Animate Agent World Modeling Benchmark

Logan Cross[1], **Violet Xiang**[2], **Nick Haber**[3], **Daniel L.K. Yamins**[1,2]
[1]Department of Computer Science, [2]Department of Psychology, [3]Graduate School of Education, Stanford University

Abstract

To advance the capacity of intuitive psychology in machines, we introduce the Animate Agent World Modeling Benchmark. This benchmark features agents engaged in a diverse repertoire of behaviors, such as goal-directed interactions with objects and multi-agent interactions, all governed by realistic physics. Humans tend to predict the future based on expected events rather than simulating step-by-step. Thus, our benchmark includes a cognitively-inspired evaluation pipeline designed to assess whether the simulated trajectories of world models capture the correct sequences of events. To perform well, models need to leverage predictive cues from the observations to accurately simulate the goals of animate agents over long horizons. We demonstrate that current state-of-the-art models perform poorly in our evaluations. A hierarchical oracle model sets an upper bound for performance, suggesting that to excel, a model should scaffold their predictions with abstractions like goals that guide the simulation process towards relevant future events.

Keywords: social cognition; cogsci-ai; world modeling; ai benchmark; intuitive psychology; abstraction; goals

Introduction

In a simple black and white film by Heider and Simmel, geometric shapes become intentional beings to observers, highlighting our tendency to attribute animacy to autonomous forms (Heider & Simmel, 1944). From infancy, we intuit that others have mental states influencing their actions (Gergely et al., 1995; Woodward, 1998). While central to our social lives, instilling this intuitive psychology into artificial agents has been challenging. Advancing this effort, we introduce the Animate Agent World Modeling (AAWM) Benchmark, which consists of a dataset of agent trajectories engaging in complex behaviors, including but not limited to goal-directed interactions with objects and dyadic multi-agent interactions. Our benchmark differs 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 evidence that humans segment the continuous perceptual stream into events (Zacks & Tversky, 2001; Kurby & Zacks, 2008). In addition, models of theory of mind propose that humans infer the intentions of other agents by inferring what goal is likely given the actions observed so far with Bayesian inference (C. Baker et al., 2011). In contrast, state of the art world models in machine learning simulate future states step by step, whereas humans simulate future states conditioned on anticipated events and goals. Similarly, traditional human trajectory prediction metrics, such as average displacement error (ADE) (Rudenko et al., 2020), do not inform why a prediction was wrong as models could suffer from distinct failure modes that lead to the same ADE performance. In social contexts in particular, understanding the structural relationships between entities and anticipating an agent's next subgoal is often more critical than precise predictions of immediate body articulations.

Therefore, a strong test of social understanding consists of assessing whether a model can simulate these goals and 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 to complement conventional trajectory prediction metrics. Baseline models failed to reliably pass these validations and showed various failure modes that we analyze. For example, 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 & Simmel, 1944; Király et al., 2003; Gergely et al., 1995; Woodward, 1998), it is our belief that incorporating this common sense knowledge into social world models is an essential foundation.

Our benchmark complements and extends a exciting line of cognitively-inspired social prediction benchmarks. 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 trajectory prediction task (Netanyahu et al., 2021). We extend the latter approach with a distinct set of behaviors, longer prediction horizons up to 50s, and a dataset two orders of magnitude larger. AGENT and the Baby Intuitions Benchmark (BIB) 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). While BIB used a grid world environment, AGENT constructed an environment in ThreeDWorld (TDW) (Gan et al., 2020) with realistic 3D physics, similar to our benchmark. The AAWM Benchmark combines the strengths of these approaches with a diverse set of behaviors and adds an event-based evaluation pipeline and analysis tools that provide a distinct test for common-sense social

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

reasoning, complementing the trajectory prediction error in PHASE and the violation of expectation paradigm in AGENT and BIB.

