An Infant-Cognition Inspired Machine Benchmark for Identifying Agency, Affiliation, Belief, and Intention

Wenjie Li
Center for Data Science
New York University
wenjieli@nyu.edu

Shannon C. Yasuda
Department of Psychology
New York University
shannon.yasuda@nyu.edu

Moira R. Dillon
Department of Psychology
New York University
moira.dillon@nyu.edu

Brenden M. Lake
Department of Psychology and
Center for Data Science
New York University
brenden@nyu.edu

Abstract

Human infants have remarkable abilities to reason about the underlying social causes driving others' actions. These abilities lay the foundation for complex cognitive development, crucial to navigating human social dynamics throughout life. In contrast, Artificial Intelligence (AI) systems continue to fall short in achieving the fundamental commonsense social knowledge present in human infancy. Recent benchmarks focusing on theory of mind and social cognition have begun to address this gap but remain limited in scope. Building on these benchmarks, we introduce eight new tasks focusing on different areas of early social competence, as informed by behavioral studies with infants. We use a self-supervised Transformer model as a baseline test of learning-driven neural-network models on our tasks. Our baseline shows improved performance on existing social-cognitive tasks compared with other deep learning models. Nevertheless, it performs sub-optimally on our new tasks, revealing the challenge of learning complex causal relationships and nuanced human social relations through visual data alone.

Keywords: Social Cognition; Theory of Mind; Deep Learning; Artificial Intelligence; Cognitive Development

Introduction

Human communication, collaboration, and learning are deeply rooted in our ability to understand and interpret the social world around us, including identifying other agents and their affiliations, beliefs, and intentions (Astington & Pelletier, 1998; Krych-Appelbaum et al., 2007; Resches & Pereira, 2007). Even human infants display remarkable proficiency in understanding the social world (J. K. Hamlin, Wynn, & Bloom, 2007; Powell & Spelke, 2013; Sommerville & Crane, 2009; Woodward, 1998). In contrast, deep learning systems often struggle with basic social-cognitive tasks (Lake, Ullman, Tenenbaum, & Gershman, 2017; Marcus & Davis, 2019). Modern deep learning architectures and training paradigms, particularly those focused on supervised learning, tend to reduce behavioral data to labels of classification problems, neglecting the nuances and complexities that factor into human reasoning (Carreira & Zisserman, 2017; Goyal et al., 2017; LeCun, Bengio, et al., 1995). Moreover, although recent large language models succeed in many language-based tasks (OpenAI, 2023), their successes are not robust, for example, to variations of classic social-cognitive tasks like theory-of-mind tasks (Ullman, 2023). By starting from infants' foundational knowledge of the social world, we can begin to identify the building blocks and inductive biases essential for the development of versatile social reasoning, highlighting key elements missing from current AI systems aiming to capture human intelligence.

Initial steps have been taken to bridge the gap between AI and infant social cognition (Gandhi, Stojnic, Lake, & Dillon, 2021; Rabinowitz et al., 2018; Shu et al., 2021). For example, the Baby Intuitions Benchmark (BIB) directly compared the performance of machines and infants on six tasks assessing an observer's inferences about individual agents' goal-directed behaviors (Gandhi et al., 2021; Stojnić, Gandhi, Yasuda, Lake, & Dillon, 2023). BIB offered an important and successful starting point for such a research program, and in doing so it also introduced a framework in which to create new tasks probing foundational social cognition not covered in its initial set of tasks.

Expanding on BIB (Gandhi et al., 2021), we introduce eight benchmark tasks that encompass different elements of infant social cognition, including infants' reasoning about other agents' goals, affiliations, beliefs, and intentions. These tasks are structurally complex, for example, challenging AI systems to track multiple agents' mental states and to differentiate among various passive, goal-directed, and socially driven behaviors. Due to this complexity, the tasks are expected to pose a significant challenge to current AI systems. Alongside our benchmark tasks, we also designed twelve background training tasks to give machines an opportunity to learn the environment and related, but not overlapping, cognitive representations to those tested in the evaluation tasks. As a baseline, we evaluate our tasks on a state-of-the-art Transformer model (Arnab et al., 2021; Vaswani et al., 2017). This model is trained on next-frame prediction, employing a self-supervised paradigm without reliance on synthetic labels or negative examples. While we found that our model performed better than previously tested baselines on BIB's tasks, it made sub-optimal predictions on our new tasks, exposing weaknesses in its abilities to understand complex causal relations and to track the mental states of multiple agents.

Below, first, we provide comprehensive explanations of each of our new tasks. Second, we describe our background training tasks. Third, we discuss our Transformer baseline

model, detailing its architecture, training methodology, and performance evaluation on both BIB's tasks and our new tasks. Finally, we discuss the broader implications of our tasks to inspire more human-like AI, considering the weaknesses and strengths of existing modeling methodology.

Benchmark Tasks

Designed within the framework of BIB (Gandhi et al., 2021), our tasks use videos of 2D shapes moving in a grid world to represent rich social interactions among animate agents (Heider & Simmel, 1944). This design eliminates the vision challenges associated with naturalistic scenes, probing a machine's ability to learn higher-level cognitive representations. (Gordon, 2016; Springer, Meier, & Berry, 1996). Moreover, our design streamlines the engineering process for synthesizing the thousands of videos necessary to train and test models.

Like BIB, the benchmark relies on a violation-of-expectation (VOE) looking-time paradigm, commonly used to test infants. For infants, the VOE paradigm uses looking time to measure their implicit predictions: Infants tend to look longer at events they find surprising (Spelke, 1985). To adopt this paradigm for our benchmark, we constructed episodes consisting of nine trials within the same environment. The first eight trials—familiarization—establish a consistent expectation, with events drawn from the same statistical distribution. The ninth trial—test—introduces an event that either conforms to or contradicts the expectation set up by familiarization. To best engage with how this paradigm is used with infants, our baseline model produces "surprisal scores" based on the discrepancy between their predicted outcome and the expected or unexpected test trial.

Our benchmark includes two tasks focused on social affiliations among agents (Approach Tasks), two focused on the attribution of goals to agents, not objects (Object Goal Tasks), two focused on the attribution of beliefs to agents (False Belief & True Belief), and two focused on agents' helping and hindering behaviors (Helping & Hindering). Each task consists of 1000 episodes. Below we provide further detail about each task, explaining their structure and criterion for success.

