated facts may remain unknown because they cannot affect the requested answer.

The benchmark contains 1,190 questions over 280 images, 287 person–image records, and 89 source groups. Source groups correspond to Ego-Exo4D takes or Nymeria sequences, with give/receive samples grouped separately by reviewed handoff event. Ego-Exo4D contributes 692 questions from 144 images, and Nymeria 498 questions from 136 images. The questions span 42 action templates covering acquisition, placement, manipulation, interpersonal transfer, and bodily repositioning. Assistance-relevant activities include CPR training, musical instrument practice, billiards, and bicycle repair. Appendix B reports answer distributions; Appendix I provides template coverage and source provenance. After construction, question identifiers, visual inputs, reference answers, and the scoring protocol were frozen for all reported comparisons.

4 Evaluation Protocol and Results

Evaluation Protocol. We evaluate 18 models zero-shot on identical per-question images and natural-language prompts, with person and object markers resolving visual references and private state annotations and action rules withheld. We report strict whole-question accuracy: responses must satisfy the JSON schema and match the complete reference answer. Overall accuracy covers all 1,190 questions. Appendices E and F detail input configurations, output schemas, and inference settings.

Baselines. We evaluate 18 VLMs, comprising six proprietary and twelve open models. The proprietary models include recently released Claude Opus 5.5 (Anthropic, 2026) and GPT-6 Astra (OpenAI, 2026b), together with GPT-6 Sol (OpenAI, 2026c), GPT-5.6 Sol and GPT-5.6 Terra (OpenAI, 2026a), and Gemini 3.8 Flash (Doshi & Popa, 2026). The open models include Qwen3.8-Flash-Next and Qwen3.8-27B (Qiu et al., 2026b); Qwen3-VL-Instruct at 2B, 4B, 8B, and 32B, plus its 8B-Thinking variant (Bai et al., 2025); InternVL3.5 at 1B and 8B (Wang et al., 2025b); MiniCPM-V-4.5 (Yu et al., 2026); Molmo2-8B (Clark et al., 2026); and Kimi-VL-A3B-Instruct (Team et al., 2025). This selection covers recent frontier models and multiple open-model families, while supporting comparisons across parameter scales within Qwen3-VL and between its 8B Instruct and Thinking configurations.

Main Comparison. Table 2 shows substantial errors even among the strongest models. Claude Opus 5.5 achieves the highest Overall accuracy of 74.96%, followed by GPT-6 Astra at 72.69%; Qwen3.8-Flash-Next leads the open models at 54.20%. Even Opus reaches only 63.85% on Action Selection and 65.70% on Blocker Diagnosis, motivating a closer examination of selection and diagnosis errors.

The evaluated proprietary models consistently outperform the open models in Overall accuracy: even the lowest-scoring proprietary model reaches 62.77%, compared with 54.20% for the strongest open model under the evaluated configurations and output protocol. Within Qwen3-VL Instruct, larger models do not improve uniformly across tasks. Moving from 4B to 32B raises Sequence Validation accuracy from 30.22% to 43.88%, but lowers Blocker Diagnosis from 31.10% to 26.16%. Similarly, Flash-Next leads the open models overall, while Qwen3.8-27B performs better on Action Selection, Readiness Reassessment, and Sequence Validation. Overall rankings therefore conceal different strengths in judging current conditions and reasoning through subsequent actions.

We use 5,000 paired bootstrap resamples over the 89 source groups to account for within-group dependence. The Opus–Astra Overall difference is 2.27 percentage points, with a descriptive 95% interval of $[-0.10, 4.56]$; this interval does not establish a clear separation between the two models. Appendix G reports comparisons and the scope of this uncertainty estimate.

5 Diagnosing the Decisions Needed for Planning

Aggregate task scores do not reveal which decisions fail. We examine whether models identify all bodily blockers, select preparations that are both applicable and effective, and track applicability through action sequences. Figure 3 summarizes the diagnostic profiles across all 18 models.

5.1 Correct Rejection Does Not Imply Complete Diagnosis

Even when models correctly judge an action as unready, they often fail to identify all unmet bodily conditions. Among schema-valid responses that correctly reject a gold-not-ready action, Opus produces a complete blocker diagnosis in 165/232 cases (71.12%), Astra in 157/229 (68.56%), and Flash-Next in 117/239 (48.95%; Figure 3). Thus, 28.88% of Opus's correct rejections within this

Table 2: Strict whole-question accuracy (%) on BODYREADY. Models are ordered by increasing Overall accuracy within each group. Bold marks the highest accuracy within each group. Task sample counts are shown in parentheses. Arch. denotes the language backbone architecture. Size follows the reported model scale; MoE sizes indicate total/active parameters. – indicates unavailable information. †Flash-Next includes 51B n-gram embedding parameters and a 4B multi-token prediction module.

Model	Arch.	Size (B)	Task accuracy (%)						
			AR (170)	AS (130)	BD (344)	RR (188)	PS (80)	SV (278)	Overall (1,190)
Open models									
InternVL3.5-1B-HF	Dense	1	54.71	0.00	0.00	45.74	0.00	6.47	16.55
Kimi-VL-A3B-Instruct	MoE	16/3	34.71	1.54	11.05	7.98	18.75	30.58	17.98
InternVL3.5-8B-Instruct	Dense	8	59.41	8.46	15.99	54.26	15.00	23.74	29.16
Qwen3-VL-2B-Instruct	Dense	2	61.18	13.08	27.03	59.57	17.50	28.78	35.29
MiniCPM-V-4.5	Dense	8.7	65.88	10.77	27.62	52.13	22.50	32.73	35.97
Molmo2-8B	Dense	8	55.88	17.69	30.23	53.72	41.25	25.90	35.97
Qwen3-VL-8B-Instruct	Dense	8	68.82	17.69	32.27	52.66	36.25	32.01	39.33
Qwen3-VL-4B-Instruct	Dense	4	71.18	29.23	31.10	52.13	40.00	30.22	40.34
Qwen3-VL-32B-Instruct	Dense	32	70.00	27.69	26.16	54.79	40.00	43.88	42.18
Qwen3-VL-8B-Thinking	Dense	8	71.76	29.23	29.36	54.26	48.75	44.60	44.20
Qwen3.8-27B	Dense	27	71.76	**30.77**	43.31	**68.09**	53.75	**52.88**	52.86
Qwen3.8-Flash-Next	MoE	125/6†	**81.18**	26.92	**47.38**	65.96	**62.50**	48.56	**54.20**
Proprietary models									
GPT-5.6 Terra	–	–	82.35	46.92	50.58	68.62	63.75	69.06	62.77
Gemini 3.8 Flash	–	–	80.00	45.38	63.95	75.53	80.00	70.50	68.66
GPT-5.6 Sol	–	–	81.76	59.23	55.23	76.06	75.00	78.78	69.58
GPT-6 Sol	–	–	**85.88**	62.31	60.47	72.87	67.50	78.06	70.84
GPT-6 Astra	–	–	83.53	56.92	63.66	75.00	80.00	**80.94**	72.69
Claude Opus 5.5	–	–	82.35	**63.85**	**65.70**	**81.91**	**83.75**	79.86	**74.96**

