CoHabit Arena: Evaluating Personal Preference Realization in Long-Horizon Human-Agent CoHabitation

Anonymous authors
Paper under double-blind review

Abstract

Completing a household task does not ensure that an embodied agent respects a user's unstated preferences. We formalize Personal Preference Realization (PPR) as the ability to infer a user's latent preferences from interaction, retain them across tasks, and express them in subsequent embodied decisions. To evaluate this capability, we introduce CoHabit Arena, a diagnostic evaluation environment built on Habitat 3.0. Each session preserves a synthetic user persona and interaction history across tasks without providing its ground-truth preference profile to the agent. Executed trajectories are retrospectively matched to preference-sensitive choices and associated with persona-conditioned reactions, allowing preference evidence to be traced to prior choices and their consequences. CoHabit Arena comprises 576 sessions of 15 tasks across 12 scenes and 36 synthetic personas, totaling 8,640 task instances derived from 1,080 source tasks. Its evaluation separates preference inference (mental alignment), preference-sensitive executed choices (behavioral alignment), and task success. We include Embodied Theory-of-Mind (E-ToM), a diagnostic baseline that maintains a persistent belief over preference axes and uses it to select subsequent plans. Across two LLM backbones, E-ToM configurations have higher recorded mental-alignment scores and lower preference-estimation error than the listed base planners, without uniform task-success gains. Missing baseline regret and unresolved measurement issues limit behavioral comparisons. CoHabit Arena thus provides a controlled setting for studying cross-task preference acquisition and testing whether inferred preferences translate into embodied behavior beyond instruction following.

1 Introduction

Agent–human interaction has become a central testbed for embodied agents. Existing arenas study household instruction following, dialogue-based task completion, human–robot collaboration, planning, and personalization (Shridhar et al., 2020; Padmakumar et al., 2022; Puig et al., 2024; Chang et al., 2025; Kwon et al., 2025). They have pushed agents toward more interactive behavior, but their primary evaluative target remains task execution, coordination, or reuse of already available user information.

Long-horizon agent–human cohabitation raises a more personal demand. A household assistant may complete a command while violating an unstated preference: entering a bedroom without asking, taking a disruptive shortcut, or placing an object in an undesirable location. Figure 1 illustrates how earlier feedback could inform a later choice. The full profile is withheld, but later corrections may be explicit. We study a stable synthetic persona within a session, not context-free or permanent real-user preferences.

We formalize this demand as **Personal Preference Realization (PPR)**: inferring an individual user's preferences from interaction, preserving them across tasks, and expressing them in later embodied decisions. Preference learning and personalization already study parts of this process (Lambert et al., 2025; Salemi et al., 2024; Zhao et al., 2025; Patel et al., 2025; Liang et al., 2026). Our focus is the evaluation of a choice–consequence–reaction relation: what the agent chose, what happened in the

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.

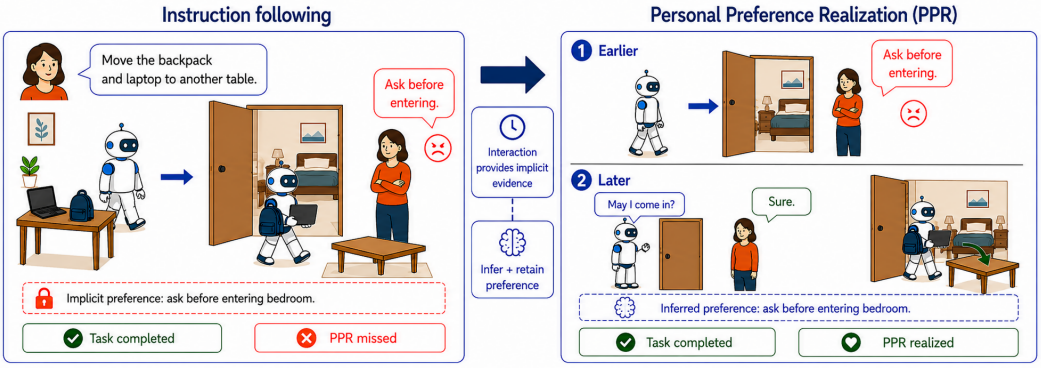


Figure 1: Personal Preference Realization (PPR) beyond instruction following: the agent infers the user's implicit preference from earlier interaction and realizes it in a later task.

environment, and how the same simulated user responded. PPR names this joint diagnostic target rather than a new general theory of preference learning.

Related work already addresses conversational preference following, embodied memory, demonstration-based personalization, and interactive feedback (Zhao et al., 2025; Kwon et al., 2025; Xu et al., 2026; Liang et al., 2026). CoHabit Arena emphasizes simulator-backed choice–outcome attribution across tasks. Table 1 is a selected qualitative comparison of design support, not an exhaustive novelty claim or evidence that every listed dimension is empirically validated. Appendix A discusses the closest settings.

We introduce **CoHabit Arena**, a diagnostic environment built on Habitat 3.0, with 12 scenes, 36 persistent synthetic personas, 576 evaluation sessions, and 8,640 task instances derived from 1,080 source tasks. Each session retains a persona and interaction history across tasks. An Embodied Theory-of-Mind (E-ToM) wrapper maintains an explicit preference state; its reported estimates improve, but information access, computation, and missing baseline behavioral scores limit the interpretation.

Table 1: Representative comparison. Y/P/N denote yes/partial/no. PPR requires implicit personal preference inference, cross-task realization, embodied consequences, and decision-attributable evidence.

Arena / Resource	Implicit personal pref.	Cross-task realization	Embodied consequence	Decision attribution
RewardBench (Lambert et al., 2025)	N	N	N	N
LaMP (Salemi et al., 2024)	P	P	N	N
FANToM (Kim et al., 2023)	P	N	N	N
TEACh (Padmakumar et al., 2022)	P	N	Y	N
PARTNR / Habitat 3.0 (Chang et al., 2025; Puig et al., 2024)	P	N	Y	N
MEMENTO (Kwon et al., 2025)	P	P	Y	P
CoHabit Arena	**Y**	**Y**	**Y**	**Y**

Our contributions are threefold: a PPR formulation separating inference, behavior, and task completion; a simulator-backed arena combining persistent personas, diagnostic choices, attribution windows, and metrics; and an initial study of recorded preference-state dynamics across planner families. The evidence does not establish general behavioral superiority or real-user generalization.

2 Personal Preference Realization

We model long-horizon agent–human cohabitation as a continuing household session between one agent and one user, rather than as independent task episodes. An evaluation instance is $\mathcal{I} = \langle s, h, \mathbf{c} \rangle$, where s is a physical scene, h is a persistent user persona, and $\mathbf{c} = (t_1, \ldots, t_L)$ is an ordered task chain. At turn i, the agent executes task t_i in scene s, producing a trajectory τ_i and receiving a user outcome y_i. The user's latent preference profile z_h is never revealed; the only available evidence before the next task is the interaction history $\mathcal{H}_{<i} = \{(t_j, \tau_j, y_j)\}_{j=1}^{i}$.

This setting turns preference learning into an evidence-attribution problem. A task may admit several feasible choices, but only some choices express the user's personal preference. We denote the relevant evidence as a choice–consequence–reaction relation (o_i, τ_i, y_i): the agent commits to a choice o_i, the environment realizes its consequence through τ_i, and the user outcome y_i constrains which latent profile z_h could explain the reaction. The agent must therefore use $\mathcal{H}_{\leq i}$ to form a user-specific estimate $\hat{z}_{i+1}$ and let that estimate influence later choices, rather than treating each outcome as isolated task feedback.

We define **Personal Preference Realization (PPR)** as inferring z_h from interaction, preserving the resulting user model across **c**, and realizing it through future choices, without requiring a particular internal representation. We evaluate high-level LLM planning in simulator-backed tasks, not a new perception or motor-control system. Physical outcomes ground the task; we do not claim equivalent state information is inexpressible in text.

3 CoHabit Arena: An Environment Simulator for Personal Preference Realization

CoHabit Arena combines physical scenes, persistent personas, and task chains into controlled interaction settings. Figure 2 summarizes construction. Algorithm 1 specifies the task-level order used to interpret the study: execute a task with the current state, match completed actions to diagnostic choices, associate feedback with those choices, update the external preference state, and use it on subsequent tasks. Retrospective processing cannot change an already completed trace. This distinguishes the proposed protocol from a demonstrated online human-intervention controller.

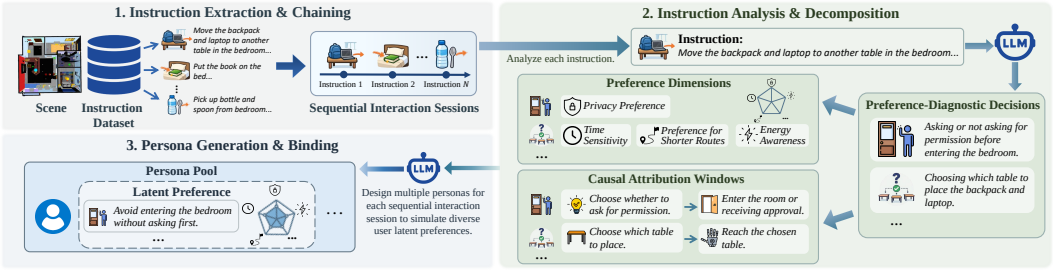


Figure 2: Construction pipeline of CoHabit Arena. The generator chains household tasks into cohabitation sessions, identifies preference-diagnostic choices and attribution windows, and binds persistent personas to create executable PPR evaluation instances.

3.1 Persistent Persona Modeling

PPR requires personal preferences to remain tied to the same user across a task chain. CoHabit Arena therefore pairs each session with a persistent, auditable synthetic persona whose diagnostic preference profile is controlled by a ground-truth vector. We construct these profiles using two components (detailed in Appendix B.3):

Preference Dimensions. We define a set of abstract axes $\mathcal{A} = \{a_1, \ldots, a_K\}$ that represent common value trade-offs (e.g., *Privacy* vs. *Helpfulness*). Here, K denotes the number of preference axes activated by the current scene or evaluation instance, rather than a fixed dimensionality shared by all scenes. Across CoHabit Arena, axes are drawn from a global allowlist, and profiles are compared after projection to the same active or common axis space.

Persona Pool. Based on these dimensions, we construct a *Persona Pool* $\mathcal{H}$. Each persona $h \in \mathcal{H}$ is associated with a ground-truth weight vector $z_h \in \Delta^K$ over the activated axes, where higher values indicate greater relative weight in the diagnostic utility. This is a controlled synthetic reference, not a population model. Mixed-type descriptive fields and utility weights are distinct objects; the exact encoding and axis scale are not recoverable from the summary statistics alone (Appendix B.3.3).

3.2 PREFERENCE EVIDENCE MODELING

PPR evidence appears when a task can be completed through multiple feasible choices, but those choices carry different personal implications. CoHabit Arena annotates such situations so that trajectories can be interpreted as preference evidence after execution, without changing the agent's normal task interface.

Preference-Diagnostic Choices. We define discrete decision points over otherwise continuous action trajectories, termed *Preference-Diagnostic Choices*. For each point p, CoHabit Arena records options $\mathcal{O}(p) = \{o_1, \ldots, o_m\}$ with different preference implications. Valid within-task alternatives must satisfy both physical feasibility and the explicit task constraints; soft preferences do not override the instructed goal. The agent executes the household task normally; after execution, the selected option $o_{\text{sel},p}$ is retrospectively inferred by matching the executed trajectory to the action bindings of options in $\mathcal{O}(p)$. Asking before entering a room illustrates a procedural preference within a fixed placement goal. The matching algorithm and acceptance thresholds require implementation records; Appendix E.2 identifies an example that does not establish valid task-equivalent alternatives.

Causal Attribution Window. To associate user outcomes with the choices that produce them, we define a *Causal Attribution Window* $W(p)$ for each decision point p. The window is specified as a logical boundary during arena construction and instantiated after execution: it opens at the planner step matched to $o_{\text{sel},p}$ and closes at the next matched decision point, a maximum-step boundary, or task termination. Reactions within $W(p)$ are assigned to p by this temporal rule. The term "causal" denotes the intended attribution relation, not identification of a causal effect from observational timing; delayed or overlapping consequences can violate the rule.

Embodied Outcome Signals. Within each instantiated attribution window, CoHabit Arena records *Embodied Outcome Signals* as standardized, persona-conditioned responses to the selected embodied option. Each signal specifies its type, severity, linked diagnostic decision, and affected preference axes, making the reaction term in choice–consequence–reaction evidence explicit and machine-readable. We model a hierarchy of signal types (Appendix B.1), from intervention-like response categories (e.g., *veto*, *takeover*) to implicit social cues (e.g., *negative sentiment*, *positive acceptance*). These are synthetic diagnostic signals, not samples from a real-human response distribution. An intervention label does not itself establish an online interruption. The wrapper uses attributed structured evidence; field-level parity with ordinary planner inputs is not documented, so this comparison cannot establish unassisted inference from natural-language feedback alone.

3.3 COHABITATION SESSION CONSTRUCTION

To evaluate cross-task realization, CoHabit Arena organizes tasks into *Cohabitation Sessions*. A session pairs a persistent persona with an ordered task chain $\mathbf{c} = (t_1, \ldots, t_L)$. Between tasks, the physical environment is reset for simulator stability, while the agent's epistemic context is preserved. Thus, when performing task t_i, the agent can use the interaction history from $(t_1, \ldots, t_{i-1})$, including past choices and user outcomes. Thus "long-horizon" refers to cross-task epistemic continuity, not uninterrupted physical-state evolution. The protocol enables testing later use of earlier evidence, but the present results do not include a matched history-swap or held-out behavior test. Details are in Appendix B.4.

3.4 MEASURING PERSONAL PREFERENCE REALIZATION

For each evaluation instance $\langle s, h, \mathbf{c} \rangle$, CoHabit Arena reports metrics that diagnose the PPR process: whether personal preferences are inferred, whether later choices realize those preferences, whether tasks remain executable, and how quickly preference evidence is acquired.

Mental Alignment. Mental Alignment measures agreement between a recorded preference estimate and the synthetic reference, not latent cognition directly. Let $z_h \in \Delta^K$ denote the ground-truth preference vector of persona h, and let z_{pred} denote the preference estimate recorded by the arena interface at a decision point. We compute Mental Alignment using cosine similarity:

$$\text{MA}(z_{\text{pred}}, z_h) = \frac{z_{\text{pred}} \cdot z_h}{\|z_{\text{pred}}\|_2 \, \|z_h\|_2}. \tag{1}$$

We report *Final MA* and *AUC MA* (Appendix B.2.1). A common readout procedure for methods without explicit vectors is not specified. Accordingly, prior-like readouts do not prove that contextual reasoning or preference-sensitive behavior is absent.