Animate Agent World Modeling Benchmark

The Animate Agent World Modeling Benchmark (AAWM) Benchmark [1] consists of 45,000+ trials of agents performing behaviors in the 3D simulation environment ThreeDWorld(Gan et al., 2020)(Figure 1A). The task is for world models to predict the trajectories of the agents in object-centric coordinate space, with an evaluation pipeline designed to examine how well the models simulate the correct events (Figure 2). The behaviors include:

- Single-step gathering - one agent picks up one of the 3 objects and delivers it to an observer agent.
- Multi-step gathering - one agent will move 3 objects.
- Collaborative gathering - two agents (a leader and a follower) collaborate to gather all 3 objects.
- Adversarial gathering - one agent does single-step gathering, while the other returns object to its starting position. Sequence repeats until the trial ends.
- Chasing - one agent chases the other agent.
- Random agent - one agent randomly moves.
- Mimicry - one agent is random agent, and mimic agent repeats their movements with a small temporal delay.
- Static agent - one agent does not move at all.

Each trial consists of two agents rendered in the 3D simulation environment. 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 tasks such as navigation and object manipulation with inverse kinematics. In addition to the two active Magnebot agents, there is an additional observer agent avatar that agents deliver objects to. This green cylindrical observer agent is always static and located in the same location in the room as a marker for the goal location. 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 (50s at 30 FPS) in simulation and downsampled to 300 timepoints for ease of computation when training world models. Note that on many evaluation trials, models must compute forward rollouts of over 250 timesteps, a relatively long horizon prediction.

The input space for world models is object-centric, such that events can be described as a function of the input space and annotated by the event labeler (Figure 1B, Figure 2A). For example, a goal consists of moving a particular object close to the goal location. Input states at timepoint t are represented as a vector $\mathbf{x}_t \in \mathbb{R}^{35}$, consisting 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 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 & Simmel, 1944; Király et al., 2003).

Baseline world models perform a trajectory prediction task, where they take in a subset of a trajectory sequence as contextual observations and generate a forward rollout to predict multiple future states of that trajectory. In evaluations, models receive a varied length of observations as context in order to observe enough predictive information about the behavior and upcoming events. Then, forward rollouts for the rest of the trajectory are generated, 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)[2].

Metrics

We've designed evaluations to assess if world models simulate accurate events. We believe that social prediction should be guided by inferring the goals and intentions of others from sparse observations, just as humans do (C. L. Baker et al., 2009; C. Baker et al., 2011). Therefore we probe whether world models can recapitulate expected events governed by predictive cues in the data distribution. Event definitions are based on two criteria: 1. Accurate labeling of the ground truth data. We have ground truth labels about events from the data generation process, and subsequently create a rule-based event labeler that can label the data with concise symbolic rules based on only seeing the the trajectory of object-centric input. 2. Exclusion of trajectories that exhibit certain failure modes, such as objects violating the laws of gravity or moving without being acted upon. This event labeling procedure affords the ability to compare the simulated data to the ground truth. Our code also includes traditional trajectory prediction metrics, average displacement error and final displacement error.

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 timepoint t_{pickup} 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 threshold that correctly labels every true goal event in the ground truth data without false positives. Our evaluation code additionally offers users the ability to toggle the difficulty of this evaluation with a positive or negative offset parameter that includes more or less contextual frames for the model to see. We evaluate

[1] Videos of these behaviors can be found here

[2] Videos of example forward rollouts can be found [here]

