Modeling Generalizable Physical Reasoning as Language-Guided Synthesis of Probabilistic Simulation Programs

Vicente Vivanco[†], **Katherine M. Collins, Joshua B. Tenenbaum, Kevin A. Smith**
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology, Cambridge, MA, USA
[†] vvc@mit.edu

Abstract

People use physical knowledge in remarkably flexible ways, yet most prior work models one aspect of physical reasoning at a time. Here we study the flexibility of people's physical reasoning, building on theories that language guides the construction of ad-hoc mental models grounded in capabilities for forming and manipulating representations of the world. We instantiate this theory in the Physical Reasoning via Interpretable Synthesis of Models (PRISM) framework, which uses structured reasoning with language models to generate task-specific programs in response to a question, relying on primitive functions for perceiving, editing, and simulating physical scenes. We compare PRISM against people's judgments on four distinct physics scenarios inspired by prior research, finding that it explains people's behavior as well as custom-written models from prior research while outperforming vision-language model baselines. PRISM thus provides a cognitively plausible framework for understanding how we assemble physical concepts to flexibly reason about the world.

Keywords: intuitive physics; physical reasoning; model synthesis architecture; vision-language models

Introduction

People are remarkably flexible in how they reason about the physical world and the kinds of questions they can ask and answer, from trying to judge "can I put one more book on this stack, or will that cause it to topple", or estimate "is it safe to cross the road", or engage in counterfactual reasoning "would Michael Jordan have made that half-court shot had he not slipped?" How is it that human physical reasoning is so flexible with such broad coverage?

Much recent research suggests that our physics knowledge relies on probabilistic simulation over structured, object-centric world models (Smith et al., 2024). This framework, with probabilistic simulation at its core, has been used to explain a variety of human physical judgments – from predictions that involve running the simulator and making judgments based on the distribution of possible future world states (Battaglia, Hamrick, & Tenenbaum, 2013; Smith & Vul, 2013; Bates, Yildirim, Tenenbaum, & Battaglia, 2019; Wang et al., 2024), to inferences that condition on observed states of the world to determine latent physical variables within the simulator (Hamrick, Battaglia, Griffiths, & Tenenbaum, 2016; Sanborn, Mansinghka, & Griffiths, 2013; Bi, Shah, Wong, Scholl, & Yildirim, 2025), to causal judgments that use simulation to generate counterfactual worlds in order to assess responsibility (Gerstenberg, Peterson, Goodman, Lagnado, & Tenenbaum, 2017; Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2021; Zhou, Smith, Tenenbaum, & Gerstenberg, 2023), to action planning that selects actions to maximize the utility expected by simulation (Allen, Smith, & Tenenbaum, 2020; Yildirim, Gerstenberg, Saeed, Toussaint, & Tenenbaum, 2017). Yet despite the common framework, all of the cognitive models used in these studies have been separately hand-designed, using simulators and reasoning frameworks customized for each experiment. This variety of models has led some to argue that a fundamental weakness of simulation-based accounts of physical reasoning is that there is no overarching theory for how we produce and use simulatable mental models in the first place (Davis & Marcus, 2015).

In parallel, there has been growing interest in how language can act as a guide to structure our thoughts to solve the variety of problems we face on a daily basis. Recent theories suggest that language processing systems in the brain have only a *shallow* understanding of semantics, and that deep understanding comes from the language system's ability to call out to other cognitive systems to assemble mental models about the scenario under discussion (Casto, Ivanova, Fedorenko, & Kanwisher, 2025; Wong et al., 2023). Brooke-Wilson (2023) and Wong et al. (2025) formalized this idea in a *Model Synthesis Architecture* (MSA) that explains how people provide answers to questions that require reasoning about open-world problems. The MSA hypothesis suggests that people combine information about the specific problem at hand with global background knowledge to construct small-scale, ad-hoc models that contain the relevant concepts, information, and variables to then use those ad-hoc models for thinking and reasoning about said questions in a resource-efficient manner. They instantiate an MSA that synthesizes probabilistic programs via a series of language model (LM)-guided generation, scoring, and filtering steps and then conducts Bayesian inference in the resulting program to answer causal reasoning questions. In this framework, language is used to construct the task-specific model, but actual reasoning is conducted via other cognitive machinery.

