Inferring the Unspoken: Aligning Embodied Agents with Implicit Preferences

Manjie Xu [1,2,3,5,7,*] **Xinyi Yang** [1,2,3,5,7,*] **Wei Liang** [4,✉] **Chi Zhang** [3,5✉] **Yixin Zhu** [2,1,5,6,7,8✉]

* Equal contribution ✉ Corresponding author

[1] Institute for Artificial Intelligence, Peking University

[2] School of Psychological and Cognitive Sciences, Peking University

[3] School of Intelligence Science and Technology, Peking University [4] Beijing Institute of Technology

[5] State Key Lab of General AI, Peking University [6] PKU-Wuhan Institute for Artificial Intelligence

[7] Beijing Key Laboratory of Brain-Computer Interface and Mental Health Modulation [8] XSVision

https://sites.google.com/view/preference-based-planning

Abstract

Instructions are never complete specifications of the behavior they request: speakers say only what their listeners cannot supply for themselves (Grice, 1975; Zipf, 2016). An agent told to "prepare an apple" must decide unaided whether to wash or cut it and where to place it. Such choices differ between people and are seldom stated, only enacted (Slovic, 1995; Lichtenstein and Slovic, 2006). Multimodal foundation models render instructions into grounded plans (Brohan et al., 2023; Driess et al., 2023), yet settle these residual choices by generic priors rather than by evidence about the person served (Zhang et al., 2024). Whether an agent can instead recover such constraints from prior behavior has not been separated from the easier question of planning once they are supplied. Here we show the failure to be largely one of acquisition rather than planning, and that stating the inferred preference before acting recovers much of what is lost. Preference-based Planning (PBP) benchmarks 290 preferences over action parameters, placement policies, and temporal orderings, asking agents to infer from a few unlabeled demonstrations which constraint governs a new underspecified instruction. Given the preference, the strongest models plan categorical tasks at near-zero error; required to infer it, performance drops sharply, and without demonstrations accuracy falls by over a third. Because a verbalized rule keeps intent while discarding perceptual particulars, preferences so represented transfer to unfamiliar scenes, whereas implicit imitation stays bound to the rooms it was learned in. The bottleneck lies not in generating actions but in the pragmatic inference preceding them, for which language proves an effective medium.

1 Introduction

Natural language is a flexible interface for human–robot collaboration, but the instructions it carries are rarely complete. By the Principle of Least Effort (Zipf, 2016) and Gricean pragmatics (Grice, 1975), speakers omit whatever they take the context to supply. In embodied settings the omission is consequential: a single instruction admits many valid executions, and which one is correct depends on both the environment and the user (Shridhar et al., 2020; Szot et al., 2021; Qiu et al., 2022; Wu et al., 2023). "Clean the kitchen" fixes no answer to where items belong, what must be hand-washed, or in what order the subtasks proceed, and individuals answer these differently (Shridhar et al., 2020; Wu et al., 2023) (Fig. 1). Personalized embodied agents must therefore move beyond literal instruction following toward *pragmatic grounding*: recovering implicit preferences from prior behavior and using them to settle what later instructions leave open.

Foundation models have advanced multimodal reasoning and high-level planning in embodied environments, letting agents translate visual observations and natural-language instructions into grounded action plans (Brohan et al., 2023; Huang et al., 2023; Driess et al., 2023; Wang et al., 2023a; Yang et al., 2025). This progress has been measured almost entirely against explicit task specifications, and personalization has lagged behind: agents fall back on generic behavioral priors and miss what is idiosyncratic to a user (Zhang et al., 2024). Closing the gap requires treating preference as a latent variable that differs across users and inferring it from what they do, a problem close to Theory of Mind (ToM) reasoning (Fawcett and Markson, 2010; Yuan et al., 2022; Ying et al., 2026). Psychology suggests why behavior is the evidence to use: preferences are seldom articulated outright, and are instead constructed or revealed in the act of choosing (Epstein, 1994; Slovic, 1995; Lichtenstein and Slovic, 2006; Simonson, 2008; Zhu et al., 2020). Embodied environments make the inference harder still, since the behavioral record available is multimodal, noisy, and few-shot.

The prevailing approach, *implicit visual imitation*, conditions action generation directly on

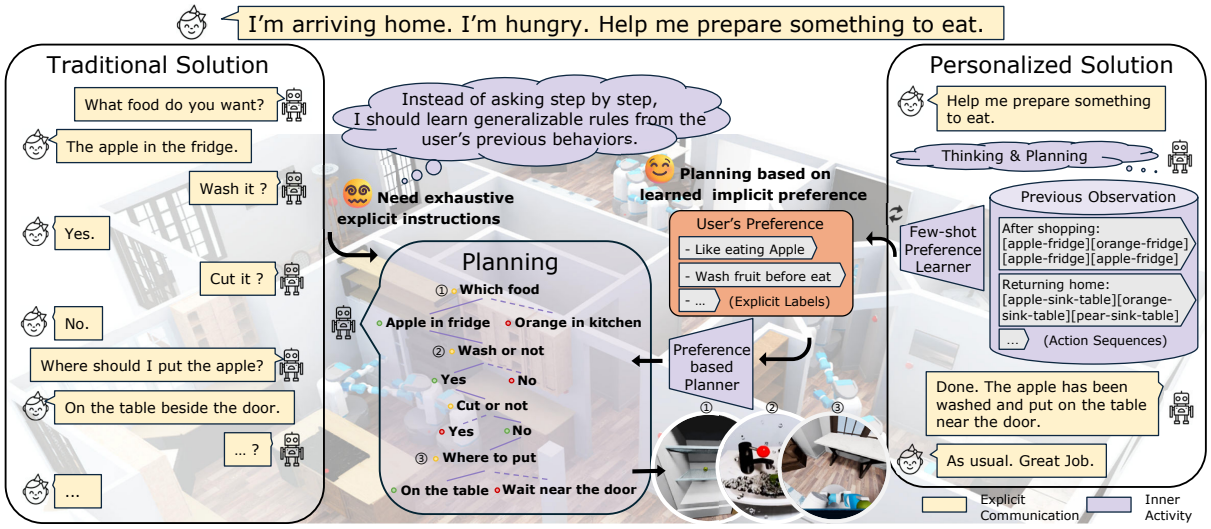


Figure 1: **The pragmatic gap in embodied planning.** An underspecified instruction such as "Prepare food" settles neither which item to use, nor whether to wash or cut it, nor where to put it down (center). A conventional agent recovers these one at a time by interrogating the user (left), whereas InTU (right) infers the governing preference from what the user has done before, verbalizes it (*e.g.*, "wash fruit before eating"), and plans on that basis. The verbalized preference is the intermediate representation bridging perception and action.

demonstration videos. In the few-shot regime this proves brittle. Demonstrations are high-dimensional and dominated by task-irrelevant variation in viewpoint, lighting, object instances, and scene layout, whereas the preference to be recovered is sparse and semantic (*e.g.*, "wash fruits before cutting," "store perishables in the fridge"). Mapping pixels to actions in a single step asks the agent to identify the user-specific constraint and ground a plan at once, so that a failure of either becomes indistinguishable from a failure of the other. Scarce and costly demonstrations sharpen the difficulty (Akgun et al., 2012). Preference-aware embodied learning has progressed in constrained domains such as rearrangement, yet remains task-specific and transfers poorly across the range of daily activities (Kapelyukh and Johns, 2022; Yuan et al., 2023; Liu et al., 2025b).

We separate the two stages with Inferring the Unspoken (InTU), a simple two-stage formulation for few-shot personalized embodied planning that first verbalizes the latent user preference from demonstrations and then plans conditioned on that statement. The factorization serves two ends: it localizes failure to acquisition or to planning, and it tests whether an explicit semantic representation improves personalization at all. Compressing noisy multimodal demonstrations into a compact, human-interpretable rule (*e.g.*, "Preference: peel apples before serving") interposes a semantic abstraction between what was observed and what is done. That explicit intermediate reasoning aids complex decision-making (Wei et al., 2022; Wang et al., 2023b); we examine language as a representation of a user's preference.

Studying this systematically calls for a testbed in which the preference, rather than the task, is what varies. We build Preference-based Planning (PBP), a benchmark for *pragmatic inference* in embodied planning, on a hierarchical taxonomy of 290 semantic preferences: Atomic Semantics fix fine-grained action parameters, Strategic Semantics fix categorical interaction and placement rules, and Temporal Semantics fix ordering and causal dependencies among subtasks. ALFRED (Shridhar et al., 2020), Habitat (Szot et al., 2021), and LoTa-Bench (Choi et al., 2024) score task completion against explicit specifications; PBP instead asks whether an unspoken user-specific constraint can be recovered from few-shot demonstrations and carried into a new scene.

Three conditions expose a clear *preference acquisition gap*: end-to-end planning from demonstrations, planning on preferences inferred through InTU, and planning on ground-truth preferences. Across State-of-the-Art (SOTA) Vision-Language Models (VLMs) and Large Language Models (LLMs), supplying the preference outright yields markedly better plans than requiring the model to infer that same preference from behavior, which places the bottleneck in acquisition rather than in downstream planning. Verbalization narrows the gap: conditioning on an inferred linguistic preference improves alignment over end-to-end con-

ditioning throughout. Performance rises with additional demonstrations and holds up better when preferences must transfer across visually distinct scenes. Explicit semantic representations thus supply a serviceable abstraction for carrying sparse user-specific constraints from behavior into action, leaving visual-to-semantic preference inference as the principal obstacle for current models.

Our contributions are threefold: **(i) PBP benchmark:** a large-scale benchmark that shifts evaluation from explicit instruction following to implicit preference inference, over 290 preferences organized by a hierarchical semantic taxonomy; **(ii) Preference acquisition gap:** a systematic diagnosis showing that current embodied agents plan effectively once user preferences are specified, but degrade substantially when the same preferences must be inferred from few-shot behavioral demonstrations; **(iii) Language-mediated preference acquisition:** evidence, through InTU, that verbalizing latent preferences before planning narrows this gap, and that language can therefore serve as an effective semantic abstraction for few-shot embodied personalization.

