Towards aligned body representations in vision models

Andrey Gizdov[1,2*†], **Andrea Procopio**[1,3*], **Yichen Li**[1], **Daniel Harari**[2], **Tomer Ullman**[1]

[1]Harvard University [2]Weizmann Institute of Science [3]Bocconi University

{andreygizdov, aprocopio, yichenli, tullman}@fas.harvard.edu, hararid@weizmann.ac.il

Abstract

Human physical reasoning relies on internal "body" representations — coarse, volumetric approximations that capture an object's extent and support intuitive predictions about motion and physics. While psychophysical evidence suggests humans use such coarse representations, their internal structure remains largely unknown. Here we test whether vision models trained for segmentation develop comparable representations. We adapt a psychophysical experiment conducted with 50 human participants to a semantic segmentation task and test a family of seven segmentation networks, varying in size. We find that smaller models naturally form human-like coarse body representations, whereas larger models tend toward overly detailed, fine-grain encodings. Our results demonstrate that coarse representations can emerge under limited computational resources, and that machine representations can provide a scalable path toward understanding the structure of physical reasoning in the brain.

Introduction

Human perception is concerned with *what* entities are present, *where* the entities are, and *how* a scene will unfold (Marr 2010; Freyd 1987). The problems of 'what', 'where, and 'how' are coupled but separate, and it is likely that they are supported by different computations and representations. In particular, for the purposes of inferring the identity of objects (e.g. telling apart a thermos and a water bottle), it may be important to have detailed object segmentations. But for the purposes of estimating *how* a physical scene will unfold, which objects will collide and where they will end up, it is sufficient and more cost-effective to have more coarse object-representations (e.g. if all that matters is catching them, a bottle and a cup can roughly be approximated by a cylinder).

The separation between coarse and fine-grain object segmentations for different purposes is widely used in engineering, in particular, in simulated environments such as games and animations. In humans, it has been suggested that the brain has a similar processing split between physics and graphics (Ullman et al. 2017; Balaban and Ullman 2025), with fine-grain meshes being used for the purposes of rendering and recognition, and coarse-grain bodies for the purposes of prediction and action. This split maps onto a neural division between the dorsal and ventral streams in human vision, and there is psychophysical evidence that humans make use of coarse body approximations (Li et al. 2023a) (see Figure 1a) in physical reasoning. Developmental studies also show that infants first carve the visual field into cohesive but rough volumetric entities with approximate spatial extent, before they infer contact relations, support, or motion trajectories (Spelke 1990; Baillargeon 2004). This coarse-object representation has proven useful in machine learning as well, as the basis for computational studies that built models of human core knowledge (Smith et al. 2019).

While psychophysical studies (Li et al. 2023a) suggest that humans rely on approximate internal representations of object bodies, the nature and structure of these representations remain largely unknown. Behavioral methods can only offer indirect, low-resolution glimpses into these internal encodings, and cannot reveal their geometric or computational form. Humans construct their understanding of objects through a largely bottom-up visual process—segmenting, grouping, and localizing entities before reasoning about their dynamics and causal relations (Spelke 1990; Baillargeon 2004). Artificial segmentation models, which are trained to perform a similar decomposition of scenes into discrete entities, therefore offer a natural computational proxy for exploring how such representations might form in the human brain (Gizdov, Ullman, and Harari 2025). If modern vision models trained for segmentation or prediction share representational structure with humans (which is *not* clear a priori), they could help reveal the hidden organization of object representations that support intuitive physics and action.

Given this potential, a central question emerges: do the latent object representations that arise in vision models resemble the coarse body representations humans rely on for physical reasoning (see Figure 1)? In nearly all segmentation models and datasets where an explicit teaching signal, reward, or loss function is used, the gold standard for accuracy

[*] Equal contributions.
[†] Corresponding author.

is ground-truth segmentation or pixel-perfect human annotation. Such fine-grained segmentation may diverge from the approximate object representations that underpin human physical intuition. This misalignment is both inefficient and risky: fine-grain representations waste computation and may lead agents to mispredict human behavior, since people act based on intuitive, coarse-grain physics rather than exact geometric detail (Li et al. 2023a). *Unlike* in the study of language models, there has been little exploration of the alignment between the body representations humans use when reasoning about physics and the object representations artificial intelligence (AI) models use for vision, which is often a first bottom-up step for physical reasoning systems.

