Computer Vision Models Show Human-Like Sensitivity to Geometric and Topological Concepts

Zekun Wang (zekun@gatech.edu)
School of Interactive Computing
Atlanta, GA 30332 USA

Sashank Varma (varma@gatech.edu)
School of Interactive Computing
School of Psychology
Atlanta, GA 30332 USA

Abstract

With the rapid improvement of machine learning (ML) models, cognitive scientists are increasingly asking about their alignment with how humans think. Here, we ask this question for computer vision models and human sensitivity to geometric and topological (GT) concepts. Under the *core knowledge* account, these concepts are innate and supported by dedicated neural circuitry. In this work, we investigate an alternative explanation, that GT concepts are learned "for free" through everyday interaction with the environment. We do so using computer visions models, which are trained on large image datasets. We build on prior studies to investigate the overall performance and human alignment of three classes of models – convolutional neural networks (CNNs), transformer-based models, and vision-language models – on an odd-one-out task testing 43 GT concepts spanning seven classes. Transformer-based models achieve the highest overall accuracy, surpassing that of young children. They also show strong alignment with children's performance, finding the same classes of concepts easy vs. difficult. By contrast, vision-language models underperform their vision-only counterparts and deviate further from human profiles, indicating that naïve multimodality might compromise abstract geometric sensitivity. These findings support the use of computer vision models to evaluate the sufficiency of the learning account for explaining human sensitivity to GT concepts, while also suggesting that integrating linguistic and visual representations might have unpredicted deleterious consequences.

Keywords: geometric reasoning; mathematical cognition; cognitive alignment; computer vision models; vision-language models

Introduction

Humans learn geometric and topological (GT) concepts through mathematics instruction. Long before they begin formal school, however, they show sensitivity to concepts such as shape, angle, rotation, and translation, leading to the proposal that they are part of *core knowledge*, i.e., innate and supported by dedicated neural circuitry (Spelke & Kinzler, 2007). This may be the consequence of evolutionary processes building in sensitivity to the mathematical and spatial structure of the world, for example to support navigation (Gibson, 1979; Shepard, 1994). Supporting evidence comes from the finding of similar sensitivities in non-human animals (Chiandetti & Vallortigara, 2007; Pasupathy & Connor, 1999; Vallortigara, 2018). The core knowledge account "explains" the emergence of GT concepts in young children who have not yet entered school and received formal mathematics instruction.

A contrasting account is that GT concepts are learned "for free", through everyday experience in the world. This account has been less studied, perhaps because a strong empirical test would require "ablation" studies where organisms are deprived of a typically rich environment in which to learn, raising ethical questions even for animal subjects.

The recent rapid rise of deep learning models provides a new way to test the sufficiency of the learning account. These models are increasingly demonstrating human-level performance in domains ranging from language acquisition (Ma, Wang, & Chai, 2024; Ma, Pan, & Chai, 2023; Chang & Bergen, 2022) to theory of mind (Ma, Sansom, Peng, & Chai, 2024; Jung et al., 2024; Y. He et al., 2023) to mathematical reasoning (Shah, Marupudi, Koenen, Bhardwaj, & Varma, 2023; Ahn et al., 2024) to concept understanding (Jin et al., 2024; Vemuri, Shah, & Varma, 2024). In addition to their fidelity at the behavioral level, there is also increasing evidence of their fidelity at the brain level. Researchers have mapped the layers of convolutional neural network (CNN) models to areas of the human visual system, specifically the ventral visual stream (K. He, Zhang, Ren, & Sun, 2015; Simonyan & Zisserman, 2015; Krizhevsky, Sutskever, & Hinton, 2012; Jacob, Pramod, Katti, & Arun, 2021; De Cesarei, Cavicchi, Cristadoro, & Lippi, 2021; Lindsay, 2021). Portelance (2022) argues that despite differences in the underling learning processes of humans and ML models, we can use these models to derive new hypotheses about human cognition, and conversely we can use cognitiv (neuro)science data to train more human-like models. We adopt a similar view.

The current study evaluates the sufficiency of the account that GT concepts are learned through experience in the world. It builds on prior studies that have used CNN models and found promising initial results. It replicates their findings and extends their approach to two new and important classes of computer vision models. The first is vision transformer models, which are notable not just for their different architecture but also for their larger size and ability to be trained on more data. The second class is vision-language models, which learn about both the visual-spatial structure of the environment and also its linguistic structure, and the mapping between the two. These models are of potential cognitive interest given that *dual-coding theory* (Paivio, 1991) suggests that processing information through both non-verbal and verbal channels should be advantageous. We compare these three classes of ML models to humans in three research questions:

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

1. Do CNNs, transformers, and/or vision-language models approach (or even exceed) the sensitivity to GT concepts shown by children and adults?

2. Which models / architectures best align with human performance, i.e., find the same classes of concepts easy vs. difficult. Is their alignment close enough to consider them viable cognitive science models?

3. How does adding language (i.e., symbols) to vision (i.e., space) affect the overall sensitivity and class-by-class alignment of vision-language models?

