Social World Modeling Benchmark

Anonymous Author(s)
Affiliation
Address
email

Abstract

Although predicting the behavior of others is a fundamental capacity of the human mind, building this intuitive psychology into machines has remained a challenge. To advance this aim, we introduce the Social World Modeling Benchmark - a dataset featuring agents engaged in a diverse repertoire of behaviors, such as goal-directed interactions with objects and multi-agent interactions, all governed by realistic physics. This benchmark introduces a unique cognitively-inspired evaluation pipeline that differs from the conventional approach in trajectory prediction to minimize step-by-step prediction errors. Humans tend to envision future states based on expected events rather than simulating step-by-step. Thus, our benchmark includes evaluations designed to assess whether the simulated trajectories of predictive world models capture the correct sequences of events. To perform well on the evaluations, world models need to leverage predictive cues from the observations in order to accurately simulate the goals of animate agents over long horizons, in addition to the relevant intermediate events that engender those goals, such as picking up an object before delivering it to a goal location. Crucially, the events are not labeled in the training dataset but are labeled post hoc in evaluation as a function of the simulated input states. Although recent developments have incorporated powerful world models into state of the art model-based reinforcement learning agents, we demonstrate that these models perform poorly in our evaluations. In particular, their iterative and autoregressive nature leads to an accumulation of errors that diverges the predictions on long horizons. Therefore, we postulate that to excel in this benchmark, a model should operate like human cognitive processes, whose predictions about the future are scaffolded by abstractions like goals that guide the simulation process towards relevant future events.

1 Introduction

Imagine you are mindlessly watching television with your friend, about to fall asleep, until an event happens: your friend gets up from the couch and walks towards the kitchen. Intuitively, you start inferring reasons for her behavior, even from this sparse observation, but cannot yet fully predict her next movements. Is she hungry and getting food, or thirsty and getting a drink? You next observe another event: she starts chopping an onion, and can now infer that she is cooking a recipe. Given this inference, you can now predict another event: she will chop the other onion on the counter when she finishes the current one, even though you do not know precisely when this will happen or the exact articulations of arm movements she will use. Next, you then decide to help her and ask which other subgoals you can help with. As humans we have sophisticated models of how other humans behave, which helps us understand their actions and guides contingent social interactions like this. Early in infancy, humans begin to develop an intuitive psychology as they learn that other people have mental states like goals and preferences that dictate their actions [Gergely et al., 1995, Woodward, 1998, Repacholi and Gopnik, 1997]. Despite its centrality to the rich social worlds we inhabit, embedding these social capacities into artificial agents has proven to be a colossal endeavor.

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

To propel this mission forward, we propose the Social World Modeling (SWM) Benchmark, which consists of two novel contributions for benchmarking social predictive models. First, SWM includes a diverse dataset of agent trajectories engaging in complex behaviors, including but not limited to goal-directed interactions with objects and dyadic multi-agent interactions. Secondly, our benchmark deviates from the conventional trajectory prediction paradigm and introduces a cognitively-inspired evaluation pipeline specifically tailored to assess the extent to which the simulated trajectories of predictive world models encapsulate the correct sequence of events.

This event-based framework is motivated by a wealth of psychological and neural evidence that humans segment the continuous perceptual stream into events [Zacks and Tversky, 2001, Kurby and Zacks, 2008]. Psychological theories posit that people use event representations in working memory to predict upcoming states, and when prediction errors occur, an event boundary is detected [Reynolds et al., 2007, Zacks et al., 2007]. Time is segmented in this fashion because within an event, future observations become more predictable and therefore can be abstracted, while observations signifying a new event are less predictable. This event-parsing mechanism regularly occurs when observing goal-directed behavior, as another agent's actions will remain relatively predictable until the goal is achieved [Zacks et al., 2007]. Goals are a useful event-based abstraction for segmenting action-sequences, both for oneself and when observing others. When planning a navigational route, the hippocampus prospectively represents goals and subgoals, demonstrating that humans and animals plan by simulating future desired states and working backwards to select actions to reach those states [Johnson and Redish, 2007, Wikenheiser and Redish, 2015, Brown et al., 2016]. Computational models of theory of mind propose that humans infer the intentions of other agents by inverting this planning generative model. Conditional on the actions one has observed so far, one infers what their goal likely is with Bayesian inference [Baker et al., 2009, 2011].

Predictive world models have led to many recent breakthroughs in model-based reinforcement learning [Hafner et al., 2020] and sequence prediction [Vaswani et al., 2017]. The primary class of world models currently used in these fields simulate future states step by step, and autoregressively bootstrap on these predictions to predict long horizons. In contrast, as previously mentioned, humans seem to simulate future states conditioned on anticipated events and goals, which guides predictions towards these events and keeps imagined futures within distribution of plausible outcomes on long timescales. Additionally, the metrics traditionally used in human trajectory prediction literature, average displacement error and final displacement error, are limited. They do not contain information about why a prediction was wrong, and they can even be misleading if a model is minizing ADE/FDE by overfitting.

