WHEN TO SHAPE A TEAMMATE: INDUCING DECISION-RELEVANT BELIEF CHANGES FOR BETTER COOPERATIVE FUTURES

Anonymous authors
Paper under double-blind review

ABSTRACT

Theory of Mind studies how agents infer others' hidden states from behavior, while legibility studies how behavior can shape such inference. In cooperative tasks, however, interpretation itself is not the objective: what matters is how it changes a teammate's behavior and future team outcomes. This raises a central question: when should an agent intentionally sacrifice its own immediate optimality to shape another agent's future behavior? A large belief change may have little effect when the teammate's decision remains unchanged, whereas a small change near a decision boundary can redirect its behavior and lead to a different cooperative future. We develop a closed-loop planning approach that evaluates candidate actions through the teammate responses and future joint outcomes they induce. The agent predicts how its behavior changes the teammate's belief and subsequent response, compares the resulting joint future with task-direct behavior, executes one action, and replans from the realized joint state. Across 400 held-out GridWorld scenes and 87 PARTNR-derived household instances, our approach reaches 97.0% and 62.1% complementary-task accuracy, compared with 95.5%/96.25% and 36.8%/46.0% for open- and closed-loop belief-based cooperative planning baselines.

1 INTRODUCTION

Cooperation sometimes requires an agent to act for what its behavior will make possible later, rather than only for what the action accomplishes immediately. Consider a soccer player who makes a run toward the sideline, not because that position is itself valuable, but because the run draws a defender away and creates space for a teammate. Similar situations arise in embodied collaboration. Moving toward an object, taking a short detour, or temporarily delaying a task can change what a teammate believes the agent intends to do, which may in turn change how the teammate behaves.

Theory of Mind (ToM) studies how latent goals, beliefs, or intentions can be inferred from observed behavior (Rabinowitz et al., 2018; Sclar et al., 2022). Legible motion and observer-aware planning consider the complementary problem of choosing behavior while accounting for how that behavior will be interpreted (Dragan et al., 2013; Dragan & Srinivasa, 2013; Miura & Zilberstein, 2021). These perspectives establish that an agent can influence another agent's inference through its behavior. In cooperation, however, changing a teammate's belief is not itself the objective. A belief change matters only when it alters the teammate's behavior, and that behavioral change is useful only when it leads to a better future joint outcome.

This leads to a central question: when should an agent intentionally sacrifice its own immediate optimality to shape another agent's future behavior? Belief change alone cannot answer this question. Two substantially different beliefs may still lead to the same decision, while a small belief change near a decision boundary can redirect the teammate toward a different object, route, or subtask. At the same time, changing the teammate's decision is not sufficient by itself. The resulting response is useful only when it leads to a better cooperative future. Whether shaping is worthwhile therefore depends on the full consequence of an action: how it changes the teammate's belief, whether that change affects the teammate's behavior, and how the resulting response changes the future joint state.

This perspective separates two quantities that are often implicitly conflated. The first is how strongly an action changes a teammate's belief. The second is how valuable the resulting change is for

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.

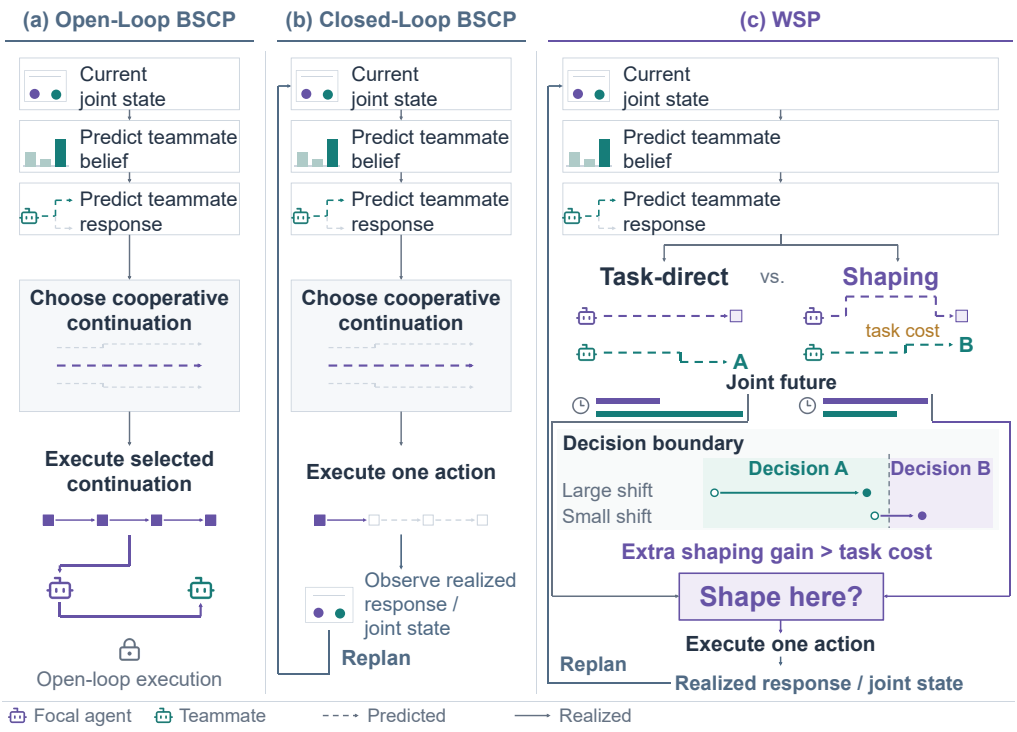

Figure 1: System-level comparison. Open-Loop BSCP plans through teammate belief and predicted response without replanning from the realized response. Closed-Loop BSCP replans after observing what the teammate actually does. WSP further treats when to shape the teammate's behavior as a state-dependent planning decision, based on the resulting future joint outcome.

cooperation. A highly informative action may have little practical effect if the teammate would make the same decision anyway. Conversely, a modest belief change can be important if it crosses a decision boundary and redirects later behavior. Even then, the value of shaping depends on where it occurs: at some states, the teammate's belief is easier to shift, so a smaller deviation can induce the same behavioral change, or even produce a more distinguishable belief update with lower cost.

We therefore evaluate candidate actions through the cooperative futures they induce. For each candidate, the agent predicts how its observable behavior changes the teammate's belief and subsequent response, then evaluates the joint continuation that follows. Shaping behavior is compared with task-direct behavior rather than rewarded for producing a large belief change by itself. This makes the timing of shaping part of the planning problem: an agent may shape immediately, wait until a later state where the same influence becomes more useful, or avoid shaping when it would not improve the joint task.

Teammate responses are also uncertain and may unfold over multiple steps. A belief update does not necessarily change the teammate's next action immediately, and a predicted response may differ from the response that is actually observed. We therefore execute this reasoning in closed loop. The agent plans prospectively over candidate continuations but executes only the current action, observes the teammate's realized response and the resulting joint state, and then replans from what actually occurred.

We develop When-to-Shape Planning (WSP) to decide when shaping a teammate is worthwhile. At each candidate state, WSP predicts how an action changes the teammate's belief and subsequent behavior, evaluates the resulting joint continuation against task-direct behavior, and uses shaping only when it leads to a better cooperative future. To isolate this decision, we compare WSP with belief-based cooperative planning baselines. Open-Loop BSCP reasons over teammate belief and predicted response but follows the selected continuation, while Closed-Loop BSCP replans after

observing the teammate's realized response. WSP uses the same belief and response reasoning, but additionally treats the choice of when to shape the teammate as a state-dependent planning decision.

Figure 1 highlights the additional decision considered in our setting. Predicting how a teammate will respond and adapting after that response are useful capabilities, but neither determines when the agent should deliberately shape the teammate's behavior. This decision depends first on whether the induced belief change is sufficient to alter the teammate's future behavior, and then on whether the resulting joint continuation is preferable to task-direct behavior at the current state.

We study this problem in two cooperative settings with asymmetric access to coordination-relevant information. In both settings, one agent has access to information that helps determine how the joint task should be divided, while its teammate must infer the appropriate complementary subtask from physical interaction. A controlled GridWorld construction task makes the relation between belief change, teammate decision, and joint outcome directly observable. A longer PARTNR-derived household setting (Chang et al., 2025) introduces navigation, delayed commitment, and longer interaction horizons, making the choice of when to shape and the need for replanning more consequential.

Our contributions are threefold:

- We develop When-to-Shape Planning (WSP) to address when shaping a teammate is worthwhile. WSP compares the future joint continuation induced by shaping with that of task-direct behavior, providing a state-dependent criterion for deciding when shaping should occur.
- We identify decision relevance as the key link between belief change and cooperative value. The magnitude of a belief change alone is insufficient: it matters only when the change alters the teammate's future behavior, making decision boundaries central to evaluating shaping actions.
- We enable WSP to operate under uncertain and potentially delayed teammate responses through closed-loop replanning. WSP predicts candidate responses, executes only the current action, observes the realized response, and replans from the resulting joint state.

Across 400 held-out GridWorld scenes and 87 PARTNR-derived household instances, our approach reaches 97.0% and 62.1% complementary-task accuracy, compared with 95.5%/96.25% and 36.8%/46.0% for matched open- and closed-loop belief-based cooperative planning baselines. Matched counterfactuals, timing controls, and response-channel ablations further show that the gains are strongly state dependent and arise through changes in teammate behavior and subsequent coordination rather than from the shaping action's direct physical effect alone.

2 Related Work

Theory of Mind and partner modeling. Machine ToM infers latent mental states from observed behavior and uses them to predict or reason about other agents (Rabinowitz et al., 2018; Sclar et al., 2022; Zhu et al., 2021; Shu et al., 2021). Recent work represents recurring behavior through explicit hypotheses or programs (Cross et al., 2025; Jha et al., 2026) and studies richer latent factors such as risk, attention, and open-ended goals (Smith & Zhang, 2025; Long et al., 2025; Ghose et al., 2026). SimpleToM emphasizes that correctly inferring a latent state and correctly predicting the behavior induced by that state are different problems (Gu et al., 2026). Our setting changes the role of belief from an inference target to a decision-relevant intermediate variable: belief changes matter only when they cross a teammate's decision boundary.