In the context of physical reasoning, this separation of language-guided model synthesis and model-based reasoning aligns with the theory that deep physical understanding does not exist within the language system, but instead is a result of language being used to construct task-specific reasoning models grounded in visual and physical concepts. Indeed,

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

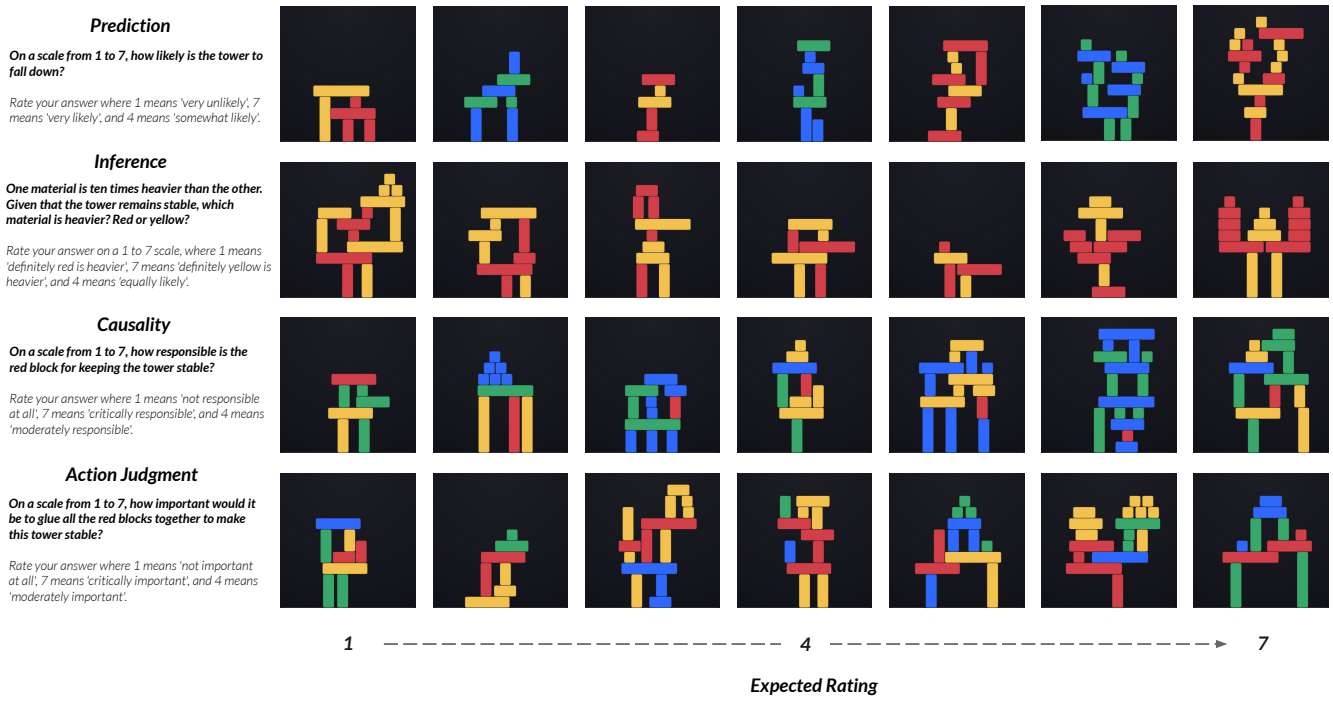

Figure 1: The four tasks tested in our experiment, spanning different types of physical judgments, with the precise question and rating scales given to people and models. Representative stimuli are selected to span the response scale as judged by the hand-built Gold Standard models.

when vision-language models answer questions about physical scenes directly, they do not behave in a human-like manner, even with fine-tuning (Buschoff, Voudouris, et al., 2025; Buschoff, Akata, Bethge, & Schulz, 2025). Instead, we argue that the crucial function of language is providing guidance on how to appropriately decide *what to simulate*, and structure the outcome of those simulations to *respond appropriately*.

Our overarching goal is to formalize the reasoning process by which people can be asked open-ended questions about arbitrary physical scenes – directly perceived or communicated via language – and produce reasonable answers. This is a similar goal to Zhang, Wong, Grand, and Tenenbaum (2023), but focuses on a different facet of the problem: they focused on how ambiguous linguistic input gives rise to a distribution over possible scenes, and demonstrated that, in a single task, reasoning involves grounding these scenes out into a probabilistic program that makes inferences based on the outcome of noisy simulations within that scene. Here we focus on a different subset of that overarching goal: how people construct bespoke programs grounded in probabilistic simulation to answer questions flexibly about observed scenes across a range of physical reasoning tasks.

To that end we introduce the *Physical Reasoning via Interpretable Synthesis of Models* (PRISM) framework as a theory for how people customize physical reasoning in response to different questions. PRISM is based on MSA, using a structured generation process to propose runnable programs that can interpret, manipulate, and simulate physical scenes. Crucial to this framework is that model synthesis *grounds out* in the ability to understand and manipulate scenes – functions for parsing a scene into objects, editing those objects, and running noisy simulations are provided as a given. This follows from the theory that components of physical reasoning are core knowledge (Spelke & Kinzler, 2007) – domain specific, early-emerging, and often cognitively impenetrable.

