MODELING THE OBJECT REPRESENTATIONS UNDERLYING HUMAN PHYSICAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Humans appear to represent objects when reasoning about physics with coarse, volumetric "bodies" that smooth concavities, trading fine visual detail for efficient physical predictions. Yet, the structure of these representations remains largely unknown. Segmentation models, in contrast, are trained for pixel-accurate masks that may misalign with such bodies. We ask whether and when these models nonetheless acquire human-like object representations. Using a time-to-collision (TTC) and change detection (CD) behavioral task with data from 178 and 50 human participants, respectively, we introduce a pipeline and an alignment metric to compare the visual representations of segmentation models to those of humans. We do this systematically on multiple architectures (DINOv2, SegFormer, DeepLabV3+, and UPerNet), varying their size and training time. We find that briefly trained models segment objects too coarsely, aligning poorly with humans, while fully trained models segment objects too finely. For each model, there is an intermediate training regime that best matches the coarse bodies observed in human behaviour, and larger models tend to reach it earlier. We show these bodies emerge under resource constraints in general-purpose vision models, providing computational support to resource-rational accounts of human cognition. This work provides a foundational framework for testing alignment between vision models and humans and shows there is a growing gap between the state-of-the-art in artificial intelligence and human cognition, driven by scaling model size and training.

1 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 computations and representations that are also coupled but separate. 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 fine-grain object segmentation, but for the purposes of estimating *how* a physical scene will unfold (which objects will collide, and where they will end up), it may be sufficient and more cost-effective to have more coarse object-representations (e.g. if all that matters is catching or ducking them up, a water bottle and thermos can be roughly approximated by a cylinder).

The separation between coarse and fine-grain object representations for different purposes is widely used in engineering, especially in simulated environments such as games and animations. Computational models of intuitive physics in humans suggest people may have a similar split between physics and graphics (Ullman et al., 2017; Balaban & Ullman, 2025), with fine-grain object *shapes* used for rendering and recognition and coarse-grain *body* representations used for physical prediction and action (see Figure 1A). This suggestion is in line with studies that show 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). Recent psychophysical studies with adults also suggest that humans use coarse body approximations (Li et al., 2023) (see Figure 1B). In particular, there's evidence that people's body approximations tend to "fill in" concave object regions (Section 2 discusses this in detail). From a neuro-scientific perspective, the physics/graphics split may map onto the neural division between the dorsal and ventral streams in human vision (Balaban & Ullman, 2025).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

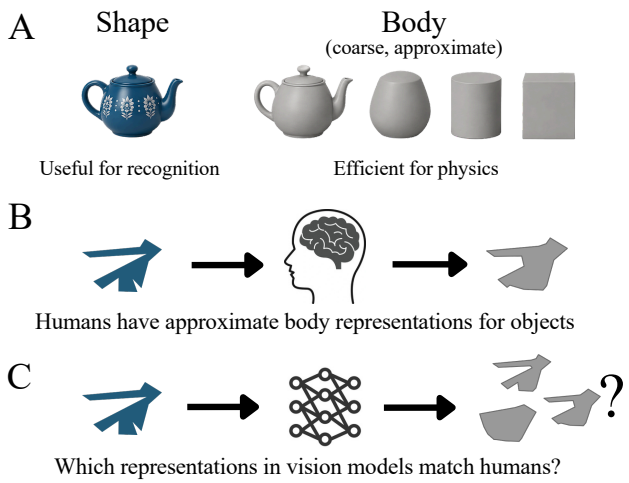


Figure 1: **Research overview. (A)** Different object representations are useful for different goals. "Body" representation is useful for physical reasoning. "Shape" representation is useful for recognition. **(B)** Prior work (Li et al., 2023) has shown evidence that humans do have such coarse body representations for physical reasoning (e.g. predicting collision times), specifically by "filling in" concave regions. **(C)** Our research asks *whether* vision segmentation models have such similar representations and under *what* conditions they develop them.

While evidence suggests humans rely on approximate representations of objects when reasoning about physics, the nature and structure of these representations in the human brain remain largely unknown. Behavioral methods offer indirect, low-resolution glimpses into these internal encodings, and do not reveal their geometric or computational form. Artificial segmentation models, which are trained to perform a similar decomposition of scenes into discrete entities, offer a computational setting for exploring how representations with similar behavioral consequences might form. If modern vision models trained for segmentation or prediction share representational structure with humans, they could provide a scalable way of testing candidate object approximations against human physical judgments.

Unlike the study of alignment in LLMs, there has been relatively little exploration of the alignment between the body representations humans use when reasoning about physics and the object representations models use for vision, which are often used as a basis for downstream physics tasks in robotics and embodied agents, where the potential benefits of such alignment can be important. Given that (i) coarse object bodies are useful for physical simulations in engineered systems, (ii) humans appear to rely on coarse bodies during physical reasoning, and (iii) segmentation networks can learn such abstractions, we ask the following question (Figure 1C): *Do segmentation models form object representations whose behavioral consequences match human physical judgments, and when?*

Contributions. In this study, we create a pipeline for such comparison and explicitly control three key factors: model architecture, size, and training time. Our contributions are as follows:

1. **A general pipeline for comparing segmentation models with human physical judgments.** We operationalize the outputs of segmentation networks and introduce a unified Time-To-Collision (TTC) and Change Detection (CD) evaluation framework using data from 178 and 50 human participants respectively (Li et al., 2023). This framework enables direct, quantitative comparison between vision model predictions and human behavior in physical reasoning tasks (see Section 3).
2. **We find an intermediate training regime whose timing varies with model size.** Across four architecture families (16 model configurations), error with respect to human judgments is non-monotonic over training: briefly trained models produce overly coarse masks, while continued training resolves increasingly fine boundary detail. Held-out participant analysis confirms that error is minimized between these endpoints, and different model sizes reach this regime at different times (see Section 4.1).
3. **We show that segmentation models develop coarse object approximations whose behavioral consequences resemble human concavity-filling effects.**

   Across models, agreement with human judgments is greatest at intermediate training stages, whose timing varies with model size. Fitted geometric operators applied to the ground-truth masks reproduce the human effect equally well (Section 4.1); the contribution of the models

is therefore not a closer fit but an account of how such geometry arises in a general-purpose learner. This motivates the hypothesis that resource constraints may contribute to similar coarse encodings in biology (see Section 5). We characterize the geometry of these model representations and when they arise during training (see Section 4).

Our results provide practical and theoretical insight. Practically, understanding when models form object approximations that align with human judgments may inform future work on vision systems that interact safely and predictably with humans, particularly as we start deploying robots and embodied agents in the physical world. Theoretically, we identify computational conditions under which human-like body representations emerge, offering a computational and visual account of how vision models can produce geometry consistent with human physical judgments.

2 Related Work

People reason efficiently about everyday physical dynamics, though they show systematic biases under certain conditions (Kubricht et al., 2017; Hartshorne & Jing, 2025). This "intuitive physics" emerges early (Spelke & Kinzler, 2007; Baillargeon, 1998), is supported by dedicated neural systems (Fischer et al., 2016; Pramod et al., 2022; 2025; Balaban et al., 2024), and is likely shared with non-human animals (Wood et al., 2024; Spelke, 2022).

There are different competing formal accounts of the mental computations that support intuitive physics. One currently prominent view suggests that people mentally simulate physical events, somewhat like the game engines or physics engine used to simulate fictional worlds in computer games and animation (Hegarty, 2004; Battaglia et al., 2013; Ullman et al., 2017). The mental-game-engine proposal has been used to investigate and explain people's reasoning about a variety of physical phenomena, including collisions, stability, tool use, liquids, rigid- and soft-body motion, physical prediction, and physical counterfactual and causal reasoning (Smith et al., 2024; Smith & Vul, 2013; Bramley et al., 2018; Sanborn et al., 2013; Hamrick et al., 2016; Bates et al., 2019; Gerstenberg & Stephan, 2021; Sosa et al., 2021; Battaglia et al., 2013).

Prior to, and proceeding alongside the mental simulation account, a variety of non-simulation accounts of intuitive physics have also been proposed. These include, but are not limited to, the use of heuristics, logical rules, bottom-up visual features, qualitative reasoning, and pre-Newtonian theory-like abstractions (Lerer et al., 2016; Baillargeon, 2002; Fragkiadaki et al., 2016; Mottaghi et al., 2016; Proffitt et al., 1990; Nusseck et al., 2007; Gilden & Proffitt, 1989; Siegler & Chen, 1998; Davis, 1990; Forbus, 1988; Piloto et al., 2022; McCloskey & Kohl, 1983). Some of these accounts point to people's errors as evidence that humans cannot be running a perfect simulation, as such a simulation would be too costly and not make errors. However, if such a mental simulation exists in the mind, it cannot be perfectly accurate. Just as engineered simulations make heavy use of various approximations and workarounds, it is likely that people's mental simulation uses approximations as well (Ullman et al., 2017; Balaban & Ullman, 2025; Wang & Ullman, 2025; Bass et al., 2021). Recent evidence further suggests that simulation and abstraction need not be competing accounts: people may blend them depending on the state and task (Sosa et al., 2025).