Behavioral Alignment. Behavioral Alignment measures whether executed choices realize the user's true preferences. Let $\mathcal{O}(p)$ denote the option set at decision point p, and let $x(o)$ denote the axis-consequence vector associated with option o, defined by the arena artifacts (Appendix B.2.2). We define the evaluator-only utility of option o under persona h as $U(o \mid z_h) = \sum_{k=1}^{K} z_h[k]\, x_k(o)$. We report *Decision Regret (D-Regret)* as

$$\text{D-Regret } (h) = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \left[\max_{o \in \mathcal{O}_p} U(o \mid z_h) - U(o_{\text{sel},p} \mid z_h) \right], \tag{2}$$

where $\mathcal{P}$ is the set of valid Preference-Diagnostic Choices in the session and $o_{\text{sel},p}$ is the option retrospectively matched to the executed trajectory at decision point p. The stated protocol excludes or separates low-confidence matches, but coverage counts are not reported. Regret is conditional on the scored choices and utility scale. Missing baseline logs prevent a uniform behavioral comparison; neither a missing entry nor an empty scored set means zero regret. $U(\cdot)$ is reserved for environment/evaluator use.

Instruction Following Ability. At the session level, we evaluate standard embodied instruction-following performance using *Success Rate (SR)* and *Success weighted by Path Length (SPL)* (Appendix B.2.3). Under the stated definitions and common aggregation, SPL cannot exceed SR. Some reported entries violate this bound; Section 5.2 and Appendix B.2.3 delimit their interpretation.

Convergence Cost. Ep@Thr is intended to count valid diagnostic decisions until error, concentration, and movement-from-prior criteria are jointly satisfied. It is a representation-dependent state diagnostic, not a general planner metric. Numerical thresholds, censoring, and index conventions are not specified sufficiently to reconcile the reported values with the definition, including the frozen-state ablation. We therefore draw no comparative convergence-speed conclusion from this column (Appendix B.2.4).

4 Preference-State Maintenance Baseline

CoHabit Arena evaluates whether evidence from earlier choices can inform later decisions. We use an Embodied Theory-of-Mind (E-ToM) wrapper as an axis-native reference baseline. It adds attributed evidence processing, an external preference state, and candidate selection to the underlying planner family. These additions need not preserve information access or inference cost. The experiment is therefore a comparison of implemented configurations, not an isolated causal test of memory architecture or a new general planning algorithm.

From arena evidence to preference observations. After task t is executed, CoHabit Arena provides the executed trajectory, the retrospectively matched preference-diagnostic choices, and the outcome signals attributed to those choices. The wrapper converts these arena outputs into a planner-independent evidence set

$$\mathcal{E}_t = \{(u_e, c_e)\}_{e=1}^{m_t}, \tag{3}$$

where m_t is the number of evidence records, $u_e \in \mathbb{R}^K_{\geq 0}$ gives axis support, and $c_e \in [0, 1]$ is an evidence weight. This is a structured observation channel. Withholding z_h alone does not rule out informative axis metadata. The exact raw-signal mapping and module permissions are not fully specified; we do not infer information parity or leakage-free implicit inference from this notation.

From observations to a persistent preference state. To retain preference evidence across tasks, the wrapper maintains a Dirichlet belief over the active preference axes, parameterized by $\alpha_t \in \mathbb{R}^K_{>0}$. The normalized mean

$$\hat{z}_{t,k} = \frac{\alpha_{t,k}}{\sum_j \alpha_{t,j}} \tag{4}$$

is the current preference estimate, while the concentration $\alpha_{0,t} = \sum_k \alpha_{t,k}$ is the total pseudo-count concentration. Given the evidence set from task t, we aggregate its weighted axis support as

$$\Delta_t = \sum_{(u_e, c_e) \in \mathcal{E}_t} c_e u_e. \tag{5}$$

The task-level belief is updated by

$$\alpha_t^{\text{end}} = \alpha_t + \mu\Delta_t, \tag{6}$$

where μ controls evidence sensitivity. The resulting state is then consolidated into cross-task memory:

$$\alpha_{t+1} = (1-\rho)\alpha_t + \rho\alpha_t^{\text{end}}, \tag{7}$$

Substitution gives $\alpha_{t+1} = \alpha_t + \rho\mu\Delta_t$. For fixed evidence, only the product $\rho\mu$ controls this update; the two steps do not establish separate fast–slow learning mechanisms or decay of old evidence. Relative mean weights can decrease as other coordinates increase. Concentration growth is not evidence of empirical calibration. The completed trace remains unchanged.

From preference state to later choices. Before the next task, the wrapper scores candidate plans under the maintained preference estimate:

$$\text{score}_{t+1}(q_i) = \sum_{k=1}^{K} \hat{z}_{t+1,k} f_k(q_i) - \lambda\,\text{cost}(q_i), \tag{8}$$

where $f_k(q_i)$ is a candidate feature score and λ weights execution cost. The feature estimator, cost scale, and candidate-generation budget are not specified by this equation. In particular, $f_k(q_i)$ must not be assumed identical to evaluator-only $x_k(o)$. Without an implementation audit and budget-matched controls, observed differences cannot be attributed solely to preference-state maintenance.

5 EXPERIMENTS

We examine the reported preference-state dynamics, task execution, and component comparisons in **CoHabit Arena**. The results are an initial diagnostic study of the listed configurations. Their interpretation is limited by incomplete behavioral coverage, undocumented cross-method readout and input parity, and measurement inconsistencies identified below. No additional trials, matched-memory controls, or real-user validation are inferred from these aggregate results.

5.1 EXPERIMENTAL SETUP

5.1.1 SETTINGS

All experiments are conducted in Habitat 3.0 (Puig et al., 2024) on CoHabit Arena. CoHabit Arena covers **12 diverse scenes**, **36 personas**, and **576 evaluation sessions**, constructed from **1,080 source household tasks** and expanded into **8,640 atomic task instances** across session-persona combinations. Each evaluation session (Section 3.3) consists of $L = 15$ sequential household tasks with preference-diagnostic choices. Sessions are the reported evaluation instances, not independent user samples; scenes, personas, and source tasks are reused. No held-out persona/scene generalization result is reported.

5.1.2 BASELINES

Planners. We report planners using the arena's metrics, with different cross-task preference representations. This does not establish identical observations or memory capacity. Specifically, we evaluate three high-level planners that span common LLM planning paradigms:

- **Reasoning-oriented planners.** *CoT* prompting (Wei et al., 2022) represents linear deliberation; *ReAct* (Yao et al., 2023b) represents interleaving reasoning with action execution; and *ToT* (Yao et al., 2023a) represents explicit branching search over alternatives. These planners primarily optimize task execution and do not maintain a persistent personal preference state.
- **ToM-style planner baseline.** We include an adaptation of *Hypothetical Minds* (HypMinds) (Cross et al., 2025), evaluated here without the E-ToM axis state. The original method maintains and refines hypotheses; absence of a Dirichlet state is not absence of memory. Results for this adaptation do not establish the limits of the full original method.
- **Reasoning-oriented planners + E-ToM.** We wrap each reasoning-oriented planner with our E-ToM-based techniques ("+ E-ToM"), retaining the planner family while adding evidence processing, persistent state, and candidate selection. No budget-matched episodic-memory control is included.

Models. To assess whether PPR gaps are visible across model families, we instantiate every planner configuration with two backbone LLMs: DeepSeek-V3.2 in *non-thinking* mode (DeepSeek-AI et al., 2025) and Gemini 3 Flash (Doshi, 2025). These backbones differ substantially in model family and deployment recipe.

5.1.3 METRICS

For each method–model configuration, we perform a single pass over the full CoHabit Arena evaluation set and report the metrics introduced in Section 3.4: Final Mental Alignment (Final MA), AUC Mental Alignment (AUC MA), Decision Regret (D-Regret), Success Rate (SR), Success weighted by Path Length (SPL), Ep@Thr, Final Prediction Error (Final L_2, i.e. L_2 at the end of the session) and the reported concentration $\alpha_0 = \sum_k \alpha_k$, initialized to 4.0. The tables retain their reported means and 95% intervals. The resampling procedure, effective sample counts, repeated seeds, and handling of early terminations are not specified, so these intervals do not establish paired significance or population-level generalization.

5.2 MAIN RESULTS

Table 2: Main results on CoHabit Arena (mean $\pm$ 95% CI over cohabitation sessions). "—" denotes unavailable matched-option logs for D-Regret, not zero regret; regret comparisons use only reported entries.

Methods	Final MA ↑	AUC MA ↑	D-Regret ↓	SR ↑	SPL ↑	Ep@Thr ↓	Final L_2 ↓	α_0 ↑
DeepSeek-V3.2								
CoT	0.832 ± 0.027	0.832	—	**0.396 ± 0.048**	**0.329 ± 0.044**	16.3	0.331 ± 0.038	4.0
CoT + E-ToM	0.884 ± 0.026	0.868	4.33 ± 1.16	0.373 ± 0.056	0.311 ± 0.052	18.4	0.277 ± 0.043	12.8 ± 1.3
ToT	0.781 ± 0.038	0.822	—	0.108 ± 0.095	0.134 ± 0.112	36.4	0.400 ± 0.050	4.0
ToT + E-ToM	**0.890 ± 0.072**	**0.885**	3.56 ± 1.53	0.068 ± 0.069	0.072 ± 0.072	12.5	**0.266 ± 0.137**	**15.4 ± 1.3**
ReAct	0.832 ± 0.027	0.832	—	0.378 ± 0.037	0.312 ± 0.039	16.3	0.331 ± 0.038	4.0
ReAct + E-ToM	0.881 ± 0.022	0.866	**3.16 ± 0.61**	0.391 ± 0.041	0.326 ± 0.037	17.4	0.281 ± 0.036	13.2 ± 1.0
Hyp-Minds	0.791 ± 0.058	0.820	—	0.109 ± 0.081	0.116 ± 0.097	**7.5**	0.387 ± 0.076	4.0
Gemini 3 Flash								
CoT	0.832 ± 0.027	0.832	—	0.274 ± 0.045	0.251 ± 0.041	16.3	0.331 ± 0.038	4.0
CoT + E-ToM	0.893 ± 0.023	0.876	3.55 ± 0.79	0.269 ± 0.049	0.243 ± 0.047	8.4	0.267 ± 0.038	17.1 ± 0.9
ToT	0.833 ± 0.028	0.833	—	0.053 ± 0.017	0.050 ± 0.018	27.1	0.330 ± 0.039	4.0
ToT + E-ToM	**0.905 ± 0.025**	**0.882**	**2.74 ± 1.09**	0.112 ± 0.031	0.096 ± 0.029	11.7	**0.254 ± 0.047**	**18.1 ± 1.0**
ReAct	0.832 ± 0.027	0.832	—	0.279 ± 0.043	0.251 ± 0.041	16.3	0.331 ± 0.038	4.0
ReAct + E-ToM	0.895 ± 0.022	0.878	3.85 ± 0.93	0.265 ± 0.048	0.239 ± 0.045	**7.8**	0.265 ± 0.037	17.1 ± 1.0
Hyp-Minds	0.827 ± 0.035	0.828	—	**0.319 ± 0.061**	**0.282 ± 0.057**	27.1	0.338 ± 0.049	4.0

Table 2 preserves the reported measurements. The following qualifications apply to its values and to the corresponding plots, including highlighted entries.

Measurement consistency. For the same samples and aggregation, the standard definitions imply $0 \leq \mathrm{SPL} \leq \mathrm{SR} \leq 1$. The DeepSeek ToT, ToT+E-ToM, and Hyp-Minds pairs $(\mathrm{SR}, \mathrm{SPL}) = (0.108, 0.134)$, $(0.068, 0.072)$, and $(0.109, 0.116)$ violate this inequality. The mismatch cannot be reconciled from the aggregate table; those pairs are not validated standard SR/SPL results. Reported Ep@Thr values also lack a reconciled relationship to the convergence definition. We retain the numbers but exclude these inconsistent measurements from performance or adaptation-speed claims.

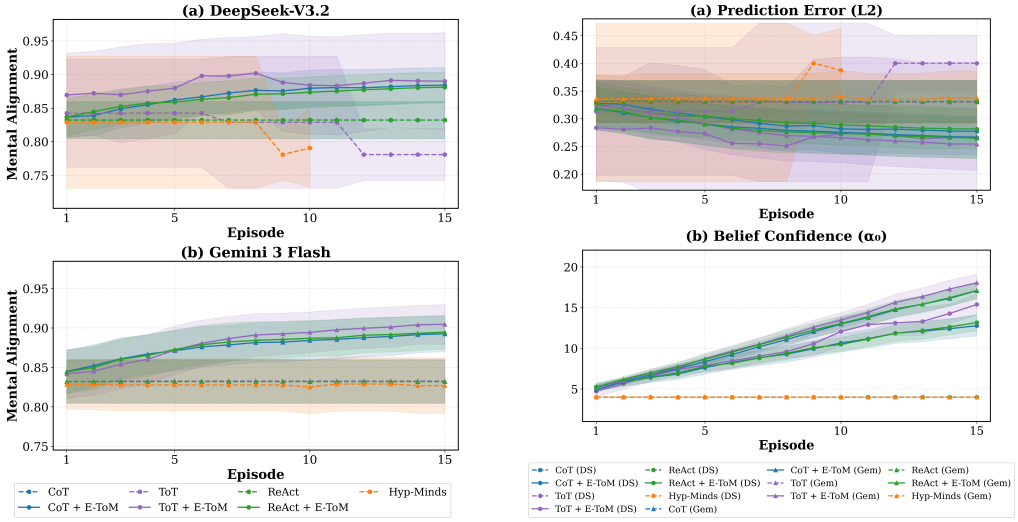
Recorded estimates and task execution are different measurements. Vanilla CoT on Gemini reports SR $= 0.274$, SPL $= 0.251$, Final MA $= 0.832$, and Final $L_2 = 0.331$. These quantities concern different outputs. A default or unchanged vector cannot reveal all information in the planner's context. The comparison motivates separate measurements, but does not quantify a correlation between task success and personal alignment or prove that standard planners have no preference knowledge.

E-ToM configurations report more accurate state estimates. The listed +E-ToM variants have Final MA 0.881–0.905 and Final L_2 0.254–0.281; Hyp-Minds reports 0.791–0.827 and 0.338–0.387, respectively. These are descriptive comparisons of recorded readouts, not representation-independent capability rankings. A common probe and same-information persistent-memory baseline are needed to isolate the source of the difference.

Behavioral realization is not established by state accuracy alone. Gemini ToT+E-ToM reports Final MA changing from 0.833 to 0.905, AUC MA from 0.833 to 0.882, and Final L_2 from 0.330 to 0.254. Its D-Regret is 2.74, but ordinary planners have no reported matched-option scores and utility scale and coverage are not characterized. We therefore do not describe this value as absolutely low or as evidence of behavioral improvement over those baselines.

Task-success changes are not uniform. Hyp-Minds has the highest reported Gemini SR (0.319), whereas +E-ToM has higher recorded MA. Such differences do not by themselves establish an acceptable personalization–execution trade-off. Preference gains require actual behavior evaluation under the same task constraints; a failure or an unscored trajectory cannot be credited as successful preference realization.