We test PRISM on a set of four physical reasoning tasks inspired by previous cognitive science experiments (Fig. 1), including prediction (Battaglia et al., 2013), inference (Hamrick et al., 2016), causal reasoning (Zhou et al., 2023), and judgments about actions (Hamrick et al., 2018). We find that PRISM can explain human judgments across all four tasks without task-specific tuning, matching the explanatory power of hand-designed probabilistic simulation programs based on models previously proposed as explanations of human physical reasoning. Additionally, it outperforms both variants of PRISM that do not include physical reasoning, and direct responses from a set of internet-scale vision-language models. Together, this provides a first step towards understanding how people can solve arbitrary physical reasoning problems, describing how language and physics can work in concert to help us understand, describe and act on the world around us.

Physical Reasoning via Interpretable Synthesis of Models

Each step of PRISM is goal-directed to answer the question at hand. Based on a natural-language question, it synthesizes a model that specifies which aspects of the scene are

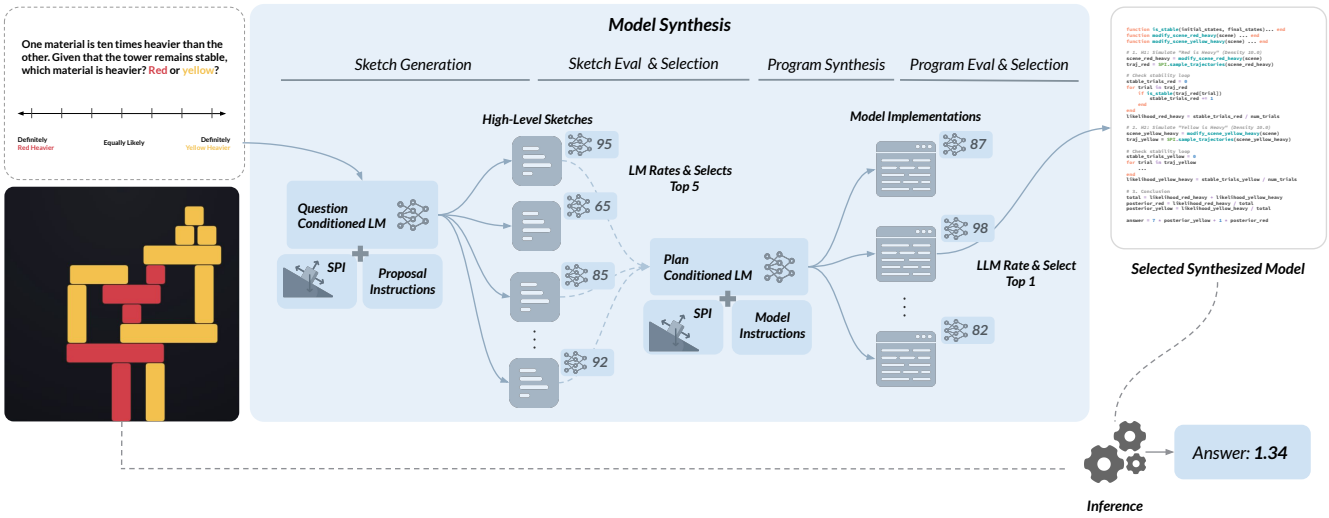

Figure 2: Illustration of the PRISM framework. A natural language query (including an output rating scale) is provided to the model synthesis framework, which starts by generating a set of program "sketches": high-level, natural language plans describing the program flow and concepts that must be defined, with knowledge of the grounded functions contained in the Scene Programming Interface (SPI). Next, these sketches are rated for how well they answer the initial query, and the top sketches are maintained. These sketches are then formalized into Julia code, which can query the SPI. Programs undergo lightweight assessment for how well they adhere to the logic laid out in the sketch, are run on sample data to ensure they can execute, and the best-scoring runnable model is selected. Finally, probabilistic inference is run in the selected program, taking in specific scene configurations, and outputting a final rating to answer the original question.

relevant, what intervention or hypothetical world to consider, which trajectories to simulate, and how to summarize the resulting distribution of outcomes into an answer, using iterative coarse-to-fine reasoning. This model synthesis is implemented as a set of structured language model (LM) queries to produce programs based on the natural-language question, designed to iteratively produce and assess more detailed program descriptions, ending in executable code (Fig. 2). Crucially, the LM is given knowledge about what is possible with grounded physical reasoning via a "Scene Programming Interface" (SPI), which is a set of functions that the model can deploy to perceive, manipulate, and simulate the scene. This pipeline confines the role of the LM to doing what we think language does: determining how to route information to other cognitive systems with deeper, domain-specific knowledge (Casto et al., 2025), rather than performing reasoning or simulation directly.