Legible behavior and observer-aware planning. Legible motion chooses behavior that makes an intended goal easier for an observer to infer (Dragan et al., 2013; Dragan & Srinivasa, 2013; Huang et al., 2017; Strouse et al., 2018). Observer-aware sequential decision making similarly includes observer belief in planning (Miura & Zilberstein, 2021). These works establish that behavior can be designed around how it will be interpreted. WSP focuses on a different question: an informative action is useful only when the induced belief change alters the teammate's future behavior and the resulting response leads to a better future team outcome.

Belief-space and interactive cooperative planning. Cooperative decision making with hidden partner state naturally admits belief-space, partially observable, and interactive formulations (Bernstein et al., 2002). Prior work goes beyond passive prediction by explicitly accounting for how an agent's own actions can affect another agent's behavior, including interactive driving (Sadigh et al., 2016), cooperative inverse reinforcement learning (Hadfield-Menell et al., 2016), and adaptive human-robot interaction (Nikolaidis et al., 2017; Sengadu Suresh et al., 2025). These works establish that cooperative planning can account for how an agent's own actions affect another agent's response, and provide the belief-aware prediction and adaptation capabilities used in interactive decision making. Our focus is instead on when such influence is worthwhile: WSP evaluates shaping as a state-dependent decision according to the teammate behavior and future joint outcome it induces.

Multi-agent and embodied collaboration. Multi-agent cooperation has been studied through decentralized control, ad hoc teamwork, social influence, learning-aware policies, and adaptation to unseen partners (Stone et al., 2010; Barrett et al., 2017; Jaques et al., 2019; Meulemans et al., 2025; Jha et al., 2025). Embodied work increasingly studies online adaptation and physical coordination (Chaudhary et al., 2026; Scofano et al., 2025; Grannen et al., 2025), while predictive models such as COMBO model multi-agent physical dynamics (Zhang et al., 2025). Habitat 3.0 and PARTNR provide richer environments for studying such embodied cooperation (Puig et al., 2024; Chang et al., 2025). We study WSP in this setting, where physical actions can change both the task state and the future behavior of other agents within an episode.

3 When-to-Shape Planning

3.1 Cooperative Decision Model

We formulate cooperation as a decentralized partially observable decision process

$$\mathcal{M} = \langle \mathcal{N}, \mathcal{X}, \{\mathcal{A}_i\}_{i\in\mathcal{N}}, P, \{\mathcal{O}_i\}_{i\in\mathcal{N}}, \Omega, r, \gamma \rangle \,, \tag{1}$$

where $\mathcal{N}$ is the agent set, $\mathcal{X}$ is the joint state space, $\mathcal{A}_i$ and $\mathcal{O}_i$ are the action and observation spaces, P and Ω are the transition and observation models, r is the shared cooperative reward, and γ is the discount factor.

Consider a focal agent i and one of its teammates j. Agent i has a coordination-relevant private variable θ_t^i, such as its selected target, role, or subtask, which is not directly observed by j. Based on its own local interaction history h_t^j, teammate j maintains

$$b_t^{j\leftarrow i}(\theta) = \Pr(\theta_t^i = \theta \mid h_t^j), \tag{2}$$

where h_t^j contains the observations and actions available to j up to time t, while o_t^j denotes only its current observation. Because agent i's behavior forms part of what j observes, actions taken by i can change this belief.

Crucially, a belief change does not necessarily change the teammate's decision. For a given interaction history, a range of beliefs may lead teammate j to the same task-relevant decision, such as selecting the same target, route, or subtask. A belief change may therefore have little behavioral consequence if this decision remains unchanged, even when the change itself is large. Conversely, a small belief change near a decision boundary can redirect the teammate's future behavior.

Importantly, crossing a decision boundary need not change the teammate's immediately next primitive action. Two trajectories may remain identical for several steps before diverging through a later target, route, or subtask decision. We therefore evaluate belief change by whether it alters the teammate's future behavior, rather than by its magnitude or immediate action change alone.

A more detailed specification of the cooperative interaction model and belief-update process is provided in Appendix A.

3.2 State-Dependent Shaping Criterion

Let ξ denote the planning information available to the focal agent, and let a be a candidate focal action. We use $Q(\xi, a)$ to denote the cooperative continuation value after executing a. To isolate the value

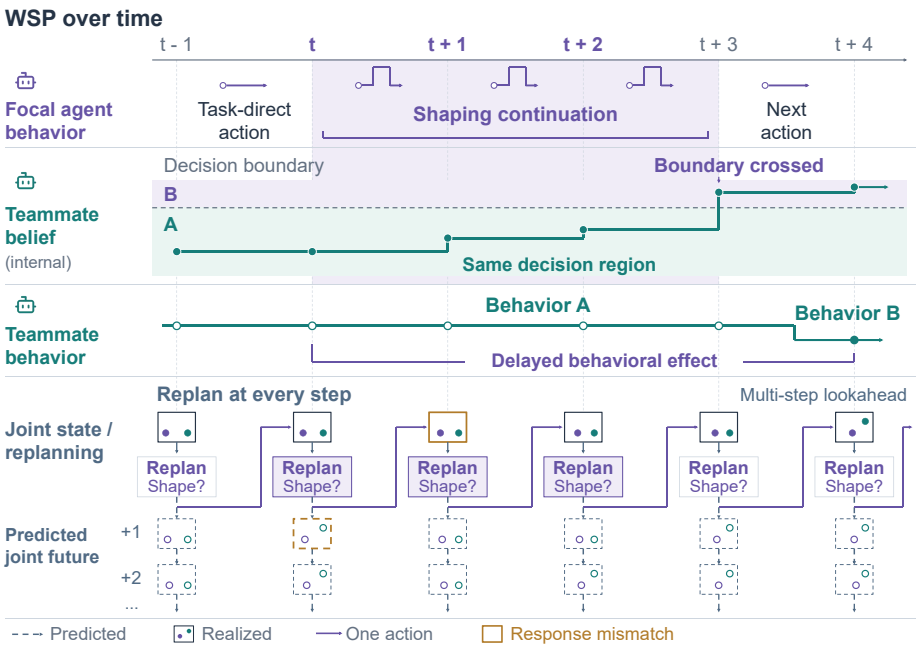

Figure 2: Closed-loop execution of WSP. The focal agent evaluates whether shaping is worthwhile, executes one action, observes the teammate's realized response and the resulting joint state, and replans. The behavioral effect of shaping may appear immediately or only after a later decision boundary is crossed.

created through shaping the teammate, we define a matched continuation $Q^-(\xi, a)$ that preserves the same focal action and immediate physical consequence, while removing the coordination evidence conveyed to the teammate.

The shaping gain of action a is

$$G_{\text{shape}}(a;\xi) = Q(\xi, a) - Q^-(\xi, a), \tag{3}$$

which captures the continuation-value difference caused by the teammate's changed belief and subsequent behavior. If the actual and matched beliefs lead to the same future teammate behavior, then $G_{\text{shape}}(a;\xi) = 0$.

We define the task-direct reference action as

$$a_{\text{task}} \in \arg\max_a Q^-(\xi, a), \tag{4}$$

and define the task cost of choosing a candidate action a instead as

$$C_{\text{task}}(a;\xi) = Q^-(\xi, a_{\text{task}}) - Q^-(\xi, a). \tag{5}$$

This cost measures how much task value is sacrificed when the focal agent deviates from task-direct behavior. Shaping is worthwhile when its additional shaping gain over the task-direct action compensates for this cost:

$$G_{\text{shape}}(a;\xi) - G_{\text{shape}}(a_{\text{task}};\xi) > C_{\text{task}}(a;\xi). \tag{6}$$

The criterion is state dependent: shaping becomes more favorable when the teammate is easier to influence, its belief is closer to a relevant decision boundary, or a useful behavioral change can be induced with lower task cost. A detailed decomposition is provided in Appendix B.

After identifying a shaping action as worthwhile, WSP executes the decision in closed loop. At each interaction state, it evaluates task-direct and shaping continuations, predicts their effects on teammate belief and future behavior, but executes only the current action. The focal agent then observes the teammate's realized response and the resulting joint state and replans from what actually occurred.

This feedback is important because teammate responses may be uncertain or delayed: a shaping action can change belief immediately while affecting behavior only after a later decision boundary is crossed. Figure 2 summarizes this execution process.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We instantiate WSP with fixed Bayesian receivers and response models in two complementary settings. In each task, one agent, denoted *Main*, knows the true complementary assignment $X^\star$, while the other, denoted *Helper*, must infer that assignment from Main's observable physical behavior.

GridWorld uses a two-agent 10×10 construction task with three candidate complementary objects and 400 held-out scenes from nine layouts. Main always delivers the same anchor object to the assembly site, whereas Helper must infer which complementary object to deliver. Because Main's physical task is unchanged across the three candidate assignments, the setting separates task efficiency from the information that Main's motion provides about $X^\star$.

The PARTNR-derived household family contains 87 paired instances from 30 validation episodes, with longer navigation and pre-commitment Helper motion (Chang et al., 2025). Each instance fixes Main's object-to-receptacle subtask and treats the remaining parallel subtasks as Helper candidates. While uncertain, Helper physically follows Main before committing to one candidate. Consequently, a shaping state affects not only what Helper believes but also where Helper will be when the assignment becomes fixed. Appendix C gives the complete task construction and likelihood models.