Temporal plots describe recorded state dynamics. Figure 3 shows increasing MA, decreasing error, and growing concentration. Its right panel plots error and concentration, not validated Ep@Thr; neither curve smoothness nor count growth establishes calibrated confidence.



(a) Mental Alignment over the sequential interaction session (shaded: 95% CI over sequential interaction session).

(b) Convergence cost over episodes (shaded: 95% CI over sequential interaction session). We denote Deepseek-V3.2 as "DS" and Gemini 3 Flash as "Gem".

Figure 3: Temporal PPR dynamics across 15-task cohabitation sessions. Left: methods with explicit preference-state maintenance improve personal preference inference over time. Right: preference-state error decreases and confidence increases across tasks. DS denotes DeepSeek-V3.2 and Gem denotes Gemini 3 Flash.

5.3 Ablation Study

Table 3: Ablation results (mean ± 95% CI over sequential interaction session). "—" denotes unavailable matched-option logs for D-Regret and is not interpreted as zero regret. We report $\Delta L_2 = L_2^T - L_2^0$ for convergence.

Variant	Final MA ↑	AUC MA ↑	D-Regret ↓	SR ↑	SPL ↑	Ep@Thr ↓	ΔL_2 ↓	α_0 ↑
E-ToM	0.882 ± 0.022	**0.868** ± 0.024	3.19 ± 0.67	0.382	0.316	3.1 ± 1.6	-0.043	13.0 ± 1.0
E-ToM w/o Retrospective Decision Mining	0.859 ± 0.025	0.852 ± 0.026	—	0.382	0.315	2.4 ± 2.2	-0.020	**20.0** ± 0.6
E-ToM w/o Window-Bound Attribution	**0.885** ± 0.020	0.868 ± 0.023	**2.88** ± 0.65	0.391	0.315	5.8 ± 3.2	-0.045	13.9 ± 1.0
E-ToM w/o Evidence Quantization	0.885 ± 0.021	0.867 ± 0.023	3.12 ± 0.62	**0.398**	**0.323**	4.3 ± 1.9	**-0.051**	5.0 ± 0.1
E-ToM w/o Belief Update	0.832 ± 0.027	0.832 ± 0.027	3.24 ± 0.67	0.387	0.314	**1.0** ± 0.0	0.000	4.0 ± 0.0
E-ToM w/o Outer-Loop Textual Refinement	0.867 ± 0.024	0.853 ± 0.026	3.13 ± 0.63	0.387	0.315	4.8 ± 3.0	-0.030	7.2 ± 0.4

Table 3 reports ReAct+E-ToM/DeepSeek-V3.2 component comparisons. Its full row differs from Table 2 (SR/SPL 0.382/0.316 versus 0.391/0.326; Ep@Thr 3.1 versus 17.4). The recorded

summaries do not establish whether runs, samples, or definitions differ. We do not pool the tables or interpret the Ep@Thr differences as validated adaptation costs.

Frozen-state negative control. Without belief updates, Final MA is 0.832, $\Delta L_2 = 0.000$, and $\alpha_0 = 4.0$. This checks that the explicit state is frozen, not that every form of preference inference requires E-ToM. Ep@Thr $= 1.0$ is incompatible with positive movement-from-prior under the stated definition and is not evidence of immediate convergence.

Retrospective matching and state concentration. Without retrospective mining, Final MA changes from 0.882 to 0.859 and concentration from 13.0 to 20.0. Higher concentration can coexist with a less accurate mean; this is not a calibration test. Missing D-Regret also prevents a behavioral comparison for this ablation.

Evidence mapping is not uniformly beneficial. Without quantization, concentration changes from 13.0 to 5.0, but Final MA is 0.885 and $\Delta L_2 = -0.051$, compared with 0.882 and -0.043 for the full row. The reported point estimates do not establish that quantization is necessary for accurate inference or that larger concentration is better.

Attribution effects are metric-dependent. Without window-bound attribution, Final MA is 0.885 versus 0.882 and reported D-Regret is 2.88 versus 3.19. These point estimates do not show uniform superiority of the full pipeline. The unreconciled Ep@Thr values cannot establish faster acquisition from the window.

Refinement is an incompletely specified component. The named outer-loop refinement ablation reports Final MA 0.867 versus 0.882 and concentration 7.2 versus 13.0. The refinement operator is not defined in Algorithm 1, and its objective, schedule, and additional calls are not documented. This row alone does not establish a specific stabilization mechanism.

5.4 Temporal Personal Preference Realization Dynamics

We analyze the temporal evolution of preference-state estimates across 15-task cohabitation sessions. Figure 3b illustrates Prediction Error (L_2) and Preference-State Confidence (α_0) over time. Ablation-level temporal curves are provided in Appendix D.4.

Recorded error decreases. The +E-ToM curves generally move from $L_2 \approx 0.33$–0.40 to 0.25–0.28. This describes the external estimate, not an independently tested change in future behavior.

Concentration is accumulated mass. Non-E-ToM values remain near 4.0, while Gemini +E-ToM reaches about 17–18 and DeepSeek about 13–15. Non-negative updates produce such growth mechanically; it does not rank reasoning quality or calibration.

Identical arenas need not yield identical evidence. Different choices can induce different reactions even within the same constructed arena. Comparing extraction on identical traces would isolate evidence processing, while end-to-end interaction measures the combined effects of planning and feedback. The present aggregate curves do not make this separation. Without matched traces, inference budgets, and repeated-run statistics, smoother curves cannot be attributed to inherently cleaner evidence processing by a backbone.

6 Conclusion

We presented **CoHabit Arena**, a simulator-backed diagnostic environment for **Personal Preference Realization (PPR)** across household task sequences. The reported E-ToM configurations exhibit more accurate recorded preference estimates without uniform task-success gains.

7 Limitations

CoHabit Arena is a controlled synthetic diagnostic, not a real-user or held-out generalization study. Missing baseline behavioral coverage, unresolved metric inconsistencies, and incomplete information, readout, and budget controls prevent stronger comparative conclusions. Recorded state improvements do not establish behavioral superiority or calibrated belief learning. Appendix F details these limitations and future work.

AI USE STATEMENT

Use in the research system. Large language models are used as the evaluated planning backbones and within the preference-state maintenance baseline described in Sections 4–5 and Appendix C. The evaluated backbones are DeepSeek-V3.2 in non-thinking mode and Gemini 3 Flash.

Assistance in analysis and manuscript preparation. Generative AI assistance included literature retrieval and discovery; feedback on the methodology and experimental interpretation; analysis of the stated equations and reported results; drafting and revising the abstract and explanatory passages; language editing; reference checking and formatting; and LaTeX preparation. ChatGPT supported these activities, and Google's Paper Assistant Tool provided automated manuscript feedback. These manuscript revisions did not add experimental runs or replace reported measurements.

Verification and responsibility. Bibliographic corrections were checked against primary publication records or explicitly identified preprint versions. The revised LaTeX source was compiled and the resulting PDF inspected. These checks concern the manuscript and bibliography, not independent reproduction of the experiments. AI-generated suggestions are not treated as empirical validation. The authors take responsibility for the final content of this work, including AI-assisted text, references, claims, and research artifacts.

ETHICS STATEMENT

CoHabit Arena is intended as a controlled evaluation resource for studying Personal Preference Realization (PPR) in simulated long-horizon human–agent cohabitation. Its positive impact is to make preference-related failures more visible: rather than measuring only task completion, the arena tests whether agents infer, retain, and realize user-specific preferences across later embodied decisions. The main risks are misuse and over-interpretation: PPR-style systems could be used to infer sensitive personal tendencies, over-personalize behavior without consent, or justify deployment in private household spaces based only on simulation performance. The study uses synthetic personas and simulator-generated interactions rather than real private user logs. Its intended information boundary separates environment/evaluator ground truth from agent inputs, but the present report does not establish a field-level access audit. We describe validity criteria and the limitations of the measurements rather than treating synthetic scores as a deployment guarantee. Personalization must not override explicit task constraints, consent, or safety requirements. Future human-subject or real-world extensions should require informed consent, privacy-preserving data handling, institutional review where applicable, and mechanisms for users to inspect, correct, or restrict learned preference models.

REFERENCES

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional AI: Harmlessness from AI Feedback. *arXiv preprint arXiv:2212.08073*, 2022. URL https://arxiv.org/abs/2212.08073.

Chris L. Baker, Rebecca Saxe, and Joshua B. Tenenbaum. Action Understanding as Inverse Planning. *Cognition*, 113(3):329–349, 2009. doi: 10.1016/j.cognition.2009.07.005. URL https://dspace.mit.edu/entities/publication/2e23673d-f720-4720-98bc-6db894ba27e1.

Chris L Baker, Julian Jara-Ettinger, Rebecca Saxe, and Joshua B Tenenbaum. Rational Quantitative Attribution of Beliefs, Desires and Percepts in Human Mentalizing. *Nature Human Behaviour*, 1: 0064, 2017. doi: 10.1038/s41562-017-0064. URL https://www.nature.com/articles/s41562-017-0064.

Daniel Brown, Wonjoon Goo, Prabhat Nagarajan, and Scott Niekum. Extrapolating Beyond Suboptimal Demonstrations via Inverse Reinforcement Learning from Observations. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 783–792. PMLR, 2019. URL https://proceedings.mlr.press/v97/brown19a.html.

Chunkit Chan, Cheng Jiayang, Yauwai Yim, Zheye Deng, Wei Fan, Haoran Li, Xin Liu, Hongming Zhang, Weiqi Wang, and Yangqiu Song. NegotiationToM: A Benchmark for Stress-testing Machine Theory of Mind on Negotiation Surrounding. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4211–4241. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.244. URL https://aclanthology.org/2024.findings-emnlp.244/.

Matthew Chang, Gunjan Chhablani, Alexander Clegg, Mikael Dallaire Cote, Ruta Desai, Michal Hlavac, Vladimir Karashchuk, Jacob Krantz, Roozbeh Mottaghi, Priyam Parashar, Siddharth Patki, Ishita Prasad, Xavier Puig, Akshara Rai, Ram Ramrakhya, Daniel Tran, Joanne Truong, John Turner, Eric Undersander, and Tsung-Yen Yang. PARTNR: A Benchmark for Planning and Reasoning in Embodied Multi-agent Tasks. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/a3cf318fbeec1126da21e9185ae9908c-Abstract-Conference.html.

Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep Reinforcement Learning from Human Preferences. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html.

Logan Cross, Violet Xiang, Agam Bhatia, Daniel Yamins, and Nick Haber. Hypothetical Minds: Scaffolding Theory of Mind for Multi-Agent Tasks with Large Language Models. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/12f483f624b378f9f3058d8ecd3c7ff5-Abstract-Conference.html.

DeepSeek-AI et al. DeepSeek-V3.2: Pushing the Frontier of Open Large Language Models. *arXiv preprint arXiv:2512.02556*, 2025. URL https://arxiv.org/abs/2512.02556.

Matt Deitke, Winson Han, Alvaro Herrasti, Aniruddha Kembhavi, Eric Kolve, Roozbeh Mottaghi, Jordi Salvador, Dustin Schwenk, Eli VanderBilt, Matthew Wallingford, Luca Weihs, Mark Yatskar, and Ali Farhadi. RoboTHOR: An Open Simulation-to-Real Embodied AI Platform. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Deitke_RoboTHOR_An_Open_Simulation-to-Real_Embodied_AI_Platform_CVPR_2020_paper.html.

Tulsee Doshi. Gemini 3 Flash: Frontier Intelligence Built for Speed. Google Blog, December 2025. URL https://blog.google/products-and-platforms/products/gemini/gemini-3-flash/. Published December 17, 2025; accessed September 23, 2026.

Jiayuan Gu, Fanbo Xiang, Xuanlin Li, Zhan Ling, Xiqiang Liu, Tongzhou Mu, Yihe Tang, Stone Tao, Xinyue Wei, Yunchao Yao, Xiaodi Yuan, Pengwei Xie, Zhiao Huang, Rui Chen, and Hao Su. ManiSkill2: A Unified Benchmark for Generalizable Manipulation Skills. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2302.04659.

Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. Cooperative Inverse Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/c3395dd46c34fa7fd8d729d8cf88b7a8-Abstract.html.

Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart Russell, and Anca Dragan. Inverse Reward Design. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/32fdab6559cdfa4f167f8c31b9199643-Abstract.html.

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

Brian Ichter, Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, Dmitry Kalashnikov, Sergey Levine, Yao Lu, Carolina Parada, Kanishka Rao, Pierre Sermanet, Alexander T. Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Mengyuan Yan, Noah Brown, Michael Ahn, Omar Cortes, Nicolas Sievers, Clayton Tan, Sichun Xu, Diego Reyes, Jarek Rettinghouse, Jornell Quiambao, Peter Pastor, Linda Luu, Kuang-Huei Lee, Yuheng Kuang, Sally Jesmonth, Nikhil J. Joshi, Kyle Jeffrey, Rosario Jauregui Ruano, Jasmine Hsu, Keerthana Gopalakrishnan, Byron David, Andy Zeng, and Chuyuan Kelly Fu. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. In *Proceedings of The 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pp. 287–318. PMLR, 2023. URL https://proceedings.mlr.press/v205/ichter23a.html.

Julian Jara-Ettinger, Hyowon Gweon, Laura E Schulz, and Joshua B Tenenbaum. The Naïve Utility Calculus: Computational Principles Underlying Commonsense Psychology. *Trends in Cognitive Sciences*, 20(8):589–604, 2016. doi: 10.1016/j.tics.2016.05.011. URL https://www.sciencedirect.com/science/article/pii/S1364661316300535.

Hong Jun Jeon, Smitha Milli, and Anca Dragan. Reward-rational (implicit) choice: A unifying formalism for reward learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 4415–4426, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/2f10c1578a0706e06b6d7db6f0b4a6af-Abstract.html.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Jiayi Zhou, Borong Zhang, Donghai Hong, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Lukas Vierling, Zhaowei Zhang, Fanzhi Zeng, Juntao Dai, Xuehai Pan, Hua Xu, Aidan O'Gara, Kwan Ng, Brian Tse, Jie Fu, Stephen McAleer, Yanfeng Wang, Mingchuan Yang, Yunhuai Liu, Yizhou Wang, Song-Chun Zhu, Yike Guo, Yaodong Yang, and Wen Gao. AI Alignment: A Contemporary Survey. *ACM Computing Surveys*, 58(5): 132:1–132:38, 2025. doi: 10.1145/3770749. URL https://doi.org/10.1145/3770749. Published online November 2025; issue date April 2026.