One major approximation in simulated environments is the use of simplified objects for the purposes of physical tracking and collision handling. While an advanced game engine may use fine-grain meshes to graphically render high-resolution images, it will often use only rough bodies for the purposes of collision detection and physics simulation. Converging theoretical and empirical work suggests that humans similarly use approximate "body" representations in physical reasoning (Ullman et al., 2017; Li et al., 2023) – volumetric approximations that prioritize properties relevant to action and physical interaction (e.g., mass, position, and rough form) while discarding fine-grained visual details (Figure 1 A, B). This idea is consistent with resource-rational accounts of cognition (Lieder & Griffiths, 2020), and the "what/where–how" functional split in visual processing (Ungerleider & Mishkin, 1982; Goodale & Milner, 1992).

In this paper, we focus on the Time-To-Collision and Change Detection experiments of Li et al. (2023, Experiments 2 and 3b) (Figure 2A). The first builds on a well-established paradigm for time-to-collision judgments with simple bodies (Rosenbaum, 1975; Tresilian, 1995; Gray & Thornton, 2001). A key behavioral signature is an asymmetry between *concave* and *convex* stimuli: different from what a random noise model would predict, human judgments seem to selectively smooth or "fill in"

concavities on objects (over-approximating inlets and gaps), predicting earlier collision for concave than convex pairings. Similarly, in the Change Detection experiment, participants are significantly worse at noticing local geometric additions when they occur inside an object's concavities compared to matched changes on convex protrusions, further supporting the idea that the human visual system mentally smooths over these internal gaps. Preliminary analyses suggested that the operative body representation is between the exact outline and the convex hull.

Segmentation networks provide a controlled computational probe of object form because their outputs expose explicit spatial masks. Yet most of these models are trained for pixel-accurate boundaries, favoring fine detail that may be misaligned with human physical prediction. Prior work has leveraged coarse bodies for modeling human expectations without direct comparison to people (Smith et al., 2019), or has used α-shape–based models to analyze human approximations without endorsing them as cognitive mechanisms (Li et al., 2023), but as a way of teasing apart different degrees of approximation.

Our work also relates to studies of how alignment with humans changes with model scale and training. In vision, more accurate object-recognition networks use visual strategies less aligned with those of humans (Fel et al., 2022) and predict neural responses in inferotemporal cortex less well (Linsley et al., 2023).In language, larger models fit human reading times worse (Oh & Schuler, 2023b), and fit is best after about two billion training tokens (Oh & Schuler, 2023a). Neural networks are also known to learn coarse, low-frequency structure before fine detail (Rahaman et al., 2019). We study what this means for object shape: which geometry segmentation models pass through as they grow or train longer, and when it matches human physical judgments.

3 METHODS

Here we outline how we compare vision models with human behavior in two psychophysical tasks: a time-to-collision (TTC) task and a change-detection (CD) task.

Human data. We use the publicly available data of Li et al. (2023) from two experiments. In the *TTC experiment*, 178 participants watched 96 test videos in which one polygon moves toward a stationary one. At a known time the stationary polygon disappears, and participants press a key when they think the two have collided (Figure 2A). The videos form 48 matched pairs (8 shapes $\times$ 6 ground-truth collision times). The two videos of a pair share the same kinematics and differ only in whether a concave or a convex side of the stationary polygon faces the collision. Humans respond earlier in the concave condition, consistent with "filling in" concavities. In the *CD experiment*, 50 participants saw a polygon I_{init} for 1 s, a blank screen for 2 s, and a second polygon I_{out} for 1 s, and reported whether it had changed. I_{out} adds a small triangle of equal area either inside a concavity or on a convex part (Appendix A). Humans detect convex changes more often than concave ones. We ask whether segmentation models reproduce both effects, and at which model sizes and training stages (Figure 2C).

Architectures. We evaluate four segmentation families (16 configurations): **SegFormer** (Xie et al., 2021) B0–B5 (3.7–84.6M parameters), **DeepLabV3+** (Chen et al., 2018) with ResNet (He et al., 2016) 18/50/101 backbones (12.3–60.4M), **UPerNet** (Xiao et al., 2018) with ConvNeXt (Liu et al., 2022) Small/Base/Large/XLarge backbones (82.0–391.0M), and **DINOv2** (Oquab et al., 2024) Small/Base/Large (27.2–310.6M) with a light convolutional decoder. SegFormer and UPerNet are initialized from checkpoints trained for semantic segmentation on ADE20K (Zhou et al., 2017), DeepLabV3+ from Cityscapes (Cordts et al., 2016), and DINOv2 from its self-supervised weights. These families cover convolutional (DeepLabV3+, UPerNet) and transformer (SegFormer, DINOv2) backbones, supervised and self-supervised pretraining, and each comes in several sizes, which lets us vary model size within an architecture. All models are fine-tuned end to end on our dataset, with three seeds per configuration.

Fine-tuning. The polygons of Li et al. (2023) are irregular shapes chosen for their low verbal association (Cooper, 1975), which limits the influence of familiarity with specific objects. The same holds for our segmentation models, which are pretrained on natural images and do not reliably segment these novel polygonal stimuli out of the box. We therefore fine-tune each model on a synthetic polygon dataset that matches the geometric and color statistics of the human stimuli but shares no shapes with them (Appendix B.2). All models are trained for 945 steps with the same

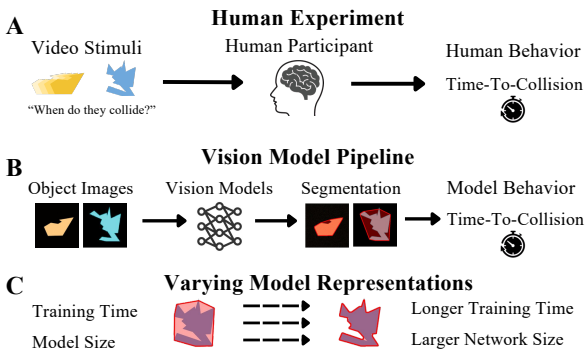


Figure 2: **Methods overview. (A)** In Li et al. (2023), participants predicted when a moving polygon would hit a stationary one that was invisible during motion. At the same true collision time, they responded earlier to hits into a concavity than onto a convex side, as if concavities were filled in. **(B)** Fine-tuned models segment the frame before motion; we move the masks at the true velocities and record when they overlap. **(C)** Training time and model size change how coarse the masks are.

recipe (Appendix B.3): AdamW (Loshchilov & Hutter, 2019) with learning rate $5 \times 10^{-5}$and a cross-entropy plus Dice loss (Milletari et al., 2016). The human stimuli are used only for evaluation. We release the dataset and its generator with our code.

3.1 Time-To-Collision

Once trained, we test the models on the same videos shown to humans (Figure 2B). Before motion starts, both objects are stationary and fully visible. We pass this initial frame $I \in \mathbb{R}^{H\times W\times 3}$ through the model, which outputs per-pixel logits $\mathbf{L}(I) \in \mathbb{R}^{H\times W\times 2}$, and take the per-pixel argmax to obtain a foreground mask:

$$\mathbf{M}(I) = \mathbf{1}\left[\arg\max_{c\in\{0,1\}} \mathbf{L}(I)_c = 1\right] \in \{0,1\}^{H\times W}. \tag{1}$$

Given that each image contains two objects, we extract the two largest connected components of $\mathbf{M}(I)$ and denote them $\mathbf{M}_1, \mathbf{M}_2 \in \{0,1\}^{H\times W}$. Those are the two polygons.

From segmentation to video kinematics. Our goal is to isolate the role of object shape in TTC prediction. We therefore provide the model with the true object velocities $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}^2$ (pixels per frame) and the frame rate Δf (frames per second) from the stimulus metadata, rather than asking it to estimate motion. $\boldsymbol{v}_2 = 0$ naturally, since one object is stationary, but we define it for notation consistency. This makes the model's TTC prediction depend only on its segmentation of the initial frame, so differences in model TTC are attributable to the geometry of the predicted masks, rather than to motion estimation.