We report complementary-task accuracy and team completion time

$$T_{\text{finish}} = \max(T_{\text{Main}}, T_{\text{Helper}}). \tag{7}$$

For each candidate continuation ρ, the controller uses the lexicographic objective

$$\mathcal{V}(\rho) = \left(P_{\text{correct}}(\rho), -\mathbb{E}[T_{\text{finish}}(\rho)]\right), \tag{8}$$

so correct complementary assignment is optimized first, while expected completion time breaks ties among equally correct continuations.

Both receivers use

$$P(a \mid X, s) = qP_{\text{exec}}(a \mid s) + (1-q)P_{\text{int}}(a \mid X, s), \tag{9}$$

with Bayesian posterior updates and $q = 0.9$ in the headline setting. GridWorld uses a Bayes-rational Helper over object choices. Household Helper follows Main with a one-step lag before commitment, commits at posterior threshold $\eta = 0.90$ or a 20-step deadline, and then executes the selected complementary task. These decision rules create ranges of posterior beliefs that leave Helper's choice unchanged and boundaries at which its target or commitment changes.

We compare six methods on identical frozen manifests. Task-Driven follows Main's task-efficient policy while Helper still infers from information naturally present in that trajectory. Open-Loop BSCP selects a belief-space cooperative continuation and executes it open loop. Closed-Loop BSCP uses the same teammate model but replans from the realized teammate response and joint state. Timing-Aware Open-Loop BSCP, abbreviated Timed Open-Loop, receives the same candidate shaping states as WSP but executes the selected continuation open loop. WSP selects whether and when to shape the teammate and replans from each realized response. Perfect Information gives $X^\star$ to Helper at the start.

For computational efficiency, candidate shaping states are proposed from local peaks of the likelihood-separation score $E(s)$, together with the initial state and a no-shaping option. WSP then evaluates these candidates according to their downstream cooperative continuation value. We later remove this proposal restriction by exhaustively evaluating every valid onset. All uncertainty estimates use 2,000 paired bootstrap resamples with seed 42, clustered by GridWorld layout or Household source episode.

4.2 OVERALL COORDINATION PERFORMANCE

Table 1 shows two different regimes. GridWorld is close to saturation once belief-aware planning is introduced: Open-Loop BSCP reaches 95.5% accuracy, Closed-Loop BSCP reaches 96.25%,

Table 1: Primary coordination results under the correctness-first objective. Accuracy is complementary-task accuracy; lower T_{finish} is better.

	GridWorld		Household	
Condition	Acc. (%)	T	Acc. (%)	T
Task-Driven	31.5	38.97	34.5	158.75
Open-Loop BSCP	95.5	22.99	36.8	156.12
Closed-Loop BSCP	96.25	22.95	46.0	149.19
Timed Open-Loop	96.25	22.86	59.8	134.76
WSP	97.0	22.83	62.1	134.43
Perfect Information	100.0	21.44	100.0	112.15

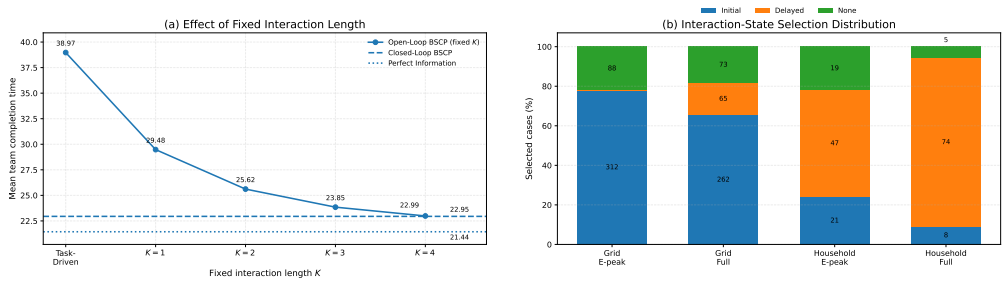

Figure 3: Temporal analysis of shaping. The left panel varies the fixed Open-Loop BSCP interaction length in GridWorld. The right panel compares shaping states proposed by local peaks of $E(s)$ with exhaustive evaluation of every valid onset.

and WSP reaches 97.0%. Open-Loop to Closed-Loop improves accuracy by only 0.75 percentage points, while the paired completion-time change is -0.04 steps with layout-clustered 95% CI $[-0.084, -0.006]$. The short task and strong evidence carried by the four-step prefix leave relatively little room for further shaping or replanning.

Household leaves substantially more room for both. Closed-Loop BSCP raises accuracy from 36.8% to 46.0%, while Timed Open-Loop reaches 59.8% and WSP reaches 62.1%. Relative to Closed-Loop BSCP, WSP improves accuracy by 16.09 points with 95% CI $[7.79, 27.16]$ and reduces mean completion time by 14.76 steps with CI $[-25.16, -7.26]$. The results show that choosing when to shape and replanning from the realized response provide complementary benefits in longer interactions.

Perfect Information remains well ahead of WSP at 100% accuracy and 112.15 completion time, showing that the Household setting is not saturated. Complete paired statistics are reported in Appendix E.

4.3 State-Dependent Timing of Shaping

The left panel explains why GridWorld is already strong under open-loop belief-space planning. Accuracy rises from 61.0% at $K = 1$ to 79.5%, 89.75%, and 95.5% at $K = 2, 3, 4$, while mean completion time falls from 29.48 to 22.99. The $K = 4$ condition exactly reproduces the headline Open-Loop BSCP result. Among scenes that are already correct, completion time changes much less, indicating that the longer prefix primarily improves complementary assignment rather than merely shortening already-correct trajectories.

The right panel directly examines when shaping should occur. In GridWorld, the primary WSP proposal selects 312 initial and 88 no-shaping cases, whereas exhaustive onset search selects 262 initial, 65 delayed, and 73 no-shaping cases. Full search improves accuracy from 97.0% to 100% and reduces mean completion time from 22.83 to 22.20.

The timing effect is stronger in Household. The primary proposal selects 21 initial, 47 delayed, and 19 no-shaping cases, whereas exhaustive search selects 8 initial, 74 delayed, and 5 no-shaping

Table 2: Matched-action intervention. "Drop" is the decrease in correct realized Helper target assignment after neutralizing coordination evidence. GridWorld reports the strict physical-match subset.

Environment	N	Normal (%)	Neutral (%)	Drop (pp)
GridWorld	391	94.63	34.27	60.36
Household	87	62.07	37.93	24.14

Table 3: Controlled comparison of Timed Open-Loop and WSP. Different onset counts instances in which replanning changes the selected shaping onset. Lower completion time is better.

Environment	Timed Acc. (%)	WSP Acc. (%)	Timed T	WSP T	Different onset
GridWorld	96.25	97.0	22.86	22.83	0 / 400
Household	59.8	62.1	134.76	134.43	21 / 87

cases. Accuracy increases from 62.1% to 69.0%, and mean completion time decreases from 134.43 to 129.03. The 6.90-point accuracy gain has episode-clustered 95% CI $[2.25, 12.90]$, and the 5.40-step completion-time reduction has CI $[1.81, 10.98]$. These results show that the most useful shaping state is often delayed and need not coincide with the state carrying the strongest local belief signal. Full enumeration statistics are reported in Appendix E.3.

4.4 Shaping Through Teammate Response

We test whether the benefit of shaping is transmitted through the teammate's changed belief and response rather than through Main's direct physical action. For each matched pair, Main's focal physical behavior is held fixed. In the normal condition, Helper updates its belief from that behavior; in the neutral condition, the same focal behavior carries no task-dependent coordination evidence. The rollouts may diverge later only because Helper responds differently.

Neutralizing the coordination evidence while holding Main's focal physical behavior fixed reduces correct Helper assignment by 60.36 percentage points in GridWorld and 24.14 points in Household. In GridWorld, the layout-clustered 95% CI is $[55.80, 65.72]$, with 238 normal-correct/neutral-wrong cases and only two reverse transitions. Completion time is unchanged on the strict physical-match subset, as expected when Main's physical effect is preserved.

In Household, the 24.14-point decrease has episode-clustered 95% CI $[13.95, 35.90]$. The changed response also propagates into physical cost: the neutral condition is 14.22 steps slower on average, with 95% CI $[5.39, 25.21]$. These matched interventions show that the cooperative benefit is transmitted through Helper's changed belief and response rather than through Main's direct physical effect alone. Appendix E.5 reports the full transition counts and replay checks.

4.5 Closed-Loop Replanning under Realized Responses

Timing-Aware Open-Loop BSCP uses exactly the same candidate shaping states as WSP but evaluates and executes the selected continuation open loop. This comparison controls for access to the same shaping opportunities and isolates the additional effect of replanning from the response that actually occurs.

In GridWorld, Timed Open-Loop and WSP choose the same onset in all 400 scenes: 312 initial-state interactions and 88 no-shaping cases. WSP nevertheless corrects three additional scenes with no reverse transition. Mean completion time changes only from 22.8575 to 22.8275, with a layout-clustered 95% interval of $[-0.080, 0]$ for the paired difference. This small gain is consistent with the near-saturated open-loop result.

Household separates shaping timing from realized-response replanning more clearly. Timed Open-Loop selects 5 initial, 47 delayed, and 35 no-shaping cases, whereas WSP selects 21 initial, 47 delayed, and 19 no-shaping cases. Replanning changes the selected onset in 21 of 87 instances; on these cases, WSP is 6.10 steps faster on average. By contrast, across the 66 same-onset cases, the mean paired difference $T_{\mathrm{WSP}} - T_{\mathrm{Timed}}$ is $+1.50$ steps. This indicates that the main value of realized-

Table 4: Household complementary-task accuracy at the headline receiver and at the belief-insensitive boundary.

q	Task	Open	Closed	Timed Open	WSP
0.90	34.48	36.78	45.98	59.77	62.07
1.00	34.48	34.48	34.48	34.48	34.48

response feedback is not to uniformly shorten trajectories, but to change which shaping continuation is preferred when the observed interaction differs from the open-loop prediction. Complete paired statistics are reported in Appendix E.6.

