Opinion: Learning Intuitive Physics May Require More than Visual Data

Ellen Su* Solim LeGris * Todd M. Gureckis Mengye Ren

New York University
{ellensu, solim.legris}@nyu.edu

Abstract

Humans expertly navigate the world by building rich internal models founded on an intuitive understanding of physics. Meanwhile, despite training on vast quantities of internet video data, state-of-the-art deep learning models still fall short of human-level performance on intuitive physics benchmarks. This work investigates whether data distribution, rather than volume, is the key to learning these principles. We pretrain a Video Joint Embedding Predictive Architecture (V-JEPA) model on SAYCam, a developmentally realistic, egocentric video dataset partially capturing three children's everyday visual experiences. We find that training on this dataset, which represents 0.01% of the data volume used to train SOTA models, does not lead to significant performance improvements on the IntPhys2 benchmark. Our results suggest that merely training on a developmentally realistic dataset is insufficient for current architectures to learn representations that support intuitive physics. We conclude that varying visual data volume and distribution alone may not be sufficient for building systems with artificial intuitive physics.

1 Introduction

A core aspect of human intelligence is the ability to engage flexibly with our environments [26, 10, 9]. This ability allows us to pursue goals, respond dynamically to changing stimuli, and navigate flexibly in new settings. Previous research in cognitive science suggests that this aspect of human intelligence emerges from the rich internal models people build of the surrounding world [11, 19, 21]. We revisit the claim that what makes our mental models *useful* in the physical world is that they are constrained by internalized physical principles which reflect real-world dynamics [6, 33, 15, 21]. As proposed by Battaglia et al. [6], humans may rely on an "internal physics engine": a cognitive mechanism that runs coarse, probabilistic simulations of how physical scenes evolve over short time periods. This mechanism possibly underpins our capacity to act and learn flexibly in novel environments. To achieve embodied AI systems with human-like abilities, we posit that the underlying world model should also demonstrate that it has internalized physical principles. However, state-of-the-art (SOTA) deep learning models still fail to approach human-level performance on intuitive physics benchmarks [8]. Despite pretraining on millions of hours of video data and achieving impressive performance on motion understanding, video question-answering, and other downstream tasks [1], recent results by Bordes et al. [8] have provided evidence that these models perform just above chance at classifying between videos which are physically possible and impossible in the real world.

For humans, decades of developmental psychology research have shown that physical reasoning is learned implicitly without any formal education and requires very little training data—infants under the age of one reliably demonstrate knowledge about object permanence and continuity through space and time [27, 28, 3, 2]. Thus, our work mainly addresses an empirical question about data distribution: **to what extent can SOTA video models internalize physical principles by training on naturalistic and developmentally realistic data?** Rather than training video models on internet video and image

*Corresponding authors. Equal contribution.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Embodied World Models for Decision Making.

datasets, we pretrain a V-JEPA model on SAYCam [32], a developmentally realistic egocentric video dataset. SAYCam captures the richness and noisiness of early visual experiences, reflects a subset of the data that infants use to learn intuitive physics, and thus presents itself as a natural option for addressing learnability of physical principles with respect to data volume and distribution. We find that the V-JEPA architecture is unable to perform well on intuitive physics benchmarks even when trained on an approximation of the visual data distribution that human infants receive during development. Next, both SOTA models and our model, despite being trained on 0.01% of the total data volume, achieve only slightly above chance performance on IntPhys2. In our analyses, we additionally show that intuitive physics tasks prove challenging for both the V-JEPA and VideoMAE learning algorithms. Our results suggest that visual datasets alone may be an insufficient training data source for current SOTA models to learn representations which support intuitive physics reasoning. In line with previous research, we propose that future work should utilize embodied datasets (egocentric, multimodal, action-annotated) and develop further innovations for model architectures.

2 Pretraining V-JEPA on a developmentally realistic dataset

While much prior work has focused on domain-specific models for intuitive physics [30, 33], recent architectural advancements have allowed for promising improvements to physical understanding in general-purpose deep learning models [13, 14, 1]. In particular, the Video Joint Embedding Predictive Architecture (V-JEPA) model [5, 1] has been hypothesized to learn flexible representations which may promote performance in physical reasoning tasks [14]. The V-JEPA model is pretrained in a self-supervised manner and learns latent representations by predicting future states of masked frames. Recent work has demonstrated that it learns useful features which are applicable to many downstream tasks and domains [5]. Thus, we opted for this model to explore our main question.