Related Work

GT concept understanding in humans

Dehaene, Izard, Pica, and Spelke (2006) conducted a pioneering and comprehensive study of human sensitivity to GT concepts. They developed a purely visual task that was appropriate even for participants with no formal mathematics education. In this task, each stimulus consists of six images. Five images embody a particular GT concept but the sixth does not. The images otherwise vary in their visual features. The participant is simply asked to indicate the "odd one out". They developed stimuli for 43 individual concepts spanning seven broader classes: Topology, Euclidean Geometry, Geometrical Figures, Symmetrical Figures, Chiral Figures, Metric Properties, and Geometrical Transformations. See Figure 1 for example stimuli for 7 concepts.

Dehaene et al. (2006) found that the Mundurucu, an indigenous Amazonian group whose members have no formal schooling, were nevertheless sensitive to 39 of the 43 concepts. Moreover, the Mundurucu adults and children performed comparably. Marupudi and Varma (2023) replicated these findings using a modified 2 alternative forced-choice version of the odd-one-out task, and furthermore showed that cognitive ability in general (i.e., fluid intelligence) and visuospatial ability in particular (i.e., mental rotation) explained only a small portion of the variability in people's performance. Izard and Spelke (2009) extended the Dehaene et al. (2006) study to people from the US. A distinctive contribution was testing both young children (ages 3–6 years) and adults. The young children also displayed sensitivity to GT concepts, displaying above-chance performance for 27 of the 43 concepts. The finding of sensitivity to GT concepts in samples that have had no formal schooling has been generally interpreted as evidence for the core knowledge account.

Interestingly, the three studies found similar variation in the relative difficulty of different GT concepts. For example, participants from all groups – Mundurucu and Western, children and adults – were highly accurate for Euclidean Geometry concepts, but found Geometrical Transformations to be difficult. Capturing this variability is an important goal.

GT concept understanding in ML models

Computer vision researchers have documented the sensitivity of CNN models to various GT concepts. Jaderberg, Simonyan, Zisserman, and Kavukcuoglu (2016) showed that CNNs with affine transformation are sensitive to parallel lines but insensitive to the affine transformation themselves because affine transformation preserves parallel lines. Laptev, Savinov, Buhmann, and Pollefeys (2016) demonstrated that training on datasets augmented with rotated images can make CNNs invariant to rotation transformations, which potentially *works against* sensitivity to the rotation concept within the Geometric Transformation class. Mumuni and Mumuni (2021) provide a comprehensive review of approaches to extending the CNN architecture to handle non-trivial geometric transformations.

The goal of this research – to develop models that are invariant to different geometric transformations, for the purpose of better generalization – makes sense for computer vision. For cognitive science, the question is whether these models learn human-like sensitivities. Hsu, Wu, and Goodman (2022) tested the sensitivity of CNNs to Euclidean geometry concepts, finding that humans outperform CNNs pre-trained on the ImageNet (Deng et al., 2009). In the direct precursor to the current study, Upadhyay, Marupudi, Varma, and Varma (2025) investigated the sensitivity of a broad range of CNN models to the 43 GT concepts of Dehaene et al. (2006). ResNet-18 (K. He et al., 2015) showed the highest overall accuracy, though it was still below that of the young children of the Izard and Spelke (2009) study. They also found medium-size correlations ($r \approx 0.55$) between the CNN models and humans in their average accuracies across the seven classes of GT concepts. Thus, there is some evidence of alignment, but also much room for improvement. In another related study, Campbell, Kumar, Giallanza, Griffiths, and Cohen (2024) explored the sensitivity of vision transformer models such as DINOv2 (Oquab et al., 2024) and CLIP (Radford et al., 2021) to the Geometric Regularity stimuli of Sable-Meyer, Ellis, Tenenbaum, and Dehaene (2022) and the Geometric Parts and Relations stimuli of Hsu et al. (2022). These newer models showed stronger alignment with human performance than the older CNN model ResNet-50 (K. He et al., 2015).

Method

Models

The current study explored the sensitivity to GT concepts and the human alignment of three classes of models – CNNs, Transformers, and Vision-Language Models (VLMs). We selected the following examples of each class:

- **CNNs**: ResNet-50, ResNet-18 (K. He et al., 2015), and EfficientNet (Tan & Le, 2020)
- **Transformers**: Vision Transformer (ViT) (Dosovitskiy et al., 2021) and DINOv2 (Oquab et al., 2024)
- **VLMs**: CLIP (Radford et al., 2021) with either a CNN or ViT vision backbone, and ALIGN (Jia et al., 2021)

Some models were selected based on prior research. In particular, Upadhyay et al. (2025) found ResNet-18 to be the CNN model that showed the greatest sensitivity to the GT concepts