4.6 Boundary of the Shaping Mechanism

The receiver mixture provides a direct boundary for the shaping mechanism. When the actual and assumed mixture weights are both $q = 1$, Eq. (9) becomes independent of X, so Main's observable behavior can no longer change Helper's task belief through the task-dependent component.

At $q = 1$, all non-perfect methods collapse to the Task-Driven accuracy of 34.48%, showing that WSP loses its coordination advantage when observable behavior can no longer affect Helper's task belief. Before this boundary, WSP obtains 60.9, 62.1, 60.9, 62.1, 64.4, and 65.5% accuracy for $q = 0.50, 0.70, 0.80, 0.90, 0.95, 0.99$, respectively.

This result complements the matched-action intervention. The matched-action experiment removes coordination evidence while preserving Main's focal physical behavior, whereas matched $q = 1$ removes the belief-dependent response channel from the receiver itself. Together, the two tests show that WSP's advantage depends on the ability of Main's behavior to alter Helper's later response. The full matched-q sweep is reported in Appendix E.7.

4.7 Robustness to Teammate-Model Mismatch

We finally evaluate robustness to teammate-model mismatch by perturbing Main's prospective teammate model while keeping the executed Helper fixed at the headline parameters. Exact model match gives WSP 62.07% accuracy and a mean completion time of 134.43. Across the tested $\hat{\beta}$ and $\hat{\eta}$ values, accuracy remains between 59.77% and 63.22%, with only small changes in completion time. WSP also maintains 62.07% accuracy for $\hat{q} \in \{0.70, 0.80, 0.90, 0.95\}$.

Even at $\hat{q} = 1$, where Main assumes that its behavior carries no task-dependent belief evidence, the actual Helper still uses $q = 0.9$. WSP therefore degrades without the complete collapse observed when the response mechanism itself is removed, retaining 58.62% accuracy and a mean completion time of 135.10. This separates teammate-model error from mechanism removal: the former changes Main's prediction of Helper, whereas the latter removes the response pathway that makes shaping possible. Full mismatch results are reported in Appendix E.8.

5 Conclusion

We introduced When-to-Shape Planning (WSP) to determine when shaping a teammate is worthwhile in cooperative decision making. Rather than treating belief change as valuable in itself, WSP evaluates whether a candidate action changes the teammate's future behavior in a way that improves the joint continuation enough to justify its task cost. This view makes decision-relevant belief change central: a large belief shift may be inconsequential if the teammate's decision remains unchanged, while a small shift near a decision boundary can redirect future behavior.

Experiments in GridWorld and PARTNR-derived household tasks show that the value of shaping is strongly state dependent. The results further show that selecting when to shape and replanning from the teammate's realized response provide gains beyond standard belief-space cooperative planning. Controlled interventions isolate the benefit transmitted through the teammate's changed belief and response from the focal action's direct physical effect. Overall, effective cooperation depends not only on predicting teammate behavior, but also on deciding when one's behavior should change it.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, literature discovery, and research discussion and ideation. The authors independently verified all cited references, technical content, experimental results, and mathematical arguments.

REFERENCES

Samuel Barrett, Avi Rosenfeld, Sarit Kraus, and Peter Stone. Making friends on the fly: Cooperating with new teammates. *Artificial Intelligence*, 242:132–171, 2017. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2016.10.005. URL https://www.sciencedirect.com/science/article/pii/S0004370216301266.

Daniel S. Bernstein, Robert Givan, Neil Immerman, and Shlomo Zilberstein. The complexity of decentralized control of markov decision processes. *Mathematics of Operations Research*, 27 (4):819–840, 2002. doi: 10.1287/moor.27.4.819.297. URL https://doi.org/10.1287/moor.27.4.819.297.

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 Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 65205–65268, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/a3cf318fbeec1126da21e9185ae9908c-Paper-Conference.pdf.

Paresh Chaudhary, Yancheng Liang, Daphne Chen, Simon Du, and Natasha Jaques. Improving human-ai coordination through online adversarial training and generative models. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 113415–113433, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/b837547467ea02deebdc9677d7f5a0be-Paper-Conference.pdf.

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 Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 6507–6546, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/12f483f624b378f9f3058d8ecd3c7ff5-Paper-Conference.pdf.

Anca Dragan and Siddhartha Srinivasa. Generating legible motion. In *Proceedings of Robotics: Science and Systems*, Berlin, Germany, June 2013. doi: 10.15607/RSS.2013.IX.024.

Anca Dragan, Kenton Lee, and Siddhartha Srinivasa. Legibility and predictability of robot motion. In *Proceedings of 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI '13)*, pp. 301 – 308, March 2013.

Debasmita Ghose, Oz Gitelson, Marynel Vázquez, and Brian Scassellati. Open-ended goal inference through actions and language for human-robot collaboration. In *Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction*, HRI '26, pp. 68–77, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400721281. doi: 10.1145/3757279.3785540. URL https://doi.org/10.1145/3757279.3785540.

Jennifer Grannen, Siddharth Karamcheti, Suvir Mirchandani, Percy Liang, and Dorsa Sadigh. Vocal sandbox: Continual learning and adaptation for situated human-robot collaboration. In Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard (eds.), *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 1–24. PMLR, 06–09 Nov 2025. URL https://proceedings.mlr.press/v270/grannen25a.html.

Yuling Gu, Oyvind Tafjord, Hyunwoo Kim, Jared Moore, Ronan Le Bras, Peter Clark, and Yejin Choi. Simpletom: Exposing the gap between explicit tom inference and implicit tom application in llms. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 50252–50297, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/524ff06d5375d76e93d8490e654a81bf-Paper-Conference.pdf.

Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart J Russell. Cooperative inverse reinforcement learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/c3395dd46c34fa7fd8d729d8cf88b7a8-Paper.pdf.

Sandy Huang, David Held, Pieter Abbeel, and Anca Dragan. Enabling robots to communicate their objectives. In *Proceedings of Robotics: Science and Systems*, Cambridge, Massachusetts, July 2017. doi: 10.15607/RSS.2017.XIII.059.

Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro Ortega, Dj Strouse, Joel Z. Leibo, and Nando De Freitas. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. 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. 3040–3049. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/jaques19a.html.

Kunal Jha, Wilka Carvalho, Yancheng Liang, Simon Shaolei Du, Max Kleiman-Weiner, and Natasha Jaques. Cross-environment cooperation enables zero-shot multi-agent coordination. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 27198–27220. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/jha25b.html.

Kunal Jha, Aydan Huang, Eric Ye, Natasha Jaques, and Max Kleiman-Weiner. Modeling others'minds as code. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 67840–67874, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/6edd46d69ef91f4555d67f7b321d6902-Paper-Conference.pdf.

Qian Long, Ruoyan Li, Minglu Zhao, Tao Gao, and Demetri Terzopoulos. Inverse attention agents for multi-agent systems. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 2776–2793, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/07c7ea6b89aa8381a8bc927e70d67fe0-Paper-Conference.pdf.

Alexander Meulemans, Seijin Kobayashi, Johannes von Oswald, Nino Scherrer, Eric Elmoznino, Blake A Richards, Guillaume Lajoie, Blaise Aguera y Arcas, and Joao Sacramento. Multi-agent cooperation through learning-aware policy gradients. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 45978–46009, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/718a3e764e63a130323c5b5abf4ac332-Paper-Conference.pdf.

Shuwa Miura and Shlomo Zilberstein. A unifying framework for observer-aware planning and its complexity. In Cassio de Campos and Marloes H. Maathuis (eds.), *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pp. 610–620. PMLR, 27–30 Jul 2021. URL https://proceedings.mlr.press/v161/miura21a.html.

Stefanos Nikolaidis, David Hsu, and Siddhartha Srinivasa. Human-robot mutual adaptation in collaborative tasks: Models and experiments. *The International Journal of Robotics Research*, 36 (5-7):618–634, 2017. doi: 10.1177/0278364917690593. URL https://doi.org/10.1177/0278364917690593. PMID: 32855581.

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 B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 15306–15336, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/430894999584d0bd358611e2ecf00b15-Paper-Conference.pdf.

Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew Botvinick. Machine theory of mind. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4218–4227. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/rabinowitz18a.html.

Dorsa Sadigh, Shankar Sastry, Sanjit A. Seshia, and Anca D. Dragan. Planning for autonomous cars that leverage effects on human actions. In *Proceedings of Robotics: Science and Systems*, AnnArbor, Michigan, June 2016. doi: 10.15607/RSS.2016.XII.029.

Melanie Sclar, Graham Neubig, and Yonatan Bisk. Symmetric machine theory of mind. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 19450–19466. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/sclar22a.html.

Luca Scofano, Alessio Sampieri, Tommaso Campari, Valentino Sacco, Indro Spinelli, Lamberto Ballan, and Fabio Galasso. Following the human thread in social navigation. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 47674–47694, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/76b010c4d21f97b7b113383ce7160cbc-Paper-Conference.pdf.

Prasanth Sengadu Suresh, Prashant Doshi, and Bikramjit Banerjee. Adaptive human-robot collaboration using type-based irl. In Silvia Chiappa and Sara Magliacane (eds.), *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pp. 4080–4091. PMLR, 21–25 Jul 2025. URL https://proceedings.mlr.press/v286/sengadu-suresh25a.html.

Tianmin Shu, Abhishek Bhandwaldar, Chuang Gan, Kevin Smith, Shari Liu, Dan Gutfreund, Elizabeth Spelke, Joshua Tenenbaum, and Tomer Ullman. Agent: A benchmark for core psychological reasoning. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9614–9625. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/shu21a.html.

