Title

Perception and Reasoning with Visual Relations in Humans and Machines

Author

Fu, Shuhao

Publication Date

2025

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Perception and Reasoning with Visual Relations in Humans and Machines

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Psychology

by

Shuhao Fu

2025

ABSTRACT OF THE DISSERTATION

Perception and Reasoning with Visual Relations in Humans and Machines

by

Shuhao Fu

Doctor of Philosophy in Psychology

University of California, Los Angeles, 2025

Professor Hongjing Lu, Chair

The world consists of objects, and narratives are built from words. Rather than perceiving the world as a list of entities, humans represent the world in a more cohesive way by apprehending and expressing the *relations* between entities. Equipped with the ability to represent and process relations, human thinking and creativity are deeply rooted in *analogy*: the ability to identify and utilize resemblances based on relations between entities. Despite decades of research on relation perception and analogy, a fundamental question remains: *how do relational representations arise from linguistic and visual inputs?* This dissertation investigates the cognitive and computational mechanisms underlying relation perception and reasoning in humans, and examines the capacities of advanced AI models across a wide range of relation tasks. Through a combination of behavioral experiments, computational modeling, and AI model evaluation, this work bridges insights from cognitive science with innovations in deep learning.

Chapter 2 investigates relation perception in humans using both realistic and synthetic stimuli, and compares human performance with that of vision-only and vision-language models, highlighting key areas where current AI falls short in accounting for human relation perception. Chapter 3 systematically evaluates the capacity for relation understanding and compositionality in multimodal generative models, revealing foundamental limitations in their ability to ground spatial and agentic relations. Chapter 4 focuses on spatial relations in 3D

object recognition, and demonstrates that hierarchical abstraction mechanisms, commonly known as local-to-global visual processing, are crucial for enabling AI models to achieve human-like robustness for 3D object recognition. Chapter 5 examines visual analogy with realistic car stimuli, showing that a part-based comparison model more closely aligns with human reasoning performance than neural networks trained specifically on analogy tasks. Chapter 6 introduces VisiPAM, a vision-based probabilistic analogical mapping model that requires no analogy-specific training, yet best accounts for human performance in a novel mapping task.

By integrating experimental and modeling approaches, this work offers novel benchmarks, cognitively-inspired design principles, and empirical evidence that reveals both the promise and limitations of current AI systems in relational tasks. These findings establish a foundation for developing models with greater generalizability, interpretability, and alignment with human relation perception and reasoning.

The dissertation of Shuhao Fu is approved.

Keith Holyoak

Philip Kellman

Ying Nian Wu

Hongjing Lu, Committee Chair

University of California, Los Angeles

2025

To my parents,

Thank you for always believing in me, encouraging me through every challenge,

and giving me the foundation to pursue my dreams.

Contents

List of Figures

Acknowledgements

Chapter 2 is a version of the manuscript: Fu, Chen, & Lu (in preparation). Chapter 3 has been published in Fu, Lee, Wang, Momennejad, Bihl, Lu, & Webb (2025). Chapter 4 has been published in Fu, Kellman, & Lu (2025). Chapter 5 has been published in Ichien, Liu, Fu, Holyoak, Yuille, & Lu (2023). Chapter 6 is an updated version of Webb, Fu, Bihl, Holyoak, & Lu (2023). This dissertation research was supported by UCLA's Dissertation Year Award.

Vita

Education

University of California, Los Angeles (UCLA) 2020 – Present

Ph.D. Candidate in Psychology (degree expected 2025)

University of California, Los Angeles (UCLA) 2024 - 2025

M.S. in Statistics

University of California, Los Angeles (UCLA) 2020 - 2021

M.A. in Psychology

The Hong Kong University of Science and Technology 2014 - 2019

B.S. in Computer Science and Mathematics

Awards

Dissertation Year Award 2024

Graduate Summer Research Mentorship Fellowship 2021

HKUST University Scholarship 2014 - 2019

Dean's List Student 2014-17

Lee Hysan Foundation Exchange Scholarship 2017

HKSAR Government Scholarship Fund – Reaching Out Award 2017

Publications

Fu, S.*, Lee, A.*, Wang, A., Momennejad, I., Bihl, T., Lu, H., Webb, T. (2025). Evaluating compositional scene understanding in multimodal generative models. *Transactions on Machine Learning Research.*

Kang, A., Chen, J. Y., Lee, Z., Fu, S. (2025). Synthetic data generation with large language models for improved depression prediction. *Proc. 47th Annual Meeting of the Cognitive Science Society.*

Fu, S., Kellman, P., Lu, H. (2025). Hierarchical Abstraction Enables Human-Like 3D Object Recognition in Deep Learning Models. *Proc. 47th Annual Meeting of the Cognitive Science Society.*

Yun, Y., Chen, Y.-C., Fu, S., Lu, H. (2025). The emergence of latent force representation in human perception of social interactions. *Proc. 47th Annual Meeting of the Cognitive Science Society.*

Wu, J., Fu, S., Dale, R., Gao, T., Liang, J. (2025). Beyond emotion: Unraveling the limited role of sentiment in extended-format communication. *Proc. 47th Annual Meeting of the Cognitive Science Society.*

Webb, T. W.*, Fu, S.*, Bihl, T., Holyoak, K. J., Lu, H. (2023). Zero-shot visual reasoning through probabilistic analogical mapping. *Nature Communication 14, 5144.*

Ichien, N., Liu, Q., Fu, S., Holyoak, K. J., Yuille, A. L., Lu, H. (2023). Two computational approaches to visual analogy: Task-specific models versus domain-general mapping. *Cognitive Science*, 47(9), e13347.

Snefjella, B., Yun, Y., Fu, S., Lu, H. (2023). Human similarity judgments of emojis support alignment of conceptual systems across modalities. *Proc. 45th Annual Meeting of the Cognitive Science Society.*

Zhang, I., Fu, S., Xu, A., Lu, H. (2022). Causal versus associative relations: Do Humans perceive and represent them differently? *Proc. 44th Annual Meeting of the Cognitive Science Society.*

Fu, S., Holyoak, K. J., Lu, H. (2021). From vision to reasoning: Probabilistic analogical mapping between 3D objects. *Proc. 44th Annual Meeting of the Cognitive Science Society.*

Ichien, N. T., Liu, Q., Fu, S., Holyoak, K. J., Yuille, A. L., Lu, H. (2021). Visual analogy: Deep learning versus compositional models. *Proc. 43rd Annual Meeting of the Cognitive Science Society.*

Fu, S., Lu, Y., Wang, Y., Zhou, Y., Shen, W., Fishman, E., Yuille, A. (2020). Domain adaptive relational reasoning for 3D multi-organ segmentation. *MICCAI.*

Fu, S., Xie, C., Li, B., Chen, Q. (2020). Attack-resistant federated learning with residual-based reweighting. *AAAI Workshops: Towards Robust, Secure and Efficient Machine Learning.*

Dreizin, D., Zhou, Y., Fu, S., Wang, Y., Li, G., Champ, K., ... Yuille, A. L. (2020). A multiscale deep learning method for quantitative visualization of traumatic hemoperitoneum at CT: Assessment of feasibility and comparison with subjective categorical estimation. *Radiology: Artificial Intelligence, 2(6).*

Chapter 1

Introduction

The world is composed of objects, but humans represent it in a more structured manner—beyond simply recognizing objects, features, and tokens, we also apprehend and express the *relations* between them. These relational representations form the foundation of human perception and reasoning. If perception lacked relational representations, the visual world would appear as a fragmented collection of object lists rather than as a cohesive scene. For example, consider looking at a photo of a kitchen. Without relations, the visual system might perceive a disconnected list: fridge, boy, cabinet, sink, window. However, with relational representations, the system provides a much richer description: a boy is opening a fridge that is next to a cabinet; the cabinet is above the sink, which is beside a window. Simply reading this description with relational context makes it far easier to imagine the scene. This thought experiment highlights the critical role that relational representations play in perception, enabling us to organize and make sense of the world by interpreting it as interconnected, meaningful scenes, and also to form a "language of vision" to communicate with cognitive systems [22, 38, 57].

This relational capacity extends far beyond basic spatial arrangements. Infants demonstrate an early understanding of physical relations between objects, such as expecting that objects continue to exist when out of sight and that they cannot pass through one another,

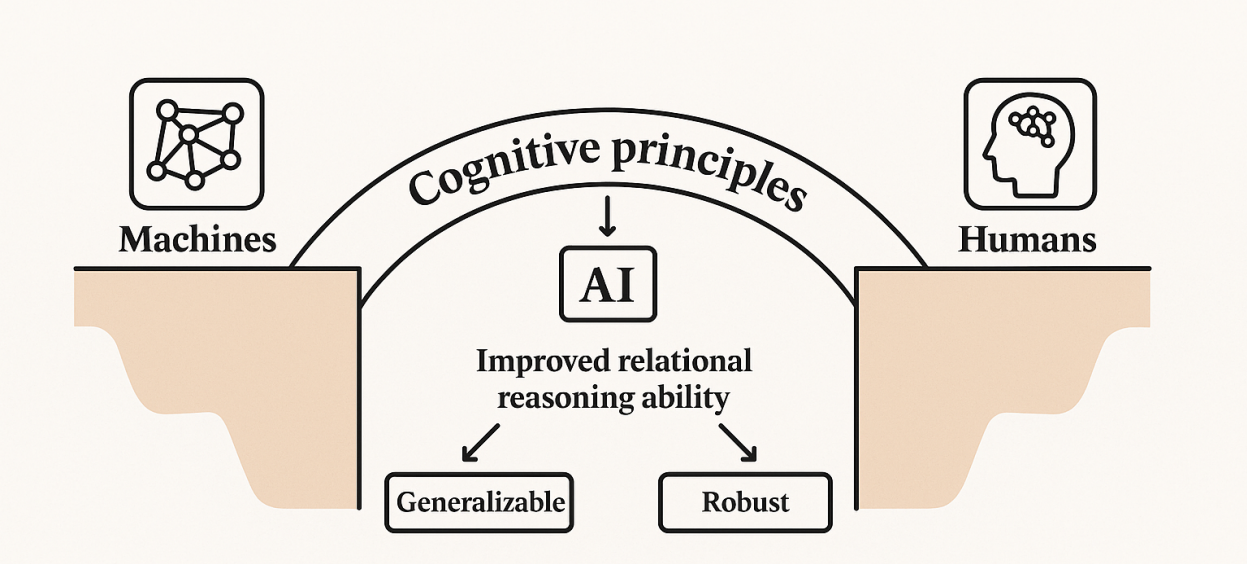


Figure 1.1: **Overview of the Dissertation Conceptual Framework.** While a significant gap remains between human and machine relational reasoning, *cognitive principles* can serve as a bridging framework. By incorporating these principles, AI systems can achieve more generalizable and robust relational reasoning capabilities, moving closer to the flexibility of human cognition.

indicating a foundational grasp of object relations in the physical world [5]. At the cognitive level of reasoning, relation representation is the core for *analogy*, a reasoning process widely recognized as central to human intelligence [134]. Humans can abstract complex relational structures from minimal input, as in analogies like "the atom is like a solar system"—despite enormous differences in physical scale and domain. Such analogical reasoning reflects a capacity to extract and manipulate structured representations that capture how parts relate to wholes, agents to objects, or entities to one another [47, 62].

Over the past decades, psychology research has largely focused on two ends of this spectrum running from perception to cognition. In vision research, studies have focused on relation representations specialized for visual processes and grounded in visual scenes. As Hafri and Firestone [55] aptly note, “relations cast no light onto our eyes." Relations themselves have no direct sensory signature; they must be inferred from sensory inputs. A growing body of work suggests that even very sophisticated relations display key signatures of automatic visual processing. Chen [28] demonstrated that participants could reliably detect topological relations such as “inside” within just 50 milliseconds of exposure, suggesting that

certain relations are computed automatically at early stages of visual processing. Similarly, Hafri et al. [56] found that people can accurately judge action-role relations (e.g., who is doing what to whom) in natural scenes with exposures as brief as 37 milliseconds. These findings highlight the perceptual fluency in detecting relations and importance of relational information in visual cognition.

Conversely, in reasoning research, relations are often treated as abstract concepts represented symbolically, typically through hand-designed inputs to computational models. This approach has shifted the focus to issues such as binding entities and relations into propositions, and finding mappings between relations during inference [37, 46, 50, 61, 62, 65]. However, I argue that both lines of research have sidestepped a critical issue: *how relational representations emerge from visual inputs*. While research on automatic visual processes sheds light on detecting relations, it does not explain how these representations interact with other processes to make humans selectively sensitive to certain relations. Similarly, while symbolic representations of relations offer advantages such as compositionality, they fail to address how such symbols are learned from experience and are grounded in sensory and linguistic inputs. Addressing this gap is essential to bridge perception and cognition in intelligent systems. Hence, new experimental paradigms and methodologies are needed to systematically investigate relation perception and reasoning in an integrated manner.

Meanwhile, the rise of artificial intelligence presents new opportunities and challenges for studying relation perception and reasoning. Despite rapid progress in computer vision and multimodal AI, current deep learning models continue to struggle with relational understanding. While convolutional neural networks (CNNs) and transformer-based architectures have reached human-level performance in object classification, segmentation, and even image captioning, they often falter when interpreting how objects relate to one another in a structured scene.

CNN-based models, for instance, can be misled by superficial context. In a now well-known example, Wang et al. [146] demonstrated that simply placing a motorbike or guitar under a

monkey in an image caused the model to classify the monkey as a human. Such mistakes suggest that the model was relying on statistical co-occurrence patterns, e.g., "humans tend to be on bikes", rather than computing explicit spatial or agentic relations. No human observer would make such an error, revealing a fundamental gap between pattern recognition and relational reasoning.

Transformer-based and multimodal large language models (MLLMs) such as CLIP [122], Flamingo [2], and GPT-4V [1] have made notable progress in aligning visual and linguistic information, yet they consistently struggle with precise relational understanding—particularly in spatial contexts. For example, these models often fail to distinguish whether a dog is under or on top of a table, revealing their inability to encode spatial configurations explicitly [73]. Recent benchmarks such as RevQA [25] and SpatialEval [147] further highlight this limitation, showing that state-of-the-art models like LLaVA-1.5 [92] can perform worse than chance when confronted with reversed spatial relationships. Similarly, studies using 3DAxiesPrompts [91] demonstrate that GPT-4V, while somewhat effective with 2D relations, struggles significantly with 3D spatial reasoning. Together, these findings suggest that current AI models are still limited in terms of the architectural capacity for structured relational reasoning, underscoring the need for models that can explicitly represent and manipulate spatial relations.

Critically, these limitations are not merely the result of data scarcity or model scale—they reflect deeper architectural gaps. Most vision or vision-language models are trained to minimize prediction error over static scenes or image-text pairs. They are not designed to represent roles, arguments, or structural regularities in ways that support compositional generalization. As a result, these models often struggle to apply relational knowledge to novel or unfamiliar configurations—something humans do with ease, even in zero-shot contexts.

To build truly intelligent systems, we must close the gap between human and machine relational reasoning. Humans possess a remarkable ability to perceive structured relations and apply them flexibly across diverse contexts—from understanding visual scenes to drawing analogies, inferring causal structure, and navigating social interactions. This relational

capacity is generalizable, compositional, and robust: we can effortlessly reason about new combinations of familiar elements, identify abstract roles and structures from sparse input, and apply prior knowledge to novel domains. Current models still lack the structured, hierarchical representations needed for true relational understanding. They often process scenes as unstructured collections of features, with little sensitivity to the global or role-based organization that underlies human perception.

This dissertation is driven by a central question: *how can we develop computational models that not only perceive visual relations as humans do, but also reason with them in a structured, flexible, and generalizable way*? To address this question, the work advances two complementary research directions. The first direction focuses on understanding how humans extract and reason with various visual relations—such as spatial, agentic, and part-whole relations—and systematically compares this capacity with that of modern machine learning systems. Despite the rapid progress in deep learning, there remains a substantial gap between human relational reasoning and the capabilities demonstrated by current AI models. Through behavioral experiments and model evaluations, this research quantifies this gap and identifies the conditions under which AI systems diverge from human performance. The second direction aims to bridge this gap by drawing on cognitive principles, particularly the use of structured object representations, to inform the design of more relationally competent AI systems. This includes leveraging part-whole hierarchies, compositional structure, and analogical mapping mechanisms that support generalization in human reasoning. By embedding these principles into computational models, the work explores how cognitive insights can lead to more interpretable and human-aligned relational inference in artificial systems, as illustrated in Figure 1.1.

These two research directions are developed across five chapters. Chapters 2 and 3 correspond to the first direction. Chapter 2 investigates relation perception in humans using both naturalistic and synthetic visual stimuli, revealing key differences between human judgments and those of vision-only and vision-language models. Chapter 3 systematically

evaluates multimodal generative models, such as DALL·E [12], showing that these systems struggle to accurately ground spatial and agentic relations.

Chapters 4 to 6 develop the second research direction. Chapter 4 focuses on 3D object recognition and demonstrates that hierarchical abstraction—processing visual input from coarse to fine—is essential for achieving human-like relational robustness in machine learning models. Chapter 5 presents a study of visual analogy using realistic car stimuli, showing that part-based comparison models more closely align with human performance than neural networks trained directly on analogy tasks. Chapter 6 introduces VisiPAM, a vision-based probabilistic analogical mapping model that requires no analogy-specific training, yet best accounts for human behavior in a novel relational mapping task. These models illustrate how cognitive principles can guide the development of more generalizable and interpretable AI systems.

Together, these investigations contribute to research in both cognitive science and artificial intelligence by advancing our understanding of perception and reasoning with visual relations. From a cognitive perspective, the dissertation provides new empirical insights into how humans perceive and generalize relational structures—including spatial, agentic, and part-whole relations—and offers formal models that capture key aspects of this flexibility. From an AI perspective, it develops systematic benchmarks for evaluating relational understanding in AI models, identifies limitations in current vision and multimodal systems, and highlights cognitively motivated design principles to support more robust and interpretable relational inference. Overall, this work bridges the gap between human and machine reasoning, providing a foundation for future models that better reflect the structured and generalizable nature of relation processings in human perception and cognition.

Chapter 2

Visual Relation Detection in Humans and Deep Learning Models

2.1 Introduction

When we look around, we don't just see individual objects like cups, tables, and chairs in isolation. Our visual system automatically processes the spatial relationships between these objects – recognizing that cups are on top of tables while chairs are below tables. This ability to detect visual relations in everyday scenes and images apppears to be effortless, but this remarkable skill should not be taken for granted.

Understanding visual relations presents a unique challenge for our perceptual system. As noted by [55], "Relations themselves cast no light onto our eyes." Unlike individual objects that have physical properties and attributes like color, texture, and shape that directly stimulate our visual receptors, spatial relations between objects and agentic relations between agents and objects exist as mental construcst that must be computed by our brain. Thus, our visual system must go beyond simply recognizing what objects are present in the scenes, it must also recognize how these objects relate to each other in both space and time.

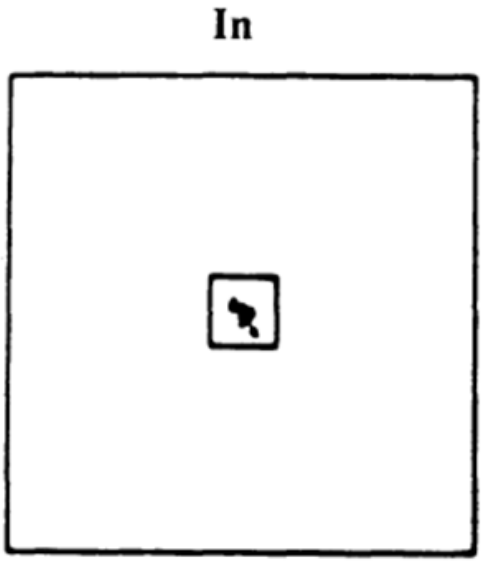
Previous research suggests that humans can detect visual relations remarkably fast in

images. Studies have shown that people can process basic spatial relationships such as containment (inside vs outside) in under 50 milliseconds [28], while more complex physical relations, including stability and support, can occur within 50-100 milliseconds [38]. Even sophisticated eventive relations, such as identifying who acted upon whom in dynamic scenes, can be extracted after masked exposures of just 37 milliseconds [56]. Social relations are equally rapid, with human social interactions being perceived in less time than it takes to make a single eye movement [69]. This rapid processing occurs within the first 500 milliseconds before higher-level object recognition can interfere [170], suggesting that relational perception operates as an automatic, early-stage visual process.

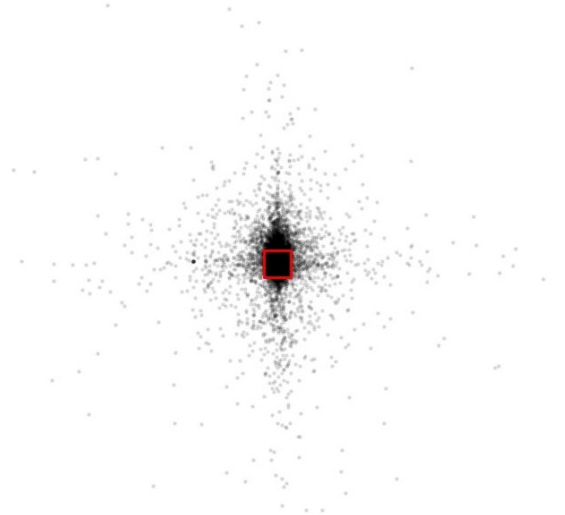
The ability to perceive visual relations represents an essential but intricate achievement of human vision that fundamentally shapes how we understand our environment. Rather than simply detecting isolated objects, our visual system constructs a rich representation of how elements relate to and interact with one another [55]. This capacity underlies critical cognitive abilities—from spatial navigation to social understanding—enabling us to predict physical events, interpret social interactions, and make rapid decisions about our surroundings. Understanding the mechanisms of relational perception thus provides crucial insights into both the sophistication of human vision and its role in higher-level cognition.

Yet this richness also poses a challenge when we try to formally define or annotate visual relations. For instance, how should we represent the relation “in” in a way that captures both its geometric regularities and its semantic variability? Lab-based studies such as [97] suggest that people hold stable spatial templates for basic relations (see Figure 2.1, top-left). However, real-world datasets like Visual Genome reveal a much broader range of usage: the same relation label may correspond to diverse spatial configurations (top-right). As illustrated in the bottom row of Figure 2.1, examples like “child in shoe” or “paper in hand” show that natural language descriptions of relations are often flexible, metaphorical, or even ambiguous.

Even with the recent advancements in artificial intelligence, visual relation detection remains a significant challenge for machine learning models. Visual relations, such as "man



Idealized spatial configuration from lab-based production task [97].

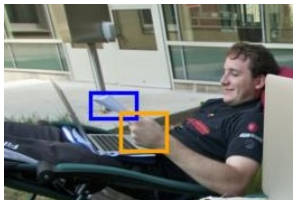


Empirical spatial distribution of the relation "in" from Visual Genome dataset [81].

Man in hat.

Child in shoe.

People in mountain.



Paper in hand.

Figure 2.1: Comparison of the spatial relation "in" across controlled and real-world contexts. (**Top left**) Spatial "template" from [97], derived from a dot-placement task. Participants were asked to indicate where a marker belongs when describing "in" relative to a central square. (**Top right**) Empirical distribution of "in" relations from the Visual Genome dataset [81]. Each black dot denotes the annotated center of a subject entity involved in an "in" relation, with the red square marking the normalized location of the reference object. The distribution reveals substantial variability, reflecting the ambiguity and looseness of spatial relation in natural images. (**Bottom**) Example images from Visual Genome demonstrating the diversity in how "in" is annotated in practice.

riding a horse" or "cup on table", are not directly observable entities but abstract constructs inferred from configurations of objects and their interactions (see detailed examination in Chapter 5). Traditional approaches to visual relation detection follow a pipeline paradigm using supervised training: first detecting individual objects using object detection models (e.g., Faster R-CNN; [129]), and then classifying the relations between pairs of objects based on their spatial features and semantic embeddings [100, 163, 167]. These models typically rely on bounding box information and object category labels, treating relation detection as a supervised classification problem [163]. Some models further incorporate external knowledge by embedding object and predicate labels into language-informed semantic spaces to benefit from advancements in natural language processing [100, 167].

More recently, multi-modal large language models (MLLMs) have demonstrated remarkable capabilities in reasoning about complex visual scenes (e.g., CLIP [122], Flamingo [2], LLaVA [93]). These models are pretrained on large-scale datasets containing images paired with captions, question-answer pairs, and other vision-language tasks, which allow them to develop a shared semantic space between text and vision. Crucially, these models do not treat relation understanding as a standalone detection task. Instead, they formulate vision-language understanding as a generative or contrastive task, where the model learns to predict relevant tokens or align embeddings, conditioned on both visual input and textual prompts. This approach aligns with the philosophical view that "relations themselves cast no light onto our eyes" [55], suggesting that relational understanding is not a perceptual primitive but a cognitive inference constructed from prior knowledge and contextual reasoning. In MLLMs, the ability to "understand" visual relations emerges as a byproduct of generalized multi-modal reasoning, rather than as a narrowly supervised task.

In this chapter, we explore two central research questions: (1) How does the human visual system extract visual relations in images? and (2) To what extent do deep learning models—especially those integrating both visual and conceptual knowledge—account for human performance in visual relation detection?

To address the first question, we conducted two complementary behavioral experiments designed to probe the temporal dynamics of relational perception. In a Temporal Order Judgment (TOJ) task, participants viewed two images flashed in rapid succession and were asked to judge which image appeared first. Critically, we manipulated whether the two images depicted the same or different visual relations. Participants were significantly more accurate when the images conveyed different relations, even when display times were as brief as 80 milliseconds. This suggests that relational information can be extracted very early in visual processing, allowing participants to reliably distinguish relation types under severe temporal constraints.

We further validated this finding using a Go/No-Go paradigm, where participants viewed a single image for 67 milliseconds and were instructed to press a key only if a specified relation (e.g., carrying, riding) was present. Across a range of relations, participants consistently performed above chance, demonstrating that relational cues are available within the first feedforward sweep of visual processing. These results support the notion that the human visual system encodes relational information rapidly and perhaps in parallel with object recognition, consistent with prior work showing early access to scene-level semantics [51, 72].

To examine how well computational models align with human relational perception, we analyzed the correspondence between human accuracy and model-derived similarity measures. Our underlying assumption is that accuracy reflects perceived similarity: participants are more likely to confuse image pairs that are conceptually or perceptually similar, leading to lower accuracy in differentiating them. Conversely, high accuracy suggests greater dissimilarity between the two relational scenes. Thus, by correlating human accuracy with model-computed distances between image or text embeddings, we can infer how closely a model's internal representation space mirrors human relational judgments.

We examined a diverse set of models that span different architectures and training regimes. These include language models like Sentence-BERT [128], which encode textual descriptions of the images; vision-only models such as ResNet-50 [58] and Vision Transformer (ViT) [36],

which are trained on object classification tasks; visual relation detection models like HL-Net [90] and VS3 [168], which are explicitly optimized to identify object relationships in images; and multi-modal vision-language models such as CLIP [122], InternVL-2.5 [30], and Qwen2-VL [148], which align visual and linguistic representations through contrastive or generative pretraining objectives on large-scale image-text pairs.

By computing cosine distances between model embeddings and correlating these with human performance across image pairs, we observed that ResNet-50, a model trained only for object recognition, showed the highest alignment with human data. This finding suggests that relational information, particularly simple spatial relations such as above or next to, may be implicitly encoded in visual representations without explicit supervision. In contrast, models trained directly for relation detection, while performing well on benchmark tasks, did not capture the relational similarity structure evident in human judgments. This discrepancy highlights a fundamental limitation in current model training objectives: explicitly classifying relation labels may not foster representations that generalize in a human-like way. Meanwhile, foundation models that learn from broad image-text corpora demonstrated intermediate alignment, likely due to their ability to integrate contextual and semantic cues from both modalities.

Taken together, these findings suggest that human-like relational understanding may emerge not through narrowly defined supervision, but through broader exposure to structured visual and linguistic contexts—a hypothesis that offers new directions for developing models that better mirror the way humans perceive and reason about visual relations.

2.2 Models

We compared eight distinct models against human performance, spanning multiple modalities and architectures: one language model (Sentence-BERT [128]), two vision models (ResNet50 [58] and ViT [36]), two visual relation detection models (HL-Net [90] and VS3 [168]),

Model type	**Model**	**Training data**	**Training task**
Language model	Sentence-BERT [128]	Text	Text prediction
Vision model	ResNet50 [58] ViT [36]	ImageNet	Object recognition
Visual relation model	HL-Net [90] VS3 [168]	Visual Genome	Relation recognition
Vision-language model	CLIP [122]	Image & text captions	Cross-modal alignment
	InternVL2.5 [30] Qwen2-VL [148]	Mixed	Text prediction about images

Table 2.1: Summary of model categories, associated architectures, training datasets, and primary training objectives used in this study.

and three vision-language models (CLIP [122], InternVL-2.5 [30], and Qwen2-VL [148]). A summary of the models, their training data, and associated tasks is provided in Table 2.1.

The language model component consisted of Sentence-BERT [128], a framework for computing dense vector representations of sentences and text that is optimized for semantic similarity tasks through siamese network fine-tuning. For vision-only models, we employed ResNet50 [58], a 50-layer residual convolutional neural network that uses skip connections to enable training of very deep networks and is widely used for image classification and feature extraction, alongside ViT [36], a transformer architecture adapted for computer vision that treats images as sequences of patches and applies self-attention mechanisms instead of convolutions.

Our evaluation included two specialized visual relation detection models with distinct capabilities. HL-Net [90] is a hierarchical layout network designed to detect and classify spatial relationships between objects in images through structured scene understanding within a closed-vocabulary setting. In contrast, VS3 [168] leverages pre-trained visual-semantic spaces to enable open-vocabulary object detection and language-supervised scene graph generation, allowing it to recognize novel objects (but not novel relations) beyond its training data through region-word alignment and text prompt grounding.

Finally, we incorporated three state-of-the-art vision-language models. CLIP [122] employs contrastive learning to jointly train vision and language encoders for understanding

relationships between images and natural language descriptions. InternVL-2.5 [30] represents a large-scale vision-language foundation model designed for multimodal understanding tasks, combining visual perception with language comprehension. Qwen2-VL [148] is a vision-language model from the Qwen series that integrates visual and textual understanding for comprehensive multimodal reasoning and generation.

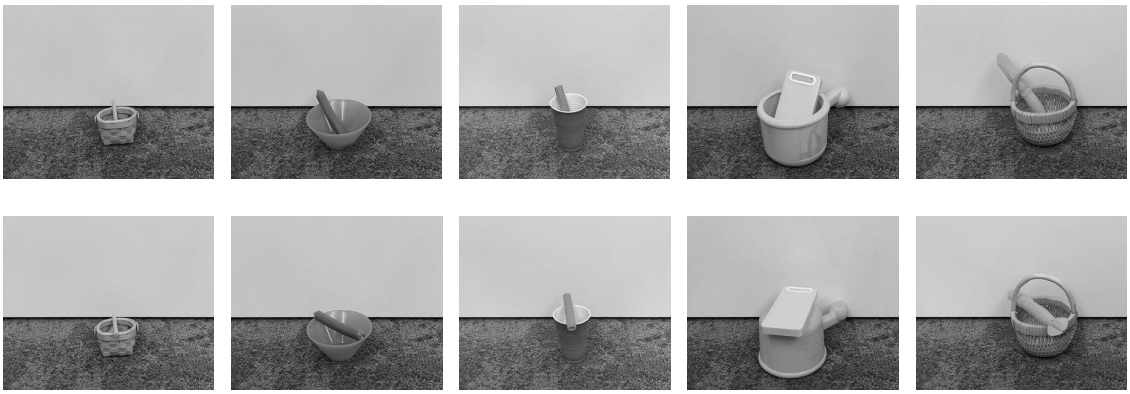
2.3 Experiments

2.3.1 Experiment 1

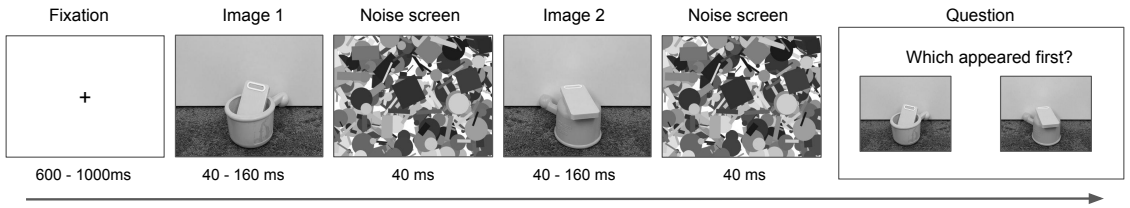
In Experiment 1, we investigated whether the rapid extraction of spatial relational information can influence image discrimination. We hypothesized that when two images depict different spatial relationships (e.g., containment vs. support), participants would find it easier to distinguish their temporal order compared to when both images depict the same spatial relationship. This advantage would suggest that relational information is extracted automatically and rapidly, facilitating the discrimination process even in a task that does not explicitly require attention to spatial relations. To test this hypothesis, we employed a temporal order judgment paradigm in which participants viewed pairs of images depicting objects in either containment or support relationships, presented at various brief durations.