Problem domain

Here, we study PRISM in the domain of reasoning about block towers (Fig. 1). Many prior works study diverse tasks for human physical reasoning – including prediction (Battaglia et al., 2013), inference (Hamrick et al., 2016), causal reasoning (Zhou et al., 2023), and action planning (Hamrick et al., 2018) – via stimuli involving block towers. This prior work provides us with models that have previously been validated as explanations of human cognition, which we refer to as Gold Standard models and adapt to the particular scenarios studied here. Thus we can use the Gold Standard models both as a yardstick to assess the performance of PRISM, and as an a priori expectation of human behavior.

Model synthesis

PRISM performs model synthesis via a series of cognitively motivated structured generation and scoring steps, combining LM proposals with simple automated checks similar to Wong et al. (2025). This procedure has four stages:

High-level program sketch generation. When faced with a new question, a reasoner may first come up with potential "sketches" for how they would go about answering that question. In PRISM, given the question and a brief description of the SPI, the LM first samples several (15) natural language 'sketches' of a program that decomposes the query into ordered steps (e.g., "load the scene; run multiple noisy simulations; estimate the fraction of trials in which the tower collapses; report that probability"), along with helper functions that would be useful to define.

Sketch evaluation and selection. With multiple proposals, a boundedly-rational reasoner then needs to decide which proposals to focus their attention on. We instantiate this filtering process by using an LM to rate (on a scale of $0-100$) each candidate program sketch for logical coherence and coverage of the question's requirements. The top five highest scoring sketches are maintained.

Granular program synthesis. Sketches are then refined into formal symbolic code that is allowed to interface with the SPI. Specifically, for each selected sketch, an LM is queried with the text of that sketch and SPI, and the LM then generates Julia code to implement the program. The synthesized program is additionally asked to provide a return value corresponding to the 1-7 response scale, but how this value is calculated is

left up to the LM.[1]

Program evaluation and selection. Here, we again assume a resource-constrained reasoner only considers one model of the world to provide their response to the query. Thus a reasoner may conduct a few lightweight checks (e.g., for semantic and syntactic sensibility) over synthesized models before choosing which to focus reasoning on. We instantiate this by having an LM rate each program separately from $0-100$ for compliance with the natural-language program and task question, including correct use of the API and well-formed outputs. Overall scores for each model are calculated by multiplying the natural-language program sketch score with the model score similar to reweighting particles in Sequential Monte Carlo algorithms. Each model is then compiled and executed on a small test scene consisting of a set of four stacked blocks – if it fails to execute, that model is eliminated.[2] Finally, the remaining model with the best score is selected and probabilistic reasoning is run in this model.

Scene Programming Interface

We endow the model with a Scene Programming Interface (SPI), which we define as a set of functions for operating on physical scenes. The SPI is meant to capture other cognitive modules that may be engaged as part of reasoning (Casto et al., 2025; Spelke & Kinzler, 2007). Each function in the SPI is defined with a name, a typed set of inputs, a typed output, and a short natural language description of its purpose.

These functions are used to: (1) *load in* ("perceive") the scene as a collection of objects (each with a position, size, and color) and output a structured world representation that consists of a set of objects with both visible properties and physically-relevant latent properties (e.g., mass or friction); (2) *select or manipulate a subset of the objects* (e.g., change object mass or remove objects), giving the synthesized program direct control over the perceived scene while not needing to operate directly over raw pixel data, and (3) provide a distribution over future states using a *stochastic physics simulator*. This simulator uses the Chipmunk2D rigid-body physics engine (Lembcke, 2013), but adds stochasticity similar to prior instantiations of probabilistic physics: there is assumed to be perceptual noise leading to small Gaussian perturbations in the initial poses of all blocks (Battaglia et al., 2013), and stochasticity is added to collisions by perturbing the restitution and direction of impulse transfer using Gaussian or von Mises noise respectively (Allen et al., 2020; Wang et al., 2024). To ensure that the model has an appropriate range of future states, we fix this function to produce 50 samples. Each sampled simulation contains one possible evolution of the scene under the noisy dynamics, including the trajectories (position and rotation) for each object.

[1]While it is possible that this output might not be properly formatted (e.g., text or a number outside the scale), in practice we find that this does not happen at runtime as all errors are caught by the 'program evaluation' phase of model synthesis.