Contributions. In this work, we make the following contributions:

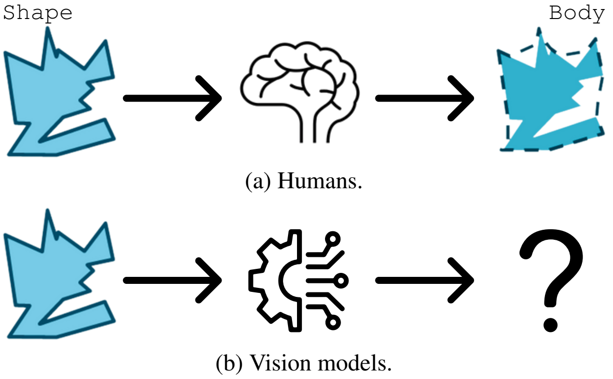
1. **Object representations in vision models are similar to the body representations in humans**. We show that artificial segmentation networks form coarse body representations similar to those discovered in humans, particularly in the context of intuitive physical reasoning. To compare the two, we propose a framework that adapts a psychophysical experiment given to 50 human participants to vision models (Section *'Human-like body representations in vision models'*).
2. **Coarse body representations emerge naturally as a consequence of small network size**. We test 6 image segmentation architectures from the same family of models, varying in size, and discover that human-like coarse representations emerge as a consequence of small network size and limited training compute. We hypothesize that the resource-constrained nature of the human brain similarly favors efficient, coarse-grain representations that balance predictive power with computational efficiency (Section *'Human-like representations as a consequence of resource constraints'*).

This work provides a comparison of object representations in humans and vision models. Such alignment is important both for safety and interpretability in human–robot interaction, and for using machine models as computational probes of the brain's internal representations, offering a scalable route to understanding how humans encode and reason about the physical world.

Related Work

People can reason efficiently about the physical dynamics of everyday objects, though they are also prone to systematic biases and errors under certain conditions (Kubricht, Holyoak, and Lu 2017). This 'intuitive physics' develops early, has a dedicated neural architecture, and is likely shared with non-human animals (Fischer et al. 2016; Spelke and Kinzler 2007; Spelke 2022). There are ongoing debates about the specific format of the representations and computations that support intuitive physics in humans, and proposals over the years have included (among others) first order logic, pre-Newtonian intuitive theories, heuristics and biases, and qualitative reasoning (Hartshorne and Jing 2025).

One prominent proposal for the computation that underlies human intuitive physics posits that people can carry out an internal mental simulation, similar to the computations that support engineered physical engines and game engines (Battaglia, Hamrick, and Tenenbaum 2013; Ullman et al. 2017). This 'mental game engine' proposal has been applied to people's reasoning about collisions, liquids, rigid- and soft-body motion, physical prediction, counterfactual and causal reasoning, and more (Smith et al. 2024). While research into mental simulation is ongoing, even if such a mental simulation exists, it cannot be perfectly accurate. Engineered simulations make heavy use of various approximations and workarounds, and it is likely that people's mental simulation uses approximations as well (Ullman et al. 2017; Balaban and Ullman 2025; Wang and Ullman 2025; Bass et al. 2021).



(a) Humans.

(b) Vision models.

Figure 1: Stimuli vs Body representations (dotted lines) in (A) humans and (B) vision models.

One major approximation in simulated environments is the use of simplified objects for the purposes of physical tracking. To clarify by example: While an advanced game may use fine-grain meshes to graphically display high-resolution images, it will often use only rough bodies for the purposes of collision detection. In line with this, there are theoretical reasons (Ullman et al. 2017) and recent empirical evidence (Li et al. 2023a) to suggest that people also make use of approximate object representations in physical reasoning.