subset still have an incorrect diagnosis. Opus omits a reach blocker in 17/145 reach-positive cases (11.72%) and adds a hand blocker in 28/101 hand-negative cases (27.72%). Correctly rejecting an action therefore does not ensure identifying the conditions that preparation must address.

The channel-level errors reveal different tendencies to omit and add bodily constraints. Across all schema-valid BD responses, hand has the highest false-positive rate of the three channels for 16 of the 17 models with valid outputs. The channel with the highest false-negative rate varies: pose for 11 models and reach for six. These comparisons use each channel's positive and negative cases as the respective denominators, rather than comparing raw error counts; pose has only 41 positive cases in the full BD set, compared with 152 for hand and 155 for reach. The results distinguish a widespread tendency to report hand constraints when none apply from model-dependent failures to recognize posture or reach constraints.

5.2 PREPARATION REQUIRES BOTH APPLICABILITY AND SUFFICIENCY

A valid preparation must itself be executable and make the target action ready upon completion. Models violate both requirements. Among schema-valid responses, Flash-Next selects 25/145 preparations that cannot begin and 8/33 that can begin but leave the target unready. Opus makes the same two types of error in 12/145 and 1/33 cases, respectively. These selections violate either the preparation's own entry conditions or the requirement that its effects make the target ready.

Finding valid preparations and rejecting invalid ones also show different patterns (Figure 3). Flash-Next answers 45/47 questions with a valid preparation correctly, but only 5/33 with none. Qwen3-

	AR		AS		BD	RR		PS		SV
	Recall: ready	Recall: not ready	Option precision*	Option recall*	Complete diagnosis*	Enabled	Still blocked	Valid option	No valid option	Both correct
Opus 5.5	82	82	80	84	71	79	86	100	61	68
GPT-6 Astra	81	86	75	85	69	66	84	100	52	65
GPT-6 Sol	81	91	78	87	70	66	80	94	30	68
GPT-5.6 Sol	78	86	74	86	63	80	72	96	45	58
Gemini 3.8 Flash	68	92	83	76	72	67	84	91	64	52
GPT-5.6 Terra	89	75	66	84	63	74	62	91	24	19
Qwen3.8-Flash	76	86	51	73	49	51	82	96	15	16
Qwen3.8-27B	59	85	53	89	46	49	89	87	6	32
Qwen3-VL-2B	29	93	35	61	25	36	86	30	0	0
Qwen3-VL-4B	55	87	44	56	34	23	83	2	94	3
Qwen3-VL-8B	85	53	40	71	34	79	24	32	42	0
Qwen3-VL-32B	51	89	45	73	27	45	66	40	39	10
Qwen3-VL-8B-T	80	64	53	75	30	74	32	79	6	23
MiniCPM-V-4.5	73	59	36	52	26	91	10	38	0	6
Molmo2-8B	98	14	45	43	39	96	8	68	3	13
InternVL3.5-8B	94	25	33	62	28	96	9	26	0	0
Kimi-VL-A3B	44	26	24	34	32	14	1	32	0	3
InternVL3.5-1B	100	9	--	--	--	85	3	0	0	0

Figure 3: Decision profiles (%, rounded; higher is better). AR: class recall; AS: option precision/recall; BD: complete diagnosis given correct rejection; RR: enabling/still-blocked edits; PS: valid/no-valid preparation; SV: both orders correct. Stars denote conditional metrics; dashes denote undefined rates. Names omit Claude and Instruct/HF; Flash abbreviates Flash-Next and T denotes Thinking. Appendix H defines metrics and denominators.

VL-4B exhibits the opposite pattern: 1/47 versus 31/33. Aggregate PS accuracy can therefore conceal a strong imbalance between identifying an effective preparation and recognizing that none of the offered candidates works. Both are necessary for selecting a valid next step.

5.3 Sequence Validity Depends on Action Order

An action changes the state against which the next action must be checked. To examine whether models track this dependency, we use 31 pairs sharing the same frame and action multiset but differing in order and reference answer. Holding the visual context and action inventory fixed makes these pairs a focused test of sensitivity to action order. A preceding action can establish a condition needed later or remove one that initially held, so the same actions can lead to different first-failure positions. We count a pair as correct only when the model identifies the reference outcome for both orders.

Opus answers both sequences correctly in 21/31 pairs, Astra in 20/31, GPT-6 Sol in 21/31, and Flash-Next in 5/31 (Figure 3). Even the strongest models therefore fail on at least one sequence in roughly one-third of these pairs. Success on one ordering does not ensure correct validation of the other, despite the unchanged visual evidence and action inventory. These results expose difficulty in determining how earlier actions establish or remove the bodily conditions required by later steps, a dependency that must be respected when arranging human actions into a sequence.

6 Conclusion

BodyReady makes human-specific visual action applicability measurable by linking a person's bodily state to action prerequisites and completion effects. Across 18 VLMs, the strongest model achieves 74.96% strict overall accuracy, yet models still produce incomplete blocker diagnoses, select preparations that cannot begin or fail to enable the target, and misjudge sequences when the same actions are reordered. Correct readiness judgments can coexist with incomplete diagnosis, while preparation and sequence tasks expose failures to respect action prerequisites and effects. BodyReady provides a controlled benchmark for studying these capabilities and a basis for investigating whether improvements translate into more reliable human-centered planning and assistance.