Yash Kant, Arun Ramachandran, Sriram Yenamandra, Igor Gilitschenski, Dhruv Batra, Andrew Szot, and Harsh Agrawal. Housekeep: Tidying Virtual Households using Commonsense Reasoning. In *Computer Vision – ECCV 2022*, volume 13699 of *Lecture Notes in Computer Science*, pp. 355–373. Springer, 2022. doi: 10.1007/978-3-031-19842-7_21. URL https://link.springer.com/chapter/10.1007/978-3-031-19842-7_21.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Le Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. FANToM: A Benchmark for Stress-testing Machine Theory of Mind in Interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14397–14413. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.890. URL https://aclanthology.org/2023.emnlp-main.890/.

Eric Kolve, Roozbeh Mottaghi, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. AI2-THOR: An Interactive 3D Environment for Visual AI. *arXiv preprint arXiv:1712.05474*, 2017. URL https://arxiv.org/abs/1712.05474v1. Version 1, December 2017.

Michal Kosinski. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45):e2405460121, 2024. doi: 10.1073/pnas.2405460121. URL https://doi.org/10.1073/pnas.2405460121.

Taeyoon Kwon, Dongwook Choi, Sunghwan Kim, Hyojun Kim, Seungjun Moon, Beong-woo Kwak, Kuan-Hao Huang, and Jinyoung Yeo. Embodied Agents Meet Personalization: Exploring Memory Utilization for Personalized Assistance. *arXiv preprint arXiv:2505.16348*, 2025. URL https://arxiv.org/abs/2505.16348v1. Version 1, May 2025.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. RewardBench: Evaluating Reward Models for Language Modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1755–1797. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-naacl.96. URL https://aclanthology.org/2025.findings-naacl.96/.

Chengshu Li, Fei Xia, Roberto Martín-Martín, Michael Lingelbach, Sanjana Srivastava, Bokui Shen, Kent Elliott Vainio, Cem Gokmen, Gokul Dharan, Tanish Jain, Andrey Kurenkov, Karen Liu, Hyowon Gweon, Jiajun Wu, Li Fei-Fei, and Silvio Savarese. iGibson 2.0: Object-Centric Simulation for Robot Learning of Everyday Household Tasks. In Aleksandra Faust, David Hsu, and Gerhard Neumann (eds.), *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pp. 455–465. PMLR, 2022. URL https://proceedings.mlr.press/v164/li22b.html.

Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Wensi Ai, Benjamin Martinez, Hang Yin, Michael Lingelbach, Minjune Hwang, Ayano Hiranaka, Sujay Garlanka, Arman Aydin, Sharon Lee, Jiankai Sun, Mona Anvari, Manasi Sharma, Dhruva Bansal, Samuel Hunter, Kyu-Young Kim, Alan Lou, Caleb R Matthews, Ivan Villa-Renteria, Jerry Huayang Tang, Claire Tang, Fei Xia, Yunzhu Li, Silvio Savarese, Hyowon Gweon, C. Karen Liu, Jiajun Wu, and Li Fei-Fei. BEHAVIOR-1K: A Human-Centered, Embodied AI Benchmark with 1,000 Everyday Activities and Realistic Simulation. *arXiv preprint arXiv:2403.09227*, 2024. URL https://arxiv.org/abs/2403.09227.

Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as Policies: Language Model Programs for Embodied Control. *arXiv preprint arXiv:2209.07753*, 2022. URL https://arxiv.org/abs/2209.07753v1. Version 1, September 2022.

Kaiqu Liang, Julia Kruk, Shengyi Qian, Xianjun Yang, Shengjie Bi, Yuanshun Yao, Shaoliang Nie, Mingyang Zhang, Lijuan Liu, Jaime Fernández Fisac, Shuyan Zhou, and Saghar Hosseini. Learning Personalized Agents from Human Feedback. *arXiv preprint arXiv:2602.16173*, 2026. URL https://arxiv.org/abs/2602.16173.

Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. CALVIN: A Benchmark for Language-Conditioned Policy Learning for Long-Horizon Robot Manipulation Tasks. *IEEE Robotics and Automation Letters*, 7(3):7327–7334, 2022. doi: 10.1109/LRA.2022.3180108. URL https://www2.informatik.uni-freiburg.de/~hermannl/.

Open X-Embodiment Collaboration. Open X-Embodiment: Robotic Learning Datasets and RT-X Models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892–6903. IEEE, 2024. doi: 10.1109/ICRA57147.2024.10611477. URL https://ieeexplore.ieee.org/document/10611477.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training Language Models to Follow Instructions with Human Feedback. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-2011. URL https://proceedings.neurips.cc/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as Operating Systems. *arXiv preprint arXiv:2310.08560*, 2024. URL https://arxiv.org/abs/2310.08560v2. Version 2.

Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. TEACh: Task-driven Embodied Agents that Chat. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 2017–2025, 2022. doi: 10.1609/aaai.v36i2.20097. URL https://ojs.aaai.org/index.php/AAAI/article/view/20097.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 2:1–2:22. Association for Computing Machinery, 2023. doi: 10.1145/3586183.3606763. URL https://doi.org/10.1145/3586183.3606763.

Maithili Patel, Xavier Puig, Ruta Desai, Roozbeh Mottaghi, Sonia Chernova, Joanne Truong, and Akshara Rai. ADAPT: Actively Discovering and Adapting to Preferences for any Task. *arXiv preprint arXiv:2504.04040*, 2025. URL https://arxiv.org/abs/2504.04040.

David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526, 1978. doi: 10.1017/S0140525X00076512. URL https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/does-the-chimpanzee-have-a-theory-of-mind/1E96B02CD9850016B7C93BC6D2FEF1D0.

Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Tsung-Yen Yang, Ruslan Partsey, Ruta Desai, Alexander Clegg, Michal Hlavac, So Yeon Min, Vladimír Vondruš, Theophile Gervet, Vincent-Pierre Berges, John Turner, Oleksandr Maksymets, Zsolt Kira, Mrinal Kalakrishnan, Jitendra Malik, Devendra Chaplot, Unnat Jain, Dhruv Batra, Akshara Rai, and Roozbeh Mottaghi. Habitat 3.0: A Co-Habitat for Humans, Avatars, and Robots. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/430894999584d0bd358611e2ecf00b15-Abstract-Conference.html.

Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew Botvinick. Machine theory of mind. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4218–4227. PMLR, 2018. URL https://proceedings.mlr.press/v80/rabinowitz18a.html.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741, 2023. doi: 10.52202/075280-2338. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. LaMP: When Large Language Models Meet Personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7370–7392. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.399. URL https://aclanthology.org/2024.acl-long.399/.

Karahan Sarıtaş, Kıvanç Tezören, and Yavuz Durmazkeser. A Systematic Review on the Evaluation of Large Language Models in Theory of Mind Tasks. *arXiv preprint arXiv:2502.08796*, 2025. URL https://arxiv.org/abs/2502.08796.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A Platform for Embodied AI Research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9339–9347, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Savva_Habitat_A_Platform_for_Embodied_AI_Research_ICCV_2019_paper.html.

Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, and Yulia Tsvetkov. Minding Language Models' (Lack of) Theory of Mind: A Plug-and-Play Multi-Character Belief Tracker. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13960–13980, Toronto, Canada, jul 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.780. URL https://aclanthology.org/2023.acl-long.780/.

Melanie Sclar, Jane Dwivedi-Yu, Maryam Fazel-Zarandi, Yulia Tsvetkov, Yonatan Bisk, Yejin Choi, and Asli Celikyilmaz. Explore Theory of Mind: Program-guided adversarial data generation for theory of mind reasoning. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/a901d5540789a086ee0881a82211b63d-Abstract-Conference.html.

Haojun Shi, Suyu Ye, Xinyu Fang, Chuanyang Jin, Leyla Isik, Yen-Ling Kuo, and Tianmin Shu. MuMA-ToM: Multi-modal Multi-Agent Theory of Mind. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 1510–1519, 2025. doi: 10.1609/aaai.v39i2.32142. URL https://ojs.aaai.org/index.php/AAAI/article/view/32142.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10740–10749, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Shridhar_ALFRED_A_Benchmark_for_Interpreting_Grounded_Instructions_for_Everyday_Tasks_CVPR_2020_paper.html.

Adil Soubki, John Murzaku, Arash Yousefi Jordehi, Peter Zeng, Magdalena Markowska, Seyed Abolghasem Mirroshandel, and Owen Rambow. Views Are My Own, but Also Yours: Benchmarking Theory of Mind Using Common Ground. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14815–14823, Bangkok, Thailand, aug 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.880. URL https://aclanthology.org/2024.findings-acl.880/.

James W. A. Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, Michael S. A. Graziano, and Cristina Becchio. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7):1285–1295, 2024. doi: 10.1038/s41562-024-01882-z. URL https://www.nature.com/articles/s41562-024-01882-z.

Andrew Szot, Alexander Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Singh Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimír Vondruš, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra. Habitat 2.0: Training Home Assistants to Rearrange their Habitat. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/021bbc7ee20b71134d53e20206bd6feb-Abstract.html.

Karthik Valmeekam, Matthew Marquez, Sarath Sreedharan, and Subbarao Kambhampati. On the Planning Abilities of Large Language Models - A Critical Investigation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 75993–76005, 2023. doi: 10.52202/075280-3320. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/efb2072a358cefb75886a315a6fcf880-Abstract-Conference.html.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-1800. URL https://proceedings.neurips.cc/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Yotam Wolf, Noam Wies, Oshri Avnery, Yoav Levine, and Amnon Shashua. Fundamental Limitations of Alignment in Large Language Models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 53079–53112. PMLR, 2024. URL https://proceedings.mlr.press/v235/wolf24a.html.

Yufan Wu, Yinghui He, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. Hi-ToM: A Benchmark for Evaluating Higher-Order Theory of Mind Reasoning in Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10691–10706.

Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.717. URL https://aclanthology.org/2023.findings-emnlp.717/.

Manjie Xu, Xinyi Yang, Wei Liang, Chi Zhang, and Yixin Zhu. Inferring the Unspoken: Aligning Embodied Agents with Implicit Preferences. *arXiv preprint arXiv:2502.00858*, 2026. URL https://arxiv.org/abs/2502.00858v4. Version 4. Earlier versions titled "Learning to Plan with Personalized Preferences".

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. In *Advances in Neural Information Processing Systems*, volume 36, 2023a. doi: 10.52202/075280-0517. URL https://proceedings.neurips.cc/paper/2023/hash/271db9922b8d1f4dd7aaef84ed5ac703-Abstract.html.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing Reasoning and Acting in Language Models. In *International Conference on Learning Representations (ICLR)*, 2023b. URL https://openreview.net/forum?id=WE_vluYUL-X.

Luyao Yuan, Xiaofeng Gao, Zilong Zheng, Mark Edmonds, Ying Nian Wu, Federico Rossano, Hongjing Lu, Yixin Zhu, and Song-Chun Zhu. In situ bidirectional human-robot value alignment. *Science Robotics*, 7(68):eabm4183, 2022. doi: 10.1126/scirobotics.abm4183. URL https://www.science.org/doi/10.1126/scirobotics.abm4183.

Zhaowei Zhang, Ceyao Zhang, Nian Liu, Siyuan Qi, Ziqi Rong, Song-Chun Zhu, and Yaodong Yang. Heterogeneous Value Alignment Evaluation for Large Language Models. In *Artificial General Intelligence*, volume 16058 of *Lecture Notes in Computer Science*, pp. 381–392. Springer, 2025. doi: 10.1007/978-3-032-00800-8_34. URL https://link.springer.com/chapter/10.1007/978-3-032-00800-8_34.

Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. Do LLMs Recognize Your Preferences? Evaluating Personalized Preference Following in LLMs. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/28a46044775d97a4efcbcf14e7f13209-Abstract-Conference.html.

Andy Zhou, Kai Yan, Michal Shlapentokh-Rothman, Haohan Wang, and Yu-Xiong Wang. Language Agent Tree Search Unifies Reasoning, Acting, and Planning in Language Models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62138–62160. PMLR, 2024a. URL https://proceedings.mlr.press/v235/zhou24r.html.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. SOTOPIA: Interactive Evaluation for Social Intelligence in Language Agents. In *International Conference on Learning Representations (ICLR)*, 2024b. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/b3075b88e583a0e98d8b24338a613060-Abstract-Conference.html.

Brian D. Ziebart, Andrew Maas, J. Andrew Bagnell, and Anind K. Dey. Maximum Entropy Inverse Reinforcement Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 1433–1438, 2008. URL https://www.cs.cmu.edu/~bziebart/publications/maximum-entropy-inverse-reinforcement-learning.html.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen,

Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. In *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 2165–2183. PMLR, 2023. URL https://proceedings.mlr.press/v229/zitkovich23a.html.

TECHNICAL APPENDICES

- Section A: Related Work
- Section B: More Details About CoHabit Arena
  - Section B.1: Embodied Outcome Signals
  - Section B.2: Evaluation Metrics Details
  - Section B.3: Persona Profiles
  - Section B.4: Cohabitation Sessions
  - Section B.5: Synthetic Artifact Validity Scope and Auditability
- Section C: Preference-State Maintenance Baseline Algorithm Details
- Section D: Supplementary Results and Analysis
  - Section D.1: Temporal Evolution of Mental Alignment
  - Section D.2: Performance at the Last Instruction
  - Section D.3: Comparative Analysis of Preference-State Maintenance Representation and Accumulation
  - Section D.4: Temporal Dynamics of Component Contributions
  - Section D.5: Compute Resources
- Section E: Case Study
- Section F: Limitations and Future Work

A RELATED WORK

A.1 EMBODIED AGENTS AND INTERACTION ARENAS.

Embodied instruction following studies how agents ground natural-language directives in perception and action to solve long-horizon tasks under partial observability. To support controlled evaluation, recent work increasingly relies on photorealistic simulators and standardized household task suites spanning navigation, manipulation, assistance, and collaboration.

Foundational platforms such as Habitat (Savva et al., 2019), Habitat 2.0 (Szot et al., 2021), and Habitat 3.0 (Puig et al., 2024) provide widely adopted infrastructure for home-scale embodied evaluation and rearrangement tasks. Building on these environments, language-grounded household benchmarks operationalize compositional everyday goals and multi-step interaction within a single episode, including ALFRED (Shridhar et al., 2020), TEACh (Padmakumar et al., 2022), and Housekeep (Kant et al., 2022). Related embodied ecosystems further expand coverage of interactive indoor scenes and manipulation, such as AI2-THOR (Kolve et al., 2017), RoboTHOR (Deitke et al., 2020), iGibson (Li et al., 2022), ManiSkill2 (Gu et al., 2023), and CALVIN (Mees et al., 2022). More recently, human-centered activity suites and collaboration-oriented arenas incorporate richer household behaviors and coordination constraints, exemplified by BEHAVIOR-1K (Li et al., 2024) and PARTNR (Chang et al., 2025).