Paralleling the interest in cognitive science, researchers in machine learning have studied the possibility of endowing machines with a sense of intuitive physics. Many datasets and challenges exist in this domain (e.g. Bear et al. 2021; Bakhtin et al. 2019; Yi et al. 2019; Riochet et al. 2021), and several different frameworks have been proposed, ranging from those that emphasize built-in structure (Smith et al. 2019) to those that emphasize learned representations (Piloto et al. 2022; Garrido et al. 2025), with various hybrid proposals in between (Duan et al. 2022).

Despite this interest in machine learning, there hasn't been direct investigation (to the best of our knowledge) of whether the approximate object representations learned by vision models match the approximations used by humans in the context of physical reasoning. Smith et al. (2019) explicitly uses approximate bodies and credits them with successful generalization, but does not compare these to people. Li et al. (2023a) used a model based on α-shapes to examine people's object approximations, but does not explicitly endorse this as a cognitive model, but as a way of teasing apart different degrees of approximation.

Aside from offering glimpses into potential parallels between human and model representations, our work takes a more detailed view of how such representations evolve during training. Rather than focusing on global alignment metrics, we investigate the micro-level dynamics of segmentation learning—how sensitivity to different geometric structures, particularly concavities, changes as models grow or train longer. This approach complements broader forecasting efforts that aggregate model performance as a function of scale or time (Hestness et al. 2017; Kwa et al. 2025; Sevilla et al. 2022), by revealing the finer representational shifts that underlie those macro-level trends

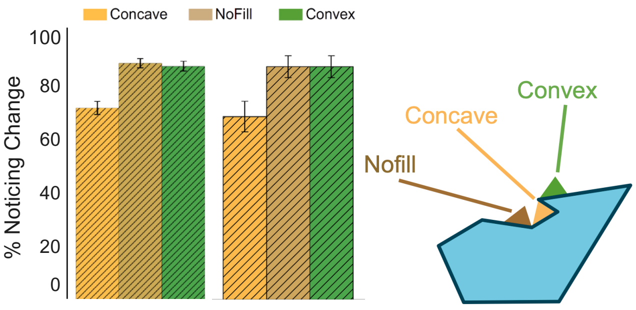
Methods

Datasets and training paradigm

Human–tested dataset. Here we revisit prior work by Li et al. (2023a) that this study builds upon. Each trial performed on a human subject (50 total participants) consists of a pair of black background RGB images of a polygon (see Figure 1): *before* (I_{init}) and *after* (I_{out}). I_{out} is the same image as I_{init}, only with a small segment added (or not) at one of three locations: CONCAVE, NOFILL, or CONVEX (see Figure 2b). Participants are shown I_{init} for 1s, then a blank screen for 2s, then I_{out} for another 1s. The participants are then asked to tell whether the polygon changed, and their accuracy is measured across the three conditions listed. We refer the reader to Li et al. (2023a), Experiment 3b, for a more detailed description of how the human data was collected. In this section, we aim to establish a pipeline for adapting pure segmentation models to the same experimental task.

Model training dataset. To compare model and human representations, we used the experimental stimuli from Li et al. (2023a) for evaluation. Because psychophysical datasets are too small to train deep networks, we generated a larger synthetic dataset designed to approximate their geometric and visual statistics. The goal was not to reproduce the exact human stimuli, but to expose models to similar shape statistics and scene composition. Each training image contains a single uniformly colored polygon on a black background, making our model more accustomed to the data shown to humans. The dataset will be released publicly upon paper acceptance.

Synthetic dataset generation. We developed a procedural polygon generator to produce geometrically diverse yet controlled stimuli. Each polygon is created by sampling: (1) a vertex count uniformly from 5–12; (2) a number of concavities (0–3); and (3) irregularity and spikiness parameters controlling local curvature and edge variance. Polygons are rendered on black backgrounds using one of 24 bright colors sampled from a palette matched to the luminance distribution of the experimental stimuli. The generator thus produces a broad range of shapes that maintain the key structural properties of the human-tested stimuli while preventing any overlap between training and evaluation data. It will be publicly available upon paper acceptance.