Mason O. Smith and Wenlong Zhang. Risk-sensitive theory of mind: Coordinating with agents of unknown bias using cumulative prospect theory. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 55999–56017. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/smith25b.html.

Peter Stone, Gal Kaminka, Sarit Kraus, and Jeffrey Rosenschein. Ad hoc autonomous agent teams: Collaboration without pre-coordination. In *Proceedings of the AAAI conference on artificial intelligence*, volume 24, pp. 1504–1509, 2010.

DJ Strouse, Max Kleiman-Weiner, Josh Tenenbaum, Matt Botvinick, and David Schwab. Learning to share and hide intentions using information regularization. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.,

2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/1ef03ed0cd5863c550128836b28ec3e9-Paper.pdf.

Hongxin Zhang, Zeyuan Wang, Qiushi Lyu, Zheyuan Zhang, Sunli Chen, Tianmin Shu, Behzad Dariush, Kwonjoon Lee, Yilun Du, and Chuang Gan. Combo: Compositional world models for embodied multi-agent cooperation. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 49996–50019, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/7d03c6bf9f07acb4038eea96c63db52d-Paper-Conference.pdf.

Hao Zhu, Graham Neubig, and Yonatan Bisk. Few-shot language coordination by modeling theory of mind. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12901–12911. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/zhu21d.html.

This appendix provides additional details supporting the main paper. Appendix A expands the cooperative interaction model and the information available to each agent. Appendix B gives the formal analysis of state-dependent shaping, including the matched counterfactual, response invariance, and the when-to-shape decomposition. Appendices C and D describe the experimental environments, teammate belief and response models, and the realization of WSP and the baselines. Finally, Appendix E reports additional quantitative results, causal interventions, mechanism-boundary analyses, and teammate-model robustness.

A Detailed Cooperative Interaction Model

The main text formulates cooperation as a decentralized partially observable decision process

$$\mathcal{M} = \langle \mathcal{N}, \mathcal{X}, \{\mathcal{A}_k\}_{k\in\mathcal{N}}, P, \{\mathcal{O}_k\}_{k\in\mathcal{N}}, \Omega, r, \gamma \rangle . \tag{10}$$

This appendix expands the components needed to describe how observable behavior changes a teammate's belief and how that belief affects later cooperative behavior.

Interaction state and actions. Let $\mathcal{N} = \{1, \dots, N\}$ denote the set of agents. The joint state is

$$x_t = \left(s_t, \theta_t^1, \dots, \theta_t^N\right) \in \mathcal{X}, \tag{11}$$

where s_t is the physical task state and θ_t^k is a coordination-relevant private variable of agent k, such as a selected subtask, role, commitment, or local plan. Agent k chooses $a_t^k \in \mathcal{A}_k$, and

$$\mathbf{a}_t = \left(a_t^1, \dots, a_t^N\right) \tag{12}$$

denotes the joint action. Alternating interaction is represented by assigning a null action to agents that are inactive at a particular interaction event.

Interaction dynamics and local history. The physical state evolves according to

$$P(x_{t+1} \mid x_t, \mathbf{a}_t). \tag{13}$$

After the transition, agents receive local observations according to the observation model Ω. Agent k receives $o_{t+1}^k \in \mathcal{O}_k$, which contains the task information locally available to that agent and may include observable evidence generated by the behavior of other agents.

The local interaction history of agent k is

$$h_t^k = \left(o_0^k, a_0^k, o_1^k, a_1^k, \dots, a_{t-1}^k, o_t^k\right), \tag{14}$$

which records the observations received by the agent and its own executed actions up to time t. Observable behavior of other agents enters through the corresponding observations rather than through direct access to their underlying action variables. Private information held by agent k is maintained separately and is available to its own controller when relevant.

Teammate belief. Consider focal agent i and teammate j. Teammate j maintains a belief about the coordination-relevant private variable of agent i:

$$b_t^{j\leftarrow i}(\theta) = \Pr\left(\theta_t^i = \theta \mid h_t^j\right), \tag{15}$$

where θ is a possible value of θ_t^i. After receiving a new observation, the belief is updated according to

$$b_{t+1}^{j\leftarrow i} = \mathcal{U}_{j\leftarrow i}\left(b_t^{j\leftarrow i}, o_{t+1}^j\right), \tag{16}$$

where $\mathcal{U}_{j\leftarrow i}$ denotes the teammate's belief-update rule. Because o_{t+1}^j can contain evidence generated by agent i's observable behavior, different focal behaviors can induce different teammate beliefs even when the underlying private variable is unchanged. The experiments instantiate $\mathcal{U}_{j\leftarrow i}$ using fixed Bayesian receivers.

Teammate response policy. Each π_k specifies how agent k acts from the information available to its controller. For teammate j,

$$a_{t+1}^j \sim \pi_j\left(\cdot \mid h_{t+1}^j, b_{t+1}^{j\leftarrow i}\right). \tag{17}$$

A belief change matters for cooperation only when it changes the teammate's decision-relevant future behavior. The behavioral effect need not appear in the immediately following primitive action; two trajectories may initially coincide and diverge only at a later target, route, commitment, or subtask decision.

Cooperative objective and planning state. The agents share cooperative reward r, and γ denotes the discount factor used for continuation values. For focal agent i, let ξ_t denote the information state used for planning. It contains the locally available task state, agent i's private coordination information, its interaction history, and the teammate-model state required to predict future responses. It does not require direct access to another agent's private variable.

The belief-update and teammate-response models together describe how observable behavior can affect future cooperation. WSP evaluates candidate continuations under these models, executes only the current action, observes the realized teammate response and resulting state, and replans from the updated information state.

B Formal Analysis of State-Dependent Shaping

This appendix formalizes the matched counterfactual used in the main text, shows when a belief change has no shaping value, and derives the state-dependent criterion for deciding when shaping is worthwhile.

B.1 Planning State and Cooperative Continuation Value

Let ξ_t denote the focal planning state defined in Appendix A, and let

$$\mathcal{B}_t = \left\{b_t^{j\leftarrow i} : i \neq j\right\} \tag{18}$$

denote the collection of teammate beliefs relevant to the current planning problem. Thus, ξ_t contains the current task information together with the belief states and other information required to evaluate future cooperative continuations.

Given joint action $\mathbf{a}_t$, the physical dynamics produce the next task state and observations, after which the teammate belief models update $\mathcal{B}_t$. These steps define the continuation kernel

$$\mathcal{K}(\xi_{t+1} \mid \xi_t, \mathbf{a}_t) = \Pr(\xi_{t+1} \mid \xi_t, \mathbf{a}_t). \tag{19}$$

For focal planning state ξ and focal action a, define

$$Q(\xi, a) = \mathbb{E}\left[\sum_{k=t}^{T} \gamma^{k-t} r_k \mid \xi_t = \xi,\, a_t^i = a,\, \mathcal{K}\right] \tag{20}$$

as the full cooperative continuation value.

B.2 Matched Belief Counterfactual

The matched continuation isolates the part of cooperative value created through the teammate's changed belief and response. Decompose teammate j's next observation as

$$o_{t+1}^{j} = \left(\bar{o}_{t+1}^{j}, e_{t+1}^{j\leftarrow i}\right), \tag{21}$$

where $\bar{o}_{t+1}^{j}$ contains the physical observation used by the task process and $e_{t+1}^{j\leftarrow i}$ denotes the task-dependent coordination evidence carried by the focal behavior. The ordinary update is

$$b_{t+1}^{j\leftarrow i} = \mathcal{U}_{j\leftarrow i}\left(b_t^{j\leftarrow i}, o_{t+1}^{j}\right). \tag{22}$$

Define a matched update operator $\mathcal{U}_{j\leftarrow i}^{-}$ that receives the same prior belief and physical observation while neutralizing only the task-dependent evidence carried by the focal behavior:

$$b_{t+1}^{-,j\leftarrow i} = \mathcal{U}_{j\leftarrow i}^{-}\left(b_t^{j\leftarrow i}, o_{t+1}^{j}\right). \tag{23}$$

For a Bayesian receiver, choosing the matched action likelihood as

$$\ell^{-}(a \mid \theta, s) = c(a, s), \tag{24}$$

for any positive $c(a, s)$ independent of θ, makes the focal task-dependent evidence cancel during posterior normalization.

Let $\mathcal{K}_a^{-}$ equal $\mathcal{K}$ except for this matched update on the selected focal-to-teammate link. The construction preserves the focal action, its immediate physical transition, immediate reward, and all nonfocal observations. Later teammate actions are generated from the matched belief, so later physical trajectories may diverge through the teammate response.

The matched continuation is

$$Q^{-}(\xi, a) = \mathbb{E}\left[\sum_{k=t}^{T} \gamma^{k-t} r_k \mid \xi_t = \xi,\ a_t^i = a,\ \mathcal{K}_a^{-}\right]. \tag{25}$$

In the matched-action experiment, $\mathcal{U}_{j\leftarrow i}^{-}$ neutralizes the relevant Helper belief update with respect to the matched Main actions during the inference window. The focal Main behavior is therefore identical at the intervention point, while later states may diverge through Helper's changed response.

The shaping gain is

$$G_{\text{shape}}(a; \xi) = Q(\xi, a) - Q^{-}(\xi, a). \tag{26}$$

It can be positive, zero, or negative depending on how the induced teammate behavior changes the cooperative continuation.