2 Related Work

Instruction following and pragmatic grounding Language-guided embodied agents have traditionally executed explicit task specifications, from Vision-and-Language Navigation (VLN) (Anderson et al., 2018; Chen et al., 2019) to household manipulation and rearrangement (Shridhar et al., 2020; Szot et al., 2021). Human instructions, however, are rarely exhaustive: speakers omit what shared context can recover, as the principle of least effort (Zipf, 2016) and pragmatic theories of communication (Grice, 1975) would predict. Embodied systems have accordingly begun reasoning about implicit goals, contextual cues, and human mental states (Jiang et al., 2022; Sarch et al., 2022; Xu et al., 2023; Ying et al., 2026). Yet these formulations resolve ambiguity from environmental context or generic commonsense, not from behavioral evidence about the particular user. We study the complementary problem of *pragmatic grounding from user history*: recovering latent user-specific preferences from a handful of behavioral demonstrations and applying them to underspecified instructions in new scenes.

Language as an intermediate representation Language increasingly serves as an intermediate representation for reasoning and planning. Chain-of-Thought and related prompting show that externalizing intermediate reasoning improves complex language reasoning (Wei et al., 2022; Wang et al., 2023b), while language and code supply structured abstractions between perception and action for task decomposition, affordance reasoning, and long-horizon planning (Brohan et al., 2023; Liang et al., 2023; Wang et al., 2023a; Xu et al., 2026; Liu et al., 2025a). Recasting multimodal observations as structured text or symbols likewise simplifies downstream planning by separating perceptual interpretation from action generation. We investigate a different intermediate variable: *user preference*. Rather than decomposing a task or describing an environment, we verbalize the latent behavioral regularity inferred from demonstrations before planning, and use this factorization as a diagnostic intervention: whether explicit preference acquisition accounts for the gap between observing user behavior and generating personalized plans.

Personalization and theory of mind in agents Personalization is well studied in recommendation and dialogue, and recent work extends it to LLMs through user-specific preferences, memories, and interaction histories (Packer et al., 2023; Wang et al., 2024; Xu et al., 2025). Embodied personalization adds a further demand: what is specific to a user must be inferred from situated interaction and behavior. Early robotic systems addressed preferences in constrained domains such as object rearrangement and household organization (Kapelyukh and Johns, 2022; Wu et al., 2023), and recent personalized embodied agents pursue memory-based user modeling and personalized context accumulation (Ziliotto et al., 2026; Kwon et al., 2026; Lee et al., 2026). These lines of work concern how user-specific information is retrieved or represented; PBP isolates a prior question, whether such preferences can be acquired from behavioral demonstrations at all and then used to resolve underspecified instructions. Separating acquisition from preference-conditioned planning gives a controlled account of where personalized embodied agents fail.

3 Problem Formulation

We cast personalized embodied planning as context-dependent preference inference: the agent must draw on prior behavioral evidence to resolve what the current instruction leaves open.

3.1 Task definition

Consider an embodied agent in a physical environment $\mathcal{E}$. The agent receives a natural-language instruction I (*e.g.*, "Prepare an apple") that is **semantically underspecified** and admits multiple valid executions. To resolve the ambiguity, it is given a few-shot behavioral context

$$\mathcal{C}_{\text{demo}} = \{X_i\}_{i=1}^{K}, \tag{1}$$

in which each X_i is an unlabeled demonstration, rendered as an egocentric video or, in the symbolic setting, as the corresponding action trace.

Demonstrations vary in behavioral pattern and task configuration, and each may exhibit several regularities at once, so no single demonstration determines the relevant preference on its own. What makes a regularity relevant is the query: given $\mathcal{C}_{\text{demo}}$, the instruction I, and the current environment $\mathcal{E}_{\text{curr}}$, the agent must recover the latent preference bearing on the task at hand,

$$\hat{S}_{\text{pref}} = \arg\max_{S} P(S \mid \mathcal{C}_{\text{demo}}, I, \mathcal{E}_{\text{curr}}), \tag{2}$$

and then act on it,

$$\hat{A} = \arg\max_{A} P(A \mid I, \hat{S}_{\text{pref}}, \mathcal{E}_{\text{curr}}). \tag{3}$$

3.2 Language-mediated preference acquisition

Planning end-to-end from demonstrations entangles two distinct problems: identifying which preference governs the current task, and grounding an action sequence that honors it. InTU separates them by making the linguistic representation S_{pref} an explicit intermediate variable. Instead of mapping behavioral history straight to actions, InTU reasons over the demonstrations in light of the current instruction and environment, states the inferred preference as a semantic constraint, and conditions planning on that statement. The two stages can then be scored apart: whether a model recovers the right preference from behavior, and whether it plans well once the preference is handed to it.

4 The PBP Benchmark

We introduce PBP, a benchmark for pragmatic inference in embodied planning. Prior embodied benchmarks score navigation or the execution of explicit manipulation goals (Shridhar et al., 2020); PBP instead requires that a latent semantic constraint be inferred from few-shot demonstrations and carried into a new, underspecified instruction.

Figure 2: **Preference levels in PBP.** The three pragmatic levels differ in what must be abstracted from behavior. (a–c) Atomic Semantics turn on a primitive action parameter, here whether the apple is microwaved, washed, or cut before eating. (d–e) Strategic Semantics turn on a categorical rule, here two incompatible schemes for arranging the same foods in a fridge. (f) Temporal Semantics turn on task order, which a single scene leaves ambiguous: the same kitchen affords eating an apple, eating cake, rearranging foods, or tidying, and only prior behavior settles which comes first.

4.1 Semantic taxonomy of preferences

PBP organizes 290 preferences into three pragmatic levels of increasing abstraction (Fig. 2). **Atomic Semantics (action level)** fix execution parameters that natural instructions leave out (*e.g.*, "fill halfway" *vs.* "fill completely"), so the difficulty lies in grounding subtle continuous state differences as discrete semantic attributes. **Strategic Semantics (option level)** fix interaction and placement policies (*e.g.*, "store perishables in the fridge"), which demands category-level abstraction: mapping instances onto abstract properties and carrying the rule across scenes. **Temporal Semantics (sequence level)** fix ordering and causal dependencies among subtasks (*e.g.*, "clean surfaces before placing items"), where the preference resides in relative action order rather than in any individual action. Section B details the dataset and the full hierarchy.

4.2 Data construction

PBP is built in OmniGibson and comprises 5,000 evaluation groups. Each group pairs (i) a behavioral history of $K=3$ unlabeled demonstration episodes spanning several tasks and scenes with (ii) a query consisting of a new scene and an underspecified instruction. Since the history encodes more than one regularity, the query is what selects the preference at issue. Gold annotations accompany every episode, including preference metadata and a canonical action sequence; neither is exposed to the model in the main preference-inference setting.

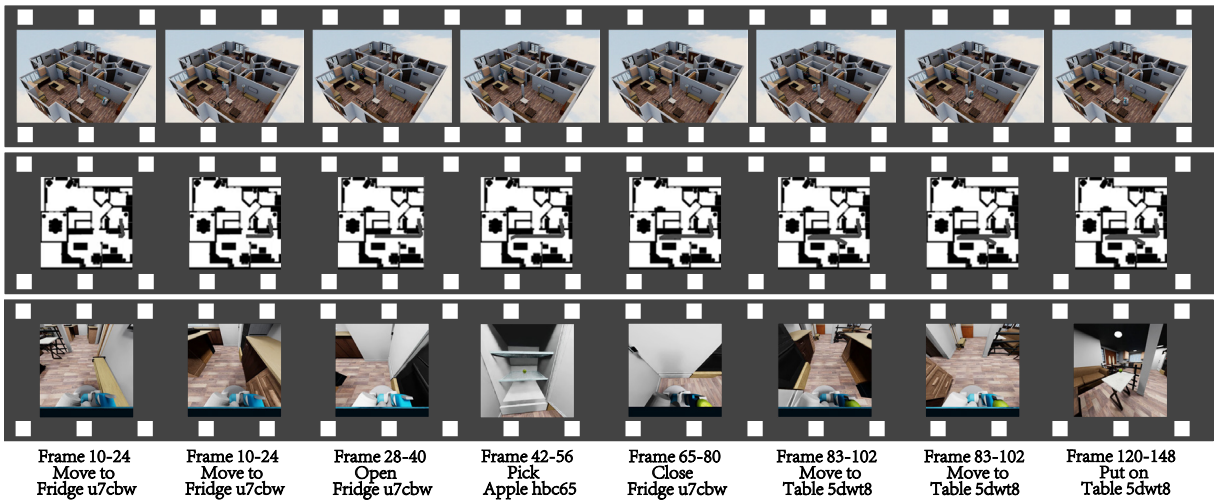


Figure 3: **Multimodal demonstration data in PBP.** Each demonstration supplies part of the context $\mathcal{C}_{\text{demo}}$ from which the latent constraint S_{pref} must be inferred (*e.g.*, "place apple on table"). The agent observes the egocentric stream (bottom) together with a bird's-eye map (middle) for spatial context; the third-person view (top) is rendered for visualization only. Frame-level action labels mark the primitives composing the episode.

4.3 Dataset statistics and diversity

PBP spans 50 interactive OmniGibson scenes (kitchens, living rooms, offices, grocery stores) and over 3,000 object assets, so that a preference cannot be recognized by appearance alone. The 290 preferences divide as:

- Action level: 75 types (continuous parameters, *e.g.*, tap temperature, slicing thickness);
- Option level: 135 types (storage and interaction strategies, *e.g.*, cabinet *vs.* counter);
- Sequence level: 80 types (long-horizon ordering constraints over 3 or more subtasks).