Stimuli

We adopted 5 pairs of images from the stimuli used in Hafri et al.'s experiment [54] as shown in Figure 2.2a, where each pair contained the same two objects arranged in either a containment or support relation. In the "contain" images, one object was positioned inside the other (e.g., a stick in a cup), while in the "support" images, one object was placed on top of the other (e.g., a stick resting on top of a cup). To create the experimental conditions, we manipulated these base images as follows: First, we horizontally flipped each image to create mirror versions. To ensure visual consistency across flipped versions, we rotated images as needed so that



Stimuli used in the Experiment 1.



Experiment 1 procedure.

Figure 2.2: (a) Example stimuli used in the experiment, adapted from Hafri et al. [54]. (b) Schematic of the experimental procedure in Experiment 1.

reference lines (such as horizontal surfaces) maintained the same orientation. This process yielded four relational conditions for each object pair: contain/contain, support/support, contain/support, and support/contain, where the first term indicates the relation in the first image and the second term indicates the relation in the second image. The contain/contain and support/support pairs were classified as "same relation" conditions, as both images within each pair depicted the same spatial relationship. The contain/support and support/contain pairs were classified as "different relation" conditions, as the two images within each pair depicted different spatial relationships.

Procedure

Participants were instructed to maintain fixation on a central fixation point at the beginning of each trial. Each trial began with the presentation of a fixation cross for a randomly varied

duration of 600-1000 ms to prevent temporal predictability. The experimental sequence then proceeded as follows: the first target image was presented for a predetermined duration, immediately followed by the second target image for an equal duration. Display durations were manipulated across trials at four levels: 40, 80, 120, or 160 ms. Each target image was immediately followed by a 40 ms noise mask to eliminate afterimage effects and prevent backward masking.

The complete trial sequence consisted of: (1) fixation cross (600-1000 ms), (2) first target image (40-160 ms), (3) first noise mask (40 ms), (4) second target image (40-160 ms), and (5) second noise mask (40 ms). The total stimulus presentation time ranged from 160 ms (for 40 ms display conditions) to 400 ms (for 160 ms display conditions), as shown in Figure 2.2b.

Following each stimulus sequence, participants made a temporal order judgment by indicating which of the two target images appeared first. Response options were presented simultaneously, and participants responded using the left and right arrow keys on the keyboard to select the image that was shown first. No time limit was imposed for the response phase.

Display durations were systematically varied across trials to examine the effect of presentation time on temporal order discrimination accuracy. Stimuli served as the first and second target images across different trials. Additionally, the spatial arrangement of response options was counterbalanced, with each image appearing in both the left and right positions on the choice screen to control for potential spatial response biases.

The experiment consisted of 320 trials in total, organized around 5 pairs of stimuli representing different spatial relationships. The stimulus pairs were categorized into four relational types: contain/contain, contain/support, support/contain, and support/support. The experimental design followed a fully crossed factorial structure with the following factors: 5 stimulus pairs $\times$ 4 relational types $\times$ 2 presentation orders (which image appeared first) $\times$ 2 spatial arrangements (which image appeared on the left in the response screen) $\times$ 4 display durations (40, 80, 120, 160 ms). The experiment was divided into four blocks of 80 trials each, with mandatory rest breaks between blocks. Participants could resume the experiment

at their discretion by pressing any key.

Participants

Participants were excluded from analysis based on the following criteria: (1) response bias, defined as selecting the same side (left or right) on more than 15 consecutive trials, indicating inattentiveness or non-task-related responding; (2) excessively fast responses, defined as more than four trials within any single block with reaction times below 300 ms, suggesting anticipatory or non-deliberative responses; (3) excessively slow responses, defined as more than one trial exceeding 120 seconds across the entire experiment, indicating task disengagement; (4) technical difficulties that compromised data integrity; (5) self-reported failure to follow task instructions during the post-experiment debriefing; and (6) failure to complete the post-experiment debriefing or survey.

A total of 37 participants were recruited through the university's subject pool. Of these, nine participants were excluded based on the criteria above: four for not completing the post-experiment survey, one for exhibiting a response bias (i.e., selecting the same response for more than 15 consecutive trials), three for having more than four responses under 300 ms within a single block, and one for correctly guessing the purpose of the experiment during the debriefing. This resulted in a final sample of 28 participants (18 female, 8 male, 1 non-binary, and 1 who preferred not to disclose their gender; M = 20.36 years, SD = 1.13) included in the data analysis.

Results

Figure 2.3 shows recognition accuracy a function of display duration for images included the same and different relations. A 2 (relation type: same vs. different) $\times$ 4 (display duration: 40, 80, 120, 160 ms) repeated-measures ANOVA revealed significant main effects for both factors.

As expected, there was a significant main effect of display duration, $F(3, 81) = 132.78$, p

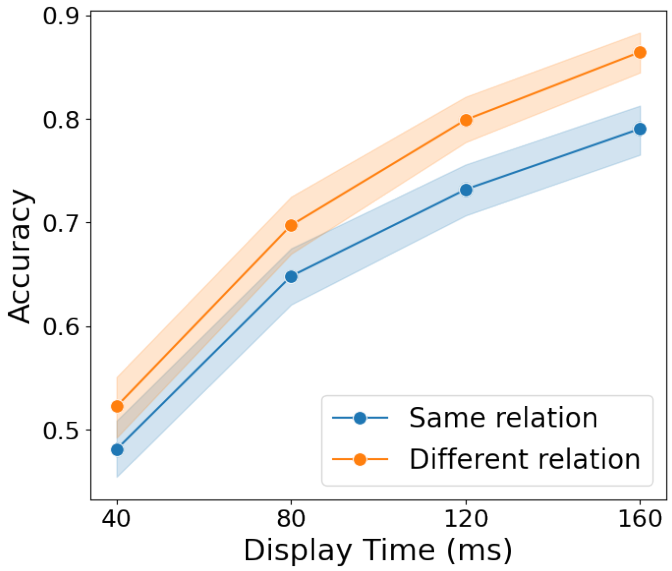


Figure 2.3: Image recognition accuracy as a function of display duration when the two target images share the same relations or show different relations. Participants showed higher recognition accuracy for stimulus pairs with different spatial relationships (orange line) compared to pairs with the same spatial relationship (blue line) across all display durations. Shaded areas represent 95% confidence intervals.

$<$ .001, with accuracy increasing systematically from 40 ms (M = 0.518, SE = 0.011) to 160 ms (M = 0.835, SE = 0.014).

More importantly, a significant main effect of relation was observed, $F(1, 27) = 31.07$, $p <$ .001, with participants demonstrating higher accuracy for stimulus pairs with different spatial relationships (M = 0.735, SE = 0.016) compared to pairs with the same spatial relationship (M = 0.677, SE = 0.015). This finding indicates that participants were able to rapidly extract and utilize spatial relational information to facilitate temporal order discrimination, even under conditions of brief stimulus presentation.

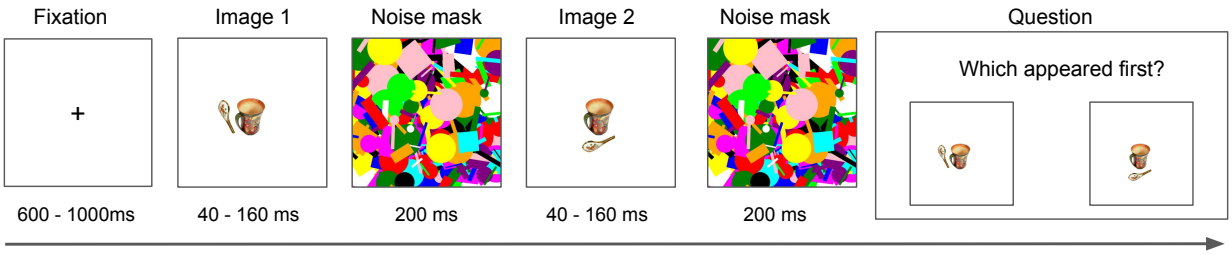
The interaction effect between display duration and relation was not significant, $F(3, 81) = 0.91$, $p = .438$, since the relation advantage remained consistent across all display durations, including the shortest 40 ms condition (same: M = 0.500, SE = 0.015; different: M = 0.537, SE = 0.014). This consistent effect across presentation times provides converging evidence that relation processing occurs rapidly and automatically. These results are consistent with previous finding that humans can recognize and utilize spatial relationships within very brief

exposure durations [28].

2.3.2 Experiment 2

In Experiment 2, we extended our investigation using a larger set of synthetically generated images to probe whether relational distinctions alone, independent of semantic content or scene context, can facilitate temporal order judgments. By precisely controlling low-level visual features, we aimed to isolate the contribution of spatial relational structure to perceptual performance. If relational perception operates independently of semantic context, we expected to observe a similar relational advantage as in Experiment 1. In addition, we leveraged the expanded stimulus set to compare human accuracy with similarity estimates from a range of computational models, allowing us to assess how well different model architectures capture human-like relational sensitivity.

(a) Example synthetic stimuli used in Experiment 2.



(b) Schematic of the experimental procedure in Experiment 2.

Figure 2.4: Overview of Experiment 2. Panels (a) and (b) correspond to the stimuli and procedure, respectively.

Stimuli

We constructed a synthetic image dataset using object images from the Big and Small Objects dataset [79], which contains real-world objects categorized by their typical physical size. From this resource, we selected 16 pairs of semantically related objects (e.g., cup–spoon, coffee–bagel). Each object was manually cropped to remove the original background and placed on a uniform canvas to form two spatial relations: "next to" and "above." In both relations, one object served as the fixed anchor, and the second object was positioned either adjacent to (left/right) or vertically aligned (above/below) without any overlap or occlusion. Example stimuli are shown in Figure 2.4a.

For each object pair, we generated four unique images: two depicting the "next to" relation (object on the left or right of the anchor) and two depicting the "above" relation (object above or below the anchor). This resulted in a total of $4 \times 16 = 64$ images.

The motivation for constructing this synthetic dataset was twofold. First, it allowed us to scale up the number of image pairs relative to Experiment 1, supporting more robust comparisons with model-based similarity estimates. Second, it provided greater experimental control by standardizing low-level visual features—such as background, alignment, and spatial frequency—across all conditions, thereby isolating the effect of relational structure.

Procedure

The procedure was largely identical to that of Experiment 1, with two key modifications. First, the grayscale mask image was replaced with a colored version, unrelated to the stimulus hue, to maintain consistency with the use of colored stimuli. Second, to minimize the perception of apparent motion arising from the rapid succession of static images, the mask was presented for a fixed duration of 200 ms. A schematic overview of the modified procedure is shown in Figure 2.4b.

Participants

Participant recruitment and exclusion criteria were identical to those used in Experiment 1. A total of 25 participants were recruited from the university's subject pool, and none were excluded from analysis. The final sample included 23 females and 2 males, with a mean age of 20.08 years (SD = 1.04).

Results

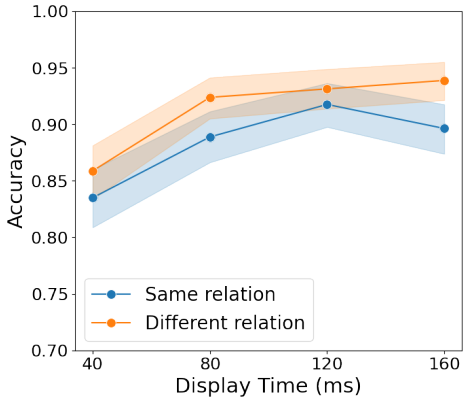
Participants' accuracy in the temporal order judgment task was analyzed using a two-way repeated-measures ANOVA with display duration (40, 80, 120, 160 ms) and relation type (same vs. different) as within-subject factors. Results are summarized in Figure 2.5a.

There was a significant main effect of display duration, $F(3, 72) = 9.79$, $p < .001$, indicating that longer exposure times led to higher accuracy. Accuracy increased from 0.847 ($SE = 0.015$) at 40 ms to 0.918 ($SE = 0.011$) at 160 ms. The analysis also revealed a significant main effect of relation type, $F(1, 24) = 11.62$, $p = .0023$. Participants performed better on trials where the spatial relation differed between the two images ($M = 0.913$, $SE = 0.009$) compared to trials with the same relation ($M = 0.884$, $SE = 0.009$), suggesting that relational changes facilitated temporal order discrimination. The interaction between display duration and relation type was not significant, $F(3, 72) = 0.93$, $p = .431$, indicating that the relational advantage was consistent across all durations.

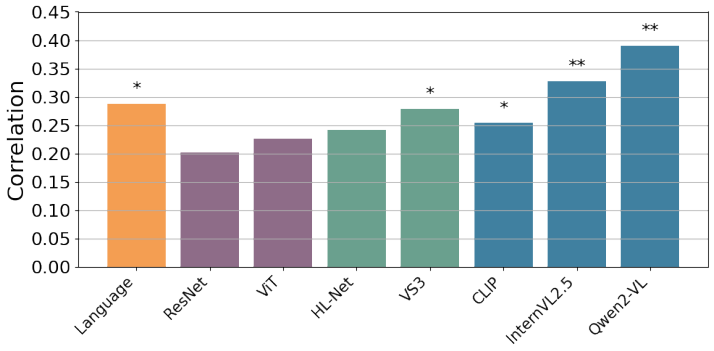
These results replicate the core findings of Experiment 1 using synthetic stimuli: both increased display time and changes in spatial relation independently contribute to more accurate temporal order judgments.

Model–Human Similarity Analysis

To evaluate the extent to which different models capture human-like sensitivity to relational structure, we compared model-based similarity estimates with human temporal order judgment accuracy at the longest display duration (160 ms). At this duration, visual encoding is



Human recognition accuracy.



Model-human correlation at 160 ms.

Figure 2.5: Experiment 2 results. (a) Accuracy in the temporal order judgment task as a function of display duration and relation type. Participants' accuracy increased with display duration, but relational type had no significant effect. Shaded regions indicate 95% confidence intervals. (b) Spearman correlation between model-estimated image similarity and human accuracy at 160 ms. Bars represent individual models, color-coded by modality: language (orange), vision-only (purple), relation detection (green), and vision-language (blue). Higher correlations reflect stronger alignment with human relational sensitivity.

presumed to be nearly complete, allowing human performance to primarily reflect higher-level relational distinctions.

We reasoned that if two images are represented as more similar in the human visual system—due to having similar relational structure—they should be harder to tell apart in a temporal order task, leading to lower accuracy. In other words, human accuracy should be negatively correlated with the perceived similarity between the two images in the human mental representation. Therefore, if a model's internal representation aligns with this human relational space, image pairs that the model considers more dissimilar should correspond to higher human accuracy. As a result, a positive correlation between model-estimated dissimilarity and human accuracy would indicate that the model captures aspects of the relational distinctions that matter to human perception.

Language Model. For the language-based Sentence-BERT model, we generated textual descriptions for each image (e.g., "a spoon next to a cup" vs. "a spoon on top of a cup") and embedded the sentences using a pretrained sentence encoder optimized for semantic similarity

tasks.

Vision Models. For ResNet50 and ViT, we extracted visual embeddings from the penultimate layer of each model using pretrained weights. Images were resized and normalized according to standard preprocessing pipelines and then passed through each model to obtain a fixed-length embedding.

Visual Relation Detection Models. For the visual relation detection models, HL-Net and VS3, we extracted relation-level embeddings rather than generic visual features. Specifically, we computed a weighted average of the predicted relation embeddings, where the weights corresponded to the confidence scores assigned to each relation by the model.

Vision-Language Models. For the vision-language models, CLIP, InternVL-2.5, and Qwen2-VL, we extracted visual embeddings from their respective vision backbones, decoupled from the language processing streams. This approach ensured that the similarity estimates were based solely on the models' visual encoding of the image stimuli, rather than any learned associations with accompanying text.

Results. We computed Spearman correlations between model-estimated image dissimilarity (1 – cosine similarity) and human temporal order accuracy across all image pairs at 160 ms. As shown in Figure 2.5b (right), models varied substantially in how well their internal representations aligned with human judgments.

Among vision-language models, Qwen2-VL showed the strongest alignment ($\rho = 0.39$, $p = .001$), followed by InternVL-2.5 ($\rho = 0.33$, $p = .008$) and CLIP ($\rho = 0.25$, $p = .043$). These results suggest that joint training on vision and language enables the development of visual representations that are sensitive to relational structure.

The visual relation detection models yielded mixed results. VS3 exhibited a significant correlation with human accuracy ($\rho = 0.28$, $p = .026$), while HL-Net reached marginal

significance ($\rho = 0.24$, $p = .054$). This discrepancy may be due to differences in architecture and training objectives: VS3's open-vocabulary framework and grounding in language space may encourage broader relational representations, while HL-Net's performance remains more closely tied to supervised classification of coarse labels.

Interestingly, the language-only Sentence-BERT model, which compared semantic similarity between image descriptions, showed a significant but modest correlation ($\rho = 0.29$, $p = .021$), suggesting that linguistic paraphrasing can capture some—but not all—of the visual distinctions relevant to human relational reasoning.

In contrast, purely visual models performed the least well in aligning with human behavior. ResNet50 ($\rho = 0.20$, $p = .108$) and ViT ($\rho = 0.23$, $p = .072$) showed only marginal correspondence with human accuracy, consistent with the view that generic visual encoders may not reliably capture abstract spatial relations without targeted supervision.

Discussion These findings suggest that models trained solely on visual relation classification tasks may not optimally encode the fine-grained spatial distinctions that guide human judgments. Although both HL-Net and VS3 are trained to recognize visual relations, only VS3 showed consistent alignment with human performance. This may be attributed to its open-vocabulary object recognition, which allows for more flexible representations of the entities involved in a relation, even though its relation labels remain fixed. Vision-language models like Qwen2-VL and InternVL-2.5 also exhibited strong sensitivity to relational differences, supporting the idea that aligning visual representations with linguistic information can help models develop more human-like relational understanding.

As discussed in the introduction, one potential reason for this alignment gap is the ambiguity in natural language relational labels. Relational phrases such as "child in shoe" or "paper in hand" (Figure 2.1, bottom row) demonstrate how verbal descriptions often deviate from literal spatial configurations [97]. In Experiment 2, we eliminated linguistic ambiguity by using synthetic stimuli with precisely defined spatial relations. Even under these controlled

conditions, models trained on natural image datasets with linguistic labels (e.g., HL-Net) did not align well with human performance. This suggests that such models may not be learning the same relational distinctions that humans rely on, despite being trained to recognize relation categories. The findings point to a gap between model predictions based on label supervision and the perceptual structure that guides human relational judgments.

2.3.3 Experiment 3

In Experiment 3, we shifted from implicit temporal judgment to explicit relation recognition using a Go/No-Go task with naturalistic images. Participants viewed briefly flashed images (67 ms) and responded with a button press only if they detected a specified target relation. This design allowed us to directly assess participants' ability to identify a range of semantic relations—including both spatial and agentic types—under time-limited conditions, providing a more ecologically valid test of rapid relational perception. We then evaluated the performance of both visual relation detection models and vision-language models on the same task to examine how well these systems approximate human relational understanding.

Stimuli

The experimental stimuli comprised 480 naturalistic images selected from the Visual Genome dataset. We focused on 12 semantic relations, evenly divided into two categories: spatial relations (above, behind, at, near, in, hanging from) and agentic relations (carrying, holding, using, playing, riding, eating). For each relation, we manually curated a set of 40 images: 20 positive examples that clearly depicted the target relation, and 20 negative examples that either did not contain the relation or presented it in a minimally salient manner.

This conservative approach to negative sampling was motivated by the observation that certain spatial relations—such as above or near—are difficult to completely eliminate from real-world scenes due to the inherent structure of natural environments.

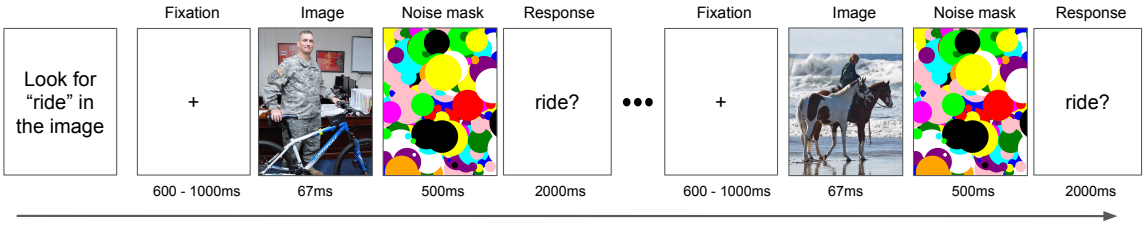


Figure 2.6: Schematic of the Go/No-Go trial procedure. Participants responded only if they judged the relation to be present in the image.

Procedure

Participants completed a Go/No-Go visual relation detection task organized into 12 blocks, with each block targeting one of the selected relations. Prior to the start of each block, participants were informed of the target relation via an instruction screen (e.g., "Please look for relation 'riding' in the image.").

Within each block, the 40 corresponding images (20 positive, 20 negative) were presented in randomized order across trials. Each trial began with a fixation cross, shown for a jittered duration randomly selected between 600 and 1000 milliseconds, followed by a target image displayed for 67 milliseconds. This was immediately succeeded by a 500-millisecond noise mask to minimize visual aftereffects.

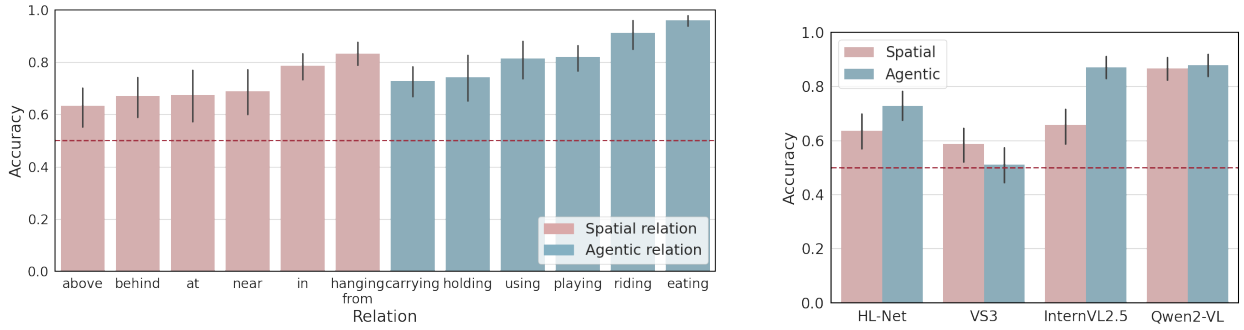
After the mask, a response prompt appeared for 2000 milliseconds, displaying the name of the relation (e.g., "riding?"). Participants were instructed to press the space bar if they believed the relation was present in the preceding image (go trials) and to refrain from responding if not (no-go trials). Responses were only recorded during the prompt phase.

A schematic overview of the trial structure is presented in Figure 2.6.

Participants

Seventeen undergraduate participants were recruited through the university's subject pool. One participant was excluded from all analyses due to self-reported lack of engagement during the post-experiment debriefing. The final sample included 16 participants (10 female, 5 male,

1 non-binary; M = 19.88 years, SD = 1.31). All participants provided informed consent and received course credit for their participation.



Human performance across 12 visual relations.

Model accuracy by relation type.

Figure 2.7: Comparison of relation detection performance between human participants and deep learning models. (a) Participants' accuracy on a Go/No-Go task shows robust detection of both spatial and agentic relations under brief presentation conditions. (b) Accuracy of visual relation detection and vision-language models across the same relation categories. Vision-language models outperform traditional models, particularly on agentic relations.

Results

Participants were able to detect both spatial and agentic relations from briefly presented images, even at a presentation duration as short as 67 ms. Accuracy rates across all 12 relations were consistently above chance, suggesting that relational information can be extracted rapidly and reliably under time-constrained viewing conditions.

As illustrated in Figure 2.7a, the highest accuracy was observed for the agentic relation "eating" (Mean = 0.961), while the lowest was for the spatial relation "above" (Mean = 0.634). All relations yielded mean accuracies well above the chance level of 0.5 (indicated by the red dashed line), confirming that participants were not responding at random. Participants demonstrated significantly higher accuracy in detecting agentic relations (e.g., riding, eating) compared to spatial relations (e.g., above, near). Specifically, mean accuracy for agentic relations was 0.830 (95% CI = [0.788, 0.872]), whereas spatial relations yielded a lower mean accuracy of 0.715 (95% CI = [0.680, 0.750]). Several factors may contribute to this disparity. First, spatial relations are often more implicit and context-dependent in natural scenes,

making them less salient and harder to explicitly evaluate within brief exposures. In contrast, agentic relations typically involve clear, goal-directed actions between agents and objects, which are perceptually more distinctive and easier to interpret rapidly.

Additionally, it is inherently difficult to eliminate spatial relations from real-world imagery, making it challenging to construct negative examples that are truly devoid of a spatial relation. This may have introduced ambiguity in some spatial trials, potentially reducing classification accuracy.

Model Results

To assess how well computational models detect visual relations under the same conditions as human participants, we evaluated both visual relation detection models (HL-Net, VS3) and vision-language models (InternVL2.5, Qwen2-VL). We excluded standard vision-only models such as ResNet or ViT from this analysis, as they do not produce relational outputs and are thus incompatible with the binary relation detection task used in this experiment.

For visual relation detection models, we checked whether the target relation was included in the top 300 predicted relations for each image. We used this higher threshold because these models often output many generic or repeated relations, and more specific ones like "riding" or "carrying" may appear later in the list. For vision-language models, we asked them the same yes/no question as in the human experiment (e.g., "Is the relation 'holding' present in the image?") and recorded their answers as either yes or no.

As shown in Figure 2.7b, all models except VS3 demonstrated above-chance accuracy for both spatial and agentic relations. Performance varied considerably across models and relation types. HL-Net and InternVL-2.5 showed a strong advantage in detecting agentic relations. VS3, however, performed poorly overall, with spatial accuracy slightly above chance and agentic accuracy near chance level. Qwen2-VL achieved the highest and most consistent performance, nearing human-level accuracy across both categories.

Interestingly, although the vision-language models were not explicitly trained for visual

relation detection, they still performed well, especially on agentic relations. This suggests that their ability to combine visual input with semantic knowledge enables a deeper understanding of image content. Their strong performance indicates that integrating vision and language may naturally support the recognition of complex relational cues in real-world scenes.

2.4 Conclusion

Across three behavioral experiments, we examined visual relation perception in humans and its alignment with computational models. In Experiment 1, participants exhibited significantly higher accuracy in image recognition when image pairs depicted different spatial relations, even with exposures as brief as 40 milliseconds. This finding suggests that relational information is rapidly and automatically extracted during early visual processing. In Experiment 2, we extended this investigation using a synthetically generated image set designed to isolate spatial relational structure while controlling for low-level visual features and semantic context. Replicating the core effect observed in Experiment 1, participants again showed significantly higher accuracy when spatial relations differed between image pairs. This demonstrates that the human visual system is sensitive to abstract relational distinctions even in the absence of naturalistic scene cues, and that spatial relations are encoded robustly and early in visual processing. In Experiment 3, participants completed a Go/No-Go task and successfully identified both spatial and agentic relations from real-world images shown for only 67 milliseconds. Higher accuracy for agentic relations suggests that dynamic, action-based cues may be particularly salient and efficiently processed by the human visual system.

Taken together, these findings provide strong evidence that relational information is encoded rapidly and reliably in human perception. The ability to detect "what is where" or "who is doing what to whom" appears to be tightly integrated with early visual processes, operating within the first feedforward sweep of perception.

Our comparison with computational models revealed substantial variability in model–human

alignment. Traditional visual relation detection models, despite being trained on relation labels, often showed limited correspondence with human behavior. Notably, HL-Net failed to capture human relational sensitivity even under controlled conditions, suggesting that supervised training on coarse relation categories may not promote human-like relational representations. In contrast, VS3 showed modest alignment, potentially due to its open-vocabulary framework that allows more flexible grounding of object entities.

Vision-only models such as ResNet-50 and ViT exhibited relatively weak alignment with human accuracy, indicating that generic visual encoders do not automatically capture abstract relational distinctions without explicit relational training. Interestingly, vision-language models—particularly Qwen2-VL and InternVL-2.5—achieved the strongest alignment with human behavior. These models appear to benefit from visual and linguistic alignment, enabling representations that are more sensitive to relational differences, even when linguistic ambiguity is absent.

These results suggest that relational understanding may not be adequately learned through narrowly defined classification objectives alone. Instead, relational understanding may emerge more naturally in models that are exposed to broad, multimodal corpora, where relations are embedded within a network of semantic and perceptual associations. This has important implications for the design of future models: rather than treating relation detection as a narrow classification task, we may need to adopt more holistic training objectives that reflect how humans learn relational structure—from rich, overlapping streams of perceptual and conceptual input.

Chapter 3

Evaluating Compositional Scene Understanding in Multimodal Generative Models

3.1 Introduction

A fundamental aspect of human visual processing is its compositional nature. By composing objects and relations in novel ways, we have the capacity to imagine and accurately perceive an infinite variety of visual scenes, even highly improbable ones [22, 55, 136]. This ability to exploit the compositional nature of visual scenes is an important priority for the development of general-purpose, human-like computer vision systems.

Recent progress in multimodal generative models has led to the development of systems with an impressively general range of visual capabilities, including systems for text-to-image generation [121], and multimodal vision-language models [1]. Notably, these systems have been touted for their compositional abilities, as evidenced by their ability to generate novel

The contents of this chapter appeared in paper [43].

and unusual scenes depicting, for instance, 'a baby daikon radish in a tutu walking a dog', or the famous 'avocado chair' [59]. On the other hand, these systems also sometimes display basic failures of compositionality. For example, it was reported that the text-to-image model DALL-E 2 was incapable of reliably generating 'a red cube on top of a blue cube' [124], and followup work found that DALL-E 2 could not reliably generate images to match prompts describing scenes involving relations [34,105]. Similarly, while the advent of multimodal vision-language models such as GPT-4 has been heralded as a major step toward the development of systems with general-purpose visual understanding [157], some recent studies have identified major shortcomings in their ability to accurately interpret visual scenes, especially for scenes involving the composition of multiple objects and relations [113, 123].

In this chapter, we assess the extent to which the current generation of multimodal generative models has made progress toward developing a capacity for compositional image understanding. Our experiments are divided into two broad sections. In the first section, we evaluate the ability of the text-to-image model DALL-E 3 to generate images based on spatial and agentic relations. We first assess the degree of improvement for relational prompts (e.g., 'a blanket covering a box') used in previous work (to test DALL-E 2) [34], and then test novel variants of these prompts, including reversed prompts that evaluate improbable scenarios (e.g., 'a rabbit chasing a tiger'), and compositional prompts involving multiple relations among several entities. We carry out online experiments with external raters to assess the extent to which the generated images accurately depict the objects and relations described in the prompts. In the second section, we evaluate the ability of multimodal vision-language models (GPT-4v, GPT-4o, Claude 3.5 Sonnet, QWEN2-VL-72B, and InternVL2.5-38B) to infer relational patterns from a set of example images. We test problems involving both synthetic images (the Synthetic Visual Reasoning Test (SVRT) [39]) and real-world scenes (using the Bongard-HOI dataset [70]), and perform a direct comparison with human behavioral data.