B.3 Decision-Relevant Response Invariance

Proposition 1 (Response invariance). Fix ξ and a. If replacing the ordinary belief update with its matched counterfactual leaves every decision-relevant teammate continuation policy unchanged almost surely, then

$$Q(\xi, a) = Q^{-}(\xi, a), \qquad G_{\text{shape}}(a; \xi) = 0. \tag{27}$$

Proof. The full and matched processes execute the same focal action and use the same immediate physical transition law. By assumption, replacing the focal belief update does not change any subsequent teammate action distribution that can affect the cooperative continuation. The two processes therefore induce the same distribution over the next decision-relevant joint state. Repeating the argument at later steps gives the same distribution over complete future joint trajectories. Since the reward function is unchanged, the expected discounted return is identical, giving $Q(\xi, a) = Q^{-}(\xi, a)$ and hence $G_{\text{shape}}(a; \xi) = 0$. □

The proposition formalizes the role of decision relevance. Two beliefs can differ substantially while producing the same teammate decision and therefore the same shaping value. Conversely, a smaller belief change can matter more when it moves the teammate across a decision boundary and changes its future behavior.

B.4 Task Cost and Exact When-to-Shape Decomposition

Let

$$a_{\text{task}} \in \arg\max_{a \in \mathcal{A}_i} Q^-(\xi, a) \tag{28}$$

be the task-direct reference action. Define its task cost for a candidate action a as

$$C_{\text{task}}(a; \xi) = Q^-(\xi, a_{\text{task}}) - Q^-(\xi, a) \geq 0. \tag{29}$$

For the detailed decomposition, define the additional shaping gain relative to the task-direct reference:

$$\Delta G_{\text{shape}}(a; \xi) = G_{\text{shape}}(a; \xi) - G_{\text{shape}}(a_{\text{task}}; \xi). \tag{30}$$

Proposition 2 (Exact when-to-shape decomposition). For every feasible action a,

$$Q(\xi, a) - Q(\xi, a_{\text{task}}) = \Delta G_{\text{shape}}(a; \xi) - C_{\text{task}}(a; \xi). \tag{31}$$

Therefore, a is strictly preferred to the task-direct reference if and only if

$$\Delta G_{\text{shape}}(a; \xi) > C_{\text{task}}(a; \xi). \tag{32}$$

Proof. From Eq. (26),

$$Q(\xi, a) = Q^-(\xi, a) + G_{\text{shape}}(a; \xi).$$

Therefore,

$$\begin{aligned} Q(\xi, a) - Q(\xi, a_{\text{task}}) &= Q^-(\xi, a) - Q^-(\xi, a_{\text{task}}) && (33)\\ &\quad + G_{\text{shape}}(a; \xi) - G_{\text{shape}}(a_{\text{task}}; \xi) && (34)\\ &= -C_{\text{task}}(a; \xi) + \Delta G_{\text{shape}}(a; \xi), && (35) \end{aligned}$$

which proves Eq. (31). The strict preference condition follows immediately. □

A useful special case follows directly. If

$$\Delta G_{\text{shape}}(a; \xi) \leq C_{\text{task}}(a; \xi) \qquad \text{for all } a \in \mathcal{A}_i, \tag{36}$$

then the task-direct reference remains optimal.

B.5 State Dependence of Shaping Value

Equation (31) makes the shaping decision state dependent. Since ξ contains both the current task state and teammate-model information, the behavioral effect produced through the teammate and the task cost required to produce it can both change during an episode.

A state can therefore be favorable for shaping because the teammate is closer to a relevant decision boundary, because the focal behavior produces more useful coordination evidence there, or because a similar behavioral change can be induced at lower task cost. The same shaping behavior may consequently be valuable at one state and unnecessary or harmful at another.

B.6 Local Likelihood Separation Is Not the Cooperative Objective

For candidate-state generation, the experiments use

$$E(s, X^\star) = \max_a \left[\log \ell(a \mid X^\star, s) - \max_{X \neq X^\star} \log \ell(a \mid X, s) \right]. \tag{37}$$

A large $E(s, X^\star)$ indicates that at least one current action is locally diagnostic of the true coordination variable. However, it contains no term for the task cost of taking that action, whether the resulting belief change alters the teammate's decision, or how the induced response changes the later joint task.

Proposition 3 (Local likelihood separation is not sufficient for cooperative value). There is no general monotone relationship between $E(s, X^\star)$ and

$$\max_a \left[\Delta G_{\text{shape}}(a; \xi) - C_{\text{task}}(a; \xi) \right]. \tag{38}$$

Proof. Consider a state s_1 in which an action strongly separates $X^\star$ from the alternatives, so $E(s_1, X^\star)$ is large, but all resulting beliefs produce the same teammate decision. By response invariance, $G_{\text{shape}} = 0$ for the focal action, so its shaping advantage cannot be positive unless it also improves the task directly.

Now consider another state s_2 with a smaller likelihood ratio but with the teammate belief close to a decision boundary. A smaller belief change can switch the teammate to a different decision that improves the future joint continuation. Thus a state with larger local likelihood separation can have smaller cooperative shaping value, and no monotone ordering is guaranteed. □

The exhaustive shaping-state experiment provides the empirical counterpart to this result: it keeps the continuation evaluator fixed while enlarging only the set of states at which downstream cooperative value is evaluated.

The quantities Q^-, G_{shape}, and C_{task} are analytical quantities. The implemented controllers optimize the correctness-first cooperative objective defined in the main text.

B.7 Belief-Insensitive Prospective Model Limit

The Household model-mismatch study exposes an informative boundary of Main's prospective teammate model. Let Main's assumed action likelihood be

$$\hat{\ell}(a \mid X, s) = \hat{q} P_{\text{exec}}(a \mid s) + (1 - \hat{q}) P_{\text{int}}(a \mid X, s), \tag{39}$$

where $P_{\text{exec}}(a \mid s)$ is independent of the coordination variable X.

Proposition 4 (Belief-insensitive prospective-model limit). For the receiver family in Eq. (39), setting $\hat{q} = 1$ makes Main's assumed posterior invariant to its focal behavior:

$$\hat{b}_{t+1}(X) = \hat{b}_t(X). \tag{40}$$

Consequently, Main's prospective model assigns zero shaping gain to this belief channel:

$$\hat{G}_{\text{shape}}(a; \xi) = 0 \qquad \text{for all } a. \tag{41}$$

Proof. At $\hat{q} = 1$, Eq. (39) becomes

$$\hat{\ell}(a \mid X, s) = P_{\text{exec}}(a \mid s)$$

for every X. Therefore,

$$\hat{b}_{t+1}(X) = \frac{\hat{b}_t(X) P_{\text{exec}}(a_t \mid s_t)}{\sum_{X'} \hat{b}_t(X') P_{\text{exec}}(a_t \mid s_t)} = \hat{b}_t(X). \tag{42}$$

Under Main's prospective model, the ordinary and matched belief updates therefore coincide, giving $\hat{Q} = \hat{Q}^-$ and $\hat{G}_{\text{shape}} = 0$. □

This prospective-model limit differs from the matched-q mechanism-removal experiment. Under teammate-model mismatch with $\hat{q} = 1$, the actual Helper still uses $q = 0.9$, so the realized belief-dependent response channel remains present and WSP can recover from observed responses through replanning.

C Experimental Environments and Receiver Models

This appendix specifies the two task families and the Bayesian receiver models used to map observable Main behavior into Helper belief updates.

C.1 GridWorld Task Construction

The GridWorld study uses the 10×10 MindZero construction environment with its original movement, pick, put, obstacle, layout, and deterministic replay mechanics. The primitive action set is

$$\mathcal{A}_{\text{GW}} = \{\text{Up}, \text{Down}, \text{Left}, \text{Right}, \text{Stay}, \text{Pick}, \text{Put}\}. \tag{43}$$

Each scene contains one anchor object A, three candidate complementary objects, and a 2×2 assembly site. Main completes the same physical subtask under every candidate assignment by delivering A to the site. Helper must deliver the true complementary object $X^\star$ to the same site. The episode ends when both assigned objects have been delivered, so the two subtasks may proceed in parallel.

Main has access to $X^\star$ and its own task, whereas Helper observes the candidate set and infers the correct member from Main's behavior. Main's shortest task path depends on A, the assembly site, obstacles, and the current state, but not on $X^\star$. The final evaluation contains 400 held-out scenes from nine layouts with seeds 5000 and above. The frozen manifest is shared by all evaluated conditions.

C.2 Bayesian Belief Update

Helper maintains a categorical belief over the candidate complementary tasks. Both environments use the mixture likelihood

$$\ell(a \mid X, s) = qP_{\text{exec}}(a \mid s) + (1-q)P_{\text{int}}(a \mid X, s), \tag{44}$$

with posterior update

$$\tilde{b}_{t+1}(X) = b_t(X)\ell(a_t \mid X, s_t), \qquad b_{t+1}(X) = \frac{\tilde{b}_{t+1}(X)}{\sum_{X'} \tilde{b}_{t+1}(X')}. \tag{45}$$

The prior is uniform. The two environments differ in how the execution component and the task-dependent evidence component are constructed.

C.3 GridWorld Execution Likelihood

The GridWorld execution component uses the MindZero ExactInference action likelihood for Main's task of delivering A to the assembly site. It is independent of the candidate X.

If the current staged planner requires a pick or put action, the required action receives probability one and the other actions receive probability zero. At movement states, the action value is

$$Q_{\text{exec}}(s, a) = \gamma V(T(s, a)) - c, \qquad \gamma = 0.95, \quad c = 0.1. \tag{46}$$

For n remaining planner steps,

$$V(n) = -c\frac{1-\gamma^n}{1-\gamma}. \tag{47}$$

The movement likelihood is

$$P^{\text{GW}}_{\text{exec}}(a \mid s) = \frac{\exp[\beta_{\text{task}} Q_{\text{exec}}(s, a)]}{\sum_{a'} \exp[\beta_{\text{task}} Q_{\text{exec}}(s, a')]}, \qquad \beta_{\text{task}} = 4. \tag{48}$$

Equivalently, the temperature is $\tau = 1/\beta_{\text{task}} = 0.25$.