2.1 SAYCam dataset

We chose the SAYCam dataset as our training data as it contains 472 hours of egocentric videos collected from the point of view of children interacting in the real world. We hypothesized that SAYCam videos would provide richer learning signal for physical reasoning than internet videos from simulations or online collections since it more closely approximates the input that infants obtain, and sometimes actively seek, from the environment. Prior work has explored learnability of vision and language from developmentally realistic and naturalistic data [24, 25]. In one instance, Orhan et al. [25] demonstrated that generic deep learning models are able to learn meaningful representations of visual objects when trained on SAYCam in a self-supervised manner. Dataset details are provided in Appendix A.

2.2 Evaluating intuitive physics

Intuitive physics understanding is difficult to measure. Without having direct access to the internal models of humans or AI systems, researchers often probe for specific knowledge or skills by analyzing behavioral patterns within a constrained and simplified environment [7]. Prior work in developmental psychology has established the violation-of-expectation paradigm, which relies on infant looking times to determine which of a pair of visual events is more surprising, and thus less aligned with the expectations of the child [22, 4]. Following suit, many machine learning benchmarks evaluate video models on intuitive physics concepts by extracting a surprise metric from model outputs for each video frame [8, 29, 34, 18]. This metric assumes that, if the predictions of the video model capture real-world physical principles, then physically plausible videos should yield lower surprise scores than implausible videos. The computation of our metrics are detailed in Appendix B. We chose to evaluate our model on the IntPhys2 benchmark, which was recently introduced by Bordes et al. [8] and improves upon previous intuitive physics understanding benchmarks by incorporating dynamic shadows and lighting, natural occlusions, and both fixed and moving camera shots. Additional benchmark details are included in Appendix C.

2.3 Implementation details

We pretrained a V-JEPA base model (215M parameters) with 16x16 patches at a spatial resolution of 224 (ViT-L/16_{224}) following the procedure set out in Bardes et al. [5] on the SAYCam dataset. Further details on model pretraining are described in Appendix D. We compared our model against two other V-JEPA models with differently sized and distributed training datasets and against a Video-MAE model trained on SAYCam [1, 5, 25]. V-JEPA-2-H-VM22M (654M parameters)[2] was pretrained on

[2] `https://huggingface.co/facebook/vjepa2-vith-fpc64-256`

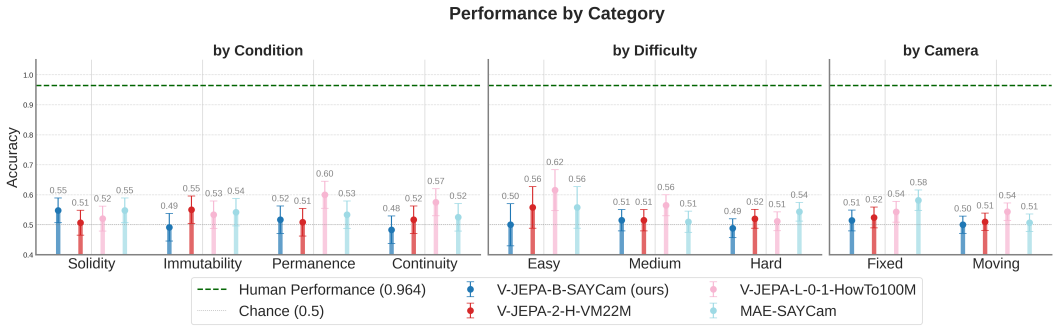

Figure 1: Breakdown of model accuracies on the IntPhys2 benchmark by physical condition, difficulty, and camera set up. All models achieve around chance performance at classifying between possible and impossible physical events with small variations in rank. Error bars mark standard error of the mean accuracy.