[2]If all models are eliminated, the top scoring model is fed back to the LM which is asked to repair the code while preserving the intended logic, and the model is tested again. This repeats twice; if the program remains erroneous the entire model synthesis pipeline is repeated.

Alternative models

We consider a set of alternative models as comparisons:

Gold standard models. These models are hand-designed for each scenario, hewing closely to the models defined in prior work. These models allow us to assess how closely PRISM matches with prior best practices.

Ablated simulation. We consider a version of PRISM that is identical except the noisy simulation function is removed from the SPI. This variant tests the contribution of physical knowledge; this model can understand scenes and produce structured programs, but cannot use simulation to reason.

Vision-language models. We compare our model against the base model (Gemini Flash Lite 2.0) used for program synthesis, as well as a more powerful model in the same family (Gemini Pro 2.5). Both of these models are provided with the task query as well as stimulus input. We test two variants of stimulus input in the prompt: either providing an image to the VLM ('Flash (Image)' and 'Pro (Image)'), or the structured JSON scene description, including brief notes on how to interpret the file (e.g., "this JSON describes a scene with a set of blocks"; we call these models 'Flash (JSON)' and 'Pro (JSON)'). These model comparisons test the necessity of explicit program synthesis and whether larger, more powerful models can make up for the structured reasoning of PRISM.

Experiment

To test how well PRISM captures human physical reasoning, we collected people's judgments on block tower stimuli and questions (Fig. 1).

We constructed four tasks that each consisted of showing participants an image of a block tower with a question that asked for a response on a scale from 1-7. The questions were similar to the inspiring experiments, with adjustments made to allow for a common response scale. For question and response scale wording, see Fig. 1.

Participants 20 participants were recruited for the experiment, with a median completion time of 24 minutes. All participants made judgments about all towers and tasks.

Procedure and stimuli We created 30 different block towers for each question. These stimuli were created such that they produced an approximately uniform distribution of judgments using the Gold Standard models, in order to induce a range of judgments from our participants.

The tasks were presented to participants in blocks, with all 30 of the same task judgments kept together but presented in a randomized order. The order of task blocks was randomized across participants. Between blocks, participants were presented with lightweight instructions, describing the new task in two sentences (e.g., "In this section you will see various block towers. Your task is to judge how likely each tower is to fall down.") and the question.

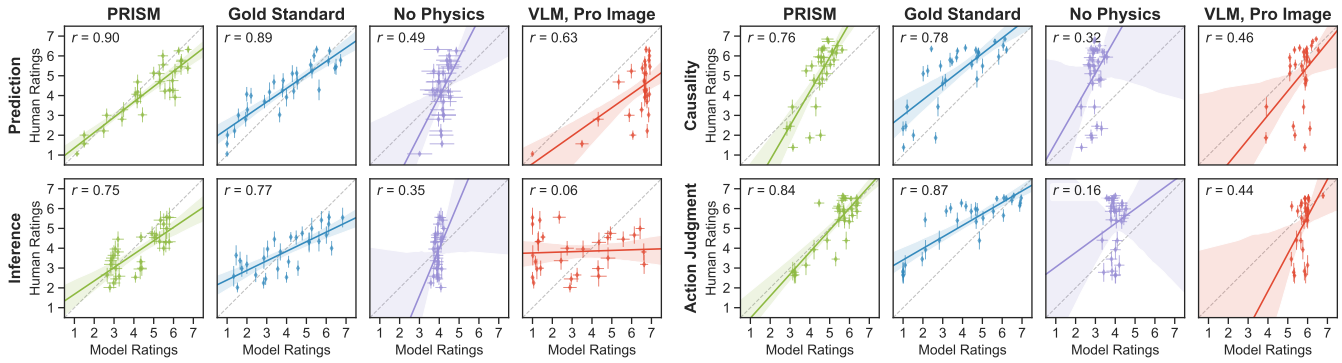

Figure 3: Trial-wise correlations between model ratings (x-axis) and human ratings (y-axis), split by model and scenario (tower). Solid lines and shaded regions represent the best fit line with 95% CI; dashed gray lines represent the identity line. Across all four scenarios, PRISM and the Gold Standard model describe people approximately as well. However, the model without simulation and the largest vision-language model tested fail to reliably capture human behavior.

Analysis Our primary measure is trial-wise correlation between human and model judgments. We averaged all 20 participant judgments for each trial. For all models except the Gold Standard models (which do not vary), we average over 50 synthesized programs or runs to account for the distribution over judgments induced by variations in model synthesis or variability in language model outputs.

