We thank all the reviewers for their insightful and constructive comments, and will revise the paper accordingly.

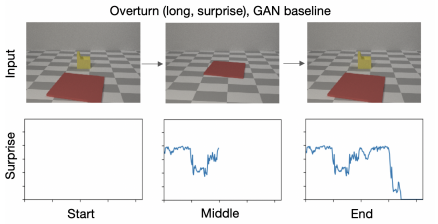
(R1) Goal. Our work has two aims. Our primary goal is to formalize human (especially infant) physical cognition, with models that can be tested by developmental psychologists. But we also aim to show how modeling infant cognition can inspire more robust AI vision systems that extract physical object representations from video and can detect violations of physical expectations to use as learning signals.*

(R1) Comparative run-times. On a standard CPU server, baselines run at 2–4s/frame, while ADEPT runs at 6–7s/frame. We currently have not optimized ADEPT for run-time performance and believe there is ample room for future improvement, such as GPU acceleration via differentiable particle filtering.†

(R1) Accuracy per-person. Calculating relative accuracy requires a comparison between surprise/control pairs, but each participant only observed one video from a pair (to avoid biases from seeing earlier near-identical scenes). We therefore must aggregate across participants to provide this distribution.

(R1) L2 loss for the GAN model. Using an L2-based surprise score, the GAN model has a relative accuracy of 0.41 on our dataset (vs. 0.63 with a discriminator-based surprise score, as shown in Table 1).

(R1) Baseline model images like Fig. 4. While baselines do not have interpretable internal belief states similar to the middle panels of Fig. 4, we can plot model surprise over time. For example, to the right we see the GAN model is not additionally surprised when the occluder fully rotates over the object (vs. ADEPT in which surprise spikes; Fig. 4b).



(R1, R2) Cognitive process underlying human surprise. ADEPT reflects a plausible hypothesis that people have a probabilistic, object-based model of intuitive physics (e.g., Battaglia et al., 2013‡), and that surprise is driven by low probability events under that model (e.g., Teglas et al.§). We plan to further explore the human/model match in future work, e.g., examining moment-by-moment surprise or neural correlates of object disappearance.

(R1, R2) ADEPT hyperparameters. We consider the difference between the sampling probabilities for surprising events and their losses a form of Importance Sampling that allows rare events to be captured with small, cognitively-plausible hypothesis sets by sampling those events more often than they should occur (see Lieder et al.¶ on this phenomenon in humans). Empirically, we also found that performance was insensitive to the exact values or tracking methods, so long as physics violations are more surprising than moderate amounts of perceptual uncertainty.

(R2) Problem setup and generalization. We designed our model to match objects based on general principles (e.g., color, shape, and size constancy). We stress that ADEPT's training was not specific to the test dataset: there were no unphysical scenes or violations in the training set. The derenderer is trained on sequences of three frames to infer the instant velocity of objects, not long-term motion patterns. The training set had motion patterns similar to the test videos to allow fair comparisons with baseline models, which do require longer sequences of motion to form predictions. We will clarify these points in revision.

(R2) Generalization to other datasets: IntPhys. We have run ADEPT on the IntPhys dataset, only retraining the derenderer to handle different visual object properties. Because the IntPhys test server is offline, we evaluate models on the validation set of scenes designed to test "object permanence," as described in Riochet et al. This set contains 90 matched sets of videos (2 plausible, 2 implausible within each set). ADEPT achieves an overall relative accuracy of 0.73, outperforming all baselines (Enc-Dec: 0.61, GAN: 0.53, LSTM: 0.65). As R2 noted, the videos of IntPhys have different visual and motion patterns (e.g., complex textures and gravitational motion), so the high performance on this second dataset suggests our model generalizes to situations where we have no control over the training and test data.

(R2) Failure cases. We agree and will move discussion of failure cases from supplemental to main text in the revision.

(R2) Dataset. We will release the dataset along with all code, human data, and model evaluations upon publication.

(R2) Occluder modeling. Our physics engine assumes motion changes require force, and so cannot capture the up-and-down motion of occluders. We chose to model them separately to avoid producing a constant surprise signal.

(R3) Additional information on metrics and fair comparisons. We follow IntPhys by using the 'relative accuracy' metric to compare plausible and implausible videos within each matched set, and also follow their choice of plausibility metrics for each baseline model. We think relative accuracy is the metric most well matched to the developmental literature, which compares reactions to surprising scenes directly to those in matched controls.

---

*Stahl, Feigenson. Observing the unexpected enhances infants' learning and exploration. Science 2015

†Jonschkowski et al. Differentiable particle filters: End-to-end learning with algorithmic priors. RSS 2018

‡Battaglia et al. Simulation as an engine of physical scene understanding. PNAS 2013

§Teglas et al. Pure reasoning in 12-month-old infants as probabilistic inference. Science 2011

¶Lieder et al. Over-representation of extreme events in decision making reflects rational use of cognitive resources. Psych. Rev. 2018