Social domains in particular are hallmarked by structural relationships between entities that are abstracted from the granularity of iteratively predicting the next step. For example, in competitive situations, anticipating your opponent's next subgoal and subverting it may be more important than accurately predicting their arm articulations in the next timestep. Additionally, the multimodality of goal-directed behavior leads to situations where you may incorrectly infer another person's goal, and therefore have large state prediction errors. But this is a natural consequence of the partial observability inherent in other agents' minds being inaccessible. Morever, even when goals are correctly inferred, humans take stochastic trajectories towards their goals, and there are hundreds of compatible paths to achieve those goals. Thus, we argue that social understanding can be aided by abstracting away from granular step-by-step actions, and representing another agent's behavior with goals and events.

Therefore, a strong test of social understanding consists of assessing whether a model can simulate these events effectively, even if the exact time course may be misaligned from the true data. Consequently, our benchmark incorporates a suite of evaluations with these criteria. For example, after observing an agent pick up an object, models must infer the agent will deliver that object to the goal location. Additionally, models are often tasked with making long horizon predictions, particularly for behaviors with multi-step goals or subgoals. Our evaluations also assess the physical plausibility of the simulated events. Models must distinguish animate agents from inanimate objects, understanding that objects remain static unless acted upon by an agent. Baseline models failed to reliably pass these validations. Models often hallucinated movements of stable objects, and struggled to pick up on predictive cues for goal-directed behavior. Given that capacities to detect animacy and attribute goals to others is learned early in life [Heider and Simmel, 1944, Király et al., 2003, Gergely et al., 1995, Woodward, 1998], it is our belief that incorporating this knowledge into social world models is an

essential foundation for the development of a theory of mind in machines. The SWM Benchmark thus serves as an important milestone to quantify this progress. Our contributions are as follows:

- A social prediction benchmark consisting of a set of behaviors varying in goal-complexity, interactivity, and stochasticity.
- A novel suite of evaluations designed to probe a world model's ability to correctly simulate events and goals.

2 Related Work

2.1 Machine Social Understanding

In recent years, several papers and benchmarks have aimed to evaluate social prediction of goal-directed agents, providing datasets and tasks that challenge machine learning models to reason about the actions and goals of agents in various scenarios. Machine Theory of Mind used meta-learning to acquire a strong prior model for agents' behavior, and subsequently predict their actions in simple 2D grid worlds [Rabinowitz et al., 2018]. The PHASE benchmark constructed a dataset of behaviors expressing social concepts such as helping and hindering, and consists of a behavior recognition task and a multi-entity trajectory prediction task [Netanyahu et al., 2021]. We extend the latter approach with a more diverse set of behaviors and longer prediction horizons. AGENT and the Baby Intuitions Benchmark utilized a developmentally-inspired violation of expectation paradigm to probe key concepts of intuitive psychology, such as action efficiency and goal preferences [Shu et al., 2021, Gandhi et al., 2021]. Whereas BIB used a grid world environment, AGENT constructed an environment in ThreeDWorld (TDW) [Gan et al., 2020] with realistic 3D physics, similarly to our benchmark. Our benchmark combines the strengths of these different approaches with a diverse set of behaviors, while also uniquely adding longer duration multi-step goals and a novel event-based evaluation pipeline that provides a stronger test for common-sense social reasoning than trajectory prediction error in PHASE or the violation of expectation paradigm in AGENT and BIB (Table 1).

More generally, there has been a long history of developing methods that model other agents in machine learning [Albrecht and Stone, 2018]. Several lines of research are action-based, including work in inverse reinforcement learning [Ng et al., 2000, Abbeel and Ng, 2004], multi-agent reinforcement learning [Raileanu et al., 2018, Lowe et al., 2017] and bayesian inverse planning [Baker et al., 2009, 2011]. Here, we examine the scenario where action labels are unavailable, and the trajectories of both objects and agents have to be inferred. Moreover, since objects and agents share the same set of features, concepts like animacy and goal-directness must be learned by the models, as opposed to built in with an inductive bias.

2.2 Human Trajectory Prediction

Several datasets and methods have also been developed concerning predicting human trajectories in domains such as human crowds [Alahi et al., 2016, Rudenko et al., 2020], sports [Kipf et al., 2018, Graber and Schwing, 2020], self-driving vehicles [Salzmann et al., 2020, Lefèvre et al., 2014], and robotics [Lasota et al., 2017]. Some lines of research involve parsing videos of pedestrians or vehicles into object-centric trajectories of the positions and velocities of the entities, which are subsequently used to train predictive models [Murino et al., 2017, Hirakawa et al., 2018]. Several methods of various architectures and objectives have been developed for these tasks [Alahi et al., 2016, Gupta et al., 2018, Tacchetti et al., 2018, Sun et al., 2022, Mangalam et al., 2021].

Our evaluation pipeline introduces a novel evaluation framework centered on event-based prediction, in contrast to the average displacement error and final displacement error metrics used in this literature. We contend that a robust measure of social understanding assesses the capacity of a model to effectively simulate higher-order events, rather than an exact replication of the chronological sequence of states.

3 The Social World Modeling Benchmark

The Social World Modeling (SWM) benchmark consists of 45,000+ trials of agents performing a repertoire of behaviors in the 3D simulation environment ThreeDWorld[Gan et al., 2020](Figure 1A).