Our results indicate that multimodal generative models have indeed made some progress toward compositional scene understanding, with DALL-E 3 displaying improvements over

DALL-E 2 in its ability to generate images from relational prompts, and the evaluated multimodal language models all displaying some capacity to infer relational patterns. These successes are tempered, however, by persistent failures of compositionality. While DALL-E 3 is capable of generating images based on more common relational prompts, performance degrades for reversed prompts involving less common scenarios, or prompts involving multiple relations. Similarly, all of the evaluated multimodal language models perform well below human participants in their ability to infer relational patterns, with performance falling to chance levels for problems involving many (> 5) objects. Overall, these results suggest that the current generation of multimodal generative models do not possess a robust capacity for compositional scene understanding.

A summary of our contributions is provided below:

- We propose two new datasets for evaluating the robustness of relational text-to-image generation through the use of reversed and compositional relational prompts.
- We perform a comprehensive human behavioral evaluation of relational image generation in a state-of-the-art text-to-image model (DALL-E 3).
- We perform a comprehensive evaluation of relational concept learning in five multimodal language models (GPT-4v, GPT-4o, Claude 3.5 Sonnet, QWEN2-VL-72B, and InternVL2.5-38B), for both real-world and synthetic images, and directly compare the performance of these models with human behavior.

3.2 Evaluating Relational Image Generation in Text-to-Image Models

We evaluated the ability of the text-to-image model DALL-E 3 to generate images based on prompts that described scenes involving relations. We performed three human experiments to evaluate the validity of images generated by text-to-Image models. Section 3.2.1 describes

an experiment evaluating DALL-E 3 on the relational prompts used in Conwell and Ullman's study of DALL-E 2 [34]. Section 3.2.2 reports an experiment involving reversed versions of these prompts, intended to test DALL-E 3's ability to generalize to unusual scenarios (e.g., 'a rabbit chasing a tiger'). Section 3.2.3 describes an experiment involving prompts with many objects and multiple relations, testing DALL-E's ability to compositionally generalize by combining relations into more complex configurations.

3.2.1 Image Generation with Basic Relational Prompts

Methods

We first assessed the ability of DALL-E 3 to generate images based on basic relational prompts, using the same prompts as those used in a previous study evaluating DALL-E 2 [34]. We adopted a design similar to the one used in previous work [34], asking human participants to judge whether images matched the corresponding text prompts.

We used a set of 75 prompts introduced by Conwell & Ullman ([34]) to evaluate relational image generation. Each prompt involved two entities with a single relation between them (e.g., 'a tiger chasing a rabbit'). We found that it was necessary to modify some prompts due to policy violations with DALL-E 3 (e.g., prompts involving 'kicking' or 'hitting'). 25 prompts were modified with slight wording changes, while respecting the criteria used to create prompts described by [34] (see Section A.2). Prompts were broadly classified into those that involved physical relations (in, on, under, covering, near, occluded by, hanging over, and tied to), and those that involved agentic relations (pushing, pulling, touching, chasing, hugging, helping, and hindering). We refer to this set of text prompts adopted from Conwell & Ullman ([34]) as *basic relational prompts*.

Images were generated by prompting DALL-E 3 [12], a text-to-image model developed by OpenAI, through the Microsoft Azure API[1] (version `2024-02-01` for all experiments). DALL-E 3 was instructed to use the exact prompt provided, and not to revise the prompt

[1] https://learn.microsoft.com/en-us/azure/ai-services/openai/

Figure 3.1: Examples of images generated by DALL-E 3 for basic relational prompts (prompts the describe scenes involving relations), using the 'natural' style. See Figure A.3 for more examples.

before image generation (prompts are sometimes automatically revised by GPT-4 before image generation). The final prompt used by DALL-E 3 was checked against the original prompt, and, if necessary, DALL-E 3 was prompted again, until the original and final prompts were identical. We generated 10 images for each prompt, with the following hyperparameters: quality set to 'standard', style set to 'natural', and image size set to 1024×1024.

We performed a human behavioral experiment to assess the extent to which the images generated by DALL-E 3 matched the prompts, using a design similar to Conwell & Ullman ([34]). Participants were presented with a prompt, along with a collection of 10 images generated by DALL-E 3, and asked to select all of the images that matched the prompt. More details can be found in Section A.3.

Results

Figure 3.2 shows the results for the experiment with basic relational prompts. 'Agreement' reflects the proportion of participants who indicated that the generated images matched with the corresponding prompts, where 1 means that all participants indicated that an image matched the prompt, and 0 means that no participants indicated that an image matched the prompt. Results are grouped according to the specific relations used in each prompt. Results from our experiment with DALL-E 3 are presented in blue, alongside the original results from Conwell & Ullman's ([34]) experiment with DALL-E 2 in red.

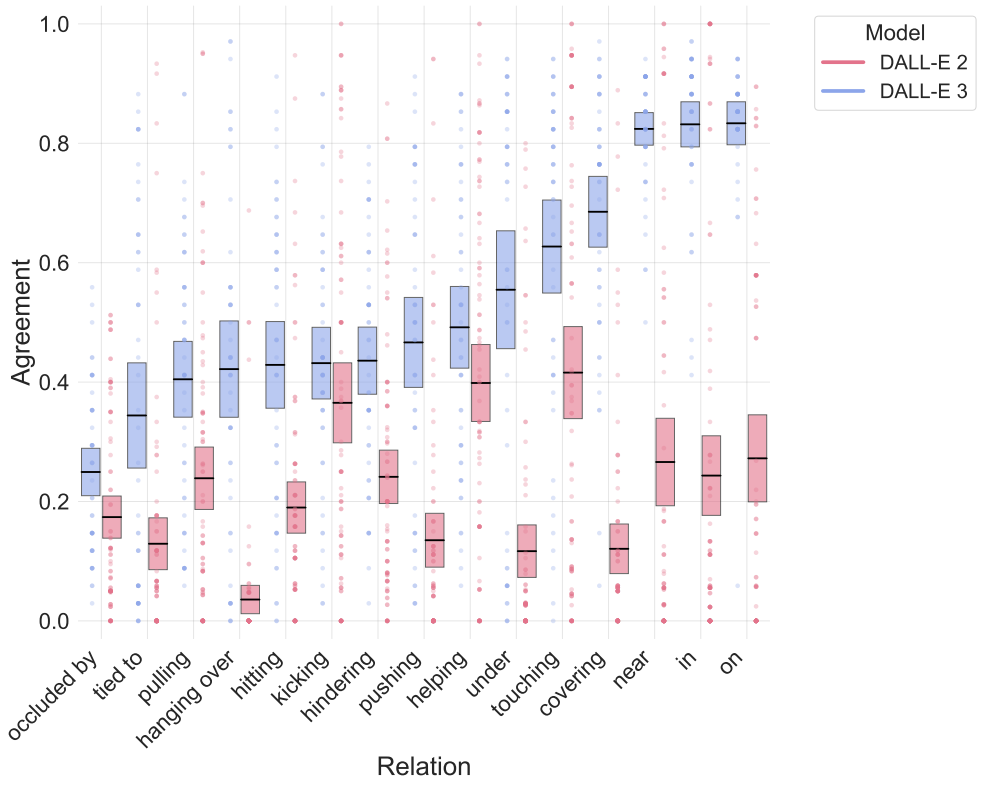


Figure 3.2: Results for basic relational prompts. The Y axis indicates the agreement between prompts and the images generated by either DALL-E 3 or DALL-E 2, as judged by human participants. Results for DALL-E 3 (in blue) are from the present study. Results for DALL-E 2 (in red) are from Conwell & Ullman ([34]). Each point reflects the average agreement for an individual image. Horizontal lines indicate the mean agreement for each relation, and boxes indicate 95% confidence intervals.

We found that DALL-E 3 displayed a significantly improved capability, relative to DALL-E 2, to generate images based on basic relational prompts (paired t-test, DALL-E 3 vs. DALL-E 2: $t(49) = 8.552$, $p < .001$), with higher agreement for all relations tested. For some relations in particular, including spatial relations such as 'on', 'near', or 'in', DALL-E 3 displayed major improvements, with average agreement of greater than 0.8, compared with the very low agreement (< 0.4) observed in images generated by DALL-E 2. We did not observe any major differences in performance between physical (in, on, under, covering, near, occluded by, hanging over, and tied to) and agentic (pushing, pulling, touching, chasing, hugging, helping, and hindering) relations. Figure 3.1 shows some examples of the images generated by DALL-E 3.

To assess the extent to which the pattern of results might be influenced by the binary decision in the human behavioral task (i.e., participants only indicated whether or not an image matched the prompt), we also performed an additional experiment involving a more fine-grained rating (a Likert scale from 1 to 7). This experiment yielded qualitatively similar results (see Figure A.2 in Section A.3). Overall, the results of our experiments with basic relational prompts indicated improved performance in DALL-E 3 relative to DALL-E 2, with some relations in particular showing major improvements.

3.2.2 Image Generation with Reversed Prompts

Methods

Next, we investigated whether DALL-E 3 was able to generate images based on prompts in which the subject and object were reversed. For instance, the basic prompt 'a tiger chasing a rabbit' was changed to 'a rabbit chasing a tiger', switching the roles of the two objects (see Section A.2 for more details about these prompts). The goal of this experiment was to evaluate whether DALL-E 3 was capable of generating relational scenes involving unusual arrangements, as this is typically taken to be a key capacity enabled by compositionality (i.e., unusual scenes can be imagined by combining familiar elements). Images were generated by querying DALL-E 3 in the same manner as the previous experiment, except that the style parameter was set to 'vivid'. We found that DALL-E 3 is more likely to generate matched images for reversed prompts using the 'vivid' style than the 'natural' style. We then performed a behavioral experiment to assess the agreement of the images with their prompts, using a similar design to the experiment with basic relational prompts. Due to the introduction of new policy warnings from reversed prompts, we made 15 more prompt revisions to 75 prompts from Experiment 1, with 7 modifications to already modified prompts and 8 modifications to new ones. This set of 75 was then filtered down to 30 final prompts along with their reversed variants (a total of 60 prompts). More details can be found in Section A.3.

Figure 3.3: Examples of images generated by DALL-E 3 for basic relational prompts and their corresponding reversed prompts using the 'vivid' style. See Figure A.4 for more examples.

Results

Figure 3.2 shows the results for the experiment with reversed prompts. DALL-E 3 displayed significantly worse agreement for reversed prompts than it did for the original basic prompts (paired t-test, basic vs. reversed prompts: $t(29) = 3.668$, $p = 0.001$), though the extent of this effect differed between different relations. Some relations displayed almost no difference in performance between the basic and reversed prompts, including 'near', 'helping', and 'touching', while other relations displayed major difference in performance, including 'on', 'in', and 'chasing'. Notably, two out of the three relations for which performance was not strongly affected by prompt reversal, 'near' and 'touching', are symmetric in nature, so reversing the order of subject and object should not affect the meaning of the prompt, whereas the three relations for which performance was strongly affected, 'on', 'in', and 'chasing', are all asymmetric. It is also worth noting that effect of prompt reversal was particularly pronounced for the relation 'chasing', likely due to the fact that the reversed prompts described very unusual scenarios (such as a rabbit chasing a tiger). Figure 3.3 shows examples generated by DALL-E 3 for basic and corresponding reversed prompts.

To systematically assess the impact of prompt likelihood on the validity of the generated images, we measured the plausibility of each text prompt using the GPT-3 language model from OpenAI (the 'davinci-002-1' engine, available through the Microsoft Azure API). We used log likelihood (LL), the sum of the log-probabilities of each token in a sentence, as a

measure of semantic plausibility [63, 74]. We performed a correlation analysis comparing LL with average agreement (between the prompt and the generated images). We found that LL had a statistically significant (but modestly sized) correlation with agreement (spearman correlation $r(598) = 0.090$, $p = 0.027$), consistent with the interpretation that lower agreement for reversed prompts was driven by the lower likelihood of these prompts in DALL-E 3's training data. These results suggest that, although DALL-E 3 shows an improved capability to generate relational scenes relative to DALL-E 2, it is not able to robustly generalize this capability to unusual scenarios that are less likely to be present in the model's training data.

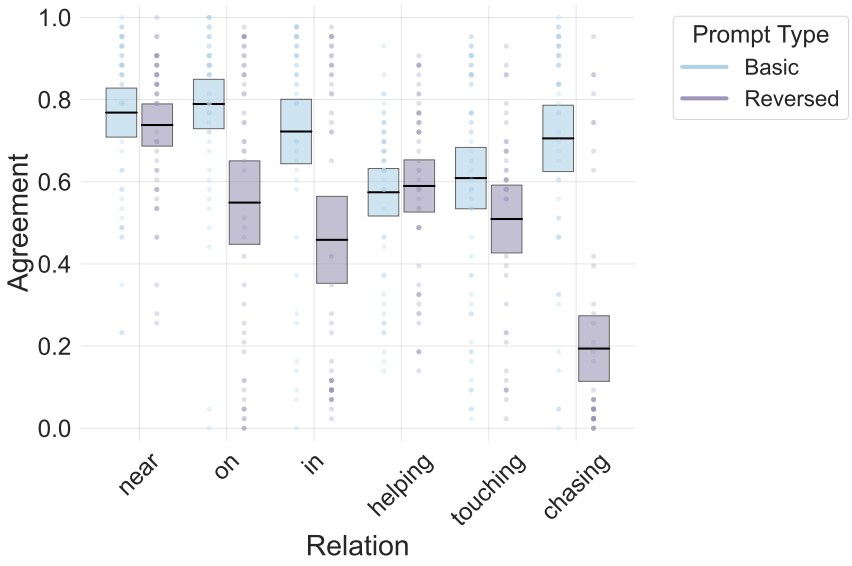


Figure 3.4: Results for compositional relational prompts. The Y axis indicates the agreement between prompts and the images generated by either DALL-E 3, as judged by human participants. Each point reflects the average agreement for an individual image. Horizontal lines indicate the mean agreement for each relation, and boxes indicate 95% confidence intervals.

3.2.3 Image Generation with Compositional Prompts

Methods

Finally, we investigated DALL-E 3's ability to generate images based on more complex prompts involving the composition of two relations. A key feature of image compositionality is the ability to generate new visual scenes by combining multiple objects connected through

various relationships. To test this capacity in DALL-E 3, we generated 60 prompts, each of which involved three entities linked by two relations (e.g., 'a spoon tied to a box, with the box near a cylinder'). These prompts were divided into five categories: 10 prompts involved two instances of the same physical relation; 10 prompts involved two different physical relations; 10 prompts involved two instances of the same agentic relation; 10 prompts involved two different agentic relations; and 20 prompts involved one physical relation and one agentic relation. For each prompt, we generated 10 images, yielding 600 images in total. Images were generated by querying DALL-E 3 in the same manner as the previous experiment. We performed a behavioral experiment to assess the agreement of the images with their prompts, using a similar design to the previous experiments (see Section A.3).

Figure 3.5: Examples of images generated by DALL-E 3 for compositional relational prompts using the 'vivid' style. See Figure A.5 for more examples.

Results

Figure 3.6 shows the results for the experiment with compositional relational prompts. Performance varied widely across relation combinations, with some combinations displaying relatively high agreement (e.g. 'in & in' or 'helping & near'). However, the majority of relation combinations displayed relatively low agreement, suggesting that DALL-E 3 had significant difficulty generating images that required the composition of multiple relations. Figure 3.5 shows examples generated by DALL-E 3 with text prompts involving multiple

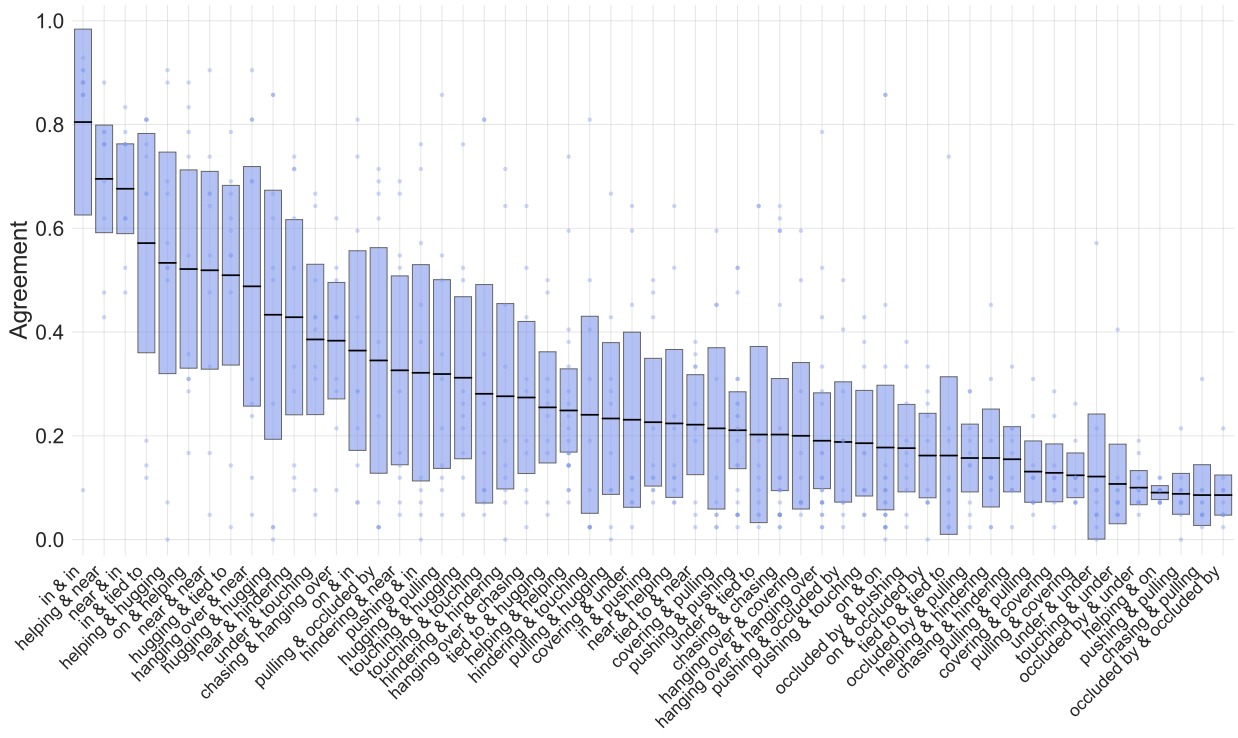


Figure 3.6: Results for compositional relational prompts. The Y axis indicates the agreement between prompts and the images generated by DALL-E 3, as judged by human participants. Each point reflects the average agreement for an individual image. Horizontal lines indicate the mean agreement for each relation, and boxes indicate 95% confidence intervals.

objects and relations.

3.2.4 Summary of Relational Image Generation Results

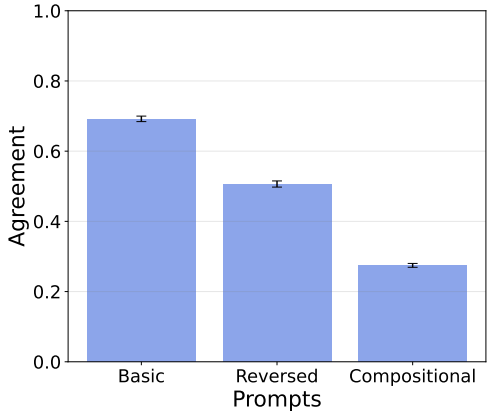


Figure 3.7: Summary of results for relational image generation experiments with DALL-E 3. Error bars reflect 95% confidence intervals.

Figure 3.7 shows a summary of the results for the relational image generation experiments. These results highlight two major findings. First, DALL-E 3 displays a notably improved ability to generate relational scenes relative to DALL-E 2. However, DALL-E 3's performance is degraded significantly for prompts in which the order of subject and object are reversed (often yielding unlikely scenes), or prompts involving the composition of two relations. These failure modes indicate that DALL-E 3's capacity for scene generation is not robustly compositional, as a compositional approach would enable entities to be combined in novel ways, and more complex scenes to be formed through the composition of multiple relations.

3.3 Evaluating Relational Concept Learning in Multimodal Language Models

We evaluated the ability of two multimodal variants of GPT-4, gpt-4-vision-preview and gpt-4o, to perform few-shot relational concept learning. We evaluated these models using two datasets: Bongard-HOI [70], involving action- and event-based relations (human-object interaction) in naturalistic scenes, and the Synthetic Visual Reasoning Test (SVRT) [39], involving visuospatial relations in synthetically generated images. Both datasets test the ability to learn abstract relational concepts from a relatively small number of demonstrations. To ensure that the learned concepts are genuinely relational, these datasets employ 'hard negatives' – incorrect answers that share object-level features with the relational concept, but differ at the relational level (e.g., for the concept 'human rides bicycles', a hard negative will involve a human and a bicycle with a different relation).

Figure 3.8: Example relational concept from Bongard-HOI. Positive instances feature a human riding a bicycle, while negative instances feature a human and a bicycle with a different relation.

3.3.1 Learning Relational Concepts from Real-World Scenes

Methods

We first evaluated relational concept learning with real-world scenes, using the Bongard-HOI dataset [70]. This dataset contains 167 distinct concepts involving human-object interactions. Each concept is defined by a subject-verb-object triplet, where the subject is always a human, the verb indicates an action (e.g., 'hold' or 'inspect'), and the object is a common item (e.g., 'bicycle' or 'apple'). Each concept is illustrated by 7 positive examples and 7 negative examples, each presented in a separate image (Figure 3.8). In the standard evaluation methodology, 6 labeled images from each class are presented (12 total), and the remaining positive and negative images are presented for classification.

We evaluated gpt-4-vision-preview and gpt-4o on problems taken from the Bongard-HOI test set, and compared their performance to human performance as reported with the original dataset, as well as the HOITrans baseline model (the best performing model in that work) [70]. These models were evaluated using the Microsoft Azure API. Due to the limited number of images that could be presented in the context window for these models (10 images maximum), evaluation was performed by presenting 9 labeled example images (randomly selecting either 5 positive examples and 4 negative examples, or vice versa) followed by a single query image (randomly selecting either a positive or negative example). Given the large size of the dataset (13,941 problems), we sampled 10 problems per concept for our evaluation (1,670 problems in

total). Some problems resulted in content policy violations. After excluding problems that triggered policy violations, gpt-4-vision-preview was tested on 1,149 problems (involving 161 concepts) and gpt-4o was tested on 1,156 problems (involving 160 concepts). Temperature was set to 0 when evaluating both models, top-p was set to 1, and the 'detail' parameter was set to 'high'.

Results

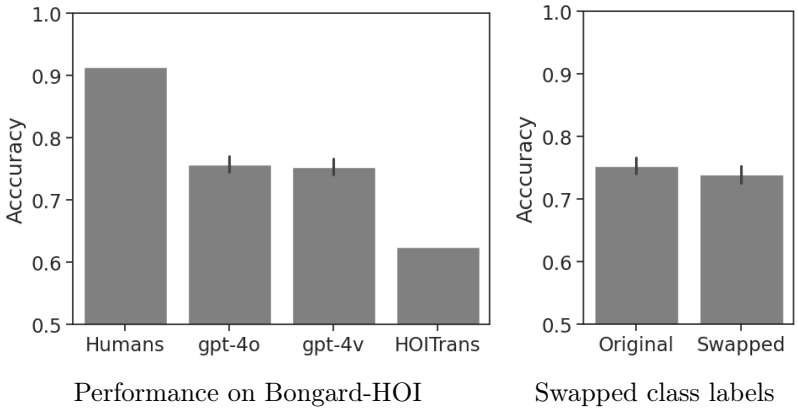


Figure 3.9: Bongard-HOI results. (a) Few-shot classification accuracy for two variants of GPT-4 (gpt-4o and gpt-4-vision-preview, shortened to gpt-4v) vs. best previous model (HOITrans) and human performance. (b) Performance of gpt-4v on original problems vs. problems with swapped class labels. Error bars reflect binomial 95% confidence intervals.

Figure 3.9a shows the results for our experiments with the Bongard-HOI dataset. Both GPT-4 variants significantly outperformed HOITrans, the best performing model from previous work (one sample proportion tests, gpt-4o vs. HOITrans: $\chi^2(1) = 80.06, p < 0.0001$, gpt-4-vision-preview vs. HOITrans: $\chi^2(1) = 75.27, p < 0.0001$), but performed significantly worse than the human participants on this task (one sample proportion task, human vs. gpt-4o: $\chi^2(1) = 335.78, p < 0.0001$, human vs. gpt-4-vision-preview: $\chi^2(1) = 352.61, p < 0.0001$).

One possible concern with these results is that the Bongard-HOI dataset is publicly available on the internet, and therefore may in principle have been present in the training data for these models. To determine whether performance on this task depends on prior training, we performed an experiment in which the labels assigned to classes were swapped

(positive examples, with a label of 1, became negative examples, with a label of 0, and vice versa). We tested gpt-4-vision-preview on the swapped-labels task, and compared performance to the standard version of the task (Figure 3.9b). There was no difference in performance for these two versions of the task (logistic regression, swapped labels vs. original labels: $z = 0.83, p = 0.41$), indicating that performance on this task was not due to memorization of a publicly available dataset. However, it is also important to consider that this dataset features common relational configurations (e.g., a man riding a bicycle) that are likely to have been present in some form in the training data for these models. Thus, while we have ruled out the possibility of literal memorization, there is still a concern that performance may depend to some extent on general familiarity with the relations and configurations present in these problems. To address this, in the next section, we present results from a synthetic dataset (SVRT) with novel relational concepts not likely to be present in the training data.

3.3.2 Learning Relational Concepts from Synthetic Scenes

Methods

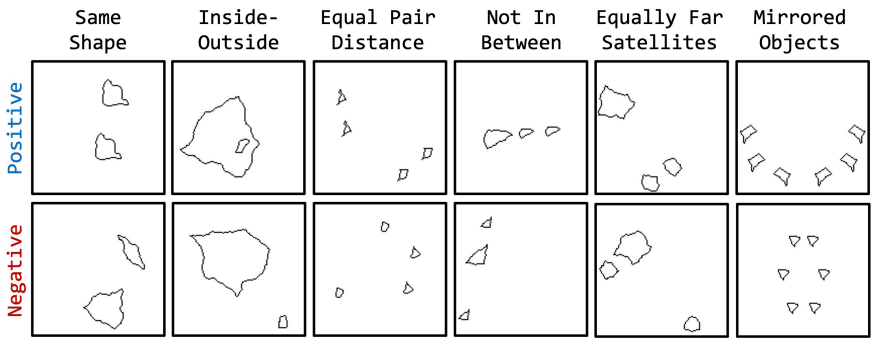


Figure 3.10: Examples of relational concepts in SVRT. Each concept shows a positive and negative instance.

We evaluated relational concept learning with synthetic scenes using the SVRT dataset [39]. This dataset features abstract relational concepts instantiated with synthetic figures, making it difficult to rely on previous experience with common real-world relational concepts. Additionally, many SVRT problems feature relational concepts involving a larger number

of objects and relations, making it possible to test for the ability to learn more complex relational concepts.

The SVRT dataset consists of 23 distinct relational concepts, each involving between 2 and 6 objects. These concepts range from simple pairwise relations (e.g., same shape) to more complex configurations of multiple objects and relations (e.g., mirrored objects) (Figure 3.10). We used the original codebase from [39][2] to generate novel positive and negative instances for each concept. This allowed us to ensure that the images would not have been present in the models' training data.

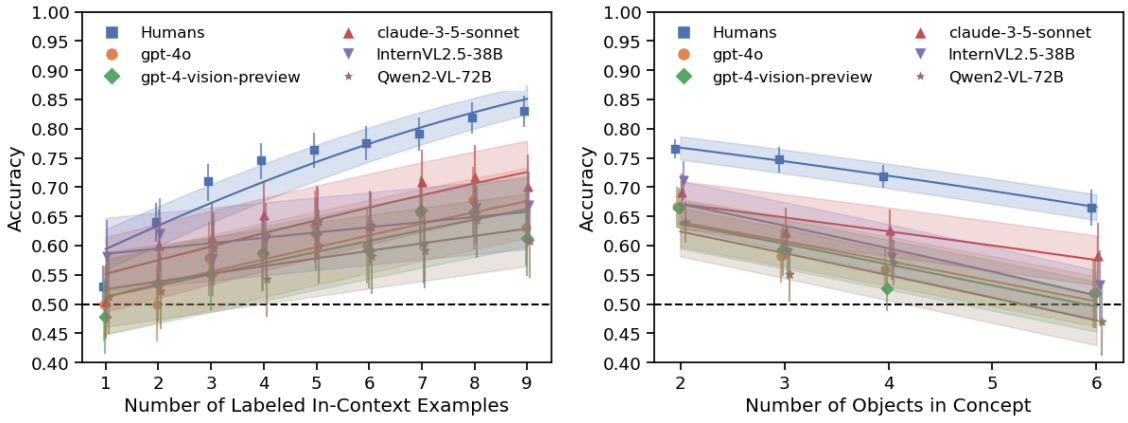
We evaluated gpt-4-vision-preview and gpt-4o on these problems using the Microsoft Azure API. To evaluate a broader set of models, we also evaluated Claude Sonnet 3.5 (through the Anthropic API), and two open-source VLMs: QWEN2-VL-72B and InternVL2.5-38B (these were evaluated locally on a workstation with 4 NVIDIA GPUs). For each problem, we presented 1-9 few-shot examples, consisting of a mixture of positive and negative instances (in random order). Each example was presented in a separate image, and was accompanied by a message indicating the class (0 or 1). After presenting the few-shot examples, we presented a target image (randomly sampling either a positive or a negative instance) and prompted the model to make a classification. For each relational concept (out of 23), and each number of few-shot examples (1-9), we evaluated each model 10 times, using completely different images each time, resulting in a total of 2,070 tests. The models occasionally made formatting errors in its responses. In these cases, we re-prompted the model until it produced a correctly formatted response. Temperature was set to 0 when evaluating both models, top-p was set to 1, and the 'detail' parameter was set to 'high'.

We also performed a comparison with human behavioral data from [86]. In that experiment, participants were presented with SVRT problems in interactive episodes. For each episode, participants received a series of positive and negative examples illustrating a relational concept, and attempted to classify each example. After each response, they received feedback

[2] https://fleuret.org/cgi-bin/gitweb/gitweb.cgi?p=svrt.git;a=summary

indicating the correct classification. The previous examples from the current episode remained on screen, and were sorted based on class. The episode continued until participants made 7 correct classifications in a row. See [86] for more details on the human behavioral experiment, and Section A.7 for more details on the analysis comparing VLMs with human performance.

Results



Performance by number of in-context examples

Performance by object count

Figure 3.11: SVRT few-shot classification accuracy for humans, gpt-4o, gpt-4-vision-preview, claude 3.5 sonnet, Qwen2-VL-72B, and InternVL2.5-38B by (a) number of in-context examples and (b) number of objects in the concept. Points are averages across all SVRT concepts and attempts, coupled with binomial 95% confidence intervals. Curves are logistic regression fits with binomial confidence interval bands. Dashed black lines indicate chance performance.