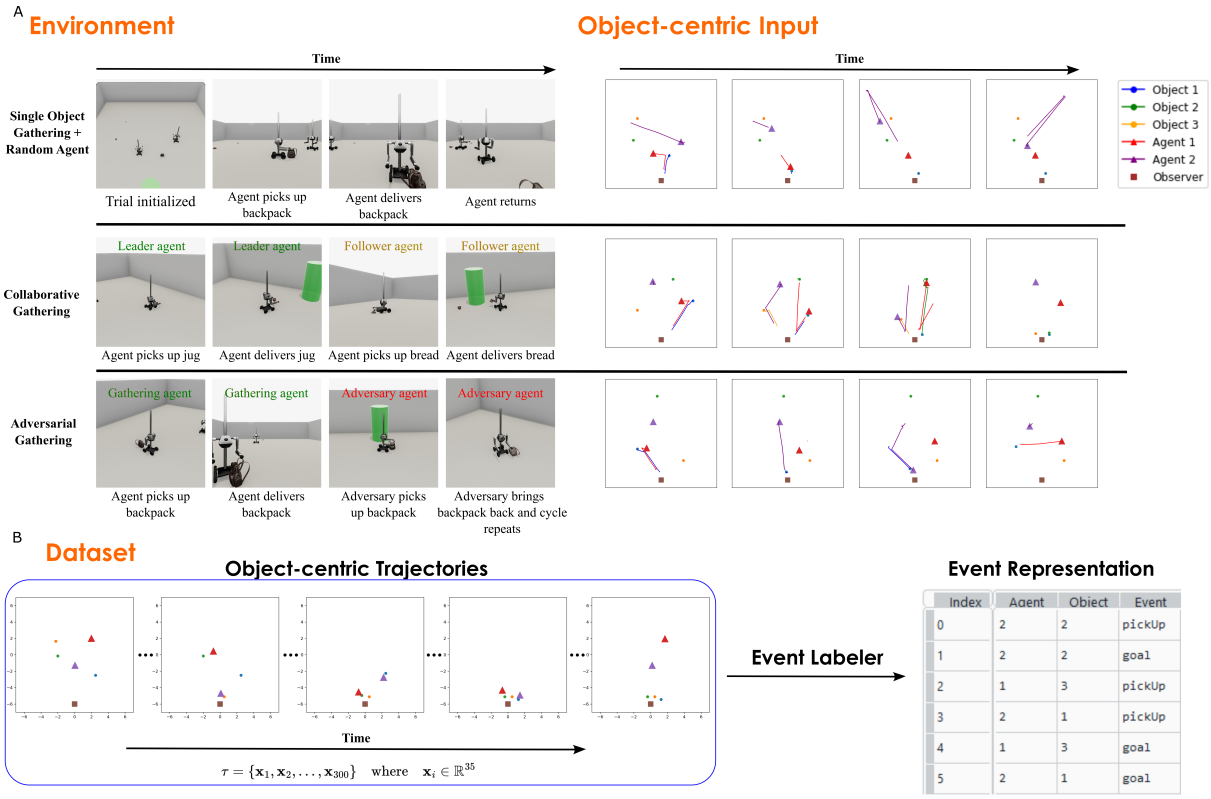

Figure 1: A. Depiction of AAWM environment. Right: Two agents are visualized as triangles, three objects as circles, and observer as a square in an abstract 2D representation of the input to world models. Solid traces reflect the future positions of the entities over the course of the next 25 steps. B. Object-centric dataset is fed to an event labeler to annotate events.

an easier version where models see 5 steps past the pick up point in Figure 3B.

Pick Up Events Evaluation: For evaluating pick up events, trials were selected where an agent gathers an object(s) to deliver to the goal location. For each trial, models were fed contextual observations until 10 steps before a true pick up event, allowing the models to perceive the agent moving towards a particular object. The models were then tasked with simulating a physically plausible pick-up event in their forward rollouts. We observed that world models frequently hallucinated irregular movements for picked-up objects, such as simulating highly fluctuating movements or moving an object without it being in close proximity to an agent.

We designed four criteria that correctly labels ground truth pick up events from the object-centric data, while punishing models that suffered from these pitfalls: 1. Height Threshold: The object must be within a realistic height range, ensuring it's neither too low nor too high; 2. Object movement: The object must be moving; 3. Close proximity to an agent. 4. Stability in height: The object's height should remain relatively stable during carrying.