VideoMix22M [1] (22 million samples), which comprises one egocentric video dataset [16], three exocentric video datasets [20, 23, 35], and 1 image dataset [12] and amounts to over 1.73 million hours of video data. Next, we evaluate V-JEPA-L-0-1-HowTo100M (326M parameters) which is pretrained by Garrido et al. [14] on a 0.1% subsample of the HowTo100M dataset[23], and receives 128 hours of unique video data. While similarly sized to SAYCam, HowTo100M is made up of exocentric and general purpose online tutorial videos, allowing us to isolate the effect of video data distribution on task performance. Finally, the VideoMAE model released by Orhan et al. [25] is trained in a self-supervised manner on the SAYCam dataset. We include the performance of this model to again isolate the effect of model architecture on performance. Our training code is available at `https://github.com/eysu35/vjepa` and our pretrained model is available for download at `https://huggingface.co/eys8549/vjepa-saycam`.

3 V-JEPA pretrained on SAYCam yields similar performance to SOTA models

We find that our model achieves similar performance to all other benchmarked models. All models are performing at or slightly above chance (V-JEPA-B-SAYCam (ours): 0.50, V-JEPA-2-H-VM22M: 0.52, MAE-SAYCam: 0.53, V-JEPA-L-0-1-HowTo100M: 0.54), suggesting that neither current large-scale internet video datasets nor human-scale naturalistic developmental video datasets are sufficient for V-JEPA architectures to learn intuitive physics concepts. The video pairs in the IntPhys2 dataset vary by physical concept, difficulty, and camera set up (see Figure 1), which allows for more fine-grained analysis compared to previous benchmarks. Despite sharing similar architectures and being trained on drastically different datasets, the models show consistently poor overall performance. In contrast, Bordes et al. [8] evaluated human participants and reported a near-perfect score of 96.44% for overall classification accuracy. Our model, the VideoMAE model, and the V-JEPA large model were fully trained on a few hundred hours of unique video data. Given that this makes up less than 0.01% of the data volume which V-JEPA-2-H-VM22M was trained on, we observe that the performance gain across intuitive physics tasks does not scale well with the size of training dataset. While pretraining on larger and more diverse video datasets may lead to richer learned representations and performance improvements in novel downstream tasks, our results highlight the fact that intuitive physics reasoning in video models remains a challenge currently unsolved by the volume of visual data alone. Comparison to V-JEPA-L-0-1-HowTo100M and MAE-SAYCam also indicate that data distribution and model architecture did not result in significant differences in task performance.

3.1 Surprise analysis

We conducted a qualitative analysis to compare the fine-grained behavioral patterns of surprise values between the V-JEPA models following the analysis in Bordes et al. [8]. The surprise patterns across video conditions shed light into whether the surprise metric captures higher-level concepts in the videos and whether the model classification is grounded in its understanding of these events. Figure 2 shows the predicted values over frames of a pair of correctly classified videos (panel A) and incorrectly classified videos (panel B). We observe that the surprise predictions from the two models trained on small-scale datasets follow similar patterns and respond to video contents more so than the model trained on VideoMix22M. Finally, for both video pairs, the surprise patterns outputted by all