C.4 GridWorld Coordination-Evidence Likelihood

The task-dependent component assigns likelihood according to how compatible Main's motion is with candidate assignment X. Let $r(s, a)$ denote the exact remaining number of Main-task steps after candidate action a. Illegal candidate actions use the configured maximum remaining cost.

Define the analytical reference distribution

$$\widetilde{P}_{\text{exec}}(a \mid s) = \frac{\exp[-\beta_{\text{task}} r(s, a)]}{\sum_{a'} \exp[-\beta_{\text{task}} r(s, a')]}, \tag{49}$$

which is used only to center the geometric pointing term.

Let $p'(s, a)$ be Main's position after action a, with illegal moves leaving the position unchanged for the geometric calculation, and let $d(p, X)$ be the obstacle-aware distance from position p to candidate object X. Define

$$\phi(a, X, s) = -d(p'(s, a), X) \tag{50}$$

$$- \sum_{a'} \widetilde{P}_{\text{exec}}(a' \mid s) \left[-d(p'(s, a'), X)\right]. \tag{51}$$

Equivalently,

$$\phi(a, X, s) = \mathbb{E}_{a' \sim \widetilde{P}_{\text{exec}}}[d(p'(s, a'), X)] - d(p'(s, a), X). \tag{52}$$

Thus $\phi > 0$ means that the action moves Main closer to X than expected under the reference task distribution.

The task-dependent action score is

$$q_{\text{int}}(a; X, s) = -r(s, a) + \lambda\phi(a, X, s), \qquad \lambda = 1, \tag{53}$$

and

$$P_{\text{int}}^{\text{GW}}(a \mid X, s) = \frac{\exp[\beta q_{\text{int}}(a; X, s)]}{\sum_{a'} \exp[\beta q_{\text{int}}(a'; X, s)]}, \qquad \beta = 2. \tag{54}$$

The headline receiver uses $q = 0.9$. GridWorld applies a minimum posterior probability of 10^{-4} before normalization. WSP does not optimize this likelihood directly; it uses the receiver to predict belief changes and then evaluates candidate shaping states through downstream cooperative value.

C.5 PARTNR-Derived Household Task Construction

The Household family uses official PARTNR validation episodes containing multiple parallel object-to-receptacle propositions. The frozen evaluation family contains 30 episodes with at least two Helper candidates.

For each atomic relocation subtask g, define deterministic task cost

$$C_i(g) = \frac{d_{\text{nav}}(p_i, o_g)}{0.25} + 8 + \frac{d_{\text{nav}}(o_g, r_g)}{0.25}, \tag{55}$$

where p_i is the initial position of agent i, o_g is the object location, r_g is the receptacle location, and d_{nav} is geodesic distance on the official navigation mesh. The fixed value 8 is the manipulation overhead used by the completion-time model.

Main's assigned subtask is frozen as

$$g_{\text{Main}} = \arg\min_{g} C_0(g), \tag{56}$$

with proposition index as the deterministic tie break. The remaining propositions form Helper's candidate set. Each candidate is treated once as $X^\star$, yielding 87 paired evaluation instances from the 30 frozen episodes. Main's assigned task remains unchanged across candidate instances derived from the same episode.

C.6 Household Planner-Based Likelihood

The Household receiver uses planner-based inverse planning over primitive navigation actions

$$\mathcal{A}_{\text{HH}} = \{\text{TurnLeft}, \text{TurnRight}, \text{MoveForward}, \text{Stay}\}. \tag{57}$$

The state used by the navigation likelihood contains Main's position and yaw. For candidate action a, analytical transition $T(s, a)$ produces a successor pose without modifying the simulator. A forward step is 0.25 m and a turn changes heading by 15°.

For candidate complementary task X, let $V_X(s')$ be the exact remaining number of primitive turn and movement actions required from successor pose s' to the first interaction location for X. Define

$$Q_X(s, a) = 1 + V_X(T(s, a)). \tag{58}$$

The task-dependent likelihood is

$$P_{\text{int}}^{\text{HH}}(a \mid X, s) = \frac{\exp[-\beta Q_X(s, a)]}{\sum_{a'} \exp[-\beta Q_X(s, a')]}, \qquad \beta = 2. \tag{59}$$

The execution component is constructed analogously from Main's own task:

$$P_{\text{exec}}^{\text{HH}}(a \mid s) = \frac{\exp[-\beta Q_{\text{Main}}(s, a)]}{\sum_{a'} \exp[-\beta Q_{\text{Main}}(s, a')]}. \tag{60}$$

The mixture in Eq. (44) uses $q = 0.9$.

D TEAMMATE RESPONSE AND WSP REALIZATION

The receiver specifies how observable Main behavior changes Helper's belief. This section specifies how that belief affects Helper's physical behavior and how the evaluated planning conditions use the realized response.

D.1 GRIDWORLD HELPER RESPONSE

GridWorld uses a frozen Bayes-rational Helper. After each observed Main action, Helper updates its posterior over the three complementary candidates, ranks the candidates according to the posterior, and uses the common task controller to deliver the selected object. The posterior therefore affects team value through Helper's object choice and its behavior before that choice is finalized.

D.2 HOUSEHOLD HELPER RESPONSE

Before commitment, Household Helper uses a one-step-lag physical response. At step t, it follows Main's previous pose while updating its Bayesian belief from Main's observed trajectory. Helper commits when its posterior reaches $\eta = 0.90$ or when the 20-step interaction window ends. If the threshold is not reached by the deadline, it selects the MAP candidate using proposition index as the fixed tie break. After commitment, the common task controller executes the selected complementary subtask.

D.3 PLANNING BASELINES AND WSP

Task-Driven follows Main's shortest or neutral task route. Helper observes this route and updates its belief from any information naturally carried by task execution.

Open-Loop BSCP selects a belief-aware cooperative continuation at the initial interaction state and executes it open loop. GridWorld uses $K = 4$ in the headline condition, while the horizon analysis evaluates $K \in \{1, 2, 3, 4\}$. Household uses a 20-step interaction horizon.

Closed-Loop BSCP uses the same prospective receiver and teammate-response model but executes one current action, observes the realized Helper response and joint state, and replans from that state under a fixed interaction schedule.

Timing-Aware Open-Loop BSCP receives the same candidate shaping states as WSP and evaluates them with an open-loop continuation. Once an onset and continuation are selected, execution remains fixed. This controls for shaping-state selection without realized-response replanning.

WSP evaluates the same candidate states using the closed-loop continuation model. GridWorld keeps at most eight proposal states including $t = 0$; Household keeps at most five including $t = 0$. A no-shaping option is appended separately. Candidate continuations are ranked lexicographically by predicted assignment correctness first and predicted completion time second.

Perfect Information gives $X^\star$ to Helper at the start while leaving the physical task unchanged.

D.4 EXHAUSTIVE SHAPING-STATE EVALUATION

The exhaustive analysis keeps the WSP continuation evaluator and correctness-first objective fixed while expanding the candidate onset set to every valid state on the common task-driven prefix. GridWorld evaluates $t = 0$ through 30. Household evaluates $t = 0$ through 20 and excludes states after Helper has already committed. The task-direct option remains available in both environments.

D.5 EVALUATION SETTINGS

GridWorld additionally fixes $\beta_{\text{task}} = 4$, $\lambda = 1$, exact search depth 3, beam width 60, posterior floor 10^{-4}, and closed-form seed 424242. Household fixes $\eta = 0.90$, uses 0.25 m forward steps and 15° turns, and assigns an eight-step manipulation overhead in the completion-time model. All methods within each environment use the same manifest, lexicographic objective, and completion-time definition.

Table 5: Core settings shared by all reported evaluations.

Setting	GridWorld	Household
Evaluation instances	400	87
Candidate tasks	3	varies by episode
Mixture weight q	0.9	0.9
Receiver inverse temperature β	2.0	2.0
Interaction horizon	$K_{\max} = 4$	$K = 20$
Bootstrap resamples	2000	2000
Bootstrap seed	42	42
Onset-state cap	8	5

E Additional Experimental Results

E.1 Complete Paired Coordination Results

For a pair A vs. B, Table 6 reports $\Delta T = T_A - T_B$, so positive values favor B. Confidence intervals use 2,000 cluster-bootstrap resamples with seed 42, resampling GridWorld layouts or Household source episodes.

Table 6: Final correctness-first paired completion-time comparisons.

Environment	Pair	ΔT	Cluster 95% CI	W/T/L (B earlier/tie/later)
GridWorld	Task-Driven vs. Open-Loop BSCP	15.9875	[14.0317,18.0359]	308/79/13
GridWorld	Open-Loop vs. Closed-Loop BSCP	0.0400	[0.0064,0.0838]	7/392/1
GridWorld	Open-Loop vs. Timed Open-Loop	0.1275	[-0.0549,0.3870]	8/390/2
GridWorld	Closed-Loop BSCP vs. WSP	0.1175	[-0.0642,0.3823]	7/391/2
GridWorld	Timed Open-Loop vs. WSP	0.0300	[0.0000,0.0797]	5/394/1
GridWorld	WSP vs. Perfect	1.3875	[1.1598,1.5688]	229/164/7
Household	Task-Driven vs. Open-Loop BSCP	2.6254	[-0.5255,7.1136]	10/71/6
Household	Open-Loop vs. Closed-Loop BSCP	6.9331	[-1.8874,16.2184]	36/35/16
Household	Open-Loop vs. Timed Open-Loop	21.3625	[10.9695,34.9898]	49/38/0
Household	Closed-Loop BSCP vs. WSP	14.7616	[7.2570,25.1645]	52/35/0
Household	Timed Open-Loop vs. WSP	0.3322	[-5.2147,3.9646]	39/46/2
Household	WSP vs. Perfect	22.2815	[12.1328,32.0571]	45/0/42

The corresponding correctness transitions show the same pattern. In GridWorld, Open-Loop to Closed-Loop BSCP corrects three scenes with no reverse transition, and Closed-Loop BSCP to WSP corrects three additional scenes with no reverse transition. In Household, Open-Loop to Closed-Loop BSCP has 11 incorrect-to-correct and three correct-to-incorrect transitions, while Closed-Loop BSCP to WSP has 14 incorrect-to-correct and no reverse transition.