In parallel, LLM-centric agent frameworks provide reusable reasoning and planning scaffolds, including Chain-of-Thought (Wei et al., 2022), ReAct (Yao et al., 2023b), Tree-of-Thoughts (Yao et al., 2023a), and LATS (Zhou et al., 2024a), though their reliability for decision-heavy planning remains debated (Valmeekam et al., 2023). For grounded control, recent systems couple LLM reasoning with skill libraries and closed-loop execution, as in SayCan (Ichter et al., 2023), Code as Policies (Liang et al., 2022), Inner Monologue (Huang et al., 2023), RT-2 (Zitkovich et al., 2023), and large-scale cross-robot data aggregation via Open X-Embodiment (Open X-Embodiment Collaboration, 2024).

Existing embodied resources provide the simulator and task infrastructure on which a preference diagnostic can be built. MEMENTO studies personalized object semantics and user routines through memory acquisition and utilization (Kwon et al., 2025). CoHabit Arena instead organizes executed choices, their consequences, and persona-conditioned reactions into a task-sequence diagnostic. This is an evaluation-layer contribution built on existing infrastructure, not a claim that prior embodied environments cannot be augmented with memory or personalization. The current experiments do not isolate the added value of physical simulation through a matched text-only control.

A.2 From Human-Value Alignment to Personal Preference Realization.

Human-value alignment studies how AI systems should act according to human values, preferences, and normative constraints. In LLMs, this is commonly instantiated through instruction tuning and preference-based optimization, including RLHF-style pipelines (Christiano et al., 2017; Ouyang et al., 2022), principle- or rule-guided self-critique such as Constitutional AI (Bai et al., 2022), and direct preference objectives like DPO (Rafailov et al., 2023). Prior work has also highlighted the gap between surface-level compliance and robust value adherence (Wolf et al., 2024), and surveyed the broader alignment landscape across objectives, methods, and evaluation challenges (Ji et al., 2025). In embodied and HRI settings, value conflicts often arise *in situ* during interaction, motivating approaches that infer and align values from collaborative behavior (Yuan et al., 2022). Recent evaluation efforts further emphasize heterogeneity across users and contexts, motivating robustness-oriented alignment benchmarks (Zhang et al., 2025).

The distinction relevant here is between receiving a complete target profile and acquiring partial preference evidence over interaction. It is not a division between supervised existing methods and an unsupervised PPR setting: CoHabit Arena includes explicit corrections and structured synthetic feedback. PrefEval already evaluates inferring, remembering, and following explicit and implicit preferences across long conversations (Zhao et al., 2025). Thus, PPR denotes a simulator-backed combination of diagnostic objectives, rather than the first study of preference acquisition and retention.

CoHabit Arena separates agreement with a synthetic preference reference, execution of preference-sensitive choices, and task completion. Its intended distinction is the association of executed trajectories with preference-diagnostic alternatives and persona-conditioned outcomes across household tasks. Whether this association supports representation-independent behavioral comparisons depends on the input, readout, matching, and coverage requirements described in Sections 3.2–3.4; the present study does not fully validate that claim.

A.3 Implicit Preference Learning and Reward Inference.

Implicit preference learning aims to infer a latent utility or reward function from behaviorally grounded signals, enabling agents to act under underspecified objectives while remaining responsive to human judgments.

Classic inverse reinforcement learning and Bayesian reward learning infer rewards from observed trajectories under rationality assumptions, including maximum-entropy IRL (Ziebart et al., 2008) and cooperative inverse RL (Hadfield-Menell et al., 2016). Subsequent work highlights reward misspecification and context dependence, motivating robust formulations such as inverse reward design (Hadfield-Menell et al., 2017) and reward-rational (implicit) choice models (Jeon et al., 2020). In deep RL, preference-based methods elicit human feedback as pairwise comparisons over trajectory segments to train a reward predictor that guides policy optimization (Christiano et al., 2017). Preference signals can also take the form of ranked demonstrations, as in T-REX, which learns rewards from ordered trajectories and extrapolates beyond suboptimal data (Brown et al., 2019). For language models, analogous preference-labeled response comparisons underpin RLHF in InstructGPT (Ouyang et al., 2022), while DPO bypasses explicit reward modeling by directly optimizing a closed-form preference objective (Rafailov et al., 2023).

Preference learning from behavior and interactive feedback is directly relevant. PbP/InTU evaluates inferring implicit preferences from demonstrations and applying them in subsequent embodied plans (Xu et al., 2026). ADAPT studies active questioning in household assistance (Patel et al., 2025); PAHF combines per-user memory, pre-action clarification, and post-action feedback, including preference changes (Liang et al., 2026). CoHabit Arena focuses on a fixed synthetic persona and retrospective choice–outcome association. It does not subsume active clarification, preference drift, or demonstration-based learning, and no empirical superiority over these systems is claimed.

Persistent memory is also not specific to E-ToM. Generative Agents stores, reflects on, and retrieves experiences for later behavior (Park et al., 2023); MemGPT manages memory tiers for long-context and multi-session interaction (Packer et al., 2024). These works motivate memory-equipped controls. Since such controls are absent from the reported experiment, the relative contribution of generic persistence and the axis-native state remains unresolved.

A.4 Theory of Mind for Interactive Agents.

Theory of Mind (ToM) models other agents' latent mental states (e.g., beliefs, desires, intentions) to predict behavior and enable coordination, and is commonly formalized as inverse planning and probabilistic mental-state inference.

Early cognitive accounts framed ToM as reasoning over unobservable mental states (Premack & Woodruff, 1978). Computational models later operationalized ToM as goal and plan inference from action (Baker et al., 2009), and refined this view through utility-based commonsense inference (Jara-Ettinger et al., 2016) and quantitative attribution of beliefs and desires (Baker et al., 2017). In machine learning, Machine Theory of Mind introduced differentiable agents that infer others' hidden states directly from behavior (Rabinowitz et al., 2018). For LLMs, recent work evaluates ToM competence and limitations via controlled tests and human comparisons (Kosinski, 2024; Strachan et al., 2024), while complementary benchmarks emphasize interactive phenomena such as mental-state tracking (Kim et al., 2023), higher-order ToM (Wu et al., 2023), negotiation (Chan et al., 2024), and common-ground reasoning (Soubki et al., 2024).

Program- and symbol-guided approaches further examine whether apparent ToM reflects genuine mental-state modeling or shortcut heuristics, including plug-and-play belief tracking (Sclar et al., 2023) and adversarial data generation (Sclar et al., 2025). Recent surveys and analyses highlight methodological variance and recurring evaluation pitfalls in LLM-based ToM research (Sarıtaş et al., 2025). In embodied and multi-agent settings, MuMA-ToM links ToM reasoning to multimodal coordination (Shi et al., 2025), and Hypothetical Minds scaffolds hypothesis-driven ToM reasoning for multi-agent reinforcement learning (Cross et al., 2025).

Hypothetical Minds uses maintained natural-language hypotheses and their evaluation, rather than merely a static ToM instruction (Cross et al., 2025). The implementation labeled Hyp-Minds in this study is an adaptation, so its results cannot establish a general limitation of the original architecture. SOTOPIA evaluates interactive social goal pursuit through open-ended role-play (Zhou et al., 2024b), illustrating that multi-turn consequence-sensitive evaluation also exists outside physical simulators. CoHabit Arena's narrower target is single-user, simulator-backed preference diagnosis across household tasks.

B More Details About CoHabit Arena

B.1 Embodied Outcome Signals

To operationalize the reaction term in PPR's choice–consequence–reaction evidence, CoHabit Arena provides a spectrum of *Embodied Outcome Signals* within each Causal Attribution Window. These signals describe how the persistent persona responds to the consequence of the retrospectively matched choice. Their natural-language form represents a user outcome, while structured fields support attribution and state updates. These roles expose an important distinction between inference from raw feedback and accumulation of supplied diagnostic metadata; field-level equality of information across methods has not been established. We categorize these signals into three levels of explicitness: *Physical Interventions*, *Instructional Cues*, and *Social Sentiment*.

B.1.1 Taxonomy of Embodied Outcome Signals

Level 1: Physical Interventions (Hard Constraints). These categories describe strong objections with interruption-like semantics. In a post-task reconstruction they annotate a completed trace; an online physical interruption requires a separately recorded event and changed execution. They are intended to represent responses to violations of a high-priority personal preference, such as safety, privacy, time sensitivity, or spatial disturbance.

- **Veto:** A stop-request category describes an objection to an action, for example by issuing a direct stop command or telling the agent not to proceed. This is treated as a strong negative reaction to the corresponding choice.
- **Takeover:** A takeover category describes the user taking responsibility for a sub-task, for example by stepping in and completing the sub-task personally. Those motives are possible

interpretations in the synthetic specification, not independently observed causes. A post-hoc label does not establish that the user actually changed the simulator trajectory.

- **Correction:** The user points out an error and demands an immediate fix, for example by indicating that the chosen object, route, or action is wrong. This acts as negative evidence for the selected option.
- **Redo:** The user requires the agent to attempt the action again, for example by asking for repeated execution. This indicates dissatisfaction with the realized outcome rather than necessarily with the original task intent.

Level 2: Instructional Cues (Soft Constraints). These signals guide the agent's behavior through specific directives or reminders without necessarily stopping the task. They often indicate personal spatial, temporal, or procedural preferences.

- **Reposition:** The user asks the agent to move or step aside, for example by requesting more personal space or clearing a blocked path. This indicates a violation of personal space or spatial blocking.
- **Wait:** The user requests a temporary pause, for example by asking the agent to briefly hold position. This often signals a preference for synchronization, lower autonomy, or reduced disturbance.
- **Constraint Repetition:** The user proactively reminds the agent of a rule (e.g., "Remember, no glass on the table"), reinforcing a persistent personal constraint.
- **Assist:** The user offers help, for example by volunteering to help with the ongoing sub-task. This may imply that the agent appears to be struggling, inefficient, or miscalibrated to the user's preferred assistance style.

Level 3: Social Sentiment (Implicit Outcome). These categories model subtler responses. Their frequency and usefulness in real households are not validated here; within the synthetic setup they are intended as evidence about whether the consequence of a choice is acceptable to the user.

- **Negative Sentiment:** The user exhibits signs of annoyance or frustration (e.g., sighing, "Ugh," "Not again"). This serves as a weak negative signal for the agent's recent behavior.
- **Positive Acceptance:** The user expresses approval or acknowledgement (e.g., "Okay," "Thanks," "Good"). In the absence of negative intervention within a diagnostic window, CoHabit Arena may also generate an implicit acceptance signal, reflecting the weak positive evidence that the choice was acceptable to the persona.

B.1.2 Embodied Outcome Signals Generation

Outcome signals are generated from the relationship between the persona's ground-truth preference profile and the consequence of the retrospectively matched option. For each preference-diagnostic choice, CoHabit Arena compares the selected option with the persona-preferred option under the corresponding axis-consequence vectors. Severe mismatches, such as violating a high-priority privacy or safety preference, lead to strong signals such as *Veto* or *Takeover*; moderate mismatches lead to *Correction*, *Wait*, or *Negative Sentiment*; and preference-consistent choices may lead to *Positive Acceptance*. When explicit trace-level feedback already exists inside the attribution window, it takes precedence; otherwise, persona-conditioned outcome signals are generated to provide standardized PPR evidence. The arena may additionally include *Freeform Outcomes* as natural-language variants of the structured signal, preventing agents from relying only on fixed keywords while preserving a reproducible metadata record for evaluation.

Generation scope. The description above gives category semantics, not a complete executable generator. Numerical mismatch thresholds, probability laws, random seeds, textual templates or generation prompts, and any backend model are not specified in the reported configuration. We therefore do not assert that this generator is deterministic, calibrated to human responses, or independent of the evaluated model families. The supplemental description does not resolve these implementation choices.

Time and information. Existing trace-level feedback and feedback synthesized after matching have different status. Only an actual online event can interrupt execution; a retrospective synthetic signal can update later beliefs but cannot change an earlier success outcome. Ground-truth profiles and option utilities may legitimately be used by the trusted environment to generate feedback.

Their direct exposure to any planner, wrapper, memory process, or candidate scorer is a separate information-access question.

B.2 Evaluation Metrics Details

B.2.1 Mental Alignment Metrics

Readout scope. The reported tables do not specify a uniform extraction procedure, prompt, timing, or failure rule for methods without an explicit preference vector. Consequently, Final MA and AUC MA describe recorded estimates rather than all information represented by a model. Substituting a prior for an unavailable readout must not be interpreted as an observed absence of learning. A common externally elicited probe, or held-out preference-sensitive choices, is required for a representation-independent comparison; that validation is not reported here.

Mental Alignment is the inference-side diagnostic for PPR. It measures how accurately a method recovers the user's latent personal preference profile. Let $z_h \in \Delta^K$ be the ground-truth preference vector of persona h, and let $z^{(t)}_{\text{pred}}$ be the preference estimate recorded by the arena interface at task or decision step t. All comparisons are made in the active or common preference-axis space defined for the current evaluation instance. The base metric is the cosine similarity:

$$\text{MA}(z_{\text{pred}}, z_h) = \frac{z_{\text{pred}} \cdot z_h}{\|z_{\text{pred}}\|_2 \, \|z_h\|_2}. \tag{9}$$

To distinguish methods that acquire the personal preference early from those that only match it near the end, we report two aggregated metrics:

Final Mental Alignment (Final MA). Measures preference-inference accuracy at the end of the cohabitation session (T), reflecting the final recovered estimate of the user's personal preference profile.

$$\text{Final MA}(h) \;=\; \text{MA}\big(z^{(T)}_{\text{pred}},\, z_h\big) \;=\; \frac{z^{(T)}_{\text{pred}} \cdot z_h}{\|z^{(T)}_{\text{pred}}\|_2 \, \|z_h\|_2} \tag{10}$$

Area Under Curve Mental Alignment (AUC MA). Measures the cumulative quality of preference inference across the session. A higher AUC indicates that the method identifies the user's preference earlier and maintains it over later tasks.

$$\text{AUC MA}(h) \;=\; \frac{1}{T}\sum_{t=1}^{T} \text{MA}\big(z^{(t)}_{\text{pred}},\, z_h\big) \;=\; \frac{1}{T}\sum_{t=1}^{T} \frac{z^{(t)}_{\text{pred}} \cdot z_h}{\|z^{(t)}_{\text{pred}}\|_2 \, \|z_h\|_2} \tag{11}$$

B.2.2 Behavioral Alignment Metrics

Behavioral Alignment is the realization-side diagnostic for PPR. It evaluates whether the agent's executed choices realize the user's true personal preferences. Let $\mathcal{P}$ be the set of valid preference-diagnostic choices encountered in a session. For each $p \in \mathcal{P}$, let $\mathcal{O}(p)$ be the option set, and let $U(o \mid z_h) = z_h^\top x(o)$ be the evaluator-only utility of option o under the ground-truth preference profile z_h, where $x(o)$ is the option's axis-consequence vector.