(a) Humans vs DETR vision model. (b)

Figure 2: (a) Change detection experiment: Humans (left) vs. Model (right). (b) A small local piece added to one of three locations: Nofill, Concave, and Convex body parts.

Models and fine-tuning

We use publicly available SegFormer models that were pre-trained on the ADE20K dataset (Zhou et al. 2017; Xie et al. 2021): a hierarchical transformer encoder with a lightweight MLP decoder. We tested six sizes (B0–B5) with parameter counts $\sim$3.8M (B0) to $\sim$84.7M (B5), covering over an order of magnitude in capacity. We then fine-tune each variant on our synthetic polygon dataset using custom training settings: AdamW optimizer with learning rate 5×10^{-5}, cosine learning rate schedule with warmup, batch size of 4, weight decay of 0.01, and 15 epochs. The training was performed on NVIDIA A100 GPUs with mixed precision (bfloat16) and TF32 optimizations enabled. The models are trained with a combination of cross-entropy and Dice loss to segment the binary segmentation task. Additionally, we tested a publicly available version of the DETR model (Carion et al. 2020) for a total of 7 architectures.

From model masks to change detection

Mask extraction. For each image $I \in \{I_{\text{init}}, I_{\text{out}}\}$ the model outputs logits $\mathbf{L}(I)$. We apply softmax and argmax to get class predictions, then extract the foreground class (1: object class; 0: background class):

$$\mathbf{M}^{(\text{pred})}(I) = \mathbf{1}[\text{argmax}(\text{softmax}(\mathbf{L}(I))) = 1]$$

The object area is computed as the sum of object class pixels:

$$A_{\text{init}} = \sum_{i,j} \mathbf{M}^{(\text{pred})}(I_{\text{init}})_{i,j}, \quad A_{\text{out}} = \sum_{i,j} \mathbf{M}^{(\text{pred})}(I_{\text{out}})_{i,j}$$

Relative Area Change (segment–normalized). Let $A_{\text{seg}}^{(\text{gt})}$ be the ground–truth pixel area of the *edited segment* (the small local piece added/removed between I_{init} and I_{out}, see Figure 2). We define the *Relative Area Change (RAC)*:

$$\text{RAC}_{\text{seg}} = \frac{A_{\text{out}} - A_{\text{init}}}{A_{\text{seg}}^{(\text{gt})}}.$$