We define a translation operator $\mathrm{Translate}(\mathbf{M}, \boldsymbol{u})$ that shifts a binary mask $\mathbf{M}$ by displacement $\boldsymbol{u}$. We simulate the future motion of the two segmented objects using only their initial masks and the true velocities: at discrete frame index $n \in \mathbb{N}$, $\mathbf{M}_i^{(n)} = \mathrm{Translate}(\mathbf{M}_i, n\boldsymbol{v}_i)$ for $i \in \{1,2\}$. The model-predicted time-to-collision is the earliest time $t = n/\Delta f$ at which the translated masks first overlap:

$$\mathrm{TTC}_{\mathrm{model}} = \min_{n\geq 0} \left\{ t = \tfrac{n}{\Delta f} \;\middle|\; \mathbf{M}_1^{(n)} \cap \mathbf{M}_2^{(n)} \neq \emptyset \right\}. \tag{2}$$

The model thus receives the same static pre-motion frame as the human observer and uses only its inferred object masks, together with the true kinematics, to produce a TTC prediction.

Comparison to human behavior. The human dataset contains matched video pairs that share the same kinematics and ground-truth collision time $\mathrm{TTC}_{\mathrm{gt}}$, and differ only in whether a concave or a convex side of one object faces the collision. For each video j and participant p, we observe a human response time $\mathrm{TTC}_{\mathrm{human}}^{(j,p)}$, defined as the time between motion onset and button press, and obtain a per-video human TTC by averaging across participants: $\mathrm{TTC}_{\mathrm{human}}^{(j)} = \frac{1}{P}\sum_{p=1}^{P} \mathrm{TTC}_{\mathrm{human}}^{(j,p)}$. Indexing videos by ground-truth collision time τ and shape condition $g \in \{\mathrm{concave}, \mathrm{convex}\}$, and letting $\mathcal{N}(\tau, g)$ denote the corresponding video set, we define the condition-averaged TTC for humans and models:

$$\mathrm{TTC}_x(\tau, g) = \frac{1}{|\mathcal{N}(\tau, g)|} \sum_{j\in\mathcal{N}(\tau,g)} \mathrm{TTC}_x^{(j)}, \qquad x \in \{\mathrm{human}, \mathrm{model}\}. \tag{3}$$

The *concavity effect* $\Delta_x(\tau)$ is the difference in TTC between the convex and concave conditions at a given ground-truth time τ:

$$\Delta_x(\tau) = \mathrm{TTC}_x(\tau, \text{convex}) - \mathrm{TTC}_x(\tau, \text{concave}). \tag{4}$$

A positive concavity effect indicates earlier predicted or reported collisions for concave than convex configurations; a negative effect indicates the reverse. For each training seed s at a given checkpoint, we define the signed discrepancy from the human concavity effect as

$$E_s = \frac{1}{|\mathcal{T}|} \sum_{\tau \in \mathcal{T}} \left[\Delta_{\text{model},s}(\tau) - \Delta_{\text{human}}(\tau) \right], \tag{5}$$

where $\mathcal{T}$ is the set of six ground-truth collision times and s indexes independently trained models (seeds). $\Delta_{\text{human}}(\tau)$ is computed from all 178 participants, except in the cross-validation of Section 4.1, where it is computed only from the participants in the relevant fold. Averaged over participants, the human effect is 31.17 ms (95% CI $[22.74, 39.60]$, t-interval over the 178 per-participant convex−concave differences). Positive E_s means a larger mean concavity effect than humans, negative E_s a smaller one; agreement occurs at $E_s = 0$. We therefore use $|E_s|$ as the alignment error of seed s, and report its mean over seeds, $\bar{E}$ (both in ms). Because the absolute value is taken after averaging over collision times, a low $\bar{E}$ means agreement in the *mean* concavity effect, not at every collision time, and does not imply that model and human representations are identical. Section 4.1 reports $\bar{E}$ across training, with the best checkpoint selected on a random half of the participants and evaluated on the other half (200 splits, both directions); Appendix D.3 reports the signed effect Δ.

3.2 Change Detection

Each trial provides two images: an initial polygon I_{init} and a second polygon I_{out} that differs from it by a small segment added to the boundary, either inside a concavity or on a convex protrusion (16 stimuli per condition; examples in Appendix A). We pass each image through the segmentation model as in Section 3.1, and denote by A_{init} and A_{out} the areas of the two predicted masks. Whereas the TTC task measures shape representation indirectly through simplified physical simulation, this task measures it directly through the model's mask sensitivity to local shape changes.

Relative Area Change (RAC). Let A_{gt} be the ground-truth pixel area of the added segment. We define the Relative Area Change as:

$$\mathrm{RAC} = \frac{A_{\text{out}} - A_{\text{init}}}{A_{\text{gt}}}. \tag{6}$$

RAC measures the *sensitivity* of a model to a change in its shape: a model that perfectly segments the ground truth has $\mathrm{RAC} = 1$, a model that misses the change has $\mathrm{RAC} \approx 0$, and $\mathrm{RAC} > 1$ means the mask grows more than the added area.

Comparison to human behavior. Humans report a binary change/no-change decision, so the metric is a detection rate (DR). To compare with humans, we label a change as detected iff the relative change in mask area, $\delta = |A_{\text{out}} - A_{\text{init}}|/A_{\text{init}}$, exceeds a threshold τ, and report $\mathrm{DR}_{\text{concave}}$ and $\mathrm{DR}_{\text{convex}}$. We fit τ once on a reference model (SegFormer-B5 at its minimum-$\bar{E}$ checkpoint, selected on the TTC task, so no CD data are used for checkpoint selection), by minimizing the RMSE to the human detection rates over $\tau \in \{0.1, 1.1, \ldots, 20.1\}\%$, which gives $\tau = 1.1\%$. We then apply it unchanged to every other configuration, which are therefore out of sample with respect to threshold selection. The convex–concave asymmetry is stable across the full range of τ (Appendix F.2).

4 Results

Section 4.1 reports the TTC results: how alignment with humans changes over training and with model size, how the models compare with fitted geometric operators, and the mask geometry behind these effects. Section 4.2 tests the same geometry on change detection.

4.1 Time-to-collision

Alignment peaks at intermediate training. In all four families, alignment with the human TTC effect is highest at an intermediate stage of training. $\bar{E}$ is high early in training, reaches a minimum

mid-training, and rises again (Figure 3). Early masks fill concavities too much and give a concavity effect larger than the human one. Late masks resolve concavities and give a smaller one. Appendix D.3 plots this signed effect across training against the human value.

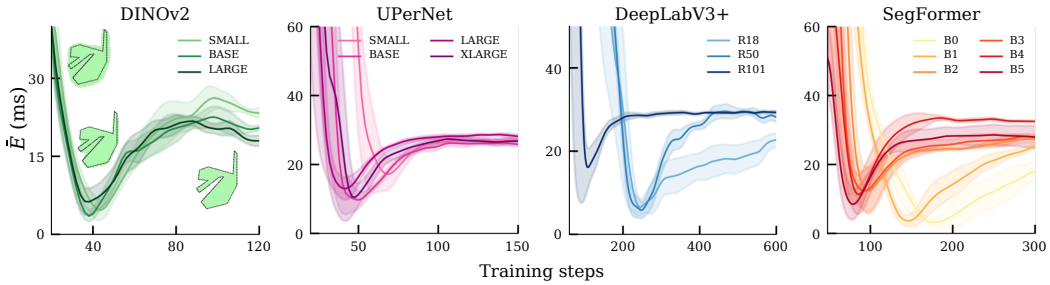


Figure 3: **Alignment with the human TTC effect peaks at intermediate training.** $\bar{E}$ (ms) across training, averaged over seeds; bands show ±SEM, and darker lines are larger models. $\bar{E}$ stays above zero because the absolute value is taken per seed before averaging (Appendix D.3).

The model effect passes through the human value. On the way from coarse to fine, the model concavity effect passes through the human value in all 16 configurations, with the seed-averaged effect lying inside the human 95% CI for a window of checkpoints.

The peak holds on held-out participants. With the checkpoint selected on the other half of the participants, the best-aligned configurations reach $\bar{E}$ of 5–8 ms. Testing all 16 configurations together, the selected checkpoint has lower $\bar{E}$ than both the first checkpoint ($t(15) = -5.17$, $p = 5.7 \times 10^{-5}$) and the last one ($t(15) = -19.6$, $p = 2.1 \times 10^{-12}$), so the trajectory is non-monotonic out of sample.

Pretraining is not required. DINOv2 trained from random initialization reaches the same held-out $\bar{E}$ as the pretrained backbone (5.5–6.3 ms; Appendix E).