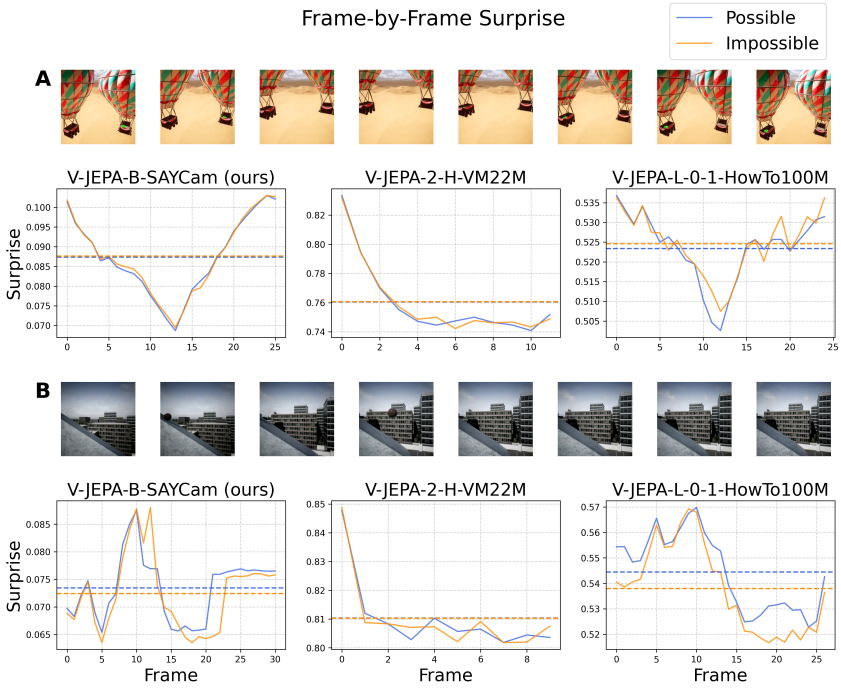

Figure 2: Model surprises for all V-JEPA models. **A)** Top: a subsample of video frames for an impossible video (violates object permanence). Bottom: Frame-by-frame surprise predictions for the corresponding possible/impossible video pair which all models classified correctly (average surprise for possible < average surprise for impossible). **B)** Top: a subsample of video frames for an impossible video (violates continuity). Bottom: Frame-by-frame surprise predictions for the corresponding possible/impossible video pair which all models classified incorrectly (average surprise for possible > average surprise for impossible). The surprise values were taken from the context which reported the best overall accuracy score. Dotted lines represent average surprise scores.

models are consistently similar for the possible and impossible videos, indicating that computing classification accuracy based on average surprise is highly sensitive to noise.

4 Discussion

Although previous results suggested that human-level intuitive physics emerges from large-scale video datasets using latent representation learning [14], recently published benchmarks suggest that physical reasoning in world models remains a machine learning challenge [8]. In this work, we explored whether the V-JEPA model could learn representations which reflect intuitive physics principles when trained on a video dataset that mimics the data a child is exposed to. Although we found that the model was able to match performance on some dimensions of the benchmark with significantly less data volume, our conclusion is that all evaluated models, which vary in size, pretraining data distribution, and architecture, are yet unable to learn representations which support intuitive physics reasoning. One limitation of our work is that the SAYCam dataset is of a drastically smaller size than the visual data children have access to and does not fully capture the range of visual experiences needed to learn intuitive physics. Thus, perhaps models would perform better on intuitive physics benchmarks if trained on a larger dataset of this distribution. Second, SAYCam only captures visual information and does not incorporate other aspects of embodiment which may be critical to learning intuitive physics. For example, perhaps children learn to understand physics by reconciling their self-directed movements with their visual fields, and thus incorporating motion or action data alongside visual data may be key to learning good representations for intuitive physics. Other datasets such as BabyView [17] might be more conducive to learning intuitive physics since they annotate egocentric videos with accelerometer data. In general, while egocentric and developmentally realistic video data are components of embodiment, we hypothesize that the signal captured in an agent's actions and motion would contribute meaningfully to internalizing physical principles in machine learning world models. The effort toward artificial intuitive physics aligns with the development of effective and efficient embodied AI systems that can operate freely in the world.

Acknowledgements

The authors gratefully acknowledge helpful contributions and input from Quentin Garrido and Brenden Lake. This work was supported in part through the NYU IT High Performance Computing resources, services, and staff expertise.

References

[1] M. Assran, A. Bardes, D. Fan, Q. Garrido, R. Howes, Mojtaba, Komeili, M. Muckley, A. Rizvi, C. Roberts, K. Sinha, A. Zholus, S. Arnaud, A. Gejji, A. Martin, F. R. Hogan, D. Dugas, P. Bojanowski, V. Khalidov, P. Labatut, F. Massa, M. Szafraniec, K. Krishnakumar, Y. Li, X. Ma, S. Chandar, F. Meier, Y. LeCun, M. Rabbat, and N. Ballas. V-jepa 2: Self-supervised video models enable understanding, prediction and planning, 2025. URL https://arxiv.org/abs/2506.09985.