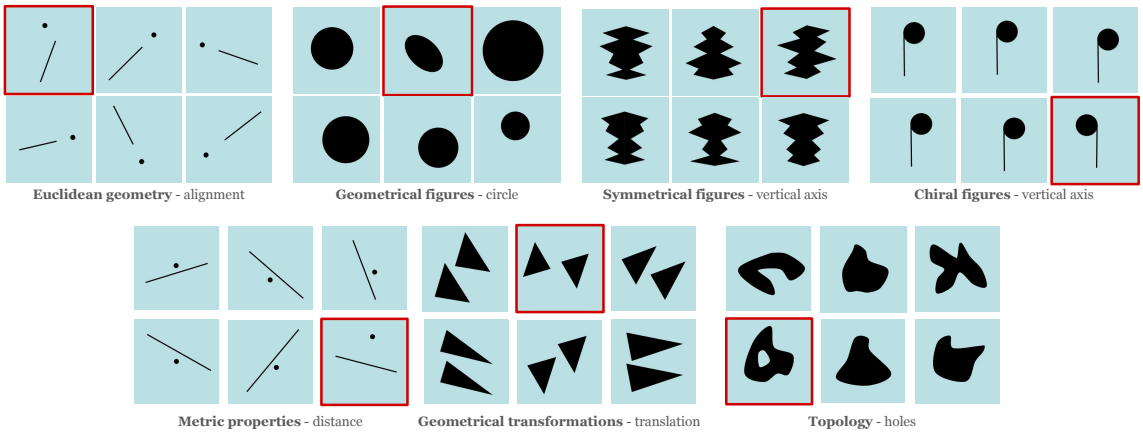

Figure 1: Example stimuli from the 7 categories from Dehaene et al. (2006). The odd-one-out is indicated by the red box.

of Dehaene et al. (2006), and Campbell et al. (2024) explored the geometric sensitivity of CLIP, DINOv2, and ResNet-50.

We use versions of these models that are publicly available on *Huggingface*. For a fair comparison to the CNN-based models, we used the "base" size variants of the Transformer and VLM models. All vision-only models (CNNs and Transformers) except for DINOv2 were trained on image classification tasks using supervised (image, class label) pairs. DINOv2 was trained in a self-supervised manner on the LVD-142M dataset (Oquab et al., 2024), a large-scale data set curated specifically for this model. It contains ImageNet-21k without requiring manual labels for its learning objective. Specifically, it uses a self-distillation loss on both image-level and patch-level representations, where a teacher model provides soft targets for a student model to learn robust representations. All CNN models were trained on ImageNet-1k, a subset of the much larger ImageNet-21k dataset used to train the Transformers. The VLM models were trained on the contrastive loss between (image, caption) pairs to maximize the similarity between the hidden representations of related images and captions. The training datasets for contrastive learning usually have to be much larger in size than those for learning image classification, and are not publicly available. Notably, the publicly available version of ALIGN was trained on a publicly available dataset, COYO (Byeon et al., 2022), and achieves comparable or superior performance to the original, non-public ALIGN model. Table 1 summarizes the key properties of the models and their training.

Stimuli and Datasets

We use the stimulus set from Dehaene et al. (2006). The set consists of 43 stimuli (or 'tasks'). Each stimulus corresponds to a GT concept (e.g., parallel lines). It is composed of 6 images, 5 of which embody that concept and 1 of which does not. The task is to select the odd-one-out image (see Figure 1). The 43 stimuli can be grouped into 7 categories (N): Topology (4), Euclidean Geometry (8), Geometrical Figures (9), Symmetrical Figures (3), Chiral Figures (4), Metric Properties (7), and Geometrical Transformations (8).

We obtained three human datasets from published studies. The first dataset comes from Dehaene et al. (2006), who measured the performance of Mundurucu adults and children. This is an indigenous Amazonian group whose members inhabit isolated communities and receive little or no formal education. (That study found comparable performance in the two age groups.) From Izard and Spelke (2009) we have the performance of Western children between the ages of 3 and 6, forming the second dataset, and from Western adults between the ages 18 and 25, forming the third dataset.

Type	Model	# Params	Dataset	# Training Data
Transformers	ViT	86M	ImageNet-21k	14M
	DINOv2	86M	LVD-142M	142M
CNN	ResNet-50	26M	ImageNet-1k	1.3M
	ResNet-18	11M	ImageNet-1k	1.3M
	EfficientNet	66M	ImageNet-1k	1.3M
VLM	CLIP-RN50	90M	WIT	400M
	CLIP-ViT	150M	WIT	400M
	ALIGN	174M	COYO	700M

Table 1: The eight models, their parameters, datasets, and training data sizes.

Evaluation on Neural Models

We followed the evaluation method of the earlier CNN study by Upadhyay et al. (2025); see also Campbell et al. (2024), Muttenthaler, Dippel, Linhardt, Vandermeulen, and Kornblith (2023), and Muttenthaler, Linhardt, et al. (2023). We first rescaled and cropped the 6 images in each stimulus to a size of 224×224. We then passed the images into the pre-trained models and collected their representations from the final hidden layer [1]. In the VLMs, the images were processed entirely by the image encoders. After collecting the image representations, we compute pairwise cosine similarities between the 6 images, and then for each image we computed its average cosine similarity to the other 5 images. The image with the lowest average cosine similarity was selected as the odd-one-out.

[1] The layer before any projection or classification.

We repeated this process for all stimuli and for all models.