Multi Goal Events Evaluation. We 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. 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 tests for physical plausibility in the simulated trajectories. The models would often hallucinate movement for stable objects, even though objects should only move when acted upon by agents. This evaluation performs rollouts on all trial types, and models were input contextual observations until a point in the trajectory where the behavior is unambiguous. 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 these event labels to the ground truth data. An object is labeled moved if the sum of its step by step displacements is greater than a tunable threshold controlling the difficulty of the evaluation. The default threshold of 4.0 was selected to tolerate the inevitable fluctuations in floating points a model

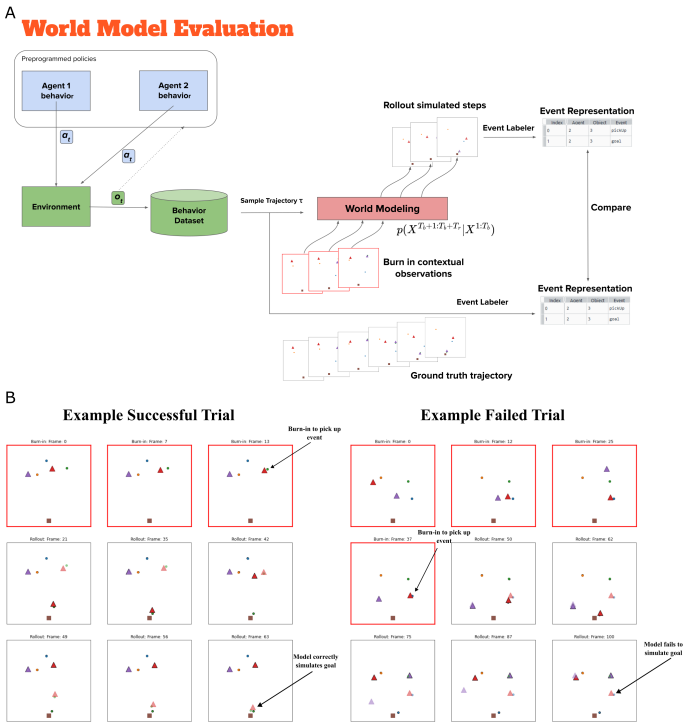

Figure 2: A. Illustration of event-based evaluation pipeline. B. Depiction of example evaluations - Goal Events Evaluation. Models are burned-in observations up until a goal object is picked up, and the remaining steps are simulated with a forward rollout. Model predictions are depicted by the more transparent 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.

will produce step by step, while also rejecting any moderate displacements that would be visible to a human. 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.

Experiments

Baseline Models

- **Dreamer/RSSM**: (Hafner et al., 2019, 2020, 2023). RSSM has deterministic and stochastic components with the RNN hidden state and variational stochastic latent state respectively. We tested both continuous (DreamerV1) and discrete (DreamverV2) versions.

- **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, as in (Doyle et al., 2023).

- **Transformer**: World model trained autoregressively as in IRIS (Micheli et al., 2022).

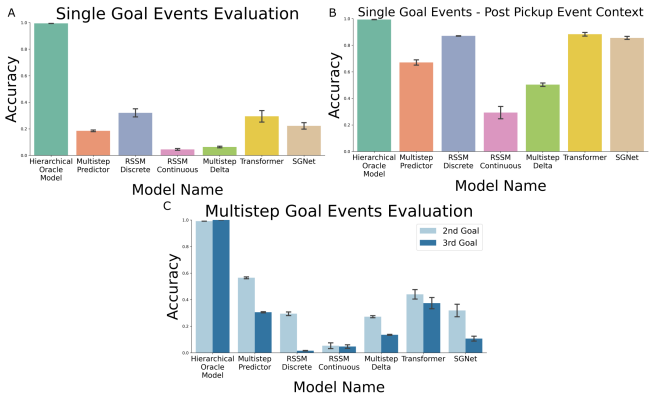

Figure 3: A. Results for the Single Goal Events Evaluation. B. Easier modified version of Single Goal Events Evaluation. Models were fed in an additional 10 steps of context for the single-step gathering trials, meaning they observe past the pickup point and see the agent carrying the object and moving towards the goal location.
C. Results for the Multistep Goal Events Evaluation. Bars reflect average across 3 seeds $\pm$ SE.

- **SGNet**: This model's objective is to predict future trajectories while simultaneously estimating step-wise goals at various time scales given the contextual observations (Wang et al., 2022). We set the step-wise goal target to the state 10 steps ahead.