Together the taxonomy and the episode structure stress fine-grained perception and high-level semantic generalization at once. The main experiments concentrate on the option and sequence levels, whose categorical and temporal demands cannot be met by direct behavioral imitation, making them the sharper test of language-mediated preference acquisition; action-level preferences are included in the benchmark for completeness.

5 Experiments

Our experiments ask whether explicit preference acquisition supports robust personalized embodied planning, along four lines. First, whether end-to-end failures originate in planning itself or in the absence of a preference representation. Second, what verbalizing the preference buys over conditioning on demonstrations implicitly. Third, how acquisition and planning scale with the number of demonstrations. Fourth, whether an inferred preference survives transfer to a visually distinct scene.

5.1 Experimental setup

Models Video-based inference is tested with Gemini-3.1-Pro, LLaVA-NeXT (Liu et al., 2024), and EILEV (Yu et al., 2023). Gemini-3.1-Pro, a closed-source VLM, receives keyframes extracted at 1 Hz and interleaved with text prompts. LLaVA-NeXT (Video) encodes the demonstrations natively as video. EILEV, built for few-shot in-context learning from embodied video, tests whether specialized pre-training helps preference inference. We add Video Vision Transformer (ViViT) (Arnab et al., 2021), a pure-Transformer video encoder trained end-to-end on spatio-temporal features; lacking any LLM component, it lower-bounds commonsense-driven pragmatic inference in PBP. To disentangle reasoning from perception we also evaluate LLMs on symbolic input (GPT-5.5, Claude-Opus-4.7, DeepSeek-v4-pro, Llama-3-8B), replacing the visual demonstrations with ground-truth scene graphs and action logs. GPT-5.5 and Claude-Opus-4.7 accept video as well; the symbolic condition is used solely to isolate reasoning under perfect perception and thus approximates an upper bound on the planning side. We further provide a symbolic causal discovery model DAG-Opt (Zheng et al., 2018).

Protocols The **End-to-End** setting has the model generate an action sequence directly from the instruction and demonstrations, with no mention of latent preferences in the prompt, as users rarely articulate every constraint in practice. The **Language-Mediated** setting has it first verbal-

Table 1: **Planning error under three sources of preference** (Levenshtein distance ↓; lower is better aligned with the user). End-to-end plans directly from demonstrations and aligns poorly. Two-stage verbalizes the inferred preference first and cuts the error roughly in half. Second-stage (gt) substitutes the ground-truth preference, so its residual is planning error alone; the distance between it and Two-stage is what acquisition costs. ViViT has no language stage and appears only as an end-to-end lower bound. Values are mean ± range over 3 independent runs.

	VIDEO-BASED INPUT				SYMBOL-BASED INPUT			
	ViViT	LLaVA-NeXT	EILEV	Gemini-3.1-Pro	Llama3-8B	DS-v4-pro	Claude-Opus-4.7	GPT-5.5
Option Level								
End-to-end	15.49±1.29	15.94±3.41	12.88±2.20	8.72±2.26	14.74±3.21	6.42±2.33	6.72±2.48	6.07±2.00
Two-stage	-	12.46±3.23	12.89±3.74	4.28±1.76	9.67±5.16	2.31±1.67	2.56±2.04	2.41±1.72
Second-stage (gt)	-	3.28±5.29	11.18±4.20	0.16±1.82	8.22±5.58	0.16±1.46	0.16±1.99	0.14±1.21
Sequence Level								
End-to-end	34.04±11.84	34.76±11.25	33.10±12.21	24.51±4.15	31.79±7.32	22.84±4.01	21.47±3.12	21.92±3.45
Two-stage	-	30.02±13.54	33.03±13.61	12.57±2.92	25.46±5.93	12.10±2.85	13.84±2.81	12.48±2.45
Second-stage (gt)	-	18.92±14.18	26.57±12.21	10.24±2.58	19.02±7.10	10.19±2.47	9.17±2.90	9.77±2.69
Overall								
End-to-end	24.76	25.35	**22.99**	16.62	23.26	14.63	14.10	14.00
Two-stage	-	21.24	22.96	8.43	17.56	7.21	8.20	7.45
Second-stage (gt)	-	11.10	18.88	5.20	13.62	5.18	4.67	4.96

ize an intermediate preference inferred from the demonstrations, then plan conditioned on that statement. Between these sits a **Preference-aware End-to-End** baseline, which asks the model to infer the preference internally but emit only the plan; comparing the three separates the benefit of reasoning about preferences from the benefit of exposing the result. Prompts follow the OpenAI Cookbook[1], and every experiment is run three times with fixed hyperparameters and no seed tuning. Section D gives implementation details and full prompts.

Isolating preference from comprehension A rule-based diagnostic confirms that every evaluated LLM satisfies the literal instruction over 95% of the time, where literal satisfaction asks only whether the plan addresses the stated objective, such as manipulating the requested object, irrespective of any hidden preference. The alignment gap we report therefore reflects preference inference and not a failure to understand the instruction.

5.2 Evaluation metrics

Two metrics track the two stages of the pipeline.

Preference inference accuracy (Acc). Whether the verbalized preference $\hat{S}_{\text{pref}}$ semantically matches the ground truth, judged by strict keyword matching together with semantic similarity checks.

Action alignment (Levenshtein distance). The edit distance between the generated sequence $\hat{A}$ and the canonical sequence A_{gt}, counting insertions, deletions, and substitutions. Being graded rather than binary, it discriminates degrees of misalignment that a success rate would collapse; lower is better. Where objects are interchangeable, an order-invariant mask keeps permutations over equivalents from being penalized. Strict ordering is enforced only for sequence-level preferences, in which the order is itself the preference.

5.3 Results and analysis

RQ1: Where does the error originate? Table 1 varies how the preference reaches the planner while holding the planner itself fixed. Handed the ground-truth preference (*Second-stage (gt)*), every model plans markedly better, and at the option level the strongest reach essentially zero edit distance. Sequence-level error persists, so specifying the preference removes much of the total error without dissolving the difficulty of long-horizon planning. Since nothing about the downstream architecture changed between conditions, the difference is attributable to acquisition. The modality comparison points the same way: Gemini-3.1-Pro closes much of the video–symbol gap left by earlier VLMs (overall end-to-end 16.62 *vs*. 24.76 for ViViT), yet still trails the symbolic models (GPT-5.5, 14.00), which locates the residual difficulty in visual-to-semantic acquisition rather than in planning.

RQ2: What does verbalization contribute? Across Table 1, stating the preference before planning lowers error consistently. Requiring a discrete constraint (*e.g.*, "the user prefers option A") relieves the planner of recovering intent and generating actions at once, and it renders acquisition errors visible rather than leaving them to surface as plan defects. Table A8 apportions the credit:

[1] https://cookbook.openai.com/

Table 2: **Accuracy of the verbalized preference** (↑), measuring whether S_{pref} is named correctly before any planning occurs. **Few-shot** supplies the $K{=}3$ demonstrations; **Ablative** withholds them, leaving only the instruction and scene. The collapse under Ablative, sharpest at the sequence level, shows that accuracy rests on the user's demonstrated behavior rather than on commonsense defaults about where things belong.

	VIDEO-BASED INPUT				SYMBOL-BASED INPUT				
	ViViT	LLaVA-NeXT	EILEV	Gemini-3.1-Pro	DAG-Opt	Llama3-8B	DS-v4-pro	Claude-Opus-4.7	GPT-5.5
Few-shot (With Context Demonstrations)									
Option Level	9.38	36.87	38.33	63.28	10.15	72.98	91.28	89.91	92.48
Sequence Level	4.24	24.85	32.69	52.10	13.49	67.18	76.48	76.92	78.01
Overall	6.81	30.86	35.51	57.69	11.82	70.08	83.88	83.42	85.25
Ablative (Zero-shot / No Context)									
Option Level	9.16	15.47	4.77	31.92	3.84	39.50	83.55	82.47	82.01
Sequence Level	4.38	8.13	0.00	14.28	1.28	6.25	18.58	18.02	18.71
Overall	6.77	11.80	2.38	23.10	2.56	22.88	51.07	50.25	50.36

the preference-aware end-to-end baseline already improves substantially over vanilla end-to-end, so reasoning about the latent preference is what matters most. Two-stage verbalization adds a further, consistent gain, largest at the sequence level, and exposes the intermediate representation for inspection and reuse.

RQ3: How much does context matter? *Preference prediction.* Table 2 isolates the first stage, where inference proves far from solved. Video-based models trail their symbolic counterparts throughout (Gemini-3.1-Pro, 57.69%, against GPT-5.5 at 85.25% and DeepSeek-v4-pro at 83.88%). The stratification is by input modality, not by model strength, which again identifies visual-to-text alignment as the binding constraint.

Few-shot vs. *zero-shot.* Withholding the demonstrations collapses accuracy across the board (Table 2, bottom): GPT-5.5 falls from 85.25% to 50.36%, Gemini-3.1-Pro from 57.69% to 23.10%. What survives is instructive. Option-level accuracy remains respectable because commonsense supplies a default, yet defaults do not settle which of several sound alternatives a particular user favors, eggs in the fridge or in the pantry. Sequence-level accuracy all but vanishes, temporal preferences being the least recoverable from generic priors.

Scaling demonstrations. Figure 4(c–d) tracks support sets of $k \in \{1, 2, 3, 5\}$. Additional demonstrations improve both prediction accuracy and downstream planning, but the returns diminish, and for EILEV the largest context ($k{=}5$) occasionally hurts, its extra frames arriving as noise. More context is not uniformly better when the stream is long and multimodal.