Approach: Social & Instrumental

Does AI expect an agent to imitate the actions of another agent it had affiliated with?

Developmental Background. Infants predict that members of the same social group will exhibit similar actions (Powell & Spelke, 2013, 2018). For example, when Powell and Spelke (2013) had 8-month-old infants observe two groups of geometric figures with eyes maintain close proximity to and perform "dance" movements with their group, the infants were surprised when one group member subsequently performed the same actions as a member of the other group instead of members of their own group. Unpublished research outlined in Spelke (2022) suggests that infants as young as 7.5 months were only surprised by such group-inconsistent actions when those actions were non-causal. When the actions included contacting an object, changing its color, infants formed no expectations about an individual agent's actions.

Familiarization Trials. An agent approaches one of two target agents to establish social affiliation (Figure 1a&b).

Test Trials. The two target agents each sequentially move in unique patterns (Figure 1a&b). The imitating agent then adopts the movement pattern of either the target agent it had previously approached or the target it had previously not approached. In the Instrumental task, a new target object and obstacles are strategically placed near the imitating agent (Figure 1b), making the movement of the agent the only efficient action to reach the target object. Here, observers should have no expectation regarding which target the imitating agent mimics, since any similarity in actions could be coincidental, stemming from the agent's goal-directed behavior. Conversely, in the Social task (Figure 1a), the agent is not constrained and a target object is not present. It is expected to mimic the target agent it previously affiliated with.

Object Goal: Agent & Object

Can an AI system recognize and attribute goals to an agent that displays self-propelled, efficient motion, but not an object that only moves upon contact with a mechanical spinner?

Developmental Background. Infants recognize that agents, but not objects, exhibit self-propelled motions (Cicchino & Rakison, 2008), have object-based goals (Woodward, 1998), and move rationally and efficiently toward their goals (Csibra, Gergely, Bıró, Koos, & Brockbank, 1999). For example, Woodward (1998) found that 5-month-old infants expected a hand, but not a mechanical claw to reach consistently for a goal object, not to a goal location.

Familiarization Trials. Each video contains a constantly rotating spinner, an ambiguous element that acts as either a goal-driven agent or passive object, and two static target objects (Figure 1c&d). In Object Goal: Agent (Figure 1c), the element, positioned a short distance away from the spinner, initiates its own movement. In Object Goal: Object (Figure 1d), the element begins moving only after contact with the spinner, in a direction perpendicular to the spinner's arm at the point of contact. In both scenarios, the element moves until it collides with one of the target objects, both of which change color upon impact. The element consistently collides with the same target object at a similar location in each episode. A gray square under the spinner ensures visual consistency between familiarization and test trials.

Test Trials. At test, (Figure 1c&d), the target objects' locations are swapped. A gray square obscures the element's initial position. This ensures ambiguity in its movement cause, requiring the element's agency to be inferred from familiarization. At the start of each trial, the element emerges from behind the square, moving straight toward either the same target object now in a new location (expected in Object Goal: Agent, no expectation in Object Goal: Object), or a different object in the location it previously approached (unexpected in Object Goal: Agent, no expectation in Object Goal: Object).

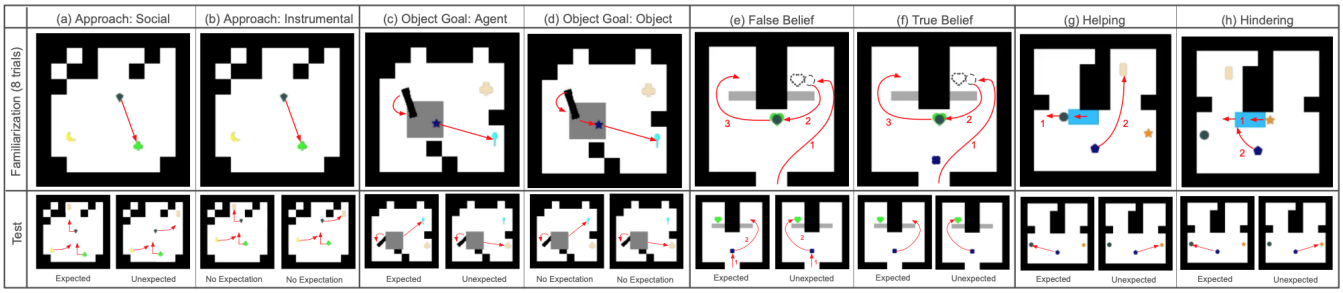

Figure 1: **Schematic Overview of the Benchmark Tasks.** Eight familiarization trials are first shown to set up an expectation about the underlying agency, affiliations, beliefs, or intentions driving the behavior of each moving entity. At test, two events are played, one consistent and one inconsistent with the previous expectations. Here, red arrows indicate the movement of entities. Where relevant, numbers indicate the sequential order of these movements. For clarity, this figure only partially represents the familiarization trials for (e) through (h).

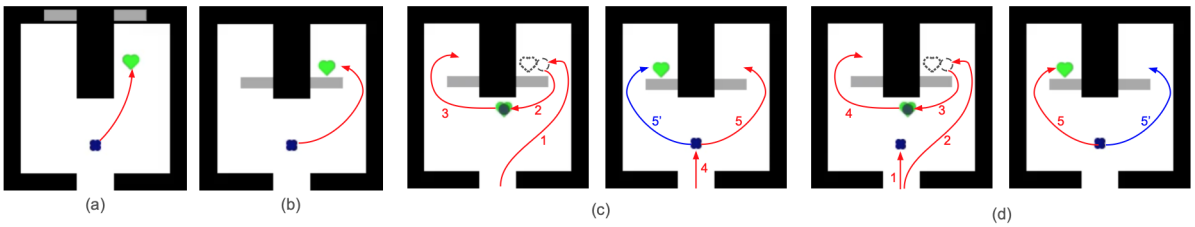

Figure 2: **Schematic of the False Belief and True Belief tasks.** Arrows indicate the direction of movements. When applicable, numbers show the orders of the actions. In the first trial (a), a clover-shaped agent approaches an observable heart-shaped goal object. In subsequent familiarization trials (b), the agent searches for its goal in the same room, even when grey occluders obstruct its view. At test, (c) in the False Belief task, when the clover-shaped agent is absent, a circular agent moves the goal to the other room (2,3) before leaving. The clover-shaped agent enters (4) and either goes to the original room, failing to find the goal object (red arrow, 5: expected), or to the room where the object was moved (blue arrow, 5': unexpected). (d) In the True Belief task, the clover-shaped agent enters the grid world (1) before the circular agent enters (2), witnessing the object change location (3,4). Hence it is expected to search the new room for its goal (red arrow, 5).