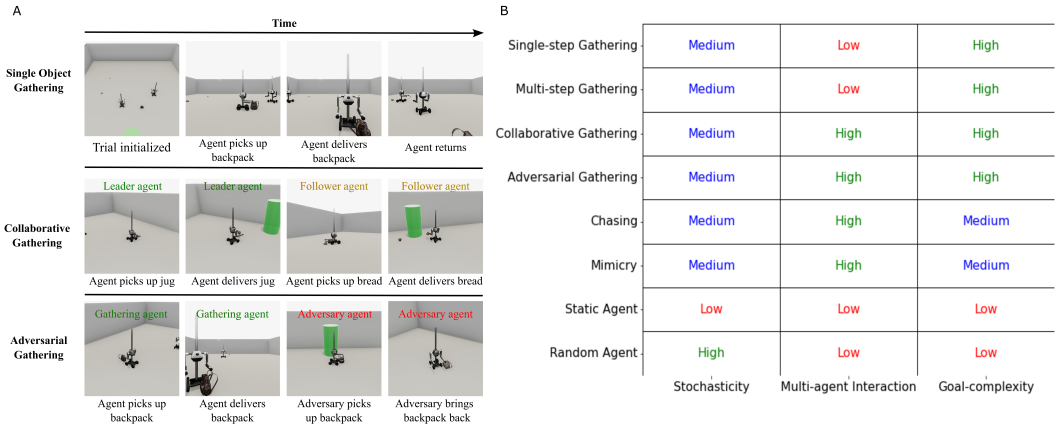

Figure 1: A. Depiction of SWM environment. Every trial starts with two agents, three objects, and a green cylindrical observer agent that agents bring objects to. Three behaviors are depicted from various POVs. Single object gathering, Collaborative gathering, and Adversarial gathering.. B. Taxonomy of the behaviors in the SWM Benchmark. Behaviors vary along important dimensions: stochasticity, multi-agent interaction, and goal-complexity.

	SWM	PHASE	AGENT	BIB
Event-based evaluation	✓	X	X	X
3D Physical Simulation	✓	X	✓	X
Multi-step Goals	✓	X	X	X
Contingent multi-agent behavior	✓	✓	X	X
Long Horizon Trajectory Prediction	✓	X	X	X
Behavioral Diversity	✓	X	X	X
Stochastic Trajectories	✓	X	X	X

Table 1: Comparison of Social Prediction Benchmarks. Social World Modeling (SWM) contains novel evaluations where world models are tested on their ability to simulate correct events, such as goals and subgoals, through the lens of long horizon trajectory prediction. It additionally contains a diverse set of behaviors, including multi-step goals and multi-agent interaction.

The task is for social world models to predict the trajectories of the agents, with a novel evaluation pipeline designed to examine how well the models simulate the correct events. A comprehensive set of behaviors was constructed to include a diverse list of behaviors that differed in important dimensions: stochasticity, multi-agent interaction, and goal-complexity (Figure 1B).

Stochasticity - There is a wide range of determinism in this set of behaviors, with the randomly moving agent being the most stochastic and the gathering agents being more determined. In line with similar recent work [Kim et al., 2020], we stress-test world models on their ability to extract signal from a noisy world. A moderate level of stochasticity exists in a majority of the animate behaviors, for example social world models will initially have uncertainty over which goal a goal-directed agent has, which will be reduced when predictive cues are observed, and uncertainty over the exact trajectory the agent will take to accomplish that goal. This latter form of stochasticity additionally adds diversity to the trajectories akin to the diversity seen in human trajectories, as the exact step-by-step path an animate agent takes while achieving goals is rarely fully predictable, even when the high-level goals and events are. Another important dimension of variation is whether the agents act independently or contingently in multi-agent interaction. **Multi-agent Interaction** - Another important dimension of variation is whether the agents act independently or contingently in multi-agent interaction. This relationship has to be inferred online as a function of the observations, providing another motivation for this set of behaviors. **Goal-complexity** - Lastly, the behaviors vary on the dimension of goal-complexity, with the gathering behaviors representing goal-directed interactions with objects, and the other behaviors operating independently of the objects. Therefore models must learn that objects will only move when acted upon, and sometimes no objects will be acted upon at all. In addition,

gathering agents will take a path to the goal object/location in a path that avoids collisions with other objects. The behaviors include:

- Single-step object gathering - one agent picks up one of the 3 objects, with its goal randomly assigned and delivers the object to a location directly in front of the observer.
- Multi-step object gathering - one agent will pick up all three objects and bring them in front of the observer in a randomly assigned sequence.
- Collaborative object gathering - one leader agent begins multi-step object gathering. Simultaneously a second follower agent will rotate its body to "look at" what the leader agent is doing, and once the leader agent picks up the first object, the follower agent begins to help out and starts gathering another object to bring to the goal location. The leader agent will then continue to gather the third object once it is finished delivering the first object.
- Adversarial object gathering - one agent is the single-step object gathering agent tasked with bringing an object from its initial position to the goal location, and the other agent has the opposite objective. Once the gathering agent brings the object to the goal location and returns to its initial position, the adversarial agent will pick up the object and bring it back to the object's original position. This sequence of events will then repeat continually until the trial ends.
- Chasing - one agent chases the other agent who attempts to navigate away from the chasing agent.
- Randomly moving agent - one agent randomly moves around the room, by selecting random location in a square bounding box around its body (with length 1/4 of the length of the room).
- Mimicry - one agent is the randomly moving agent, and mimic agent repeats the random movements of the randomly moving agent with a small temporal delay.
- Static agent - one agent does not move at all.