Larger models reach the aligned regime earlier. In size–time space, low $\bar{E}$ forms a band that shifts to earlier training steps as model size grows (Figure 4). For example, SegFormer B0 reaches its minimum at step 202 and B5 at step 102, and DINOv2 Small at step 42 and Large at step 36 (Appendix Table 3). Per-family slopes are in Appendix D.4.

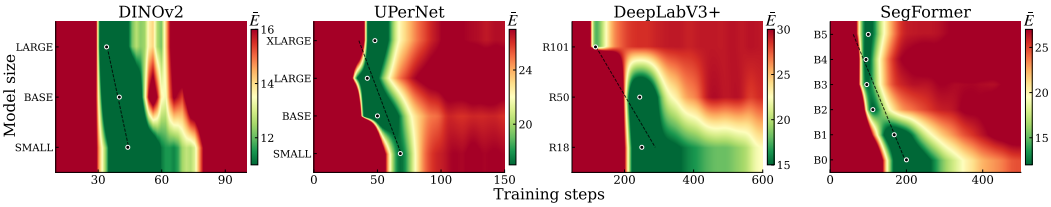


Figure 4: **Larger models reach low alignment error earlier.** $\bar{E}$ (ms) over training steps (horizontal) and model size (vertical) for each family. Low $\bar{E}$ forms a band that occurs earlier in training for larger models.

Fitted geometric operators describe the human effect equally well. We applied classical operators to the ground-truth masks, fitted their single parameter to the human data, and scored them with the same TTC pipeline. Morphological closing reaches $\bar{E} = 1.6$ ms, α-shapes 3.2 ms, and dilation 4.8 ms. These values are within the 0.8–14.5 ms range of the models at their minimum-error checkpoints (Appendix D.5). The convex hull, which fills every concavity, overshoots by 412.8 ms. This is consistent with a human body representation between the exact outline and the convex hull (Li et al., 2023). The two approaches differ in what they receive. The operators start from the correct segmentation, with a smoothing amount chosen on human responses. The models see only the image and no human data, and they pass through this geometry while learning to segment.

Concavities are resolved last. At the aligned stage, convex parts are already sharp while concavities are only partly resolved. As training proceeds, convex corners become sharp first. Concavities stay partly filled for longer: the mask cuts across the mouth of each notch, and the probability map shows a

wide low-confidence band inside it (Figure 5). With more training, this band shrinks and the boundary moves into the notch; larger models show the same progression at a fixed step (Appendix C). Under translation, a partly filled concavity is reached earlier than a convex edge, which gives $\Delta > 0$. The same geometry predicts that small additions inside a concavity change the mask less than additions on a convex part, which we test next.

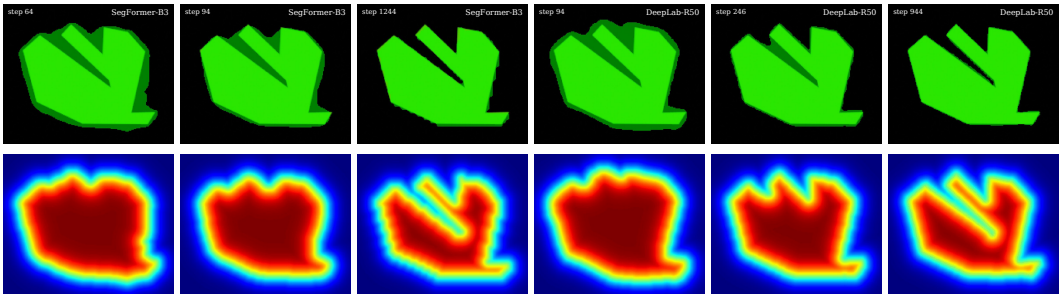
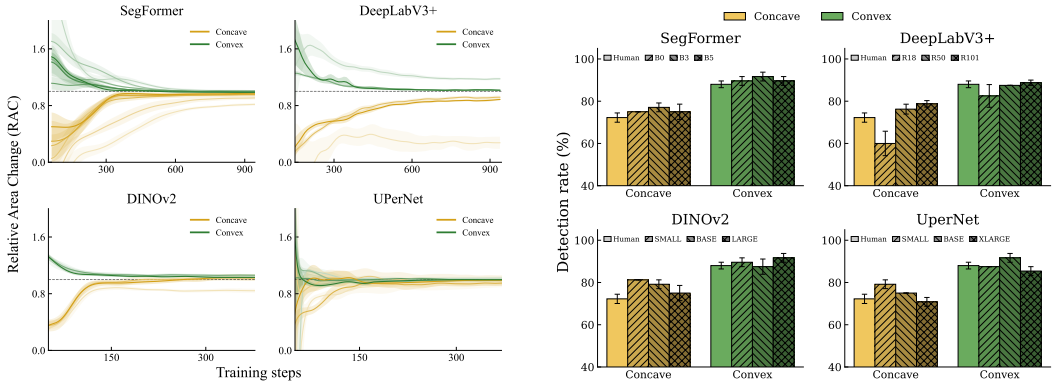


Figure 5: **Concavities are resolved after convex parts.** Masks (top) and foreground softmax probability maps (bottom) for SegFormer-B3 and DeepLabV3+-R50 at an early, an intermediate, and a late training step. Convex parts sharpen first; concave notches stay partly filled at intermediate steps and are resolved late.

4.2 Change detection

Models are less sensitive to concave changes. Like humans, the models are less sensitive to changes inside concavities than to changes on convex parts. Concave RAC stays below convex RAC for most of training and approaches it as concavities are resolved, earlier for larger models (Figure 6A).



(A) (B)

Figure 6: **Models are less sensitive to concave than to convex changes, like humans. (A)** RAC across training; darker lines are larger models, and the dotted line marks RAC $= 1$ (the added area is fully captured). Concave RAC is below convex RAC for most of training. **(B)** Detection rates with one threshold ($\tau = 1.1\%$) fitted on SegFormer-B5 and applied to all configurations shown.

Detection rates follow the human ordering. Converted to binary decisions, the models detect convex changes more often than concave ones in all four families, as humans do (Figure 6B). The detection threshold $\tau = 1.1\%$ is fitted once, on SegFormer-B5 at its TTC-selected checkpoint, and applied unchanged to the other 11 configurations. The ordering holds over a wide range of τ (Appendix F).

5 DISCUSSION

People often rely on approximate "body" representations when making physical judgments: coarse, volumetric approximations that discard fine detail and texture. Fitted geometric operators reproduce the resulting concavity effect (Section 4.1), but they do not explain how such geometry could be learned. We therefore asked whether similar geometry arises in deep segmentation models during standard training.

We find that it does. Model masks simplify concavities while preserving convex detail, mirroring the asymmetry in human responses. Early in training, masks smooth over concavities; with continued training, they resolve increasingly precise boundaries. Alignment with humans peaks in between: mid-training models attain the lowest alignment error, and larger models reach this regime earlier, a pattern consistent across families. Some architectures and training regimes produce effects resembling human judgments, and we illustrate candidate geometries consistent with those effects. Object approximations consistent with human concavity effects thus emerge from segmentation training alone, without any human behavioral supervision, and segmentation accuracy and human alignment peak at different points along a coarse-to-fine trajectory.

Why would humans rely on such approximations? The human visual system develops under metabolic and informational constraints, time pressure, and limited perceptual resolution, which may favor efficient, lower-complexity representations. Our results fit this resource-rational view (Lieder & Griffiths, 2020): with limited training, models also favor coarse geometry over fine detail, and this is the stage at which their behavior best matches human judgments. Segmentation models therefore offer a concrete, controllable setting for studying how such representations can arise.

Precise segmentation remains essential for tasks that depend on exact contours, such as instance annotation or fine-grained recognition. When a model's judgments should agree with people's, however, intermediate coarseness gives the closest match, consistent with the coarse-body account of Li et al. (2023). Checkpoint selection is therefore a simple way to obtain masks that better match human physical judgments from standard segmentation models, with potential value for future work in applications sensitive to agreement with human judgment, such as domestic human–robot interaction.

6 LIMITATIONS AND CONCLUSION

Our experiments use 2D polygon stimuli and controlled kinematics: models segment a single static frame and receive ground-truth velocities. This choice is deliberate, because it isolates object representations from motion estimation and learned dynamics. Humans are also sensitive to the ground-truth time to collision: their concavity effect grows from 20.7 to 53.0 ms as collision time increases. Our models process one static frame and cannot adapt to this dependence, so we aggregate across collision times. Finally, fitted geometric operators applied to the ground-truth masks match the models on the alignment metric (Section 4.1). The models should therefore not be read as a better fit to the human summary; their value lies in showing how such representations could arise in a general-purpose perceptual system. Future work should test these effects in realistic 3D environments and in more advanced model paradigms, such as vision-language models.