[2] R. Baillargeon. The acquisition of physical knowledge in infancy: A summary in eight lessons. In U. Goswami, editor, *Blackwell Handbook of Childhood Cognitive Development*, pages 46–83. Blackwell, Oxford, 2002.

[3] R. Baillargeon and J. DeVos. Object permanence in young infants: Further evidence. *Child Development*, 62(6):1227–1246, dec 1991. ISSN 0009-3920. doi: 10.2307/1130803.

[4] R. Baillargeon, E. S. Spelke, and S. Wasserman. Object permanence in five-month-old infants. *Cognition*, 20(3):191–208, 1985. doi: 10.1016/0010-0277(85)90008-3.

[5] A. Bardes, Q. Garrido, J. Ponce, X. Chen, M. Rabbat, Y. LeCun, M. Assran, and N. Ballas. Revisiting feature prediction for learning visual representations from video. In *International Conference on Learning Representations (ICLR)*, 2024.

[6] P. W. Battaglia, J. B. Hamrick, and J. B. Tenenbaum. Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 110(45):18327–18332, 2013. doi: 10.1073/pnas.1306349110.

[7] D. Bear, E. Wang, D. Mrowca, F. Binder, H.-Y. Tung, P. T, C. Holdaway, S. Tao, K. Smith, F.-F. Sun, F.-F. Li, N. Kanwisher, J. Tenenbaum, D. Yamins, and J. Fan. Physion: Evaluating physical prediction from vision in humans and machines. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/d09bf41544a3365a46c9077ebb5e35c3-Paper-round1.pdf.

[8] F. Bordes, Q. Garrido, J. T. Kao, A. Williams, M. Rabbat, and E. Dupoux. Intphys 2: Benchmarking intuitive physics understanding in complex synthetic environments, 2025. URL https://arxiv.org/abs/2506.09849.

[9] R. A. Brooks. Intelligence without representation. *Artificial Intelligence*, 47(1-3):139–159, 1991.

[10] A. Clark. *Being There: Putting Brain, Body, and World Together Again*. MIT Press, 1997.

[11] K. J. W. Craik. *The Nature of Explanation*. Cambridge University Press, Cambridge, UK, 1943.

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[13] C. Finn, I. J. Goodfellow, and S. Levine. Unsupervised learning for physical interaction through video prediction. *CoRR*, abs/1605.07157, 2016.

[14] Q. Garrido, N. Ballas, M. Assran, A. Bardes, L. Najman, M. Rabbat, E. Dupoux, and Y. LeCun. Intuitive physics understanding emerges from self-supervised pretraining on natural videos, 2025. URL https://arxiv.org/abs/2502.11831.

[15] T. Gerstenberg and J. B. Tenenbaum. Intuitive theories. *Oxford handbook of causal reasoning*, pages 515–546, 2017.

[16] R. Goyal, S. E. Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fründ, P. Yianilos, M. Mueller-Freitag, F. Hoppe, C. Thurau, I. Bax, and R. Memisevic. The "something something" video database for learning and evaluating visual common sense. *CoRR*, abs/1706.04261, 2017.

[17] C.-C. Hsu, Z. Meng, Y. Wang, O. Tslil, M. Z. Shou, A. Pentland, C. Breazeal, and D. Yamins. BabyView: A Large-Scale Real-World Egocentric-Vision Dataset of Infants in Their First Three Years of Life. In *European Conference on Computer Vision (ECCV)*, 2024.

[18] S. Jassim, M. Holubar, A. Richter, C. Wolff, X. Ohmer, and E. Bruni. Grasp: A novel benchmark for evaluating language grounding and situated physics understanding in multimodal language models, 2024.

[19] P. N. Johnson-Laird. *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Harvard University Press, 1983.

[20] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017.

[21] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman. Building machines that learn and think like people, 2016.

[22] F. Margoni, L. Surian, and R. Baillargeon. The violation-of-expectation paradigm: A conceptual overview. *Psychological Review*, 131(3):716, 2024.

[23] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*, 2019.

[24] A. E. Orhan and B. M. Lake. Learning high-level visual representations from a child's perspective without strong inductive biases, 2023.