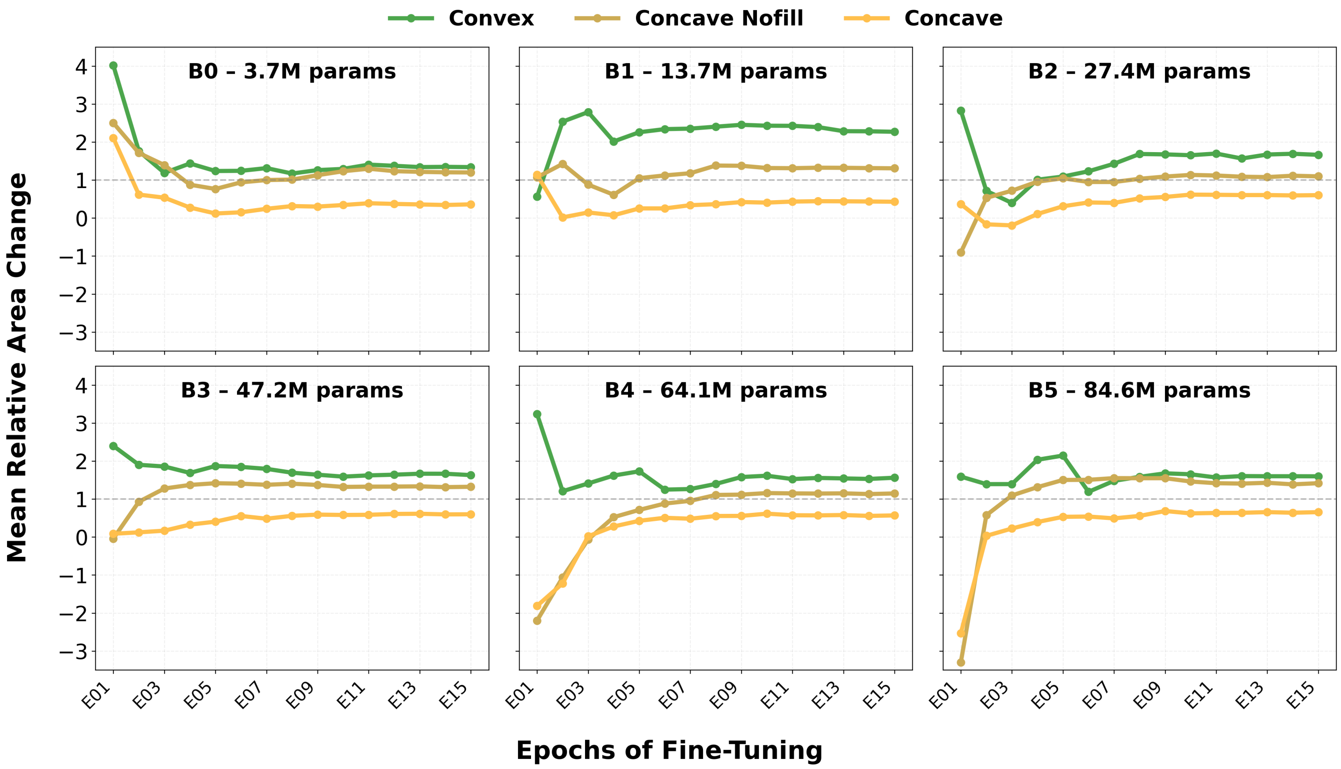


Figure 3: Mean $\mathrm{RAC}_{\mathrm{seg}}$ during fine-tuning per category across models.

$\mathrm{RAC}_{\mathrm{seg}} > 0$ for additions (mask grows), $\mathrm{RAC}_{\mathrm{seg}} < 0$ for removals (mask shrinks), and $\mathrm{RAC}_{\mathrm{seg}}$ measures how strongly the mask responds *locally* at the manipulated region. A simple way to think about this is the local *resolution* or *sensitivity* of a model. A model that perfectly segments the ground truth will always have $\mathrm{RAC}_{\mathrm{seg}} = 1$, while a model incapable of perceiving a given local change will always have $\mathrm{RAC}_{\mathrm{seg}} \approx 0$.

Change/no–change decision. We label a pair $\{I_{\mathrm{init}}, I_{\mathrm{out}}\}$ DETECTED iff $\mathrm{RAC}_{\mathrm{seg}} > \tau$, with $\tau \in \{0.1, 0.2, 0.3, 0.4, 0.5, ..., 20\}\%$. Each τ allows us to derive a change/no-change decision for each condition: CONCAVE, NOFILL, CONVEX.

Human-like representations in vision models

We begin this section by stating an important takeaway from the above work by Li et al. (2023b); people's representation of concave body parts appears more coarse than their representation of convex body parts. That is, people tend to ”fill in” or ”diffuse” concavities in the context of physical reasoning, which doesn't appear to be the case with convex body parts (see Figure 1). In the context of the change/no-change experiment discussed above, this is particularly evident in Figure 2a (left). We see that the same holds for image segmentation models following the pipeline described above in Figure 2a (right). This result is dependent on $\tau = 1\%$, which is the optimal threshold value that minimizes the RMSE with human data, but given that we apply *the same* τ to concave, nofill, and convex changes, we are simply showing that the model perceives changes in concavities as smaller, analogous to the more coarse representations formed by humans.