- **Hierarchical Oracle Model**: Model with hierarchical structure to demonstrate the value of using future event-based representations and multiple timescales. This model is the Multistep Predictor model that takes in the ground truth future end-state of the trajectory as input, in addition to the normal sequences of states. Since this model uses privileged information it cannot be directly compared to the other models.

Results

Baseline Models Fail to Leverage Predictive Information to Simulate Goal Events. The performance on Single Goal Events Evaluation can be visualized in Figure 3A. No baseline models performed with better than 50% accuracy on the evaluation, even though goal events are completely deterministic from the context observed by the models (burn-in until the pick up point). The RSSM Discrete model performed the best on this evaluation, followed by the Transformer and SGNet. The Multistep Delta and RSSM Continuous models struggled to correctly simulate events less than 10% of the validation trials. With privileged information about the end state, the Hierarchical Oracle model sets an upper bound for performance on this benchmark if world models could correctly infer the goal of the agent. This result highlights a substantial gap in performance between this model and the other baseline models while concurrently pinpointing event prediction as the key bottleneck and challenge for world models. With the oracle model we assume a perfect end-state predictor in order to separate the event prediction/end-state prediction problem from the computational problem of guiding short, moderate, and long horizon predictions with an anticipated future state. The notable performance of SGNet in the goal events evaluations also indicates the value of predicting a future goal state and conditioning forward rollouts on this prediction.

To further identify exactly what the models struggle with on the Single Goal Events Evaluation, we modified the evaluation such that they perceived 5 more frames of context/burn-in, meaning that models observe past the pickup point and see the agent start to bring the object towards to goal location. The time between the pickup event and the goal event was on average 16.3 ± 9.7 std steps, thus 5 additional steps gives models observations of the agent and object traversing towards the goal. Performance dramatically improved for all baseline models, demonstrating that the models struggle at the pick up point (Figure 3B). Since the vector of velocity in the trajectory can dramatically change directions at this change point, models often simulated movement in the same direction as previously observed or predicted that the agent would continue to stay still after picking the object up. Therefore, these models struggled to make the correct inference that a pick up event precedes an agent moving that object to the goal, even though this inference would be common sense to a human. This result additionally illustrates that the world models are able to extrapolate movement along the same vector of motion and correctly simulate the goal when the goal event is imminent.

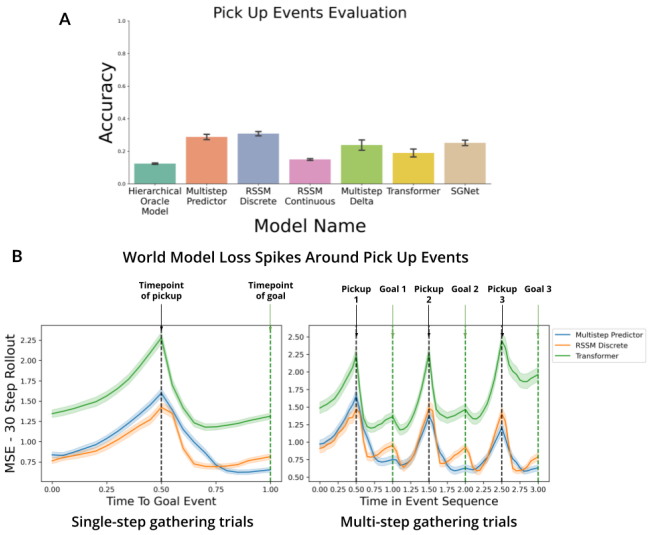

Figure 4: A. Results for the Pick Up Events Evaluation. Bars reflect average across 3 seeds $\pm$ SE. B. Models were tested with 30 step forward rollouts on various timepoints in single-step and multi-step gathering trials to evaluate how world model loss evolves over time around these events. Three representative models are shown for ease of visualization.