Results

Overall sensitivity

For each model, we computed the overall accuracy averaged across the 43 concepts. Figure 2 shows these measures along with the overall accuracy in each of the human datasets. All but one of the models achieve higher overall accuracies than would be predicted by chance (16.67%), $zs > 2.3$, $ps < .02$; the only exception is ALIGN ($z = 1.187, p = .118$).

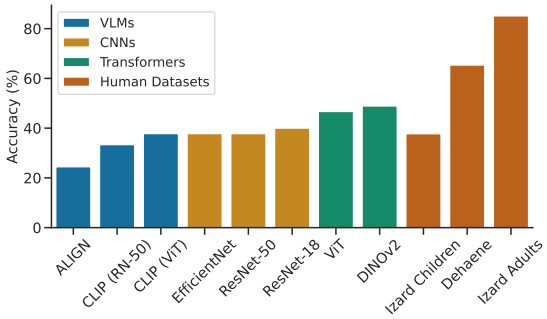

Figure 2: Average accuracy on the odd-one-out task for both the ML models and the human participants.

The Transformers outperform both the CNNs and VLMs, achieving 46.67% (ViT) and 48.89% (DINOv2). By comparison, the best-performing CNN, ResNet-18, achieves an accuracy of 40%, and the best-performing VLM, CLIP (ViT), achieves an accuracy of 37.78%.

Notably, only the Transformers show higher overall accuracies than the 3-6 year old children of Izard and Spelke (2009) (37.72%). However, their performance is still far below the 65.34% accuracy of the Mundurucu adults and children (Dehaene et al., 2006) and the 85.10% accuracy of the Western adults (Izard & Spelke, 2009).

The CNNs show comparable overall accuracies to the 3-6 year old children of Izard and Spelke (2009). Interestingly, the two deeper models (EfficientNet and ResNet-50) show slightly lower sensitivity to GT concepts (both 37.78%) than the shallower model (ResNet-18, 40%), paralleling the findings of Upadhyay et al. (2025).

The VLMs show the least sensitivity to GT concepts. The best VLM, CLIP (ViT), achieves an overall accuracy of 37.78%, while the ResNet-50 variant of CLIP, CLIP (RN-50), achieves an overall accuracy of only 33.33%. ALIGN is the worst-performing model, achieving an overall accuracy of only 24.44% despite being trained on the largest dataset.

Sensitivity by class

To further probe the sensitivities of the models, we evaluated their average accuracy for each of the 7 classes; see Figure 3. A number of notable patterns are visible at this finer-grain level of analysis.

First, the high overall accuracy of the Transformers is not driven by superior performance on one or two classes. Rather, ViT and DINOv2 outperform all of the other models on 6 of the 7 classes, with a few exceptions: The Transformers perform poorly on Symmetrical Figures concepts, and the CNN model ResNet-50 achieves a higher accuracy for the Geometric Transformations class.

Second, the Transformers do not just achieve higher overall accuracies than the 3-6 year old children of Izard and Spelke (2009) (see Figure 2). They are as accurate or more accurate than the children on all 7 of the classes. Conversely, they are less accurate than the two samples that include adult participants on all 7 of the classes.

Third, the Euclidean Geometry concepts are the easiest, with both models and humans achieving their highest accuracies for this class. Conversely, the Geometrical Transformations and Symmetrical Figures concepts are the most difficult, with especially the models achieving their lowest accuracies for these classes. The biggest discrepancy between the models and humans is on Symmetrical Figures, for which no model achieves greater than 33% accuracy. Also notable is that even the Transformers, the best-performing of all models, struggle with Symmetrical Figures. However, it must be noted that this class contains the fewest concepts, which make the data patterns here somewhat tentative.

Alignment to human profiles

That a model is highly accurate across the concepts of a class is not useful if, in fact, humans find that class to be difficult, given that our goal is to find alignment between model and human performance. To most directly evaluate this alignment, we compute the Pearson correlation (r) between the profile of accuracies across the 7 classes for each of the 8 models and the 3 human datasets; see Figure 4. There are a number of patterns to notice.

First, the class accuracy profiles of the Transformers achieve the closest alignment to those of the human datasets. This is particularly striking for the GT sensitivities of the 3-6 year old children in the Izard and Spelke (2009) study ($rs > 0.90$). Thus, the Transformers offer the best account of the human data at all three levels: overall accuracy, by-class accuracy, and by-class accuracy profile (i.e., human alignment).

Second, among the other models, the CNN model EfficientNet also achieves close alignment with the human datasets, particularly the adults in the Izard and Spelke (2009) study ($r = 0.86$). This raises the question of whether there is continuity in the development of GT concepts, i.e., given a model architecture, more parameters results in better alignment with children's performance, giving way to better alignment with adult performance. Or whether there is a qualitative shift across development, that is, a shift from a transformer-like architecture to a CNN-like architecture.

Third, the CLIP (ViT) model shows the worst alignment to the class accuracy profiles of adults, and the CLIP (RN5-50) model shows the worst alignment to the 3-6 year old children in the Izard and Spelke (2009) study.