In addition to traditional evaluation metrics for trajectory prediction (average displacement error and final displacement error), we designed a battery of evaluations in order to assess whether world models could simulate the correct events in their forward rollouts. Taking inspiration from cognitive science, we believe that social prediction in world models should be guided by inferring the goals and intentions of others based on the available observations [Baker et al., 2009, 2011]. Therefore the focus of our evaluation pipeline is to probe whether world models can recapitulate expected events governed by predictive cues in the data distribution. Successful evaluations hinge on generating physically plausible events with intact abstractions, such as the correct goals being achieved, with tolerance for the wide range of step-by-step atomic actions that animate agents can take to accomplish those goals.

Each trial consists of two agents rendered in the 3D simulation environment, TDW [Gan et al., 2020]. Thus, the behaviors of the agents are randomly selected from a list of compatible pairs, including examples of dyadic multi-agent interactions such as chasing or collaborative gathering, and agents acting independently (ie. gathering + random). The agents are embodied by "Magnebot" avatars: sophisticated robot-like agents that can perform complex tasks such as navigation, object manipulation, and social interaction. The Magnebot has a cylindrical body with four wheels that allow it to move in any direction, and two arms with three joints and a magnet at the end for grasping and manipulating objects. It additionally can slide its body along the y-axis to reach objects at different heights. In addition to the two active Magnebot agents, there is an additional observer agent avatar that agents deliver objects to. This observer agent is always static and located in the same location in the room as a marker for the goal location, and it is represented by a cylindrical embodied avatar in TDW. Each trial additionally includes three objects randomly selected from a set of six (jug, purse, bread, jar, backpack, and vase). Each trial is generated for 1,500 timepoints in simulation and downsampled to 300 timepoints for ease of computation when training world models. On each trial, the positions of the agents and objects are randomly initialized in a square environment.

The input space for world models is object-centric, such that events can be described as a function of the input space. For example, a goal consists of moving a particular object close to the goal location. Input states consist of 7 features for each of the 5 moveable entities, the x,y,z positions (with y being the height) and 4 features for rotation. Since agents and objects share the same set of features, in

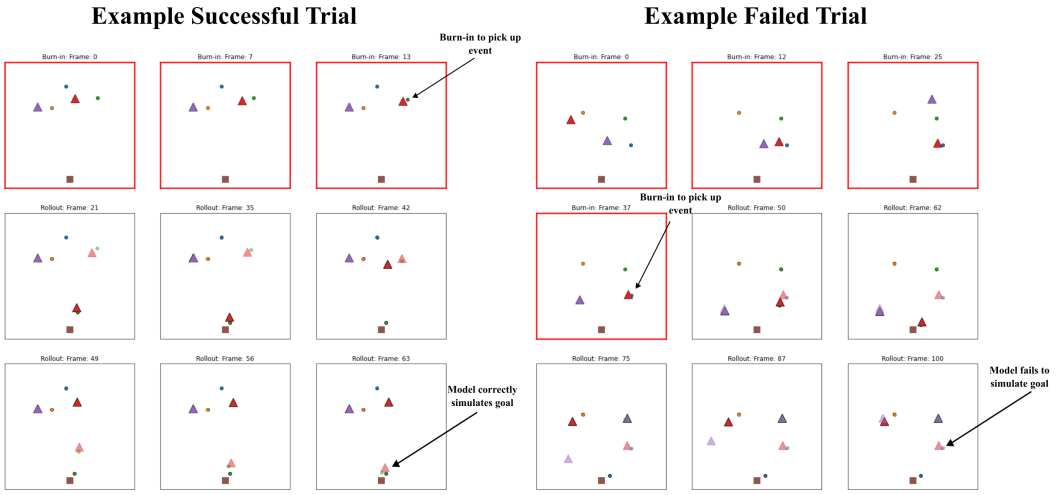

Figure 2: Depiction of example evaluation - Goal Events Evaluation. Two agents are visualized as triangles and three objects as circles in an abstract 2D representation of the 3D physical simulation environment. Models are burned-in observations up until a goal object is picked up, and the remaining trajectory is simulated with a forward rollout. Model predictions are depicted by the more transparent shapes. Left - the red agent picks up the green object and delivers it to a goal location, which is simulated successfully by the model with a delay. Right - the red agent picks up the blue object and delivers it to the goal, and the model fails to simulate this event.

order to predict their future trajectories, world models need to learn to distinguish animate agents from inanimate objects, just as human infants do early in life [Heider and Simmel, 1944, Király et al., 2003].