Figure 3C depicts results for the Multi Goal Event Evaluation. Some models such as the Multistep Predictor and Transformer tended to better simulate the delivery of the 2nd object with more context, while the RSSM models did not. Thus, feeding in more observational steps had differential effects on each model with some performing better and others worse. As expected, most models performed significantly worse on the 3rd object with the longer horizon. The Multistep Predictor model performed the best on the 2nd goal in this evaluation, and the Transformer performed the best on the 3rd goal.

Event segments are structured around high world model loss during pick up events. Baseline models performed less than 40% accuracy on the Pick Up Event Evaluation (Figure 4A). 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. Even the Hierarchical Oracle model struggles on this evaluation, as it has no additional information to guide the simulation of pick up events. The RSSM Discrete and Multistep Predictor performed the best on the evaluation.

In order to evaluate how world model loss varied by across

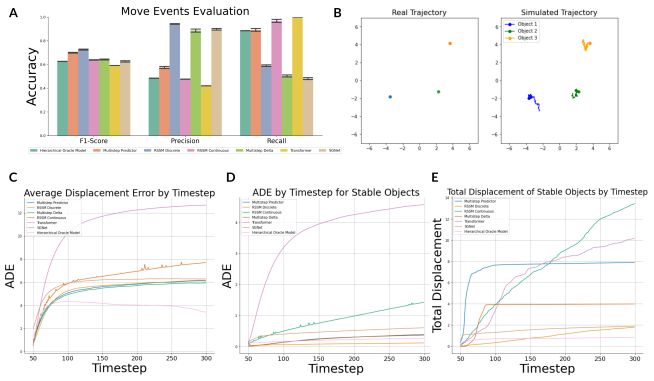

Figure 5: A. Move Events Results. Bars reflect average across 3 seeds ± SE. B. Example of hallucinated movement - mimicry trial for RSSM Continuous. Circles are initial positions of the objects, traces reflect displacement throughout the trial. C. ADE averaged by rollout timestep. D. ADE by timestep only for stable objects. E. Total displacement of stable objects by timestep.

a trial and by event, we computed forward rollouts on single-step gathering and multi-step gathering validation trials at various timepoints in a trial, with timepoints normalized by their distance to pick up and goal events (Figure 4B, see Supplementary Material for more information). For visualization clarity, we depict the MSE curves of three representative models, and all baseline models show similar patterns.

These plots demonstrate that prediction error dynamics of models are structured around these events. Both the peaks/troughs of these curves indicate useful information that can be summarized in three important takeaways from these plots. 1. There are high peaks in the world model loss of 30 step rollouts around pick up events, likely due to models struggling during and shortly after the pick up event. Prediction error spikes are caused both by the inherent increase in complexity of producing physically plausible pick up events and the inability to leverage predictive information to correctly simulate the goal after the pick up event. 2. There are troughs in the curves before the goal delivery events. This illustrates that predicting the rest of the event segment becomes trivial if a goal is correctly identified (Figure 3B). 3. There is a rise in model prediction error near and after the goal events. This further demonstrates that models struggle at change points and directional changes as previously discussed. The patterns here are analogous to the research in human psychology that suggests humans segment time into events based on the predictable nature of the states within the event segment and prediction errors signaling event boundaries. (Reynolds et al., 2007; Zacks et al., 2007).

Models hallucinate movement and suffer from error accumulation. In the move events evaluation, models are test on moving only the correct objects in simulated trajectories. Given unambiguous context, models should be able to simulate rollouts where the appropriate objects will be picked up by agents and the rest will remain stationary. The event labeler annotates the rollouts, comparing object movements to ground truth. Precision, recall, and f1-score metrics are plotted on Figure 5A. The RSSM Discrete has the highest f1-score, striking the best balance between minimizing false positives and false negatives or all the models. The Multistep Delta model also shows high precision, likely due to its objective to predict the difference between the current state and the next state. This stability bias results in lower recall scores, a pattern also seen in RSSM Discrete and SGNet. Transformer and RSSM Continuous models, conversely, display high recall but suffer from poor precision and rampant false positives by predicting constant motion as demonstrated by an example of hallucinated movement in Figure 5B.