AI USE STATEMENT

AI tools assisted with manuscript drafting and editing, methodological feedback, interpretation of analysis results, code preparation, and figure generation. Automated proposals also supported the annotation workflow described in Section 3.3; visual-state review and rule-based reference construction are distinguished there from evaluated-model predictions. The authors retain responsibility for the paper's claims, data handling, and final verification.

REFERENCES

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.

Anthropic. Introducing Claude Opus 5.5. https://www.anthropic.com/claude-opus-5-5, 2026. Accessed: 2026-09-25.

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*, 2025.

Yuxuan Cai, Jiangning Zhang, Zhenye Gan, Qingdong He, Xiaobin Hu, Junwei Zhu, Yabiao Wang, Chengjie Wang, Zhucun Xue, Chaoyou Fu, et al. Humanvideo-mme: Benchmarking mllms for human-centric video understanding. *arXiv preprint arXiv:2507.04909*, 2025.

Chien-Yi Chang, De-An Huang, Danfei Xu, Ehsan Adeli, Li Fei-Fei, and Juan Carlos Niebles. Procedure planning in instructional videos. In *ECCV (11)*, volume 12356 of *Lecture Notes in Computer Science*, pp. 334–350. Springer, 2020.

Yi Chen, Yuying Ge, Yixiao Ge, Mingyu Ding, Bohao Li, Rui Wang, Ruifeng Xu, Ying Shan, and Xihui Liu. Egoplan-bench: Benchmarking multimodal large language models for human-level planning. *International Journal of Computer Vision*, 134(3):118, 2026.

Sijie Cheng, Zhicheng Guo, Jingwen Wu, Kechen Fang, Peng Li, Huaping Liu, and Yang Liu. Egothink: Evaluating first-person perspective thinking capability of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14291–14302, 2024.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Rohun Tripathi, Sangho Lee, Mohammadreza Salehi, Jason Ren, Chris Dongjoo Kim, Yinuo Yang, et al. Molmo2: Open weights and data for vision-language models with video understanding and grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28652–28668, 2026.

Tulsee Doshi and Raluca Ada Popa. Introducing Gemini 3.8 Flash and 3.8 Flash Cyber. https://blog.google/innovation-and-ai/models-and-research/gemini-models/3-8-flash-and-3-8-flash-cyber/, 2026. Accessed: 2026-09-25.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19383–19400, 2024.

Atharva Gundawar, Som Sagar, and Ransalu Senanayake. Pac bench: Do foundation models understand prerequisites for executing manipulation policies? In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025. doi: 10.52202/085713-3671. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/9ecafb09de180aaad7b7205be7eb24a4-Paper-Datasets_and_Benchmarks_Track.pdf.

Rishi Hazra, Brian Chen, Akshara Rai, Nitin Kamra, and Ruta Desai. Egotv: Egocentric task verification from natural language task descriptions. In *ICCV*, pp. 15371–15383. IEEE, 2023.

Baoxiong Jia, Ting Lei, Song-Chun Zhu, and Siyuan Huang. Egotaskqa: Understanding human tasks in egocentric videos. *Advances in Neural Information Processing Systems*, 35:3343–3360, 2022.

Harsha Kokel, Michael Katz, Kavitha Srinivas, and Shirin Sohrabi. Acpbench: Reasoning about action, change, and planning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 26559–26568, 2025.

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 *NeurIPS*, 2024.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024.

Yuansen Liu, Haiming Tang, Jinlong Peng, Jiangning Zhang, Xiaozhong Ji, Qingdong He, Wenbin Wu, Donghao Luo, Zhenye Gan, Junwei Zhu, et al. Human-mme: A holistic evaluation benchmark for human-centric multimodal large language models. *arXiv preprint arXiv:2509.26165*, 2025.

Ziang Liu, Yunting Yan, Christy Sum Yu Cheung, Tailai Ying, Bodong Liu, Shiqin Tong, Alexander Orkwis, Katherine Dimitropoulou, and Tapomayukh Bhattacharjee. Knowing When Not to Help: Active Estimation of Human Reachability for Just-Right Robot Assistance. In *Proceedings of Robotics: Science and Systems*, Sydney, Australia, July 2026. doi: 10.15607/RSS.2026.XXII.119.

Lingni Ma, Yuting Ye, Fangzhou Hong, Vladimir Guzov, Yifeng Jiang, Rowan Postyeni, Luis Pesqueira, Alexander Gamino, Vijay Baiyya, Hyo Jin Kim, et al. Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In *European Conference on Computer Vision*, pp. 445–465. Springer, 2024.

Léopold Maillard, Francis Engelmann, Tom Durand, Boxiao Pan, Yang You, Or Litany, Leonidas Guibas, and Maks Ovsjanikov. Sceneteract: Agentic functional affordances and vlm grounding in 3d scenes. *arXiv preprint arXiv:2603.29798*, 2026.

Kaichun Mo, Leonidas J Guibas, Mustafa Mukadam, Abhinav Gupta, and Shubham Tulsiani. Where2act: From pixels to actions for articulated 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6813–6823, 2021.

Tushar Nagarajan, Christoph Feichtenhofer, and Kristen Grauman. Grounded human-object interaction hotspots from video. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8687–8696. IEEE, 2019.

OpenAI. GPT-5.6 System Card. https://deploymentsafety.openai.com/gpt-5-6, 2026a. Accessed: 2026-09-25.

OpenAI. GPT-6 Astra System Card. https://deploymentsafety.openai.com/gpt-6-astra, 2026b. Accessed: 2026-09-25.

OpenAI. Introducing GPT-6 Sol and Luna. https://openai.com/index/introducing-gpt-6-sol-and-luna/, 2026c. Accessed: 2026-09-25.

Dhruvesh Patel, Hamid Eghbalzadeh, Nitin Kamra, Michael Louis Iuzzolino, Unnat Jain, and Ruta Desai. Pretrained language models as visual planners for human assistance. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15256–15268. IEEE, 2023.

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.

Lu Qiu, Yi Chen, Yuying Ge, Yixiao Ge, Ying Shan, and Xihui Liu. Egoplan-bench2: A benchmark for multimodal large language model planning in real-world scenarios. *International Journal of Computer Vision*, 134(5):222, 2026a.