Baseline world models were trained by sampling 80 step sequences from the training data, feeding in the first 50 steps of the sequence as context/burn-in and predicting the next 30 steps of the sequence. In evaluations, a variable length of observations were fed into the models as context in order for the models to observe enough predictive information about the behavior and upcoming events. Then, models computed forward rollouts for the rest of the trajectory. For Dreamer models, this rollout was done in the imagined latent space, and these latents were subsequently decoded back into the input space. Next, the simulated input data was fed into an event labeler (described in more detail for each evaluation below), and the labeled events were compared to the events from the true data that was fed into the event labeler. Our evaluations examine both whether the simulated events are correct (the correct goal(s) were achieved), and precise (only the goal objects were moved).

3.1 Metrics

Average displacement error and final displacement error. We compute conventional trajectory prediction metrics, both average displacement error and final displacement error. For this metric, models were burned-in 50 steps of context for every trial in the validation data and rolled out for the rest of the trajectory.

Single Goal Events Evaluation For evaluating single goal events, all single-step gathering trials in the validation set were selected. For each trial, models were burned-in observations up until the point where the agent picked up the goal object, and the rest of the trajectory was simulated with a forward rollout (Figure 2). Then we ask whether the object that was picked up was properly delivered to the goal location in the simulated/imagined trajectory. The event labeler would label this as 'True' if that specific object was moved to a location with a Euclidean distance less than 2.0 from the observer/goal location at any point, a liberal threshold that correctly labels every true goal event in the original data. Our evaluation code additionally allows one to toggle on a harder criterion, where the object must be at the goal location at the end of the trial. World models that simulate the delivery of the object to the goal location but deteriorate the state-space while continuing to iterate on its own predictions across many steps perform even worse with this criterion included.

Pick Up Events Evaluation. On the evaluation for pick up events, trials were selected where an agent gathers an object (or multiple objects) to deliver to the goal location. For each trial, models

Figure 3: Results for the Conventional Evaluation Metrics. Bars reflect average across 3 seeds $\pm$ SE.

were fed contextual observations until a point in the trajectory several steps before a pick up event so the models see the agent moving towards a particular object. Models are then tasked with simulating a physical plausible pick up event in their forward rollouts. We noticed that world models often tended to hallucinate irregular movements for picked up objects, such as simulating high fluctuations in the y-coordinate (height) of the object when it was picked up and carried by the gathering agent. Objects would also commonly move without being in close proximity to the carrying agent. Therefore, the event labeler will label a pick up event if the trajectory meets the following criteria: 1. the object is above a threshold 0.6 in the y-coordinate plane 2. the object is moving 3. the object is close to the carrying agent during this sequence 4. the object does not highly fluctuate in the y-coordinate plan. Concretely for criteria 4, the largest displacement in the y-dimension should be in the first 10% or last 10% of the carrying sequence, when the object is picked up or dropped.

Multi Goal Events Evaluation. We additionally adapted the single goal evaluation for multi-step goals, which occurred with the multi-step gathering agent and with the collaborative gathering agents. All three objects in the environment were delivered to the observer in these trials. Models were burned-in to the time point where the 2nd object was picked up. Thus, at this point in the trajectory it is evident that all three objects will be delivered to the goal location. As in the other evaluations, the models then simulated the result of the trajectory, and these simulated input states were assessed by the event labeler. Similarly to single goal event evaluation, we ask whether the 2nd and 3rd objects were properly delivered to the goal location in the forward rollouts.

Move Events Evaluation. The move events evaluation additionally tests for physical plausibility in the simulated trajectories. The baseline models would often hallucinate movement for stable objects, even though objects should only move when acted upon by agents. This evaluation performed rollouts on all trial types, and models were input contextual observations until a point in the trajectory where the behavior should be unambiguous. For single-step gathering, this involves burning in until a few steps past the time point of goal delivery. For multi-step gathering, this involves burning in until the time point where the 2nd object is picked up, identically to the multi goal event evaluation. And for the other behaviors, 50 steps of observations were included as context. Thus, for the non goal-oriented behaviors such as chasing and mimicry, no objects should be moved in the entire trajectory. The event labeler takes in the simulated rollouts and detects which of the three objects were moved on every trial, and compares this boolean array to the ground truth data. An object is labeled moved if the sum of its step by step displacements is greater than a threshold of 4.0. This threshold is tunable to control the difficulty of this evaluation, and was selected to tolerate the inevitable fluctuations in floating points a model will produce step by step, while also rejecting any moderate displacements. The goal is to minimize the false positive rate, while also appropriately detecting when objects should be moved. To pass this evaluation, models need to differentiate animate agents from inanimate objects and learn that objects will not move unless acted upon by an agent.

4 Experiments

4.1 Baseline Models

We implemented and tested the following baselines:

- **Dreamer/RSSM**: [Hafner et al., 2019, 2020, 2023]. It includes both deterministic and stochastic components with the RNN hidden state and variational stochastic latent state respectively. We tested both continuous (DreamerV1) and discrete (DreamverV2) versions of the RSSM.
- **Multistep Predictor**: This model processes the input with and LSTM and MLP to compute predicted states for the next 30 steps, supervised with L^2 loss.
- **Multistep Delta**: Identical in architecture to the Multistep Predictor but computes the predicted difference between the current state and the next state, similarly to recent work [Doyle et al., 2023].
- **Transformer**: Transformer based world model trained autoregressively as in models like IRIS [Micheli et al., 2022].