We completed additional analyses to identify how simulated rollouts unfold over time. Figure 5C plots the average displacement error (ADE) per timestep. Baseline models show error accumulation, particularly in the first 50 steps of the rollout. Interestingly, the oracle model has minimal error accumulation and shows a decrease in ADE post-timestep 100, hinting that hierarchical structure can tether rollouts to anticipated states and reduce errors. Figure 5D illustrates ADE by timestep while isolating the features that represent the positions of static objects in a trial. RSSM Discrete and Multistep models do well in keeping them stable, while stable objects tend to drift for the Transformer and RSSM Continuous. Keeping an object near its initial position doesn't guarantee stability in move event evaluations; an object can still accumulate large total displacements by jittering around its initial position. Total displacement by time is shown in Figure 5E. The Multistep Predictor exhibits such jittering, while Transformer and RSSM Continuous tend to drift, increasing both ADE and total displacement. Multistep Delta is less prone to jitter due to its state difference output. SGNet and RSSM Discrete perform impressively, near oracle levels, indicating that a hierarchical inductive bias or temporal prior can reduce movement issues.

Discussion

Here, we introduced the Animate Agent World Modeling Benchmark, which includes a comprehensive set behaviors built in a realistic physical simulation environment, and a event-based evaluation pipeline. We tested commonly used world models on the pipeline and demonstrate that they fall short on reliably passing our evaluations. 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 fall short in acquiring common-sense social understanding. Additionally, the exceptional performance of the Hierarchical Oracle Model on the goal events evaluations suggest that directly building an event-based or goal-based inductive bias into world models should be useful.

References

Baker, C., Saxe, R., & Tenenbaum, J. (2011). Bayesian theory of mind: Modeling joint belief-desire attribution. In

Proceedings of the annual meeting of the cognitive science society (Vol. 33).

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349.

Doyle, C., Shader, S., Lau, M., Sano, M., Yamins, D. L., & Haber, N. (2023). Developmental curiosity and social interaction in virtual agents. *arXiv preprint arXiv:2305.13396*.

Gan, C., Schwartz, J., Alter, S., Mrowca, D., Schrimpf, M., Traer, J., ... others (2020). Threedworld: A platform for interactive multi-modal physical simulation. *arXiv preprint arXiv:2007.04954*.

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.

Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition*, *56*(2), 165–193.

Hafner, D., Lillicrap, T., Ba, J., & Norouzi, M. (2019). Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*.

Hafner, D., Lillicrap, T., Norouzi, M., & Ba, J. (2020). Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193*.

Hafner, D., Pasukonis, J., Ba, J., & Lillicrap, T. (2023). Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*.

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

Király, I., Jovanovic, B., Prinz, W., Aschersleben, G., & Gergely, G. (2003). The early origins of goal attribution in infancy. *Consciousness and cognition*, *12*(4), 752–769.

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

Micheli, V., Alonso, E., & Fleuret, F. (2022). Transformers are sample efficient world models. *arXiv preprint arXiv:2209.00588*.

Netanyahu, A., Shu, T., Katz, B., Barbu, A., & Tenenbaum, J. B. (2021). Phase: Physically-grounded abstract social events for machine social perception. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 35, pp. 845–853).

Reynolds, J. R., Zacks, J. M., & Braver, T. S. (2007). A computational model of event segmentation from perceptual prediction. *Cognitive science*, *31*(4), 613–643.

Rudenko, A., Palmieri, L., Herman, M., Kitani, K. M., Gavrila, D. M., & Arras, K. O. (2020). Human motion trajectory prediction: A survey. *The International Journal of Robotics Research*, *39*(8), 895–935.

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).

Wang, C., Wang, Y., Xu, M., & Crandall, D. J. (2022). Stepwise goal-driven networks for trajectory prediction. *IEEE Robotics and Automation Letters*, *7*(2), 2716–2723.

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

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., & Reynolds, J. R. (2007). Event perception: a mind-brain perspective. *Psychological bulletin*, *133*(2), 273.

Zacks, J. M., & Tversky, B. (2001). Event structure in perception and conception. *Psychological bulletin*, *127*(1), 3.