THINKING WITH SPATIAL–SOCIAL CODE FOR SOCIAL REASONING IN VIDEOS

Anonymous authors
Paper under double-blind review

ABSTRACT

Real-world social reasoning is essential for intelligent systems to understand human behavior, anticipate needs, and coordinate actions with people. Inferring beliefs, goals, and intentions requires connecting the visual context of an interaction with each participant's history of observations and actions. The central challenge is bridging observable interaction and latent social states: visual events acquire social meaning through a participant's access to information, prior experience, and ongoing goals. A representation must therefore capture not only the scene, but also how it informs different participants' perspectives. We argue that explicit code offers a computational medium for organizing and using these dependencies through stable references, agent-specific records, and update operations. We introduce Spatial–Social Code (SSC), which links physical situations, visual evidence, and inferred social states in a persistent, updateable representation. Spatial Code organizes entities, spatial relations, physical events, and perceptual-access evidence. Social Code links these records to participants' information and hypotheses about their beliefs, goals, and intentions. Their connections make explicit how situated experiences inform social inference, supporting both queryable reasoning and social memory over time. We evaluate SSC on cross-agent video question answering, collaborative visual grounding, and embodied interaction, demonstrating improvements across multiple tasks. Together, these capabilities provide a computational interface for social reasoning grounded in evolving physical situations and individual experience.

1 INTRODUCTION

Real-world social reasoning is a foundational capability for intelligent systems that operate alongside people. Human actions are guided by beliefs, goals, and intentions; understanding these states is essential to explaining behavior, anticipating needs, and coordinating joint activity. For an embodied assistant, such understanding shapes consequential decisions: what information to provide, which action to take next, and when to intervene. Interactive assistance already brings these demands together through tasks such as mistake detection, hand forecasting, and intervention prediction (Wang et al., 2023), while multimodal Theory of Mind (ToM) benchmarks examine the belief and goal inferences underlying social understanding (Jin et al., 2024; Shi et al., 2025). The central challenge is to interpret the physical world in relation to the people acting within it: what they have experienced, what they believe, and what they are trying to achieve. This makes the connection between visual situations and individual experience a central requirement for real-world social intelligence.

Much of the relevant evidence is visual. People do not continuously narrate what they have noticed, remembered, or intended. Their locations, actions, orientations, and responses provide evidence about their engagement with a situation. Yet social states are not directly visible in the same way as an object relocation. A camera may record an event that a participant misses; an agent may retain an earlier belief after the world changes; and the same search action can support different goals in different contexts. Long, multi-view recordings compound this difficulty by distributing evidence across people and time (Kim et al., 2026). A successful account must connect *what happened*, *what each participant experienced*, and *what those experiences support about the participant*.

The representational challenge is to connect observable interaction with the latent states that explain human behavior. Social meaning depends on a relation between the physical situation and an in-

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.

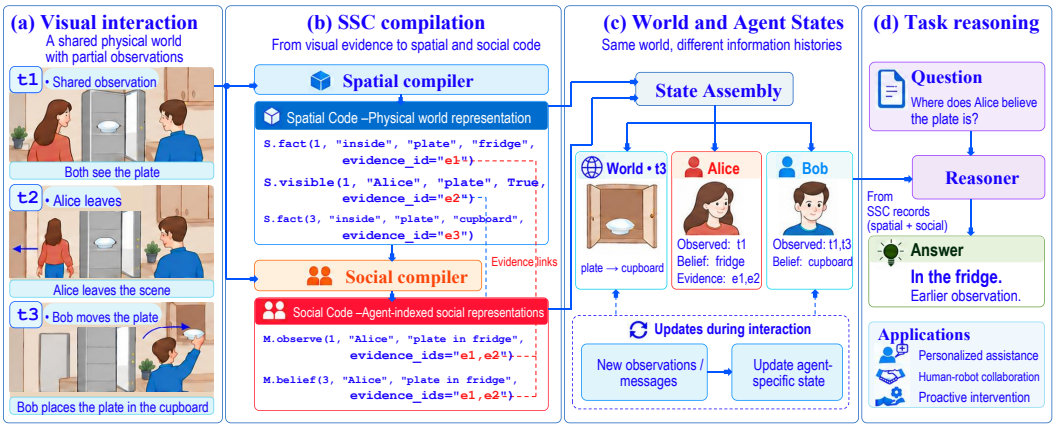

Figure 1: **Spatial–Social Code: from visual interaction to agent-specific reasoning.** (a) Participants acquire different observation histories in a shared physical situation. (b) Spatial and Social compilers produce code linked by evidence references. (c) State assembly organizes linked records into world and agent states, preserving time and evidence references; observations and messages update them during interaction. (d) A reasoner uses the records for task-specific inference and can also access the original visual evidence (connection omitted for clarity). The illustrative belief example assumes persistence and no intervening communication; personalized assistance, human–robot collaboration, and proactive intervention are potential applications, while online updates are adapter-specific.

dividual's experience: an event informs a belief through the person's access to it, while an action supports a goal interpretation through its context and history. A scene-centered account alone leaves these relations unspecified, and a mental-state description detached from visual evidence leaves its interpretation ungrounded. Direct multimodal prediction and free-form summaries do not explicitly require these relations to be represented or maintained, leaving their reconstruction to each downstream inference. **We argue that visual social reasoning requires an explicit connection between physical situations, individual experience, and social-state interpretation.** This connection is the organizing principle of our representation.

Code provides a **computational medium** for expressing and operating on such relations. Visual programs compose perception with executable operations (Gupta & Kembhavi, 2023; Surís et al., 2023), and code-based policies connect perception to embodied control (Liang et al., 2023). Spatial Code exposes geometric variables for video reasoning (Chen et al., 2026), while PDDL-Mind verifies state evolution in textual ToM (Yi et al., 2026). These works demonstrate how explicit representations make intermediate information available for computation. However, geometric representations alone do not specify participants' information and inferred mental states, while text-based state tracking does not address constructing those states from visual interaction. Real-world social reasoning calls for a representation that connects both: the physical conditions and visual evidence of an interaction, and their implications for individual participants. Stable references, agent scope, compositional queries, and explicit updates make code a suitable medium for preserving and using these connections.

We introduce **Spatial–Social Code (SSC)**, which represents these connections through two linked stores (Figure 1). *Spatial Code* describes an evolving interaction situation: entities, relations, events, and the perceptual conditions under which agents participate. *Social Code* describes agent-specific information and hypotheses about beliefs, goals, and intentions. Shared entity and event references connect the stores through information-access and behavioral-evidence relations. Restricted execution materializes records that a reasoner can inspect; an online adapter maintains private state from observations, action feedback, and confirmed messages. Persisting these records also provides a task-specific memory in which a world update need not erase another agent's older belief.