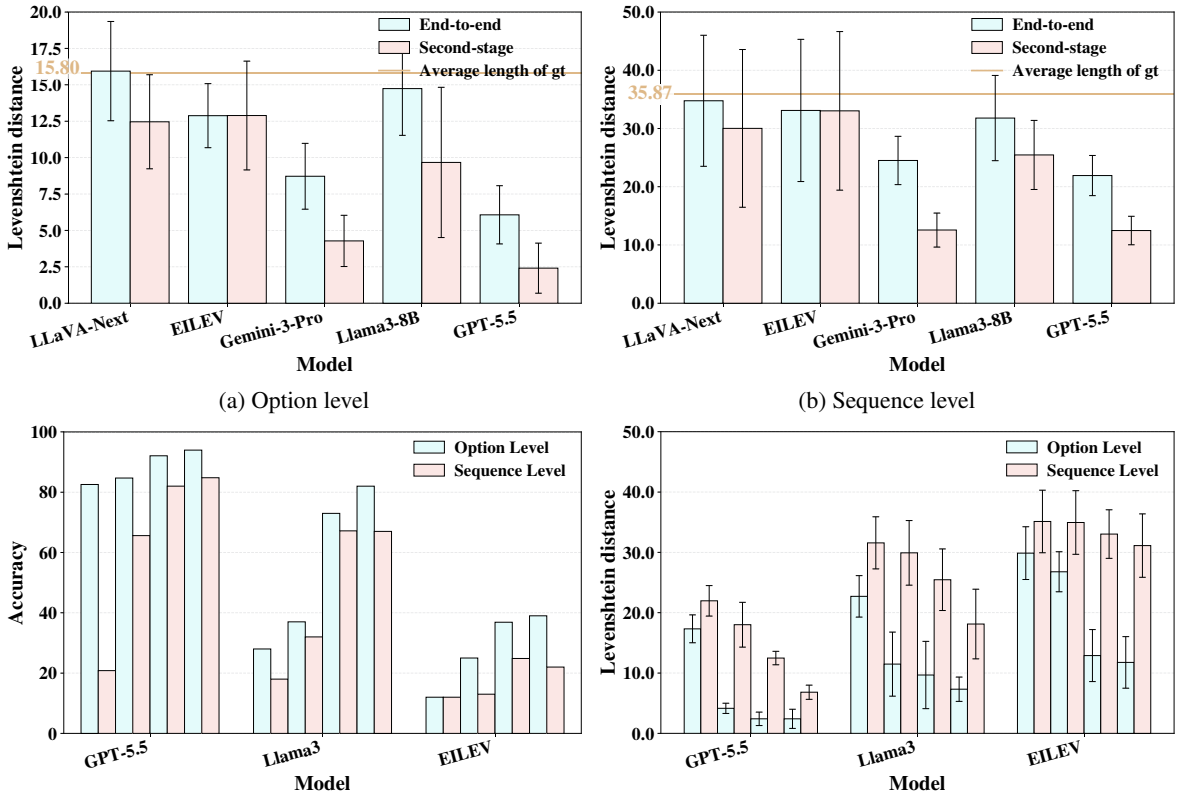
Diagnostic takeaway. The option level probes category abstraction, the sequence level temporal-causal reasoning, and the latter costs consistently 2–3× more error (Gemini-3.1-Pro, 8.72 *vs*. 24.51). Abstracting temporal structure from visual streams is the sharpest remaining difficulty.

Table 3: **Preference inference across scene shift** (accuracy ↑). Under *direct*, demonstration and test episodes share a room and its objects; under *gen*, they share only the preference. The symbolic models hold most of their accuracy across the shift, whereas the vision-based models give up more of theirs, having partly bound the preference to the scene it was demonstrated in.

	Option Level		Sequence Level	
Model	*direct*	*gen*	*direct*	*gen*
LLaVA-NeXT	33.25	36.87	33.12	24.85
EILEV	46.93	38.33	37.53	32.69
Gemini-3.1-Pro	65.89	63.28	68.21	52.10
DS-v4-pro	95.18	91.28	86.42	76.48
GPT-5.5	95.01	92.48	86.90	78.01

RQ4: Does an inferred preference survive a change of scene? A preference worth learning must outlast the room it was learned in. We therefore contrast a *direct* setting, in which demonstration and test episodes are rendered in the same room with the same objects, against a *generalization* (*gen*) setting, in which they share the preference but nothing of the scene. Table 3 separates the two families cleanly at the sequence level, where the ordering constraint cannot be read off any single frame. The symbolic models surrender little (GPT-5.5, 86.90 to 78.01; DeepSeek-v4-pro, 86.42 to 76.48): once a preference has been put into words it is invariant to pixels. The vision-based models give up more, Gemini-3.1-Pro falling from 68.21 to 52.10 and EILEV from 37.53 to 32.69, which suggests their few-shot "learning" attaches partly to background texture and object layout rather than to the behavioral rule, so that changing the scene breaks the correlation. Option-level accuracy is more stable for every model, commonsense priors carrying some of the burden there; LLaVA-NeXT even scores marginally higher under *gen*, though at roughly a third accuracy its two conditions are close to indistinguishable.

Figure 5 resolves this per sample. Certain scenes



(a) Option level (b) Sequence level

(c) Preference learning (d) Action planning

Figure 4: **Effect of verbalization and of context size.** (a–b) Verbalization narrows the alignment gap at both the (a) option and (b) sequence level: End-to-end generates the sequence directly from prior observations, whereas Two-stage first receives a predicted preference label. The — line marks the average ground-truth sequence length (15.80 for option level, 35.87 for sequence level), against which the residual distances should be read. (c–d) Enlarging the support set ($k \in \{1, 2, 3, 5\}$) generally improves both stages, (c) preference learning and (d) action planning, on Option Level and Sequence Level tasks alike, though the gain saturates in the visual encoders. Higher is better in (c), lower in (d).

defeat all models alike, a property of the environment rather than of any architecture. More tellingly, the vision-based models succeed on largely disjoint samples under *direct* and *gen*: what they acquire holds in the room where it was shown and lapses elsewhere. Verbalization counteracts precisely this, forcing scene-specific detail to be dropped and only the transferable rule retained.

5.4 Further Discussion

Language serves here not as a channel for communication but as a compact semantic interface between behavioral evidence and planning. Raw demonstrations carry a great deal of perceptual variation that bears on nothing; the preference within them is sparse and semantic. Verbalization compresses the one into the other, keeping intent and shedding trajectory detail, and it is this compression that carries preferences across environments.

Recovering the preference in the first place remains hard. Gemini-3.1-Pro, among the strongest VLMs available, names only about 63% of option-level and 52% of sequence-level preferences correctly (Table 2), so abstracting latent intent from behavior is far from solved. The shortfall is of a piece with weak ToM reasoning, and argues for training objectives that reward pragmatic inference from trajectories rather than fidelity to the trajectories themselves.

Qualitative analysis The failures are legible in individual episodes (Fig. 5). Given demonstrations of "wash fruits before cutting," end-to-end Gemini-3.1-Pro frequently omits the brief washing behavior and proceeds directly to cutting, responding to the salient objects in view rather than to the regularity they instantiate. The two-stage pipeline first commits to a constraint, *e.g.*, "the user goes to the sink before using the knife," and planning conditioned on that statement is far likelier to emit the requisite `toggle_sink` action. Verbalization steadies the preference representation and keeps subtle constraints from being forgotten over a long horizon.

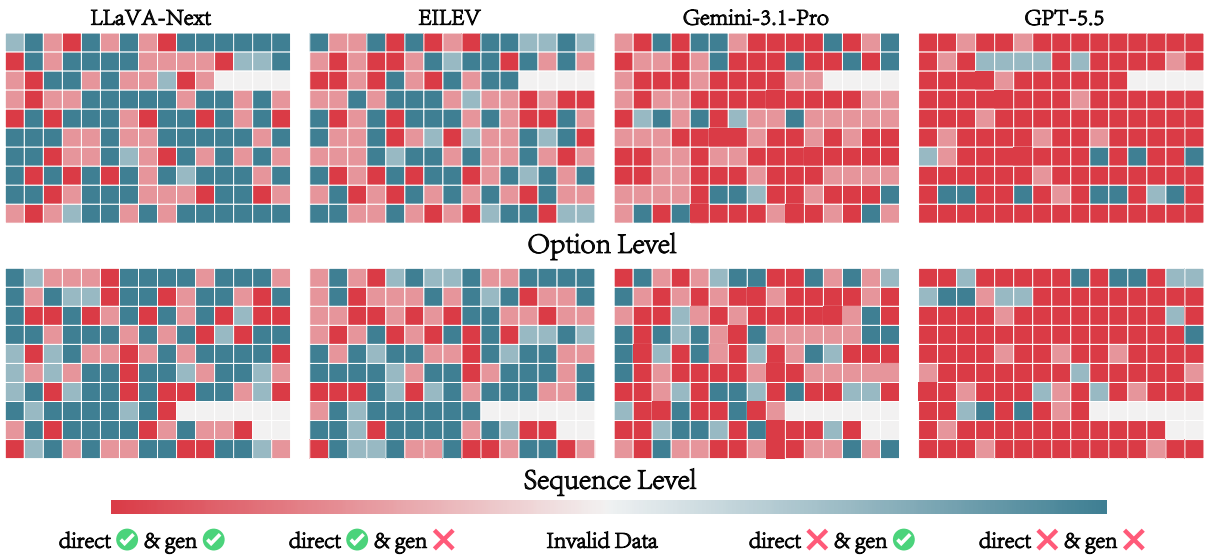


Figure 5: **Per-sample outcomes under *direct* and *gen*.** Each row is a distinct scene and each cell a test sample, colored by whether the model succeeds in both settings, in neither, or in only one. Two patterns recur: some scenes defeat every model, and the vision-based models succeed on largely disjoint samples across the two settings, a signature of scene-dependent rather than rule-based inference.

6 Conclusion

We have recast personalized embodied planning as a problem of pragmatic reasoning. PBP evaluates implicit semantic alignment rather than instruction following, and reveals a substantial gap between what agents can do once a preference is given and what they can recover on their own. InTU shows the gap to be narrowable: an agent that states the preference before acting aligns with human intent more reliably than one imitating demonstrations directly. Language, on this evidence, is a workable bridge from observed behavior to personalized action, and pragmatics, language, and embodied intelligence meet on ground worth further study.