Zihan Qiu, Zekun Wang, Xiao Li, Yanpeng Li, Yang Xu, Yixuan Wang, Huaqing Zhang, Rui Men, Bochao Mao, Chengruidong Zhang, et al. On the design of qwen3. 8-next architecture: Evaluation, efficiency, and training stability. *arXiv preprint arXiv:2608.30320*, 2026b.

Keisuke Sakaguchi, Chandra Bhagavatula, Ronan Le Bras, Niket Tandon, Peter Clark, and Yejin Choi. proscript: Partially ordered scripts generation. In *EMNLP (Findings)*, volume EMNLP 2021 of *Findings of ACL*, pp. 2138–2149. Association for Computational Linguistics, 2021.

Richard S. Sutton, Doina Precup, and Satinder Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artif. Intell.*, 112(1-2):181–211, 1999. doi: 10.1016/S0004-3702(99)00052-1. URL https://doi.org/10.1016/S0004-3702(99)00052-1.

Shixiang Tang, Cheng Chen, Qingsong Xie, Meilin Chen, Yizhou Wang, Yuanzheng Ci, Lei Bai, Feng Zhu, Haiyang Yang, Li Yi, et al. Humanbench: Towards general human-centric perception with projector assisted pretraining. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21970–21982. IEEE, 2023.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025.

Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. Planbench: An extensible benchmark for evaluating large language models on planning and reasoning about change. *Advances in Neural Information Processing Systems*, 36:38975–38987, 2023.

Mrinal Verghese, Brian Chen, Hamid Eghbalzadeh, Tushar Nagarajan, and Ruta P Desai. User-in-the-loop evaluation of multimodal llms for activity assistance. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 1144–1154, 2025.

Siyin Wang, Zhaoye Fei, Qinyuan Cheng, Shiduo Zhang, Panpan Cai, Jinlan Fu, and Xipeng Qiu. World modeling makes a better planner: Dual preference optimization for embodied task planning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21518–21537, 2025a.

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*, 2025b.

Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B. Tenenbaum, and Chuang Gan. STAR: A benchmark for situated reasoning in real-world videos. *CoRR*, abs/2405.09711, 2024.

Kaige Xie, Ian Yang, John Gunerli, and Mark Riedl. Making large language models into world models with precondition and effect knowledge. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 7532–7545, 2025.

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 *ICML*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Ranchi Zhao, et al. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11704–11715, 2026.

Ce Zhang, Yale Song, Ruta Desai, Michael Louis Iuzzolino, Joseph Tighe, Gedas Bertasius, and Satwik Kottur. Enhancing visual planning with auxiliary tasks and multi-token prediction. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4190–4200. IEEE, 2026a.

Hongxin Zhang, Zeyuan Wang, Qiushi Lyu, Zheyuan Zhang, Sunli Chen, Tianmin Shu, Behzad Dariush, Kwonjoon Lee, Yilun Du, and Chuang Gan. Combo: Compositional world models for embodied multi-agent cooperation. In *International Conference on Learning Representations*, volume 2025, pp. 49996–50019, 2025a.

Jiahan Zhang, Muqing Jiang, Nanru Dai, Taiming Lu, Arda Uzunoglu, Shunchi Zhang, Yana Wei, Jiahao Wang, Vishal Patel, Paul Liang, et al. World-in-world: World models in a closed-loop world. In *International Conference on Learning Representations*, volume 2026, pp. 55660–55699, 2026b.

Yichi Zhang, Xin Luna Dong, Zhaojiang Lin, Andrea Madotto, Anuj Kumar, Babak Damavandi, Joyce Chai, and Seungwhan Moon. Proactive assistant dialogue generation from streaming egocentric videos. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 12055–12079, 2025b.

Ethics Statement

This study reuses recorded human activities from existing datasets and evaluates model responses rather than intervening on participants. Person and object references serve action-grounding tasks. The readiness abstraction does not establish an individual's physical capacity or the safety of a real instruction, and scores should not be used for such decisions. Any redistribution of source imagery remains subject to the source datasets' access and usage conditions.

A Scope and Limitations

The benchmark uses a compact, curated visual pool with repeated questions per image and uneven template frequencies. It measures visible entry conditions within fixed operation boundaries, not force, dexterity, intention, execution safety, or all real-world requirements. The final version was developed with earlier model results available; it is not an independent model-blind test. Dynamic answers follow fixed symbolic completion rules rather than observed future outcomes. Because detailed effect conventions are not included in the evaluated requests, dynamic scores can also reflect disagreement about an action's completion endpoint. We do not isolate perception from action-semantic inference, demonstrate that every answer depends on the image, or measure downstream assistance success. Output budgets and provider implementations differ in documented ways. Future extensions can address broader visual coverage and closed-loop assistance; the present evidence already identifies complete constraint diagnosis and immediately usable preparation as distinct targets for improvement.

B Task Composition and Reference Answers

Table 3 records the label composition of the frozen evaluation. These counts describe the constructed questions, not the prevalence of actions or bodily constraints in daily life. In particular, the BD readiness component has a substantial majority-class baseline, and PS includes both answerable and empty-option cases.

Task relationships. The six tasks test different questions over shared state and action abstractions; they are not interchangeable measures on a single difficulty ladder. AR/AS query current admissibility, BD asks for the complete diagnosis, and RR specifies a hypothetical state edit. PS additionally requires an available action that repairs the target, whereas SV checks a supplied order without asking the model to generate a plan. Neither task evaluates unrestricted goal achievement or closed-loop assistance.

Table 3: Reference-answer composition. Sets contain all and only the qualifying options or failed channels.

Task	n	Composition
AR	170	Ready: 85; not ready: 85.
AS	130	Correct-set sizes $0/1/2/3/4/5$: $48/28/35/15/3/1$. Offered-set sizes $2/3/4/5$: $12/24/32/62$.
BD	344	No blocker: 75; hand: 75; reach: 76; pose: 39; hand+reach: 77; pose+reach: 2.
RR	188	All targets initially not ready; after the supplied edit, 98 ready and 90 still not ready. Hand/pose/reach edits: $92/3/93$.
PS	80	One correct option: 47; empty set: 33. Offered-set sizes $2/3/4/5$: $31/37/8/4$, totaling 225 candidate slots.
SV	278	First-failure indices $0/1/2/3$: $80/80/79/39$. Two-step sequences: 120; three-step sequences: 158.