Results

Across all four scenarios, PRISM explains participants' judgments well, better than all ablations and VLMs (see Fig. 3). There is no evidence of differences in explanatory power between PRISM and the Gold Standard models (all ps>0.75 across all four scenarios by Fisher's z-test). Alternate models capture substantially less variance in people's judgments (no-physics ablations: all ps<0.05; all VLMs: all ps<0.067, except for Pro (JSON) on the causal task: $p = 0.2$). While all VLMs are moderately correlated with people for the prediction and action judgment queries, the models are highly "jagged" across queries, depending on the prompt (JSON prompting yields a better fit to people for the inference queries than directly giving the image) and across model type (Flash is substantially worse across prompt types for the causality query).

We next focus on comparisons between PRISM and the Gold Standard models. There is strong correlation between the average judgments of PRISM and the average judgments of the Gold Standard models, similar to the correlations between those models and human judgments (all ps>0.75 by Fisher's z-test), and approaching the noise ceiling of human split-half correlations (Fig. 4). Nonetheless, PRISM does explain some variance in human judgments not captured by the Gold Standard models (partial correlations for prediction: $r_{partial} = 0.46$; inference: $r_{partial} = 0.16$; causal reasoning: $r_{partial} = 0.40$; action judgment: $r_{partial} = 0.43$). This suggests that PRISM might be picking up on small differences in the definition or use of physics – e.g., different definitions of stability or different counterfactual hypotheses – that people are using but the Gold Standard models by definition do not.

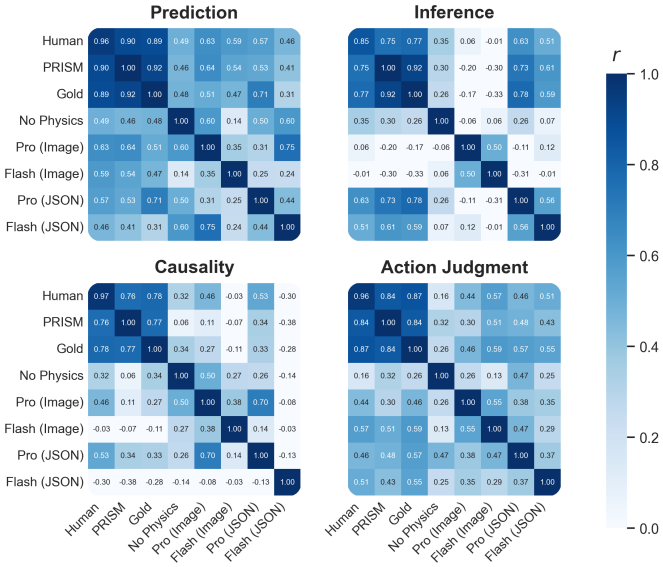

Figure 4: Trial-wise correlations between model responses and participants or other models. We find strong clustering between PRISM, people, and the Gold Standard models, suggesting that PRISM is recovering similar programs to those that have previously been shown to explain people's behavior.

As each run of PRISM produces an interpretable program for each scenario (Fig. 6A) we can analyze how well each generated model or each participant can be explained by other participants. This allows us to compare variability across people and generated models, as well as to assess model generation errors. People were generally well-correlated with each other (with some exceptions in the inference task), though with some variability (Fig. 5A). PRISM follows a similar distribution, though slightly more peaked, and includes more outliers, such as models with highly negative correlations and some generated models that produced constant predictions across all stimuli (Fig. 5B).

We next investigate outliers in the model distributions to

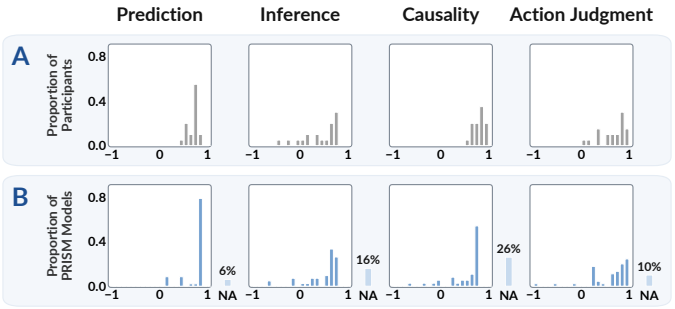

Figure 5: Histogram of correlations across trials between **(A)** each participant and the average of all other participants and **(B)** each of the 50 models proposed by PRISM and the average participant judgment. N/A next to the PRISM correlation indicates the proportion of synthesized models that produce constant responses across all trials and thus the correlation is undefined.

understand discrepancies between PRISM and people. We focus on two particular errors: strong anti-correlation with people, and no variation in responses (leading to undefined correlations; Fig. 5). The undefined correlations are often the result of errors in the definition of stability – either defining it in such a way that everything is stable, or nothing is stable. Similarly, the anti-correlations are often driven by sign errors, e.g., reversing the order of the Likert scale (Fig. 6B). In these cases, the PRISM framework is proposing reasonable program structures, but failing to implement those programs in a human-like way.