4.2 Results

Average displacement error and final displacement error are computed for every behavior type and every model (Figure 3). Note that this represents the average prediction error for that behavior across all trials including that behavior, which will include trials paired with different agent types. For example, the random agent can be paired with the mimic, single-step gathering, or multistep gathering agents. The RSSM Discrete had the lowest ADE for all the behaviors and the lowest FDE for single-step gathering, mimicry, collaborative gathering, and adversarial gathering. Multistep Predictor and Multistep Delta had similarly low metrics, with the Multistep Delta model showing the lowest FDE for chasing, random, and multi-step gathering.

The performance on Single Goal Events Evaluation can be visualized in Figure 4. No baseline models performed with better than 50% accuracy, even though goal events are completely deterministic from the context observed by the models (burn-in until the pick up point). The transformer world model performed the best, followed by the RSSM Discrete and Multistep Predictor models. Thus, there exists a substantial room for improvement for world models to infer future states at the right level of abstraction in order to simulate goal events correctly.

Baseline models performed less than 60% accuracy on the Pick Up Event Evaluation. Models would regularly hallucinate irregular movements in the y-dimension, or simulate objects teleporting without being near an agent to carry it. These results suggest the models were not able to learn the physical principles of the environment effectively.

Figure 4 also depicts results for the Multi Goal Event Evaluation. Some models tended to better simulate the delivery of the 2nd object with more context, while the RSSM models did not. As expected, models performed worse on the 3rd object with the longer horizon. The Multistep Predictor model performed the best on this evaluation.

Precision, recall, and F1-Score metrics were calculated in the Move Events Evaluation based on the model's ability to simulate movement of goal objects while minimizing the false positive rate. Multistep Delta models had a high precision score but a low recall score, suggesting that the model often objects static, but often incorrectly kept them static. Since the Multistep Delta model predicts the difference in between the current state and the next state, these results make sense because the model can keep objects static by predicting zeros. RSSM Continuous and Transformer models suffered from the opposite problem, unable to keep any of the objects static in simulation, thereby achieving maximum recall and the minimum possible precision. The RSSM Discrete model achieved the greatest balance and thereby had the highest F1-Score.

5 Discussion

Here, we introduced the Social World Modeling Benchmark, which includes both a comprehensive set behaviors built in a realistic physical simulation environment, and a event-based evaluation pipeline.

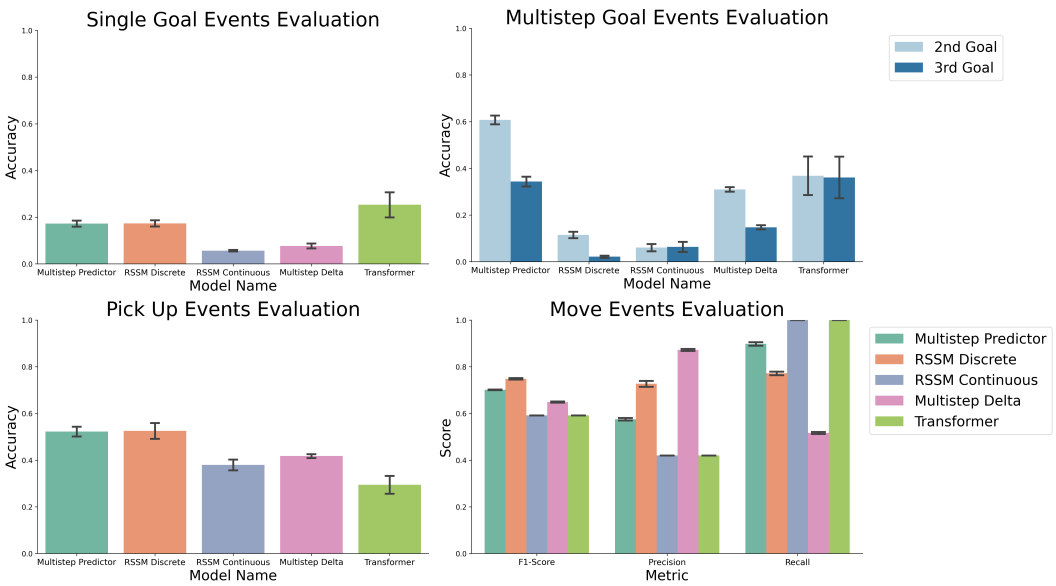

Figure 4: Results for the Event-based Evaluation Metrics. Bars reflect average across 3 seeds $\pm$ SE.

We tested world models from recent developments in model-based reinforcement learning on the pipeline and demonstrate that they fall short on reliably passing our evaluations. Models perform particulary poorly on simulating goal events, while performing better on shorter horizon evaluations such as the Pick Up Event Evaluation. Additionally, on the Move Events Evaluation, a majority of models hallucinated movement for stable objects, while the Multistep Discrete model suffered from the opposite problem, with high precision and low recall.