We evaluate SSC across cross-agent video question answering, collaborative visual grounding, and embodied interaction, examining both social inference and the use of persistent state. In the full

video-QA evaluation, the largest category-level improvement is on Theory of Mind: accuracy increases from 41.28% to 47.66% over a same-backbone, same-clip baseline. Component comparisons and online experiments further examine the roles of representation content and state updates. Our contributions are:

- **A representation and reasoning framework for visual social understanding.** We introduce Spatial–Social Code, linking physical situations and visual evidence to agent-specific information and inferred social states. The framework supports evidence-linked reasoning and continued state updates for social memory.
- **An empirical evaluation across environments and tasks.** We study SSC in cross-agent video question answering, collaborative visual grounding, and embodied interaction. The evaluation shows a 6.4% improvement in video-based Theory-of-Mind accuracy and uses component comparisons and online experiments to examine representation content and state updates.

2 Related Work

Visual social reasoning and assistance. Social-IQ studies social understanding from video (Zadeh et al., 2019); MMToM-QA and MuMA-ToM evaluate multimodal belief and goal reasoning (Jin et al., 2024; Shi et al., 2025). HoloAssist connects egocentric perception to task guidance and intervention (Wang et al., 2023). These settings motivate reasoning about participants rather than only recognizing actions. Inverse planning and agent modeling explain behavior using latent goals and beliefs (Baker et al., 2009; Rabinowitz et al., 2018; Shu et al., 2021). SSC supplies an explicit evidence interface for such interpretations; it does not introduce a new general theory of human behavior or claim to infer true private mental states.

Code and explicit representations. VisProg and ViperGPT generate programs that compose visual tools (Gupta & Kembhavi, 2023; Surís et al., 2023); program-aided reasoning delegates operations to an interpreter (Gao et al., 2023; Chen et al., 2023); Code as Policies generates embodied controllers (Liang et al., 2023). Thinking with Spatial Code represents video using explicit three-dimensional variables such as bounding boxes and orientations (Chen et al., 2026). SSC uses a restricted language to construct relational evidence and agent-state records. Its Spatial layer concerns the spatiotemporal situation in which information and action occur; it does not require a learned metric 3D reconstruction module.

Perspective taking and state tracking. SimToM filters context through a target perspective (Wilf et al., 2024), AutoToM constructs agent models (Zhang et al., 2025), and PDDL-Mind verifies symbolic state transitions while tracking perceptual access (Yi et al., 2026). PDDL-Mind evaluates textual inputs; its language can in principle accept grounded predicates, so modality alone is not an expressivity distinction. SSC investigates constructing and retaining links from visual evidence to agent-specific states. Belief and perspective benchmarks (Nematzadeh et al., 2018; Kim et al., 2023; Wu et al., 2023; Xu et al., 2024) and literal/functional ToM distinctions (Gu et al., 2026; Juneja et al., 2026) also motivate separating a correct verbal answer from successful behavior.

3 Problem Formulation

Visual social reasoning. Let $O_{\leq t}$ denote the visual interaction evidence available to a reasoning system up to time t, together with any supplied language or communication records. Given a query q, the system predicts an answer $\hat{y} = R(O_{\leq t}, q)$ about participants' information, beliefs, goals, or intentions. In an interactive setting, these judgments instead inform an action toward a task goal. Both settings require relating observable events to the perspectives of the agents involved.

Physical situation and individual experience. A current physical configuration does not uniquely determine a participant's social state. Consider two histories h and h' in which an object reaches the same location, but a participant witnesses its relocation only in h. Under a belief-persistence assumption, their belief answers can differ:

$$W_t(h) = W_t(h'), \qquad y^*(h, q) \neq y^*(h', q). \tag{1}$$

An encoder that collapses these histories into the same representation cannot support this distinction on its own. The relevant information is relational: what happened, how a participant encountered it, and what interpretation that experience supports. The representation must preserve this connection across time, rather than retaining only the latest scene configuration.

Linked representation. For agents $\mathcal{A} = \{1, \ldots, N\}$, we introduce

$$C_t = \big(P_t, \{S_t^i\}_{i \in \mathcal{A}}, L_t\big), \tag{2}$$

where P_t represents physical situations, events, and perceptual evidence; S_t^i represents information attributed to agent i and hypotheses about its social state; and L_t connects these records through participants, time, and evidence references. These are estimates constructed from available evidence. A video analyst can represent multiple participants, while a decentralized agent maintains its own state C_t^i from the information available to it.

Representation requirements. This formulation yields three requirements. **Agent attribution** preserves whose information a record describes. **Evidence linkage** connects social interpretations to the observations, events, or communications that support them. **Temporal updating** distinguishes changes in the physical situation from changes in a participant's information. Spatial–Social Code realizes these requirements through typed records and explicit operations, providing a common interface for downstream reasoning and state maintenance.

4 SPATIAL–SOCIAL CODE

Spatial–Social Code (SSC) links the shared physical setting of visual interaction with individual perspectives arising from different experiences, providing explicit context for social reasoning. Spatial Code describes physical events, spatial relations, and perceptual evidence; Social Code represents participants' information and interpretations of their beliefs, goals, and intentions. Participant, time, and evidence references connect these layers (Equation 2). Following Figure 1, we explain their design, use in reasoning, and maintenance during interaction.

4.1 REPRESENTATION DESIGN

Spatial Code and Social Code share entity, event, and time references while preserving each record's meaning and source. An object's location can appear as a world estimate, a participant's belief, or a received report. Links capture how events enter individual experience and support social interpretations.

Spatial Code is a structured representation of the physical interaction context, including entities, spatial relations, time-indexed events, and perceptual conditions. Social understanding requires how participants encounter information, beyond scene contents. An object's containment, the cabinet opening, and an observer's presence jointly support judging whether the observer acquired location information. Spatial Code therefore retains the conditions of observation and action beyond position or geometry alone. It provides a traceable physical basis for social inference without requiring complete metric 3D reconstruction.

Social Code represents participant-specific experience and social states, distinguishing attributed observations and received information from hypotheses about beliefs, goals, and intentions. An event visible to the system may not be experienced by every participant, and an observed action does not directly reveal its purpose. Social records therefore retain agent attribution, time, and supporting evidence, allowing an interpretation to be inspected separately from its basis. Different agents can hold different information; an agent's earlier belief can coexist with a newer physical fact. States without sufficient support can remain unknown.