C Action Boundaries and Prerequisites

This appendix defines the operation boundaries and bodily prerequisites of the 42 action templates. A grounded action $a = \langle \tau, \rho \rangle$ specifies an operation template τ and its bindings ρ to the designated person and relevant objects. Table 4 defines the starting stages and included movements. Table 5 associates each template with its boundary and requirements on H_p, P_p, and Q_p. Object roles are named by their function in the operation.

Reading the requirements. Human readiness concerns whether the designated person p can directly begin the specified operation. Hand prerequisites are evaluated against H_p: a hand can be available when free or in releasable contact, while holding an unrelated object occupies it. Control denotes an operation-relevant relation to the bound object; image overlap or arbitrary contact does not establish control. An unspecified-side one-hand requirement can be satisfied by either qualifying hand; two-hand requirements must be satisfied by distinct physical hands. Established control may imply reachability of the same object; reachability of a different destination must be assessed separately through Q_p. Posture requirements are evaluated against P_p. A reach failure does not by itself constitute a separate posture failure unless the operation also has an independent posture requirement. In Table 5, St, Be, Si, Sq, and Kn denote standing, bent forward, sitting, squatting, and kneeling; All permits all five. None in a prerequisite column means no explicit requirement in that channel, not that the corresponding visual fact is known to be absent.

Operation boundaries. The boundary descriptions in Table 4 specify what belongs to each operation. A preparation outside that boundary cannot be silently completed before the target operation begins. The descriptive boundary names are shared across templates with the same starting-stage convention; the individual hand, posture, and reach requirements remain those in Table 5.

Table 4: Operation boundaries shared by the action templates.

Boundary	**Included operation and starting stage**
Direct	Reaching and first contact belong to the operation. Direct two-handed operation includes establishing two-hand contact with the target, without first obtaining a separate tool. Clearing occupied hands, changing posture, or relocating beforehand remains separate.
Acquire	Reaching, initial contact, and acquisition of the bound object belong to the operation. Clearing hands, changing posture, or relocating beforehand remains separate.
Use/place	The source object or tool is controlled before the operation begins; first acquisition is separate. Movement toward the bound destination belongs to the operation, subject to the current reach requirement.
Aside	The source object is controlled, even if it still rests on a support. First acquisition is separate. No destination is bound by this template.
Package	One hand controls the supported package and the other opens it. First obtaining the package is separate.

Continued on next page

Table 4 continued

Boundary	Included operation and starting stage
Stabilize	The specified two-hand operation begins with distinct hands controlling the two named roles. First acquiring either object is separate.
Tear	Reaching, first steadying contact with the roll, and tearing form one operation; prior roll control is not required. Releasing another held object or relocating beforehand is separate.
Carry	The plate is already controlled with both hands. Movement and straightening as the plate begins to move are included; first acquiring the plate is separate.
Walk	Walking to the bound target is the operation. Being at the target is not a prerequisite; other preparations remain separate.
Posture	The named posture transition is included. Additional relocation or other preparation is separate.
Sit	Lowering, in-place turning or pivoting, and small aligning foot adjustments are included. Walking to or around the seat beforehand is separate.
Receive	Reaching, grasping, and taking the offered object are included in an ongoing handoff. Clearing a hand or moving position beforehand is separate.
Give	Handing over the controlled object is included in the ongoing handoff. First acquiring the object is separate.
Controlled use	Both hands begin in operation-appropriate contact with or control of the mobile object. First acquisition is separate.

Object-level eligibility and role binding. Object-level eligibility is determined by action–role compatibility and the object states recorded in F. An object-directed action must be appropriate for the bound entities and their current states before its bodily prerequisites are evaluated. These checks are separate from hand, posture, and reach prerequisites. Posture-only transitions have no object-state gate. Searching behind a cushion does not require first removing it; gluing concerns a part steadied by the other hand; the seat must be directly sittable rather than merely reachable by hand; two-handed acquisition must be appropriate for the instance; and receiving or giving requires a handoff that is in progress rather than complete. In handoff rows, p is the receiver or giver as specified. Roles in the tables refer to the entities bound by ρ; a contextual role does not necessarily introduce a reach prerequisite.

Table 5: All 42 instantiated action templates and their bodily prerequisites. Roles identify bound entities; the reach column lists only roles with an explicit reach requirement.

Action and roles	Boundary	Hand prerequisites	Posture	Reach
Open hinged storage Roles: storage	Direct	One available hand	All	storage
Close hinged storage Roles: storage	Direct	One available hand	All	storage
Open a drawer or storage compartment Roles: storage compartment	Direct	One available hand	All	storage compartment
Close a drawer or storage compartment Roles: storage compartment	Direct	One available hand	All	storage compartment
Search upper storage by hand Roles: storage	Direct	One available hand	All	storage

Continued on next page

Table 5 continued

Action and roles	Boundary	Hand prerequisites	Posture	Reach
Search lower storage by hand Roles: storage	Direct	One available hand	All	storage
Search behind a cushion by hand Roles: cushion	Direct	One available hand	All	cushion
Place an object aside Roles: object	Aside	One hand controlling object	All	None
Take an object from upper storage Roles: object and storage	Acquire	One available hand	All	object
Take an object from lower storage or a support Roles: object and storage/support	Acquire	One available hand	All	object
Pick up an object from a surface Roles: object; optional surface	Acquire	One available hand	All	object
Place an object on a surface Roles: object and surface	Use/place	One hand controlling object	All	surface
Place an item into a container Roles: container and object	Use/place	One hand controlling object	All	container
Attach a sticky note to a surface Roles: sticky note and surface	Use/place	One hand controlling sticky note	All	surface
Open a supported package with the other hand Roles: package	Package	One hand controlling package and a distinct available hand	All	None
Empty a controlled container onto a surface Roles: surface and container	Use/place	One hand controlling container	All	surface
Pick up a designated item from a surface Roles: surface and item	Acquire	One available hand	All	item
Apply glue while steadying a part Roles: glue and part	Stabilize	One hand controlling glue and a distinct hand controlling part	All	part
Attach a part while steadying a base Roles: base and part	Stabilize	One hand controlling part and a distinct hand controlling base	All	base
Place a pan on a stove Roles: pan and stove	Use/place	One hand controlling pan	All	stove
Insert a controlled item into a receptacle Roles: item and receptacle	Use/place	One hand controlling item	All	receptacle
Operate an appliance control Roles: control	Direct	One available hand	All	control
Pour oil into a pan Roles: oil container and pan	Use/place	One hand controlling oil container	All	pan