Discussion

How do people reason so flexibly and so fast about such a wide range of problems in the physical world? Here, we extend recent work positing that people reason about arbitrary causal inference problems by synthesizing and conducting probabilistic inference in bespoke models (Wong et al., 2025; Brooke-Wilson, 2023), to describe how people might synthesize models that ground out in the rich, structured knowledge people have about the physical world. We present the Physical Reasoning via Interpretable Synthesis of Models framework, which captures people's judgments across a range of queries one may ask about the physical world, reaching the level of "Gold Standard" models hand-tuned to describe this behavior, and outperforming large-scale vision-language models. Crucially, we support prior work that suggests that probabilistic simulation is the foundation for physical reasoning like the questions studied here, as a variant of PRISM without this capability cannot explain the graded physical judgments that our participants make.

Yet, despite these successes, we observe two main discrepancies between people and PRISM. While people all tend to make judgments that are reasonably correlated with each other, the synthesized models are jagged: though most successfully capture human judgments, some make errors in implementing the synthesized models that diverge from people. There is also less diversity in the outputs of the successful

A) Success Case (r=0.75)

```
function is_stable(initial_states, final_states)
    # Check if any block moved more than a small threshold
    threshold = 0.1 * min_block_size
    for i in 1:length(blocks)
        dist = distance(initial_states[i], final_states[i])
        if dist > threshold
            return false
        end
    end
    return true
end

function modify_scene_red_heavy(scene) ... end
function modify_scene_yellow_heavy(scene) ... end

# 1. H1: Simulate "Red is Heavy" (Density 10.0)
scene_red_heavy = modify_scene_red_heavy(scene)
traj_red = SPI.sample_trajectories(scene_red_heavy)

# Check stability loop
stable_trials_red = 0
for trial in traj_red
    if is_stable(traj_red[trial])
        stable_trials_red += 1
    end
end
likelihood_red_heavy = stable_trials_red / num_trials

# 2. H2: Simulate "Yellow is Heavy" (Density 10.0)
scene_yellow_heavy = modify_scene_yellow_heavy(scene)
traj_yellow = SPI.sample_trajectories(scene_yellow_heavy)

# Check stability loop
stable_trials_yellow = 0
for trial in traj_yellow
    ...
end
likelihood_yellow_heavy = stable_trials_yellow / num_trials

# 3. Conclusion
total = likelihood_red_heavy + likelihood_yellow_heavy
posterior_red = likelihood_red_heavy / total
posterior_yellow = likelihood_yellow_heavy / total

answer = 7 * posterior_yellow + 1 * posterior_red
```

A tower is unstable if ***any*** blocks moved beyond some threshold

Tests first hypothesis: *red blocks are heavier*

Tests second hypothesis: *yellow blocks are heavier*

Translates ***posterior probabilities*** of the set of hypotheses to the 1 to 7 scale

B) Failure: Sign Error (r=-0.68)

```
rating = 0.0
if posterior_h1 > posterior_h2
    rating = 1 + 6 * (posterior_h1 - 0.0) / (1.0 - 0.0)
else
    rating = 7 - 6 * (posterior_h2 - 0.0) / (1.0 - 0.0)
end
```

Previous calculations are correct, ***flips the scale*** when translating from probabilities

Figure 6: Example code generated by PRISM for the Inference task. **(A)** One example synthesized model that correlates well with people. This model appropriately sets up two separate hypotheses (either red blocks are heavier or yellow blocks are heavier), calculates the posterior conditioning on stability, and maps that probability to the 1-7 scale. **(B)** Code snippet from a poorly fitting model, where the mapping between posteriors and the 1-7 scale is reversed, though all other parts of the model work well. This causes the model to have highly negative correlations with participants.

generated models than in human judgments, suggesting additional variability in either the model generation process, or the types or amounts of simulation used to support people's reasoning. Future research will focus on extending the framework to address these gaps, for instance by extending the model to better account for individual differences in judgments via different methods of reasoning or different amounts of resources used (e.g., different numbers of simulations). Nonetheless, PRISM is an important first step towards explaining general, flexible physical reasoning in people.