Conclusion. Behavioral evidence suggests that people rely on coarse object bodies, not precise shapes, when predicting everyday physical events. We show that segmentation models pass through a regime whose behavioral consequences resemble these human effects, at intermediate stages of training and earlier in larger models. Theoretically, this suggests that a general visual system, not a hand-engineered geometry, can develop human-like object representations when training or size is limited, consistent with resource-rational accounts of cognition (Lieder & Griffiths, 2020). Practically, we provide a framework for measuring how closely model object representations align with human ones. The framework shows that higher segmentation accuracy does not imply closer alignment with human judgments.

REPRODUCIBILITY STATEMENT

Code and the synthetic fine-tuning dataset are available in an anonymized repository at `https://anonymous.4open.science/r/shape-alignment-3C2D`. The repository includes the procedural polygon generator (Appendix B.2), the fine-tuning configuration for all four model families (Appendix B.3), the TTC and CD evaluation pipeline (Sections 3.1 and 3.2), and the analysis scripts for the held-out checkpoint selection, the non-monotonicity tests, and the geometric baselines (Sections 4.1 and 4.2). The human behavioral data are the publicly released data of Li et al. (2023); the participant-level re-analysis that yields Δ_{human} and its confidence interval is included in the repository.

AI USE STATEMENT

We used generative AI tools (large language models) for editing and rewriting text; checking consistency of numbers across sections; drafting LaTeX for tables; and drafting and debugging analysis code. All experiments, analyses and results were produced by the authors, who verified every AI-assisted change and take full responsibility for the content.

REFERENCES

Renée Baillargeon. Infants' understanding of the physical world. In Michel Sabourin, Fergus Craik, and Michèle Robert (eds.), *Advances in Psychological Science, Vol. 2: Biological and Cognitive Aspects*, pp. 503–529. Psychology Press, 1998.

Renée Baillargeon. The acquisition of physical knowledge in infancy: A summary in eight lessons. In Usha Goswami (ed.), *Blackwell Handbook of Childhood Cognitive Development*, pp. 47–83. Blackwell, 2002.

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

Halely Balaban and Tomer D. Ullman. Physics versus graphics as an organizing dichotomy in cognition. *Trends in Cognitive Sciences*, 29(11):985–996, 2025.

Halely Balaban, Kevin A. Smith, Joshua B. Tenenbaum, and Tomer D. Ullman. Electrophysiology reveals that intuitive physics guides visual tracking and working memory. *Open Mind*, 8:1425–1446, 2024.

Ilona Bass, Kevin A. Smith, Elizabeth Bonawitz, and Tomer D. Ullman. Partial mental simulation explains fallacies in physical reasoning. *Cognitive Neuropsychology*, 38(7–8):413–424, 2021.

Christopher J. Bates, Ilker Yildirim, Joshua B. Tenenbaum, and Peter W. Battaglia. Modeling human intuitions about liquid flow with particle-based simulation. *PLoS Computational Biology*, 15(7):e1007210, 2019.

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

Neil R. Bramley, Tobias Gerstenberg, Joshua B. Tenenbaum, and Todd M. Gureckis. Intuitive experimentation in the physical world. *Cognitive Psychology*, 105:9–38, 2018.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 801–818, 2018.

Lynn A. Cooper. Mental rotation of random two-dimensional shapes. *Cognitive Psychology*, 7(1):20–43, 1975.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3213–3223, 2016.

Ernest Davis. *Representations of Commonsense Knowledge*. Morgan Kaufmann, 1990.

Herbert Edelsbrunner, David G. Kirkpatrick, and Raimund Seidel. On the shape of a set of points in the plane. *IEEE Transactions on Information Theory*, 29(4):551–559, 1983.

Thomas Fel, Ivan Felipe, Drew Linsley, and Thomas Serre. Harmonizing the object recognition strategies of deep neural networks with humans. In *Advances in Neural Information Processing Systems*, volume 35, pp. 9432–9446, 2022.

Jason Fischer, John G. Mikhael, Joshua B. Tenenbaum, and Nancy Kanwisher. Functional neuroanatomy of intuitive physical inference. *Proceedings of the National Academy of Sciences*, 113(34):E5072–E5081, 2016.

Kenneth D. Forbus. Qualitative physics: Past, present, and future. In Howard E. Shrobe (ed.), *Exploring Artificial Intelligence: Survey Talks from the National Conferences on Artificial Intelligence*, pp. 239–296. Morgan Kaufmann, 1988.

Katerina Fragkiadaki, Pulkit Agrawal, Sergey Levine, and Jitendra Malik. Learning visual predictive models of physics for playing billiards. In *International Conference on Learning Representations (ICLR)*, 2016.

Jennifer J. Freyd. Dynamic mental representations. *Psychological Review*, 94(4):427–438, 1987.

Tobias Gerstenberg and Simon Stephan. A counterfactual simulation model of causation by omission. *Cognition*, 216:104842, 2021.

David L. Gilden and Dennis R. Proffitt. Understanding collision dynamics. *Journal of Experimental Psychology: Human Perception and Performance*, 15(2):372–383, 1989.

Melvyn A. Goodale and A. David Milner. Separate visual pathways for perception and action. *Trends in Neurosciences*, 15(1):20–25, 1992.

Rob Gray and Ian M. Thornton. Exploring the link between time to collision and representational momentum. *Perception*, 30(8):1007–1022, 2001.

Jessica B. Hamrick, Peter W. Battaglia, Thomas L. Griffiths, and Joshua B. Tenenbaum. Inferring mass in complex scenes by mental simulation. *Cognition*, 157:61–76, 2016.

Joshua K. Hartshorne and Mengguo Jing. Insights into cognitive mechanics from education, developmental psychology and cognitive science. *Nature Reviews Psychology*, 4(4):277–291, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Mary Hegarty. Mechanical reasoning by mental simulation. *Trends in Cognitive Sciences*, 8(6):280–285, 2004.

James R. Kubricht, Keith J. Holyoak, and Hongjing Lu. Intuitive physics: Current research and controversies. *Trends in Cognitive Sciences*, 21(10):749–759, 2017.

Adam Lerer, Sam Gross, and Rob Fergus. Learning physical intuition of block towers by example. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 430–438, 2016.

Yichen Li, YingQiao Wang, Tal Boger, Kevin A. Smith, Samuel J. Gershman, and Tomer D. Ullman. An approximate representation of objects underlies physical reasoning. *Journal of Experimental Psychology: General*, 152(11):3074–3086, 2023.

Falk Lieder and Thomas L. Griffiths. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, 43:e1, 2020.

Drew Linsley, Ivan F. Rodriguez Rodriguez, Thomas Fel, Michael Arcaro, Saloni Sharma, Margaret Livingstone, and Thomas Serre. Performance-optimized deep neural networks are evolving into worse models of inferotemporal visual cortex. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11976–11986, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

David Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. MIT Press, 2010. Originally published 1982 by W. H. Freeman.

Michael McCloskey and Deborah Kohl. Naive physics: The curvilinear impetus principle and its role in interactions with moving objects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 9 (1):146–156, 1983.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *Proceedings of the International Conference on 3D Vision (3DV)*, pp. 565–571, 2016.

Roozbeh Mottaghi, Hessam Bagherinezhad, Mohammad Rastegari, and Ali Farhadi. Newtonian scene understanding: Unfolding the dynamics of objects in static images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3521–3529, 2016.

Manfred Nusseck, Julien Lagarde, Benoit Bardy, Roland Fleming, and Heinrich H. Bülthoff. Perception and prediction of simple object interactions. In *Proceedings of the 4th Symposium on Applied Perception in Graphics and Visualization (APGV)*, pp. 27–34, 2007.

Byung-Doh Oh and William Schuler. Transformer-based language model surprisal predicts human reading times best with about two billion training tokens. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 1915–1921, 2023a.

Byung-Doh Oh and William Schuler. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, 11:336–350, 2023b.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Luis S. Piloto, Ari Weinstein, Peter W. Battaglia, and Matthew Botvinick. Intuitive physics learning in a deep-learning model inspired by developmental psychology. *Nature Human Behaviour*, 6(9):1257–1267, 2022.

R. T. Pramod, Michael A. Cohen, Joshua B. Tenenbaum, and Nancy Kanwisher. Invariant representation of physical stability in the human brain. *eLife*, 11:e71736, 2022.