Limitations The benchmark is synthetic, and no simulator reproduces the full complexity of real environments. Preference labels are human-defined and may pass over finer variation in how people actually behave. We also supply the relevant demonstrations as context, whereas a deployed agent would first have to retrieve them from long-term memory, a step this work does not address.

7 Conclusion

In this work, we reframed the problem of personalized embodied planning as a pragmatic reasoning task. We introduced PBP, a benchmark for evaluating implicit semantic alignment, and demonstrated that current agents exhibit a substantial gap between preference acquisition and preference-conditioned planning. Through experiments, we show that explicit verbalization of preferences enables agents to align with human intent more robustly than implicit visual imitation. We hope this work inspires further research into the intersection of pragmatics, language, and embodied intelligence.

Limitations Our benchmark relies on synthetic data, and simulator environments cannot fully capture real-world complexity. Human-defined preference labels may also overlook subtle variations in user behavior. In addition, we assume that relevant demonstrations are provided as context; real-world agents must additionally retrieve relevant experiences from long-term memory.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China (32595491,62376009), the Beijing Nova program, the NVIDIA Academic Grant Program using Spark and Thor, the State Key Lab of General AI at Peking University, the PKU-BingJi Joint Laboratory for Artificial Intelligence, the Wuhan Major Scientific and Technological Special Program (2025060902020304), the Hubei Embodied Intelligence Foundation Model Research and Development Program, and the National Comprehensive Experimental Base for Governance of Intelligent Society, Wuhan East Lake High-Tech Development Zone.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Baris Akgun, Maya Cakmak, Karl Jiang, and Andrea L Thomaz. 2012. Keyframe-based learning from demonstration: Method and evaluation. *International Journal of Social Robotics*, 4:343–355.

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

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. 2021. Vivit: A video vision transformer. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, and 1 others. 2023. Do as i can, not as i say: Grounding language in robotic affordances. In *Conference on Robot Learning (CoRL)*.

Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. 2019. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jae-Woo Choi, Youngwoo Yoon, Hyobin Ong, Jaehong Kim, and Minsu Jang. 2024. Lota-bench: Benchmarking language-oriented task planners for embodied agents. In *Proceedings of International Conference on Learning Representations (ICLR)*.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, and 1 others. 2023. Palm-e: An embodied multimodal language model. In *Proceedings of International Conference on Machine Learning (ICML)*.

Seymour Epstein. 1994. Integration of the cognitive and the psychodynamic unconscious. *American Psychologist*, 49:709.

Christine A Fawcett and Lori Markson. 2010. Children reason about shared preferences. *Developmental Psychology*, 46:299.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64:86–92.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, and 1 others. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*.

Herbert Paul Grice. 1975. Logic and conversation. *Syntax and Semantics*, 3:43–58.

Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, and 1 others. 2023. Inner monologue: Embodied reasoning through planning with language models. In *Conference on Robot Learning (CoRL)*.

Kaiwen Jiang, Annya Dahmani, Stephanie Stacy, Boxuan Jiang, Federico Rossano, Yixin Zhu, and Tao Gao. 2022. What is the point? a theory of mind model of relevance. In *Annual Meeting of the Cognitive Science Society (CogSci)*.

Ivan Kapelyukh and Edward Johns. 2022. My house, my rules: Learning tidying preferences with graph neural networks. In *Conference on Robot Learning (CoRL)*.

Taeyoon Kwon, Dongwook Choi, Hyojun Kim, Sunghwan Kim, Seungjun Moon, Beong-woo Kwak, Kuan-Hao Huang, and Jinyoung Yeo. 2026. Embodied agents meet personalization: Investigating challenges and solutions through the lens of memory utilization. In *Proceedings of International Conference on Learning Representations (ICLR)*.

Jeongeun Lee, Chanyoung Park, and Dongha Lee. 2026. Personalizing embodied multimodal large language model agents over long-term user interactions. *arXiv preprint arXiv:2605.26256*.

Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Michael Lingelbach, Jiankai Sun, and 1 others. 2023. Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation. In *Conference on Robot Learning (CoRL)*.

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

Sarah Lichtenstein and Paul Slovic. 2006. The construction of preference: An overview. *The Construction of Preference*, 1:1–40.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jiahong Liu, Zexuan Qiu, Zhongyang Li, Quanyu Dai, Wenhao Yu, Jieming Zhu, Minda Hu, Menglin Yang, Tat-Seng Chua, and Irwin King. 2025a. A survey of personalized large language models: Progress and future directions. *arXiv preprint arXiv:2502.11528*.

Runze Liu, Chenjia Bai, Jiafei Lyu, Shengjie Sun, Yali Du, and Xiu Li. 2025b. Vlp: Vision-language preference learning for embodied manipulation. In *Annual Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G Patil, Ion Stoica, and Joseph E Gonzalez. 2023. Memgpt: Towards llms as operating systems. *arXiv preprint arXiv:2310.08560*.

Shuwen Qiu, Sirui Xie, Lifeng Fan, Tao Gao, Jungseock Joo, Song-Chun Zhu, and Yixin Zhu. 2022. Emergent graphical conventions in a visual communication game. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*.

Gabriel Sarch, Zhaoyuan Fang, Adam W Harley, Paul Schydlo, Michael J Tarr, Saurabh Gupta, and Katerina Fragkiadaki. 2022. Tidee: Tidying up novel rooms using visuo-semantic commonsense priors. In *Proceedings of European Conference on Computer Vision (ECCV)*.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2020. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*.

Itamar Simonson. 2008. Will i like a "medium" pillow? another look at constructed and inherent preferences. *Journal of Consumer Psychology*, 18:155–169.

Paul Slovic. 1995. The construction of preference. *American Psychologist*, 50:364.

Ioan A Sucan, Mark Moll, and Lydia E Kavraki. 2012. The open motion planning library. *IEEE Robotics & Automation Magazine*, 19:72–82.

Andrew Szot, Alex Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Chaplot, Oleksandr Maksymets, and 1 others. 2021. Habitat 2.0: training home assistants to rearrange their habitat. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023a. Voyager: An open-ended embodied agent with large language models. In *NeurIPS 2023 Workshop IMOL*.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023b. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Yu Wang, Yifan Gao, Xiusi Chen, Haoming Jiang, Shiyang Li, Jingfeng Yang, Qingyu Yin, Zheng Li, Xian Li, Bing Yin, Jingbo Shang, and Julian Mcauley. 2024. MEMORYLLM: Towards self-updatable large language models. In *Proceedings of International Conference on Machine Learning (ICML)*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*.

Jimmy Wu, Rika Antonova, Adam Kan, Marion Lepert, Andy Zeng, Shuran Song, Jeannette Bohg, Szymon Rusinkiewicz, and Thomas Funkhouser. 2023. Tidybot: Personalized robot assistance with large language models. *Autonomous Robots*, 47:1087–1102.

Manjie Xu, Cheng Chen, Xin Jia, Jingyi Zhou, Yongji Wu, Zejian Wang, Chi Zhang, Kai Zuo, Yibo Chen, Xu Tang, and 1 others. 2025. Cross-scenario unified modeling of user interests at billion scale. *arXiv preprint arXiv:2510.14788*.

Manjie Xu, Guangyuan Jiang, Wei Liang, Chi Zhang, and Yixin Zhu. 2023. Active reasoning in an open-world environment. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*.

Manjie Xu, Isabella Yin, Xinyi Tu, Chi Zhang, and Yixin Zhu. 2026. Code over words: Overcoming semantic inertia via code-grounded reasoning. In *Findings of the Association for Computational Linguistics: ACL 2026*.

Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, Marziyeh Movahedi, Manling Li, and 1 others. 2025. Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. In *Proceedings of International Conference on Machine Learning (ICML)*.

Lance Ying, Xinyi Li, Shivam Aarya, Yizirui Fang, Yifan Yin, Jason Xinyu Liu, Stefanie Tellex, Joshua B Tenenbaum, and Tianmin Shu. 2026. Pragmatic embodied spoken instruction following in human-robot collaboration with theory of mind. In *International Conference on Robotics and Automation (ICRA)*.

Keunwoo Peter Yu, Zheyuan Zhang, Fengyuan Hu, and Joyce Chai. 2023. Efficient in-context learning in

vision-language models for egocentric videos. *arXiv preprint arXiv:2311.17041*.

Luyao Yuan, Xiaofeng Gao, Zilong Zheng, Mark Edmonds, Ying Nian Wu, Federico Rossano, Hongjing Lu, Yixin Zhu, and Song-Chun Zhu. 2022. In situ bidirectional human-robot value alignment. *Science robotics*.

Yuan Yuan, Huandong Wang, Jingtao Ding, Depeng Jin, and Yong Li. 2023. Learning to simulate daily activities via modeling dynamic human needs. In *Proceedings of the ACM Web Conference*.

Hongxin Zhang, Weihua Du, Jiaming Shan, Qinhong Zhou, Yilun Du, Joshua B Tenenbaum, Tianmin Shu, and Chuang Gan. 2024. Building cooperative embodied agents modularly with large language models. In *Proceedings of International Conference on Learning Representations (ICLR)*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, and 1 others. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. 2018. Dags with no tears: Continuous optimization for structure learning. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*.

Yixin Zhu, Tao Gao, Lifeng Fan, Siyuan Huang, Mark Edmonds, Hangxin Liu, Feng Gao, Chi Zhang, Siyuan Qi, Ying Nian Wu, and 1 others. 2020. Dark, beyond deep: A paradigm shift to cognitive ai with humanlike common sense. *Engineering*.