The layers are connected through visual evidence. The Spatial compiler extracts physical relations, events, and perceptual conditions; the Social compiler combines these records with available visual interaction evidence to construct attributed experiences and social interpretations (Figure 1(b)). An event's implications for an agent depend on evidence of access, such as presence, inspection, or received communication. In Figure 1(a–c), relocating the plate supports updating the world state, while Alice's earlier observation supports her retained location belief. Task adapters can incorporate

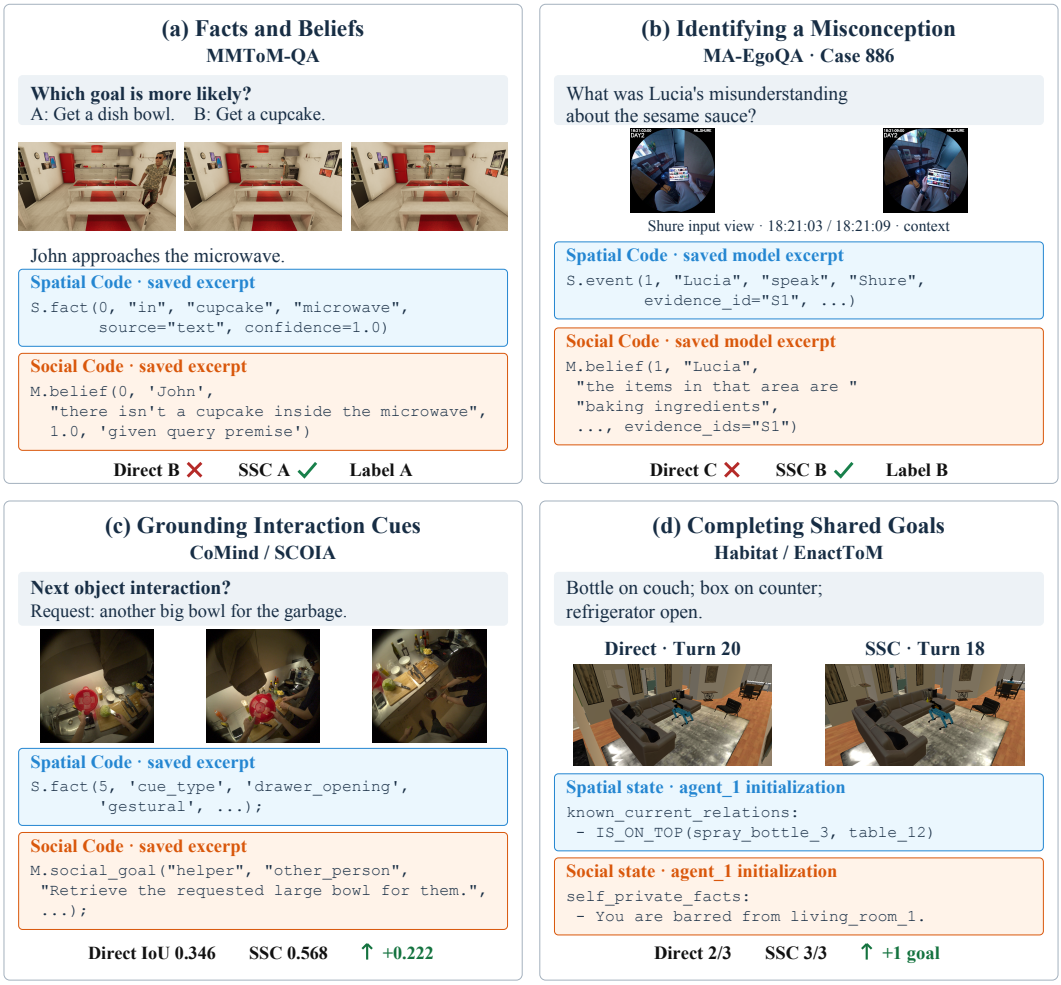

Figure 2: **Qualitative examples across four environments.** (a) MMToM-QA: text-supplied facts and belief premise, frames `0/96/224`, and candidate-conditioned readout. (b) MA-EgoQA: saved model records link Lucia's inferred misconception to S1, also cited in answer B; Shure input frames provide context. (c) CoMind: The example illustrates how an interaction-informed object description can guide localization, with higher IoU for SSC in this case. (d)Habitat: Episode 1566 pairs initialization records with final replay crops to illustrate the task context and resulting physical outcomes, including more completed goals with SSC in this run (Savva et al., 2019; Juneja et al., 2026). Together, the panels illustrate how physical context and participant information can be organized for question answering, grounding, and action.

additional perception tools: collaborative grounding uses SSC object candidates to guide region localization, as detailed in Appendix D. Video compilation protocols appear in Appendix B.

Figure 2 illustrates the resulting physical and social records across task adapters; Section 5.5 analyzes their associated outcomes.

4.2 Reasoning with Linked Physical and Social States

SSC makes physical context, individual experience, and social interpretation jointly accessible. World states supply behavioral context, agent states supply perspectives, and evidence links identify supporting experiences. The reasoner reads their time, ownership, and sources directly.

Belief judgments combine previously acquired information with subsequent experiences, rather than reporting the current physical state. In Figure 1, the world record places the plate in the cupboard,

while Alice's record retains her earlier observation of it in the fridge. This distinction supports agent-specific answers.

Goal and intention inference interprets behavior through physical conditions and individual experience. Repeatedly opening containers may support a search goal, but what an agent seeks and why it continues searching depend on available objects, earlier actions, and its information. Spatial Code supplies this context; Social Code supplies experiences and existing interpretations. Evidence links allow the reasoner to assess candidate explanations against relevant observations, rather than assigning a fixed intention to an isolated action.

State assembly organizes generated records by world, participant, time, and evidence reference for this use (Figure 1(c)). The model predicts visual and mental-state contents; the runtime checks permitted record structure and, when enabled, reference membership. The online updater instead applies deterministic action and delivery rules (Appendix A). The reasoner receives these records and a task query to produce an answer or action judgment (Figure 1(d)). Full-video pipelines retain raw visual input for revisiting details absent from the records.

4.3 State Updates during Interaction

Continuing interaction requires judgments to change with experience. SSC retains linked records across steps, updating them through observations, action feedback, and delivered communication according to which agent acquired the information. For agent i, the update loop in Figure 1(c) is

$$C^i_{t+1} = U_i(C^i_t, o^i_{t+1}, f^i_t, d^i_{t+1}), \tag{3}$$

where o^i denotes observations, f^i action feedback, and d^i confirmed delivered communication. Equation 3 separates physical changes from agent-specific information changes: moving an object does not inform every participant of its new location, and delivering a message updates the recipient's available information. Agent attribution and evidence sources allow these changes to be recorded separately.