In future work, as world models are further developed, our benchmark can be extended to include more complex behaviors and evaluations within the same framework. Our code contains motor primitives for picking up and delivering objects, allowing multi-step and collaborative gathering behaviors to be expanded with more objects and subsequently longer horizon predictions. Additionally, goal locations can be varied to add more difficulty to the goal simulation process.

Recent work has demonstrated the power of using world models in deep reinforcement learning agents in environments with both extrinsic rewards [Hafner et al., 2020] and intrinsic rewards [Doyle et al., 2023, Kim et al., 2020]. Another aim of our benchmark is to isolate the world models used in these agents in a 3rd person social setting and characterize their quality. Developments in this arena can then subsequently be plugged back into the RL loop, leading to more intelligent capacities for autonomous agents. Although Dreamer has exhibited impressive performance in various domains such as Atari and Minecraft [Hafner et al., 2020, 2023], our results suggest that its world model may not be good enough to learn common-sense social understanding and be deployed in interactive social settings. Even though behavioral repertoires are consistent throughout our dataset, social settings are hallmarked by dynamic change and contingent interaction that becomes a moving target. As one social task reaches a point of predictability or simplicity, new social tasks are built on top of it to generate fresh new challenges in an emergent curriculum [Leibo et al., 2019]. Recent research has also demonstrated that Dreamer suffers in dynamically shifting environments without modifications to its replay buffer [Kauvar et al., 2023]. Additionally, world model quality has large ramifications when deploying world model based curiosity signals [Pathak et al., 2017, Sekar et al., 2020, Doyle et al., 2023, Kim et al., 2020]. If the model does not learn the right concepts, it will fail to explore the most interesting parts of the state-space or actively seek the best information to minimize uncertainty (ie. looking at someone's gaze to infer their goal). In coordination with their use in curiosity signals, world model prediction errors can be leveraged to segment the world into events as humans do [Reynolds et al., 2007, Zacks et al., 2007], catalyzing more world model learning progress. The limitations of the baseline models we tested also suggest that directly building an event-based or goal-based inductive bias into world models may be useful, and we plan to investigate this in future work.

References

Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 1, 2004.

Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 961–971, 2016.

Stefano V Albrecht and Peter Stone. Autonomous agents modelling other agents: A comprehensive survey and open problems. *Artificial Intelligence*, 258:66–95, 2018.

Chris Baker, Rebecca Saxe, and Joshua Tenenbaum. Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the annual meeting of the cognitive science society*, volume 33, 2011.

Chris L Baker, Rebecca Saxe, and Joshua B Tenenbaum. Action understanding as inverse planning. *Cognition*, 113(3):329–349, 2009.

Thackery I Brown, Valerie A Carr, Karen F LaRocque, Serra E Favila, Alan M Gordon, Ben Bowles, Jeremy N Bailenson, and Anthony D Wagner. Prospective representation of navigational goals in the human hippocampus. *Science*, 352(6291):1323–1326, 2016.

Chris Doyle, Sarah Shader, Michelle Lau, Megumi Sano, Daniel LK Yamins, and Nick Haber. Developmental curiosity and social interaction in virtual agents. *arXiv preprint arXiv:2305.13396*, 2023.

Chuang Gan, Jeremy Schwartz, Seth Alter, Damian Mrowca, Martin Schrimpf, James Traer, Julian De Freitas, Jonas Kubilius, Abhishek Bhandwaldar, Nick Haber, et al. Threedworld: A platform for interactive multi-modal physical simulation. *arXiv preprint arXiv:2007.04954*, 2020.

Kanishk Gandhi, Gala Stojnic, Brenden M Lake, and Moira R Dillon. Baby intuitions benchmark (bib): Discerning the goals, preferences, and actions of others. *Advances in Neural Information Processing Systems*, 34:9963–9976, 2021.

György Gergely, Zoltán Nádasdy, Gergely Csibra, and Szilvia Bíró. Taking the intentional stance at 12 months of age. *Cognition*, 56(2):165–193, 1995.

Colin Graber and Alexander Schwing. Dynamic neural relational inference for forecasting trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1018–1019, 2020.

Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2255–2264, 2018.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019.

Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193*, 2020.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Fritz Heider and Marianne Simmel. An experimental study of apparent behavior. *The American journal of psychology*, 57(2):243–259, 1944.

Tsubasa Hirakawa, Takayoshi Yamashita, Toru Tamaki, and Hironobu Fujiyoshi. Survey on vision-based path prediction. In *Distributed, Ambient and Pervasive Interactions: Technologies and Contexts: 6th International Conference, DAPI 2018, Held as Part of HCI International 2018, Las Vegas, NV, USA, July 15–20, 2018, Proceedings, Part II 6*, pages 48–64. Springer, 2018.

Adam Johnson and A David Redish. Neural ensembles in ca3 transiently encode paths forward of the animal at a decision point. *Journal of Neuroscience*, 27(45):12176–12189, 2007.

Isaac Kauvar, Chris Doyle, Linqi Zhou, and Nick Haber. Curious replay for model-based adaptation. *International Conference on Machine Learning*, 2023.