Probability maps. We refer the reader to Figure 4. On the *top row*, segmented masks show clear outward diffusion around concavities, while corners and convex edges remain stable. The same pattern appears in the *bottom row* of probability maps, where activations spread beyond concave boundaries. This consistent “filling-in” effect suggests that models, like humans, simplify concave regions into smoother, coarser body representations.

Training dynamics. As an additional test, we measured the *average* $\mathrm{RAC}_{\mathrm{seg}}$ on the test set for each change type throughout fine-tuning (Figure 3). Concave changes consistently remain at lower $\mathrm{RAC}_{\mathrm{seg}}$ values (y-axis), even with extended training, indicating that models remain less responsive to changes inside concavities. We hypothesize that this reflects a representational bias toward compact, convex-like encodings. Representing objects through their convex hulls—or approximations close to it—requires fewer vertices and less spatial detail (Duan and Lafarge 2015), reducing both memory load and computational cost. Moreover, such coarse representations generalize more effectively across object categories by capturing global shape structure

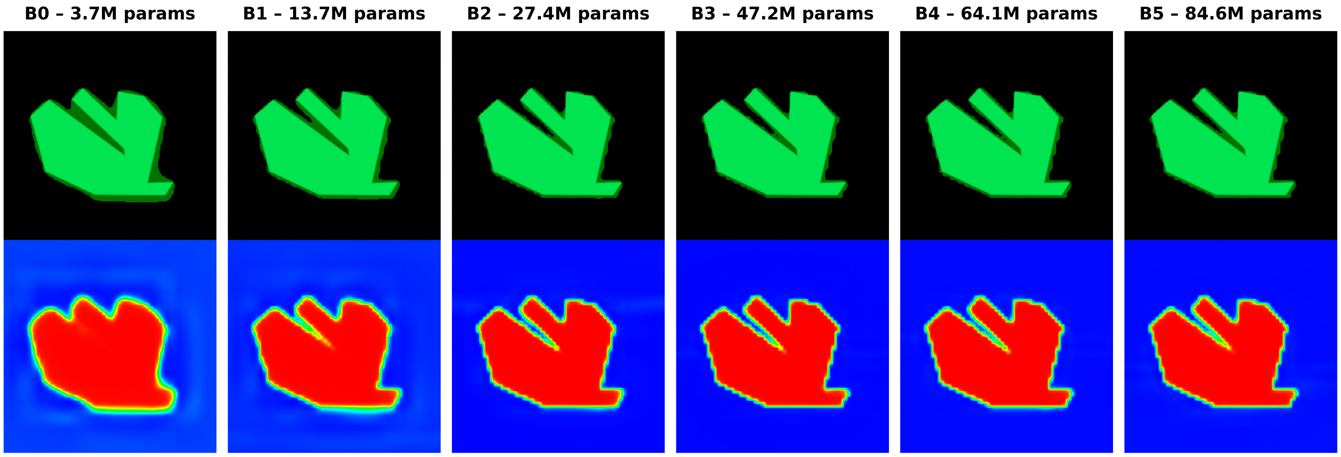


Figure 4: Mask overlays (first row) and probability heatmaps after 10 epochs of training across models.

rather than local irregularities. This consistent pattern between humans and vision models may hint at the underlying mechanisms and specific geometric forms that approximate body representations take in the human brain.

Human-like representations as a consequence of resource constraints

Probability maps. We refer the reader to Figure 4. Notably, the "filling-in" effect discussed above is much stronger in smaller models, whereas larger models preserve sharper boundaries and show reduced diffusion in concavities.

Training dynamics. Similarly, we measured the *average* $\mathrm{RAC}_{\mathrm{seg}}$ for each change type throughout the fine-tuning (Figure 3). The gap between concave and convex changes diminishes both with increased training compute and with larger model size. As models grow and train longer, they become more sensitive to local geometric variations and less reliant on coarse, convex approximations. This pattern supports our hypothesis that human-like body representations emerge partly as a consequence of resource constraints: smaller or less-trained models favor compact, efficient representations that smooth concavities, while larger, more capable systems can afford finer geometric detail.