Fourth, among the datasets, the ML models generally achieve closer alignment to the children's data of Izard and

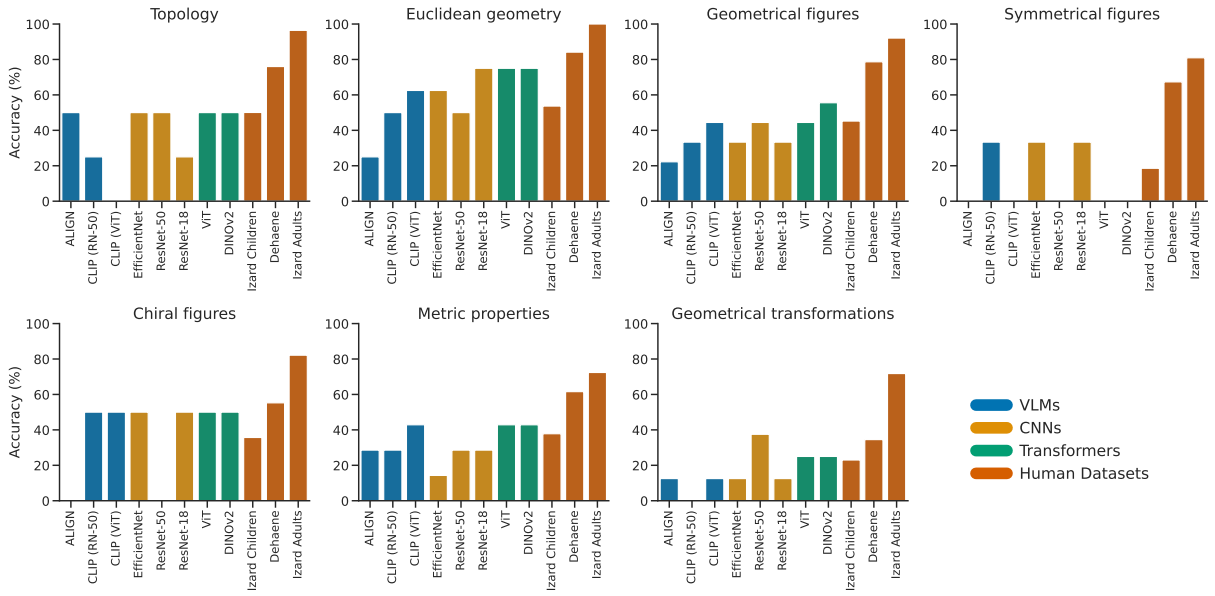

Figure 3: Accuracy profiles of the models and humans for each of the 7 classes of GT concepts.

Spelke (2009) than to the other two datasets, both of which contain adult data.

Discussion

Our results find that the Transformers show more promise as a cognitive science models of geometric and topological concepts than either CNNs or VLMs. At a coarse-grain level, they show the highest overall accuracies – higher even than the 3-6 year old children from the Izard and Spelke (2009) study; see Figure 2. At the finer-grain level of the 7 classes, their accuracy profiles most closely align with that of the 3-6 year old children from the Izard and Spelke (2009) study; see Figure 3 and Figure 4. An unexpected finding is that the VLMs trained with contrastive loss give both the worst overall and class-level performance, and also show the worst alignment to human profiles. This is surprising conceptually given the findings that support dual-coding theory (Paivio, 1991). This is also surprising technically: the sizes of the VLM models and the datasets on which they are trained are much larger than is the case for the Transformers and CNNs. This may suggest that current modality alignment techniques, such as maximizing the similarity between the hidden representation of both textual and visual information in a shared vector space, fail to fuse the two streams of information in a way which builds sensitivity to GT concepts, and that more human-like alignment methods are needed.

Transformer models show higher sensitivity to GT concepts

Figures 2 and 3 demonstrate that the Transformers, which have more parameters than CNNs and are trained on a larger dataset (e.g., ImageNet-21k with 14 million training samples compared to ImageNet-1k with 1.3 million training samples used for CNNs), exhibit higher sensitivity to GT concepts, and even surpass the overall accuracy of the 3-6 year old children from the Izard and Spelke (2009) study. These findings support the claim by Upadhyay et al. (2025) that GT concepts can be learned "for free" as a consequence of training on more generic visual processing tasks such as image classification. This contrasts with the "core knowledge" view that humans possess innate knowledge of mathematical concepts which requires minimal external input to activate (Spelke & Kinzler, 2007).

Transformers are generally larger than CNNs and trained on more data. Beyond this, we hypothesize that Transformers show higher sensitivity to GT concepts than CNNs for two reason. First, Transformers relax the constraint of transformation-invariants (Raghu, Unterthiner, Kornblith, Zhang, & Dosovitskiy, 2022). Second, their self-attention mechanism may learn both global and local object geometry whereas CNNs cannot (Cordonnier, Loukas, & Jaggi, 2020); the result is a more robust representation than CNNs. As a result, Transformers can acquire the sensitivity to object transformations and geometric/topological concepts. A direction for future research is to analyze the salience maps to each GT concept stimulus in Transformers and CNNs, to better understand the aspects that drive their sensitivities to GT concepts. For instance, one could collect the attention weights for each layer of a Transformer model and directly use them as the salience map. This is because the attention score tells the importance of current image patch with respect to other patches when making classification decisions. For CNNs, which do not possess attention mechanisms, we can compute the gradients of the target class score with respect to the feature maps of a convolutional layer, weighting these feature maps by the averaged gradients, and then summing them up (Selvaraju et al., 2019). A higher gradient score means that a small change in the feature map would cause a significant change in the