Kuno Kim, Megumi Sano, Julian De Freitas, Nick Haber, and Daniel Yamins. Active world model learning with progress curiosity. In *International conference on machine learning*, pages 5306–5315. PMLR, 2020.

Thomas Kipf, Ethan Fetaya, Kuan-Chieh Wang, Max Welling, and Richard Zemel. Neural relational inference for interacting systems. In *International conference on machine learning*, pages 2688–2697. PMLR, 2018.

Ildikó Király, Bianca Jovanovic, Wolfgang Prinz, Gisa Aschersleben, and György Gergely. The early origins of goal attribution in infancy. *Consciousness and cognition*, 12(4):752–769, 2003.

Christopher A Kurby and Jeffrey M Zacks. Segmentation in the perception and memory of events. *Trends in cognitive sciences*, 12(2):72–79, 2008.

Przemyslaw A Lasota, Terrence Fong, Julie A Shah, et al. A survey of methods for safe human-robot interaction. *Foundations and Trends® in Robotics*, 5(4):261–349, 2017.

Stéphanie Lefèvre, Dizan Vasquez, and Christian Laugier. A survey on motion prediction and risk assessment for intelligent vehicles. *ROBOMECH journal*, 1(1):1–14, 2014.

Joel Z Leibo, Edward Hughes, Marc Lanctot, and Thore Graepel. Autocurricula and the emergence of innovation from social interaction: A manifesto for multi-agent intelligence research. *arXiv preprint arXiv:1903.00742*, 2019.

Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, 30, 2017.

Karttikeya Mangalam, Yang An, Harshayu Girase, and Jitendra Malik. From goals, waypoints & paths to long term human trajectory forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15233–15242, 2021.

Vincent Micheli, Eloi Alonso, and François Fleuret. Transformers are sample efficient world models. *arXiv preprint arXiv:2209.00588*, 2022.

Vittorio Murino, Marco Cristani, Shishir Shah, and Silvio Savarese. *Group and crowd behavior for computer vision*. Academic Press, 2017.

Aviv Netanyahu, Tianmin Shu, Boris Katz, Andrei Barbu, and Joshua B Tenenbaum. Phase: Physically-grounded abstract social events for machine social perception. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 845–853, 2021.

Andrew Y Ng, Stuart Russell, et al. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, page 2, 2000.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, pages 2778–2787. PMLR, 2017.

Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, SM Ali Eslami, and Matthew Botvinick. Machine theory of mind. In *International conference on machine learning*, pages 4218–4227. PMLR, 2018.

Roberta Raileanu, Emily Denton, Arthur Szlam, and Rob Fergus. Modeling others using oneself in multi-agent reinforcement learning. In *International conference on machine learning*, pages 4257–4266. PMLR, 2018.

Betty M Repacholi and Alison Gopnik. Early reasoning about desires: evidence from 14-and 18-month-olds. *Developmental psychology*, 33(1):12, 1997.

Jeremy R Reynolds, Jeffrey M Zacks, and Todd S Braver. A computational model of event segmentation from perceptual prediction. *Cognitive science*, 31(4):613–643, 2007.

Andrey Rudenko, Luigi Palmieri, Michael Herman, Kris M Kitani, Dariu M Gavrila, and Kai O Arras. Human motion trajectory prediction: A survey. *The International Journal of Robotics Research*, 39(8):895–935, 2020.

Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 683–700. Springer, 2020.

Ramanan Sekar, Oleh Rybkin, Kostas Daniilidis, Pieter Abbeel, Danijar Hafner, and Deepak Pathak. Planning to explore via self-supervised world models. In *International Conference on Machine Learning*, pages 8583–8592. PMLR, 2020.

Tianmin Shu, Abhishek Bhandwaldar, Chuang Gan, Kevin Smith, Shari Liu, Dan Gutfreund, Elizabeth Spelke, Joshua Tenenbaum, and Tomer Ullman. Agent: A benchmark for core psychological reasoning. In *International Conference on Machine Learning*, pages 9614–9625. PMLR, 2021.

Fan-Yun Sun, Isaac Kauvar, Ruohan Zhang, Jiachen Li, Mykel J Kochenderfer, Jiajun Wu, and Nick Haber. Interaction modeling with multiplex attention. *Advances in Neural Information Processing Systems*, 35:20038–20050, 2022.

Andrea Tacchetti, H Francis Song, Pedro AM Mediano, Vinicius Zambaldi, Neil C Rabinowitz, Thore Graepel, Matthew Botvinick, and Peter W Battaglia. Relational forward models for multi-agent learning. *arXiv preprint arXiv:1809.11044*, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Andrew M Wikenheiser and A David Redish. Hippocampal theta sequences reflect current goals. *Nature neuroscience*, 18(2):289–294, 2015.

Amanda L Woodward. Infants selectively encode the goal object of an actor's reach. *Cognition*, 69 (1):1–34, 1998.

Jeffrey M Zacks and Barbara Tversky. Event structure in perception and conception. *Psychological bulletin*, 127(1):3, 2001.

Jeffrey M Zacks, Nicole K Speer, Khena M Swallow, Todd S Braver, and Jeremy R Reynolds. Event perception: a mind-brain perspective. *Psychological bulletin*, 133(2):273, 2007.