[25] A. E. Orhan, W. Wang, A. N. Wang, M. Ren, and B. M. Lake. Self-supervised learning of video representations from a child's perspective, 2024.

[26] R. Pfeifer and J. C. Bongard. *How the Body Shapes the Way We Think: A New View of Intelligence*. MIT Press, 2006.

[27] J. Piaget. *The Origins of Intelligence in Children*. International Universities Press, New York, 1952.

[28] J. Piaget. *The Construction of Reality in the Child*. Basic Books, 1954.

[29] R. Riochet, M. Y. Castro, M. Bernard, A. Lerer, R. Fergus, V. Izard, and E. Dupoux. Intphys 2019: A benchmark for visual intuitive physics understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5016–5025, 2022. doi: 10.1109/TPAMI.2021.3083839.

[30] K. Smith, L. Mei, S. Yao, J. Wu, E. Spelke, J. Tenenbaum, and T. Ullman. Modeling expectation violation in intuitive physics with coarse probabilistic object representations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[31] E. S. Spelke. Preferential-looking methods as tools for the study of cognition in infancy. In G. Gottlieb and N. A. Krasnegor, editors, *Measurement of Audition and Vision in the First Year of Postnatal Life: A Methodological Overview*, pages 323–363. Ablex Publishing, 1985.

[32] J. Sullivan, M. Mei, A. Perfors, E. H. Wojcik, and M. C. Frank. Saycam: A large, longitudinal audiovisual dataset recorded from the infant's perspective. *Open Mind : Discoveries in Cognitive Science*, 5:20 – 29, 2020.

[33] T. D. Ullman, E. Spelke, P. Battaglia, and J. B. Tenenbaum. Mind games: Game engines as an architecture for intuitive physics. *Trends in Cognitive Sciences*, 21(9):649–665, 2017. ISSN 1364-6613. doi: https://doi.org/10.1016/j.tics.2017.05.012.

[34] L. Weihs, A. Yuile, R. Baillargeon, C. Fisher, G. Marcus, R. Mottaghi, and A. Kembhavi. Benchmarking progress to infant-level physical reasoning in AI. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856.

[35] R. Zellers, J. Lu, X. Geng, S. Poria, C. Xiong, R. Krishna, G. Bar-Haim, E. Botzer, O. Shapira, Y. Bitton, R. Rinott, and Y. Bengio. MERLOT Reserve: Neural Script Knowledge through Vision and Language and Sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18458–18469, 2022.

A SAYCam dataset

SAYCam is a longitudinal audiovisual dataset of head cam recordings collected from three young children (nicknamed S, A, and Y) from the ages of 6 to 32 months [32]. Each infant wore a head cam for approximately 2 hours per week over the course of approximately 2.5 years. In total, SAYCam contains 472 hours of video, with 194, 141, and 137 hours of video from S, A and Y, respectively. The data for each child consist of a set of continuous, natural and uninstructed recordings, each between 1 to 2 hours. All of the videos take place in the home environments of the children and include activities such as crawling, being held, lying, or sitting [32].

B Violation of Expectation Metric

The violation-of-expectation paradigm has been extensively used in developmental psychology to evaluate psychological constructs in preverbal infants [4, 22]. In a typical experimental setup, infants are presented with two similar visual scenes following habituation, where one of the scenes contains an event which is hypothesized to appear surprising only if the infant has some understanding of the underlying property being tested. Surprise is usually obtained by measuring relative looking time [31], and is used to determine whether a concept violation has occurred.

In recent machine learning literature, this notion of surprise has been adapted to serve as a proxy for model understanding [14]. Both types of models we evaluated (pixel and latent prediction methods) can be evaluated in the same way, with the only difference being how the target of the prediction is encoded. For V-JEPA, we use the latent representations of the future obtained by encoding the video and then only keeping the future frames, while for the VideoMAE trained on SAYCam, the target is simply the normalized future of the video. Considering a video V with frames $1, ..., T$, a context encoder f_θ handling C frames, a target encoder g_ψ producing the ground truth M future frames from the video, and a predictor predicting M frames in the future, we can measure surprise at time t as