R. T. Pramod, Elizabeth Mieczkowski, Cyn X. Fang, Joshua B. Tenenbaum, and Nancy Kanwisher. Decoding predicted future states from the brain's "physics engine". *Science Advances*, 11(22):eadr7429, 2025.

Dennis R. Proffitt, Mary K. Kaiser, and Susan M. Whelan. Understanding wheel dynamics. *Cognitive Psychology*, 22(3):342–373, 1990.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5301–5310, 2019.

David A. Rosenbaum. Perception and extrapolation of velocity and acceleration. *Journal of Experimental Psychology: Human Perception and Performance*, 1(4):395–403, 1975.

Adam N. Sanborn, Vikash K. Mansinghka, and Thomas L. Griffiths. Reconciling intuitive physics and Newtonian mechanics for colliding objects. *Psychological Review*, 120(2):411–437, 2013.

Robert S. Siegler and Zhe Chen. Developmental differences in rule learning: A microgenetic analysis. *Cognitive Psychology*, 36(3):273–310, 1998.

Kevin A. Smith and Edward Vul. Sources of uncertainty in intuitive physics. *Topics in Cognitive Science*, 5(1): 185–199, 2013.

Kevin A. Smith, Lingjie Mei, Shunyu Yao, Jiajun Wu, Elizabeth S. Spelke, Joshua B. Tenenbaum, and Tomer D. Ullman. Modeling expectation violation in intuitive physics with coarse probabilistic object representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Kevin A. Smith, Jessica B. Hamrick, Adam N. Sanborn, Peter W. Battaglia, Tobias Gerstenberg, Tomer D. Ullman, and Joshua B. Tenenbaum. Intuitive physics as probabilistic inference. In Thomas L. Griffiths, Nick Chater, and Joshua B. Tenenbaum (eds.), *Bayesian Models of Cognition: Reverse Engineering the Mind*, pp. 369–393. MIT Press, 2024.

Felix A. Sosa, Tomer D. Ullman, Joshua B. Tenenbaum, Samuel J. Gershman, and Tobias Gerstenberg. Moral dynamics: Grounding moral judgment in intuitive physics and intuitive psychology. *Cognition*, 217:104890, 2021.

Felix A. Sosa, Samuel J. Gershman, and Tomer D. Ullman. Blending simulation and abstraction for physical reasoning. *Cognition*, 254:105995, 2025.

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

Elizabeth S. Spelke. *What Babies Know: Core Knowledge and Composition, Volume 1*. Oxford University Press, 2022.

Elizabeth S. Spelke and Katherine D. Kinzler. Core knowledge. *Developmental Science*, 10(1):89–96, 2007.

J. R. Tresilian. Perceptual and cognitive processes in time-to-contact estimation: Analysis of prediction-motion and relative judgment tasks. *Perception & Psychophysics*, 57(2):231–245, 1995.

Tomer D. Ullman, Elizabeth S. Spelke, Peter W. Battaglia, and Joshua B. Tenenbaum. Mind games: Game engines as an architecture for intuitive physics. *Trends in Cognitive Sciences*, 21(9):649–665, 2017.

Leslie G. Ungerleider and Mortimer Mishkin. Two cortical visual systems. In David J. Ingle, Melvyn A. Goodale, and Richard J. W. Mansfield (eds.), *Analysis of Visual Behavior*, pp. 549–586. MIT Press, 1982.

YingQiao Wang and Tomer D. Ullman. Resource bounds on mental simulations: Evidence from a liquid-reasoning task. *Journal of Experimental Psychology: General*, 154(8):2105–2124, 2025.

Justin N. Wood, Tomer D. Ullman, Brian W. Wood, Elizabeth S. Spelke, and Samantha M. W. Wood. Object permanence in newborn chicks is robust against opposing evidence. *arXiv preprint arXiv:2402.14641*, 2024.

Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 418–434, 2018.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. SegFormer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5122–5130, 2017.

APPENDIX OVERVIEW

- **Appendix A: Human data.** The two experiments of Li et al. (2023), the CD stimuli, and the summary statistics of the human effect.
- **Appendix B: Models and training.** Architectures, the synthetic fine-tuning dataset, and the training settings.
- **Appendix C: Mask geometry across model size.** How masks change with model size at a fixed training step.
- **Appendix D: Time-to-collision analyses.** Statistics at the minimum-error checkpoints, held-out checkpoint selection with per-family tests, the signed effect across training, the size–timing slopes, and the geometric baselines.
- **Appendix E: Pretraining ablation.** DINOv2 trained from random initialization.
- **Appendix F: Change-detection analyses.** Threshold selection and robustness to the threshold.

A HUMAN DATA

All human data are the publicly released data of Li et al. (2023). The human stimuli are used only for evaluation; no model is trained on them.

Time-to-collision (TTC). 178 participants watched 96 videos in which one polygon moves toward a stationary one. At a known time the stationary polygon disappears, and participants press a key when they think the two have collided. The response time is measured from motion onset to the key press. The videos form 48 matched pairs (8 shapes $\times$ 6 ground-truth collision times). The two videos of a pair share the same kinematics and differ only in whether a concave or a convex side of the stationary polygon faces the collision.

Human TTC effect. Each participant contributes one convex$-$concave difference. Averaged over participants, the concavity effect is $\Delta_{\text{human}} = 31.17$ ms (95% CI $[22.74, 39.60]$, t-interval over the 178 per-participant differences). The between-participant standard deviation is 57.0 ms, so the effect corresponds to Cohen's $d = 0.55$. The effect depends on the ground-truth collision time: it grows from 20.7 to 53.0 ms as collision time increases. Our models process a single static frame and cannot express this dependence, so all model–human comparisons average over collision times (Equation 5).

Change detection (CD). 50 participants saw a polygon I_{init} for 1 s, a blank screen for 2 s, and a second polygon I_{out} for 1 s, and reported whether it had changed. I_{out} adds a small triangle of equal area either inside a concavity or on a convex part of I_{init} (Figure 7), with 16 stimuli per condition (8 shapes $\times$ 2 change sizes). Li et al. (2023) also include a third condition, in which the added area lies inside a large concavity without filling it; we use only the concave and convex conditions. Humans detect convex changes more often than concave ones. This is Experiment 3b of Li et al. (2023).

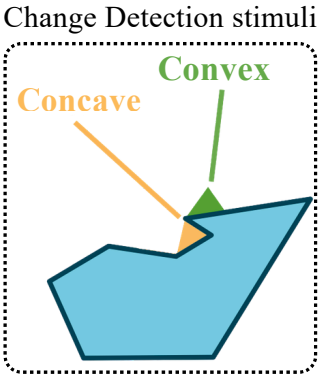


Figure 7: **Change-detection stimuli.** The same area is added either inside a concavity (orange) or on a convex part (green).

B MODELS AND TRAINING

B.1 ARCHITECTURES

We evaluate 16 configurations from four families (Table 1). The families cover convolutional backbones (DeepLabV3+ with ResNet, UPerNet with ConvNeXt) and transformer backbones (SegFormer, DINOv2), each at several sizes. All models are fine-tuned end to end on the synthetic dataset of Appendix B.2 with the same recipe (Appendix B.3). Per-configuration parameter counts are listed in Table 3.

Family	Backbones	Params (M)	Initialization
SegFormer (Xie et al., 2021)	B0–B5	3.7–84.6	ADE20K (Zhou et al., 2017)
DeepLabV3+ (Chen et al., 2018)	ResNet-18/50/101	12.3–60.4	Cityscapes (Cordts et al., 2016)
UPerNet (Xiao et al., 2018)	ConvNeXt S/B/L/XL	82.0–391.0	ADE20K
DINOv2[†] (Oquab et al., 2024)	S/B/L	27.2–310.6	Self-supervised

Table 1: **Model families.** Parameter ranges span the smallest to the largest backbone in each family. [†]With a light convolutional decoder.

B.2 SYNTHETIC FINE-TUNING DATASET

Models pretrained on natural images do not reliably segment the polygonal stimuli of Li et al. (2023). To adapt them without contaminating evaluation, we built a synthetic dataset that matches the geometric and photometric statistics of the human stimuli but shares no shapes with them. The dataset and its generator are in the anonymized repository linked in the Reproducibility Statement.

Canvas. Each image is a 1000×400 (width $\times$ height) RGB canvas with a single uniformly colored polygon centered on a black background. The aspect ratio matches the stimuli of Li et al. (2023).