Figure 3.11 shows the comparison of human participants with GPT-4 on the SVRT. Human participants significantly outperformed all models (logistic regression, gpt-4-vision-preview: $\beta = 0.63, z = 12.01, p < 0.0001$; gpt-4o: $\beta = 0.66, z = 12.65, p < 0.0001$; claude 3.5 sonnet: $\beta = 0.44, z = 8.18, p < 0.0001$; Qwen2-VL-72B: $\beta = -0.70, z = 13.52, p < 0.0001$; InternVL2.5-38B: $\beta = -0.52, z = 9.82, p < 0.0001$). Both human participants and all VLMs displayed improved performance with more in-context examples (human: $\beta = 0.17, z = 15.57, p < 0.0001$, gpt-4-vision-preview: $\beta = 0.078, z = 4.47, p < 0.0001$, gpt-4o: $\beta = 0.085, z = 4.88, p < 0.0001$; claude 3.5 sonnet: $\beta = 0.096, z = 5.32, p < 0.0001$; Qwen2-VL-72B: $\beta = 0.053, z = 3.08, p = 0.0021$; InternVL2.5-38B: $\beta = 0.038, z = 2.18, p <$

0.030). (Figure 3.11a), and degraded performance as a function of the number of objects in a relational concept (human: $\beta = -0.13, z = 6.18, p < 0.0001$, gpt-4-vision-preview: $\beta = -0.18, z = 5.12, p < 0.0001$, gpt-4o: $\beta = -0.16, z = 4.69, p < 0.0001$; claude 3.5 sonnet: $\beta = -0.1144, z = 3.29, p = 0.0010$; Qwen2-VL-72B: $\beta = -0.16, z = 4.61, p < 0.0001$; InternVL2.5-38B: $\beta = -0.19, z = 5.62, p < 0.0001$) (Figure 3.11b). Notably, performance for both GPT-4 variants and the two open-source VLMs was statistically indistinguishable from chance levels for problems with 6 objects (gpt-4-vision-preview: $\chi^2(1) = 0.3, p = 0.58$; gpt-4o: $\chi^2(1) = 0.3, p = 0.58$; Qwen2-VL-72B: $\chi^2(1) = 0.83, p = 0.36$; InternVL2.5-38B: $\chi^2(1) = 1.070, p = 0.30$), although Claude Sonnet showed performance above chance for these problems ($\chi^2(1) = 6.85, p = 0.0089$). Thus, while the VLMs that we investigated displayed some capacity to learn relational concepts in this task, and was sensitive to the same factors as human participants (number of in-context examples and number of objects per image), performance was nevertheless well below human level, and they were generally not capable of solving more complex problems involving a greater number of objects and relations. In the Appendix, we also present additional experiments that explore the role of chain-of-thought prompting (section A.8), and interactive testing of GPT-4 (section A.8.1), finding that neither of these factors improve performance on this task.

3.4 Related Work

3.4.1 Datasets for Evaluating Compositionality in Neural Networks

A number of previous studies have investigated the ability of neural networks to process compositional scenes. Prior to the development of large-scale pre-trained models, several datasets were proposed for evaluating compositional visual processing and reasoning in deep learning models, including Visual Genome [81], CLEVR [71], PGM [8], and OOD [45], typically emphasizing out-of-distribution and compositional generalization (but with large, task-specific training sets).

3.4.2 Evaluating Compositionality in CLIP and Text-to-Image Models

More recent research has focused on evaluating the zero-shot and few-shot capability of pre-trained multimodal generative models to process and generate compositional scenes. [88] investigated the compositional processing capabilities of Contrastive Language-Image Pre-Training (CLIP) [121], on which both text-to-image models and multimodal language models are based, finding that CLIP could compose concepts at the single-object level, but could not reliably compose multiple objects or relations. [105] and [34] investigated the ability of DALL-E 2 to generate images based on compositional and relational prompts. Our study adds to this work by investigating the ability of DALL-E 3 to generate compositional scenes, finding that it has improved performance relative to DALL-E 2, but is still not robustly compositional, especially in settings with many objects and relations. Concurrent work from [33] also shows that DALL-E 3 struggles with prompts that involve logical operators such as negation.

3.4.3 Visual Reasoning in Multimodal Language Models

Other recent studies have investigated the ability of multimodal language models (e.g., GPT-4v) to process compositional scenes. Both [162] and [113] evaluated GPT-4v on visual analogy problems. [162] found limited success (GPT-4v was able to solve visual analogies for certain types of relations), whereas [113] found that GPT-4v performed very poorly on problems from the Abstraction and Reasoning Corpus (ARC) [32], even for problems that human participants perform nearly perfectly on. Our work adds to these studies by investigating the ability of these models to infer relational concepts in a few-shot setting (whereas analogy involves the inference of a relational pattern from just a single example), finding limited success in this domain.

It is also worth noting that previous work has found that both convolutional neural

networks and vision transformers can perform nearly perfectly on the SVRT with a very large number of training examples [77, 108, 109]. In this work we were interested specifically in the ability to solve SVRT problems with a relatively small number of examples (as is done with human participants), which remains a challenging setting [114, 143, 152].

3.5 Discussion

In this work, we have presented a broad-ranging evaluation of compositional scene understanding in multimodal generative models. In our experiments with text-to-image models, we found that DALL-E 3 shows an improved capability to generate relational images relative to the previous generation of these models (DALL-E 2), but displays a lack of robustness for images involving reversed (i.e., novel) relational configurations, and more complex scenes involving many objects and more than one relation. In our experiments with multimodal generative models, we found that all of the models we evaluated (including two multimodal variants of GPT-4, Claude 3.5 Sonnet, and two open-source models) displayed some capability for few-shot learning of relational visual concepts, for both real-world scenes and abstract visuospatial problems, but underperformed relative to human participants, and did not show any capacity to learn relational concepts involving a larger number of objects. Overall, these results present a complex picture, with the current generation of multimodal generative models displaying some capacity for compositional and relational processing, while still falling short of the robustness and generality exhibited by human visual processing.

These results raise the question of how to reconcile the complex pattern of successes and failures observed both in the present set of experiments and other recent work. A major theme throughout all of these studies is that the current generation of models seem to display especially poor performance in more complex multi-object settings. [88] found that CLIP was capable of composing concepts at the single-object level, but could not do so reliably for multi-object scenes. This is echoed by our findings with text-to-image

models, where performance is especially degraded for prompts involving a large number of objects. In the case of multimodal generative models, both our results and the results of [162] indicate some capacity for processing of visual relations, whereas [113] find very poor performance on the ARC dataset. This discrepancy may also be explainable by a difficulty with multi-object processing: the SVRT and Bongard-HOI (both investigated in this work), and KiVA (investigated by [162]) all involve a relatively small number of objects per image, whereas ARC problems typically involve a much larger number of objects. In the few cases that we investigated that did involve a larger number of objects (for some SVRT problems), performance was very poor, with only one out of the five models that we evaluated showing performance better than chance.

Interestingly, similar results have also been observed for tasks that do not involve relations, such as counting [123, 125], or multi-object scene description [19]. Indeed, [19] found that multimodal language models were able to solve visual analogy problems when they were decomposed into object-level representations, but could not do so reliably when the same problems were presented within a single image. This pattern of results suggests that the poor performance observed in the current generation of multimodal models may be due primarily to a more basic difficulty with parsing of multi-object scenes, rather than a difficulty with relational processing per se. This would also explain the discrepancy between the relatively poor performance of multimodal models on visual analogy problems and the more robust capability of language models on text-based analogy problems ([150], [117], [151], cf. [87]).

Finally, it is worth considering how the compositional and multi-object capabilities of multimodal generative models might be further improved. [19] argued that these difficulties are related to the classic 'binding problem' from cognitive science [41, 52, 139], which arises any time that a shared set of representational resources must be used to represent multiple objects, but without a mechanism for binding features at the level of objects. In recent years, object-centric representation learning methods, most notably including slot-based approaches [18, 96], have arisen as a solution to the binding problem, and have been shown

to dramatically improve performance in visual reasoning tasks that require multi-object compositionality [35, 114, 115, 152]. This suggests that a promising direction for future work may be to combine multimodal generative models with object-centric visual processing methods.

Chapter 4

Hierarchical Abstraction Enables Human-Like 3D Object Recognition in Deep Learning Models

4.1 Introduction

Objects in the natural world are three-dimensional and possess physical properties such as geometric shape, material, and volume. From a brief glance, the human visual system excels in extracting visual attributes from pixel-level information in images to infer these object properties. Among the object properties, recognizing the three-dimensional shape of objects is considered fundamental for most daily life tasks involving navigation and interaction with the external world. Humans utilize depth cues and spatial relationships to understand and identify objects. Considerable research [95, 106, 145] indicate that humans construct mental representations of objects that incorporate 3D structural information, facilitating recognition from different perspectives. Human 3D object recognition is both accurate and remarkably robust. A key example of this robustness is the ability to perceive and interpret 3D objects even with limited information, such as point cloud displays that are sparsely sampled along

object surfaces [53, 116, 140, 141, 144]. Such high accuracy and robustness of human vision system appears particularly tuned to 3D object recognition. Previous research shows that 3D shape perception arises early in human infancy [75], and that human toddlers begin to develop a strong shape bias between 18 and 24 months, shifting from an early reliance on texture and other perceptual features to generalizing object names based on shape as their vocabulary expands [160]. While sensitivity to texture evolves, it is secondary compared to shape-based object recognition.

These empirical evidence in human development is in contrast with the default strategies used in deep convolutional neural networks (DCNN) [58, 82, 133]. Baker et al. [6] found that DCNNs struggled to classify objects in images based on their global shape. In one experiment, they presented CNNs with object that preserved the global shape but were filled with textures from other objects. The networks showed a strong bias for classifying based on textures rather than shapes. Further experiments revealed that CNNs could not reliably classify objects based on outlines alone, indicating a reliance on local features rather than global shapes. These findings suggest that while CNNs can access local shape features in images, they do not form global shape representations crucial for human-like object recognition.

While the challenges of object recognition in 2D images are well-documented, the transition to 3D object recognition in the point cloud display introduces additional complexities and opportunities for both human and machine perception research. Recent advancements in deep learning have led to novel architectures designed specifically for 3D point clouds. PointNet, for example, introduced innovative ways to directly process point clouds by learning spatial features from raw 3D data [120]. Another significant development is Dynamic Graph CNN (DGCNN), which builds on the idea of graph neural networks to construct local neighborhoods and learn local geometric features dynamically [149]. These models have shown human-like performance in various 3D recognition tasks, but whether they acquire 3D representations similar to humans remains unknown. The most recent advancement is to adopt the visual transformer architecture in 3D object recognition, such as the Point Transformer [169]. These

transformer-based models also reach human-level recognition performance for 3d object recognition.

In this chapter, we aim to compare the 3D object recognition in humans and deep learning models. Through a set of experimental manipulations, we analyze how both humans and models recognize 3D objects, particularly in challenging conditions where local features are disrupted or the objects are presented from unusual perspectives. We then conducted ablation studies to pin down the core computational mechanisms underlying 3D object recognition.

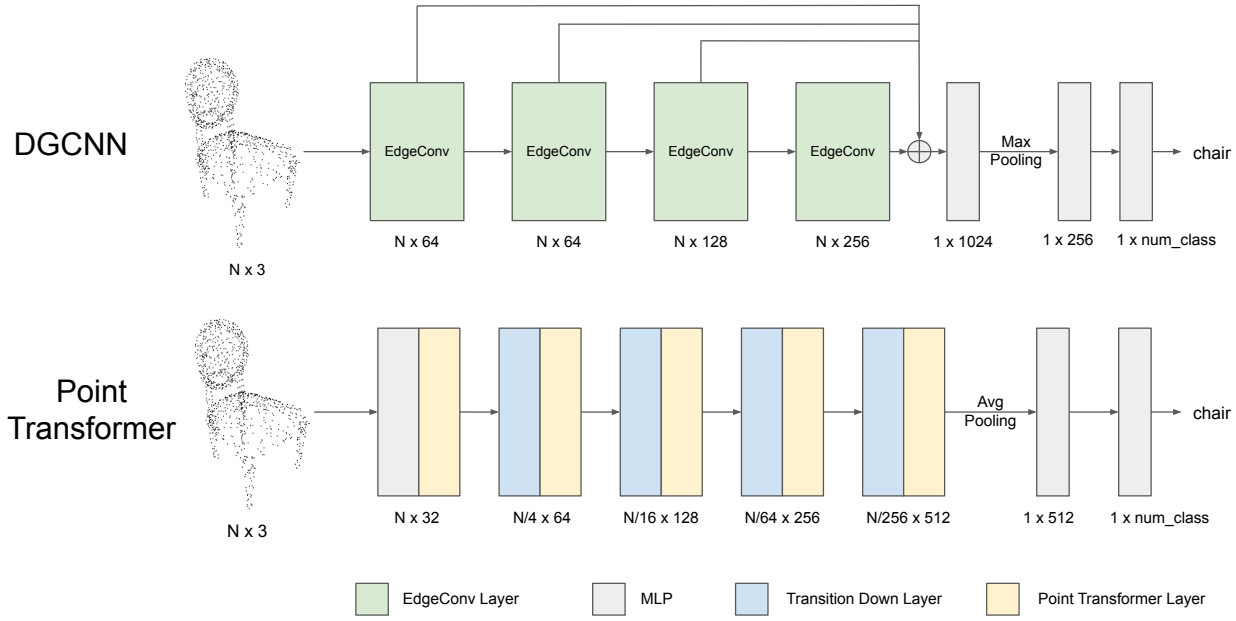


Figure 4.1: The architectures of DGCNN (top) and Point Transformer (bottom). MLP: Multi-layer perceptron, consisting of multiple fully connected layers. $\oplus$: concatenation.

4.2 Modeling Methods

4.2.1 Dataset

We used stimuli selected from the publicly available point cloud dataset, ModelNet40 [158]. The ModelNet40 dataset includes a set of 3D CAD models from 40 object categories. There are 12,311 3D CAD objects in total, with 9,843 objects used for training and 2,468 for testing.

The point clouds consist of 1,024 points sampled uniformly from the surface of each 3D CAD model.

4.2.2 Deep Learning Models

We evaluated two deep learning models for 3D object recognition from point cloud data: a convolution-based model (DGCNN; [149]) and a transformer-based model (Point Transformer; [169]). Both models take 3D coordinates of sampled points and generate feature embeddings for object classification. A more detailed illustration of the model architecture is in Figure 4.1 and a summary of the comparison between the two models are in Table 4.1.

DGCNN processes point clouds as graphs using EdgeConv layers, which extract local geometric features by comparing each point to its neighbors in feature space. Unlike traditional CNNs that operate on regular grids, DGCNN dynamically updates the neighborhood structure in each layer, enabling it to capture complex 3D shapes through local feature aggregation and global max pooling. Our implementation used 1,024 points and 20 nearest neighbors, based on the publicly available pretrained model.

Point Transformer adopts a self-attention mechanism inspired by transformer architectures in NLP and vision. Each Point Transformer layer integrates information from neighboring points using attention weights based on spatial and feature similarity. Transition Down layers perform hierarchical downsampling, enabling the model to capture both local and global shape features. This design allows for adaptive focus on informative regions of the input, enhancing recognition performance.

While both models operate on point cloud inputs, DGCNN emphasizes local geometric structure, whereas Point Transformer combines hierarchical pooling and attention to capture context-dependent relationships. To align with human experiments, we trained both models on the ModelNet40 dataset and extracted logits corresponding to the ten object categories shown to participants, enabling a direct comparison of recognition performance. Both models were trained using the same data augmentation procedures, including random point dropout,

random scaling, and random shifting of the input point clouds.

DGCNN	**Point Transformer**
Feature embeddings of a reference point are modulated by adding the weighted feature differences from the neighbor points	Feature embeddings of a reference point are computed using feature embeddings of neighbor points weighted by similarity
Neighbor points are defined in feature space, changing from layer to layer	Neighbor points are defined based on distance in 3D space
No spatial resolution change	Downsampling: spatial resolution change from fine to coarse
No position (3D coordinates) encoding	Position encoding added to each layer

Table 4.1: Comparison and key differences between a convolution-based model (DGCNN) and a transformer-based model (Point Transformer).

4.3 Experiments

4.3.1 General Human Study Procedure

We use the same general procedure across all experiments unless otherwise specified. Participants were instructed that they would view a rotating point cloud object during each trial. The stimulus was displayed for 3 seconds, after which ten buttons, each labeled with a different object name, appeared for selection. Their task was to select the object category that best matched the presented point cloud object. The ten object categories were airplane, bottle, bowl, chair, cup, lamp, person, piano, stool, and table.

Participants first completed a practice trial showing a rotating point cloud of a plant. They had to select the correct object category before proceeding to the experimental trials. If participants selected a wrong object category during practice, then the trial was repeated, and a hint message was displayed below the ten category buttons, directing participants to select the "Plant" button. This practice trial aimed to familiarize participants with the point-cloud display and ensure they understood the recognition task. The object category in the practice trial was not included in the subsequent experimental trials.

The experimental trials were similar to the practice trial, except that no feedback was

provided. At the end of the experiment, demographic information was collected, and participants were presented with debriefing information about the study.

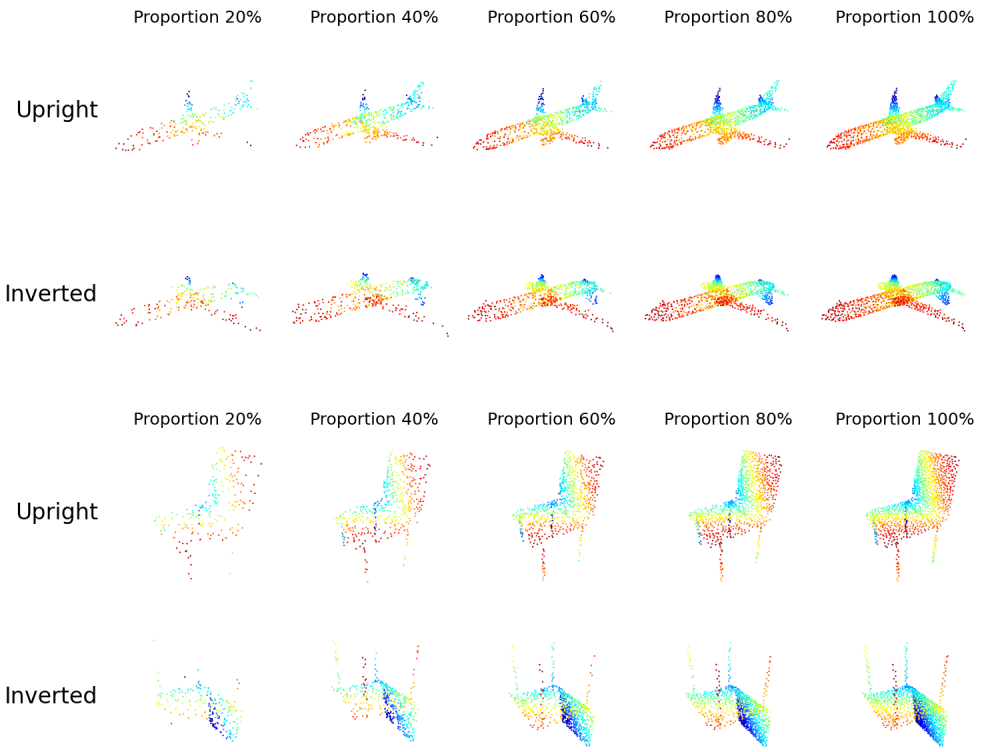


Figure 4.2: Downsampled and inverted point cloud stimuli used in Experiment. A random subset of points from the point cloud is retained in varying proportions. Colors represent depth, with red indicating proximity and blue indicating distance. In the actual experiments, the stimuli were presented as black points with rotation in depth, viewed from a horizontal viewpoint.

4.3.2 Experiment 1: Point density and object orientation

In the first experiment, we investigated the recognition performance of both human participants and neural network models under two conditions: (1) varying point cloud sparsity and (2) presentation of point clouds in an inverted (upside-down) orientation. By combining these two conditions, we examined the robustness of human and model recognition across different levels of detail and atypical viewpoints.

Participants

Two groups of participants were recruited through the UCLA Subject Pool. For the sparse point cloud condition, 56 participants were recruited (45 female, 11 male), with one participant excluded for reporting a lack of seriousness, resulting in a final sample of 55 participants (mean age = 19.8, SD = 1.4). For the inverted point cloud condition, 47 participants were recruited (40 female, 7 male), with a mean age of 20.4 years (SD = 1.5). The average completion time for both conditions was approximately 10 minutes, with a slight variance between groups.

Stimuli

The stimuli were selected from the test set of the ModelNet40 dataset to ensure a fair comparison between human participants and deep neural network (DNN) models. We selected 7 objects from each of the ten categories, where each stimuli was transformed under two conditions 1) sparse point clouds: each point cloud were randomly downsampled to seven proportions (20%, 30%, 40%, 50%, 60%, 80%, and 100% of 1024 points). 2) Inverted point clouds: each stimulus from condition 1 were flipped upside down. Therefore, we have 7 objects x 7 downsampling ratios x 10 categories = 490 stimuli for the upright condition and the inverted condition. Each point cloud was presented as a GIF rotating 10 degrees per frame around the vertical axis. The GIF is displayed at 10 frames per second, completing a full 360-degree rotation in 3.6 seconds.

In the experiment, the participants are split into two groups, where one group viewed the upright point clouds and the other viewed the inverted point clouds. Each participant viewed one object exemplar only once, with a random permutation of proportions. In other words, each participant viewed all seven objects from one category, and each object was displayed in the point cloud format with a different downsampling proportion. This resulted in a total of 70 trials per participant. The trials were randomized for each participant. For the models, we used all 490 objects x 2 conditions = 980 objects for testing.

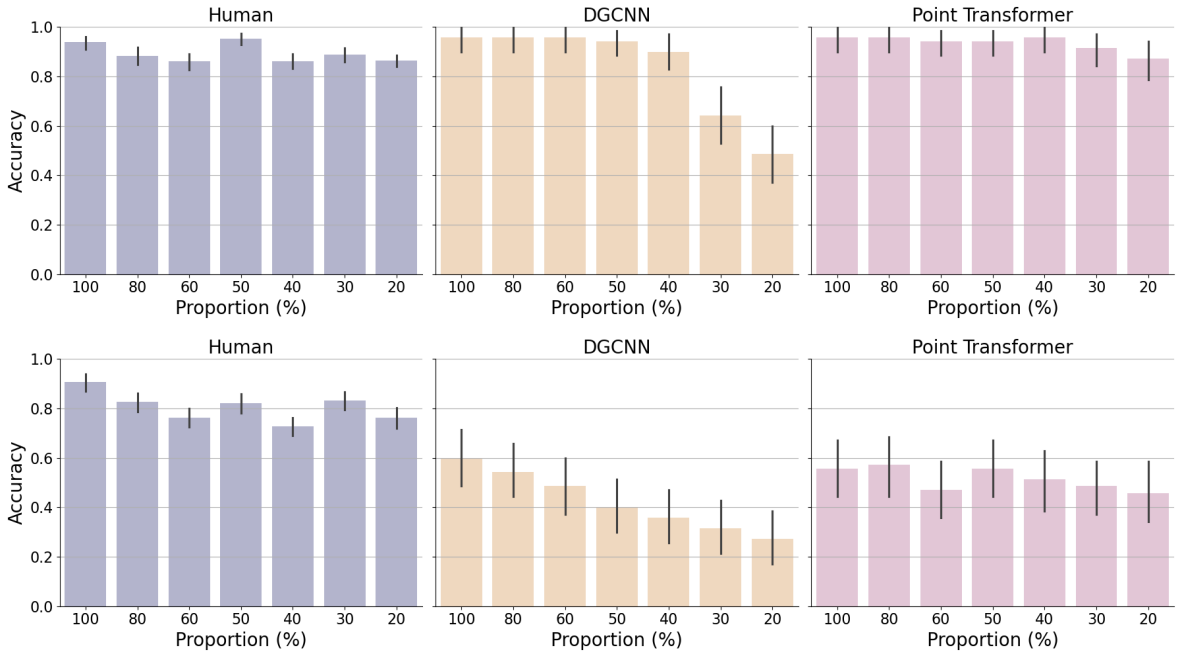


Figure 4.3: Accuracy of human participants and models as a function of point density for upright point clouds (top) and inverted point clouds (bottom). Error bars represent the 95% confidence interval.

Results

The results of the experiment are presented in Figure 4.3. Despite the sparse information provided in point cloud display, human participants consistently demonstrated high accuracy across all levels of point density. Their performance ranged from 86.2% to 95.3%, with only a slight decline as the point density decreased. For instance, when the point clouds were downsampled to 20% of the original points, the mean accuracy was 86.4% (CI = [83.5%, 89.2%]), and at 30%, the mean accuracy was 88.9% (CI = [86.3%, 91.5%]). This suggests that human participants are highly resilient to reduced point density, maintaining reliable recognition performance even with sparsely represented 3D objects.

The performance of the DGCNN model, however, was markedly affected by the reduction in point density. While the model's accuracy approached human performance at higher point densities (above 30%), it declined sharply at lower proportions. Specifically, the accuracy dropped to 64.3% at 30% point density and 48.6% at 20%. In contrast, the Point Transformer

model exhibited greater robustness to sparsity. Its accuracy remained high across all levels of point density, ranging from 94.3% at 50% to 87.1% at 20%, showing similar performance robustness as humans.

In the inverted condition, human participants showed the inversion effect with lower accuracy than upright condition, but still significantly greater than the chance level. Furthermore, humans consistently outperformed the machine learning models, highlighting their adaptability to changes in perspective. Note that the Point Transformer model performed better than the DGCNN model overall and also showed less reduction in the inverted condition, particularly at lower point densities.

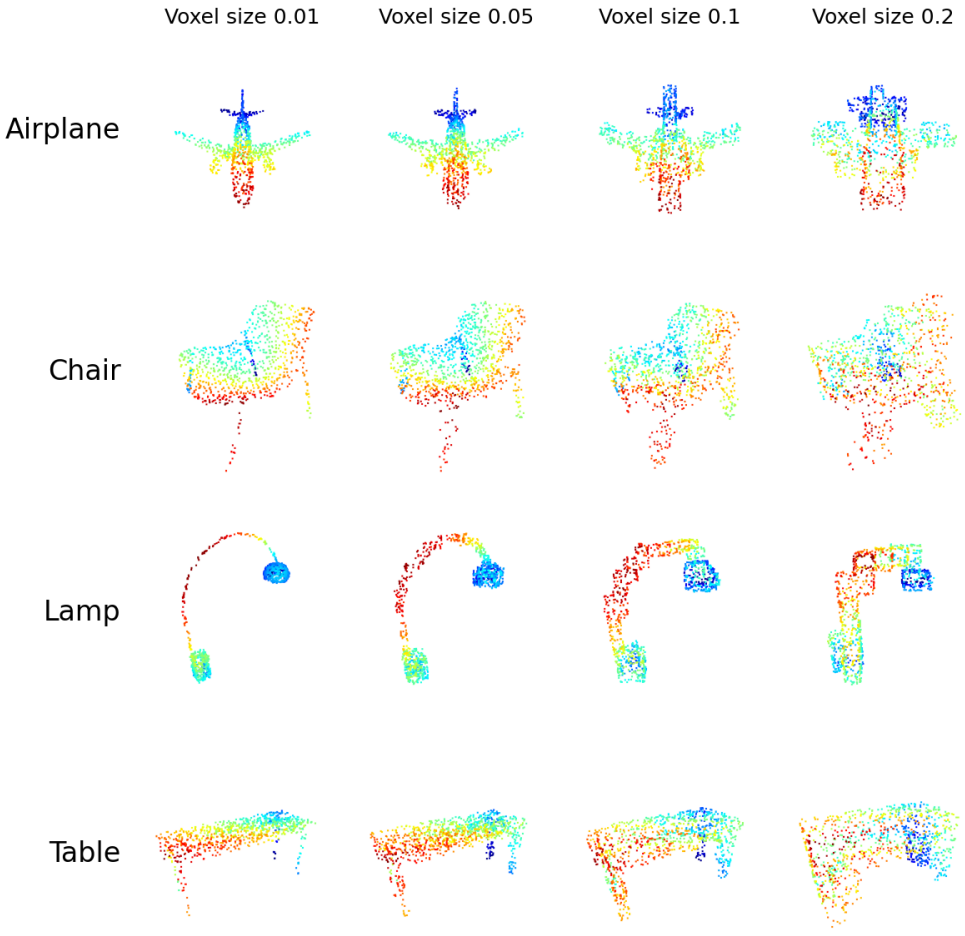


Figure 4.4: Lego-like point cloud stimuli for Experiment 2. Points were uniformly sampled from voxel surfaces converted from point clouds, with varying voxel sizes determining resolution.

4.3.3 Experiment 2: Lego-Like Point Clouds

In Experiment 2, we aimed to disrupt the local geometric features of point clouds while preserving their global shape of 3D objects. This was achieved by converting point clouds into voxel grid displays, analogous to reducing the resolution of an image. The larger the voxel size, the lower the spatial resolution of the point cloud. The process involved generating a voxel grid from the point cloud, sampling points on the voxel surfaces, and normalizing the sampled points. The stimuli section below details the methodology and the corresponding implementation. The idea of introducing Lego-like point clouds was inspired by the sawtooth images from Baker et al. (2018), where sawtooth effects were added to silhouette images to disrupt local contour features. In our 3D point cloud display, we similarly aimed to disrupt local curvature features by converting point clouds into Lego-like representations while keeping the global shape almost unchanged.

Participants

A total of 60 participants were recruited through the university's Subject Pool. The sample comprised 49 females, 9 males, 1 non-binary individual, and 1 participant who preferred not to disclose their gender. The mean age of the participants was 20.6 years (SD = 3.9). The average completion time for the experiment was 6.3 minutes (SD = 2.3).

Stimuli

We first converted each point cloud into a voxel grid representation using the Open3D library. A point cloud consists of discrete data points capturing an object's surface geometry, while voxels are small cubic units dividing 3D space into a regular grid, analogous to pixels in 2D images but extending into the third dimension. To convert a point cloud into a voxel grid, we superimposed a voxel grid over the point cloud, determine voxel occupancy by checking which voxels contain points, and then created a structured, blocky representation of the object.

After obtaining the voxel grid representation of the point clouds, we sampled points

uniformly on the surface of this grid. These sampled points were aggregated to form a new point cloud representing the voxel grid. Varying the voxel size allowed us to sample the point cloud at different resolutions, with larger voxel sizes resulting in lower resolution. A visualization of the Lego-like point cloud stimuli across different voxel sizes is presented in Figure 4.4.

The resulting point cloud from the voxel sampling was then normalized to center it at the origin and scale it to fit within a unit sphere. This normalization does not affect the GIFs presented to human participants but is crucial for ensuring consistent input for machine learning models.

For the stimuli, we used the same 70 objects from 10 categories as in Experiment 1. Each object was sampled with four different voxel sizes: 0.01, 0.05, 0.1, and 0.2. Each participant viewed four random but non-overlapping objects from each voxel size within each category, constituting a total of 40 stimuli per participant.

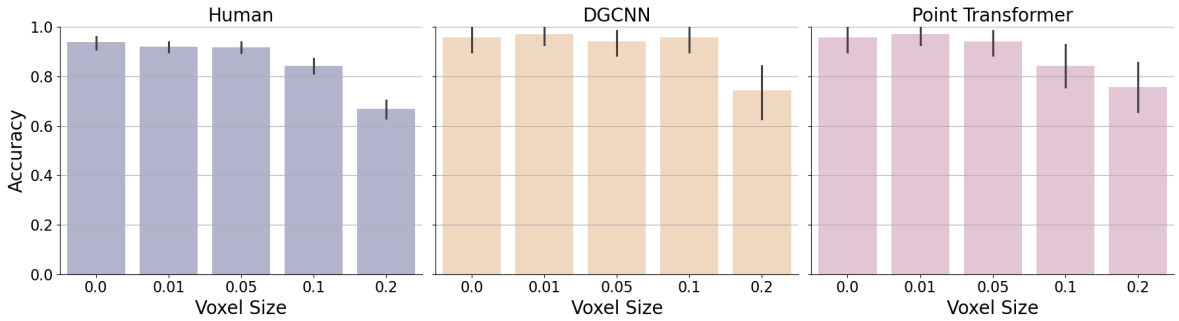


Figure 4.5: Accuracy of human participants and models on Lego-like point clouds with varying voxel sizes. Error bars represent the 95% confidence interval.