False Belief & True Belief

Does AI understand that an agent can hold and act on a belief that is inconsistent with reality?

Developmental Background. Young toddlers predict that an agent will act on an object based on their beliefs about where that object is (Onishi & Baillargeon, 2005; Scott & Baillargeon, 2017), and toddlers prefer agents who intend to help based on their beliefs about the state of the world, whether those beliefs are true or false. For example, Woo & Spelke (2023) showed that 15-month-old toddlers prefer agents who intend to help other agents even if their intention is not fulfilled (i.e., the helping agent had a false belief about the location of the target agent's preferred object).

Familiarization Trials. In the first trial, an agent moves toward an observable goal object located in one of two rooms (Figure 2). In the following trials, the goal object is always placed within the same room, but depending on the placement of two grey occluders, the agent may or may not be able to observe the object's location (Figure 2a&b). The agent always moves to the same room to find the object, establishing that the agent looks for the goal object where it was last seen.

Test Trials. The goal object is initially located within the same room, but a second agent switches its location to the other room. In the False Belief Task, the first agent observes the object change location. In the True Belief Task, the first agent is not present until after the switch, failing to observe the object change location. The first agent searches for its goal in the original location (Expected for False Belief, Unexpected for True Belief), or the new location (Unexpected for False Belief, Expected for True Belief).

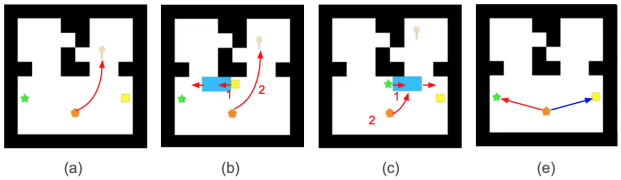

Figure 3: **Schematic of the Helping and Hindering tasks.** Arrows indicate the direction of movements. Where present, numbers show the sequence of actions. In the depicted scenario, a pentagonal agent moves toward a spoon-shaped goal (a). In the Helping task (b), a clover-shaped agent assists by removing an obstacle, facilitating the pentagonal agent's goal attainment. Conversely, in the Hindering task (c), a star-shaped agent impedes the pentagonal agent by placing an obstacle in its path. In both tasks (e), the pentagonal agent is expected to approach the clover-shaped agent (red arrow) and it is unexpected to approach the star-shaped object (blue arrow).

Helping & Hindering

Can AI infer that a goal-directed agent prefers to socially interact with another agent who helps it, as opposed to one that hinders it, in achieving its goal?

Developmental Background. Infants predict that agents will approach others who help them achieve their goals (Fawcett & Liszkowski, 2012; J. Hamlin, 2015; J. K. Hamlin & Wynn, 2011; K. J. Hamlin, Wynn, & Bloom, 2010; Kuhlmeier, Wynn, & Bloom, 2003; Lee, Yun, Kim, & Song, 2015; Premack & Premack, 1997). For example, J. K. Hamlin (2013) found that 10-month-old infants were more likely to reach for a puppet who removed a door blocking a target puppet's preferred object instead of reaching for a different puppet who also removed a door, but one blocking the target puppet's nonpreferred object. Moreover, infants showed this preference only when the two puppets moving the doors saw which object the target puppet preferred, not when they were naive of the target puppet's preference.

Familiarization Trials. During the first four familiarization trials (Figure 3a), two agents observe a goal-directed agent approach a goal object which changes color upon contact. In the next four familiarization trials, a barrier is placed in the environment. In the Helping task (Figure 3b), a helping agent removes the barrier obstructing the goal-directed agent from reaching its goal, while the other agent does nothing. In the Hindering task (Figure 3c), a hindering agent moves the barrier to obstruct the entrance, preventing the agent from reaching its goal, while the other agent does nothing.

Test Trials. At test, the goal and barriers are removed. It is expected that the goal-directed agent moves toward the helper in the Helping task and moves toward the stationary agent who does not hinder in the Hindering task.

Background Training Set

Infants bring knowledge to developmental tasks, unlike deep learning systems that start with much more limited inductive biases. We provide a background training set to offer machines a learning opportunity. We will show four example tasks in this section. See Figure A.1 for the full set. This set presents elements of the environment and task setup, providing understanding essential to solve the benchmark, but irrelevant to social cognition. For example, a rotating spinner that changes the motion of an object it contacts (Figure 4a&b). The single object collection (Figure 4c) and occluded goal migration tasks (Figure 4c) demonstrate navigation constraints.

Background tasks also introduce the cognitive abilities required to succeed at the evaluation tasks. These training tasks are carefully designed to contain low-level visual signals and movement patterns that are distinct from evaluation, preventing trivial task solution through pattern matching. For instance, the Occlusion: No-Navigation Preference task (Figure 4b) demonstrates an agent's consistent preference for an object, which it approaches without navigational maneuvers. The Single Object Collect task illustrates efficient navigation, avoiding obstacles. Employing this knowledge, a machine may infer that a goal-driven agent is likely to navigate with efficiency toward a preferred object among several options.

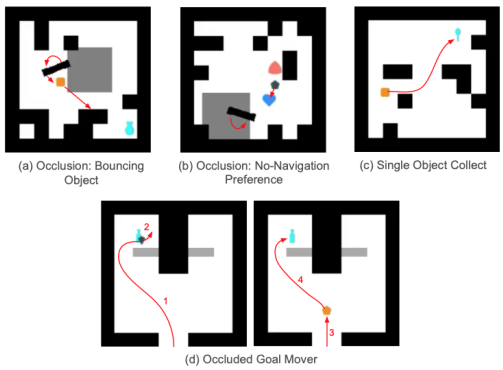

Figure 4: **Selected Background Training Tasks.** Models are trained on 12 background tasks that introduce the environmental dynamics and cognitive skills necessary to solve evaluation tasks.