Polygon generator. Each polygon is defined by:

- a vertex count sampled uniformly from $\{5, \ldots, 12\}$;
- a base radius $r_{\text{base}} \sim \mathcal{U}(58, 62)$ pixels, which keeps polygon area in a narrow band;
- an irregularity coefficient $\sim \mathcal{U}(0.15, 0.30)$ for the angular jitter between successive vertices;
- a spikiness coefficient $\sim \mathcal{U}(0.15, 0.25)$, which sets the radial variation around r_{base}.

After this polar layout, each vertex is moved by an independent Cartesian offset of magnitude $\sim \mathcal{U}(0.20, 0.40) \cdot r_{\text{base}}$, so that shapes are not star-convex. Independent x and y scale factors $\sim \mathcal{U}(0.85, 1.15)$ then remove any remaining circular bias.

Concavities. Each polygon receives 1 to 4 concave notches (uniform). A notch is inserted on a random edge by adding interior vertices pulled 35–55% of the way toward the centroid, with small lateral jitter. Three notch styles are mixed:

- one interior vertex near the edge midpoint (50% of notches);
- two interior vertices at $1/3$ and $2/3$ of the edge (30%);
- two interior vertices close together around the midpoint (20%).

A polygon is rejected and resampled if it leaves the canvas (10 px margin) or if any two non-adjacent edges intersect. Together with the offsets and scaling above, this covers the concave geometry of the human stimuli without reproducing any of them.

Colors. Fill colors are drawn uniformly from 24 saturated hues, luminance-matched to the stimuli of Li et al. (2023): RGB channels are capped near $V \approx 81\%$, saturation is near 100%, and per-color luminance falls in high (140–191), medium (85–139), and low (17–64) bands in a 40/40/20 ratio. Mean luminance is ≈ 112, close to that of the human stimuli. Polygons are rendered at $4\times$ supersampling and downsampled with Lanczos resampling, giving antialiased edges.

Splits and ground truth. The dataset has 500 training and 200 validation images, split with a fixed seed. Shapes and colors are sampled independently per image, so no polygon is shared across splits, and no frame of Li et al. (2023) appears in either split. Ground-truth masks are obtained by thresholding the rendered image at intensity > 10, which separates the polygon from the background while keeping antialiased boundary pixels. No data augmentation is used; training and validation losses agree closely throughout fine-tuning.

B.3 Fine-tuning hyperparameters

All configurations share the settings in Table 2.

Setting	Value
Optimizer	AdamW (weight decay 0.01)
Learning rate	5×10^{-5}
Scheduler	Cosine annealing with linear warmup (5% of total steps)
Loss	Cross-entropy + Dice ($\lambda = 0.5$)
Effective batch size	8 (gradient accumulation)
Training length	15 epochs (945 steps)
Training / validation images	500 / 200
Data augmentation	None
Image size	400×1000 px (height $\times$ width)
Precision	bfloat16 (mixed)
Gradient clipping	Max norm 1.0
Seeds	3 per configuration
Hardware	NVIDIA A100 (80 GB)

Table 2: **Fine-tuning hyperparameters,** shared by all 16 configurations.

C Mask geometry across model size

Figure 8 shows masks at a fixed training step for three sizes in two families. Smaller models produce coarser masks that leave concave notches partly filled. Larger models resolve the notches further at the same step, while convex parts are sharp at all sizes. This is the same progression that training time produces (Figure 5).

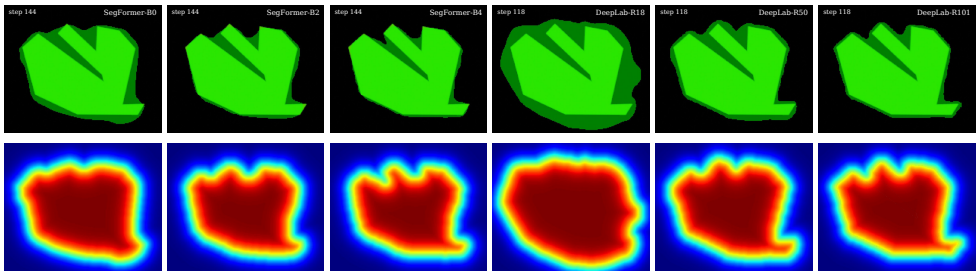


Figure 8: **Masks across model size,** at a fixed training step (step 144 for SegFormer B0/B2/B4; step 118 for DeepLabV3+ R18/R50/R101). Top: mask overlays. Bottom: foreground probability maps, obtained applying a softmax to the logits.

D Time-to-collision analyses

D.1 Statistics at minimum-error checkpoints

Table 3 describes each configuration at the checkpoint with minimum $\bar{E}$ on the full participant sample. These numbers are in-sample; held-out results are in Appendix D.2. $\bar{E}$ and Δ are given as mean $\pm$ SEM across seeds, with Δ first averaged over the six collision times. *Pairwise* is the number of the

48 matched concave–convex pairs (8 shapes $\times$ 6 collision times) with $\mathrm{TTC}_{\mathrm{convex}} > \mathrm{TTC}_{\mathrm{concave}}$, averaged across seeds. *Test* is a one-sided paired t-test over matched pairs, $H_1 : \mu_{\mathrm{convex}} > \mu_{\mathrm{concave}}$ ($^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$).

Every configuration shows a positive concavity effect at its minimum-error checkpoint. Over the whole trajectory, the seed-averaged effect lies inside the human 95% CI for a window of checkpoints in all 16 configurations (Table 4).

Model	Params (M)	Best step	$\bar{E}$ (ms)	Δ (ms)	Pairwise	Test
SegFormer B0	3.7	202	2.4 ± 0.5	31.3 ± 1.8	45/48	***
SegFormer B1	13.7	164	5.3 ± 1.4	32.4 ± 3.3	43/48	***
SegFormer B2	27.4	102	5.0 ± 3.4	28.6 ± 4.3	44/48	***
SegFormer B3	47.2	98	4.9 ± 2.7	27.0 ± 3.1	46/48	***
SegFormer B4	64.1	92	3.4 ± 1.3	31.0 ± 2.3	46/48	***
SegFormer B5	84.6	102	3.0 ± 1.8	32.1 ± 2.5	46/48	***
DeepLabV3+ R18	12.3	240	3.8 ± 1.6	34.3 ± 2.2	39/48	***
DeepLabV3+ R50	41.3	248	2.5 ± 1.1	29.3 ± 1.7	31/48	**
DeepLabV3+ R101	60.4	94	6.0 ± 5.3	26.0 ± 5.8	35/48	**
DINOv2 Small	27.2	42	3.2 ± 1.9	29.6 ± 2.8	44/48	***
DINOv2 Base	**92.9**	**38**	$\mathbf{0.8 \pm 0.3}$	$\mathbf{31.2 \pm 0.6}$	**46/48**	***
DINOv2 Large	310.6	36	3.1 ± 2.3	28.7 ± 2.7	47/48	***
UPerNet Small	82.0	68	14.5 ± 5.9	18.5 ± 7.8	38/48	***
UPerNet Base	122.0	38	3.3 ± 1.8	34.4 ± 1.8	41/48	***
UPerNet Large	235.0	40	11.1 ± 4.0	20.0 ± 4.0	42/48	***
UPerNet XLarge	391.0	42	4.3 ± 2.7	26.8 ± 2.7	42/48	***
Human				31.17 [22.74, 39.60]		

Table 3: **TTC statistics at minimum-error checkpoints (in-sample).** The best configuration is in bold. The human row gives the mean effect and its 95% CI.

D.2 Held-out checkpoint selection

Protocol. We split the 178 participants into two equal folds. The minimum-$\bar{E}$ checkpoint is selected on one fold, with Δ_{human} computed from that fold only, and $\bar{E}$ at that checkpoint is measured on the other fold. This is repeated 200 times, in both directions.

Results. On held-out participants, the best-aligned configuration of each family keeps $\bar{E}$ of 5–8 ms (SegFormer B0 5.3, UPerNet XLarge 5.6, DINOv2 Base 5.6, DeepLabV3+ R50 7.6 ms). Selecting on held-out participants instead of on the evaluated ones raises $\bar{E}$ by 0.4–5.4 ms across the 16 configurations, less than the 8.4 ms half-width of the human 95% CI.

Test of non-monotonicity. On each held-out fold, we compare $\bar{E}$ at the selected checkpoint with $\bar{E}$ at the first and last checkpoints, which are fixed in advance. A non-monotonic trajectory requires the selected checkpoint to be lower than both. The main text reports the test across all 16 configurations. Table 4 also gives the tests within each family. With three to six configurations per family, these tests have low power. The comparison with the last checkpoint is significant in all four families, and the comparison with the first checkpoint in SegFormer, UPerNet and DeepLabV3+.