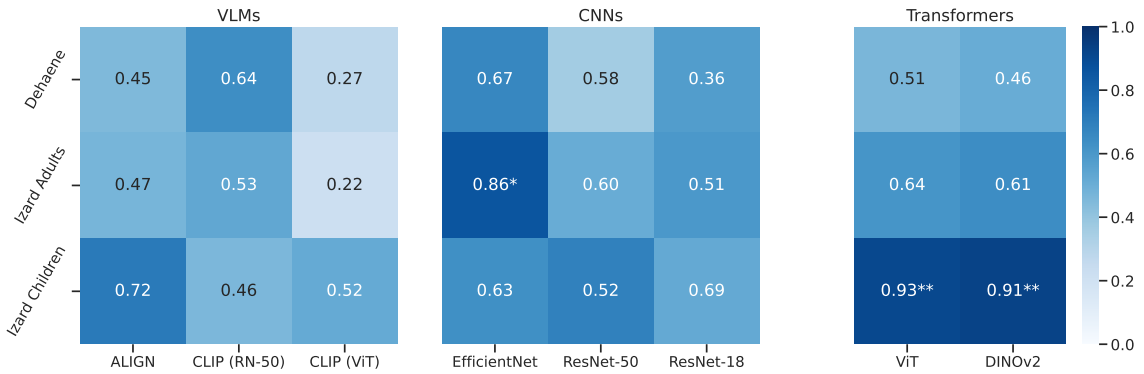

Figure 4: Heatmap of Pearson r coefficients between human participants' sensitivity profiles and the models' sensitivity profiles across the 7 classes of GT concepts. Note that * denotes $p < 0.05$ and ** denotes $p < 0.01$.

output, indicating that the feature map is highly relevant to the model's performance.

Modality alignment with text reduces the sensitivity to GT concepts

Humans process information through both verbal and non-verbal channels. Paivio (1991) argues that these channels work independently and interactively to enhance understanding and memory, and enable more robust learning because the information encoded in both formats provides multiple pathways for retrieval. Recent research in multi-modal alignment indeed is consistent with this view. For example, CLIP (Radford et al., 2021) significantly outperforms CNNs in image retrieval and classification tasks. It is therefore surprising, perhaps, that the CLIP models show lower sensitivity to GT concepts and worse alignment to human accuracy profiles compared to CNNs and Transformers. We argue that the effectiveness of current VLMs is limited by their core alignment strategy: optimizing similarity between image and caption representations. This approach struggles with inputs where the visual content is hard to describe naturally in text, such as the abstract synthetic images found in odd-one-out tasks. In these cases, VLMs fail to build robust representations that use both vision and text. Our findings (Figure 3) support this, showing reduced VLM sensitivity to abstract concepts like those of Topology and Geometrical Transformation compared to vision-only models – categories that are difficult to map textually onto specific pixels. In contrast, VLM performance on Metric Properties and Geometric Figures is less impacted, as descriptions involving distance, relative position, or basic shapes map more easily to pixel-level information. Conwell and Ullman (2022) also argues that failures in image generation guided by CLIP embeddings reveal a lack of a true compositional understanding of textual relations. Their results suggest that these models often reproduce statistically common image-caption pairings from training data rather than accurately constructing the specific relationship requested in the prompt, suggesting that CLIP models fail to capture representations for out-of-distribution text and image.

From cognitive alignment to developmental alignment

Transformers exceed the overall accuracy of the 3-6 year old children of the Izard and Spelke (2009) study (see Figure 2), and their performance profiles across the 7 classes of GT concepts are strongly correlated with those of the children (see Figure 4). In both respects, they exceed the cognitive alignment of the other evaluated models A key question is whether Transformer models show not just strong cognitive alignment but also strong developmental alignment. A key advantage of computational models over human participants is their inspect-ability and manipulate-ability. Future research can measure their sensitivity to GT concepts over training and evaluate whether this tracks the growing sensitivity shown by children over development. For example, we can track when during training a Transformer model becomes sensitive to shapes before transformations, and compare this ordering of unfolding sensitivities to that of the developing child. At a finer-grain level, we can analyze how representations of object parts, boundaries, or spatial relationships might evolve in the "developing" parameters of Transformers.

We can also conduct experiments on Transformers to test the causality of the learning account. A controlled, long-term ablation study in human participants – systematically withholding certain types of stimuli or training at particular times during development – would be both unethical and practically infeasible. By contrast, it is straightforward to perform parallel computational studies of computer vision models or study their development over different training "curricula". Such studies will potentially lead to novel hypotheses about *when* and *how* young children acquire specific cognitive competencies (Evanson, Lakretz, & King, 2023; Ma, Wang, & Chai, 2024). The strong alignment especially between the Transformer models and and the young children in their sensitivity to GT concepts suggests their potential for cognitive and developmental science.