Notably, similar to infants' experiences, the training set consists of expected examples only. While it is challenging to match the complexity of infants' real-world experiences, we believe this training set offers a limited yet reasonable foundation for meaningful comparison. We encourage the usage of additional data sources to enhance machines' performance.

Baseline Model

We propose a baseline model aimed at predicting the subsequent frame in a test trial, given the preceding frames and a selected familiarization trial as context. We choose a Transformer architecture for its efficacy in processing sequential data, as illustrated in Figure 5 (Vaswani et al., 2017; Dosovitskiy et al., 2020; Arnab et al., 2021). The core challenge for the Transformer is to learn to represent temporal-spatial continuities, causal relationships, and key concepts in social reasoning (Zhou, Dong, & El Saddik, 2020; Gandhi et al., 2021; Hein & Diepold, 2022).

Data Preparation

In the dataset, each task is contained in one video, which is segmented into nine trials–eight familiarization trials followed by one test trial. At inference, each of the eight familiarization trials is paired with its test trial, resulted in eight predictions. During one training iteration, one pair of a familiarization trial and a test trial are randomly chosen from any of the nine trials in a video, providing the model with a wider range of training data. Frames are sampled from each trial with a stride of 25 frames, with an upper limit of 20 frames for a single trial.

Model Architecture

As shown in Figure 5, a three-layer CNNs transforms each video frame into a series of patch embeddings. These embeddings are then augmented with sinusoidal temporal-spatial

positional encodings to preserve their original context within the video. The processed patches from the familiarization trial are then fed into a Transformer encoder, establishing the context for subsequent predictions in the test trial. To predict the t^{th} frame in the test sequence, the Transformer decoder receives the patch embeddings from all preceding test frames up to the $(t-1)^{th}$ frame and uses cross-attention to incorporate the encoded patches from the familiarization trial. Finally, a two-layer deconvolutional network transforms the decoder output into RGB image. Refer to Appendix B for comprehensive architectural details and configurations.

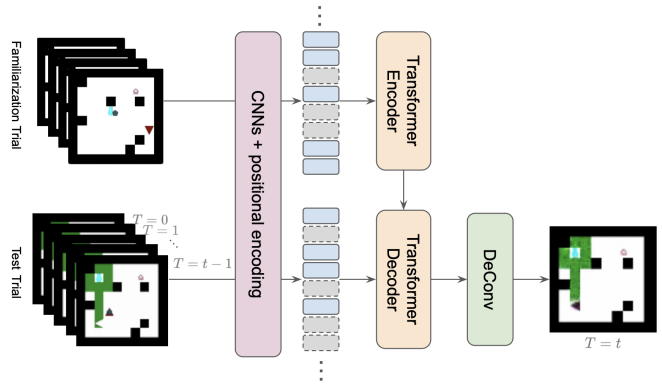

Figure 5: **Model Architecture and Training Procedure.** The Transformer model predicts the test frame at time t, given the frames from a familiarization trial and test frames from time 0 to $t-1$.

Training

The training objective for the network is to minimize the mean square error (MSE) loss between the predicted frames and target frames on a pixel-wise basis. The model is trained with the background training sets in BIB as well as ours. Refer to Appendix B for learning rate, weight decay, and other hyperparameters. After training for 100 epochs, the model yields a final loss of 5.5×10^{-4} on the validation set, underscoring the model's predictive accuracy. For comparison, the baseline MSE between two successive frames averaged 2.6×10^{-3}. Visualizations of model predictions on the validation set (Figure 6a) reveal that the model is capable of predicting the color, approximate shape, and step size of an agent. It also learns some higher-level concepts, including that agent has object goal instead of location goal. (Figure 6b).

Figure 6: **Example Prediction on a Held-out Task.** a clover-shaped agent navigates to the bottle-shaped goal object. Left: previous frame, Middle: current frame (target), Right: model prediction.

Alternative Models

To demonstrate the capacity of our baseline Transformer model, we first compare its performances with models previously tested on BIB. These models employ different architectures, training approaches, oracle information and built-in knowledge (Table 1). BC-MLP and BC-RNN are behavioral cloning models with the training objective to predict future coordinates of the main agent (Gandhi et al., 2021). Conversely, Video-RNN processes and constructs a pixel image of the next frame. The VT model uses Transformer-like attention mechanisms to predict both the next frame and coordinates (Hein & Diepold, 2022). Finally, HBToM predicts the coordinates of the agent by leveraging a hierarchical Bayesian approach to construct relevant cognitive functions in a probabilistic model (Zhi-Xuan et al., 2022). Prediction accuracies of these models are shown in Table 2.

Notably, Video-RNN and Transformer use only video frames as input. This provides versatility facilitating easy adaptation to new tasks (Cholakov & Kolev, 2021; Girdhar & Grauman, 2021). The increased complexity in our tasks poses a challenge to models tailored for specific, predefined tasks in BIB. For example, BC-MLP, BC-RNN, HBToM, and VT rely on input or predictions based on coordinates of one agent, making it difficult to adapt to a multi-agent environment. Additionally, the VT model in Hein and Diepold (2022) limits its processing to only three patches with the most significant changes to manage its computational demands. However, this method falls short of capturing multiple dynamic elements within a frame. Finally, to solve the new tasks, HBToM in Zhi-Xuan et al. (2022) requires a new set of built-in common sense knowledge related to the new tasks.

Evaluation

Each of a task's eight familiarization trials is paired with its test trial, resulting in eight prediction losses. A prediction is deemed successful if the average loss for the expected video is lower than that for the unexpected video. The performance of the baseline model on the new tasks is recorded in Table 3, alongside its performance on BIB in Table 2. We discuss the predictions of the baseline model in detail to illustrate its capabilities and limitations.

Goal Preference. VT and the baseline model significantly outperform MLP- and RNN-based models in goal preference and inaccessible goal tasks. As shown in Table 2, VT scores accuracies of 82.1% and 89.8%, respectively, and our Transformer model recorded accuracies of 73.7% and 78.8%. This underscores the strength of the attention mechanism in representing element-wise relations.