For example, a confirmed Pick changes the acting agent's holding state; a failed Place records failure without asserting a new location. A confirmed placement updates that agent's location record, while other agents receive the information only through their observations or delivered messages. A received message retains the sender as its source rather than becoming a direct observation. The planner reads this attributed state on its next step. Model-based interpretation and deterministic bookkeeping thus play distinct roles: the former estimates social content, while the latter maintains access, provenance, and update status. Result-specific implementation versions are identified in Appendix C.

5 Experiments

Our experiments examine SSC's central design: linking physical context with participants' experiences to support social understanding and interactive decision-making. We first evaluate overall performance across tasks, testing whether this representation helps models use information distributed across participants and time. We then use component comparisons and controlled information-history experiments to examine the roles of spatial information, social information, and their connections, with particular attention to distinguishing changes in the physical world from what different individuals know. For ongoing interaction, we compare retaining initial records with dynamically updating state to investigate how new observations, action feedback, and communication inform subsequent decisions. Finally, case studies and error analysis trace how recorded information enters answers and actions, illustrating SSC's practical benefits and areas for further improvement.

5.1 Experimental Setup

We evaluate SSC across four environments to connect task performance with the linked representation and memory operations introduced in Section 4. MMToM-QA tests belief and goal inference (Jin et al., 2024), while MA-EgoQA tests cross-agent understanding in egocentric video (Kim et al., 2026; Yang et al., 2025); both use answer accuracy. We also evaluate collaborative grounding on CoMind, counting localization as correct at IoU > 0.5 (Gavryushin et al., 2026), and embodied

Table 1: **Task results under the stated evaluation protocols.** Percentages (↑). B includes published reference baselines; A reports MMToM-QA results; C and D show local comparisons; component controls appear in Section 5.3. Bold scores mark higher values within matched comparisons in B and D, and between the SAM-equipped rows in C.

A. MMToM-QA ($N = 600$)

Backbone	Method	Overall	Goal
Gemini 3 Pro	Direct	73.33	49.33
Gemini 3 Pro	SSC (Ours)	78.33	75.67

B. MA-EgoQA ($N = 1{,}741$) (Kim et al., 2026)

Method	Overall	EI	SI	TC	TR	ToM
All captions / Gemini-2.5-Flash	36.93	33.98	41.22	36.36	46.59	24.26
BM25 / Qwen3-VL-8B-Instruct	36.01	30.64	44.68	37.60	33.45	30.21
WorldMM-8B	27.63	22.56	29.26	34.50	25.09	21.70
EgoMAS / Gemini-2.5-Flash	41.41	48.19	41.49	41.32	39.37	33.62
Direct	48.71	44.01	50.80	52.27	51.92	41.28
SSC (Ours)	**51.92**	**44.57**	**54.52**	**53.93**	**57.84**	**47.66**

C. CoMind (Gavryushin et al., 2026)

Method	JA-L	JA-R	SCOIA	HO loc.
Direct + SAM	18.00	16.00	**10.00**	**11.00**
SSC + SAM (Ours)	**22.00**	**23.00**	9.00	10.00

D. Habitat / EnactToM

Backbone / state protocol	Method	Success	Completion
Gemini 3 Flash / v2	Direct	22.22	57.18
	SSC without updates	27.78	65.51
	SSC (Ours)	**66.67**	**83.10**

JA-L/R denote joint-attention views; HO loc. denotes handover localization (IoU > 0.5). SSC without updates retains only the initial state. Published baselines retain their original protocols; model versions, evaluation splits, and budgets are detailed in Appendices B–D.

interaction in Habitat/EnactToM, measured by task success and physical-goal completion (Savva et al., 2019; Juneja et al., 2026).

We use Direct as the baseline without SSC records. For online interaction, Static retains initialization and Dynamic updates during the task. We compare individual layers and lossless prose to examine representation content and readout, and supply SAM to both CoMind localization arms. All comparisons retain failures in their denominators.

5.2 Main Results across Environments

MMToM-QA asks what a participant believes or intends, requiring observed actions to be interpreted in their physical and informational context. SSC makes these dependencies available through linked event and participant records. Across 600 questions, it reaches 78.33% accuracy versus Direct's 73.33%, with goal inference improving from 49.33% to 75.67% (Table 1(A)). This goal-inference gain is consistent with linking actions to physical context and participant experience. Category-level results and detailed evaluation protocols are provided in Appendix B.

MA-EgoQA brings this challenge to egocentric recordings, where answering across participants requires connecting what happens over time with whose experience is relevant to the question. This is the role of SSC's participant, time, and evidence references. Using the same Gemini 3 Pro back-

bone and clips, SSC improves accuracy from 48.71% to 51.92% across 1,741 questions, answering 56 more questions correctly (Table 1(B)). The largest gains occur in ToM (+6.38 points) and temporal reasoning (+5.92 points; Figure 3(c)). Their concentration in perspective- and time-sensitive questions supports the usefulness of organizing video evidence along these dimensions, while the component analysis below examines how much can be attributed to the representation itself.

CoMind tests whether interaction context can guide visual localization, including identifying the object involved in joint attention or handover. SSC supplies object descriptions to SAM, connecting contextual reasoning to a visual grounding module. With the same SAM selection and coordinate rules, SSC raises joint-attention localization from 18% to 22% in one view and from 16% to 23% in the other (Table 1(C)). These gains illustrate a useful role for structured context in selecting a shared referent. Complete results and statistical analyses appear in Appendix D.

Habitat/EnactToM moves from interpreting recorded events to choosing actions as the environment changes. Here, a useful representation must incorporate new observations, action outcomes, and received messages into subsequent decisions. Across twelve development tasks and three seeds, Dynamic SSC succeeds in 24/36 runs (66.67%), compared with Direct's 8/36 (22.22%); physical-goal completion improves from 57.18% to 83.10% (Table 1(D)). Under shared private text inputs and inference limits, this is the strongest evidence for SSC as a persistent state interface for interaction. Section 5.4 tests the importance of continued updating beyond the initial records.