Decision Regret (D-Regret). We calculate regret as the difference between the utility of the best available option and the utility of the option $o_{\text{sel},p}$ retrospectively matched to the executed trajectory. We report the average regret across all valid diagnostic choices in the session:

$$\text{D-Regret}(h) \;=\; \frac{1}{|\mathcal{P}|}\sum_{p\in\mathcal{P}} \left[\max_{o\in\mathcal{O}(p)} U(o \mid z_h) \;-\; U(o^{(p)}_{\text{sel}} \mid z_h)\right]. \tag{12}$$

Lower regret indicates stronger preference realization. Note that $U(\cdot)$ is used strictly for evaluation and is never observed by the agent.

Utility scale and selection. The domain and common scaling of $x(o)$ are not specified by the current artifact summary. Thus the absolute size of D-Regret is not interpretable as a standardized effect. For

a bounded common scale $x_k(o) \in [a, b]$ and simplex weights, the defining average lies in $[0, b-a]$; this mathematical fact does not determine the scale used in the experiment. Physically feasible options must also respect the explicit task goal, or be scored against a documented updated instruction.

Coverage and evaluator independence. Regret conditions on the set of matched diagnostic choices. Counts for encountered, unvisited, unmatched, excluded, and scored choices are not reported, so different methods may be averaged over different subsets. An empty set makes the formula undefined, not zero. Evaluation should match every method's traces independently of whether that method contains a mining module. The missing baseline and mining-ablation entries show that the current report does not provide that uniform comparison.

B.2.3 Instruction-Following Metrics

We adopt standard embodied instruction-following metrics to evaluate task-level execution ability, following prior work. Specifically, we report Success Rate (SR) and Success weighted by Path Length (SPL), which jointly measure task completion and navigation efficiency. These metrics are computed over the cohabitation session to ensure that PPR is evaluated alongside, rather than instead of, basic functional competence.

Equation 13 shows the standard SR and SPL formulas used in embodied navigation.

$$\begin{aligned} \mathrm{SR} &= \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{\mathrm{success}_i\} \\ \mathrm{SPL} &= \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{\mathrm{success}_i\} \cdot \frac{L_i}{\max(P_i,\, L_i)} \end{aligned} \tag{13}$$

where N is the number of tasks, $\mathbf{1}\{\mathrm{success}_i\}$ indicates whether the i-th task is successful, L_i is the shortest path length, and P_i is the actual path length traveled.

Consistency of the reported columns. For non-negative lengths, each success-weighted path term is at most its success indicator. Therefore identical samples and weights require $\mathrm{SPL} \le \mathrm{SR}$, including after consistent macro-aggregation. The three DeepSeek rows identified in Section 5.2 violate this condition. Different denominators, a partial-completion measure, or an implementation error could explain a discrepancy, but none is established by the available aggregate record. We do not silently rename the values, alter the definition, or infer corrected measurements; the affected pairs remain unresolved.

B.2.4 Convergence Cost

Convergence Cost is a belief-state diagnostic for methods that expose an explicit preference state. We operationalize convergence as the earliest task or diagnostic-decision index at which the preference-state estimate is accurate, confident, and no longer near its initial prior. We report this cost as $\mathrm{Ep@Thr}$.

Concretely, $\mathrm{Ep@Thr}$ is jointly determined by the prediction error to the ground-truth personal preference profile, the confidence of the preference-state estimate, and the amount of effective movement from the initial prior. This prevents methods that do not update their preference estimate from being counted as converged merely because the prior is accidentally favorable.

Prediction Error (L_2). We measure the Euclidean distance shown in Equation 14 between the predicted preference vector and the ground-truth personal preference vector:

$$L_2(t) \;=\; \left\| z_{\text{pred}}(t) - z_h \right\|_2 . \tag{14}$$

where $z_{\text{pred}}(t)$ is the predicted preference vector at task or decision step t, z_h is the ground-truth personal preference vector, and $\|\cdot\|_2$ denotes the Euclidean norm.

Preference-State Confidence (α_0). We use the Dirichlet concentration shown in Equation 15 as a scalar proxy for accumulated preference-state confidence:

$$\alpha_0(t) \;=\; \sum_{k=1}^{K} \alpha_k(t), \tag{15}$$

where $\alpha^{(t)} = (\alpha_1(t), \ldots, \alpha_K(t))$ are the Dirichlet parameters at step t, and K is the number of active preference axes.

The term "confidence" is a proxy within this pseudo-count representation. Without a likelihood model and held-out calibration assessment, its numerical concentration is not a calibrated probability that the inferred profile is correct. Non-negative evidence updates mechanically increase total mass, even for noisy or biased evidence. Equal concentration across methods does not imply equal normalized means or equal latent knowledge.

Effective Preference Movement. To avoid treating an unchanged prior as convergence, we additionally require the preference estimate to move sufficiently from its initial value:

$$M(t) \;=\; \left\| z_{\text{pred}}(t) - z_{\text{pred}}(0) \right\|_2 . \tag{16}$$

Episode-to-convergence (Ep@Thr). We say the preference-state estimate has converged when L_2 is below the error threshold, α_0 is above the confidence threshold, and $M(t)$ exceeds a minimum movement threshold within a stability window of size w:

$$\text{Ep@Thr} \;=\; \min\Big\{\, t \;:\; \forall i \in [t-w+1,\, t],\;\; L_2(i) \le \tau_L \;\wedge\; \alpha_0(i) \ge \tau_\alpha \;\wedge\; M(i) \ge \tau_M \Big\}. \tag{17}$$

where τ_L is the error threshold, τ_α is the concentration threshold, τ_M is the minimum effective-movement threshold, w is the stability window size, and t is the task or diagnostic-decision index. The **Convergence Cost** is reported as Ep@Thr; lower values indicate faster acquisition of a usable personal preference estimate within the cohabitation session.

Unreconciled reported convergence. An unchanged estimate has $M(t) = 0$ and cannot satisfy a positive τ_M, regardless of a favorable prior. Thus the frozen-state row with Ep@Thr $= 1.0$ is incompatible with this definition. Other entries cannot be validated without the thresholds, choice of task versus diagnostic-decision index, stability-window convention, and censoring rule. A 15-task session can contain more than 15 diagnostic choices, so a larger index alone is not a contradiction. The 17.4 and 3.1 values for nominally corresponding full configurations are also not reconciled. No time-to-convergence ranking is supported by the present column.

B.3 Persona Profiles

CoHabit Arena constructs a scene-grounded pool of **36 persistent synthetic personas** to support controlled evaluation of Personal Preference Realization (PPR). Each persona profile encodes an individual user's implicit personal preferences as a reference signal, enabling systematic assessment of whether an agent can infer these preferences from interaction and realize them in later household decisions.

Concretely, a profile specifies three components: (1) a natural-language persona description, (2) an implicit personal preference specification together with an interaction tone, and (3) behavioral regularizers in the form of constraints and habitual patterns. All components are grounded in the originating scene and linked to preference-diagnostic choices that make the persona behaviorally identifiable.

B.3.1 Profile Representation

To ensure reproducibility and auditability, we represent the persona pool in a versioned, machine-readable format that serves as a stable interface between arena generation and evaluation. Each profile record is uniquely identified by an ID and an archetype label, and contains structured fields for implicit preferences, interaction style, and behavioral rules, with optional extensions for Theory-of-Mind attributes and reward specifications. Figure 4 illustrates an example profile.

B.3.2 Archetype Coverage

The persona pool comprises **36 persona profiles**, distributed across distinct persona archetypes that capture qualitatively different household value systems and interaction preferences. Table 4 lists how the **36 persona profiles** are distributed across archetypes: a small set of frequent archetypes forms the head and the remainder populate a long tail. The labels describe the constructed pool, not a

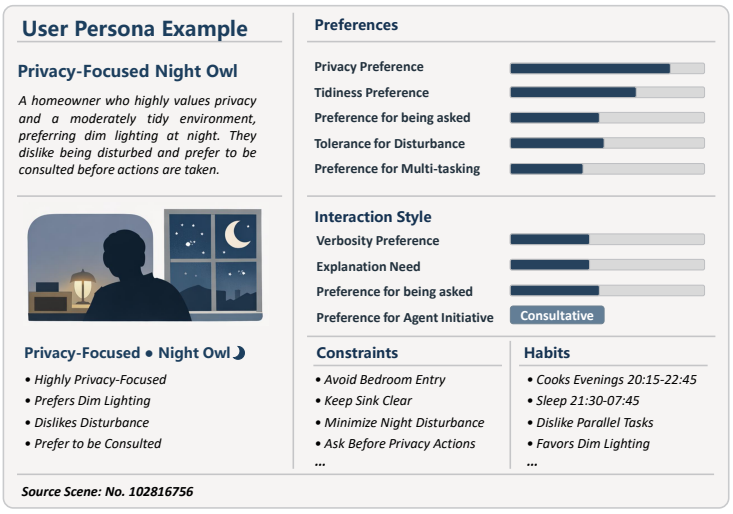


Figure 4: An example persona profile shown as a card.

validated sample of a resident population. They span different preference emphases (e.g., privacy, efficiency orientation, disturbance aversion) and interaction attitudes. Archetype names alone do not establish independent latent preferences, demographic coverage, or generalization to new users.

Table 4: Distribution of the 36 persistent persona profiles across user archetypes.

User Archetype	Count
Privacy-Focused Consultative Resident	3
Privacy-Aware Tidy Autonomous Resident	3
Efficiency-Oriented Minimalist	2
Privacy-Prioritizing Resident	2
Privacy-Prioritizing Low-Disturbance Resident	2
Disturbance-Averse Route Planner	1
Efficiency-First Minimalist	1
Efficiency-Focused Minimalist	1
Efficient Low-Disturbance Resident	1
Multi-tasking Autonomous Disturbance-Tolerant Resident	1
Multi-tasking Autonomous Tidy Resident	1
Multi-tasking Tidy Autonomous Resident	1
Privacy-Focused Consultative Night Owl	1
Privacy-Focused Night Owl	1
Privacy-Focused Consultative Night Dweller	1
Privacy-Preferring Autonomous Resident	1
Privacy-Focused Autonomous Night Dweller	1
Privacy-Focused Consultative Homebody	1
Privacy-Focused Disturbance-Averse Resident	1
Relaxed Low-Privacy Resident	1
Speed-Oriented Resident	1
Step-Minimizing Home Manager	1
Step-Minimizing Resident	1
Tidy Low-Disturbance Collaborative Resident	1
Tidy Privacy-Focused Autonomous Homebody	1
Tidy Privacy-Focused Collaborative Resident	1
Tidy Privacy-Focused Low-Disturbance Resident	1
Tidy-Focused Collaborative Resident	1
Tidy-Focused Homebody	1

B.3.3 PREFERENCE DIMENSIONS

Each persona profile encodes a mixed-type personal preference specification over a global allowlist of axes. For evaluation, each scene or session activates a subset of these axes and maps the persona to a scene-conditioned preference vector $z_h \in \Delta^K$, where K denotes the number of active axes for the current evaluation instance. Figure 5 (a) summarizes the distribution of preference fields, showing core axes shared across many personas alongside optional axes that introduce controlled heterogeneity.

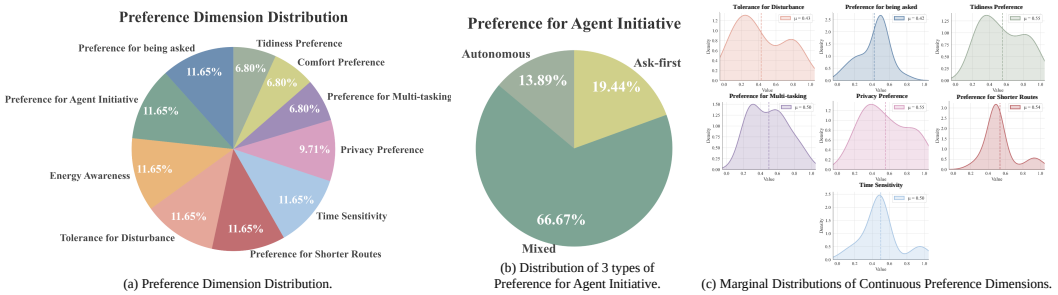
Representation boundary. The descriptive record may contain categorical values, independently scaled preferences, habits, and structured comfort settings. The vector z_h is instead a relative weighting used by the diagnostic utility. The present report does not provide the exact conversion for every field, including encoding, active-axis selection, direction, and normalization. The stated simplex must therefore not be interpreted as a literal model of the entire mixed-type profile or a population Dirichlet distribution. A normalization that loses absolute strength can change the meaning of a preference; no unreported conversion is supplied here.

Identifiability. Finite diagnostic choices may not distinguish all reference vectors. If two profiles give the same feedback distribution and option ordering on every encountered choice, interaction evidence cannot uniquely identify them. MA and L_2 can still distinguish their coordinates, creating a mismatch between recoverability and scoring. Axis coverage and held-out behavioral predictions are needed to establish identifiability; the reported field-frequency plots do not provide that test.

We organize the preference fields into three groups:

- **Categorical:** *Preference for Agent Initiative* indicates whether the user expects the agent to act independently (*Autonomous*), ask before acting (*Ask-first* or *Consultative*), or balance between the two (*Mixed*). As shown in Figure 5 (b), all categories are represented with non-trivial frequency, ensuring evaluation spans distinct autonomy regimes.
- **Continuous (0–1 scale):** Dimensions such as *Energy Awareness*, *Time Sensitivity*, *Preference for Shorter Routes*, *Tolerance for Disturbance*, *Tidiness Preference*, *Preference for Multi-tasking*, and *Privacy Preference* quantify persona trade-offs along efficiency, interaction overhead, environmental tolerance, and household norms. In the current release, *Energy Awareness* is fixed to a neutral value of 0.5 for all users, preventing effort-related signals from being conflated with other efficiency axes (e.g., time or path length) while preserving compatibility with future extensions. Figure 5 (c) reports their marginal distributions, showing broad support across the range rather than collapsing to a single operating point.
- **Structured:** *Comfort Preference* captures persona comfort over environment states (e.g., lighting, temperature, noise level) using an extensible key–value map, allowing additional state factors to be incorporated without changing the overall schema.

This mixed-type preference specification provides a compact yet expressive interface for representing diverse resident preferences. It supports PPR evaluation by providing a controllable reference for personal preference inference while allowing agents to be tested on whether such inferred preferences are realized in later choices.



(a) Preference Dimension Distribution.

(b) Distribution of 3 types of Preference for Agent Initiative.

(c) Marginal Distributions of Continuous Preference Dimensions.

Figure 5: **Personal preference field distribution.**

B.4 COHABITATION SESSIONS

To evaluate PPR in a measurable and controlled manner, CoHabit Arena must go beyond isolated household instructions. Interaction traces should (1) accumulate personal preference evidence across tasks, and (2) support consistent scoring without revealing the latent preferences to the agent.

To this end, we construct *Cohabitation Sessions* by chaining scene-grounded household instructions derived from PARTNR episodes under a shared sequential protocol. Each session specifies embodied contexts and instructions, together with static difficulty descriptors and a structured evaluation rubric. This design enables controlled comparison across methods under identical task chains, while supporting fine-grained assessment of personal preference inference and realization over time.