Efficiency. All models successfully predicted that agents efficiently navigate to their goals, with accuracy above 90%. However, the Irrational Agent task is much more challenging for many models. Notably, the transformer baseline achieves an accuracy of 80.4% on this task, far surpassing the other data-driven models including the Video-RNN (50.1%) and VT (29.5%). One explanation is that many models learn to

Table 1: **Oracle information used by different models.**

	Deep Learning Models					Bayesian Principled Model
Privileged Information	BC-MLP	BC-RNN	Video-RNN	VT	Transformer (Ours)	HBToM
Environment meta-data (element type, coordinates, etc.)	x	x		x		x
Built-in inductive biases				x		x

Table 2: **Comparisons of model prediction accuracy (%) on BIB.**

	Deep Learning Models					Bayesian Principled Model
Task Name	BC-MLP	BC-RNN	Video-RNN	VT	Transformer (Ours)	HBToM
Preference	26.3	48.3	47.6	82.1	73.7	99.7
Multi-Agent	48.7	48.3	50.3	49.1	50.2	99.2
Inaccessible Goal	73.3	80.7	71.8	78.9	78.9	99.7
Efficiency: Path Control	94.0	92.8	99.2	96.0	96.0	94.9
Efficiency: Time Control	99.1	99.1	99.9	99.0	99.9	97.2
Efficiency: Irrational Agent	73.3	55.7	50.1	29.5	80.4	96.6
Instrumental: No Barrier	98.8	98.8	99.7	98.7	97.9	98.8
Instrumental: Inconseq Barrier	55.2	78.2	77.0	96.9	57.3	97.0
Instrumental: Blocking Barrier	47.2	56.6	62.5	82.1	21.4	99.7

Table 3: **Baseline prediction accuracy (%) on the new social cognition tasks.**

Task Name	Baseline
Approach: Social	38.6
Approach: Instrumental	50.2
Goal Attribution: Agent	40.8
Goal Attribution: Object	53.5
True Belief	97.5
False Belief	2.4
Helping	60.7
Hindering	58.3

solve the Efficiency tasks independently of the familiarization trials (Gandhi et al., 2021; Hein & Diepold, 2022). Given the set of background training tasks we provide, it is possible for the model to learn efficient navigation as a general knowledge, applying it to all predictions without context. This strategy, while effective for background training trials, fails for the Irrational Agent task where the context is critical. One possible explanation for the relative success of our Transformer model is in the way the background training trials are sampled: During training, any trial, including the first eight trials, can be treated as the "test trial," potentially guiding the model to learn a more complete representation of the familiarization trials.

Instrumental Action. The baseline model makes accurate predictions (97.9%) in the No Barrier task, slightly above chance (57.3%) in the Inconsequential Barrier task and below chance (21.4%) in the Blocking Barrier task. An examination of these complementary tasks suggests that the model fails to capture the sequential and causal relations in the actions. For example, in the model predictions of an Instrumental: Blocking Barrier task (Figure 7), the green wall starts fading before the key insertion. This is because the model fails to capture the causal relations during background training, instead associating the fading of the green wall with frame numbers. This heuristic fails in evaluation because the agent is generated farther away from the key, which is inserted in later frames. This results show the weakness of Transformer to understand and represent causal relations.

Figure 7: **Example Predictions on an Instrumental: Blocking Barrier Task.** In the second predicted frame of this trial (left), a cloved-shape agent moves towards the triangular key; In the third frame (middle), the agent picks up the key, and the lower-right corner of the green wall is gone; In the fifth frame (right), the agent and the key get closer to the lock but the key has not been inserted yet. The lower-right corner of the green wall is again missing.

Helping & Hindering. As shown in Table 3, in the Helping and Hindering tasks, the model achieves an accuracy of 60.7% and 58.3%, respectively. A closer examination of the frame prediction suggests that the model might be capable of identifying the target agent, but fails to render it correctly to match the target frame. As shown in Figure 8a, in the target frame (center), the blue pentagonal agent is at the lower-right corner of the orange circular agent, because the pentagonal agent approaches from below (left). In the predicted frame (right), the model correctly renders the blue pentagonal agent near the orange circular agent, but at the upper-right corner instead of the lower-right corner in the target. This discrepancy can be traced back to the lack of variety in some background training tasks. In the Trapped Agent: Hindering task from the training set, for example, a goal-directed

agent always starts from the top part of the grid world and approach one of the two social agents from above (Figure 8b). This shows a lack of diversity in the background training set, as well as Transformer's challenges with generalization from limited data.

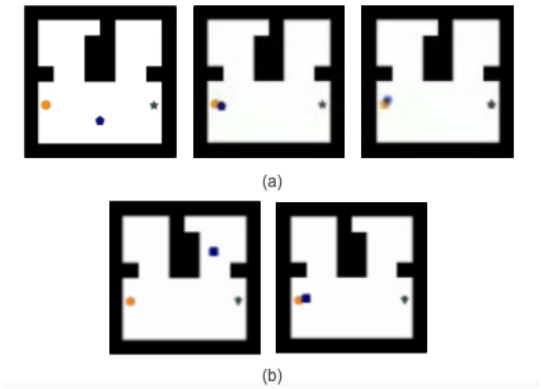

Figure 8: **Example Prediction on a Hindering Task.** In a Hindering task in evaluation (a), a pentagonal agent starts from the lower part of the grid world (a, left), moves towards and finally collides with a circular agent from below (a, center). In the model prediction (a, right), the pentagonal agent appears near the upper-left corner of the circular target agent. In a similar training episode of the Trapped Agent: Hindering task (b), a clove-shaped agent starts from the upper-right room of the grid world (b, left) and approach a circular target agent from above (b, right).

True & False Belief. The baseline Transformer achieves accuracy of 97.5% and 2.4% respectively, indicating that it expects the agent to always approach the goal behind the wall, regardless of its knowledge about the relocation. This could be attributed to either a genuine ignorance of false belief, or a failure to understand the role of the occluder to create an environment with only partial knowledge.

Discussion

Machine benchmarks focusing on human social cognition create valuable opportunities to develop AI systems with human-like and human-compatible competencies. By comparing the predictions of AI and infants, we can align the goals of AI systems with the foundational cognitive building blocks of human cognition. Our present work builds on previous work, in particular on the Baby Intuitions Benchmark (Gandhi et al., 2021; Stojnić et al., 2023), by introducing eight new evaluation tasks that explore various social-cognitive abilities present from human infancy, including reasoning about agency, affiliation, belief, and intention. We also generated twelve background training tasks to provide machines an opportunity to learn the environmental dynamics and necessary social cognitive groundwork.