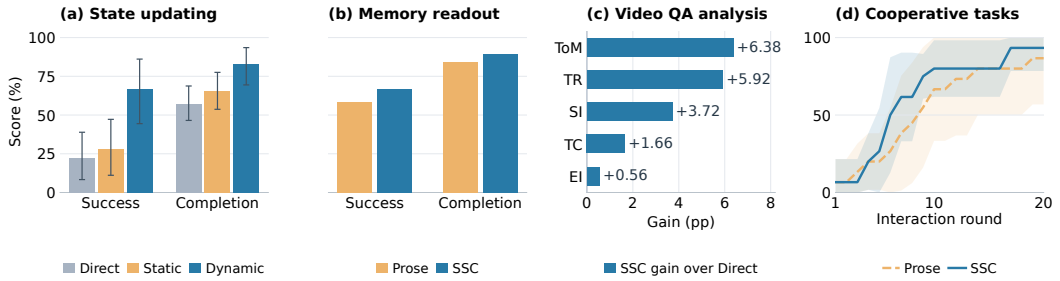

Figure 3: **Maintaining physical and participant information for reasoning and interaction.** (a) Dynamic SSC achieves higher task success and physical-goal completion than retaining only the initial records, supporting the value of incorporating observations, action feedback, and communication throughout a task. (b) With the same updater, SSC achieves higher task success than prose (8/12 vs. 7/12), showing promise as a structured memory interface for planning. (c) MA-EgoQA gains are largest in ToM and temporal reasoning, consistent with organizing evidence by participant and time. (d) In an exploratory five-task cooperative subset, SSC reaches higher final completion, illustrating progress over successive interactions. Bars in (a,b) show success and completion; (a) uses 95% task-bootstrap intervals. Curves in (d) show means $\pm$ one task SD, clipped to [0,100]. Protocols and full-task results appear in Appendix C.

5.3 Representation and Component Analysis

We examine spatial and social contributions using a historical MMToM-QA diagnostic with Qwen3-VL-32B-Instruct. Across 70 questions, joint SSC reaches 71.43%, exceeding Spatial-only by 32.86 points and Social-only by 5.71 points. The combined pipeline thus answers more questions correctly than either component, illustrating the potential of relating physical context to participant information. Pipeline differences and paired statistics appear in Appendix B.

We further conduct a controlled information-history experiment to test whether SSC distinguishes world changes from changes in individual knowledge (Equation 1). Forty text episodes each retain the same physical events across four histories: witnessed or unseen movement, and delivered or undelivered reports. For example, a moved object's actual location is fixed, but only an informed agent should know its new location. We query the world, target agent, and unrelated agent. Paired joint correctness requires all three answers to be correct in both histories of each contrast.

Table 2 shows that SSC preserves world-location accuracy at 100% while correctly changing the target judgment in 97.5% of contrasts. Removing links or attribution retains high world accuracy but

Table 2: **Tracking who knows what while physical events stay fixed.** Forty text episodes, four histories, 160 answers per condition; all failures count. Paired correctness requires correct world, target, and unrelated-agent judgments in both contrasted histories. Deletions remove information; prose preserves all fields. Values are percentages.

	SSC (Ours)	Prose	No links	No attribution
World location correct	100.0	99.4	94.4	98.8
Target judgment changes correctly	97.5	93.8	1.9	0.0
Paired joint correctness	**95.0**	92.5	0.0	0.0

reduces correct target changes to 1.9% or 0%. This separation motivates SSC's explicit information-access relations: knowing where an object is does not establish who knows its location. These deletions test information availability; the lossless prose control retains all fields. SSC achieves 95.0% paired correctness versus prose's 92.5%, using 22.2% fewer reported tokens. This offers an encouraging accuracy–token tradeoff in this diagnostic. With a shared updater and planner in twelve online task pairs, SSC also succeeds in 8/12 versus prose's 7/12 (Figure 3(b)). Together, these results show that SSC can preserve distinctions between world and participant states in a representation usable for both answering and planning; the online comparison tests readout with shared updates.

5.4 Continued State Updating

We compare Static and Dynamic SSC to test the value of incorporating experience after initialization. Dynamic succeeds in 24/36 Habitat runs versus Static's 10/36, a 38.89-point gain, with more successes in every seed (Figure 3(a)). This supports a central design choice: observations, action feedback, and communication should continually revise the state available to the planner. Dynamic also uses fewer calls than Direct (53.36 versus 64.56). These results support maintaining state throughout interaction. Figure 3(d) illustrates this interaction process in five exploratory cooperative tasks, where SSC reaches higher final completion than prose with the same updater. Continued updating improves on initialization; the trajectories illustrate how maintained records support successive decisions.

5.5 Case Studies and Error Analysis

The cases connect stored information to subsequent action. In Habitat 588, a report delivered at turn 2 enters the attributed state and remains in the next planner input before successful navigation (Appendix Figure 4). Delivery changes the recipient's information without changing the spatial state, illustrating SSC's separation of physical events from individual experience. The CoMind case shows object descriptions guiding the same SAM selection rule.On MA-EgoQA, SSC corrects 210 Direct errors. Beyond aggregate gains, SSC's explicit records provide concrete places to inspect evidence coverage, participant attribution, and the information supplied to a decision.

6 Conclusion

Understanding social interaction requires connecting what happens in the world with what each person has experienced. SSC makes this connection explicit and maintains it through interaction, supporting both behavioral interpretation and action. Our evaluations demonstrate practical benefits from organizing physical and participant information and keeping it current. In a shared environment, a change in object location need not change every participant's understanding, and a message can change that understanding without altering the world. Preserving this distinction gives an assistant a basis for deciding when to provide information and when to act. It also makes the evidence behind a participant-specific judgment available for inspection and revision as the interaction develops. For longer collaborations, this interface could help distinguish missing observations from outdated assumptions and guide whether the next step should involve clarification, communication, or action. This offers a foundation for agents that carry social context across decisions, with the longer-term potential to adapt assistance and collaboration to what different people know and need.

AI USE STATEMENT

We used OpenAI Codex to assist with literature comparison, experimental design suggestions, interpretation of results, manuscript and LaTeX editing, translation, and the development of local audit scripts. Generative AI was also used to create conceptual illustrations for Figure 1. AI-assisted checks included reference verification, consistency checks between reported results and archived records, and LaTeX compilation checks. The model-based experimental methods are described in the main text and supplementary material. The authors take responsibility for the final manuscript, including all claims, results, and artifacts produced with AI assistance.

ETHICS STATEMENT

This work studies benchmark-based inference about observations, beliefs, and intentions. Inferred states are model hypotheses rather than verified private mental states. Use and redistribution of egocentric recordings must respect dataset permissions and privacy terms.

REPRODUCIBILITY STATEMENT

The supplement specifies representation interfaces, evaluation protocols, task-level online results, and artifact availability. Source code, summaries, selected records, and deterministic exports support the analyses. The Flash archive additionally provides all paired MMToM outputs and all 108 Habitat rollouts with scoring records; rerunning the experiments requires the corresponding model services, raw data, and simulator assets.

REFERENCES

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://doi.org/10.1016/j.cognition.2009.07.005.

Jieneng Chen, Wenxin Ma, Ruisheng Yuan, Yunzhi Zhang, Jiajun Wu, and Alan Yuille. Thinking with Spatial Code for Physical-World Video Reasoning. *arXiv preprint arXiv:2603.05591*, 2026. URL https://arxiv.org/abs/2603.05591.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. Program of Thoughts Prompting: Disentangling Computation from Reasoning for Numerical Reasoning Tasks. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=YfZ4ZPt8zd.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. PAL: Program-aided Language Models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 10764–10799, 2023. URL https://proceedings.mlr.press/v202/gao23f.html.