Continued on next page

Table 5 continued

Action and roles	Boundary	Hand prerequisites	Posture	Reach
Crack an egg into a pan with both hands Roles: egg and pan	Use/place	Two hands controlling egg	All	pan
Cook food in a pan with a spatula Roles: pan and spatula	Use/place	One hand controlling spatula	All	pan
Steady a roll and tear off a sheet Roles: paper-towel roll	Tear	Two available hands	All	roll
Wipe a pan with a paper towel Roles: pan and paper towel	Use/place	One hand controlling paper towel	All	pan
Transfer food from pan to plate with a spatula Roles: pan and plate and spatula	Use/place	One hand controlling spatula	All	pan, plate
Season food with a bottle Roles: food and seasoning bottle	Use/place	One hand controlling seasoning bottle	All	food
Carry a plate with both hands Roles: plate	Carry	Two hands controlling plate	St, Be	None
Walk to a target Roles: target	Walk	None	St	None
Stand up Roles: no object	Posture	None	Sq, Kn, Si	None
Sit down on a seat Roles: seat	Sit	None	St, Be	seat
Squat down Roles: no object	Posture	None	St, Be	None
Kneel down Roles: no object	Posture	None	St, Sq, Be	None
Pick up an object with both hands Roles: object	Acquire	Two available hands	All	object
Directly operate an object with both hands Roles: object; optional context	Direct	Two available hands	All	object
Place an object on a surface with both hands Roles: object and surface	Use/place	Two hands controlling object	All	surface
Receive an offered object with one hand Roles: receiver; giver and object	Receive	One available hand	All	object
Receive an offered object with both hands Roles: receiver; giver and object	Receive	Two available hands	All	object

Continued on next page

Table 5 continued

Action and roles	Boundary	Hand prerequisites	Posture	Reach
Give an offered object with one hand Roles: giver; object and receiver	Give	One hand controlling object	All	receiver
Use a controlled object with both hands Roles: object	Controlled use	Two hands in operation-appropriate contact with or control of object	All	None

D Construction, Action Effects, and Scope

The construction workflow separates proposal, human assessment, and deterministic label derivation (Table 6). The review of visual state does not imply that every final QA pair was separately relabeled by a human. The final answers are derived from the frozen state annotations and rule versions. Earlier model evaluations were available during benchmark development. For the final diagnosis expansion, we used reference category quotas rather than selecting individual questions according to model correctness.

Table 6: Roles of the six construction phases.

Phase	Purpose	Output used downstream
Action definition	Specify action identities, boundaries, and bodily requirements.	Fixed action catalog and reference rules.
Candidate proposal	Prepare frames and propose entities, candidate actions, and visual state.	Initialized records for human review; proposals are not final labels.
Visual review	Review relevant visual facts and grounded roles.	Person-specific annotations with uncertainty.
Reference derivation	Bind actions and evaluate frozen requirements.	Readiness, blockers, unknowns, and non-body exclusions.
Question construction	Construct task-specific questions and reference answers.	Candidates with source provenance and derivations.
Evaluation	Freeze model inputs and evaluate outputs.	Exact requests, answers, and strict scores.

Reference effects. Effects supplement entry requirements only for tasks that simulate completed actions. They do not follow automatically from $R(a)$. A completion convention fixes the endpoint needed to interpret a hypothetical successor state. Representative conventions include the following:

- Acquiring a specified object establishes the participating hand's control of that object; it does not free a different occupied hand.
- Placing an object and releasing it makes the participating hand available. The placed object is treated as reachable at that immediate completion instant, before any subsequent posture or location change.
- For opening and closing storage, the reference completion state makes the operating hand available, without retaining its prior control relations.
- Using a held tool retains control of it unless the specified operation includes release. Pouring does not silently put the container aside.
- A specified posture transition updates posture; environmental reach relations affected by that change require a determined update rather than automatic inheritance.
- Moving to a designated target establishes the declared target relation. It does not make every object in the destination region reachable.

These are reference-model conventions, not guarantees about all physical executions of the same verb. Only questions with sufficiently determined successor facts for their requested answer are admitted. Actions whose endpoint cannot support a determinate update cannot occupy a simulated position merely because their current readiness is known. For SV, once a step cannot begin, later successful effects are never applied.

Body versus environment. The bodily representation abstracts hand–object relations, posture, and reach. Object facts and compatibility remain necessary to ground meaningful queries: a mismatched object or an inapplicable object state is not evidence that the person's body needs preparation. These checks are kept separate from the three blocker channels. The abstraction does not include strength, pain, skill, fatigue, risk, or a complete geometric collision model.

E Model Requests and Output Contract

The evaluated inputs use natural-language actions and images annotated with bounding boxes to identify the referenced people and objects. When numerous object boxes obscure visual details, the original unannotated image is also provided alongside the annotated version. For each question, all models receive the same images and text. Private current-state annotations, action-level entry rules, and gold derivations are withheld. RR includes its hypothetical edit as part of the question. The internal action-effect catalog defines the reference progression. The final PS/SV requests ask for ordinary successful completion but supply neither the full catalog nor a separate per-step completion-convention paragraph. Consequently, detailed reference endpoints are construction assumptions rather than explicit premises in those evaluated requests. The evaluation therefore concerns grounded judgments under the given task wording, rather than a text-state planning task with the entire domain model provided.

Table 7: Output schemas. Values shown illustrate schema syntax, not a benchmark answer. The action IDs refer to displayed options.

Task	Required object
AR / RR	`{"ready": true}` or `{"ready": false}`.
AS / PS	`{"selected": ["A1", "A3"]}`, with the complete set in displayed order; an empty array is permitted.
BD	`{"ready": false, "blockers": ["hand", "reach"]}`; blocker channels use the prescribed hand, pose, reach order.
SV	`{"first_failure": 2}`, with an integer in $0, \ldots, L$. Zero denotes all steps passing.