We supply a lower-bound for machine performance with a Transformer baseline. The encoder-decoder model is trained with a self-supervised learning paradigm. Its non-task-specific architecture and training procedure provide an adaptable pipeline for future datasets structured around the VOE paradigm. Despite showing promise on the existing BIB with minimal oracle guidance (Table 1), its performance on our new tasks highlights the necessity for more powerful AI systems capable of reasoning about complex environments and nuanced social dynamics.

What can we learn from existing models of social cognition, and how do we create AI that can identify agency, affiliation, belief, and intention like infants do? Our results point to two complemetary strategies: creating more realistic training data and integrating cognitive inductive biases into model architectures. The disparity between the simplistic, passive learning environment we provided and the rich, multi-modal, and interactive experiences that shape infant learning is pronounced. Efforts to bridge this gap have included capturing infants' sensory experiences through head-mounted cameras (Vong, Wang, Orhan, & Lake, 2024; Emin Orhan, Wang, Wang, Ren, & Lake, 2024; Orhan, Gupta, & Lake, 2020; Sullivan, Mei, Perfors, Wojcik, & Frank, 2021), eye-tracking (Sheybani, Hansaria, Smith, & Tiganj, n.d.; Mendez, Yu, & Smith, n.d.; Candy et al., 2023), and simulating interaction with the environment via embodied agents (Wykowska, Chaminade, & Cheng, 2016). Our benchmark is poised to serve as a critical testing ground for models trained on these datasets.

Moreover, the innate knowledge and biases that infants possess facilitate flexible adaptation and efficient generalization, a balance that existing models of commonsense reasoning struggle to achieve. Existing models fall into two categories: structured Bayesian models, such as BIPaCK and HBToM in Shu et al. (2021), and deep learning models, such as neural networks for behavioral cloning (Gandhi et al., 2021) and video prediction(Gandhi et al., 2021; Hein & Diepold, 2022). Structured Bayesian models can perform very well on synthetic benchmarks, but they usually rely on task-specific inductive biases and features that would generalize poorly to the fully complexity of real applications. On the other hand, deep learning models adopt an end-to-end approach, offering greater robustness in noisy environments, but they are data intensive and generally under-perform when confronted with out-of-distribution data, as we find in our study (Table 2). With the long-term goal of developing machines that have infant-like social reasoning, we hope that our benchmark stimulates new work that expands both modeling approaches, as well as new approaches that combine the strengths of existing ones.

References

Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., & Schmid, C. (2021). Vivit: A video vision transformer. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 6836–6846).

Astington, J. W., & Pelletier, J. (1998). The language of mind: Its role in teaching and learning. *The handbook of education and human development: New models of learning, teaching and schooling*, 569–593.

Candy, T. R., Dalessandro, A., Tellez, V., Biehn, S., Mestre, C., Haaff, T., ... Smith, L. (2023). The distribution of gaze positions of human infants in natural behavior. *Journal of Vision*, *23*(9), 4999–4999.

Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the ieee conference on computer vision and pattern recognition* (pp. 6299–6308).

Cholakov, R., & Kolev, T. (2021). *Transformers predicting the future. applying attention in next-frame and time series forecasting.*

Cicchino, J. B., & Rakison, D. H. (2008). Producing and processing self-propelled motion in infancy. *Developmental Psychology*, *44*(5), 1232.

Csibra, G., Gergely, G., Bıró, S., Koos, O., & Brockbank, M. (1999). Goal attribution without agency cues: the perception of 'pure reason'in infancy. *Cognition*, *72*(3), 237–267.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... others (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Emin Orhan, A., Wang, W., Wang, A. N., Ren, M., & Lake, B. M. (2024, January). Self-supervised learning of video representations from a child's perspective. *arXiv e-prints*, arXiv:2402.00300. doi: 10.48550/arXiv.2402.00300

Fawcett, C., & Liszkowski, U. (2012). Observation and initiation of joint action in infants. *Child Development*, *83*(2), 434–441.

Gandhi, K., Stojnic, G., Lake, B. M., & Dillon, M. R. (2021). Baby intuitions benchmark (bib): Discerning the goals, preferences, and actions of others. *Advances in neural information processing systems*, *34*, 9963–9976.

Girdhar, R., & Grauman, K. (2021, October). Anticipative video transformer. In *Proceedings of the ieee/cvf international conference on computer vision (iccv)* (p. 13505-13515).

Gordon, A. (2016). Commonsense interpretation of triangle behavior. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 30).

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., ... He, K. (2017). Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*.

Hamlin, J. (2015). The case for social evaluation in preverbal infants: gazing toward one's goal drives infants' preferences for helpers over hinderers in the hill paradigm. *Frontiers in psychology*, *5*, 1563.

Hamlin, J. K. (2013). Moral judgment and action in preverbal infants and toddlers: Evidence for an innate moral core. *Current Directions in Psychological Science*, *22*(3), 186–193.

Hamlin, J. K., & Wynn, K. (2011). Young infants prefer prosocial to antisocial others. *Cognitive development*, *26*(1), 30–39.

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature*, *450*(7169), 557–559.

Hamlin, K. J., Wynn, K., & Bloom, P. (2010). Three-month-olds show a negativity bias in their social evaluations. *Developmental science*, *13*(6), 923–929.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American journal of psychology*, *57*(2), 243–259.

Hein, A., & Diepold, K. (2022). Comparing intuitions about agents' goals, preferences and actions in human infants and video transformers. In *Svrhm 2022 workshop @ neurips.*

Krych-Appelbaum, M., Law, J. B., Jones, D., Barnacz, A., Johnson, A., & Keenan, J. P. (2007). "i think i know what you mean": The role of theory of mind in collaborative communication. *Interaction Studies*, *8*(2), 267–280.

Kuhlmeier, V., Wynn, K., & Bloom, P. (2003). Attribution of dispositional states by 12-month-olds. *Psychological science*, *14*(5), 402–408.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and brain sciences*, *40*, e253.

LeCun, Y., Bengio, Y., et al. (1995). Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, *3361*(10), 1995.

Lee, Y.-e., Yun, J.-e. E., Kim, E. Y., & Song, H.-j. (2015). The development of infants' sensitivity to behavioral intentions when inferring others' social preferences. *PLoS One*, *10*(9), e0135588.