References

Allen, K. R., Smith, K. A., & Tenenbaum, J. B. (2020). Rapid trial-and-error learning with simulation supports flexible tool use and physical reasoning. *Proceedings of the National Academy of Sciences*, *117*(47), 29302–29310.

Bates, C. J., Yildirim, I., Tenenbaum, J. B., & Battaglia, P. (2019). Modeling human intuitions about liquid flow with particle-based simulation. *PLoS computational biology*, *15*(7), e1007210.

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332.

Bi, W., Shah, A. D., Wong, K. W., Scholl, B. J., & Yildirim, I. (2025). Computational models reveal that intuitive physics underlies visual processing of soft objects. *Nature Communications*, *16*(1), 6303.

Brooke-Wilson, T. (2023). *Bounded rationality as a strategy for cognitive science* (Unpublished doctoral dissertation). Massachusetts Institute of Technology, Cambridge, MA.

Buschoff, L. M. S., Akata, E., Bethge, M., & Schulz, E. (2025). Visual cognition in multimodal large language models. *Nature Machine Intelligence*, *7*(1), 96–106.

Buschoff, L. M. S., Voudouris, K., Akata, E., Bethge, M., Tenenbaum, J. B., & Schulz, E. (2025). Testing the limits of fine-tuning to improve reasoning in vision language models. *arXiv preprint arXiv:2502.15678*.

Casto, C., Ivanova, A., Fedorenko, E., & Kanwisher, N. (2025). What does it mean to understand language? *arXiv preprint arXiv:2511.19757*.

Davis, E., & Marcus, G. (2015). The scope and limits of simulation in cognitive models. *arXiv preprint arXiv:1506.04956*.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2021). A counterfactual simulation model of causal judgments for physical events. *Psychological review*, *128*(5), 936.

Gerstenberg, T., Peterson, M. F., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2017). Eye-tracking causality. *Psychological science*, *28*(12), 1731–1744.

Hamrick, J. B., Allen, K. R., Bapst, V., Zhu, T., McKee, K. R., Tenenbaum, J. B., & Battaglia, P. W. (2018). Relational inductive bias for physical construction in humans and machines. In *Proceedings of the 40th annual conference of the cognitive science society.*

Hamrick, J. B., Battaglia, P. W., Griffiths, T. L., & Tenenbaum, J. B. (2016). Inferring mass in complex scenes by mental simulation. *Cognition*, *157*, 61–76.

Lembcke, S. (2013). *Chipmunk 2D Physics Engine.* Howling Moon Software.

Sanborn, A. N., Mansinghka, V. K., & Griffiths, T. L. (2013). Reconciling intuitive physics and newtonian mechanics for colliding objects. *Psychological review*, *120*(2), 411.

Smith, K. A., Hamrick, J. B., Sanborn, A. N., Battaglia, P. W., Gerstenberg, T., Ullman, T. D., & Tenenbaum, J. B. (2024). Probabilistic models of physical reasoning. In *Bayesian models of cognition: Reverse-engineering the mind.* MIT Press.

Smith, K. A., & Vul, E. (2013). Sources of uncertainty in intuitive physics. *Topics in cognitive science*, *5*(1), 185–199.

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental Science*, *10*(1), 89–96.

Wang, H., Jedoui, K., Venkatesh, R., Binder, F. J., Tenenbaum, J., Fan, J. E., ... Smith, K. A. (2024). Probabilistic simulation supports generalizable intuitive physics. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 46).

Wong, L., Collins, K. M., Ying, L., Zhang, C. E., Weller, A., Gerstenberg, T., ... Brooke-Wilson, T. (2025). Modeling open-world cognition as on-demand synthesis of probabilistic models. *arXiv preprint arXiv:2507.12547*.

Wong, L., Grand, G., Lew, A. K., Goodman, N. D., Mansinghka, V. K., Andreas, J., & Tenenbaum, J. B. (2023). From word models to world models: Translating from natural language to the probabilistic language of thought. *arXiv preprint arXiv:2306.12672*.

Yildirim, I., Gerstenberg, T., Saeed, B., Toussaint, M., & Tenenbaum, J. (2017). Physical problem solving: Joint planning with symbolic, geometric, and dynamic constraints. *arXiv preprint arXiv:1707.08212*.

Zhang, C. E., Wong, L., Grand, G., & Tenenbaum, J. B. (2023). Grounded physical language understanding with probabilistic programs and simulated worlds. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 45, pp. 3476–3483).

Zhou, L., Smith, K. A., Tenenbaum, J. B., & Gerstenberg, T. (2023). Mental jenga: A counterfactual simulation model of causal judgments about physical support. *Journal of Experimental Psychology: General*, *152*(8), 2237.