Results

The results of Experiment 2 are depicted in Figure 4.5. Human participants demonstrated relatively stable accuracy across different voxel sizes. Human performance remained almost the same when voxel size increased from 0.0 (mean accuracy 93.8%, CI = [91.8%, 95.8%]) to 0.05 (mean accuracy 91.8%, CI = [89.5%, 94.1%]). Then accuracy gradually reduced as the voxel size increased to 0.1 (mean accuracy 84.3%, CI = [81.3%, 87.3%]) and 0.2 (mean

accuracy 66.8%, CI = [62.9%, 70.7%]). This suggests that human participants can effectively recognize objects even when local geometric features are jittered, provided the global shape remains intact.

Both models achieved performance levels comparable to human observers at lower voxel sizes, suggesting their ability to generalize well with minimal distortion. However, as voxel size increased, performance began to decline across all groups. Notably, the Point Transformer exhibited a gradual decline in accuracy that closely mirrored the human trend, suggesting that it captures a similar sensitivity to degradation in spatial resolution. In contrast, DGCNN maintained stable performance up to voxel size 0.1 but showed a sharp drop at 0.2, indicating a threshold-like effect leading to brittle performance. Specifically, DGCNN's accuracy dropped from 95.71% at voxel size 0.1 to 74.29% at voxel size 0.2.

We further analyzed the correlation between model and human accuracy patterns across all experimental conditions, in the next section titled "Model and Human Performance Correlation" (also see Figure 4.8). This section provides a more detailed quantitative comparison of model alignment with human behavior across the two experiments.

Investigating Mechanisms Underlying Global Shape Bias

To identify which computational mechanism in the Point Transformer model contribute to its global shape bias, we systematically evaluated three primary differences between the Point Transformer and the DGCNN: (1) the Attention mechanism, (2) Position Encoding, and (3) the Downsampling supporting hierarchical pooling and abstraction of 3D shapes. By selectively removing each component, we developed multiple variants of the Point Transformer and assessed their performance on the same downsampled point cloud stimuli used in Experiment 1.

Figure 4.6 illustrates the accuracy of different Point Transformer variants as a function of point density, stimuli used in Experiment 1. The original Point Transformer model, the "Original" variant in the plot, achieves the highest performance across all proportions. The

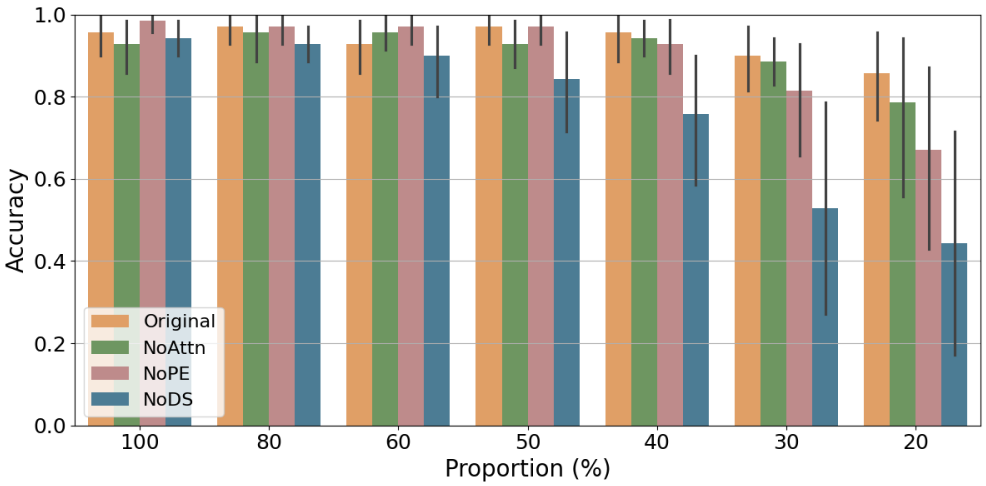


Figure 4.6: Accuracy of variants of Point Transformers on downsampled point clouds. Original = the original Point Transformer, NoAttn = No attention, NoPE = No position encoding, NoDs = No downsampling.

"NoAttn" and "NoPE" variants, which removes the attention mechanism and the position encoding mechanism respectively, experiences a mild accuracy drop compared to the original. This suggests that while attention and position encoding are beneficial, the model maintains substantial effectiveness even in their absence.

In contrast, the variant lacking the downsampling mechanism ("NoDS") demonstrated performance comparable to the original model at higher point densities but exhibited significant accuracy declines at lower point densities. This result underscores the critical role of the downsampling component in generalizing across varying degrees of point sparsity. Downsampling facilitates hierarchical representation within the model, enabling a global integration of shape information and reducing dependence on local features.

Interestingly, this effect mirrors the visual pathways in the human brain, particularly the hierarchical processing in ventral pathway, indicated by increased receptive fields of neurons from the low-level to high-level visual areas. Similarly, downsampling in Point Transformer enforces a global shape bias, helping the model recognize objects even when the number of available points is significantly reduced. By progressively reducing the number of points covering the entire objects, downsampling allows the model to integrate local features into a

coherent global representation, making it more robust to various data transformations.

4.3.4 Enhancing DGCNN with Downsampling

Given the critical role identified for the downsampling mechanism in fostering a global shape bias in the point transformer model, we next examined whether integrating this computational mechanism into the DGCNN architecture could enhance its performance. We introduced the Transition Down layer from the Point Transformer into each EdgeConv layer of the DGCNN, matching the structural depth of the original Point Transformer.

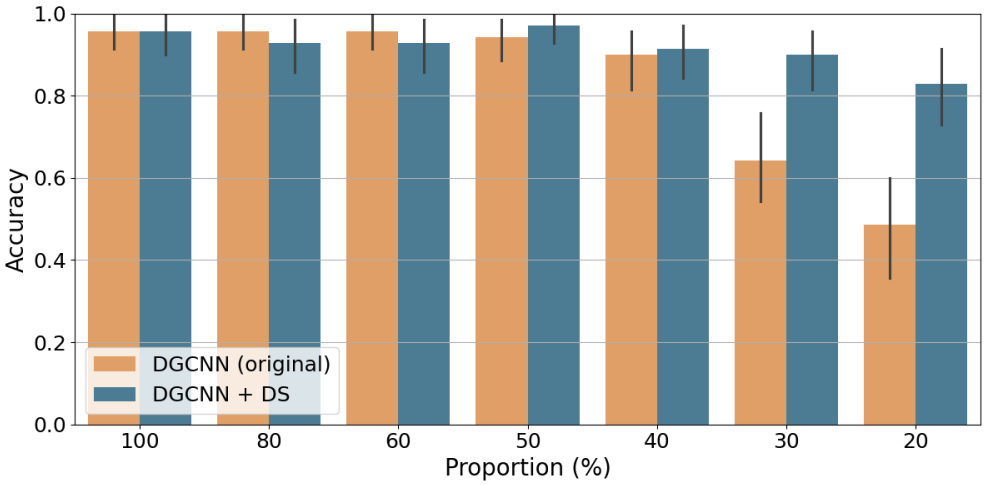


Figure 4.7: Accuracy of variants of DGCNN model on downsampled point clouds. DGCNN + DS = DGCNN model with additional Downsampling layer after each EdgeConv layer.

As demonstrated in Figure 4.7, incorporating the Transition Down layer significantly improved the DGCNN's robustness to point density. The original DGCNN exhibited significant performance degradation at point densities less than 40% . In contrast, the modified DGCNN with the Transition Down layer ("DGCNN + DS") increased the the accuracy for the low point density condition. For example, performance increased from 0.486 to 0.829 for the 20% point density, aligning closely with both human performance and the original Point Transformer.

The addition of the Transition Down layer effectively compels the model to construct more abstract, hierarchical representations of 3D objects, shifting its focus towards global

shape characteristics rather than relying excessively on detailed local features. Crucially, this demonstrates that downsampling can effectively bridge the gap between seemingly distinct model architectures, transformer-based and convolution-based, highlighting a shared computational mechanism toward robust shape recognition. While it is commonly assumed that the attention mechanism in transformer-based architectures primarily enables better generalization, our results clearly demonstrate that it is actually the downsampling mechanism driving this effect. The success of downsampling across both transformer and convolution-based models challenges conventional assumptions about their inherent differences, suggesting that hierarchical abstraction strategies may universally benefit deep learning models.

4.3.5 Model and Human Performance Correlation

We compared the accuracy patterns of the DGCNN and Point Transformer, along with their respective variants, DGCNN + DS and Point Transformer noDS, to human accuracy patterns across experimental conditions in the two experiments. By pooling performance data across all tested conditions, we computed Pearson correlations between each model's accuracies and human responses.

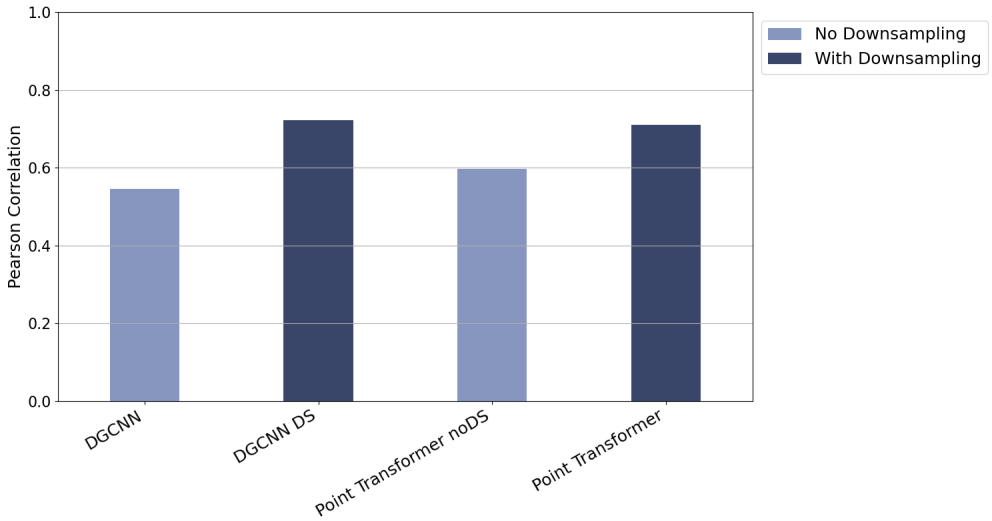


Figure 4.8: Correlation between model and human accuracy patterns across conditions. Models with downsampling show significantly stronger alignment with human.

As shown in Figure 4.8, both model variants that incorporate the downsampling mechanism (DGCNN + DS and Point Transformer) exhibited substantially higher correlations with human performance than their non-downsampling counterparts (DGCNN and Point Transformer noDS). Specifically, the correlation between human performance and DGCNN + DS was r = 0.721 (p <0 .001), and for the original Point Transformer it was r = 0.711 ($p < 0.001$), both significantly higher than the correlation with the original DGCNN (r = 0.545, p = 0.016). These differences were statistically supported by comparisons such as DGCNN vs. DGCNN + DS (z = -15.570, $p < 0.0001$) and DGCNN vs. Point Transformer (z = -8.163, $p < 0.0001$), indicating that the introduction of downsampling yields not only improved model robustness but also closer alignment to human accuracy profiles.

This finding provides converging evidence that downsampling plays a central role in eliciting human-like recognition of 3D shapes. The enhanced human-model correspondence across two structurally distinct architectures further underscores the broad applicability of this mechanism.

4.4 Conclusions and Discussions

Across two behavioral experiments and a series of modeling analyses, we demonstrated that humans exhibit strong robustness in recognizing 3D objects, even under challenging conditions such as sparse input, inversion, or disrupted local geometry. While current deep learning models—DGCNN and Point Transformer—achieved comparable accuracy to humans under standard conditions, their robustness under distortion varied substantially.

In particular, the Point Transformer model consistently mirrored human performance trends, showing gradual declines under increasing point sparsity and local geometry disruption. In contrast, the DGCNN model was more brittle, with sharp drops in performance in the untrained conditions. These findings indicate that models relying predominantly on local geometric features, such as DGCNN, lack the holistic processing strategies characteristic of

human perception.

Our ablation studies further revealed that the downsampling mechanism in Point Transformer is the key component supporting its robustness and global shape bias in 3D object recognition. By introducing this mechanism into DGCNN, we significantly improved its performance and alignment with human responses. Importantly, our findings challenge the common assumption that attention mechanisms are the main contributors to strong generalization in transformer-based models; instead, hierarchical abstraction through downsampling plays the more critical role for 3D shape recognition.

Taken together, our results highlight the importance of global shape representations in achieving human-like 3D object recognition. Integrating cognitively inspired mechanisms, such as hierarchical downsampling, into deep learning models can improve their generalization and robustness, bringing them closer to the perceptual strategies employed by the human visual system.

Chapter 5

Visual Analogy Using Part-Based Representation

5.1 Introduction

Despite the many recent advances in work on artificial intelligence (AI), capturing core characteristics of human thinking remains an elusive goal. One of the canonical examples of human intelligence is the ability to reason by analogy—recognizing similarities based on relations between entities, even when the entities themselves are dissimilar. Computational models of analogical reasoning developed in cognitive science and AI fall into two broad classes (see Figure 5.1). An early proposal in cognitive science construed analogy-making as "high-level-perception", a highly context-dependent process in which building representations, and the processes operating over those representations, are in principle inseparable [23]. Early advocates never developed a computational model that actually implemented this theoretical position, instead relying on hand-coded representations tailored to toy problems [61]. However, contemporary AI work on visual analogy has succeeded in modeling the acquisition of

The contents of this chapter appeared in paper [67, 68].

perceptual representations suitable for analogical reasoning by end-to-end manner training from raw inputs of stimuli coded as image pixels [132]. Such task-specific models generally assume that training and testing data consist of independent and identically distributed (i.i.d.) samples from an unknown probability distribution. This approach has been applied with some success to solving visual analogy problems inspired by Raven's Progressive Matrices (RPM) [127]. After extensive training with RPM-like problems, deep neural networks have achieved human-level performance on test problems with similar basic structure [60, 166].

Early critics of this high-level-perception approach to analogy articulated a number of theoretical reasons to doubt that it accurately characterizes mechanisms underlying human analogical reasoning [40]. However, contemporary AI models based on end-to-end learning offer much more complete and powerful instantiations of the basic approach. In the present chapter we evaluate two leading examples of these models, focusing on a general limitation that such models face: difficulty in generalization. Specifically, the success of these models depends on high similarity between training problems and test problems (due in part to their assumption of i.i.d samples), and on datasets of massive numbers of RPM-like problems (e.g., 1.42 million problems in the PGM dataset [8], and 70,000 problems in the RAVEN dataset [165]. For example, Zhang and colleagues [166] used 21,000 training problems from the RAVEN dataset, and 300,000 from the PGM dataset.

This dependency on direct training in a reasoning task using big data makes the end-to-end approach qualitatively different from human analogical reasoning. In order for the RPM task to be of interest as a measure of fluid intelligence—the ability to manipulate novel information in working memory—extensive pretraining on RPM-like problems must necessarily be avoided [134]. When the RPM task is administered to a person, "training" is limited to general task instructions. Human analogical reasoning is a prime example of an out-of-task situation involving zero-shot or few-shot learning—the ability to make inferences with minimal prior exposure to structurally similar problems, a core characteristic of human intelligence [84]. There is evidence that humans can deal with both out-of-task and

out-of-distribution situations. For example, after observing a handful of examples of a visual concept, people can classify (out-of-distribution) and generate (out-of-task) new instances of the concept by recursive application of a rule [83].

An alternative approach to visual analogy maintains that the reasoning process is domain-general, and separable from the acquisition of perceptual representations (see Figure 5.1 bottom panel). In this domain-general approach, analogical mapping is viewed not as a task to be trained directly, but rather as an inference problem based on comparisons between representational structures. The core of analogical reasoning is assumed to be based on domain-general processes that operate on representations of entities, their attributes, and relations between entities [65, 98]. For visual problems, the effective representations are those that support human perception of a three-dimensional world consisting of interrelated objects and their parts, segregated by boundaries between elements [14]. These representations provide compositional structures that are not created to solve one particular task; rather, they are used to solve multiple computational problems that arise for intelligent systems. (See [11] for a similar contrast between goal-specific and goal-general representations in human vision.)

Some recent computational models of visual processes have this task-general character. For example, a model trained only on unoccluded objects can classify objects when they are occluded (out-of-distribution) [174]. Similarly, a model trained to classify objects robustly to occlusion can, as a side effect, estimate the amodal boundary of an object (out-of-task transfer) [137]. Such representations appear to be closely linked to the human capacity to learn from few examples (sometimes zero-shot or one-shot learning) and to generalize learning to novel situations [7]. Once task-general perceptual representations have been acquired, analogical reasoning can be performed by comparing relations across analogs to assess their similarity. Analogy thus emerges from the ability to represent relational structures and compute their similarity [104].

Although theoretically appealing, analogy models based on structural comparison have usually side-stepped the problem of representation learning, often relying on representations

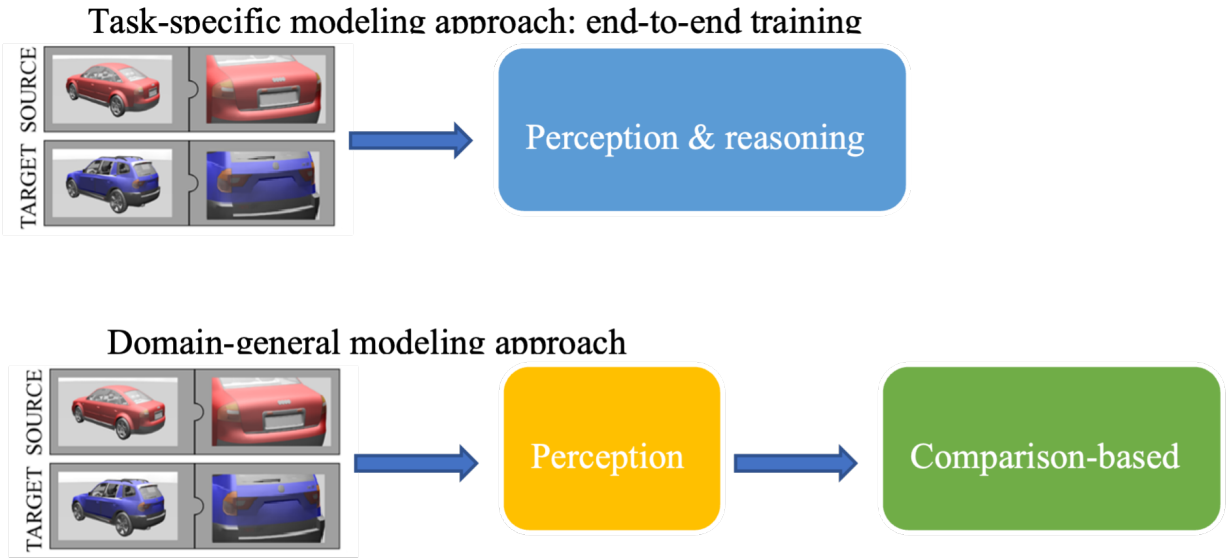


Figure 5.1: Overview of the two modeling approaches. The task-specific modeling approach uses end-to-end training with a large number of analogy problems. The domain-general modeling approach replies on perceptual representations learned for visual tasks, coupled with comparisons between representational structures.

hand-coded by the modeler. A practical disadvantage of this reliance is that such models have been unable to extract representations directly from input images or texts, making it difficult to compare their performance to that of end-to-end models using the same training and test data. To compare the task-specific and domain-general computational approaches, the present chapter introduces a new database based on pixel-level images of realistic 3D objects—cars—which can be used to generate massive numbers of training data for deep neural networks, and also to create analogy problems suitable for both humans and models. After obtaining benchmark data from human experiments, we compared human performance both to that of task-specific analogy models based on deep neural networks and to that of a new domain-general model that we developed, which extracts and compares representations of cars and their component parts. To create these representations, we use the dataset to train a vision model to identify parts of cars, so that the relations between a whole object and its parts can be identified. Once these representations have been learned, we show that analogy problems can be solved simply by computing relational similarity between analogs,

without any additional training on the reasoning task itself (thus instantiating out-of-task testing).

The structure of this chapter is as follows: In Experiment 1, we begin with a comparison of task-specific and domain-general models with respect to their ability to reproduce human-like response patterns on traditional, four-term A:B::C:D visual analogy problems. In Experiment 2, we go on to compare these modeling approaches using more open-ended mapping problems. In both experiments, all problems required finding analogical correspondences based on realistic images of cars. To foreshadow our results: Across both experiments, we show that our model implementing a domain-general approach to analogical reasoning matches human performance more closely than two alternative models implementing a task-specific approach.

5.2 Experiment 1: Four-Term Visual Analogies with Objects and their Parts

In previous studies of visual analogy, researchers have often employed simplified visual stimuli to generate problems out of combinations of geometric shapes in RPM-like problems [98] and other meaningless two-dimensional forms [15,39]. Due to the high complexity of these problems, solving them often involves interleaved processes of representation-building and comparison, in which reasoners iteratively re-represent the same stimulus to support increasingly more informative comparisons [20]. In order to provide a clear comparison between task-specific and domain-general approaches to analogical reasoning, we sought to avoid any ambiguity in our stimuli that might promote iterative re-representation. Such ambiguity might add complexities beyond the scope of the models we are assessing here, and thus blur any performance differences between models instantiating each approach.

In order to focus on a contrast between intermixing the processes of representation-building and comparison (in a task-specific model) or separating them (in a domain- and task-general model), we used visual stimuli intended to evoke unambiguous representations. Specifically,

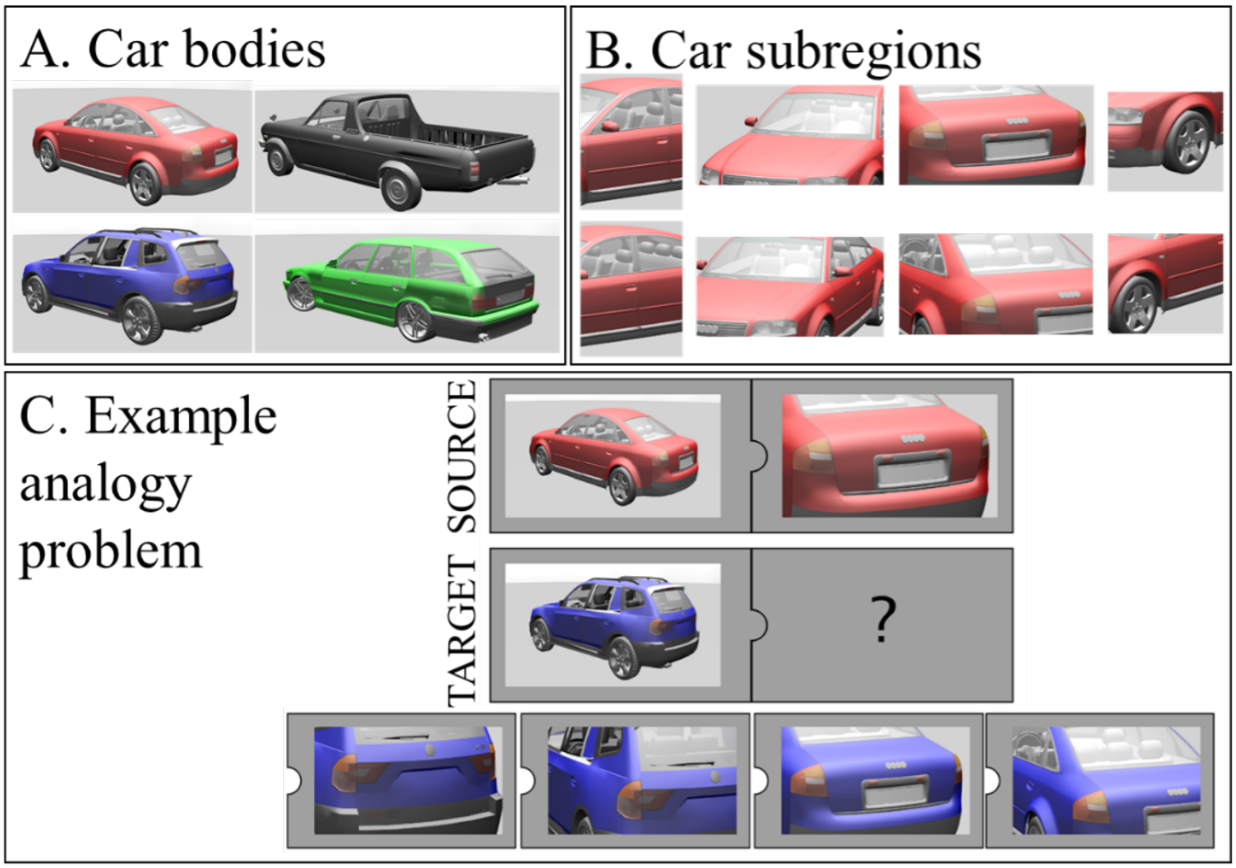


Figure 5.2: Experimental stimuli. Panel A: 3D car bodies, a sedan (top left), truck (top right), SUV (bottom left), and station wagon (bottom right). Panel B: subregions of the sedan car body, a door and window (leftmost column), hood and windshield (second column), trunk and bumper (third column), and headlight and front wheel (rightmost column). These subregions can be entire parts (top row), or pieces that do not correspond to entire parts (bottom row). Panel C: an example analogy problem in A:B::C:? format, constructed out of some of the images shown in panels A and B. The row of answer options includes the correct answer (leftmost), a wrong-subregion distractor (second), a wrong-car distractor (third), and a both-wrong distractor (rightmost).

we used visually rich stimuli depicting familiar objects in 3-D, namely cars (see examples in Figure 5.2 below). Their visual detail enabled variation in the texture, shading, and angle of the depicted objects without compromising experimental control over the analogy problems that they were used to generate. Moreover, the stimulus set is sufficiently extensive to provide a massive amount of data for training both deep learning models of analogy and a model based on comparison of compositional structures. The three computational models examined in the present chapter all avoid hand-coding of representations, instead taking raw pixel-level images as inputs to solve analogy problems.

The analogy problems in this dataset focus on part-whole relations. In general, human perception and thinking show sensitivity to part-whole relations across both visual and semantic domains. Structural description theories of object recognition, which include part relations, can account for viewpoint-invariant recognition in human vision [13]. Members of basic level categories are unified by having a large number of shared parts. For example, Tversky and Hemenway [142] showed that people converged in assessing the "goodness" of meaningful car components, such as headlights and doors. Critically, part-whole relations also permit early analogical reasoning. For example, children as young as four years old can map parts of a human body to their corresponding locations on a tree or mountain [46].

In the current study, the source analog (A:B) in each problem was an image of a car body (Figure 5.2A) paired with a subregion of it (Figure 5.2B). Each problem was presented as a four-term analogy in A:B::C:? format (Figure 5.2C). The target was a whole-car image of a different car type than the source, paired with a set of four alternative images of subregions as options to complete the analogy (see Methods). The problems systematically varied the four subregion images: [1] the spatial alignment between images of source and target cars, which were generated by using the same or different orientations of the whole-car images and the subregion images in an analog; [2] whether or not images of analogous subregions depicted unitary car components (part versus piece conditions); and [3] the visibility of analogous subregions in the whole-car images (visible-component versus invisible-component conditions). These experimental manipulations enabled us to test whether the same manipulations to experimental stimuli that affect human performance also has the similar impacts on model performance. This approach to model evaluation can provide greater insight than merely correlating model performance with human data without systematically manipulating independent variables in experiments.

We predicted that each of these manipulations would affect human performance and thus provide qualitative effects against which to ultimately compare alternative modeling approaches. It is known that spatial alignment between analogs helps children and adults

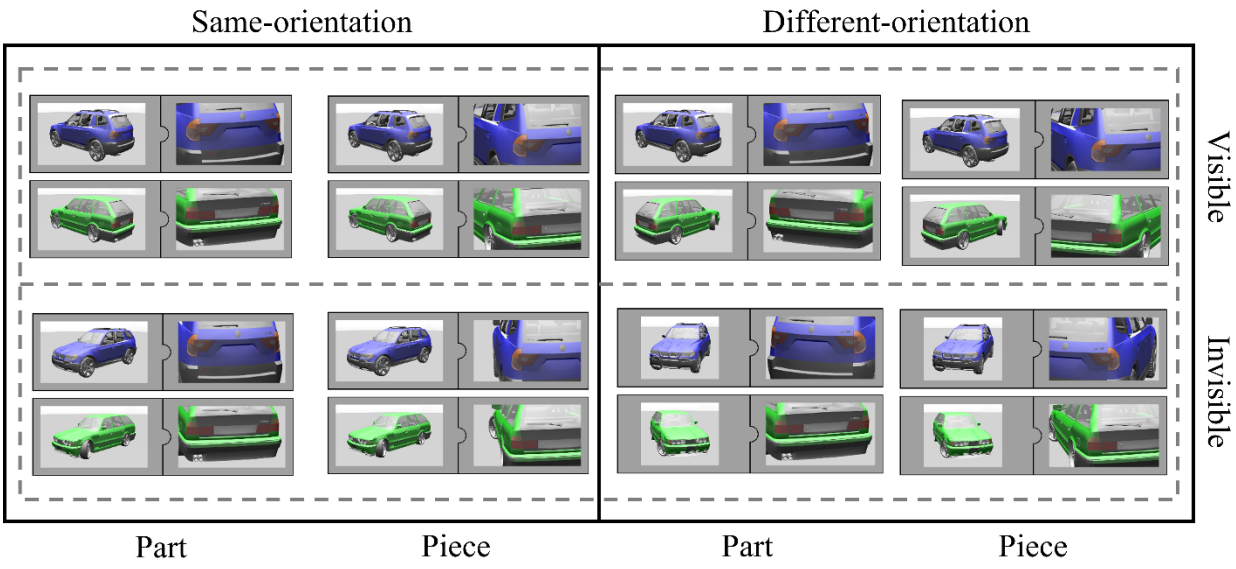


Figure 5.3: Examples of each of the 8 problem types in the four-term visual analogy dataset tested in Experiment 1. For each problem, the correct completion (D term) is shown; A:B (source) appears on top, and C:D appears on bottom.

find visual similarities and differences between them [107, 171]; accordingly, we predicted that people would be more successful in comparing vehicles with the same rather than different orientations. Based on the work of Tversky and Hemenway [142], we also predicted that mappings based on natural parts on which people agree (e.g., trunk, front door) would be more accurate than mappings based on pieces that cut across part boundaries (e.g., top-back region, side region). Finally, solving analogy problems when the component is not visible in the whole-car image presumably depends on knowledge of 3D structure. Such problems were expected to be more difficult than those in which the analogous components are visible in the whole-car images, because the visible condition is supported by direct perceptual information, whereas the invisible condition depends on knowledge not provided in the immediate input.

5.2.1 Four-term Analogy Task

We tested human participants and computational models on the same visual analogy dataset. Each problem in this dataset was a four-term analogy in A:B::C:? format. The source analog (A:B) was an image of a car body paired with a subregion of it. The target was a different-style

car body, paired with a set of four alternative images of subregions as options to complete the analogy. The problems systematically varied (1) the spatial alignment between images of source and target cars, (2) whether or not images of analogous subregions depicted unitary car components, and (3) the visibility of analogous subregions in the whole-car images.

Car images in analogy problems were generated using 3D car models taken from the ShapeNet 3D Core dataset [24]. Four subtypes of cars were selected from the dataset: sedan, SUV, station wagon, and truck, each represented by a single 3D car model, and Figure 5.2A shows these car bodies. These subtypes yielded a counter-balanced design including the following four pairs of car bodies, in which each body used twice (once as a source and once as a target): sedan (source) – SUV (target), SUV – station wagon, truck – sedan, station wagon – truck. Car subregion images were generated using the UDA-Part dataset [94]. This dataset provides detailed part annotations for 3D CAD models of vehicles. Each surface mesh on the CAD models was assigned a label from a set of 31 vehicle parts (e.g., back left door, front right door, front left windows, right mirror, bumper). Eight vehicle parts were selected from the dataset, each of which were spontaneously generated by participants in a part-labeling task: door, window, hood, windshield, trunk, bumper, headlight, and wheel. These eight parts were rated as 'good' or characteristic parts of cars [142]. Car subregions used in our visual analogy problems either fully (parts condition; top row in Figure 5.2B) or partially (pieces condition; bottom row in Figure 5.2B) depicted one of mentioned car parts. The images were rendered with Blender software using randomly selected textures. The virtual camera had a resolution of 1024×2048 and field of view of 90 degrees. Figure 5.2B shows examples of different car subregions used to construct the analogy problems.