B.4.1 SESSION REPRESENTATION

Each session corresponds to a household task chain instantiated in a fixed scene and initial state. In addition to the natural-language instructions, sessions include static descriptors characterizing task scope and expected effort, as well as a multi-component scoring rubric used for evaluation. This representation separates task specification from scoring, allowing the same session to be reused across different agents and evaluation runs while preserving consistent attribution and comparability.

B.4.2 SCENE DIVERSITY

Sessions are instantiated across **12** distinct photorealistic indoor environments, each representing a complete household with fixed layout, geometry, and object placement. As illustrated in Figure 6, these scenes vary substantially in floorplan scale, room composition, and connectivity patterns, ranging from compact layouts to more expansive, highly connected spaces.

The scenes offer different layouts and object arrangements. This is descriptive coverage, not evidence that a planner cannot overfit or that preferences transfer to unseen homes. No separately identified held-out-scene or held-out-persona result is reported. Reusing source tasks and personas across sessions also means that task-instance count is not the number of independent users or independent experimental conditions.



Figure 6: **Scene visualization.** Top-down renderings of the **12** indoor environments used in the session pool. Scenes vary in scale, room composition, and connectivity, providing diverse spatial contexts for household cohabitation sessions.

B.4.3 TASK CATEGORY COVERAGE

The session pool spans **four** task categories inherited from the PARTNR taxonomy (Chang et al., 2025), which groups household tasks by their primary interaction structure. These include rearrangement tasks, spatially constrained navigation tasks, temporally ordered tasks, and tasks involving discrete object-state changes.

Figure 7 reports the empirical distribution of categories. While rearrangement tasks constitute the majority—reflecting the prevalence of object relocation in household activities—the remaining

categories introduce complementary interaction patterns. Together, this mix ensures that PPR is evaluated across qualitatively different task structures rather than a single manipulation routine.

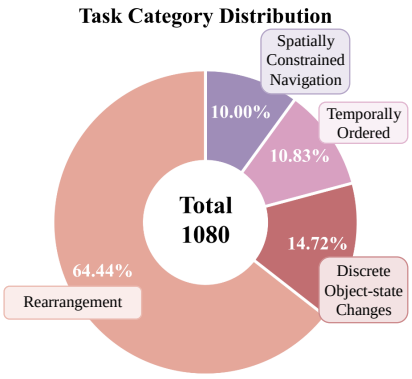


Figure 7: **Task category distribution.**

B.4.4 Object Coverage and Frequency

Sessions reference a broad range of household-relevant object categories, reflecting the diversity of objects encountered in realistic indoor instructions. Object mentions follow a pronounced head–tail distribution: a small set of common objects appears frequently, while a long tail of less frequent objects remains consistently present.

Frequent and rare objects provide different contexts for diagnostic choices. Figure 8 summarizes their coverage and frequencies. A long tail in this constructed pool does not by itself establish open-world generalization; such a conclusion requires an explicit held-out evaluation and control of shared tasks and scenes.

B.4.5 Session Complexity

Sessions vary in expected execution horizon and interaction effort, reflecting differences in navigation distance, number of required actions, and revisits. We summarize this variation using a static *effort proxy* that characterizes per-session difficulty independently of agent behavior.

Concretely, the effort proxy is computed as

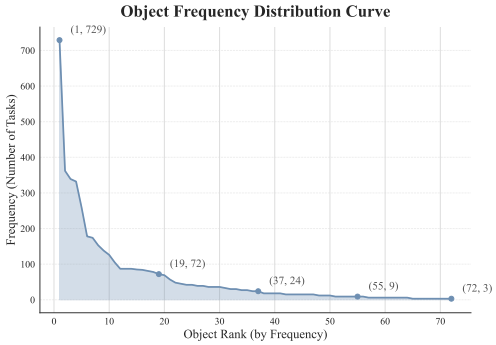
$$\textit{energy} = s_{\min} \cdot (1 + 0.15 \cdot n_{\text{obj}}), \tag{18}$$

where $s_{\min}$ estimates the minimal number of atomic actions required by the task specification, and n_{obj} denotes the number of target objects involved.

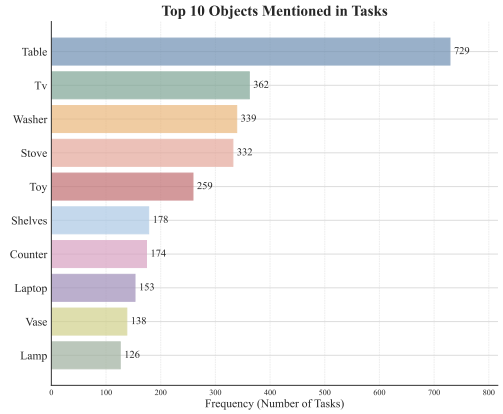
The $s_{\min}$ term is initialized to a base value of 2 and incremented according to task attributes: it increases by 1 if the task involves object state changes (e.g., opening, switching, or moving objects), by 1 if temporal ordering constraints are present, and by 1 if spatial relation constraints are required. The n_{obj} term is derived from the number of target objects specified in the task, defaulting to 1 when a single object is involved.

This formulation is intentionally simple and task-derived. The $s_{\min}$ term is a task-attribute heuristic for execution effort, not a proven lower bound on action count or shortest path. The multiplicative factor $(1 + 0.15 \cdot n_{\text{obj}})$ accounts for the increased coordination and interaction burden introduced by multi-object tasks, while ensuring that object count acts as a secondary modulator rather than the dominant driver of session difficulty.

We emphasize that this effort proxy is not intended to estimate true physical energy consumption or optimal execution cost. Instead, it serves as a lightweight, agent-agnostic normalization signal that stratifies sessions by relative interaction complexity. Its deterministic and transparent construction enables reproducible analysis and avoids coupling difficulty estimates to specific planners, policies, or simulator dynamics.



(a) Object mention counts as a function of object rank (sorted by frequency), illustrating a pronounced head–tail distribution.



(b) Counts of the top-10 most frequently mentioned object categories.

Figure 8: **Object count and frequency.** (a) The rank–frequency curve confirms a pronounced head–tail structure. (b) The top-10 plot enumerates the most frequently referenced household objects. Together, these statistics demonstrate that CoHabit Arena sessions are grounded in a diverse object space with both high-frequency "common objects" and a meaningful long tail, supporting realistic variability and generalization-oriented PPR evaluation.

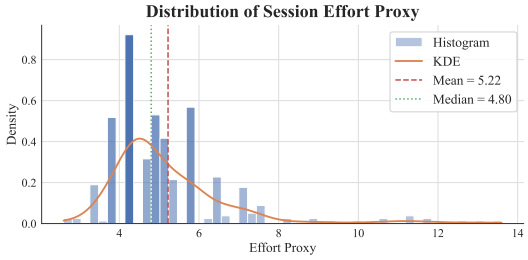


Figure 9: **Session effort proxy distribution.** Distribution of the effort proxy with mean/median indicators, showing a mid-range mass and a non-trivial high tail.

Figure 9 shows variation in the static effort proxy. This does not imply variable task-chain length: the reported evaluation uses 15 tasks per session. The proxy describes task attributes, not measured physical energy or validated interaction duration, and does not itself establish ecological validity.

B.5 Synthetic Artifact Validity Scope and Auditability

CoHabit Arena uses synthetic personas, option utilities, and standardized outcomes as a controlled diagnostic target. The profile, choice, and outcome schemas permit inspection of persona coherence, task feasibility, preferential distinctness, and reaction consistency. These are validation criteria, not evidence that a completed independent human study has confirmed them. The report does not specify an audit sample, raters, procedure, agreement, or quantitative outcomes. No population-validity or completed human-validation claim is made on that basis.

Feedback shortcuts and robustness. Because the same axes inform preference generation, structured feedback, updates, and evaluation, a method could exploit the synthetic coding rather than acquire a transferable user model. Withholding the exact reference vector does not eliminate this possibility. Alternative generators, hidden diagnostic metadata, noisy or delayed feedback, mismatched histories, and independently judged choices are informative controls, but results for them are not reported. Synthetic consistency is therefore not equivalent to robustness under real reactions.

Intended information boundary. Ground-truth preference vectors, option utilities, and axis-consequence vectors belong to trusted environment/evaluator processing, which includes legitimate

feedback generation. The boundary concerns all components of an evaluated agent, not only the LLM prompt. The wrapper described in Section 4 receives attributed, axis-level evidence, while the complete field permissions of ordinary planners and candidate scorers are not documented. Accordingly, the current results cannot establish either equal information or absence of privileged-feature use. This is an unresolved audit question, not an assertion that leakage occurred.

Experimental independence. Reported intervals over sessions do not reveal whether scene/persona clustering, reused task chains, early termination, or stochastic model calls were accounted for. The record does not identify a separate development split, locked held-out set, repeated seeds, or exact session-start memory reset across runs. Thus neither effective independent sample size nor out-of-sample generalization can be reconstructed from the aggregate counts.

Reproducibility scope. The current description provides conceptual schemas, equations, pseudocode, and aggregate results. A fully reconstructable experiment additionally needs instantiated profiles and choice bindings, generation thresholds and seeds, exact prompts and permissions, candidate budgets, per-task logs, and the metric code that produced the tables. These implementation records are not supplied by the manuscript source package; source files for typesetting are not an experimental code release.

C Preference-State Maintenance Baseline Algorithm Details

Algorithm 1 summarizes the reference state-update order: complete an execution, obtain matched decisions and associated feedback, map them to evidence, and update cross-task memory. It is schematic rather than a complete implementation. The named mining, window, collection, mapping, and execution functions do not fix their prompts, thresholds, action interface, or error handling.

For fixed task evidence, the sequential additions and consolidation reduce to $\alpha_{t+1} = \alpha_t + \rho\mu\Delta_t$. This is weighted pseudo-count accumulation. No explicit forgetting term, calibrated observation likelihood, or separate fast–slow inference theorem is established. A coordinate's normalized weight may decrease as other counts grow, but this is not the same as retracting erroneous evidence.

The "Outer-Loop Textual Refinement" row names an experimental variant without a specified operator in this algorithm. Its inputs, optimization objective, update schedule, and additional model calls are not recoverable from the pseudocode. Likewise, the exact candidate set and feature scores in Equation 8 cannot be inferred from the call to `ExecuteAgent`. We do not retrofit an unreported implementation or claim these missing choices have been reproduced.

D Supplementary Results and Analysis

D.1 Temporal Evolution of Mental Alignment

Figure 3a illustrates the temporal progression of Mental Alignment (MA) over a 15-task cohabitation session for both DeepSeek-V3.2 and Gemini 3 Flash backbones. This visualization provides a session-level view of how different methods acquire personal preference evidence over time.

Recorded Preference-State Trends. The explicit-state configurations generally show increasing MA over the reported sequence, while several ordinary configurations remain near the reported prior-level readout. This visual difference describes recorded state estimates. It does not establish that a model with a flat external vector fails to encode user information internally. The final values should be read from Table 2, not inferred from approximate visual endpoints or a universal claim of scores above 0.90.

Backbone-Conditional Curves. The plotted trajectories differ across DeepSeek-V3.2 and Gemini 3 Flash. Smoothness can reflect trajectories, averaging, sample coverage, or inference settings as well as extraction behavior. Without trace-matched inputs and repeated-run statistics, the curves do not isolate evidence quality or justify ranking intrinsic backbone reliability.

Readout and Termination Scope. Near-prior curves restrict conclusions to the recorded interface state. An unspecified common probe cannot establish the absence of unstructured learning. The description also notes early termination for a Hyp-Minds trajectory, but the report does not specify

Algorithm 1: Preference-State Workflow for Personal Preference Realization

Input: Sequential Interaction Session $\mathbf{c} = (t_1, \ldots, t_L)$, Prior Belief α_{init}, Parameters μ, ρ
Output: Estimated User Profile $\hat{z}$

```
// Initialize persistent memory
α_mem ← α_init;
for each task t_k ∈ c do
    // Initialize fast belief for current task
    α ← α_mem;
    // Execute task using current belief
    Trace τ ← ExecuteAgent(t_k, α);
    // -- Retrospective Decision Mining --
    P ← MineDiagnosticDecisions(τ);
    for each decision p ∈ P do
        Selected Option ô ← p.selected_option;
        // -- Outcome-to-Evidence Mapping --
        Attribution Window W(p) ← IdentifyWindow(τ, p);
        Signals E ← CollectSignals(τ, W(p));
        Evidence (u, c) ← MapToEvidence(E);
        // -- Fast Update (Eq. 6) --
        α ← α + μ · c · u;
    end
    // -- Slow Update (Eq. 7) --
    α_mem ← ρ · α + (1 − ρ) · α_mem;
    // Update Preference Estimation
    ẑ ← α_mem / Σ α_mem;
end
return ẑ
```

censoring and survivor aggregation. Apparent late-session improvement must therefore not be interpreted as a matched comparison over guaranteed identical completed sessions.

The temporal plots make explicit-state evolution inspectable. They do not replace a common readout, matched behavior tests, or verification of how missing and terminated trajectories enter each aggregate. The measurement qualifications in Section 5.2 apply throughout this appendix.

D.2 Performance at the Last Instruction

Figure 10 compares end-of-session performance across two backbone models (DeepSeek-V3.2 and Gemini 3 Flash) and four planning paradigms. This cross-sectional view complements the temporal curves by showing which capabilities are visible at the final task of each cohabitation session.

Configuration-Level State Readouts. Across the listed planner families, +E-ToM is associated with higher Final MA and lower reported L_2. This shows that the reference state mechanism can be attached to different planner families. It does not isolate its representation from structured inputs, extra calls, or generic persistent memory, and is not a representation-independent proof of PPR.

Concentration Is Not a Quality Ranking. The reported ordinary configurations remain at $\alpha_0 = 4.0$, while +E-ToM reports larger values. This comparison concerns the presence and scale of the pseudo-count state, not all uncertainty in the underlying model. Larger values for one backbone do not prove cleaner evidence, more accurate calibration, or better reasoning.

Task Execution Remains a Separate Measurement. The reported task scores do not increase uniformly with the state-estimation scores. This observation does not demonstrate statistical independence or a desirable trade-off: the inconsistent SR/SPL pairs require reconciliation, and preference benefits must be demonstrated on valid executed choices rather than inferred from state accuracy alone.