Alexey Gavryushin, Dingxi Zhang, Zhao Huang, Alexandros Delitzas, Jiaqi Chen, Ben Ellis, Cedric Zöllner, Manthan Patel, Manuel Kaufmann, Marc Pollefeys, and Xi Wang. CoMind: Understanding Collaborative Human Activity from Multiple Minds and Views. *arXiv preprint arXiv:2607.06691*, 2026. URL https://arxiv.org/abs/2607.06691.

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 *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2410.13648.

Tanmay Gupta and Aniruddha Kembhavi. Visual Programming: Compositional Visual Reasoning Without Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14953–14962, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Gupta_Visual_Programming_Compositional_Visual_Reasoning_Without_Training_CVPR_2023_paper.html.

Chuanyang Jin, Yutong Wu, Jing Cao, Jiannan Xiang, Yen-Ling Kuo, Zhiting Hu, Tomer Ullman, Antonio Torralba, Joshua B. Tenenbaum, and Tianmin Shu. MMToM-QA: Multimodal Theory of Mind Question Answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16077–16102, 2024. doi: 10.18653/v1/2024.acl-long.851. URL https://aclanthology.org/2024.acl-long.851/.

Gurusha Juneja, Dylan Lu, Saaket Agashe, Parth Diwane, Edward Gunn, Jayanth Srinivasa, Gaowen Liu, William Yang Wang, Yali Du, and Xin Eric Wang. EnactToM: An Evolving Benchmark for Functional Theory of Mind in Embodied Agents. *arXiv preprint arXiv:2605.09826*, 2026. URL https://arxiv.org/abs/2605.09826.

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, 2023. doi: 10.18653/v1/2023.emnlp-main.890. URL https://aclanthology.org/2023.emnlp-main.890/.

Kangsan Kim, Yanlai Yang, Suji Kim, Woongyeong Yeo, Youngwan Lee, Mengye Ren, and Sung Ju Hwang. MA-EgoQA: Question Answering over Egocentric Videos from Multiple Embodied Agents. *arXiv preprint arXiv:2603.09827*, 2026. URL https://arxiv.org/abs/2603.09827v2.

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. In *IEEE International Conference on Robotics and Automation*, 2023. URL https://arxiv.org/abs/2209.07753.

Aida Nematzadeh, Kaylee Burns, Erin Grant, Alison Gopnik, and Thomas L. Griffiths. Evaluating Theory of Mind in Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2392–2400, 2018. doi: 10.18653/v1/D18-1261. URL https://aclanthology.org/D18-1261/.

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, pp. 4218–4227, 2018. URL https://proceedings.mlr.press/v80/rabinowitz18a.html.

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*, 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.

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.

Tianmin Shu, Abhishek Bhandwaldar, Chuang Gan, Kevin A. Smith, Shari Liu, Dan Gutfreund, Elizabeth Spelke, Joshua B. Tenenbaum, and Tomer D. Ullman. AGENT: A Benchmark for Core Psychological Reasoning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 9614–9625, 2021. URL https://proceedings.mlr.press/v139/shu21a.html.

Dídac Surís, Sachit Menon, and Carl Vondrick. ViperGPT: Visual Inference via Python Execution for Reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11888–11898, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Suris_ViperGPT_Visual_Inference_via_Python_Execution_for_Reasoning_ICCV_2023_paper.html.

Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, Neel Joshi, and Marc Pollefeys. HoloAssist: an Egocentric Human Interaction Dataset for Interactive AI Assistants in the Real World. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20270–20281, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Wang_HoloAssist_an_Egocentric_Human_Interaction_Dataset_for_Interactive_AI_Assistants_ICCV_2023_paper.html.

Alex Wilf, Sihyun Lee, Paul Pu Liang, and Louis-Philippe Morency. Think Twice: Perspective-Taking Improves Large Language Models' Theory-of-Mind Capabilities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8292–8308, 2024. doi: 10.18653/v1/2024.acl-long.451. URL https://aclanthology.org/2024.acl-long.451/.

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

Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. OpenToM: A Comprehensive Benchmark for Evaluating Theory-of-Mind Reasoning Capabilities of Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8593–8623, 2024. doi: 10.18653/v1/2024.acl-long.466. URL https://aclanthology.org/2024.acl-long.466/.

Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, Bei Ouyang, Zhengyu Lin, Marco Cominelli, Zhongang Cai, Bo Li, Yuanhan Zhang, Peiyuan Zhang, Fangzhou Hong, Joerg Widmer, Francesco Gringoli, Lei Yang, and Ziwei Liu. EgoLife: Towards Egocentric Life Assistant. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28885–28900, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yang_EgoLife_Towards_Egocentric_Life_Assistant_CVPR_2025_paper.html.

Qiutong Tony Yi, Wang Bill Zhu, Robin Jia, and Jesse Thomason. PDDL-Mind: Large Language Models are Capable on Belief Reasoning with Reliable State Tracking. *arXiv preprint arXiv:2604.17819*, 2026. URL https://arxiv.org/abs/2604.17819.

Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-IQ: A Question Answering Benchmark for Artificial Social Intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8807–8817, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Zadeh_Social-IQ_A_Question_Answering_Benchmark_for_Artificial_Social_Intelligence_CVPR_2019_paper.html.

Zhining Zhang, Chuanyang Jin, Mung Yao Jia, Shunchi Zhang, and Tianmin Shu. AutoToM: Scaling Model-based Mental Inference via Automated Agent Modeling. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4318. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/bc3b87855af00be8e63a6d036869cb84-Abstract-Conference.html.

SUPPLEMENTARY MATERIAL CONTENTS

Appendix **Page**

A Implementation and Update Semantics 14
B Video QA Experimental Details 14
C Online Interaction: Protocol and Full Task Results 14
D Symmetric Collaborative Grounding 15
E Controlled Information-History Diagnostic 15
F Traceable Case Examples 16
G Aligned Video, Temporal SSC, and Proxy Rendering 17

A Implementation and Update Semantics

SSC records contain typed values, entity or participant scope, time, status, and evidence references. Video adapters construct records through permitted literal calls; online adapters maintain versioned private stores. Models infer visual and mental-state contents. The runtime checks call structure and, when enabled, evidence-ID membership; it does not verify semantic support.