Filippo Ziliotto, Jelin Raphael Akkara, Alessandro Daniele, Lamberto Ballan, Luciano Serafini, and Tommaso Campari. 2026. PersONAL: Towards a comprehensive benchmark for personalized embodied agents. In *International Conference on Robotics and Automation (ICRA)*.

George Kingsley Zipf. 2016. *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio Books.

A Dataset Card

We follow the datasheet proposed in Gebru et al. (2021) for documenting our proposed PBP:

1. **Motivation**

   (a) **For what purpose was the dataset created?**
   The benchmark was created to evaluate existing learning agents on their ability to understand and adapt to various human preferences. Specifically, it aims to test the agents' proficiency in few-shot learning from demonstrations, where they must respond to ambiguous task instructions and formulate adaptive task plans based on limited examples of user preferences. The benchmark is designed to highlight the challenges and gaps in current AI systems' capabilities in planning activities and abstracting human preferences, ultimately driving advancements towards developing more intelligent and personalized embodied agents.

   (b) **Who created the dataset and on behalf of which entity?**
   N/A.

   (c) **Who funded the creation of the dataset?**
   N/A.

   (d) **Any other Comments?**
   None.

2. **Composition**

   (a) **What do the instances that comprise the dataset represent?**
   Each instance contains an egocentric video of an agent's activity, its bird's-eye-view map of the position of the agent, and a frame-level textual annotation of the current action, as shown in Fig. 3. Additionally, we provide a rendered third-person view of the entire process.

   (b) **How many instances are there in total?**
   15,000.

   (c) **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?**
   No. The dataset contains a set of demonstrations rendered within the simulator, The users can render more diverse instances if they want. We have provided the rendering instructions.

   (d) **What data does each instance consist of?**
   The instances that comprise the benchmark represent various types of human preferences applied to different tasks within a realistic embodied scene. Each instance is designed to challenge the learning agents to understand and adapt to these preferences based on a few demonstration examples, reflecting the diverse and hierarchical nature of user preferences in real-world scenarios. See above for data details.

   (e) **Is there a label or target associated with each instance?**
   Yes.

   (f) **Is any information missing from individual instances?**
   No.

   (g) **Are relationships between individual instances made explicit?**
   Yes.

   (h) **Are there recommended data splits?**
   No.

   (i) **Are there any errors, sources of noise, or redundancies in the dataset?**
   No.

   (j) **Is the dataset self-contained, or does it link to or otherwise rely on external resources (*e.g.*, websites, tweets, other datasets)?**
   Self-contained.

   (k) **Does the dataset contain data that might be considered confidential (*e.g.*, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)?**
   No.

   (l) **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?**
   No.

   (m) **Does the dataset relate to people?**
   No.

(n) **Does the dataset identify any subpopulations (*e.g.*, by age, gender)?**
No.

(o) **Is it possible to identify individuals (*i.e.*, one or more natural persons), either directly or indirectly (*i.e.*, in combination with other data) from the dataset?**
No.

(p) **Does the dataset contain data that might be considered sensitive in any way (*e.g.*, data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?**
No.

(q) **Any other comments?**
None.

3. **Collection Process**

(a) **How was the data associated with each instance acquired?**
We render PBP using NVIDIA's Omniverse and OmniGibson simulation environment (Li et al., 2023).

(b) **What mechanisms or procedures were used to collect the data (*e.g.*, hardware apparatus or sensor, manual human curation, software program, software API)?**
The data for each instance in the benchmark was acquired by sampling preferences from a predefined set and constructing tasks paired with a few demonstrations that shared high-level preferences but differed in specific objects and scenes. Each sampled preference was randomly assigned to one of the 50 scenes provided by OmniGibson, with relevant objects sampled within the scene. Egocentric observation and action sequences of an embodied agent were generated as the agent performed tasks guided by a rule-based planner using planning primitives like inverse kinematics for grasping and the A* algorithm for movement.

(c) **If the dataset is a sample from a larger set, what was the sampling strategy (*e.g.*, deterministic, probabilistic with specific sampling probabilities)?**
N/A.

(d) **Who was involved in the data collection process (*e.g.*, students, crowdworkers, contractors) and how were they compensated (*e.g.*, how much were crowdworkers paid)?**
N/A.

(e) **Over what timeframe was the data collected?**
N/A.

(f) **Were any ethical review processes conducted (*e.g.*, by an institutional review board)?**
The dataset raises no ethical concerns.

(g) **Does the dataset relate to people?**
No.

(h) **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (*e.g.*, websites)?**
N/A.

(i) **Were the individuals in question notified about the data collection?**
N/A.

(j) **Did the individuals in question consent to the collection and use of their data?**
N/A.

(k) **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?**
N/A.

(l) **Has an analysis of the potential impact of the dataset and its use on data subjects (*e.g.*, a data protection impact analysis) been conducted?**
Yes.

(m) **Any other comments?**
None.

4. **Preprocessing, Cleaning and Labeling**

(a) **Was any preprocessing/cleaning/labeling of the data done (*e.g.*, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal**

of instances, processing of missing values)?
N/A.

(b) **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (*e.g.*, to support unanticipated future uses)?**
N/A.

(c) **Is the software used to preprocess/clean/label the instances available?**
N/A.

(d) **Any other comments?**
None.

5. **Uses**

(a) **Has the dataset been used for any tasks already?**
No, the dataset is newly proposed by us.

(b) **Is there a repository that links to any or all papers or systems that use the dataset?**
No, the dataset is new.

(c) **What (other) tasks could the dataset be used for?**
This dataset could be used for research topics like embodied AI and human-computer interaction.

(d) **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?**
N/A.

(e) **Are there tasks for which the dataset should not be used?**
N/A.

(f) **Any other comments?**
None.

6. **Distribution**

(a) **Will the dataset be distributed to third parties outside of the entity (*e.g.*, company, institution, organization) on behalf of which the dataset was created?**
No before it is made public.

(b) **How will the dataset be distributed (*e.g.*, tarball on website, API, GitHub)?**
On our project website upon acceptance.

(c) **When will the dataset be distributed?**
Upon acceptance.

(d) **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?**
Under CC BY-NC [1] license.

(e) **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?**
No.

(f) **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?**
No.

(g) **Any other comments?**
None.

7. **Maintenance**

(a) **Who is supporting/hosting/maintaining the dataset?**
The authors.

(b) **How can the owner/curator/manager of the dataset be contacted (*e.g.*, email address)?**
N/A.

(c) **Is there an erratum?**
Future erratum will be released through the website.

(d) **Will the dataset be updated (*e.g.*, to correct labeling errors, add new instances, delete instances')?**
Yes.

(e) **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (*e.g.*, were individuals in question told that their data would be retained for a fixed period of time and then deleted)?**
N/A. The dataset does not relate to people.

(f) **Will older versions of the dataset continue to be supported/hosted/maintained?**
Yes.

(g) **If others want to extend/augment/build on/contribute to**

[1] https://creativecommons.org/licenses/by-nc/4.0/

the dataset, is there a mechanism for them to do so?
Yes. We will release the source code as well as a licence on our project website after acceptance.

(h) **Any other comments?**
None.

B Dataset Statistics

Simulation lengths in the dataset range from 1 to 5 minutes, depending on the tasks recorded. And the videos are recorded at 30 fps.

B.1 Preferences

See Table A1 for the preference statistics in PBP. See Fig. A1 for the hierarchical organization of user preferences.

Table A1: **Dataset Statistics in PBP.**

	Action	Option	Sequence
Preference Num	75	135	80
Episode Num	5000	5000	5000
Sub-task Num	1	2-3	2-3

B.2 Actions

See Table A2 for the action statistics in PBP. We implement 17 action primitives in PBP to assist with model planning and dataset rendering. These action primitives have parameters that simplify tasks and are considered the lowest-level actions. Each sub-task contains 8 to 20 such lowest-level actions. Generally, most of these actions consist of two parts: the robot movement part and the arm (gripper) execution part. For robot movement, we use the A* algorithm to find paths and avoid collisions. We build a connection map during scene initialization for navigation, taking the robot's width into consideration. For the arm (gripper) execution, we primarily use the Inverse Kinematics (IK) algorithm to compute arm movements. However, since IK cannot handle complex tasks, such as picking objects from the fridge, we also leverage the Open Motion Planning Library (OMPL) planner (Sucan et al., 2012) with forward planning to assist in planning the arm positions.

B.3 More Dataset Details and Discussion

Dataset production In summary, we follow the order of "sample preference - sample scene - sample objects to be manipulated - generate actions guided by a rule-based planner".

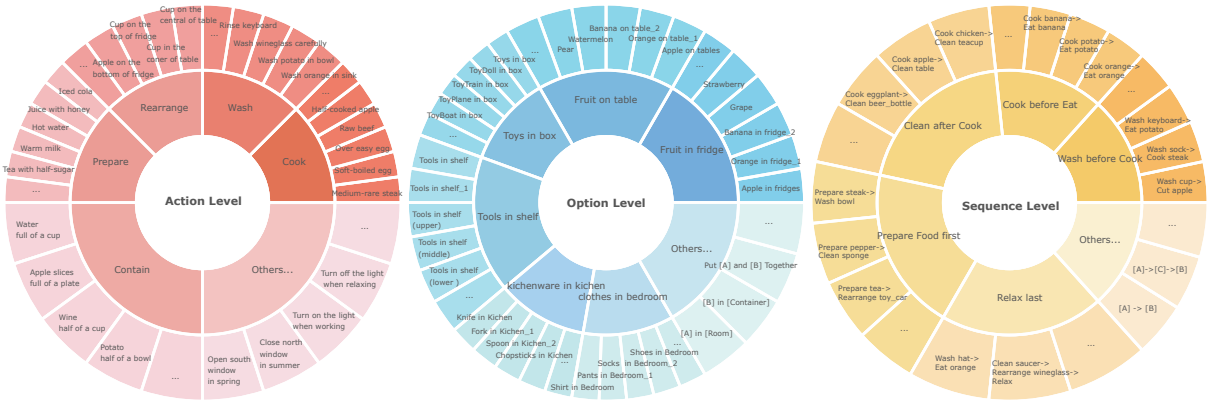
Length and FPS of the simulations The length of the simulations ranges from 1 to 5 minutes, depending on the tasks recorded. The videos are recorded at 30 fps.