Strict scoring accepts the specified JSON object, including ordinary surrounding JSON whitespace, but not a Markdown code fence, prefatory explanation, extra fields, or a repaired value. Schema validity and task correctness are distinct: a valid object with the wrong set remains incorrect. All questions remain in the denominator. A relaxed diagnostic reading of existing responses, reported separately below, does not change official predictions or scores.

F Recorded Evaluation Configurations

Table 8 summarizes the recorded settings. Model names identify the evaluated configurations as recorded in the result files. Token limits and provider behaviors differ; the comparison evaluates these configurations on shared questions and does not isolate architecture, compute budget, or parameter count. The same frozen input is used for a given question across models. The input bank combines 1,040 existing inputs with 150 additional blocker-diagnosis inputs, without model-specific rewrites. There is no system message or provider-specific structured-output constraint. This does not imply identical hidden provider infrastructure or a uniform single-example latency setup.

G Additional Aggregate Analyses

Formatting sensitivity. Table 9 separates strict performance from an auxiliary reading of the same raw responses. These diagnostic numbers cannot be substituted into the main table, whose scoring

Table 8: Generation configurations. An exhausted budget does not trigger answer repair in the official score.

Models	Configuration	Relevant qualification
GPT-5.6 Terra / Sol	Medium reasoning effort; 8,192-token maximum.	Existing Sol responses reused only when per-question input hashes matched.
GPT-6 Sol / Astra	Medium reasoning effort; 1,024-token maximum.	17 / 1 questions exhausted the budget without a complete answer.
Claude Opus 5.5 / Gemini 3.8 Flash	Medium reasoning effort; 1,024-token maximum; provider-default sampling.	Provider implementations are not assumed equivalent.
Local open models	vLLM on four B200 GPUs; temperature 0; usually 256 output tokens.	Typically one replica per GPU; Qwen3.8-Flash-Next uses four-way tensor parallelism.
Qwen3-VL-8B-Thinking	Reasoning parser; 8,192-token maximum.	75 questions exhausted the budget without a complete final answer.

contract includes formatting. In particular, a large relaxed gain measures sensitivity to the answer interface; it does not by itself identify the underlying visual reasoning quality. For Opus, only four responses violate the output schema. Even counting all four as correct would raise overall accuracy from 74.96% to just 75.29%, showing that output-format failures account for only a small fraction of its errors.

Table 9: Auxiliary relaxed reading of existing outputs; no new model calls. The official comparison uses only the strict column.

Model	Strict	Relaxed	Relaxed BD	Still unparsed
InternVL3.5-8B	29.16	34.96	32.56	0
Kimi-VL-A3B	17.98	33.45	21.80	18
InternVL3.5-1B	16.55	29.08	16.86	84

Source-group uncertainty. The reported paired bootstrap uses 89 source groups, 5,000 resamples, and seed 20260923. Each sampled group carries its associated questions. A single original recording may contribute multiple groups because handoff events are grouped separately; dependence across those groups is not captured. Descriptive Overall differences in percentage points are Opus–Astra $+2.27$ ($[-0.10, 4.56]$), Opus–GPT-6 Sol $+4.12$ ($[1.19, 6.92]$), and Astra–GPT-6 Sol $+1.85$ ($[-0.61, 4.26]$). They address within-bank source dependence; neither unseen-domain generalization nor model-development selection uncertainty is captured.

State-update and source strata. On SV questions whose answers depend on state updates versus the remaining questions, Opus scores 108/137 versus 114/141, Astra 118/137 versus 107/141, and Flash-Next 53/137 versus 82/141. The direction varies across models, and these subsets also differ in sequence length and content. Similarly, Opus scores 358/498 on Nymeria and 534/692 on Ego-Exo4D, and 519/702 on single-image versus 373/488 on two-image requests. Source and image-format groups were not randomly assigned; these comparisons do not isolate a domain or additional-image effect. No no-image intervention was completed for this release.

H Diagnostic Definitions and Model-Wise Counts

The main table uses strict whole-question accuracy throughout. Additional diagnostics answer specific questions about observable errors (Table 10); they are not averaged into a replacement benchmark score. Invalid outputs remain incorrect in official metrics. For conditional analyses, the valid or correctly judged subset is explicitly counted, and an empty denominator is reported as undefined rather than zero.

Table 10: Diagnostic questions for the six tasks. These distinguish output behavior, not uniquely identified internal causes.

Task	Diagnostic distinction	Denominator / interpretation
AR	False acceptance versus false rejection.	Ready and not-ready recall: 85 cases each; invalid outputs count as incorrect within their reference class.
AS	Extra invalid options versus missed valid options.	Precision/recall pool option decisions over schema-valid questions; whole-set accuracy remains the primary metric.
BD	Wrong readiness versus incomplete or spurious blockers.	Conditional complete diagnosis uses gold-not-ready, schema-valid, readiness-correct cases. Channel omissions use true-blocker opportunities; additions use absent-blocker opportunities.
RR	Rejecting an enabling edit versus accepting an insufficient edit.	98 enabled and 90 still-blocked cases; a supplied edit is not an executed action.
PS	Missing a valid preparation versus selecting an unready or ineffective candidate.	Question exactness: 47 positive and 33 empty cases. Candidate error rates use schema-valid opportunities in each class.
SV	Premature failure, missed/delayed failure, and paired order errors.	First-failure confusion includes all-pass as a separate outcome; pair exactness requires both answers correct out of 31 pairs.

Table 11: Denominators for Figure 3. AS valid counts schema-valid questions; precision and recall are option-level ratios. BD is complete diagnosis among correctly rejected gold-not-ready actions. SV requires both answers in a pair to be correct. Model-dependent conditioning does not yield identical subsets across rows.