1. **Compile and answer.** The Spatial compiler receives frames and protocol-permitted question context. The Social compiler receives the resulting records and available visual evidence. The answerer reads both traces, the query, and the original clips in the full-video condition. Candidate scoring is specific to the historical MMToM pipeline in Appendix B.
2. **Observe and act.** Initialize separate agent stores from public and private inputs. Dynamic SSC replaces the latest private observation while archiving its previous value. Successful Pick/Place invalidates the acting agent's old object-location relations and writes supported holding/release/location effects; Navigate records proximity and Open/Close records joint state. Failure is logged without asserting success or invalidating locations. Static retains initialization; Direct omits the state block.
3. **Receive and read.** Only confirmed recipients receive a message record. Mentioned entities index the sender-attributed report; reports are not promoted to observations. Same-key replacements archive prior values and preserve supersession links. The next planner input contains active records with source, turn, status, and confidence.

These are the archived Flash v2 rules. They handle specified relations and replacements, not general semantic contradiction resolution. False visibility, explicit unknown, and an omitted record remain distinct. Source modules are `runtime.py`, `updaters.py`, and `state_store.py`.

B Video QA Experimental Details

MA-EgoQA. Direct and SSC share Gemini 3-Pro-Preview and selected clips. Selection and compilation use question/option information; SSC answers with both traces and original clips, followed by an evidence-sparse raw-video audit.

MMToM-QA main evaluation. Inputs comprise RGB frames, initial-state/action text, and options. SSC uses candidate-conditioned evidence, rule-based bridging, Bayesian scoring, and conditional frame refinement.

MMToM-QA component diagnostic. Section 5.3 uses Qwen3-VL-32B-Instruct on 70 questions from 30 episodes, ten per type, distinct from the main Gemini evaluation. Spatial-only answers from frames and a saved spatial trace; Social-only compiles separately and uses an answer posterior; joint SSC uses the saved runtime prediction.

C Online Interaction: Protocol and Full Task Results

Evaluation and scoring. The Flash suite uses `gemini-3-flash-preview`, state variant v2, seeds 891011/891012/891013, low reasoning, and an 8,192-token output cap per call. Task success requires every physical goal; completion is the fraction satisfied. Placement requires the intended relation and object release in all three samples, separated by ten physics steps of $1/60$ second. Open/closed checks use the default articulated link and threshold 0.10.

Costs and uncertainty. Direct/Static/Dynamic use 176,859/194,312/218,227 logged tokens per rollout. These exclude unlogged failed network attempts. Their per-seed successes are 2/4/2, 4/3/3, and 8/9/7. Paired intervals use 50,000 task-bootstrap draws (seed 20260921), retaining all three seeds per sampled task. Dynamic-minus-Direct intervals are [13.89,72.22] success points and [6.02,43.98] completion points. Figure 3(a) uses shared draws with seed 20260923 for marginal intervals. Table 3 includes all tasks, including regressions.

Table 3: **Gemini 3 Flash: all twelve tasks.** Strict successes out of three seeds and mean physical completion.

	Successes / 3			Mean completion (%)		
Episode	Direct	Static	Dynamic	Direct	Static	Dynamic
1781	1	1	0	66.67	33.33	22.22
1013	1	1	1	58.33	83.33	75.00
623	0	1	3	44.44	77.78	100.00
744	0	0	3	33.33	55.56	100.00
1166	0	0	3	33.33	44.44	100.00
772	1	2	3	83.33	91.67	100.00
588	3	3	2	100.00	100.00	77.78
983	1	2	0	66.67	88.89	66.67
618	0	0	2	33.33	44.44	88.89
619	0	0	2	55.56	66.67	88.89
1526	0	0	2	55.56	44.44	77.78
1566	1	0	3	55.56	55.56	100.00

Figure 3(d) includes all five cooperative tasks (1781, 772, 588, 618, 619) at seed 891011, selected as an exploratory subgroup. Progress is measured after each action, carrying terminal states to turn twenty. Shading is one between-task SD. Mean progress over rounds is SSC/prose 62.42/54.25% in this subgroup.

Implementation and scoring audit. The frozen Flash source includes successful-move invalidation, inactive history, and invalid-measurement marking after failed physics advancement. All 1,885 archived physical measurements across 108 rollouts are marked valid with successful advancement; final scores match the published CSV (`strict_physical_v2`). The audit links configuration and source-file hashes, not independently recovered per-run Git commits. It is an offline consistency check, not a simulator rerun. Earlier updater/protocol results remain separate in the research archive.

D Symmetric Collaborative Grounding

The local CoMind evaluation uses 300 inspected train-development examples (100 per task). Direct and SSC share multi-view RGB frames, task text, and the preceding ten-second transcript. The frozen Gemini 3 Flash generations use temperature zero and a 16,384-token output cap; SSC first compiles Spatial/Social records. SAM 3.1 is applied symmetrically to both frozen final predictions, using the same identity fallback and proposal-selection rules, without gold-label selection. Only boxes change. Joint-attention and handover boxes use the half-scale horizontal mosaic; SCOIA uses helper-view coordinates. Depth is unused.

All examples and generation failures remain included. Localization requires IoU > 0.5. Four-field localization means are Direct 0.75%, Direct+SAM 13.75%, SSC 0.25%, and SSC+SAM 16.00%. The SAM-equipped difference is +2.25 points. All 300 examples pass offline proposal/coordinate replay. Table 4 reports every field, including social outputs. The superseded one-sided SAM batch remains archived rather than mixed with this table.

E Controlled Information-History Diagnostic

We evaluate forty constructed text episodes under four information histories and four conditions, yielding 640 answers. Entity vocabulary and source IDs come from EnactToM; no videos or simulator trajectories are replayed. An event state machine defines world and agent-location targets. The protocol and split were frozen before formal inference. Prose preserves all fields through deterministic conversion; it is not an independently updated memory. The diagnostic uses a factorized StateStore adapter rather than the historical planner serialization.

As reported in Table 2, paired joint correctness is 95.0% for SSC, 92.5% for prose, and 0.0% for both deletion controls. With 10,000 paired bootstrap resamples over base episodes, the SSC–prose differ-

Table 4: All fields in the symmetric CoMind comparison (%). SAM changes boxes only; non-localization fields therefore match their respective ungrounded arms.

Field	Direct	Direct+SAM	SSC	SSC+SAM
Joint attention / left localization	1.00	18.00	0.00	22.00
Joint attention / right localization	1.00	16.00	0.00	23.00
Joint attention / cue type	35.00	35.00	41.00	41.00
Joint attention / category L1	0.00	0.00	13.00	13.00
SCOIA / localization	1.00	10.00	1.00	9.00
SCOIA / cue type	30.00	30.00	21.00	21.00
SCOIA / verb	8.00	8.00	8.00	8.00
SCOIA / noun L1	29.00	29.00	27.00	27.00
Handover / localization	0.00	11.00	0.00	10.00
Handover / delivery flow	54.00	54.00	48.00	48.00
Handover / initiator	52.00	52.00	48.00	48.00
Handover / initiation type	25.00	25.00	26.00	26.00
Handover / category L1	0.00	0.00	0.00	0.00
Handover / time within 0.25 s	5.00	5.00	7.00	7.00

ence is 2.5 percentage points; differences against both deletion controls are 95.0 points (Bonferroni-adjusted 97.5% intervals: [87.5,100.0]). Malformed outputs count as incorrect.