Each analogy problem (see an example in Figure 5.2C) consisted of an incomplete 2 x 2 image array with the source whole-car image in the top left corner, a subregion of that car body in the top right corner, the target whole-car in the bottom left corner, and a question mark in the bottom right corner. Each problem was based on one of four pairs of car bodies: sedan (source) and SUV (target; pictured in Figure 5.2C), SUV and wagon, wagon and

truck, or truck and sedan. Below the array was a set of four answer options presented in a randomized order, and the task was to select the option that best fit the bottom right cell of the 2 x 2 array. In addition to the correct option (e.g., leftmost in Figure 5.2C bottom row), one option depicted a disanalogous subregion of the target car (wrong subregion) (e.g., second in Figure 5.2C), one option depicted an analogous subregion of the source car (wrong car) (e.g., third in Figure 5.2C), and a fourth option depicted a disanalogous subregion of the source car (both wrong) (e.g., rightmost in Figure 5.2C). When the correct answer was a part subregion (e.g., top left image depicting a door and window in Figure 5.2B), the wrong-subregion distractor depicted the corresponding piece subregion (e.g., bottom left image depicting a partial view of a door and window in Figure 5.2B). These assignments were reversed when the correct answer was a piece subregion.

Our entire test set consisted of 128 analogy problems that varied three factors: orientation of source and target cars (same vs. different), subregion type (part vs. piece), and subregion visibility in the corresponding whole car images (visible vs. invisible). Figure 5.3 includes examples of different problem types. Object orientations of cars in source and target images were either same or different in the analogy problems. For part problems, the source and target subregions fully depicted corresponding car components, whereas for piece problems these subregions included multiple partial components of car parts. For visible problems, the source and target subregions were visible from the images respectively depicting the source and target car bodies, whereas for invisible problems these subregions were occluded in the corresponding images. Crossing each of these factors yielded 8 problem types (see examples in Figure 5.3). Both humans and models completed these analogy problems, and we detail human data collection the next section.

5.2.2 Human Participants and Experiment Methods

Seventy-nine participants were recruited from Amazon Mechanical Turk (M_{age} = 42, SD_{age} = 11, age range = $[24, 73]$, 41 female, 38 male). Participants provided online consent

in accordance with the UCLA Institutional Review Board and were compensated with a monetary reward. In order to avoid excessive fatigue, each participant was asked to complete 32 problems out of the possible 128. To that end, we created four 32-problem subsets, each consisting of 4 problems falling into one of the 8 problem types. The 4 problem subsets were distributed evenly across participants, and varied which of the 4 car body pairs were combined with which of the 4 car subregions on each of the 8 problem types.

Before starting the task, participants were given a practice analogy problem with line drawings of a gas pump and a lawnmower and a battery and a flashlight (instantiating the common relation x powers y). For this practice problem and for the car analogy problems, participants were instructed "to select which option contains a picture that is related to the picture in the left bottom box in the same way as the pictures in the top two boxes." No training or practice on solving the car analogies was provided. No feedback was provided in the experiment.

5.2.3 Computational Models

We implemented models to instantiate task-specific and task-general modeling approaches, respectively. In order to instantiate a task-specific modeling approach to analogical reasoning based on extensive training with highly similar reasoning problems, we implemented two deep learning models, a Siamese Network [16] and a Relation Network [132]. A Siamese Network, which has been successfully applied to visual detection tasks [16] and to visual and verbal analogy tasks [130, 131], contains two or more identical subnetworks, emphasizing the role of comparisons among multiple inputs in forming comparable feature embeddings as visual representations (see top panel of Figure 5.4). The embeddings are compared to assess the similarity between inputs. A Relation Network was developed specifically to learn to solve relational reasoning problems (see bottom panel of Figure 5.4). This class of models has been successfully applied to RPM-like analogy problems [8] and query tasks [132].

We compare each of these networks to a new part-based comparison model (PCM), which

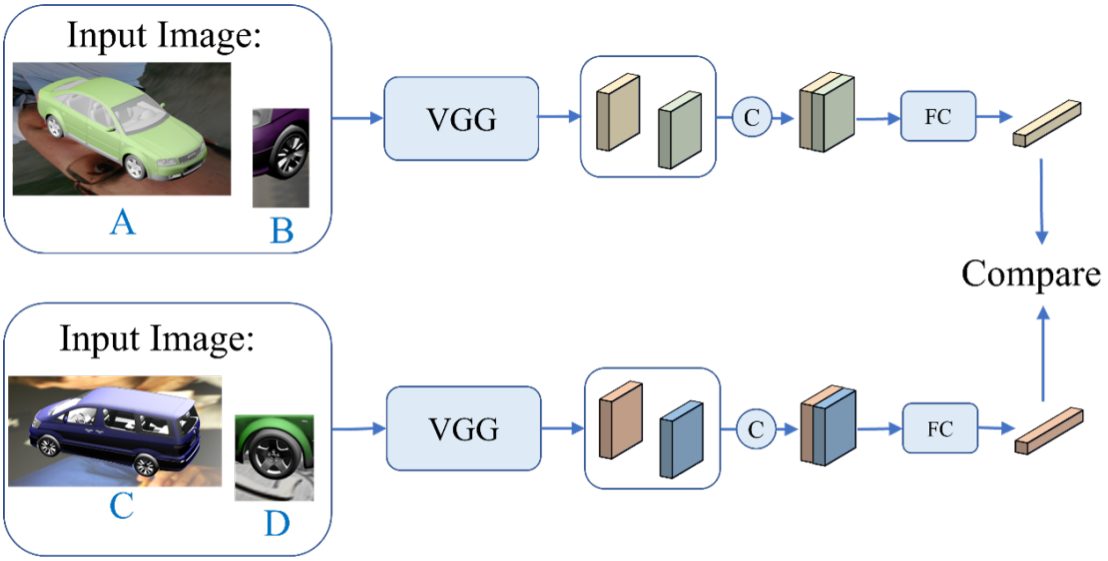
we introduce to instantiate the task-general representational comparison approach to analogy. This model employs image segmentation algorithms to extract visual features representing part-based structures for 3D objects (see Figure 5.5 below), and then compares representations by computing a generic measure of relational similarity.

Training Task-Specific Models

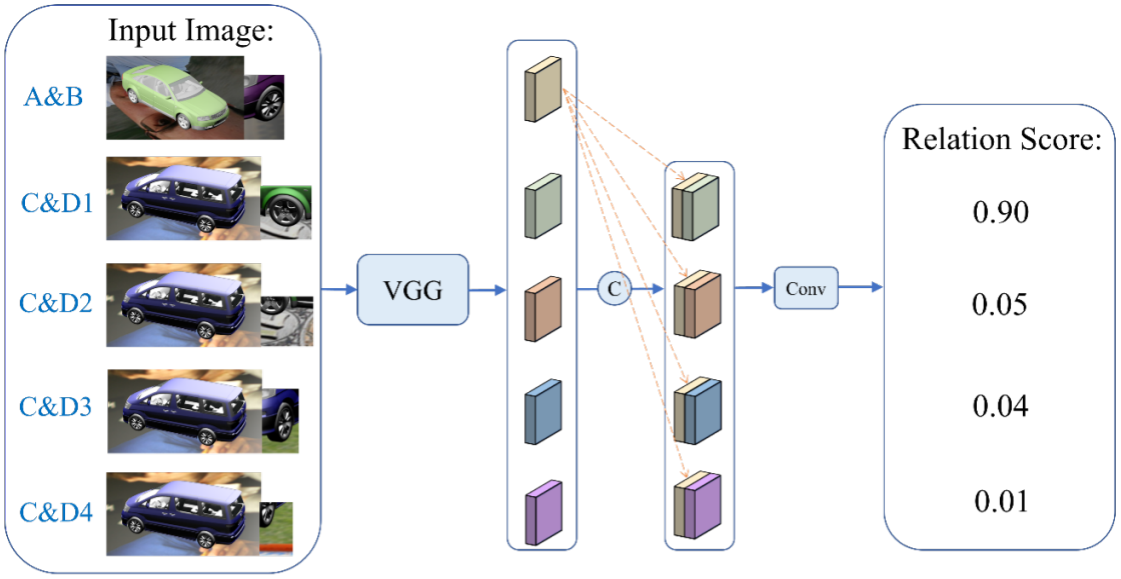
To train the deep learning models, 30,000 analogy problems were created using part labels to generate images of subregions. We divided the set of 30,000 analogy problems into 27,000 for training and 3,000 for test. For each analogy problem, we sampled two whole-car images from the following five different car types of the 3D CAD models (denoted as A, C): sedan, SUV, station wagon, truck, and minivan. Then we randomly selected part or piece subregions of the car in the A image to generate the B image. The correct option for the D image was the unique image that yielded the same whole-part relation for the C:D image pair as for the A:B pair. For each analogy question, the other three alternative D images were one other part/piece images from the car in the C image and two subregion images from the car in the A image. Whole-car images and car subregion images had random background, lighting, camera position, and object texture/color. None of the images used to train the networks were included in the visual analogy task used to compare model and human performance.

Siamese Network

We adapted a Siamese Network to learn to solve our car analogy problems. Figure 5.3A shows the model architecture for this network, which employs a VGG-16 network to translate pixel-level images to features. Features of whole-car images (A and C images) are processed by spatial pooling to form embeddings of size 4x4x512; subregion image features (B and D1-8 images) are pooled to form 1x1x512 embeddings. The feature embedding for image B is then expanded to the same size as that for image A, and then concatenated with A in the channel dimension separately. Feature embeddings for D images are similarly expanded



Siamese Network



Relation Network

Figure 5.4: Siamese Network (top panel) and Relation Network (bottom panel) architectures for answering car analogy problems. "C" in the figure indicates concatenation operation, "FC" indicates fully connected, and "Conv" indicates convolution operation.

and concatenated with the feature embedding for the C image. The concatenated features are passed through another convolutional layer and three fully-connected layers to obtain the final embedding of an image pair, which is a vector of size 512. We use contrastive loss with margin 1 to minimize the distance between concatenated feature embeddings from A:B images and embeddings from C:D images to choose the D image that best completes the analogy. More details about training all models are provided in Section 1 of Supplementary Materials.

Relation Network

We adopted the implementation by [138] to set up the Relation Network to learn to solve the car analogy problems; Figure 5.3B shows the model architecture for this implementation. The input to the Relation Network is pairs of images including a whole-car image and their corresponding subregions (e.g., A&B images, C&D1-8 images). To train the network to solve analogy problems, we padded the subregions images B and D images to the same image size as their corresponding whole-car images A and C, and then concatenated the whole-car image with each of the subregion images in the channel dimension, forming 6-channel input image pairs (i.e., the first three channels from color channels of a whole-car image and the second three channels from padded subregion images). The input image-pairs are then passed through the VGG network separately to get a fusion relation feature of A&B and C&D1-8 images. The extracted features of the A&B image pair are then concatenated with the features of the C&D image pairs in the channel dimension. The concatenated features are then passed through two convolutional layers and two fully connected layers to estimate a relation score indicating the probability that a candidate C&D pair instantiates the same relation as the question pair of A&B images. Cross entropy loss is used to train the Relation Network.

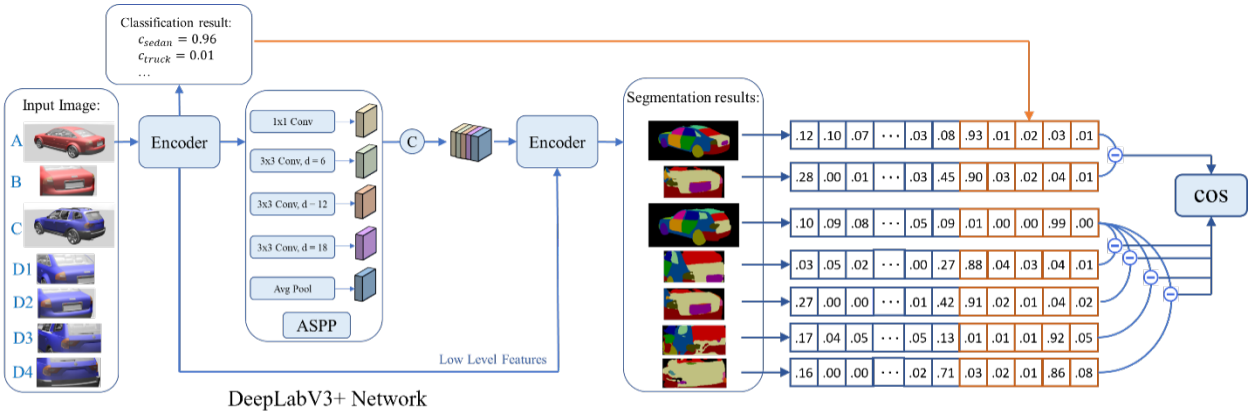


Figure 5.5: PCM model architecture, which consists of DeepLabv3+ Network Architecture for identification and segmentation of synthetic car images (left), and a comparison-based reasoning module (right).

Part-Based Comparison Model (PCM)

The part-based comparison model (PCM) is our instantiation of the domain- and task-general approach to visual analogy. However, the image representation components in PCM were trained to recognize cars and segment their parts. The training dataset was based on the same UDA-Part dataset used to generate images of car subregions in the four-term visual analogy problems. We rendered 40,000 synthetic images (30,000 for training and 10,000 for test; none of these were used in the analogy task) with automatically-generated part segmentation ground-truth. Each training image depicted one of the five car subtypes that were used to train the task-specific models: sedan, SUV, station wagon, truck, and minivan. We rescaled these images to .5 of their original size, which yielded the best performance on part classification and on the analogy problems in Experiment 1.

The image representation component in PCM uses a variant of the DeepLabv3+ architecture, a deep neural network that is widely used for semantic segmentation in computer vision [27]. The left side of Figure 5.5 shows this architecture, which includes encoder layers, an atrous spatial pyramid pooling (ASPP) module, and decoder layers. An input image first is processed through an encoder module composed of a ResNet101 [58], yielding a feature map with 2048 channels and size (height and width) down-sampled by 16. Features extracted

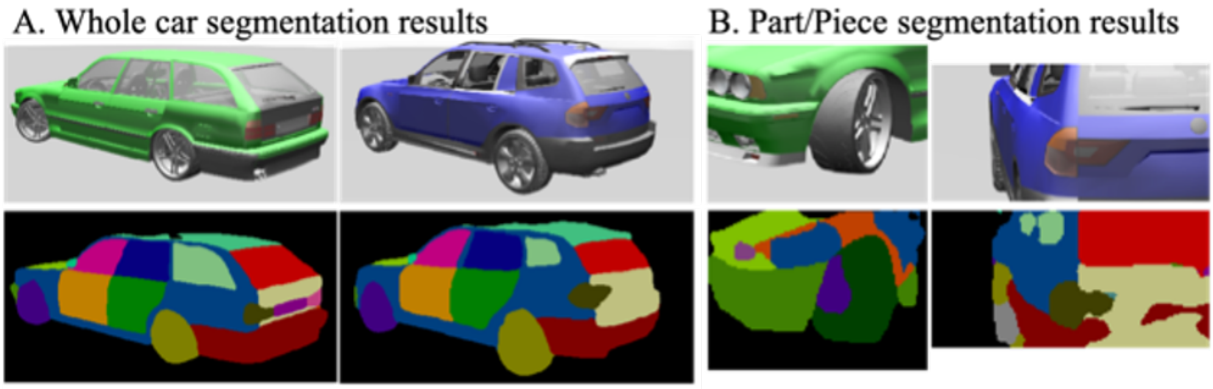


Figure 5.6: PCM segmentation results for images used in analogy problems. The model never saw these images during training.

from ResNet101 are commonly used for image segmentation. The ASPP module then samples the input features at multiple spatial rates to gather information at different spatial scales. The outputs from different spatial scales are concatenated in the channel dimension and passed through the decoder layers. The output is a mask with the same height and width as the input image. Each pixel of the mask is assigned a label to indicate whether it belongs to the background or each of the 31 parts labeled in the dataset. We added a second output branch after the encoder layers of DeepLabv3+ to predict the car type label, which formed a regular object classifier. Training was conducted using two standard cross-entropy losses: one for segmentation, and one for classification of car types.

Before feeding the images into the network, standard data augmentation was applied (e.g., translation, scaling, cropping). We controlled the augmentation parameters to ensure the network was only trained with whole car images (i.e., no partial car of subregion images were used during training). Car images used in the human experiment were excluded from the training set. The model achieved high performance for both part segmentation (mean intersect over union (mIoU) = 0.57) and subtype classification (accuracy = 0.99) on the test set. Figure 5.6 shows the segmentation results for the trained model.

After training, we applied the network to images used in the analogy problems to obtain segmentation and classification predictions. When applied to the untrained whole-car images

used in human experiment (images A and C), the segmentations were reasonable, and the classification accuracy was perfect (Figure 5.6A). When applied to the part/piece subregion images used in the experiment (i.e., image B and the alternative D images), which provide incomplete visual input, the segmentations yielded small errors, and the classification accuracy dropped to .71 (Figure 5.6B). These results suggest that the image representation component in PCM based on visual deep learning models of semantic segmentation and object recognition shows a substantial degree of generalization to untrained images for these visual tasks.

Next, we created compositional descriptions of car images based on extracted parts using the segmentation and classification results. We first converted the pixel-level labels in the resulting segmentation map to a low-dimensional feature vector. Specifically, we counted the number of pixels that were parsed into each of the part segments, and computed the proportion for each part (i.e., number of pixels in a part divided by the total number of pixels in the resulting segmentation map). The dimensions of feature vectors were defined using a taxonomy with 13 parts for cars, adapted from a more generalized segmentation scheme used for parsing complex, real-world scenes in the PASCAL-Part dataset [29]. The 13 car parts aggregate over subsets of the 31 segments in the UDA-Part dataset on which the part segmentation model was trained. These 13 parts are back part (including back bumper, tail lights and trunk segments), front part (including front bumper and hood), left frame, right frame, roof, door (including back left door, back bright door, front left door and front right door), window (including all window in the car), wheel (including all four wheels), back license plate, front license plate, left mirror, right mirror, and head light.

In addition to part-related features, we included the object information of car subtypes by concatenating the part-proportion vector with the car-type classification result, yielding the 18-dimensional feature vector $\boldsymbol{f} = \text{cat}(\mathbb{P}(\boldsymbol{m}), \boldsymbol{c})$, where $\boldsymbol{m}$ is the resulting segmentation map, $\boldsymbol{c}$ is the car-type prediction, $\sum f_{1:13} = 1$ and $\sum f_{14:18} = 1$. Thus, for each analogy question, PCM represents images $\boldsymbol{I}_A, \boldsymbol{I}_B, \boldsymbol{I}_C, \boldsymbol{I}_{D_1}, \boldsymbol{I}_{D_2}, \boldsymbol{I}_{D_3}, \boldsymbol{I}_{D_4}$ as feature vectors $\boldsymbol{f}_A, \boldsymbol{f}_B, \boldsymbol{f}_C, \boldsymbol{f}_{D_1}, \boldsymbol{f}_{D_2}, \boldsymbol{f}_{D_3}, \boldsymbol{f}_{D_4}$, respectively. Finally, to solve the visual analogy problem, a

decision is derived by selecting the best D image $\in \{D_1, D_2, D_3, D_4\}$ such that the relation that holds between image A and image B is similar to the relation between the two images C and D. We computed the difference between feature vectors $\boldsymbol{f}_A$ and $\boldsymbol{f}_B$, and between $\boldsymbol{f}_C$ and $\boldsymbol{f}_D$, and used cosine distance to measure the dissimilarity of the two difference vectors. The same approach has been used in the Word2vec model [172], and has proved effective in modeling visual analogical reasoning [103]. The preferred answer $\hat{D}$ is defined as the D image that generates minimum cosine distance between difference vectors:

$$\hat{D} = \underset{D \in \{D_1, D_2, D_3, D_4\}}{\operatorname{argmin}} \Big[1 - \cos\big(\boldsymbol{f}_B - \boldsymbol{f}_A,\ \boldsymbol{f}_D - \boldsymbol{f}_C\big)\Big].$$

5.2.4 Human Performance on Four-Term Visual Analogies

Participants achieved an overall mean proportion correct of .61 (SD = .21, range = [.09, .97], greatly exceeding the chance level of .25 correct. Figure 5.7 provides a breakdown of response selections for 8 experimental conditions. When participants did not select the correct answer, they more often selected one of the two distractors that included an element of the correct answer (correct subregion or else correct car body). They very seldom selected the option that was completely incorrect (wrong subregion of wrong car body). A three-way repeated measures ANOVA revealed that accuracy was higher on same-orientation problems ($M_{\text{acc}} = .65, SD_{\text{acc}} = .22$) than on different-orientation problems ($M_{\text{acc}} = .56, SD_{\text{acc}} = .23$), $F(1, 75) = 34.20, p < .001$, on part problems ($M_{\text{acc}} = .64, SD_{\text{acc}} = .23$) than on piece problems ($M_{\text{acc}} = .58, SD_{\text{acc}} = .22$), $F(1, 75) = 14.37, p < .001$, and on visible-subregion problems ($M_{\text{acc}} = .63, SD_{\text{acc}} = .23$) than on invisible- subregion problems ($M_{\text{acc}} = .59, SD_{\text{acc}} = .22$), $F(1, 75) = 7.95, p = .006$. No interaction effects were significant.

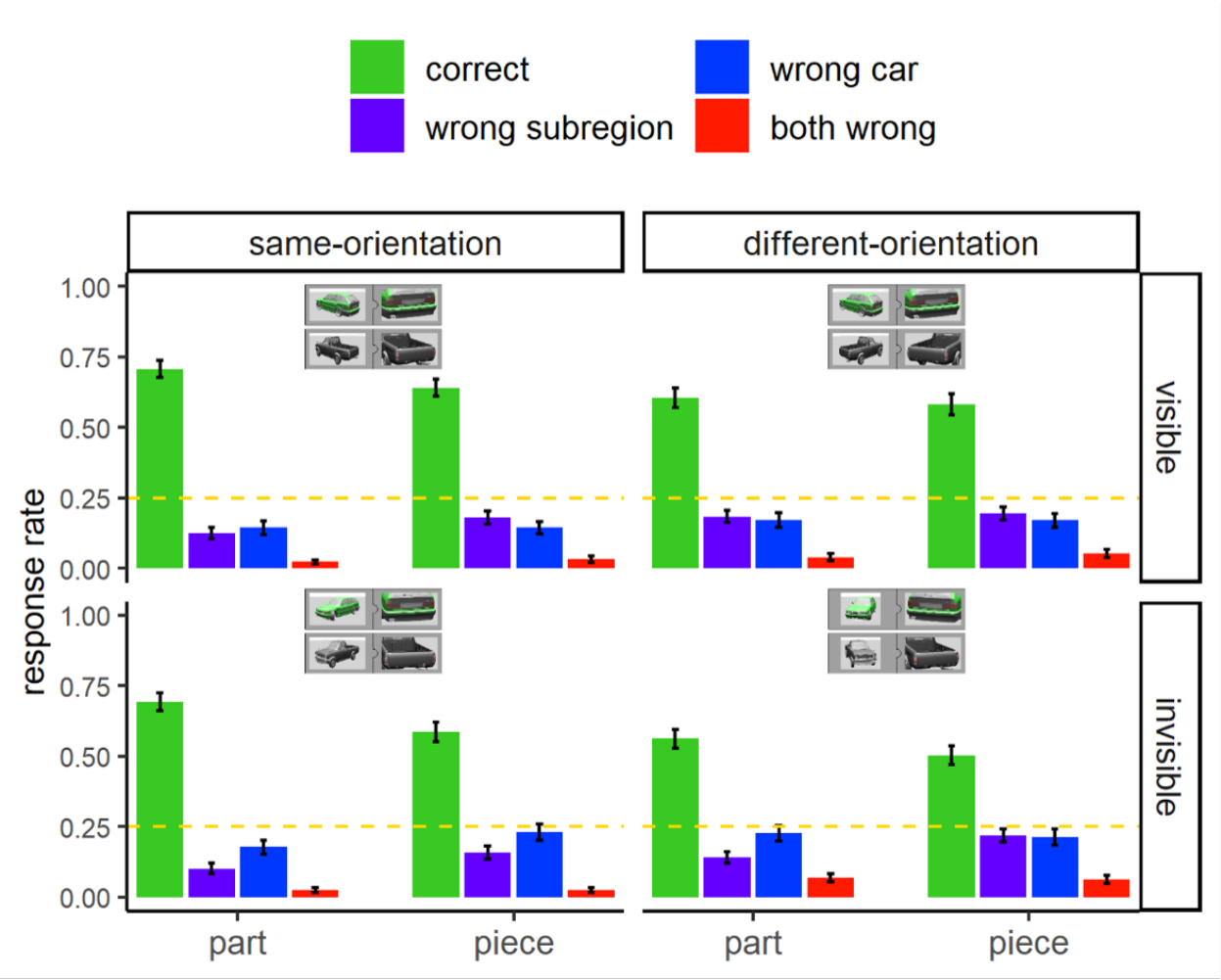


Figure 5.7: Human response selections for each problem type. Examples in each quadrant depict an experimental condition (same versus different orientation between the whole and subregion images for the left and right column panels; subregion images visible versus invisible in the corresponding whole images for the top and bottom row panels). Dashed line reflects chance performance. Error bars reflect ± 1 standard error of the mean.

5.2.5 Model Performance on Four-Term Visual Analogies

All three models (Siamese Network, Relation Network, PCM) were tested on the same 128 visual analogy problems used in the human experiment. These involve car images never used in training for any of the models. Overall proportion correct was .38 for the Siamese Network, .56 for the Relation Network, and .61 for PCM. Figure 5.8 plots model and human performance, broken down according to each experimental condition. The PCM model matched overall human accuracy most closely (.61). More importantly, only PCM captures

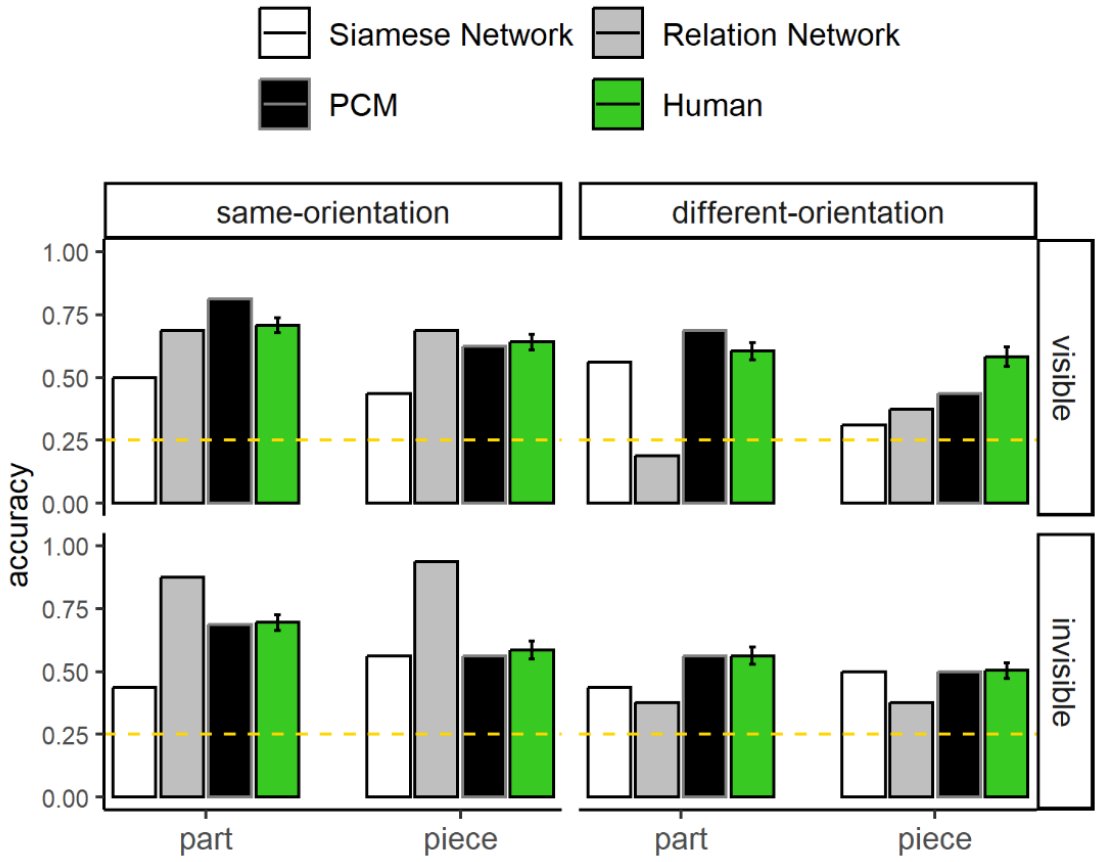


Figure 5.8: Model and human performance on the visual analogy task broken down by problem type. Dotted lines indicate chance performance. Error bars reflect ± 1 SEM for human data.

the qualitative differences in human accuracy across all conditions (Figure 5.9). PCM, like humans, shows higher accuracy on same-orientation problems than on different-orientation problems (Figure 5.9 top), on part problems than on piece problems (Figure 5.9 middle), and on visible-component problems than on invisible-component problems (Figure 5.9 bottom). While the two task-specific models also reproduced the effect of orientation, neither was able to capture all three qualitative effects, as PCM did. In order to quantitatively assess the fits of the models to human data, we computed the root mean squared deviation (RMSD) between model accuracy and mean human accuracy for each of the 8 types of analogy problems. This measure indicates how much each model deviates from human performance, with a lower RMSD indicating closer fit to human data. PCM yielded an RMSD of .07, achieving by far the closest fit to human data. RMSD was .17 for the Siamese Network and .23 for the

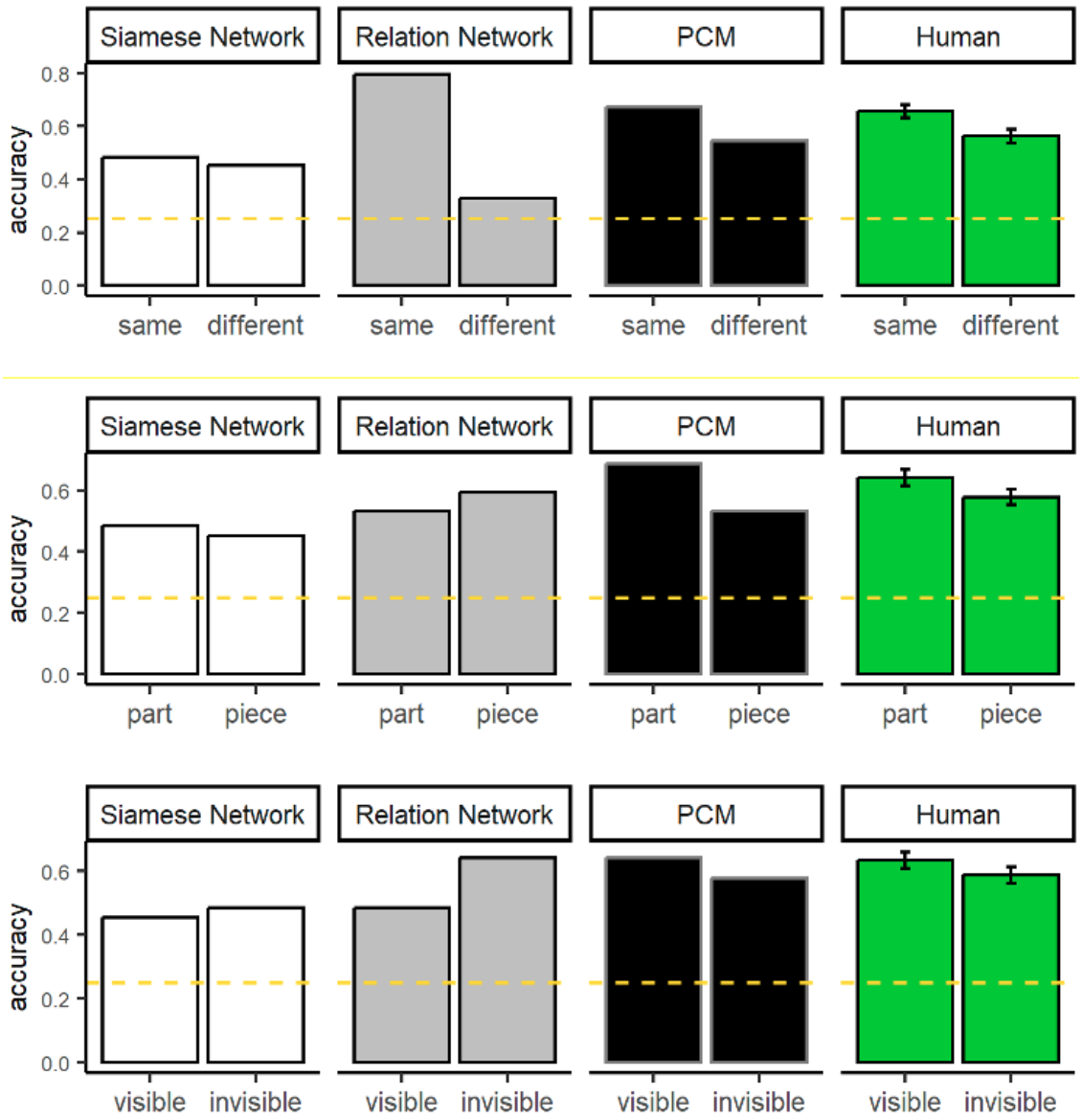


Figure 5.9: Model and human performance on the visual analogy task plotted separately for each of the following manipulations: Orientation (top), subregion type (middle), and visibility (bottom). Dotted lines indicate chance performance. Error bars reflect ± 1 SEM for human data.