$$S_t = ||p_\phi(f_\theta(V_{t:t+C})) - g_\psi(V_{t:t+C+M})||_1 \tag{1}$$

Next, prior work reports different design choices for computing model accuracies from surprise values. Building on this work, we define our accuracy as the fraction of video pairs in which the average surprise (across all video frames) for a possible video is lower than the impossible video. As the evaluation is conducted over multiple context lengths as a hyperparameter, our final performance score is taken from the context which maximized the model accuracy for each condition following Bordes et al. [8] and Garrido et al. [14]. The average surprise per context is computed as,

$$\texttt{AvgSurprise} = \frac{1}{T} \sum_{t \in \{1, 1+s, ..., T-(C+M)\}} S_t \tag{2}$$

where s is a stride parameter for the sliding window, reducing the amount of compute used to evaluate each video clip. For all our evaluations, we use $s = 2$.

C Benchmark details

The IntPhys2 benchmark [8] builds off of the first version by including pixel-to-pixel aligned pairs of videos simulating physically impossible and possible events. The creators of this benchmark generated all video pairs with a photorealistic simulation engine which avoids challenges like data leakage and the appearance of spurious features from video stitching.

The benchmark contains contains three main splits: the "Debug" (30 pairs), "Main" (506 pairs) and "Held-out" (172 pairs) sets. Metadata containing labels, scene and difficulty among other attributes were only released for the debug and main splits. This new collection of generated videos are photorealistic and include dynamic shadows and lighting, natural occlusions, and both fixed and moving camera shots. All these components add richness to the dataset and establishes the benchmark as a more comprehensive and realistic evaluation benchmark for intuitive physics understanding. Following Bordes et al. [8], we evaluate our model on the "Main" data split which contains metadata annotations. Model performance is measured across object permanence, object immutability, spatio-temporal continuity, and solidity.

D Implementation Details

The pretraining procedure we followed was heavily influenced by the code and configurations set up in Bardes et al. [5] and Garrido et al. [14]. As we were comparing our pretrained model to the V-JEPA version 1 model trained on a large-scale video dataset, we mimicked the original training process as closely as possible. However, the drastic reduction in dataset size required us to make a few changes: reducing the training time from 200 (60,000 steps) to 40 epochs (12,000 steps), reducing the warm up period from 40 to 10 epochs, and fixing the weight decay at 0.04 rather. In addition, we partitioned the data from each child in SAYCam 80/20 into a train and validation set, monitored the loss curves of both to ensure convergence, and implemented early stopping to mitigate overfitting. Finally, we trained our model across 2 NVIDIA A100-SXM4-40GB GPUs.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims in our abstract and introduction reflect our empirical results and identify key areas for future research.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: We discuss the limitations of our pretraining dataset (size) and the missing components between egocentric visual data and embodiment.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: Our paper presents empirical results, but we justify the assumptions we made in defining our experiment.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer:

   Justification: We used open source models and datasets and will release the repository containing our code as well as the checkpoints for our pretrained model in the final version of this paper. They are currently omitted to preserve anonymity.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We include sufficient detail on hyperparameters, implementation configurations and compute resources such that our pretraining job can be reproduced. Our main experimental results can also be reproduced by running our model (with public checkpoints) on a publicly available benchmark.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.

- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All details are included in our appendix D and in our methods section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We include error bars which reflect the standard error of the average model accuracies for each model. Since all models performed close to chance, we did not consider statistical testing between model performances to be useful.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

Justification: We include the compute resources we used in appendix D and followed the distributed training protocol set out in prior work (cited in the methods and appendix).

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: Our research did not directly involve human research participants nor did we disclose any private or sensitive information about human subjects involved in the dataset we used. Our work has no potential harmful downstream consequences.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Our work contributes to a growing literature in world models and machine learning for physical reasoning. These efforts will enable future embodied artificial intelligence and the development of many useful technologies for society.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our work poses no risks for data leakage or other security concerns.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We use the SAYCam dataset, published under a CC BY 4.0 license, and the V-JEPA model, published under a CC BY-NC 4.0 license.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We do not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our work does not include crowdsourcing or research from human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our work does not include crowdsourcing or research from human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: This work does not involve LLMs as part of the core method development.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.