References

Ahn, J., Verma, R., Lou, R., Liu, D., Zhang, R., & Yin, W. (2024). *Large language models for mathematical*

reasoning: Progresses and challenges. Retrieved from https://arxiv.org/abs/2402.00157

Byeon, M., Park, B., Kim, H., Lee, S., Baek, W., & Kim, S. (2022). *Coyo-700m: Image-text pair dataset.* https://github.com/kakaobrain/coyo-dataset.

Campbell, D., Kumar, S., Giallanza, T., Griffiths, T. L., & Cohen, J. D. (2024). *Human-like geometric abstraction in large pre-trained neural networks.* Retrieved from https://arxiv.org/abs/2402.04203

Chang, T. A., & Bergen, B. K. (2022). Word acquisition in neural language models. *Transactions of the Association for Computational Linguistics*, *10*, 1–16.

Chiandetti, C., & Vallortigara, G. (2007, July). Is there an innate geometric module? effects of experience with angular geometric cues on spatial reorientation based on the shape of the environment. *Animal Cognition*, *11*(1), 139–146. Retrieved from http://dx.doi.org/10.1007/s10071-007-0099-y doi: 10.1007/s10071-007-0099-y

Conwell, C., & Ullman, T. (2022). *Testing relational understanding in text-guided image generation.* Retrieved from https://arxiv.org/abs/2208.00005

Cordonnier, J.-B., Loukas, A., & Jaggi, M. (2020). *On the relationship between self-attention and convolutional layers.* Retrieved from https://arxiv.org/abs/1911.03584

De Cesarei, A., Cavicchi, S., Cristadoro, G., & Lippi, M. (2021, June). Do humans and deep convolutional neural networks use visual information similarly for the categorization of natural scenes? *Cogn. Sci.*, *45*(6), e13009.

Dehaene, S., Izard, V., Pica, P., & Spelke, E. (2006, January). Core knowledge of geometry in an amazonian indigene group. *Science*, *311*(5759), 381–384. Retrieved from http://dx.doi.org/10.1126/science.1121739 doi: 10.1126/science.1121739

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 ieee conference on computer vision and pattern recognition* (p. 248-255). doi: 10.1109/CVPR.2009.5206848

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... Houlsby, N. (2021). *An image is worth 16x16 words: Transformers for image recognition at scale.* Retrieved from https://arxiv.org/abs/2010.11929

Evanson, L., Lakretz, Y., & King, J.-R. (2023). *Language acquisition: do children and language models follow similar learning stages?* Retrieved from https://arxiv.org/abs/2306.03586

Gibson, J. J. (1979). The ecological approach to visual perception.. Retrieved from https://api.semanticscholar.org/CorpusID:33656271

He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Deep residual learning for image recognition.* Retrieved from https://arxiv.org/abs/1512.03385

He, Y., Wu, Y., Jia, Y., Mihalcea, R., Chen, Y., & Deng, N. (2023). *Hi-tom: A benchmark for evaluating higher-order theory of mind reasoning in large language models.* Retrieved from https://arxiv.org/abs/2310.16755

Hsu, J., Wu, J., & Goodman, N. D. (2022). *Geoclidean: Few-shot generalization in euclidean geometry.* Retrieved from https://arxiv.org/abs/2211.16663

Izard, V., & Spelke, E. S. (2009, January). Development of sensitivity to geometry in visual forms. *Hum. Evol.*, *23*(3), 213–248.

Jacob, G., Pramod, R. T., Katti, H., & Arun, S. P. (2021, March). Qualitative similarities and differences in visual object representations between brains and deep networks. *Nat. Commun.*, *12*(1), 1872.

Jaderberg, M., Simonyan, K., Zisserman, A., & Kavukcuoglu, K. (2016). *Spatial transformer networks.* Retrieved from https://arxiv.org/abs/1506.02025

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., ... Duerig, T. (2021). *Scaling up visual and vision-language representation learning with noisy text supervision.* Retrieved from https://arxiv.org/abs/2102.05918

Jin, M., Yu, Q., Huang, J., Zeng, Q., Wang, Z., Hua, W., ... Zhang, Y. (2024). *Exploring concept depth: How large language models acquire knowledge at different layers?* Retrieved from https://arxiv.org/abs/2404.07066

Jung, C., Kim, D., Jin, J., Kim, J., Seonwoo, Y., Choi, Y., ... Kim, H. (2024). *Perceptions to beliefs: Exploring precursory inferences for theory of mind in large language models.* Retrieved from https://arxiv.org/abs/2407.06004

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, & K. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 25). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper$_f$iles/paper/2012/file/c*Paper.pdf*

Laptev, D., Savinov, N., Buhmann, J. M., & Pollefeys, M. (2016). *Ti-pooling: transformation-invariant pooling for feature learning in convolutional neural networks.* Retrieved from https://arxiv.org/abs/1604.06318

Lindsay, G. W. (2021, September). Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of Cognitive Neuroscience*, *33*(10), 2017–2031. Retrieved from http://dx.doi.org/10.1162/jocn$_{a0}$1544 doi: 10.1162/jocn$_{a0}$1544

Ma, Z., Pan, J., & Chai, J. (2023). World-to-words: Grounded open vocabulary acquisition through fast mapping in vision-language models. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 524–544).