Relation Network.

Recall that we trained task-specific models using 30,000 analogy problems, and we trained PCM's classifier using 30,000 whole-car images. We examined each model's performance on our analogy problems as a function of size of the training set. We varied the number of analogy problems used to train the Siamese Network and Relation Network: 5,000, 10,000, 15,000, and 30,000 problems. Figure 5.10 shows the two networks' training (top left) and

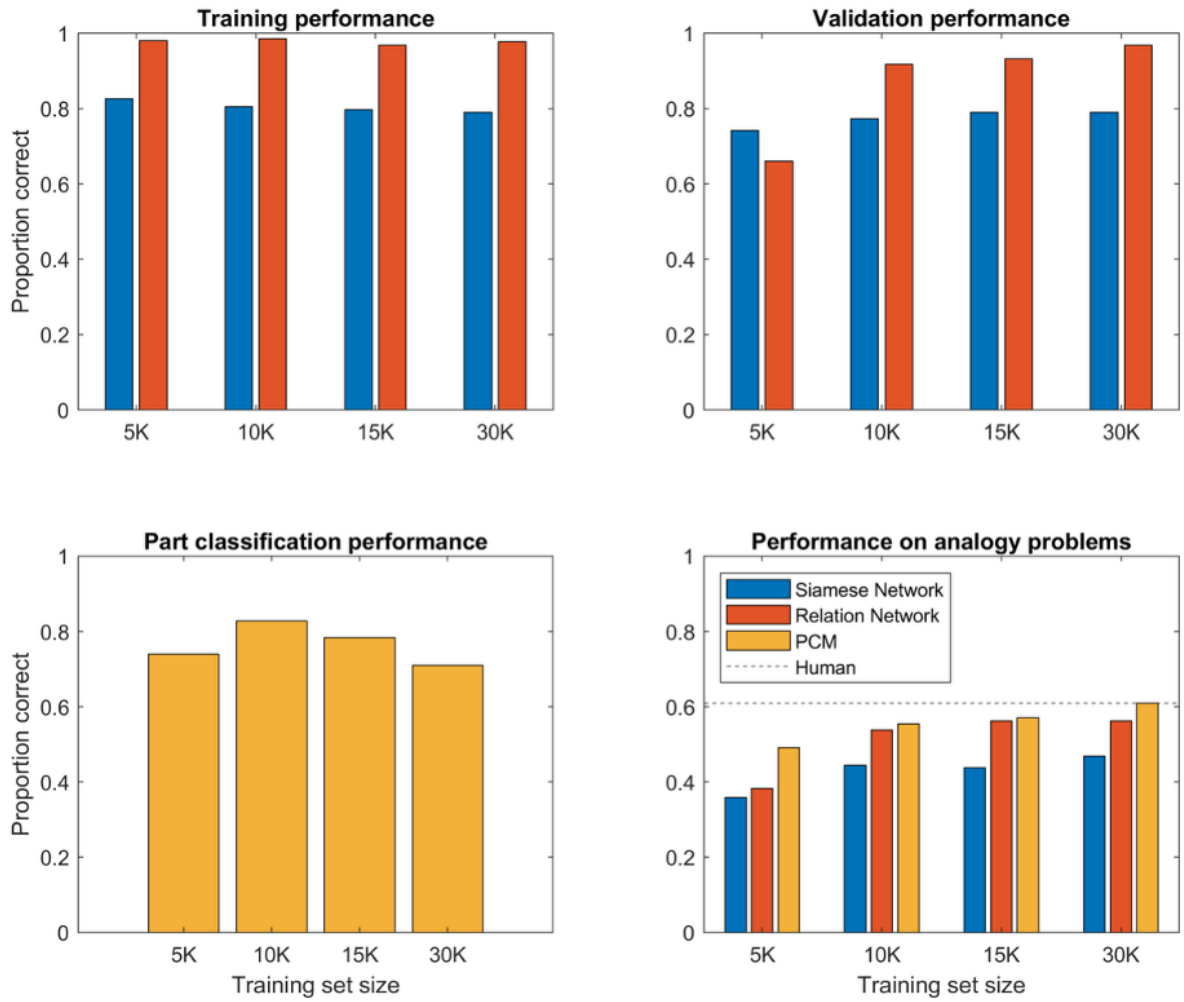


Figure 5.10: Siamese Network and Relation Network's performance on training problems (top left) and on held-out validation problems (top right), PCM's part classification performance (bottom left), and all models' overall performance on visual analogy problems (bottom right) all as a function of training set size. Human performance on analogy problems is represented by the dotted line in the bottom right panel.

validation performance (top right), as well as their performance on Experiment 1's four-term analogies (bottom right) as a function of their training set size. Across all three performance metrics, performance of both models plateaus at or before 10,000 training examples.

We also varied the number of images used to train PCM's classifier. PCM's part classification performance on the images used to construct our visual analogy problems showed an inverted U, indicating some overfitting on its training images as set size increased beyond 10,000 images (see Figure 5.10 bottom left panel). However, increasing the training set size

beyond 10,000 images appeared to aid analogy performance, as the model's accuracy on the visual analogy problems steadily increased as a function of its classifier's training set size (Figure 5.10 bottom right panel).

To further evaluate the models as accounts of human analogical reasoning, we considered a possible non-analogical shortcut strategy for answering the problems. Using the problem in Figure 5.2C as an example, some problems in the experiment could be answered correctly by selecting the D image most visually similar to the B image but not identical to it. We tested the possibility that any of the three models might have exploited this shortcut strategy, by providing only the subregion images (B image and D images) without showing the whole-car images (A and C images). The Siamese Network and PCM performed at chance on this test, indicating they did not exploit the shortcut strategy. However, Relation Network made the exact same responses when the "analogy" problem was reduced to just two of the four terms, implying that the model did not actually learn how to reason by analogy (i.e., by comparing a whole-part relation in source images to a similar relation in target images). Judging by its performance breakdown across conditions, Relation Network's clear advantage on same-orientation problems relative to different-orientation problems (Figure 5.9 top, grey bars) suggests that the model exploited the visual similarity between spatially-aligned car bodies and subregions to select the correct answer among the four response options. In Section 3 of the Supplementary Materials, we report the results of control simulations that replace Relation Network's pixel-level input with the low-dimensional output of PCM's classifier that served as the input to PCM's reasoning module. While providing Relation Network with this alternative input did not improve overall performance on the visual analogy problems in Experiment 1, it did prevent the model's acquisition of this non-analogical shortcut strategy.

The finding that the Relation Network was able to correctly answer roughly half of the four-term analogy problems using a demonstrably non-analogical strategy raises concerns about the extent to which this four-term analogy problem set assesses analogical reasoning. Accordingly, we performed Experiment 2 to compare model and human performance on a task

that placed stronger demands on analogical reasoning, rather than some possible shortcut strategies guided by purely visual features.

5.3 Experiment 2: Open-Ended Visual Analogies between Objects and their Parts

In Experiment 2, we asked participants to complete an open-ended visual analogy task, in which they were given an image of a car body with two colored markers on that car and were asked to mark the analogous spots on a different car body. We compared humans and models in their ability to identify and match analogous parts across car types in this open-ended analogy task. None of the models received any additional training beyond that described in connection with Experiment 1. Experiment 2 thus allowed us to compare task-specific and task-general approaches to analogical reasoning on out-of-task generalization. Moreover, examining each model's responses on this more open-ended task will reveal its response strategies, including the non-analogical strategies learned by the Relation Network.

5.3.1 Open-Ended Visual Analogy Task

In order to produce these open-ended analogy problems, we used the CGPart dataset [94] that contains the five car subtypes used to train models: sedans, SUVs, wagons, trucks, and minivans. From these 3D models, we generated 120 unique problems, each consisting of a source and target image showing different car bodies. A few examples of the stimuli are shown in Figure 5.8. The CGPart dataset provides detailed pixel-level annotations of vehicles and their component parts, which enabled us to place colored markers on source images so that they were centered at car parts of interest, either on wheels, bumpers, or roofs. Across problems, we manipulated the spatial alignment of the car bodies in the source and target images (the sole manipulation that in Experiment 1 consistently affected performance of both humans and all three models). The two car bodies in an image pair were shown either in

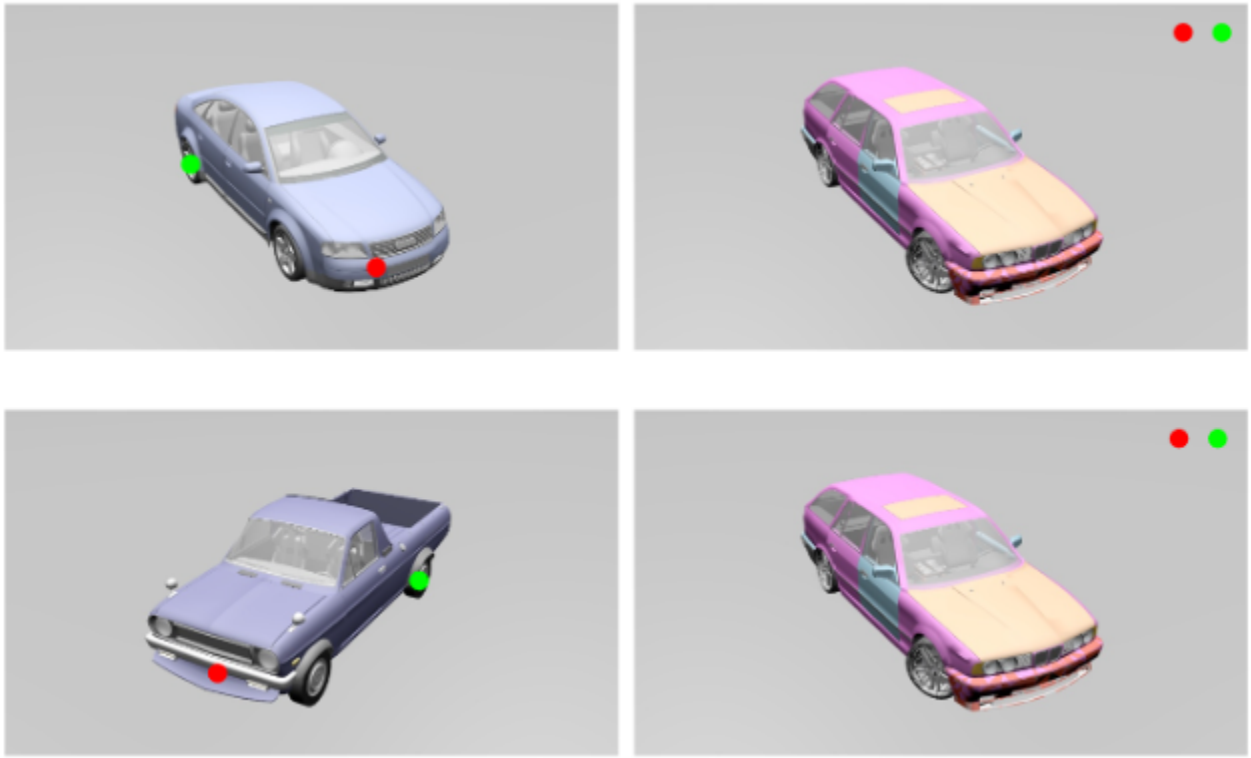


Figure 5.11: Two example trials (top and bottom rows) for the open-ended visual analogy task. Given a source image with two colored dots (left), participants were asked to place corresponding dots on the target image (right).

the same orientation or different orientations, 120 degrees apart. When car bodies were not spatially aligned, all marked parts (e.g., bumpers) remained visible in both images.

Each trial of this task consisted of a pair of images—a source image and a target image—each depicting a different car body. Two different-colored markers, one red and one yellow, were placed at the center of two parts on the car body in the source image, and participants were asked to use their mouse to drag a second pair of yellow and red markers to the analogous locations on the car body depicted in the target image. Specifically, this marker-placement task was to place the yellow marker in the target image so that it mapped to the yellow marker in the source image, and similarly for the red marker.

5.3.2 Human Participants and Experiment Methods

Forty-nine participants ($M_{\text{age}} = 19.53, SD_{\text{age}} = 1.43$, 40 female, 9 male) were recruited from the Psychology Department subject pool at UCLA. Participants provided online consent in accordance with the UCLA Institutional Review Board and were compensated with course credit.

On each trial, participants were presented with one image pair on a computer screen as shown in Figure 5.11, and completed a marker-placement task. For each of two colored markers, they were asked to "move the marker on the top right corner in the target image to the corresponding location that maps to the same-color marker in the source image." If the participant did not think there was an analogy between the two images, they were allowed to move the markers back to the top right corner and leave them there. No time constraint was imposed. On each trial, the exact location of each marker placement was recorded.

Data from 7 of the 49 participants were excluded from analyses either because they indicated they were not serious during the experiment or because they attempted to move the markers fewer than half of the trials. Thus, data from the remaining 42 participants were included in analyses.

5.3.3 Model Simulations

In order to test models on open-ended mapping problems, we adopted a simple modification to the simulations for the four-term analogy problems in Experiment 1. Because our models could only pinpoint image patches with the same relation to the whole image, we adopted a sliding window method to find the marker locations for the target image that were most analogous to those provided in the source image. From the source image, we used the image to be the whole-car image (image A), and we cropped out an image patch centered around the marker location to generate the subregion image (image B). We used the target image as the whole-car image for image C, and then slid the patch window on the target image with stride of 5 pixels. To generate subregion image D, we started by cropping an image patch

from the top-left corner of the target image, and then moved 5 pixels rightward to crop the next image patch from the target image. When we reached the rightmost position of the image, we moved the window back to the leftmost position and 5 pixels downward from its previous position, and then repeated the last operation.

This sliding window procedure enabled us to sample hundreds of image patches from the target image to generate the subregion images (image D). We then used each model to identify which patch subregion in the target image was most analogous to the patch subregion centered at the marker location of the source image. Thus the main difference between the testing procedure in Experiment 2 and that used in Experiment 1 is the number of answer options. Whereas the four-term analogy problems in Experiment 1 each had four answer options of subregion images, the sliding window procedure for the more open-ended analogy problems tested in Experiment 2 yielded hundreds of answer options from patches sampled from the target image. We did not test model predictions on problems where more than 30% of participants failed to find an analogy between the source and target, and therefore opted out of responding. This excluded 20 out of 120 problems in the following analyses.

5.3.4 Human and Model Performance on Open-Ended Visual Analogy Task

For each trial, we calculated the mean location of colored marker placements, averaged across participants. We then computed the spatial distance (in pixels) from the marker location provided by each individual participant to the overall mean location. This measure of individual distance to the mean marker location provided a quantitative assessment of human variability in mapping judgments, with smaller distance values indicating higher consistency across participants for their marked locations in the same analogy problem. For the same-orientation image pairs, the average distance to the mean marker location was around 5.7 pixels, whereas different view pairs had an average distance of around 14.9 pixels. Note that relative to the object sizes (average height of 202 pixels and width of

317 pixels), even 15-pixel derivation represents a small spatial distance, indicating strong agreement of people's judgments in analogical mapping. A repeated-measures ANOVA of trial-level deviation from the mean marker placements with one within-subjects factor (same- vs. different-orientation for source and target images) yielded a reliable main effect of orientation, $F(1, 243) = 54.5, p < .001$.

When comparing model predictions to human responses, we found that the responses of different model were variably affected by the size of each patch window. Table 1 shows the pixel-wise deviation between model predictions and mean human marker placements for different window sizes on same-orientation and different-orientation problems. Whereas our task-general model PCM was robust to changes in window size, both task-specific models were highly sensitive to such changes. Regardless, PCM's predictions were substantially closer to human marker placements across all window sizes, achieving the best performance using window size of 100.

Table 5.1: The mean distance of model predictions to the human marker center measured in pixels. Smaller values indicate more accurate model predictions for human marker locations.

Window Size	PCM		Relation Network		Siamese Network	
	same	different	same	different	same	different
100	17.7	28.0	60.7	84.8	93.0	93.7
110	17.0	30.4	56.1	79.4	98.3	107.5
120	18.0	30.6	45.6	63.2	77.3	91.4
130	19.2	33.1	42.5	65.5	83.2	85.1
140	19.5	38.5	36.6	73.0	79.0	72.6
150	18.1	43.0	40.3	72.0	66.4	76.4

We were also able to much more precisely characterize the short-cut strategy learned by Relation Network when it was trained on the type of four-term analogies used in Experiment 1. We depicted the scores of patches at each location on the target image, which are plotted as a heatmap in Figure 5.12. These scores were generated using a window size of 140, which showed the best fit of the Relation Network to human performance. Scores reflect performance on same-orientation problems, but the model's prediction for different-orientation problems

was similar. When the source image, the source patch subregion centered at a marker location, and the target image were passed in as inputs, the Relation Network generated a reasonable score distribution over the car in the target image. As shown in the top row of Figure 5.12, the network focuses on the bumper when the source marker patch is a bumper (left) and focuses on the wheels when the source marker patch is a wheel (right). However, when given the source marker patch without the whole source and target images, the Relation Network still produced the same score distribution (second row in Figure 5.12). When given the whole-car source and target images and a blank image in place of the source marker patch, the network focused on the background, which was most visually similar to the blank image (third row in Figure 5.12). Thus, despite being trained to solve four-term visual analogy problems, Relation Network ultimately learned to match subregion images based on visual similarity. In contrast, both the Siamese Network and PCM made use of the whole-car images in their predictions, resulting in very different response distributions when presented with and without whole-car images.

5.4 Discussion

We compared two computational approaches to visual analogical reasoning. One approach applies task-specific knowledge to reasoning problems, acquired from extensive exposure to highly similar problems. The other approach applies a domain-general comparison procedure to representations of entities and their component structure. The fundamental difference between these two approaches is that the former approach completely merges formation of perceptual representations and reasoning with them, whereas the latter first learns perceptual representations, to which a domain-general comparison process is then applied.

We introduced the Part-Comparison Model (PCM) to instantiate the latter approach. We trained this model to segment object parts and identify 3D objects from images in order to ultimately generate representations of object structure in terms of part-whole

relations. In Experiment 1, by applying a domain-general comparison procedure to part-based difference vectors, PCM could account for qualitative patterns of human accuracy on four-term visual analogies between images of cars and their subregions. Specifically, PCM's analogy performance, like that of humans, was sensitive to the spatial alignment between analogs, the integrity of analogous subregions as natural parts, and the visibility of analogous subregions in whole-car images.

In Experiment 2, PCM closely approximated human responses on an open-ended analogy task in which participants were asked to freely map parts across car bodies. Since PCM was not trained to solve any analogy problems at all, its performance on the four-term (Experiment 1) and the open-ended task (Experiment 2) both constitute out-of-task testing. In contrast, two popular deep learning models that have been applied to visual analogies—a Siamese Network and a Relation Network—deviated seriously from human performance when confronted with both out-of-sample (but within-task) testing in Experiment 1, and with out-of-task testing in Experiment 2. Indeed, after extensive direct training on four-term analogy problems, the Relation Network did not learn to reason by analogy at all. Instead, the Relation Network acquired a non-analogical shortcut strategy based on visual similarity. An important methodological implication of these findings is that simply matching (or even exceeding) human overall accuracy on some benchmark task is not a sufficient criterion for inferring that a computational model is simulating human cognitive mechanisms. It is necessary to compare model and human performance in greater detail at the level of controlled manipulations of problem types.

PCM relies on visual processes to learn representations of object parts, offering a parsimonious account of representation learning. Previous approaches to learning similar types of representations have used more computationally intensive (though more efficient) learning processes, such as analogical structure mapping [26]. Unlike such models, PCM does not presuppose that analogical reasoning is already available for use in acquiring representations of entity structure. Rather, we assume analogical reasoning operates on relational representa-

tions that can be created by more basic learning mechanisms that operate on nonrelational inputs [104]. The representations that PCM acquires are expressive enough to provide the basis for a successful approach to visual analogy, based simply on comparisons between relations derived from the model's learned representations.

The failures of the task-specific models in accounting for human performance in the present experiments illustrate a general limitation of treating analogy as the application of task-specific knowledge acquired by simultaneously learning representations of task constituents and also decision procedures for responding on that same task. Such an approach is unlikely to achieve out-of-task generalization. Instead of learning perceptual representations that might be generally useful in multiple tasks, these models acquire representations tailored to idiosyncrasies of the specific task used in training. A possible way to augment this task-based approach is to incorporate meta-learning, in which a model is trained to solve (and explicitly represent) multiple distinct tasks, and then transfers its acquired knowledge to solve novel tasks, based on their similarity to prior learned tasks [85]. However, it remains to be seen whether meta-learning across a range of analogy tasks can allow a task-based approach to account for the breadth of human analogical reasoning.

The more promising approach, illustrated by PCM, is to first learn componential structure (here, part-whole relations for 3D objects) for visual representations that are generally useful in distinctively different tasks (e.g., object recognition and segmentation) [11]. By training with varied visual tasks, the learned representations acquire multi-task consistency. This approach is broadly consistent with other deep learning models that emphasize the acquisition of representations for use in a range of downstream visual reasoning tasks [44, 153, 155].

In the verbal domain, early word-embedding models such as Word2vec acquired lexical representations after having been trained to predict text in sequence within large corpora; these representations were then used to solve simple four-term verbal analogies [110] . More recent Large Language Models (LLMs) have been particularly effective in exploiting this approach [17], especially when training on text prediction is combined with subsequent human

feedback reinforcement learning (HFRL). These methods yield representations of verbal inputs that support a wide range of capacities, including complex analogical reasoning [150]. However, current LLMs are inapplicable to visual analogy, and the problem of building representations is arguably more difficult in the case of vision. Language modeling takes discretized text tokens as inputs, whereas visual perception lacks such well-defined units.

Although PCM requires extensive training to learn its compositional representation for visual inputs, it requires no training at all to solve analogies. Rather, PCM achieves superior performance on our visual analogy tasks simply by comparing its representations of entity structure, using a domain-general similarity measure. Humans can (and machines perhaps may) achieve analogical reasoning by learning representations that encode relational structure, coupled with efficient computation of relational similarity.

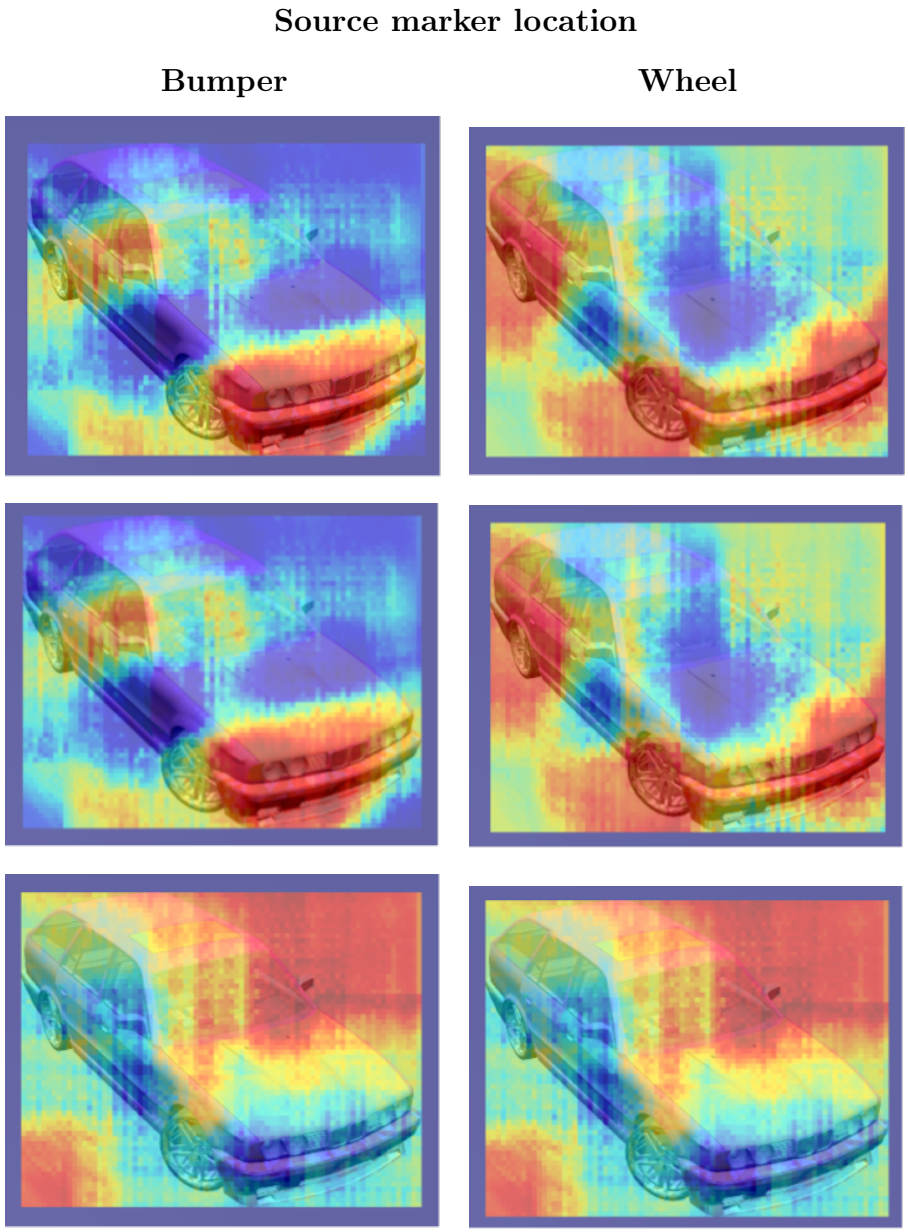


Figure 5.12: Prediction scores for Relation Network when the moving patch is centered on each location under various conditions. Left: the source patch is centered on a bump; right: the source patch is centered on a wheel. Top row: all whole-car images and the source marker patch are passed to the network; middle row: only the source marker patch is passed to the network; bottom row: the whole-car source and the target images, along with a blank image for source patch images, are passed to the network. Redness in the heatmap means the network yields a high score whereas blue indicates the network gives a low score.

Chapter 6

VisiPAM

6.1 Introduction

If a tree had a knee, where would it be? Human reasoners (without any special training) can provide sensible answers to such questions as early as age four, revealing deep understanding of spatial relations across different object categories [46]. How such abstraction is possible has been the focus of decades of research in cognitive science, leading to several proposed theories of analogical reasoning [37, 47, 50, 61, 62, 65]. The basic elements of these theories include structured representations involving bindings between entities and relations, together with some mechanism for mapping the elements in one situation (the *source*) onto the elements in another (the *target*), based on the similarity of those elements. However, while this work has succeeded in accounting for a range of empirical signatures of human analogy-making, a significant limitation is that the structured representations at the heart of these theories have typically been hand-designed by the modeler, without providing an account of how such representations might be derived from real-world perceptual inputs [23]. This concern is particularly notable in the case of visual analogies. Unlike linguistic analogies in which relations are often given as part of the input (via verbs and relational phrases), visual analogies

The contents of this chapter appeared in paper [42, 154].

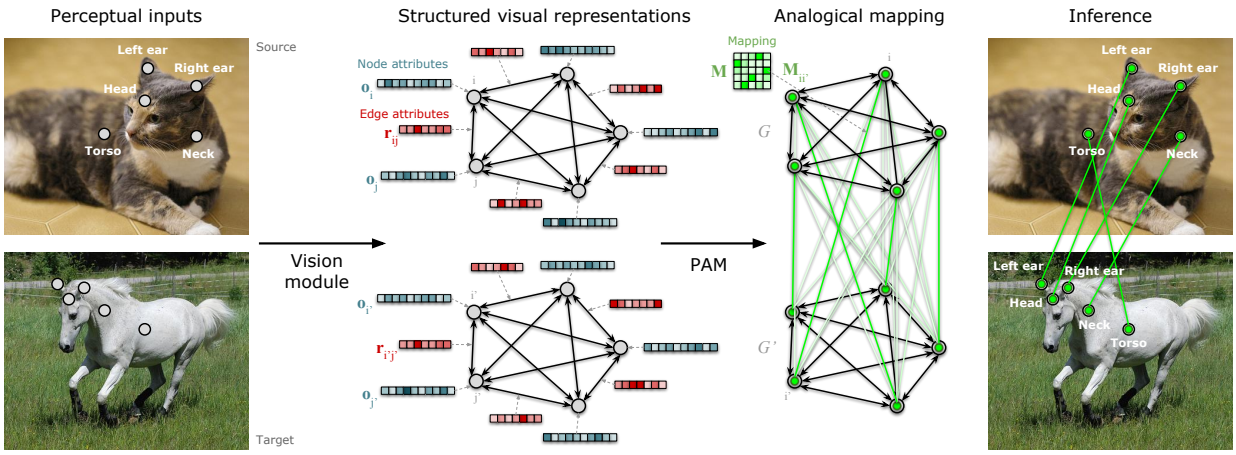


Figure 6.1: **Overview of visiPAM.** VisiPAM contains two core components: a vision module and a reasoning module. The vision module receives visual inputs in the form of either 2D images, or point-cloud representations of 3D objects, and uses deep learning components to extract structured visual representations. These representations take the form of attributed graphs, in which both nodes $\mathbf{o}_{1..N}$ (corresponding to object parts) and edges $\mathbf{r}_{1..N(N-1)}$ (corresponding to spatial relations between parts) are assigned attributes. The reasoning module then uses Probabilistic Analogical Mapping (PAM) to identify a mapping $\mathbf{M}$ from the nodes of the source graph G to the nodes of the target graph G', based on the similarity of mapped nodes and edges. Mappings are probabilistic, but subject to a soft isomorphism constraint (preference for one-to-one correspondences).

more obviously depend on a mechanism for the *eduction of relations* [135]: the extraction of relations from non-relational inputs, such as pixels in images.