F TRACEABLE CASE EXAMPLES

Figure 4 traces two selected cases. In Habitat 588 (job 0032, seed 891011), a turn-2 message enters state v7–v8 and remains in planner input v11 before successful navigation, while the spatial layer is unchanged.

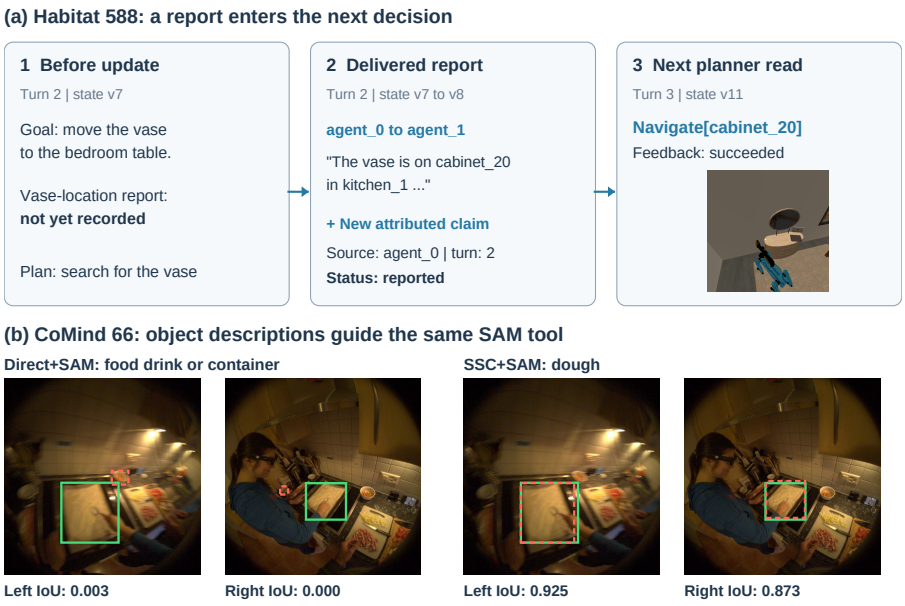

Figure 4: **State maintenance and grounding examples.** (a) Habitat 588: a delivered report remains in the next planner input before successful navigation (Savva et al., 2019; Juneja et al., 2026). (b) CoMind 66: object descriptions guide the same SAM rule (Gavryushin et al., 2026). Green: gold; red dashed: prediction.

G ALIGNED VIDEO, TEMPORAL SSC, AND PROXY RENDERING

Figure 5 aligns four observations from the CoMind `032159` demonstration: visible objects, identity retention through a view shift, colander interaction, and drawer contents becoming visible. Frames 0, 102, 198, and 300 of the supplied 30 fps video match SSC observations 0, 17, 33, and 50 at 5 fps. Each original/proxy pair comes from one decoded frame; center cards excerpt the matching saved record.

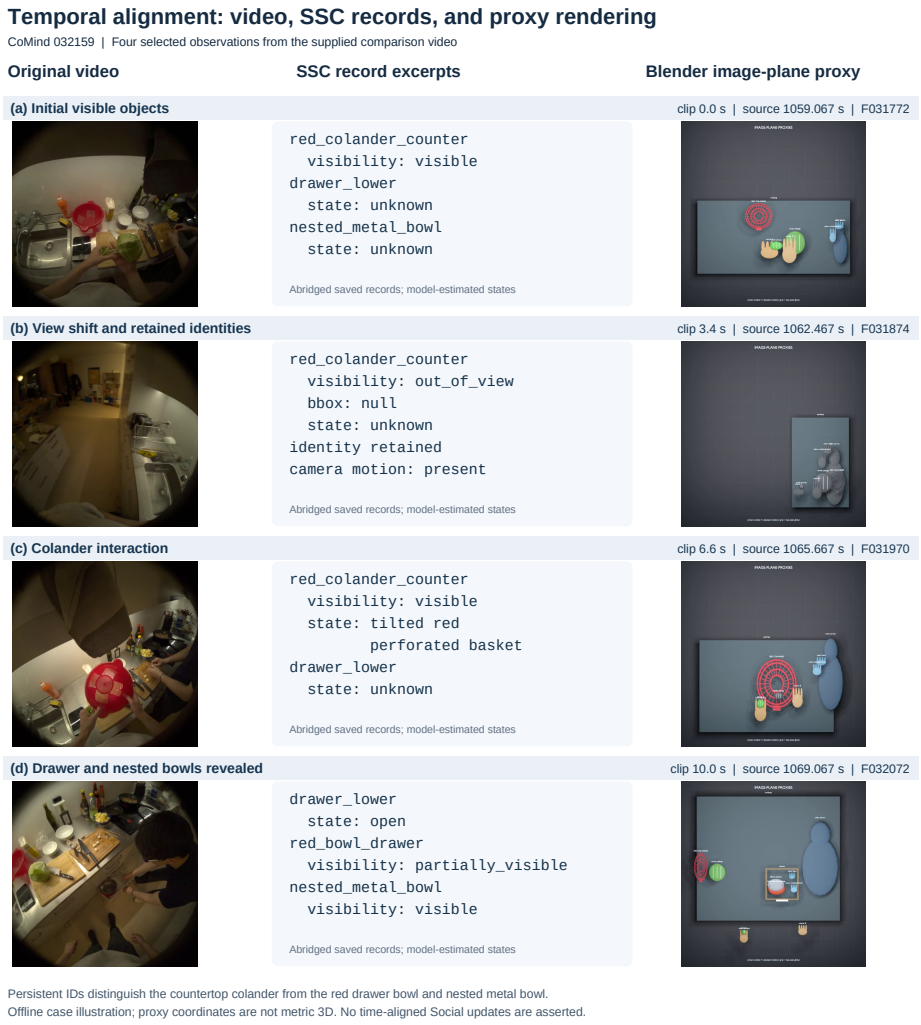

Figure 5: **Aligned video, SSC records, and Blender proxy views.** Rows correspond to clip times 0.0, 3.4, 6.6, and 10.0 s, with source timestamps and evidence IDs. Image panels are extracted without generative modification. Records are offline model estimates; the 2.5D image-plane proxy is not metric reconstruction. Untimed transcript context provides no timed Social updates. This is a qualitative illustration, not a new benchmark result.