Ma, Z., Sansom, J., Peng, R., & Chai, J. (2024). *Towards a holistic landscape of situated theory of*

mind in large language models. Retrieved from https://arxiv.org/abs/2310.19619

Ma, Z., Wang, Z., & Chai, J. (2024). *Babysit a language model from scratch: Interactive language learning by trials and demonstrations.* Retrieved from https://arxiv.org/abs/2405.13828

Marupudi, V., & Varma, S. (2023, March). Graded human sensitivity to geometric and topological concepts. *Cognition*, *232*, 105331. Retrieved from http://dx.doi.org/10.1016/j.cognition.2022.105331 doi: 10.1016/j.cognition.2022.105331

Mumuni, A., & Mumuni, F. (2021, June). Cnn architectures for geometric transformation-invariant feature representation in computer vision: A review. *SN Computer Science*, *2*(5). Retrieved from http://dx.doi.org/10.1007/s42979-021-00735-0 doi: 10.1007/s42979-021-00735-0

Muttenthaler, L., Dippel, J., Linhardt, L., Vandermeulen, R. A., & Kornblith, S. (2023, May). Human alignment of neural network representations. In *11th international conference on learning representations, iclr 2023.* Kigali, Rwanda: OpenReview.net.

Muttenthaler, L., Linhardt, L., Dippel, J., Vandermeulen, R. A., Hermann, K., Lampinen, A., & Kornblith, S. (2023). Improving neural network representations using human similarity judgments. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in neural information processing systems* (Vol. 36, pp. 50978–51007). Curran Associates, Inc.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., ... Bojanowski, P. (2024). *Dinov2: Learning robust visual features without supervision.* Retrieved from https://arxiv.org/abs/2304.07193

Paivio, A. (1991, September). Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology / Revue canadienne de psychologie*, *45*(3), 255–287. Retrieved from http://dx.doi.org/10.1037/h0084295 doi: 10.1037/h0084295

Pasupathy, A., & Connor, C. E. (1999, November). Responses to contour features in macaque area v4. *Journal of Neurophysiology*, *82*(5), 2490–2502. Retrieved from http://dx.doi.org/10.1152/jn.1999.82.5.2490 doi: 10.1152/jn.1999.82.5.2490

Portelance, E. (2022). *Neural network approaches to the study of word learning*. Unpublished doctoral dissertation, Stanford University.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... Sutskever, I. (2021). *Learning transferable visual models from natural language supervision.* Retrieved from https://arxiv.org/abs/2103.00020

Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., & Dosovitskiy, A. (2022). *Do vision transformers see like convolutional neural networks?* Retrieved from https://arxiv.org/abs/2108.08810

Sable-Meyer, M., Ellis, K., Tenenbaum, J., & Dehaene, S. (2022). A language of thought for the mental representation of geometric shapes. *Cognitive Psychology*, *139*, 101527. doi: 10.1016/j.cogpsych.2022.101527

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2019, October). Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, *128*(2), 336–359. Retrieved from http://dx.doi.org/10.1007/s11263-019-01228-7 doi: 10.1007/s11263-019-01228-7

Shah, R. S., Marupudi, V., Koenen, R., Bhardwaj, K., & Varma, S. (2023). *Human behavioral benchmarking: Numeric magnitude comparison effects in large language models.* Retrieved from https://arxiv.org/abs/2305.10782

Shepard, R. N. (1994, March). Perceptual-cognitive universals as reflections of the world. *Psychonomic Bulletin amp; Review*, *1*(1), 2–28. Retrieved from http://dx.doi.org/10.3758/BF03200759 doi: 10.3758/bf03200759

Simonyan, K., & Zisserman, A. (2015). *Very deep convolutional networks for large-scale image recognition.* Retrieved from https://arxiv.org/abs/1409.1556

Spelke, E. S., & Kinzler, K. D. (2007, January). Core knowledge. *Developmental Science*, *10*(1), 89–96. Retrieved from http://dx.doi.org/10.1111/j.1467-7687.2007.00569.x doi: 10.1111/j.1467-7687.2007.00569.x

Tan, M., & Le, Q. V. (2020). *Efficientnet: Rethinking model scaling for convolutional neural networks.* Retrieved from https://arxiv.org/abs/1905.11946

Upadhyay, N., Marupudi, V., Varma, K., & Varma, S. (2025, February). Alignment of cnn and human judgments of geometric and topological concepts. In *Proceedings of the 39th annual aaai conference on artificial intelligence (aaai'25)* (pp. XXX–XXX). Philadelphia, PA.

Vallortigara, G. (2018, January). Comparative cognition of number and space: the case of geometry and of the mental number line. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1740), 20170120. Retrieved from http://dx.doi.org/10.1098/rstb.2017.0120 doi: 10.1098/rstb.2017.0120

Vemuri, S. K., Shah, R. S., & Varma, S. (2024). *How well do deep learning models capture human concepts? the case of the typicality effect.* Retrieved from https://arxiv.org/abs/2405.16128