Marcus, G., & Davis, E. (2019). *Rebooting ai: Building artificial intelligence we can trust*. Vintage.

Mendez, A. H., Yu, C., & Smith, L. B. (n.d.). Controlling the input: How one-year-old infants sustain visual attention. *Developmental Science*, e13445.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *science*, *308*(5719), 255–258.

OpenAI. (2023). *ChatGPT*. (https://chat.openai.com/chat)

Orhan, E., Gupta, V., & Lake, B. M. (2020). Self-supervised learning through the eyes of a child. *Advances in Neural Information Processing Systems*, *33*, 9960–9971.

Powell, L. J., & Spelke, E. S. (2013). Preverbal infants expect members of social groups to act alike. *Proceedings of the National Academy of Sciences*, *110*(41), E3965–E3972.

Powell, L. J., & Spelke, E. S. (2018). Third-party preferences for imitators in preverbal infants. *Open Mind*, *2*(2), 61–71.

Premack, D., & Premack, A. J. (1997). Infants attribute value±to the goal-directed actions of self-propelled objects. *Journal of cognitive neuroscience*, *9*(6), 848–856.

Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S. A., & Botvinick, M. (2018). Machine theory of mind. In *International conference on machine learning* (pp. 4218–4227).

Resches, M., & Pereira, M. P. (2007). Referential communication abilities and theory of mind development in preschool children. *Journal of Child Language*, *34*(1), 21–

52.
Scott, R. M., & Baillargeon, R. (2017). Early false-belief understanding. *Trends in Cognitive Sciences*, *21*(4), 237–249.
Sheybani, S., Hansaria, H., Smith, L. B., & Tiganj, Z. (n.d.). Curriculum learning with infant egocentric videos..
Shu, T., Bhandwaldar, A., Gan, C., Smith, K., Liu, S., Gutfreund, D., … Ullman, T. (2021). Agent: A benchmark for core psychological reasoning. In *International conference on machine learning* (pp. 9614–9625).
Sommerville, J. A., & Crane, C. C. (2009). Ten-month-old infants use prior information to identify an actor's goal. *Developmental science*, *12*(2), 314–325.
Spelke, E. S. (1985). Preferential-looking methods as tools for the study of cognition in infancy.
Spelke, E. S. (2022). *What babies know: Core knowledge and composition volume 1* (Vol. 1). Oxford University Press.
Springer, K., Meier, J. A., & Berry, D. S. (1996). Nonverbal bases of social perception: Developmental change in sensitivity to patterns of motion that reveal interpersonal events. *Journal of Nonverbal Behavior*, *20*, 199–211.
Stojnić, G., Gandhi, K., Yasuda, S., Lake, B. M., & Dillon, M. R. (2023). Commonsense psychology in human infants and machines. *Cognition*, *235*, 105406.
Sullivan, J., Mei, M., Perfors, A., Wojcik, E., & Frank, M. C. (2021). Saycam: A large, longitudinal audiovisual dataset recorded from the infant's perspective. *Open mind*, *5*, 20–29.
Ullman, T. (2023). Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.
Vong, W. K., Wang, W., Orhan, A. E., & Lake, B. M. (2024). Grounded language acquisition through the eyes and ears of a single child. *Science*, *383*(6682), 504-511. doi: 10.1126/science.adi1374
Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition*, *69*(1), 1–34.
Wykowska, A., Chaminade, T., & Cheng, G. (2016). Embodied artificial agents for understanding human social cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1693), 20150375.
Zhi-Xuan, T., Gothoskar, N., Pollok, F., Gutfreund, D., Tenenbaum, J. B., & Mansinghka, V. K. (2022). Solving the baby intuitions benchmark with a hierarchically bayesian theory of mind. *arXiv preprint arXiv:2208.02914*.
Zhou, Y., Dong, H., & El Saddik, A. (2020). Deep learning in next-frame prediction: A benchmark review. *IEEE Access*, *8*, 69273–69283.

A. Full Set of Background Training Tasks

Like the evaluation tasks, each background training task is consisted of eight familiarization trials and one test trial. Each task type is described below.

a. Occlusion: Object Collection As shown in Figure A.1a, an agent navigates efficiently, avoiding walls and a rotating spinner, to reach a goal object, which changes color at impact. A grey square-shaped occluder appears under all elements in familiarization and above at the last trial.

b. Occlusion: Bouncing Object As shown in Figure A.1b, a spinner rotates and comes into contact with a passive object. At impact, the object which was previously still begins moving in a straight line perpendicular to the spinner. The object stops moving when it hits a wall or when it contacts another goal object, which changes color at impact. A grey square-shaped occluder lays under all elements in familiarization and on top of them at test time.

c. Occlusion: No-Navigation Preference As illustrated in Figure A.1c, an agent consistently reaches one of two proximal target objects, with varying locations across trials, demonstrating a preference to a goal object, instead of a goal location. Upon contact, both objects change color, eliciting cues for causal efficacy that distinguish goal-approach behavior from a pro-social approach. A spinning object is in the scene and do not interact with other elements. A grey square-shaped occluder lays under all elements in familiarization and on top of them at test time.

d. Occlusion: No-Navigation Bouncing Object As shown in Figure A.1d, a rotating spinner propels an element toward one of two target objects, which change color upon contact. The target object varies between trials, depending on the spinner's initial rotation and the element's position, illustrating that passive movement driven by another object does not necessarily signify goal preference.

e. Single Object Collect An agent navigates through obstacles to reach a single target object, which changes color upon contact (Figure A.1e). This task demonstrates obstacle navigation and goal pursuit. In some but not all episodes, the agent returns to the starting point after interacting with the object.

f. Moving Object As shown in Figure A.1f, an agent approaches a goal object and picks it up, which appears as the agent above and in the center of the object. The agent and object pair then travel together to a different location. After they stop moving, the agent drops off the object. This task shows how an object can be picked up and moved around by an agent in the environment.