Together, these findings suggest that human-like coarse body representations may reflect an efficient encoding strategy that emerges naturally when computational or biological systems must balance representational detail with resource efficiency.

Discussion

People reason about the physical world using approximate rather than exact representations of objects. These coarse body representations are efficient for predicting how objects move or interact, without encoding every geometric detail. In this work, we asked whether such approximations naturally emerge in vision models trained for segmentation.

Across experiments, we found consistent parallels between human and model behavior. Both humans and models tend to simplify concavities, effectively "filling in" missing regions and producing smoother object representations. Quantitatively, models show lower sensitivity to local changes inside concavities, similar to the coarse representations observed in people's reasoning about intuitive physics.

Importantly, these effects vary with model capacity and training compute. Smaller networks and shorter training produce stronger concavity-smoothing effects, while larger or more extensively trained models develop sharper, more fine-grained boundaries. The same effect is *not* observed for convex object parts, which are more stable and detailed. This pattern supports our hypothesis that human-like coarse body representations can emerge as an efficient solution under resource constraints: when capacity is limited, both biological and artificial systems favor compact, convex-like encodings that balance accuracy with computational efficiency.

Understanding this tradeoff may help clarify how efficient object representations arise in both minds and machines. While precise segmentations are essential for many applications, there is also value in representations that are "not too fine, not too coarse" — the kind that support intuitive, generalizable physical reasoning. This balance, much like in the tale of Goldilocks, may reflect the sweet spot of perception.

References

Baillargeon, R. 2004. Infants' reasoning about hidden objects: Evidence for event-general and event-specific expectations. *Developmental Science*, 7(4): 391–414.

Bakhtin, A.; van der Maaten, L.; Johnson, J.; Gustafson, L.; and Girshick, R. 2019. Phyre: A new benchmark for physical reasoning. *Advances in Neural Information Processing Systems*, 32.

Balaban, H.; and Ullman, T. D. 2025. Physics versus graphics as an organizing dichotomy in cognition. *Trends in Cognitive Sciences*.

Bass, I.; Smith, K. A.; Bonawitz, E.; and Ullman, T. D. 2021. Partial mental simulation explains fallacies in physical reasoning. *Cognitive Neuropsychology*, 38(7-8): 413–424.

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

Bear, D. M.; Wang, E.; Mrowca, D.; Binder, F. J.; Tung, H.-Y. F.; Pramod, R.; Holdaway, C.; Tao, S.; Smith, K.; Sun, F.-Y.; et al. 2021. Physion: Evaluating physical prediction from vision in humans and machines. *arXiv preprint arXiv:2106.08261*.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229.

Duan, J.; Dasgupta, A.; Fischer, J.; and Tan, C. 2022. A survey on machine learning approaches for modelling intuitive physics. *arXiv preprint arXiv:2202.06481*.

Duan, L.; and Lafarge, F. 2015. Image Partitioning Into Convex Polygons. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Fischer, J.; Mikhael, J. G.; Tenenbaum, J. B.; and Kanwisher, N. 2016. Functional neuroanatomy of intuitive physical inference. *Proceedings of the national academy of sciences*, 113(34): E5072–E5081.

Freyd, J. J. 1987. Dynamic mental representations. *Psychological review*, 94(4): 427.

Garrido, Q.; Ballas, N.; Assran, M.; Bardes, A.; Najman, L.; Rabbat, M.; Dupoux, E.; and LeCun, Y. 2025. Intuitive physics understanding emerges from self-supervised pretraining on natural videos. *arXiv preprint arXiv:2502.11831*.

Gizdov, A.; Ullman, S.; and Harari, D. 2025. Seeing More with Less: Human-like Representations in Vision Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4408–4417.

Hartshorne, J. K.; and Jing, M. 2025. Insights into cognitive mechanics from education, developmental psychology and cognitive science. *Nature Reviews Psychology*, 1–15.