Actions contained in each simulation The number of actions in simulations varies among different preference levels. There is 1 subtask for action-level, 2-3 subtasks for option-level, and 2-3 subtasks for sequence-level preferences. Each subtask contains 8-20 actions.

Scenes and rooms Each scene contains various types of rooms. The main differences between scenes are the type, number, and layout of both rooms and furniture. Additionally, each room may contain different objects and have unique layouts. Details of the scenes and rooms can be found in Omnigibson's official documentation (https://behavior.stanford.edu/omnigibson/), as we directly adopt these scenes from the open-sourced project.

290 preference types Considering that preferences in household activities are not only multi-dimensional but also hierarchical, we first define a hierarchy of preferences from the perspective of how things happen in a life scenario, that is, from each specific action to a sub-task consisting of several actions, and then to the sequence combining these sub-tasks. The next step is to expand each level with typical tasks and actions.

The egocentric view Collecting both egocentric observations and third-person views is feasible in PbP or similar environments built on simulators like iGibson. However, in real-world scenarios, it is generally easier to gather egocentric observations of human daily activities, as these can be efficiently captured through wearable devices. Additionally, there are numerous egocentric-view datasets available, such as Ego4D(Grauman et al., 2022), which further facilitate this approach. While third-person views can provide a different perspective, they often encounter issues such as occlusion. Although research based on third-person views is essential for applications involving real robots, focusing on egocentric views in the current work allows for a more straightforward exploration of preference learning and planning. Nevertheless, third-person view data can be obtained by integrating additional cameras, as outlined in our provided code.



(a) Action level (b) Option level (c) Sequence level

Figure A1: **Hierarchical organization of user preferences.** Our framework organizes preferences in a three-tiered structure, visualized through sunburst diagrams: (a) Action level captures fine-grained execution details within specific tasks, from quantity preferences in "Contain" (*e.g.*, "half a cup" *vs.* "full cup") to environmental controls (*e.g.*, lighting and window operations). (b) Option level represents spatial preferences for object categories, encoding both storage decisions (*e.g.*, table *vs.* fridge for fruits) and organizational choices (*e.g.*, shelf levels and boxes for tools/toys). (c) Sequence level defines temporal relationships between tasks, encompassing both basic preparation sequences (*e.g.*, "Prepare Food first") and conditional orderings (*e.g.*, "Clean after Cook," "[A]->[B]"). Each diagram's hierarchical structure branches from general categories to specific instances, revealing detailed preference patterns upon closer inspection. (Vector graphics; zoom in for details.)

Action ground truth In experiments involving vision input, we do not explicitly provide the action sequence of the user. In the symbolic-based experiment, we provide the action sequence to reduce the perception cost to concentrate more effectively on the inference and planning aspects of the study.

C Experiment Details

C.1 Archived Experimental Results

The following tables record experimental results from previous model iterations that have been superseded in the current version of this work. These results are retained for completeness and archival reference.

C.2 Case Study

We also provide a case with preference *Put fruit on the bed* in the following table Section C.2. We present a simplified version of the demonstrations, where all video outputs have been translated into symbol-based action sequences for ease of understanding. Video-based models such as LLaVA-Next and GPT-4V struggle with comprehending preferences and tend to replicate certain action patterns from the video demonstration, such as "move to" and "pick up". Llama3 demonstrates a partial understanding and execution of the preference. It correctly moves to each fruit (grape, banana), picks them up, and places them on the bed. However, it also interacts with the pencil and places it on the bed, which is not required by the preference. Ideally, the pencil should be placed on the table, similar to the pen. On the other hand, GPT-4.1(Symbol) accurately interacts with the grape and banana by moving to each fruit, picking it up, and placing it on the bed. This demonstrates a better understanding and execution of the preference compared to the other models.

D Baseline Details

We evaluate two families of models corresponding to two practically relevant settings. **Video-based models** operate directly on egocentric observations and therefore reflect the full challenge of extracting user preferences from raw behavioral evidence. **Symbol-based models** receive textual action traces and scene descriptions, removing most perceptual uncertainty and providing a controlled setting for diagnosing preference reasoning and downstream planning. Across both settings, we use a common two-stage evaluation protocol whenever supported by the model: preference prediction followed by preference-conditioned planning.

D.1 Video-Based Models

LLaVA-NeXT We evaluate LLaVA-NeXT-Video-7B-DPO (Liu et al., 2024), an open-source multimodal model with native video support. Following the official implementation, we use the `vicuna_v1` conversation template, sample 32 frames per video, and use a spatial pooling stride of 2. LLaVA-NeXT represents a practical open-source setting in which the agent must infer

Table A2: Action Primitives in PBP.

Action List	Explanation
Move_to_[]	Move to a specified location, or a specified room, or a specified object
Rotate_to_[]	Rotate to a specified orientation or a specified object
Pick_[]	Pick up an object using the gripper, *e.g.*, "Pick_apple"
Place_[]	Place an object at a location, *e.g.*, "Place_apple_on_table"
Fill_[]_with_[]	Fill a container with a substance, *e.g.*, "Fill_glass_with_water"
Pour_[]	Pour a substance from a container, *e.g.*, "Pour_milk"
Open_[]	Open an object, *e.g.*, "Open_door"
Close_[]	Close an object, *e.g.*, "Close_fridge"
Cut_[]	Cut an object, *e.g.*, "Cut_carrot"
Cook_[]	Cook an item, *e.g.*, "Cook_pasta"
Wash_[]	Wash an object, *e.g.*, "Wash_dishes"
Clean_[]	Clean a surface or object, *e.g.*, "Clean_counter"
Cover_[]	Cover an object, *e.g.*, "Cover_bowl"
Uncover_[]	Uncover an object, *e.g.*, "Uncover_bowl"
Toggle_on_[]	Turn on a device, *e.g.*, "Toggle_on_light"
Toggle_off_[]	Turn off a device, *e.g.*, "Toggle_off_stove"
Wait_[]	Wait some time

Table A3: **Archived Results — Levenshtein Distance (Superseded Models).** Results for GPT-4V (Video), DeepSeek-R1, GPT-4.1, and GPT-5.2 from Table 1. These have been superseded by updated experiments with Gemini 3 Pro, DeepSeek-v4-pro, Claude Opus 4.7, and GPT-5.5.

	DEPRECATED MODELS				
	GPT-4V	Llama3-8B	DeepSeek-R1	GPT-4.1	GPT-5.2
Option Level					
End-to-end	15.63±2.31	14.74±3.21	8.73±3.03	7.42±2.67	7.41±2.78
Two-stage	8.37±2.19	9.67±5.16	3.19±2.19	2.26±2.03	2.18±2.33
Second-stage (gt)	1.26±2.55	8.22±5.58	1.76±1.89	0.15±2.85	0.17±2.46
Sequence Level					
End-to-end	33.75±11.15	31.79±7.32	28.72±4.12	26.48±3.25	26.00±3.34
Two-stage	27.52±9.48	25.46±5.93	18.61±3.25	14.19±3.01	12.61±2.77
Second-stage (gt)	11.36±8.05	19.02±7.10	14.10±3.76	10.31±2.98	10.17±3.32
Overall					
End-to-end	24.69	23.26	18.72	16.95	16.70
Two-stage	17.94	17.56	10.90	8.22	7.40
Second-stage (gt)	6.31	13.62	7.93	5.23	5.17

Table A4: **Archived Results — Preference Inference Accuracy (Superseded Models).** Results for GPT-4V (Video), DeepSeek-R1, GPT-4.1, and GPT-5.2 from Table 2. These have been superseded by updated experiments with Gemini 3 Pro, DeepSeek-v4-pro, Claude Opus 4.7, and GPT-5.5.

	DEPRECATED MODELS				
	GPT-4V	Llama3-8B	DeepSeek-R1	GPT-4.1	GPT-5.2
Few-shot (With Context Demonstrations)					
Option Level	48.48	72.98	86.02	88.91	89.78
Sequence Level	37.50	67.18	71.21	70.28	73.34
Overall	42.99	70.08	78.62	79.60	81.56
Ablative (Zero-shot / No Context)					
Option Level	29.42	39.50	78.19	75.29	78.24
Sequence Level	0.00	6.25	15.29	14.11	18.44
Overall	14.71	22.88	46.74	44.70	48.34

Table A5: **Archived Results — Generalization (Superseded Models).** Results for GPT-4V, DeepSeek-R1, GPT-4.1, and GPT-5.2 from Table 3. These have been superseded by updated experiments with Gemini 3 Pro, DeepSeek-v4-pro, Claude Opus 4.7, and GPT-5.5.

	GPT-4V	DeepSeek-R1	GPT-4.1	GPT-5.2
Option	53.24	86.02	88.91	89.78
Option *gen*	48.48	84.98	86.57	86.57
Sequence	39.42	71.21	70.28	73.34
Sequence *gen*	37.50	70.16	69.28	69.28

Case	ID	Scene	Actions
Context	Demo 1	Beechwood	move to orange, pick up orange, move to bed with orange, place orange on bed, move to apple, pick up apple, move to bed with apple, place apple on bed, move to dvd, pick up dvd, move to table with dvd, place dvd on table.
	Demo 2	Rs	move to banana, pick up banana, move to bed with banana, place banana on bed, move to club sandwich, pick up club sandwich, move to countertop with club sandwich, place club sandwich on countertop, move to orange, pick up orange, move to countertop with orange, move to bed with orange, place orange on bed.
	Demo 3	Merom	move to pen, pick up pen, move to table with pen, place pen, move to strawberry, pick up strawberry, move with to bed with strawberry, place strawberry on bed, move to textbook, pick up textbook, move with textbook, move to bookshelf with textbook, place textbook in bookshelf.
Test Scene		Beechwood	grape, pencil, banana
Answer	LLaVA-Next	Beechwood	move to grape, move to bed with grape, move to pencil, pick up pencil, move to banana, pick up banana
	GPT-4V	Beechwood	move to grape, pick up grape, move to bed with grape, place grape on bed
	Llama3	Beechwood	move to grape, pick up grape, move to bed with grape, place grape on bed, move to pencil, pick up pencil, move to bed with pencil, place pencil on bed, move to banana, pick up banana, move to bed with banana, place banana on bed
	GPT-4.1(Symbol)	Beechwood	move to grape, pick up grape, move to bed with grape, place grape on bed, move to banana, pick up banana, move to bed with banana, place banana on bed

Table A6: Case Study with preference *Put fruit on the bed*.

preferences directly from temporally extended egocentric observations.