g. & j. No-Navigation Approach: Instrumental & Social This task involves three agents and a goal object. In the familiarization phase, an agent consistently approaches one of two stationary agents in proximity, demonstrating a social affiliation (left panels of Figure A.1g&j). At test time, the two previously stationary agents each move in a unique movement. Following that, the previously approaching agent also performs a movement. In No-Navigation Approach: Instrumental (right panel of Figure A.1g, the movement is identical to one of the other two agents which might or might not have been approached before. In addition, the agent interact with a goal object at the end of the move, causing it to change color. The route taken is the only efficient path to the goal object, cueing a goal-directed behavior. In No-Navigation Approach: Social (right panel of Figure A.1j), the previously approaching agent moves in the same way as the approached agent in familiarization.

h. & k. No-Relocation: True & False Belief This task features an occluder that might or might not appear between a goal-directed agent and its goal (Figure A.1h&k). Specifically, it is always in between on the first familiarization trial, obstructing the agent from identifying the side of the room where the goal is located. The agent randomly enters a room and either succeeds in or fails to find the goal object, giving machines an opportunity to learn that the occluder interferes with the perception of the rational goal-directed agent. If the agent finds the goal object in the first familiarization trial, it would consistently go to the same side of the room to find the object in the later familiarization trials, where the occluder randomly shows up between the agent and object or behind the environment. If the agent did not find the goal object in the first familiarization, it would later consistently go to the other side of the room to find the goal object. This task communicates two assumptions: 1) the agent believes a goal object is always at where it was last seen. 2) if the goal object is absent on one side of the room, it must be on the other side. At test time, a different agent moves the goal object around within the same room, either in the absence (False Belief, Figure A.1h) or the presence (True Belief, Figure A.1k) of the first agent. The task then shows the initial agent returning to the same room to retrieve the goal object.

j. & l. Trapped Agent: Helping & Hindering This task (Figure A.1j&l) employs a similar setup as the Helping task and the Hindering task in evaluation (Figure 3). However, the locations of the goal-directed agent and the goal object are swapped. In Helping (Figure A.1j), a social agent move away an obstacle to help the main agent approach its goal object while the other social agent does nothing. In Hindering (Figure A.1l), a social agent blocks the goal-directed agent from approaching its goal while the other social agent watches. At test, the goal-directed agent approaches the agent who helps, or the agent who does nothing when the other agent hinders.

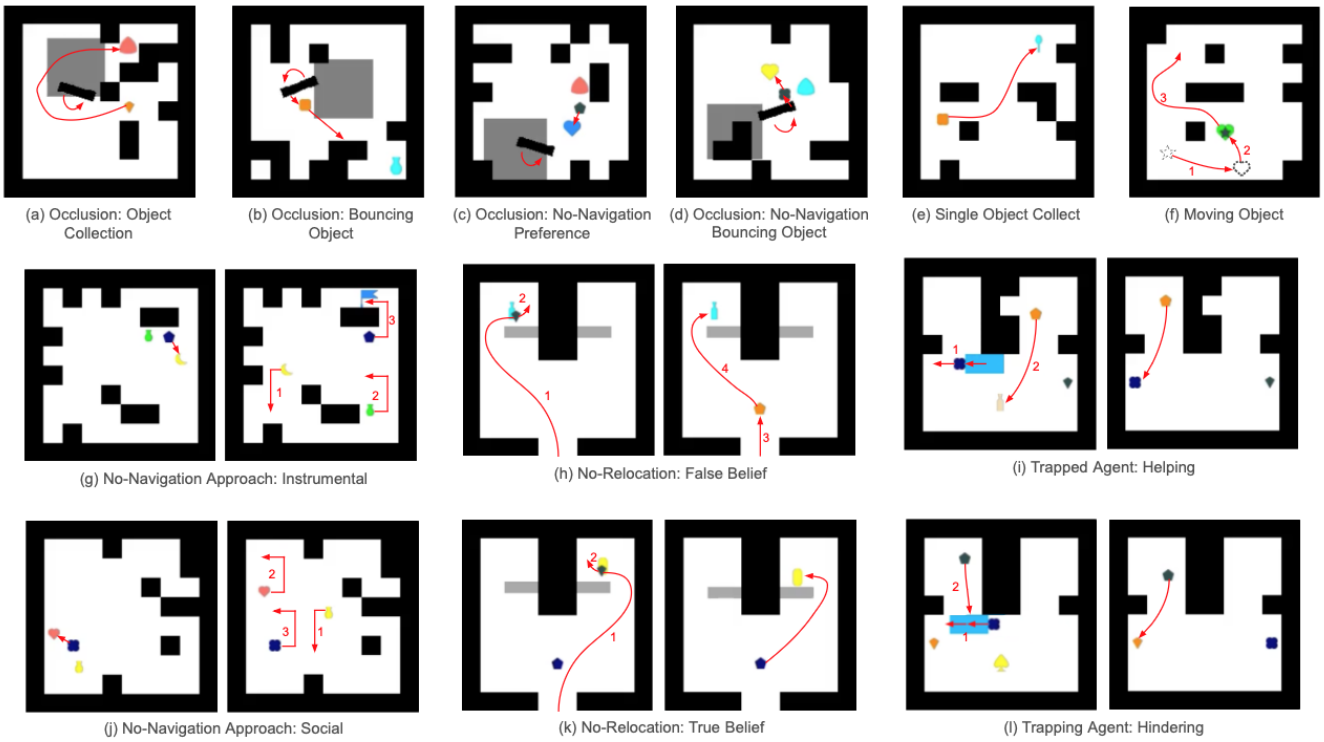

Figure A.1: **Overview of All Background Training Tasks.** AI models undergo training across various tasks to capture environmental dynamics and essential cognitive abilities for solving the benchmark. This figure illustrates the schematics of both familiarization and test trials. For tasks (a)-(f), a single panel depicts each task, where the familiarization trials and a test trial are drawn from the same distribution. For tasks (g)-(l), the left panel represents a familiarization trial, while the right panel illustrates a test trial. Red arrows indicate the direction of movement, and, where applicable, numbers indicate the orders at which the actions take place.

B. Model Hyperparameters

Each video frame is resized to 84 x 84 pixels and then split into 49 patches of 12 x 12 pixels. A three-layer CNN creates a 256-dimensional embedding for each patch, with sinusoidal positional encoding. The row, column, frame, and trial numbers are separately encoded and each takes up 64 dimensions of the positional encoding. The Transformer encoder and decoder each has eight attention heads and five layers. The model is trained on two A100 GPUs for 100 epochs with a batch size of 32. The learning rate and weight decay are both set at 1×10^{-4}