E.2 GridWorld Open-Loop Saturation Diagnostic

Let T_{stable} be the first step at which the correct assignment becomes the MAP hypothesis and remains so for the rest of inference.

Table 7: GridWorld posterior diagnostic. Stable-correct rate is the fraction of scenes that reach a stable correct MAP assignment.

Condition	Stable-correct (%)	Mean T_{stable}	Median T_{stable}
Task-Driven	31.50	9.071	6
Open-Loop BSCP	94.75	3.881	1
Closed-Loop BSCP	95.50	4.008	2

Open-Loop BSCP adds 256 stable-correct scenes relative to Task-Driven. Among scenes stable-correct under both Open- and Closed-Loop BSCP,

$$T_{\text{stable}}^{\text{Open}} - T_{\text{stable}}^{\text{Closed}} = -0.145$$

on average, with layout-clustered 95% CI $[-0.241, -0.059]$. This confirms that GridWorld is already close to a belief-resolution ceiling before closed-loop refinement.

E.3 Exhaustive Shaping-State Results

Table 8: Primary WSP proposal versus exhaustive shaping-state evaluation under the same correctness-first continuation evaluator.

	Accuracy (%)		T_{finish}		Initial/Delayed/None	
Environment	WSP	Full	WSP	Full	WSP	Full
GridWorld	97.0	100.0	22.83	22.20	312/0/88	262/65/73
Household	62.1	69.0	134.43	129.03	21/47/19	8/74/5

For GridWorld, exhaustive evaluation improves accuracy by 3.0 points with cluster 95% CI $[1.40, 4.27]$ and reduces mean completion time by 0.625 steps with CI $[0.424, 0.780]$. It corrects 12 scenes with no reverse transition.

For Household, exhaustive evaluation improves accuracy by 6.90 points with CI $[2.25, 12.90]$ and reduces mean completion time by 5.40 steps with CI $[1.81, 10.98]$, correcting six instances with no reverse transition. Relative to Closed-Loop BSCP, exhaustive onset search improves Household accuracy by 22.99 points and reduces completion time by 20.16 steps.

These results show that local likelihood separation is useful for proposing candidate states but does not by itself identify the state with the highest downstream cooperative value.

E.4 Fixed Open-Loop Interaction Horizon

Table 9: GridWorld Open-Loop BSCP as the fixed interaction horizon K varies.

K	Accuracy (%)	Mean T	Median T	$T \mid$ correct
1	61.00	29.4775	26.0	23.3852
2	79.50	25.6150	23.5	22.8082
3	89.75	23.8475	23.0	22.6128
4	95.50	22.9850	22.0	22.6309

Accuracy increases rapidly with the interaction horizon, while completion time conditioned on already-correct assignments changes much less. The $K = 4$ condition exactly reproduces the headline Open-Loop BSCP result.

E.5 Matched-Action Teammate-Response Intervention

GridWorld. The intervention contains 400 unique scenes. Main's factual action sequence is reproduced in all 400, and 391 also pass the strict matched-physical-effect check. On this strict subset, normal Helper target assignment is correct in 94.63% of scenes and the evidence-neutral condition in 34.27%, a 60.36-point degradation with layout-clustered 95% CI $[55.80, 65.72]$.

The paired transitions are 238 normal-correct/neutral-wrong, 2 normal-wrong/neutral-correct, 132 correct in both, and 19 wrong in both. Completion time is identical on the strict physical-match subset. The evidence-neutral receiver is ordering-sensitive under posterior ties, so the causal claim uses paired realized target-assignment degradation rather than absolute neutral MAP accuracy.

Household. All 87 instances pass the strict physical-match check. Normal target assignment is correct in 62.07% and the evidence-neutral condition in 37.93%, a 24.14-point degradation with episode-clustered 95% CI $[13.95, 35.90]$.

The paired transitions are 26 normal-correct/neutral-wrong, five reverse, 28 correct in both, and 28 wrong in both. The neutral rollout increases mean completion time by 14.22 steps with 95% CI

$[5.39, 25.21]$. Assignment reproduction is exact in all 87 factual replays; 18 normal replays differ only in completion time, so assignment degradation is the primary Household causal outcome.

Together, these interventions isolate the cooperative effect transmitted through Helper's changed belief and response while holding Main's focal physical behavior fixed.

E.6 Timing-Controlled Closed-Loop Replanning

In GridWorld, Timing-Aware Open-Loop BSCP and WSP choose the same onset in all 400 scenes: 312 initial and 88 no-shaping cases. WSP improves accuracy from 96.25% to 97.0%, while mean completion time changes from 22.8575 to 22.8275.

In Household, Timing-Aware Open-Loop BSCP selects 5 initial, 47 delayed, and 35 no-shaping cases, while WSP selects 21 initial, 47 delayed, and 19 no-shaping cases. The two controllers choose different onsets in 21 of 87 instances.

Across those 21 different-onset cases, WSP is 6.10 steps faster on average. Across the 66 same-onset cases, its mean completion time is 1.50 steps higher. The aggregate WSP advantage is therefore concentrated in states where the realized teammate response changes the relative value of candidate shaping continuations.

E.7 Matched-q Shaping-Mechanism Boundary

When the actual and assumed mixture weights are varied together, the belief-dependent response channel changes in the realized interaction itself rather than only in Main's prospective model.

Table 10: Household accuracy under matched $q_{\mathrm{actual}} = \hat{q}$. No monotonicity is assumed; $q = 1$ is the shaping-mechanism boundary.

q	Task	Open-Loop	Closed-Loop	Timed Open-Loop	WSP
0.50	34.48	36.78	44.83	57.47	60.92
0.70	34.48	35.63	43.68	59.77	62.07
0.80	34.48	36.78	42.53	59.77	60.92
0.90	34.48	36.78	45.98	59.77	62.07
0.95	34.48	36.78	42.53	62.07	64.37
0.99	34.48	35.63	49.43	62.07	65.52
1.00	34.48	34.48	34.48	34.48	34.48

At $q = 1$,

$$P(a \mid X, s) = P_{\mathrm{exec}}(a \mid s)$$

for every X. The observed action therefore carries no task-dependent likelihood ratio across candidate assignments, and all non-perfect methods collapse to the Task-Driven accuracy of 34.48%. This is a mechanism-removal condition rather than a teammate-model prediction error.

E.8 Robustness to Teammate-Model Mismatch

We evaluate teammate-model mismatch by keeping the executed Household Helper fixed at

$$(q, \beta, \eta) = (0.90, 2.0, 0.90)$$

while perturbing one parameter in Main's prospective teammate model at a time.

Exact model match reproduces the headline results: Closed-Loop BSCP achieves 45.98% accuracy with mean completion time 149.19, while WSP achieves 62.07% accuracy with mean completion time 134.43.

Moderate mismatch leaves WSP close to its exact-model performance. Across the tested $\hat{\beta}$ and $\hat{\eta}$ values, WSP accuracy remains between 59.77% and 63.22%, while the corresponding completion-time changes remain small. For

$$\hat{q} \in \{0.70, 0.80, 0.90, 0.95\},$$

Table 11: Household robustness to teammate-model mismatch. The executed Helper remains fixed at $(q, \beta, \eta) = (0.90, 2.0, 0.90)$, while Main perturbs one assumed parameter at a time. ΔT is the change in mean completion time relative to exact model match.

	Closed-Loop BSCP		WSP	
Assumed setting	Acc. (%)	ΔT	Acc. (%)	ΔT
Mixture weight $\hat{q}$				
$\hat{q} = 0.70$	42.53	$+4.02$	62.07	-1.40
$\hat{q} = 0.80$	43.68	$+0.07$	62.07	-0.45
$\hat{q} = 0.90$	45.98	0.00	62.07	0.00
$\hat{q} = 0.95$	42.53	$+4.08$	62.07	-0.30
$\hat{q} = 1.00$	34.48	$+9.79$	58.62	$+0.67$
Inverse temperature $\hat{\beta}$				
$\hat{\beta} = 1.0$	45.98	-2.21	62.07	-1.65
$\hat{\beta} = 1.5$	43.68	-0.75	59.77	$+0.14$
$\hat{\beta} = 2.0$	45.98	0.00	62.07	0.00
$\hat{\beta} = 3.0$	41.38	$+2.10$	60.92	-0.97
$\hat{\beta} = 4.0$	40.23	$+1.20$	60.92	-1.16
Commitment threshold $\hat{\eta}$				
$\hat{\eta} = 0.70$	45.98	-0.46	62.07	-2.34
$\hat{\eta} = 0.80$	42.53	$+0.95$	63.22	-2.35
$\hat{\eta} = 0.90$	45.98	0.00	62.07	0.00
$\hat{\eta} = 0.95$	45.98	-1.22	63.22	-0.76

WSP maintains 62.07% accuracy and decreases to 58.62% only at $\hat{q} = 1$.

Importantly, $\hat{q} = 1$ here is a teammate-model mismatch rather than removal of the shaping mechanism. Main's prospective model assumes that its behavior provides no task-dependent evidence, but the actual Helper remains fixed at $q = 0.9$ and continues to update its belief from Main's observed behavior. The realized belief-dependent response channel therefore remains present.

Under this mismatch, Closed-Loop BSCP reaches 34.48% accuracy and mean completion time 158.98, whereas WSP retains 58.62% accuracy and mean completion time 135.10. WSP changes its selected shaping onset in 25 of 87 instances relative to exact $\hat{q} = 0.9$, showing how closed-loop replanning can partially compensate for errors in the prospective teammate model.