Hestness, J.; Narang, S.; Ardalani, N.; Diamos, G.; Jun, H.; Kianinejad, H.; Patwary, M. M. A.; Yang, Y.; and Zhou, Y. 2017. Deep Learning Scaling is Predictable, Empirically.

Kubricht, J. R.; Holyoak, K. J.; and Lu, H. 2017. Intuitive physics: Current research and controversies. *Trends in cognitive sciences*, 21(10): 749–759.

Kwa, T.; West, B.; Becker, J.; Deng, A.; Garcia, K.; Hasin, M.; Jawhar, S.; Kinniment, M.; Rush, N.; Arx, S. V.; Bloom, R.; Broadley, T.; Du, H.; Goodrich, B.; Jurkovic, N.; Miles, L. H.; Nix, S.; Lin, T.; Parikh, N.; Rein, D.; Sato, L. J. K.; Wijk, H.; Ziegler, D. M.; Barnes, E.; and Chan, L. 2025. Measuring AI Ability to Complete Long Tasks.

Li, Y.; Wang, Y.; Boger, T.; Smith, K. A.; Gershman, S. J.; and Ullman, T. D. 2023a. An approximate representation of objects underlies physical reasoning. *Journal of Experimental Psychology: General*, 152(11): 3074–3086.

Li, Y.; Wang, Y.; Boger, T.; Smith, K. A.; Gershman, S. J.; and Ullman, T. D. 2023b. An Approximate Representation of Objects Underlies Physical Reasoning. *Journal of Experimental Psychology: General*.

Marr, D. 2010. *Vision: A computational investigation into the human representation and processing of visual information*. MIT press.

Piloto, L. S.; Weinstein, A.; Battaglia, P.; and Botvinick, M. 2022. Intuitive physics learning in a deep-learning model inspired by developmental psychology. *Nature human behaviour*, 6(9): 1257–1267.

Riochet, R.; Castro, M. Y.; Bernard, M.; Lerer, A.; Fergus, R.; Izard, V.; and Dupoux, E. 2021. Intphys 2019: A benchmark for visual intuitive physics understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9): 5016–5025.

Sevilla, J.; Heim, L.; Ho, A.; Besiroglu, T.; Hobbhahn, M.; and Villalobos, P. 2022. Compute Trends Across Three Eras of Machine Learning. In *2022 International Joint Conference on Neural Networks (IJCNN)*.

Smith, K.; Mei, L.; Yao, S.; Wu, J.; Spelke, E.; Tenenbaum, J.; and Ullman, T. 2019. Modeling expectation violation in intuitive physics with coarse probabilistic object representations. *Advances in neural information processing systems*, 32.

Smith, K. A.; Hamrick, J. B.; Sanborn, A. N.; Battaglia, P. W.; Gerstenberg, T.; Ullman, T. D.; and Tenenbaum, J. B. 2024. Probabilistic models of physical reasoning. *Reverse engineering the mind: Probabilistic models of cognition*.

Spelke, E. 2022. *What babies know: Core knowledge and composition volume 1*, volume 1. Oxford University Press.

Spelke, E. S. 1990. Principles of object perception. *Cognitive Science*, 14(1): 29–56.

Spelke, E. S.; and Kinzler, K. D. 2007. Core knowledge. *Developmental science*, 10(1): 89–96.

Ullman, T. D.; Spelke, E.; Battaglia, P.; and Tenenbaum, J. B. 2017. Mind games: Game engines as an architecture for intuitive physics. *Trends in cognitive sciences*, 21(9): 649–665.

Wang, Y.; and Ullman, T. D. 2025. Resource bounds on mental simulations: Evidence from a liquid-reasoning task. *Journal of Experimental Psychology: General*.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Yi, K.; Gan, C.; Li, Y.; Kohli, P.; Wu, J.; Torralba, A.; and Tenenbaum, J. B. 2019. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*.

Zhou, B.; Zhao, H.; Puig, X.; Fidler, S.; Barriuso, A.; and Torralba, A. 2017. Scene Parsing through ADE20K Dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.