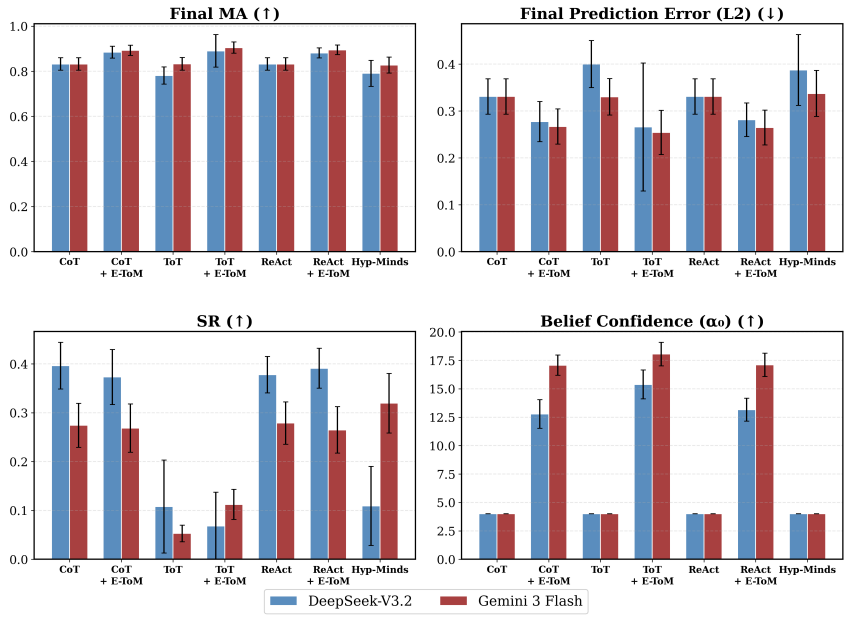


Figure 10: End-of-session results on CoHabit Arena (mean ± 95% CI over cohabitation sessions).

D.3 Comparative Analysis of Preference-State Maintenance Representation and Accumulation

Figure 10 compares recorded estimates for different configurations. The following distinctions describe their representations; they are not causal explanations established by budget- and information-matched controls.

Context-Based Reasoning (CoT, ReAct, ToT). These planner families can condition on prior interaction in the LLM context. Context dilution is a possible explanation for limited adaptation, but has not been isolated here. ToT adds branching search; its low reported task success does not support calling that search a demonstrated improvement in this arena. Absence of an explicit Dirichlet concentration must not be equated with absence of memory.

Natural-Language Hypotheses (Hyp-Minds). The original Hypothetical Minds framework maintains and evaluates hypotheses (Cross et al., 2025). Our adapted configuration lacks the E-ToM axis state, but its constant recorded α_0 is not a measurement of uncertainty in the original method. Without a common preference readout and a precise adaptation specification, the results cannot establish that textual hypotheses are inherently unsuitable for personalization.

Explicit Preference-State Maintenance (+E-ToM). The external state is persistent and exposes a numerical mean and concentration. These are useful diagnostic properties, not evidence that this representation is necessary or superior to other persistent memories. Comparisons with episodic retrieval, textual summaries, and simple weighted accumulation would separate representation from persistence and supplied axis information. Those comparisons are not part of the present results.

D.4 Temporal Dynamics of Component Contributions

Figure 11 provides a longitudinal view of the ablation study, showing Prediction Error (L_2) and preference-state confidence (α_0) across a 15-task cohabitation session. This analysis shows how different components affect the reliability and stability of PPR evaluation.

Frozen-State Control. The no-update variant retains its explicit prior, with no change in ΔL_2 or concentration. This is a negative control for the implemented state update, not a universal statement that preference acquisition requires this update rule. Its reported Ep@Thr cannot be reconciled with the positive-movement condition.

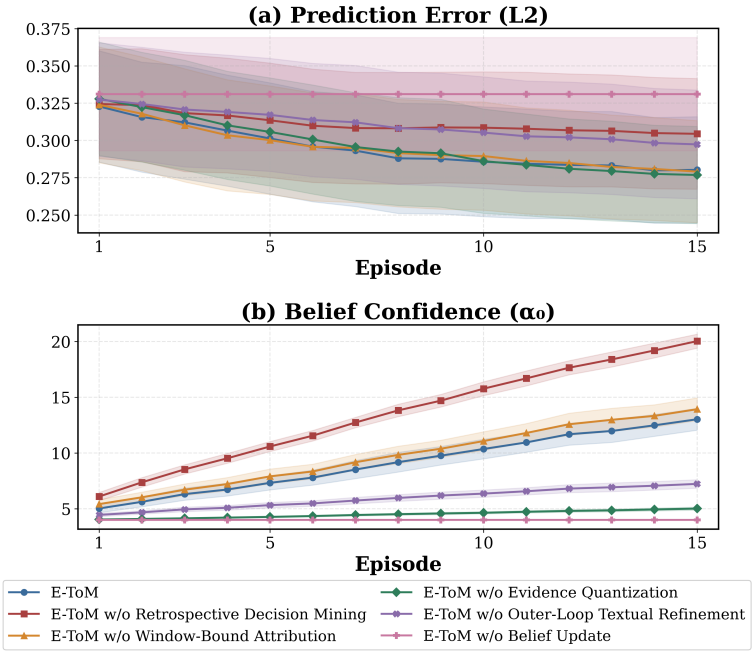


Figure 11: Temporal dynamics in the ablation study (mean ± 95% CI over cohabitation sessions). We denote DeepSeek-V3.2 as "DS" and Gemini 3 Flash as "Gem".

Mining and Evidence Accumulation. The no-mining row has larger concentration but worse Final MA than the full row. This illustrates why mean accuracy and count mass must be distinguished. It does not establish calibrated overconfidence, and missing behavior coverage prevents a conclusion about its executed preference realization.

Quantization Does Not Uniformly Improve Accuracy. The no-quantization row has lower concentration but a slightly higher Final MA and a more negative ΔL_2 than the full row. Thus the current point estimates do not establish that raw signals are unusable without this mapping, or that increased concentration is a benefit. A calibration assessment would be needed for an uncertainty claim.

Attribution and Refinement Must Be Read Separately. Removing refinement reports lower Final MA and concentration. Removing the window instead reports concentration 13.9, above the full row's 13.0, and better point estimates on some other metrics. The results therefore do not support a shared claim that both removals reduce confidence growth. Unspecified refinement implementation and unreconciled convergence values also preclude mechanism-level and speed conclusions.

The ablations reveal different reported state patterns, not uniform dominance of the full pipeline. The interpretation is restricted to the recorded estimates and excludes unverified calibration, causal component necessity, and behavioral superiority. Additional controlled tests would be needed to select a general-purpose personalization design from these observations.

D.5 Compute Resources

The reported experiments were run on a local compute server with 8 NVIDIA V100 GPUs, 32GB memory per GPU, and 376GB system memory. DeepSeek-V3.2 inference used the official DeepSeek API, while Gemini 3 Flash inference used an Azure-hosted endpoint. Local compute was therefore used primarily for simulator execution, orchestration, logging, metric computation, and artifact generation, while LLM inference was served by external API backends. The full reported experimental suite, including the main evaluation and ablation runs, required approximately 48 wall-clock hours under this setup.

These resource totals do not specify per-method token counts, candidate numbers, generation parameters, endpoint versions, or retry and termination policies. They therefore cannot establish compute

parity or quantify the incremental cost of E-ToM. No missing cost measurement is inferred from the hardware inventory or wall-clock total.

E CASE STUDY

The cases below retain the reported tasks, choices, signals, and measurements to illustrate the diagnostic representation. They are not independent validation of behavioral improvement. In particular, an input–output trace, attribution rule, and numeric evidence mapping are not interchangeable with an empirical counterfactual test.

These examples make parts of the intended choice–consequence–reaction relation inspectable. Their numerical summaries do not establish the correctness of the underlying generator, completeness of the matched alternatives, or a causal change in future behavior. Where an example conflicts with task feasibility or retrospective timing, that conflict is stated explicitly rather than repaired by inventing a new trajectory.

The three illustrations concern diagnostic alternatives, attributed state updates, and a recorded preference trajectory. They provide examples of the schema, not controlled evidence that each component is necessary, that the same user behaves consistently across contexts, or that a learned profile causes a particular later action.

A stronger behavioral test would keep the future task fixed and compare correct, absent, or mismatched user histories. Such a test is not reported here, so differences between illustrative users cannot by themselves distinguish personalization from sampling or task-context differences.

E.1 PREFERENCE-DIAGNOSTIC CHOICES EXPOSE PERSONAL PREFERENCE EVIDENCE

The toy-placement example reports different choices and state changes for different synthetic users. It does not include matched histories, repeated sampling, or a controlled future task. The example illustrates how records are represented; it does not establish that the choice differences were caused by acquired preferences.

To quantify this process, we track changes in two metrics: Preference Error (L_2), which measures the Euclidean distance between the predicted and ground-truth personal preference profiles (Eq. 14), and Preference-State Confidence (α_0), which records the pseudo-count concentration rather than validated confidence (Eq. 15). Detailed formal definitions are provided in Appendix B.2.4.

Case Study 1: Same Task and Preference-Diagnostic Choice, Different Users and Choices

In the same toy-placement task and facing the same placement choice, the agent makes different placements when cooperating with different users. The different outcome signals provide evidence across preference dimensions and drive preference-state updates.

Instruction: Put the toy animal, toy vehicle, and toy swing on the bedroom shelves.
Preference-Diagnostic Choice Type: Placement Choice
Option A: Put all the toys on shelf A.
Option B: Put all the toys on shelf B.
Option C: Put some of the toys on shelf A and the others on shelf B.

Divergent choices with different outcome signals under the same preference-diagnostic choice.

User	Archetype	Opt	Outcome Signal (Preference Dimensions)	$\Delta L_2 \downarrow$	$\Delta\alpha_0 \uparrow$
U1	Multi-tasking Routine	**B**	Veto (*Time Sensitivity*)	+0.0019	+0.22
U2	Tidy & Low-Disturbance	**A**	Acceptance (*Privacy Preference*)	-0.0147	**+0.49**
U3	Autonomous Hands-off	**C**	Veto (*Time Sensitivity*)	**-0.0165**	+0.24

E.2 CAUSAL ATTRIBUTION WINDOWS MAKE PERSONAL PREFERENCE REALIZATION EVIDENCE TRACEABLE

Window $W(p)$ assigns outcomes to earlier choices by a temporal rule. The retained example below has two validity limits: its kitchen-cabinet instruction is not equivalent to the living-room or

unrestricted-cabinet options, and retrospective attribution cannot alter a completed task's success. The example is therefore not used as evidence that the window causes a task-completion improvement.

Case Study 2: Avoiding Overconfident but Poor Execution

This retained example reports completion values of 0.9 and 0.1 under two attribution labels. Their run linkage and evaluation unit are not documented, so they cannot be interpreted as a change to the same completed interaction or as a validated causal effect. The instruction requires the kitchen cabinet, while the listed alternatives do not both preserve that requirement. The numeric entries are preserved as reported, but this example does not establish valid within-task preference realization.

Instruction: Move the toy airplane, toy bee, and plant saucer from the bedroom table to the shelves. Place them next to each other on the shelves. After that, move them to the kitchen cabinet and place them next to each other on the cabinet.
Preference-Diagnostic Choice Type: Placement Choice
Option A: Put the toys and plant saucer into the cabinet in the living room.
Option B: Put the toys and plant saucer into any cabinet in the house freely.
Observed signal: Veto directing the agent to choose Option B to reduce time cost.

Same interaction, different attribution.

	With $W(p)$	Without $W(p)$
Evidence confidence c	**0.950**	0.243
Activated dimensions	*Time Sensitivity*	*Privacy Preference, Time Sensitivity, Preference for Shorter Routes, Tolerance for Disturbance*
Preference-state concentration α_0	6.91	**7.92**
Task completion	**0.9**	0.1

E.3 Embodied Outcome Signals Drive Preference-State Updates

Veto, correction, and acceptance categories can supply synthetic evidence for the reference update. Case Study 3 plots one recorded state trajectory. It does not include a matched no-history execution or independent future-choice evaluation, so decreasing vector error and increasing concentration are not alone evidence of successful behavioral realization.

Case Study 3: Salient Signals Drive Preference-State Updates

Over a 15-task cohabitation session, the agent uses salient outcome signals (veto, acceptance, and correction) to update its preference-state estimate, reducing preference error while increasing confidence.

User Archetype: Step Minimizer
User Preference Axes: Preference for Shorter Routes = 0.61 Tolerance for Disturbance = 0.26 Time Sensitivity = 0.13
Cohabitation Session Length: 15 tasks

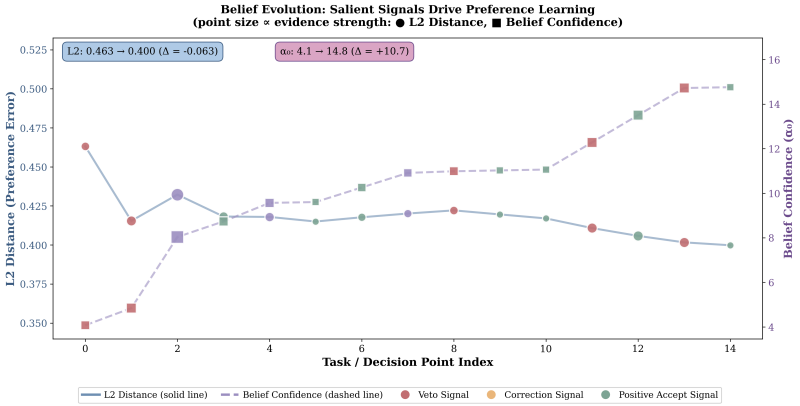


Figure 12: Case Study 3: Over a 15-task cohabitation session, salient embodied outcome signals drive preference-state updates—preference error (L_2) decreases while preference-state confidence (α_0) steadily increases.

F Limitations and Future Work

CoHabit Arena is a controlled synthetic diagnostic, not a population model. Fixed within-session personas omit context-dependent preferences, clarification, drift, and multi-user negotiation. No fully specified independent human audit is reported.

The evidence is limited by unreconciled SR/SPL and Ep@Thr measurements, differing nominal full configurations, missing baseline regret and coverage, and incomplete information, readout, and compute controls. Higher recorded MA therefore does not establish unassisted inference, causal memory benefits, calibrated confidence, or behavioral superiority.

Mixed-type encoding, utility scales, matching, signal generation, and refinement also require implementation records. Finite choices may not identify every reference vector, and reused scenes or personas do not establish held-out generalization. These limits are detailed in Sections 5.2–5.4 and Appendices B–E; retaining measurements and limiting their interpretation do not resolve them.

Future work should extend PPR evaluation in four directions. First, richer preference models should capture context-dependent trade-offs, preference composition, uncertainty calibration, and temporal drift while preserving enough structure for diagnosis. Second, attribution mechanisms should move beyond fixed windows toward causal and counterfactual credit assignment over longer interaction histories, including ambiguous or delayed reactions. Third, the arena should be expanded to more diverse scenes, tasks, personas, interaction modalities, and multi-user settings where preferences may conflict. Fourth, sim-to-real and human-in-the-loop studies are needed to test whether preference-diagnostic choices, embodied outcome signals, and PPR metrics remain valid under real sensor noise, safety constraints, privacy requirements, and naturally evolving user behavior.