Model	AS valid	AS precision	AS recall	BD conditional	SV paired
Claude Opus 5.5	130/130	135/169	135/160	165/232	21/31
GPT-6 Astra	130/130	136/181	136/160	157/229	20/31
GPT-6 Sol	127/130	137/175	137/158	152/216	21/31
GPT-5.6 Sol	130/130	138/187	138/160	131/207	18/31
Gemini 3.8 Flash	95/130	90/108	90/119	172/240	16/31
GPT-5.6 Terra	130/130	135/204	135/160	112/179	6/31
Qwen3.8-Flash-Next	130/130	117/230	117/160	117/239	5/31
Qwen3.8-27B	130/130	142/266	142/160	109/238	10/31
Qwen3-VL-2B-Instruct	129/130	95/273	95/157	29/114	0/31
Qwen3-VL-4B-Instruct	130/130	89/201	89/160	71/211	1/31
Qwen3-VL-8B-Instruct	130/130	113/284	113/160	53/158	0/31
Qwen3-VL-32B-Instruct	130/130	117/258	117/160	69/253	3/31
Qwen3-VL-8B-Thinking	119/130	109/204	109/145	54/181	7/31
MiniCPM-V-4.5	130/130	84/233	84/160	38/146	2/31
Molmo2-8B	130/130	69/153	69/160	34/87	4/31
InternVL3.5-8B-Instruct	130/130	99/296	99/160	11/40	0/31
Kimi-VL-A3B-Instruct	32/130	12/51	12/35	38/119	1/31
InternVL3.5-1B-HF	0/130	–	–	–	0/31

Matched-instance judgments. Action IDs identify 60 AR–AS and 120 AR–BD exact action-instance pairs. For AR–AS, Opus, Astra, and Flash-Next disagree on readiness in 6/59, 5/60, and 12/60 pairs with two schema-valid answers. For AR–BD the corresponding counts are 5/118, 9/120, and 10/120. These compare the same initial action query across question formats; they do not isolate a causal effect of explanation or candidate count.

Size and configuration comparisons. Within-family size comparisons use Qwen3-VL Instruct at 2B, 4B, 8B, and 32B. The 8B Thinking configuration is shown separately: its reasoning mode and 8,192-token limit differ from the Instruct settings. Other model families provide complementary

Table 12: Preparation Selection across all models. The first two result columns retain the full strict question denominators. The last two give selected invalid candidates over schema-valid opportunities of the relevant type; denominators can shrink when a model fails the output contract.

Model	Valid option	Empty set	Unready A selected	Ineffective A selected
Claude Opus 5.5	47/47	20/33	12/145	1/33
GPT-6 Astra	47/47	17/33	15/145	2/33
GPT-6 Sol	44/47	10/33	23/145	4/33
GPT-5.6 Sol	45/47	15/33	17/145	3/33
Gemini 3.8 Flash	43/47	21/33	5/127	1/27
GPT-5.6 Terra	43/47	8/33	27/145	3/33
Qwen3.8-Flash-Next	45/47	5/33	25/145	8/33
Qwen3.8-27B	41/47	2/33	28/145	12/33
Qwen3-VL-2B-Instruct	14/47	0/33	61/145	9/33
Qwen3-VL-4B-Instruct	1/47	31/33	4/145	0/33
Qwen3-VL-8B-Instruct	15/47	14/33	39/145	9/33
Qwen3-VL-32B-Instruct	19/47	13/33	25/145	5/33
Qwen3-VL-8B-Thinking	37/47	2/33	23/120	5/25
MiniCPM-V-4.5	18/47	0/33	56/145	8/33
Molmo2-8B	32/47	1/33	41/145	11/33
InternVL3.5-8B-Instruct	12/47	0/33	40/145	20/33
Kimi-VL-A3B-Instruct	15/47	0/33	37/82	13/30
InternVL3.5-1B-HF	0/47	0/33	–	–

profiles, not controlled scaling experiments. Aggregate group differences must not be interpreted as effects of weight availability or a single architecture choice.

I Action Coverage and Provenance

The final questions span 42 action templates, counted across all action positions. Table 13 distinguishes questions containing a template, action slots (allowing repeated occurrences within one question), and direct target occurrences in AR/BD/RR/PS. These counts need not agree, and their question totals must not be summed into a visual sample count. The final 2,255 action slots include 737 posture or movement slots (32.68%); 368/1,190 questions (30.92%) include at least one such action. The remaining 822 questions (69.08%) contain none of these five templates. The five are walking to a target, standing up, squatting, kneeling, and sitting down. This is a template-presence partition, not a claim that the remaining questions are independent of posture. Low-frequency templates cannot support reliable standalone performance claims.

Table 13: Final template coverage. A question may involve multiple templates.

Template	Questions	Slots	Direct targets
Walk to a target	247	260	91
Acquire an object with both hands	156	159	83
Stand up	152	152	0
Squat down	142	142	0
Move an object aside	115	128	42
Kneel down	109	109	20
Direct two-handed operation	106	108	60
Place a held object	94	104	34
Place an object with both hands	78	84	36
Close a hinged storage door	76	77	32
Sit down on a seat	74	74	42
Pick an object up from a surface	70	74	33

Table 13 (continued)

Template	Questions	Slots	Direct targets
Place an item in a container	69	70	23
Search behind a cushion	67	67	20
Operate an appliance control	65	65	20
Open a hinged storage door	59	60	18
Search upper storage by hand	51	51	17
Search lower storage by hand	46	46	17
Close a drawer	42	45	14
Open a drawer	40	40	14
Empty a held package	36	36	17
Retrieve an object from below	32	32	18
Carry a plate with both hands	30	30	12
Retrieve an object from above	30	30	13
Tear a paper towel	30	30	11
Open a held package	25	25	8
Pick up a toy part	20	20	7
Use a controlled object with both hands	20	20	12
Place a pan on a stove	18	18	7
Insert bread into a toaster	14	14	9
Attach toy parts	13	13	5
Give an offered object with one hand	13	13	13
Attach a sticky note	12	12	6
Apply glue to a part	10	10	3
Receive an offered object with both hands	9	9	9
Receive an offered object with one hand	6	6	6
Crack an egg into a pan	5	5	3
Pour oil into a pan	5	5	3
Wipe a pan with a towel	4	4	2
Cook with a spatula	3	3	1
Season food	3	3	1
Transfer food to a plate	2	2	0

The final evaluation uses 280 physical images and 287 person-specific records. The source annotation pool contains 293 images and 300 records; the additional records arise from separately evaluating both participants in seven handoff images. The final question counts by source are 498 for Nymeria and 692 for Ego-Exo4D, covering 141 and 146 person-specific records respectively. Related frames and questions are linked by source group rather than treated as independent observations.