EILEV We additionally evaluate EILEV (Yu et al., 2023), a multimodal model specifically designed for few-shot in-context learning from embodied videos. Our implementation uses OPT-2.7B (Zhang et al., 2022) as the language backbone. Because EILEV is pretrained on Ego4D (Grauman et al., 2022), it provides a useful baseline for testing whether egocentric pretraining improves preference inference from behavioral demonstrations. Unlike the other models, EILEV expects interleaved video–text examples in a specific in-context format; we therefore preserve its native demonstration ordering while keeping the semantic content of the prompt unchanged.

GPT-4V and Gemini-3.1-Pro We use GPT-4V (Achiam et al., 2023) as an archived closed-source multimodal baseline and Gemini-3.1-Pro as the stronger model in the main experiments. For GPT-4V, we use the Azure OpenAI API with `gpt-4-turbo-2024-04-09` and a temperature of 0.05. Because of image-count limitations, each video is temporally subsampled to a compact set of frames. For Gemini-3.1-Pro, we use the Google GenAI API with model id `gemini-3.1-pro-preview`, also with temperature 0.05. We extract keyframes at 1 Hz and interleave them with text delimiters that preserve demonstration boundaries. These models approximate a practical deployment scenario in which a general-purpose multimodal foundation model observes user behavior directly and must recover preference-relevant semantics without access to symbolic action annotations.

D.2 Symbol-Based Models

To isolate preference reasoning from visual perception, we evaluate LLMs on symbolic versions of the same demonstrations. Each video is replaced by a textual scene description and action trace. This setting is not intended to represent a more realistic interface; instead, it serves as a diagnostic upper-bound setting in which the relevant behavioral events are explicitly observable.

Llama3-8B We evaluate Meta-Llama-3-8B-Instruct (Touvron et al., 2023) using the official implementation with temperature 0.05. Llama3-8B provides a relatively lightweight open-source language baseline for preference inference and planning from symbolic trajectories.

DeepSeek-R1 and GPT-4.1 DeepSeek-R1 and GPT-4.1 are retained as archived symbolic baselines. For DeepSeek-R1, we use a self-hosted 671B

model with temperature 0.05 and suppress the explicit reasoning trace using an empty <think> block, as this setting did not improve performance in our experiments. For GPT-4.1, we use `gpt-4.1-2025-04-14` with temperature 0.05. Their archived results are reported in Section C.1.

GPT-5.5, Claude-Opus-4.7, and DeepSeek-v4-pro The main symbolic experiments use GPT-5.5, Claude-Opus-4.7, and DeepSeek-v4-pro. GPT-5.5 is accessed through the OpenAI API, Claude-Opus-4.7 through the Anthropic API, and DeepSeek-v4-pro through a self-hosted deployment. All are evaluated with temperature 0.05 using the same symbolic-input protocol. Because these models receive ground-truth action traces rather than pixels, their performance primarily reflects preference inference and planning rather than visual recognition.

D.3 Prompting Protocol

We use two preference-inference prompt formats for different evaluation purposes.

Setting A: Query-Conditioned Preference Inference This is the main setting used throughout our experiments. The model receives a set of unlabeled behavioral demonstrations together with the current task context. The demonstrations may reflect different aspects of the user's behavior and need not correspond to a single identical preference. The model must identify which behavioral regularities are relevant to the current query and infer the preference that should guide the present task.

```
"Preference Inference"

You are a robot assistant that infers a user's
↪  preference from prior behavior.

All possible preferences are:
{ALL POSSIBLE PREFERENCES}

Here are several previous behavioral
↪  demonstrations:
[DEMO_1]
[DEMO_2]
[DEMO_3]

The current instruction is:
[INSTRUCTION]

The current scene is:
[SCENE_DESCRIPTION]

Based on the user's previous behavior, infer the
↪  preference
that is relevant to the current task.

Question:
What preference should be applied in the current
↪  situation?
Choose from the preferences listed above.
```

We use this prompt for the main personalized preference-acquisition setting, where the model must infer the latent preference relevant to the current task by jointly reasoning over unlabeled behavioral history, the instruction, and the current environment.

Setting B: Labeled Few-Shot Preference Recognition For completeness, we provide an alternative labeled few-shot prompt template for preference recognition under explicit in-context supervision. This protocol is not used for the main experimental results reported in this paper.

```
"Few-Shot Preference Recognition"

You are a robot assistant that recognizes user
↪  preferences from behavior.

All possible preferences are:
{ALL POSSIBLE PREFERENCES}

Here are several labeled behavioral examples:
[DEMO_1] The preference is [PREFERENCE_1]
[DEMO_2] The preference is [PREFERENCE_2]
[DEMO_3] The preference is [PREFERENCE_3]

Now observe the following unlabeled query
↪  behavior:
[TEST_CASE]

Question:
What preference is expressed by the query
↪  behavior?
Choose from the preferences listed above.
```

Stage Two: Preference-Conditioned Planning To evaluate whether explicit preference representations improve planning, we provide the inferred preference directly to the planner:

```
"Stage Two / Preference-Conditioned Planning"

You are a robot assistant that generates an
↪  action sequence
consistent with the user's preference.

The inferred user preference is:
[PREDICTED_PREFERENCE]

The current instruction is:
[INSTRUCTION]

The current scene is:
[SCENE_DESCRIPTION]

All possible actions are:
{ALL POSSIBLE ACTIONS}

Generate an action sequence that satisfies the
↪  instruction
while respecting the inferred preference:
```

For the oracle *Second-stage (gt)* condition, we use the same template but replace `[PREDICTED_PREFERENCE]` with `[GROUND_TRUTH_PREFERENCE]`. This controlled substitution allows us to measure how much planning error is attributable to preference acquisition rather than downstream action generation.

End-to-End Planning The end-to-end condition removes the explicit preference representation and asks the model to generate a plan directly from behavioral context:

```
"End-to-End Planning"

You are a robot assistant that generates an
↪  action sequence
based on the observed user behavior.

Behavioral context:
[DEMONSTRATIONS]

The current instruction is:
[INSTRUCTION]

The current scene is:
[SCENE_DESCRIPTION]

All possible actions are:
{ALL POSSIBLE ACTIONS}

Generate the action sequence:
```

The main evaluation therefore consists of three complementary components: *unlabeled preference induction*, *preference-conditioned planning*, and *end-to-end planning*. Together, they ask whether the model can infer a latent preference from behavior, whether it can plan effectively once that preference is explicit, and whether it can solve both problems jointly without an explicit intermediate representation.

D.4 Implementation Summary

For reproducibility, Table A7 summarizes the models used in the main experiments. Archived models are retained only for historical comparison.

Table A7: Model versions and access methods used in the main experiments.

Model	Version / ID	Access
ViViT	official Scenic implementation	local
LLaVA-NeXT	LLaVA-NeXT-Video-7B-DPO	local
EILEV	EILEV + OPT-2.7B	local
Gemini-3.1-Pro	`gemini-3.1-pro-preview`	Google API
Llama3-8B	Meta-Llama-3-8B-Instruct	local
DeepSeek-v4-pro	deepseek-v4-pro	self-hosted
Claude-Opus-4.7	`claude-opus-4-7`	Anthropic API
GPT-5.5	`gpt-5.5`	OpenAI API

E Explicit Preference Reasoning

To distinguish whether the improvement of InTU comes from explicitly reasoning about latent preferences or from exposing the preference as a separate intermediate representation, we introduce a Preference-aware End-to-End baseline. Unlike the vanilla End-to-End setting, which directly generates actions from behavioral demonstrations without being prompted to consider latent preferences, this baseline explicitly asks the model to infer the user's preference internally and directly produce the final action sequence without outputting the intermediate preference. We compare this setting with Two-stage Verbalization, where the inferred preference is explicitly verbalized before planning. This experiment allows us to separate the benefit of preference reasoning from the additional advantages of an interpretable and controllable intermediate representation.

Table A8: **Effect of Explicit Preference Reasoning.** We compare vanilla End-to-End planning, Preference-aware End-to-End reasoning (P-aware), and Two-stage Verbalization. Preference-aware End-to-End asks the model to infer preferences internally before generating actions, while Two-stage Verbalization explicitly exposes the inferred preference as an intermediate representation. Lower Levenshtein distance indicates better alignment.

Model	Vanilla ↓	P-aware ↓	Two-stage ↓
Option Level			
Llama3-8B	14.74 ± 3.21	9.24 ± 4.11	9.67 ± 5.16
Gemini-3.1-Pro	8.72 ± 2.26	4.56 ± 1.96	4.28 ± 1.76
Sequence Level			
Llama3-8B	31.79 ± 7.32	28.13 ± 5.34	25.46 ± 5.93
Gemini-3.1-Pro	24.51 ± 4.15	15.73 ± 3.55	12.10 ± 2.85
Overall			
Llama3-8B	23.26	18.68	17.56
Gemini-3.1-Pro	16.62	10.14	8.19