Seed variability. At the selected checkpoint, the standard deviation of $\bar{E}$ across training seeds is 4.3 ms on average, below the 8.4 ms half-width of the human CI. Per-configuration seed spread is the SEM column of Table 3.

Effect sizes. In units of the human between-participant standard deviation (57.0 ms), the human effect is $d = 0.55$, and the model effects at their selected checkpoints span $d = 0.32$ to 0.60.

D.3 Signed concavity effect across training

Figure 9 shows the signed effect $\Delta_{\mathrm{model},s}(\tau)$, averaged over collision times and then over seeds. Across families, Δ starts above the human value, passes through it, and then falls further as concave

Table 4: **Per-family tests of the training trajectory.** The aligned checkpoint is selected on one half of the participants, and $\bar{E}$ is compared on the other half with the first and last checkpoints (one-sided paired t-tests across the n backbones of each family, $df = n - 1$). Δ in human CI: configurations whose seed-averaged concavity effect lies inside the human 95% CI for a window of checkpoints. $\hat{\beta}$: OLS slope of model-size rank on the minimum-$\bar{E}$ training step, in size ranks per step; negative values mean larger models reach the minimum earlier (two-sided p, $df = n - 2$; Appendix D.4).

		Aligned < first		Aligned < last			Size slope	
Family	n	t	p	t	p	Δ in human CI	$\hat{\beta}$	p
SegFormer	6	-2.68	.022	-27.4	<.0001	6/6	-0.037	.020
UPerNet	4	-4.15	.013	-8.03	.002	4/4	-0.090	.216
DINOv2	3	-2.54	.063	-31.6	<.001	3/3	-0.197	.073
DeepLabV3+	3	-5.25	.017	-12.6	.003	3/3	-0.012	.308
All	16	-5.17	5.7×10^{-5}	-19.6	2.1×10^{-12}	16/16	–	–

boundaries are resolved. In all 16 configurations, the seed-averaged effect lies inside the human 95% CI for a window of checkpoints.

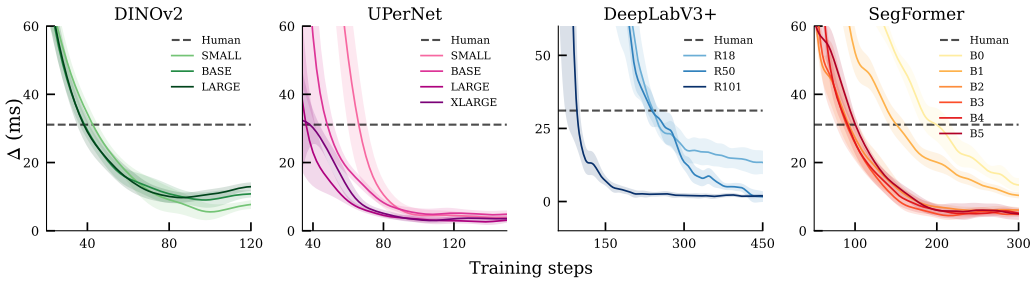


Figure 9: **Signed TTC concavity effect across training.** Positive values mean earlier predicted collision for concave than for convex configurations. Lines: mean across seeds; bands: ±SEM; dashed line: human effect.

Why $\bar{E}$ does not reach zero. The signed error of each seed does pass through zero. Without smoothing, the minimum of $|E_s|$ over checkpoints is 0.17–1.03 ms across the SegFormer backbones, and individual seeds reach 0.01 ms. The plotted curves stay higher for two reasons:

- $\bar{E}$ takes the absolute value per seed and then averages. Seeds cross zero at different steps, so the average does not touch zero even when each seed does. Averaging the signed errors first would give 0.04–1.54 ms.
- Display smoothing adds a further 1.1–7.0 ms.

Seed averaging is the larger of the two. In addition, checkpoints are sampled at discrete steps, so none needs to fall exactly at a zero crossing.

D.4 Model size and timing of alignment

For each family, $\hat{\beta}$ is the OLS slope of model-size rank on the minimum-$\bar{E}$ training step (Table 4).

The fitted steps are the per-row minima of the Figure 4 heatmaps (seed-averaged $\bar{E}$, smoothed), not the unsmoothed best steps of Table 3. It is negative in all four families and significant for SegFormer ($n = 6$, $p = .020$), but not for UPerNet ($p = .216$), DINOv2 ($p = .073$) or DeepLabV3+ ($p = .308$). With three to six sizes per family, these per-family tests have little power.

D.5 GEOMETRIC BASELINES

We applied four operators to the ground-truth polygon masks of the TTC stimuli: morphological closing, α-shapes (Edelsbrunner et al., 1983), dilation, and the convex hull. The free parameter of each operator was fitted to the human data, and the resulting masks were scored with the same TTC pipeline and $\bar{E}$ as the models (Table 5). Both the operator parameters and the model checkpoints are selected on the full participant sample, so the comparison is in-sample on both sides.

Operator	Fitted parameter	$\bar{E}$ (ms)
Morphological closing	$r = 36$	1.6
α-shape	$\alpha = 24$	3.2
Dilation	$r = 6$	4.8
Convex hull	none	412.8
Models, minimum-error checkpoints (Table 3)		0.8–14.5

Table 5: **Geometric operators applied to ground-truth masks,** with parameters fitted to the human data and scored with the same $\bar{E}$ as the models.

E PRETRAINING ABLATION

Motivation. All models start from weights learned on natural images: SegFormer, UPerNet, and DeepLabV3+ from semantic-segmentation checkpoints (ADE20K or Cityscapes), and DINOv2 from self-supervised weights (Table 1). The coarse-to-fine trajectory could therefore be inherited from this pretraining rather than produced by learning to segment the polygons.

Setup. We reran the full DINOv2 pipeline from random initialization: fine-tuning on the synthetic dataset (Appendix B.2) and TTC evaluation on the human stimuli. Both the pretrained and the randomly initialized backbones were scored with the held-out participant protocol of Appendix D.2.

Result. Held-out $\bar{E}$ is 5.5–6.3 ms for both the pretrained and the randomly initialized backbones. Natural-image pretraining is therefore not needed to reach the level of alignment reported in the main text.

F CHANGE-DETECTION ANALYSES

F.1 FROM MASK AREAS TO DETECTION DECISIONS

Section 3.2 uses two area measures. RAC $= (A_{\text{out}} - A_{\text{init}})/A_{\text{gt}}$ normalizes by the area of the added segment and measures local sensitivity (main-text Figure 6A). To obtain binary decisions comparable to human responses, we instead use the relative change in mask area, $\delta = |A_{\text{out}} - A_{\text{init}}|/A_{\text{init}}$, and label a change as detected iff $\delta > \tau$.

Threshold selection. The threshold is fitted once, on a reference configuration fixed in advance: SegFormer-B5 (ADE20K, seed 123) at its minimum-$\bar{E}$ checkpoint. That checkpoint is selected on the TTC task, so no CD data are used for checkpoint selection. We choose τ by minimizing the RMSE between model and human detection rates over $\tau \in \{0.1, 1.1, \ldots, 20.1\}\%$, which gives $\tau = 1.1\%$. This value is then applied unchanged to every other configuration (main-text Figure 6B), which are therefore out of sample with respect to the threshold.

F.2 DETECTION RATE VS. THRESHOLD

The convex–concave ordering does not depend on the chosen threshold. In all four families, convex changes are detected more often than concave ones over a wide range of θ (Figure 10). This figure uses the final training step, not the minimum-$\bar{E}$ checkpoint.

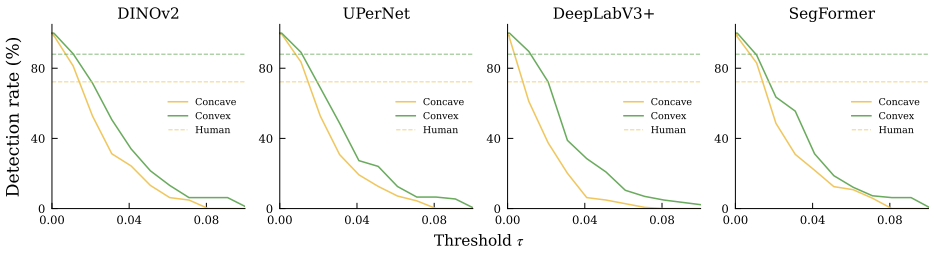


Figure 10: **Detection rate as a function of the threshold τ on δ**, at the final training step, averaged over backbones and seeds. Solid lines: model detection rates for concave (yellow) and convex (green) changes. Dashed lines: human detection rates.