More recently, a separate line of research inspired by the success of deep learning in computer vision and other areas has aimed to develop neural network algorithms with the ability to solve analogy problems. This work tackles the challenge of solving analogy problems directly from pixel-level inputs; however, it has typically done so by eschewing the structured representations and reasoning operations that characterize cognitive models, focusing instead on end-to-end training directly on visual analogy tasks. These approaches have generally relied on datasets with massive numbers (sometimes more than a million) of analogy problems, and typically yield algorithms that cannot readily generalize to problems with novel content [8, 112, 165]. More recent efforts have attempted to address this limitation, proposing algorithms that can learn from fewer examples or display some degree of out-of-distribution generalization [3, 60, 76, 153, 155]. But this work still maintains the basic paradigm

of treating analogy problems as a task on which a reasoner will receive at least some direct training. This is in stark contrast to human reasoners, who can often solve analogy problems *zero-shot* – that is, with no direct training on those problem types. When visual analogy problem sets are administered to a person, 'training' is typically limited to general task instructions with at most one practice problem. This format is critical to the use of analogy problems to measure *fluid* intelligence [21, 134], the ability to reason about novel problems without prior training or exposure.

Here we propose a synthesis that combines the strengths of these two approaches. Our proposed model, visiPAM (*visual Probabilistic Analogical Mapping*), combines learned representations derived directly from pixel-level or 3D point-cloud inputs, together with a reasoning mechanism inspired by cognitive models of analogy. VisiPAM addresses two key challenges that arise in developing such a synthesis. The first issue is how structured representations might be extracted from unstructured perceptual inputs. VisiPAM employs attributed graph representations that explicitly keep track of entities, relations, and their bindings. We propose two methods for deriving such representations, one for 2D images and one for point-cloud inputs representing 3D objects. Second, representations inferred from real perceptual inputs are necessarily noisy and high-dimensional, posing a challenge for any reasoning mechanism that must operate over them. To address this problem, visiPAM employs a recently proposed Probabilistic Analogical Mapping (PAM) [102] method that can efficiently identify patterns of similarity governing noisy, real-world inputs. We evaluated visiPAM on a part-matching task with naturalistic images, where it outperformed a state-of-the-art deep learning model by a large margin (30% relative reduction in error rate) – despite the fact that, unlike the deep learning model, visiPAM received no direct training on the part-matching task. We also performed a human experiment involving visual analogies between 3D objects from disparate categories (e.g., an analogy between an animal and a man-made object), where we found that visiPAM closely matched the pattern of human behavior across conditions. Together, these results provide a proof-of-principle for an approach that weds the representational power

of deep learning with the similarity-based reasoning operations that characterize human cognition.

6.2 Computational framework

Figure 6.1 shows an overview of our proposed approach. VisiPAM consists of two core components: 1) a vision module that extracts structured visual representations from perceptual inputs corresponding to a source and a target, and 2) a reasoning module that performs probabilistic mapping over those representations. The resulting mappings can then be used to transfer knowledge from the source to the target, to infer the part labels in the target image based on the part labels in the source.

The inputs to the vision module consisted of either 2D images, or point-cloud representations of 3D objects. The vision module uses deep learning components to extract representations in the form of *attributed graphs*. These graph representations consist of a set of nodes and directed edges, each of which is associated with a set of attributes (i.e., a vector). In the present work, nodes corresponded to object parts, and edges corresponded to spatial relations, though other arrangements are possible (e.g., nodes could correspond to entire objects in a multi-object scene, and edges could encode other visual or semantic relations in addition to spatial relations). For 2D image inputs, we used iBOT [173] to extract node attributes that captured the visual appearance of each object part. iBOT is a state-of-the-art, self-supervised vision transformer that was pretrained on a masked image modeling task. Masked image modeling shows promise as a suitable pretraining objective for capturing general visual appearance, rather than only information relevant to a particular task, such as object classification. For 3D point-cloud inputs, we used a Dynamic Graph Convolutional Neural Network (DGCNN) [149] trained on a part segmentation task in which each point was labelled according to the object part to which it belonged. We hypothesized that this objective would encourage the intermediate layers of the DGCNN (which were used

to generate node embeddings) to represent the local spatial structure of object parts. We evaluated visiPAM on analogies involving 3D objects from either the same superordinate category (e.g., both the source and target were man-made objects) or different superordinate categories (e.g., the source was a man-made object, and the target was an animal). Note that the DGCNN was only trained on part segmentation for man-made objects, not for animals. Edge attributes encoded either the 2D or 3D spatial relations between object parts. We denote the node attributes of each graph as $\mathbf{o}_{i=1..N}$, and the edge attributes as $\mathbf{r}_{1..N(N-1)}$, where N is the number of nodes in the graph. We denote the source graph as G and the target graph as G'.

After these representations were extracted by the vision module, we used PAM to identify correspondences between analogous parts in the source and target, based on the pattern of similarity across both parts and their relations. Formally, this corresponds to a graph-matching problem, in which the objective is to identify the mapping $\mathbf{M}$ that maximizes the similarity of mapped nodes and edges. PAM adopts Bayesian inference to compute the optimal mapping:

$$p(\mathbf{M}|G, G') \propto p(G, G'|\mathbf{M})p(\mathbf{M}) \tag{6.1}$$

where $\mathbf{M}$ is an $N \times N$ matrix in which each entry $\mathbf{M}_{ii'}$ corresponds to the strength of the mapping between node i in the source and node i' in the target. The posterior in Equation 6.1 is based on the following log-likelihood:

$$\log(p(G, G'|\mathbf{M})) = (1-\alpha)\frac{\sum_i \sum_{j\neq i} \sum_{i'} \sum_{j'\neq i'} \mathbf{M}_{ii'}\mathbf{M}_{jj'} \operatorname{sim}(\mathbf{r}_{ij}, \mathbf{r}_{i'j'})}{N(N-1)} + \alpha\frac{\sum_i \sum_{i'} \mathbf{M}_{ii'} \operatorname{sim}(\mathbf{o}_i, \mathbf{o}_{i'})}{N} \tag{6.2}$$

in which $\operatorname{sim}(\mathbf{r}_{ij}, \mathbf{r}_{i'j'})$ is the cosine similarity between edge $\mathbf{r}_{ij}$ (the edge between nodes i and j in the source) and edge $\mathbf{r}_{i'j'}$ (the edge between nodes i' and j' in the target), and $\operatorname{sim}(\mathbf{o}_i, \mathbf{o}_{i'})$ is the cosine similarity between node $\mathbf{o}_i$ in the source and node $\mathbf{o}_{i'}$ in the target.

Intuitively, the optimal mapping is one that assigns high strength to similar nodes, and to nodes attached by similar edges. The relative influence of node vs. edge similarity is controlled by the parameter α. In addition, PAM incorporates a prior $p(\mathbf{M})$ that favors isomorphic (one-to-one) mappings. In other words, a mapping function with one-to-one correspondences between the source and the target is assigned a higher prior probability. At the algorithm level, given that an exhaustive search over possible mappings was not feasible (we address problems with up to 10 object parts, corresponding to $10! \approx 3.6$ million possible one-to-one mappings), we used a graduated assignment algorithm [49]. In this approach, $\mathbf{M}$ is initialized as a uniform mapping, in which each source node is equally likely to map to any target node, and this mapping is then iteratively updated based on the likelihood in Equation 6.2. This iterative procedure is also guided by the prior $p(\mathbf{M})$ to encourage convergence to an isomorphic (one-to-one) mapping. It should be noted that this approach is fully differentiable, and therefore the entire model, including the vision module, could in principle be learned end-to-end based on the mapping task. However, since our goal in the present work was to account for the human capacity for zero-shot analogical reasoning, we instead use visual representations provided from pretrained deep learning models, such that no direct training on the mapping task is required.

6.3 Analogical mapping with 2D images

We evaluated visiPAM on a part-matching task developed by Choi et al. [31]. In this task two images are presented, each together with a set of coordinates corresponding to the locations of various object parts (as in Figure 6.1). The task is to label the object parts in the target image, based on a comparison with the labelled object parts in the source image. Importantly, the test set for this dataset consists only of object categories not present in the training set, so that it is not possible to solve the task by learning to classify the object parts in the target image directly. Choi et al. [31] also developed the Structured Set Matching Network

(SSMN), which they found outperformed a range of other methods after training directly on the part-matching task with a set of separate object categories. Here we go beyond this benchmark by removing any direct training on the part-matching task, relying instead on visiPAM's similarity-based reasoning mechanism to align the object parts between source and target.

	Within-category		Between-category
	Animals	Vehicles	Animals
VisiPAM	**63.2% (59.1%)**	**69.5% (58.8%)**	**67.9% (59.9%)**
VisiPAM (nodes only)	55.5% (50.6%)	61.6% (48.1%)	62.3% (52.9%)
VisiPAM (edges only)	47.8% (42%)	63.7% (50.9%)	51.5% (39.4%)
SSMN	46.6% (40.7%)		
Random	10%	26%	20%

Table 6.1: **Analogical mapping with 2D images.** Mapping accuracy on part-matching task proposed by Choi et al. [31]. Original task in [31] involved evaluation on within-category animal comparisons after training with 37,330 mapping problems. VisiPAM significantly outperformed SSMN, the previous state-of-the-art, despite having no direct training on mapping. VisiPAM also performed well on new problems involving within-category vehicle comparisons, and between-category animal comparisons (e.g., mapping from cat to horse). VisiPAM performed best when mapping was based on both node and edge similarity. 'Random' denotes chance performance (determined by average number of part comparisons). Values in parentheses reflect chance-normalized performance (percentage of the range between chance performance and 100% accuracy). VisiPAM's performance is roughly comparable across all conditions once chance performance is taken into account.

Table 6.1 shows the results of this evaluation. Accuracy was computed based on the proportion of parts that were mapped correctly across all problems (i.e., the model could receive partial credit for mapping some but not all of the parts correctly in a given image pair). As originally proposed, the part-matching task involved within-category comparisons of animals (comparing either two images of cats, or two images of horses). In that setting, visiPAM significantly outperformed SSMN, resulting in a 30% reduction relative to SSMN's error rate, despite the fact that SSMN, but not visiPAM, received direct training (37,330 problems) on this task. We also developed extensions of this task involving other object categories. These included within-category mapping of vehicles (involving either two images of planes, or two images of cars), and between-category mappings of animals (mapping from

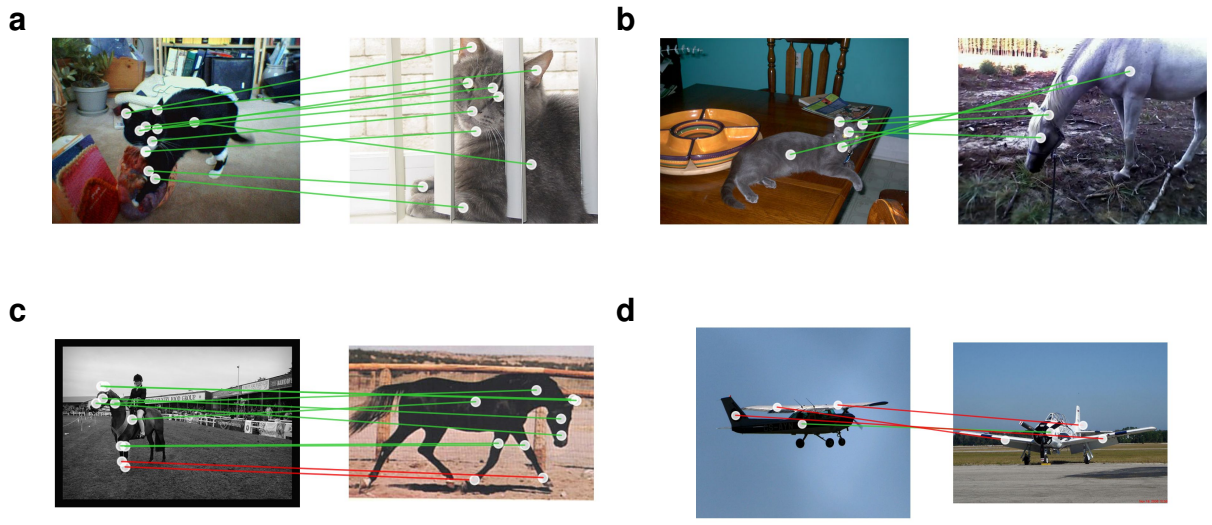


Figure 6.2: **VisiPAM part mappings with 2D images.** Examples of part mappings identified by visiPAM for 2D image pairs. **(a)** Successful within-category animal mapping involving 10 separate parts. Note the significant variation in visual appearance between source and target. **(b)** Successful between-category animal mapping. **(c)** Within-category animal mapping illustrating a common error in which the left and right feet are mismapped. **(d)** Within-category vehicle mapping. Mapping between images of planes was especially difficult, likely due to significant variation in 3D pose.

an image of a cat to an image of a horse). We found that visiPAM performed comparably well across all three of these conditions (within- and between-category animal mapping, within-category vehicle mapping).

A key element of our proposed model is that it performs mapping based on the similarity of both object parts *and* their relations. To determine the importance of this design decision, we performed ablations on either node or edge similarity components, by setting α to either 0 or 1. We found that both of these ablations significantly impaired the performance of visiPAM. This pattern aligns with findings from studies of human analogy-making, which show that human reasoners are typically sensitive to similarity of both entities and relations [80].

We also performed ablations that targeted different aspects of visiPAM's edge embeddings. Specifically, visiPAM's edge embeddings contain information based on both angular distance and relative location (vector difference). We found that both of these sources of information contributed to visiPAM's performance (Supplementary Table B.1). Furthermore, we found

that augmenting visiPAM's edge embeddings with topological information (i.e., whether two parts are connected) can lead to further improvements in mapping performance (Supplementary Table B.2). Thus, the specific types of spatial relations employed by the model play an important role in its performance, and the addition of new sources of relational information will likely lead to continued improvement.

Figure 6.2 shows some examples of the mappings produced by visiPAM. These include some impressive successes, including mapping of a large number of object parts across dramatic differences in background, lighting, pose, and visual appearance (Figure 6.2a), as well as between objects of different categories (Figure 6.2b). Notably, we also found that visiPAM's performance was unaffected when the target image was horizontally reflected relative to the source (Section 6.3.1 and Table 6.2), suggesting that the model is robust to low-level image manipulations.

Our analyses also illustrated some of the model's limitations. Figure 6.2c shows an example of an error pattern that we commonly observed, in which visiPAM confused corresponding left and right parts (in this case the left and right feet in a comparison of two horses). A more comprehensive analysis is shown in Figures 6.3, where it can be seen that this kind of confusion of corresponding lateralized parts accounts for a large percentage of visiPAM's errors. We also found that visiPAM particularly struggled with mapping images of planes (achieving the model's lowest within-category mapping accuracy of 42.5%), as shown in Figure 6.2d. This is likely due to the limitation of the 2D spatial relations used in this version of the model, which are particularly ill-suited to objects such as planes that can appear in a wide variety of poses and viewpoints. Thus, while visiPAM shows an impressive ability to perform mappings between complex, real-world images using only these 2D spatial relations, we suspect that accurate 3D spatial knowledge is likely necessary to solve some challenging visual analogy problems at a human level. To address this concern, we next sought to evaluate the performance of visiPAM in the context of 3D visual inputs.

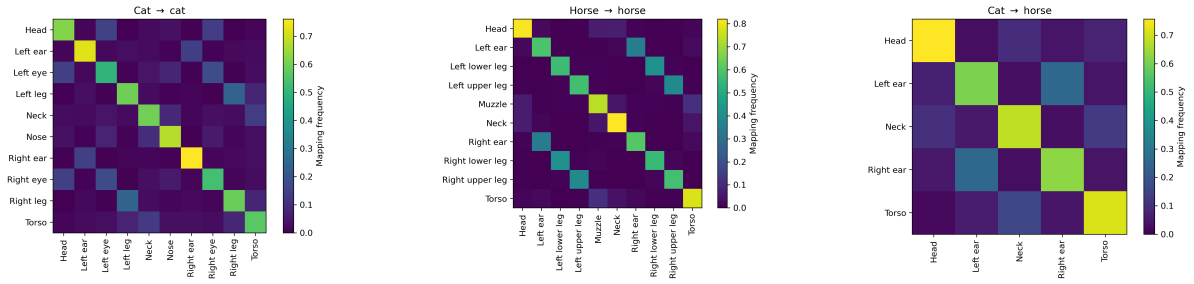


Figure 6.3: Error patterns in animal part mappings. The left two panels show error patterns for *within-category* part mapping (cat-to-cat and horse-to-horse), while the right panel shows *between-category* mapping from cat to horse. Each heatmap represents the frequency of mapping between source parts (rows) and target parts (columns). Perfect mapping corresponds to all values being 1 along the diagonal.

6.3.1 Error analysis

We conducted additional analyses to better understand visiPAM's strengths and limitations in the context of the 2D image mapping task. First, we performed a test in which the target image was horizontally reflected relative to the source image. This experiment allowed us to test the extent to which visiPAM's performance might depend on canonical orientation. We found that this manipulation did not affect visiPAM's performance for either within-category or between-category animal mapping (Table 6.2), suggesting that the model is robust to these kinds of low-level image manipulations.

	Within-category	Between-category
	Animals	
VisiPAM (default)	63.2% (59.1%)	67.9% (59.9%)
VisiPAM (target horizontally reflected)	62.7% (58.6%)	67.9% (59.9%)
Random	10%	20%

Table 6.2: **VisiPAM's performance is robust to image manipulations.** Mapping accuracy for both within-category (e.g., mapping from cat to cat, or horse to horse) and between-category (e.g., mapping from cat to horse) animal comparisons was largely unaffected by horizontal reflection of the target image. 'Random' denotes chance performance (determined by average number of part comparisons in each condition). Values in parentheses reflect chance-normalized performance (percentage of the range between chance performance and 100% accuracy).

We also evaluated the frequency with which visiPAM makes specific types of errors.

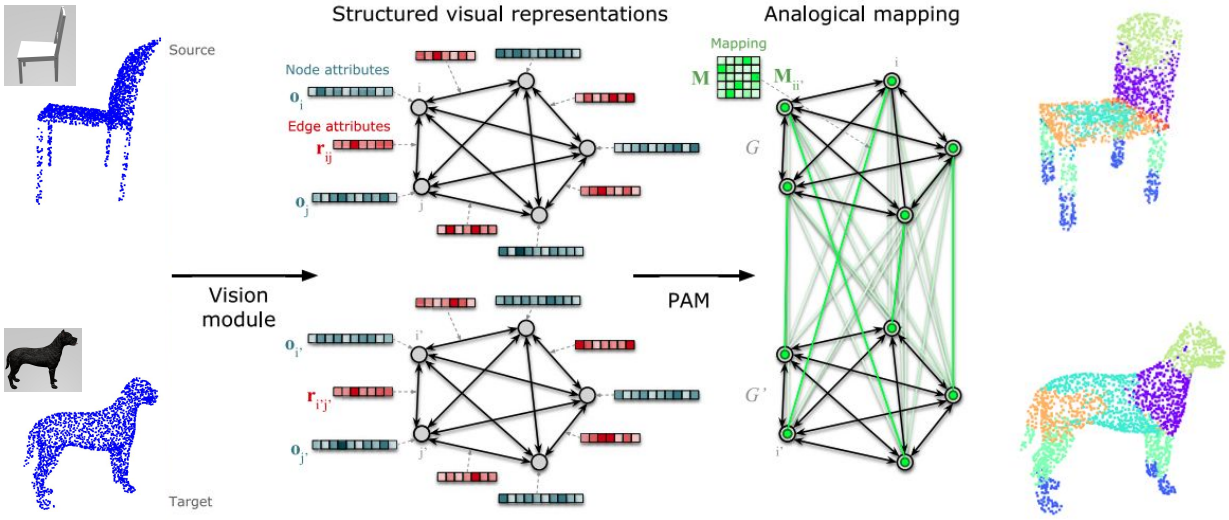


Figure 6.4: **Overview of visiPAM for 3D object mapping.** Parts of a 3D object are clustered based on point features to form graph nodes. The final mapping results are shown on the right, where parts with the same color indicate a correspondence between the source and target objects.

Specifically, we computed the frequency with which each source part was mapped to each target part. The results of this analysis for both within-category and between-category animal comparisons are shown in Figures 6.3. The most common type of error involved confusion of corresponding lateralized parts (e.g., confusion of left and right ears, or left and right legs). There were also occasionally errors related to spatial proximity of parts (e.g., confusion of the neck and head).

6.4 Analogical mapping between 3D objects

We evaluated visiPAM on analogies involving point-cloud representations of 3D objects, as illustrated in Figure 6.4, and compared visiPAM's performance with the results of a human experiment. In that experiment, we presented human participants with analogy problems consisting either of images from the same superordinate category (e.g., an analogy between a dog and a horse), or two different superordinate categories (e.g., an analogy between a chair and a horse). Human behavioral data was especially important as a benchmark in the latter case, since there is not necessarily a well-defined, correct mapping between the parts of the

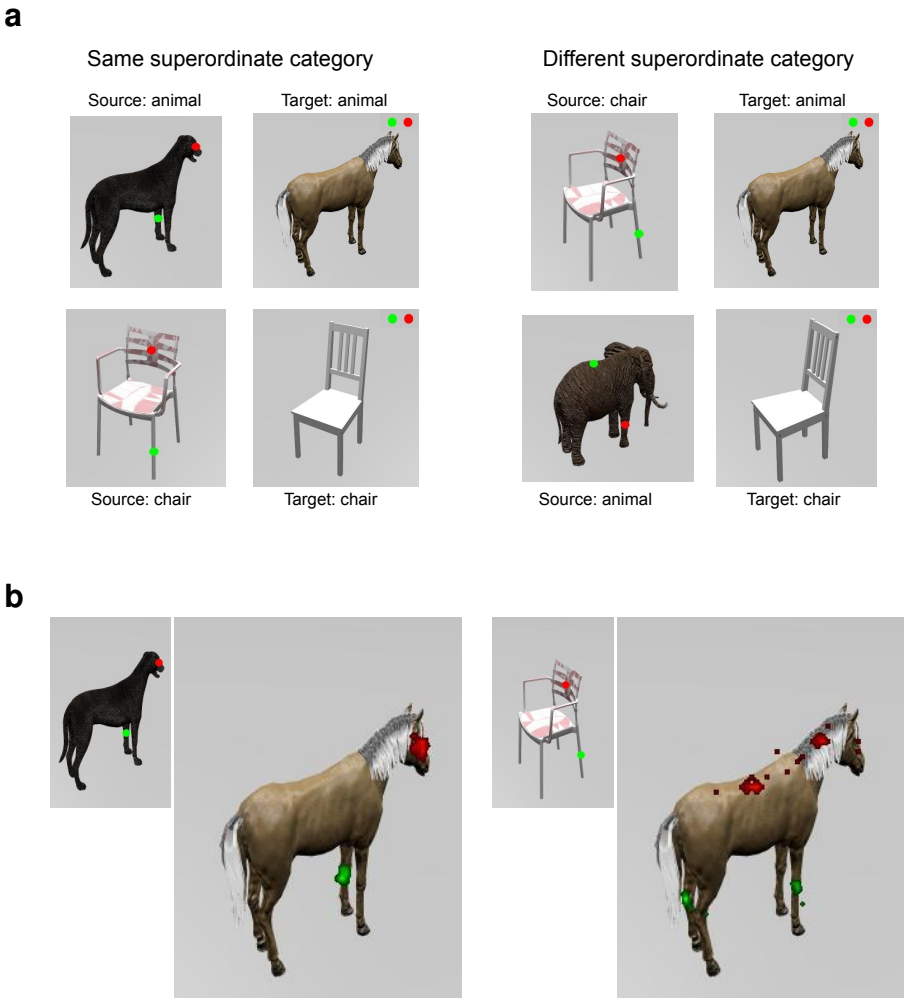


Figure 6.5: **Experiment measuring human performance for mapping between 3D objects.** **(a)** Sample stimuli used in human experiment. Participants were instructed to move markers in target image to locations corresponding to same-color markers in source image. Left panel: example trials with source and target images from the same superordinate object category. Right panel: example trials with images from different superordinate categories. **(b)** Example heatmaps of human marker placements on target images for two comparisons. The intensity of the color indicates the proportion of participants who placed the marker in that location. Source images have been reduced in size for the purpose of illustration. Left panel: marker placements were highly consistent across subjects when source and target came from same superordinate category. Right panel: marker placement showed more variation across participants when source and target came from different superordinate categories.

two objects.

Figure 6.5a shows the analogy task used in the human experiment, which we adapted from a classic task employed by Gentner [46]. On each trial, participants received a source

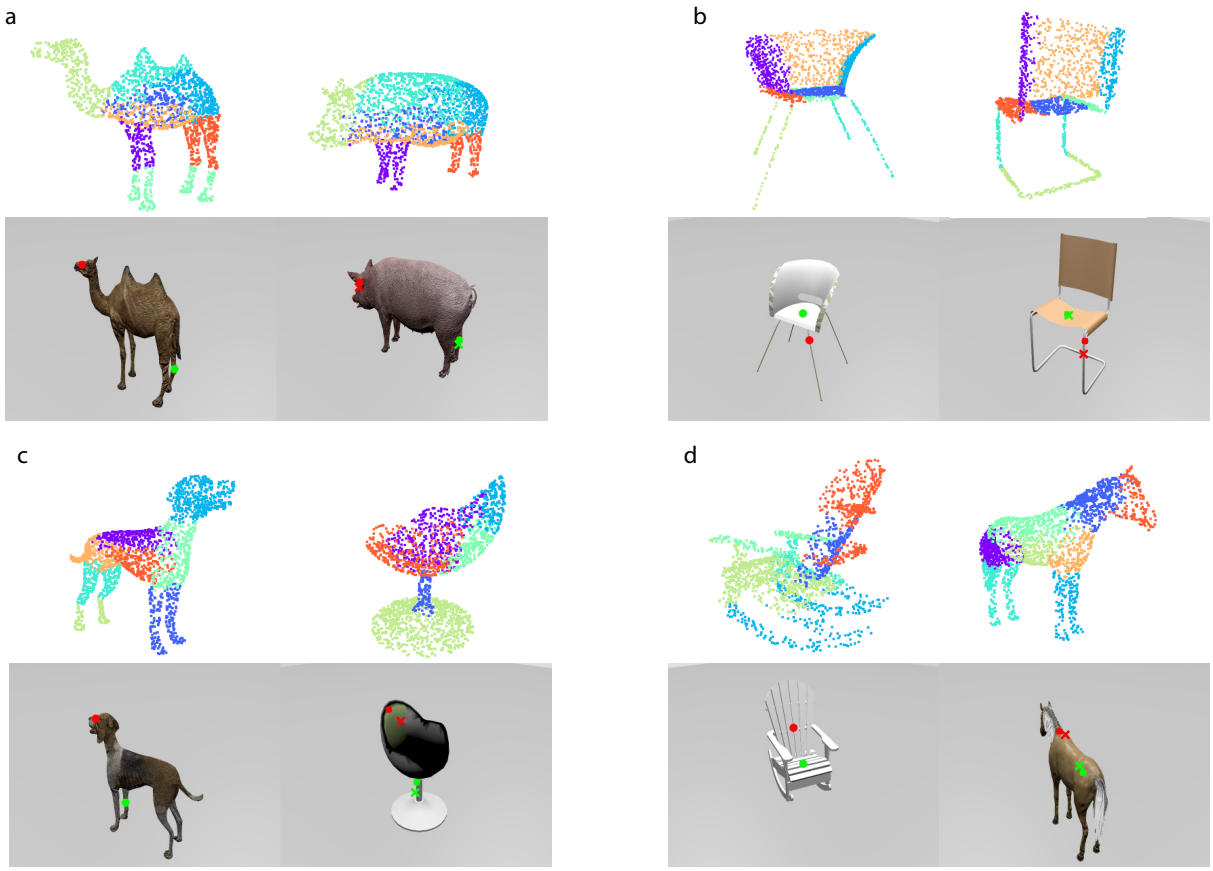


Figure 6.6: **VisiPAM part mappings between 3D objects.** Examples of part mappings for 3D objects represented as point-clouds. Colored regions of point-clouds (upper panels in each figure) indicate the mappings identified by visiPAM for each cluster in the source and target. Images (lower panels in each figure) depict the source marker locations (red and green circles in the lower left panels of each figure), the target marker locations identified by visiPAM (red and green circles in the lower right panels of each figure) and the average of the target marker locations identified by human participants (red and green crosses). **(a)** and **(b)** show examples of problems in which the source and target are from the same superordinate category. **(c)** and **(d)** show examples of problems in which the source and target are from different superordinate categories.

image and a target image. The source image had two colored markers placed on the object, while the target image had two markers that appeared in the upper right corner. Participants were instructed to 'move the marker on the top right corner in the target image to the corresponding location that maps to the same-color marker in the source image.' Figure 6.5b shows two representative examples of human responses. Marker placements from different participants are shown as a heatmap on the target image. When presented with source and target images from the same superordinate categories, there was generally strong agreement

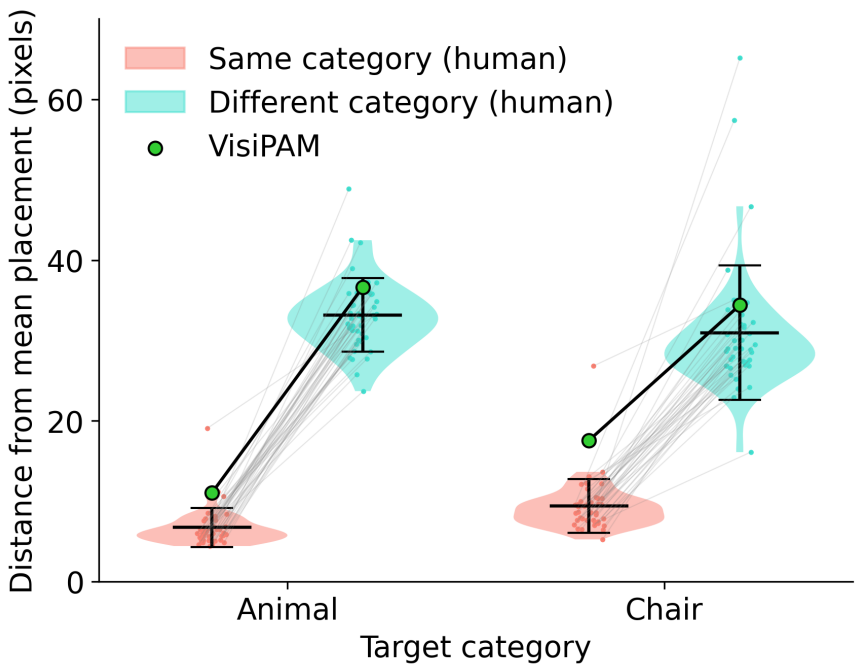


Figure 6.7: **Comparison of visiPAM with human behavior.** Violin plot comparing human placements and visiPAM predictions. Strip plots (small colored dots) indicate average distance of marker locations from the overall human mean, one point for each participant. Thin gray lines show the within-participant differences for the same- vs. different-superordinate-category conditions. Violin plots exclude data points greater than 2.5 standard deviations away from mean, though these individual data points are included in strip plots. Black horizontal lines indicate mean human distances in each condition, and error bars indicate standard deviations. Large green dots indicate visiPAM predictions (i.e., the distance between model predicted location and human mean location). Both human and visiPAM mappings were more variable when mapping objects from different superordinate categories.

between participants. Responses were significantly more variable when participants were presented with source and target images from different superordinate categories (distance from mean marker placement of 32.06 pixels in the different-superordinate-category condition vs. 8.08 pixels in the same-superordinate-category condition; 2 (target category, animal vs. man-made object) X 2 (category consistency, different- vs. same-superordinate-category) repeated-measures ANOVA, main effect of category consistency $F_{1,40} = 625.37$, $p < 0.0001$). We also found a significant interaction, such that the effect of category consistency was larger for problems where the target category was an animal vs. a man-made object ($F_{1,40} = 19.29$, $p < 0.0001$). There was no significant main effect of target category. For problems involving mapping between different superordinate categories, responses in some trials were bimodal,

with some participants preferring one mapping (e.g., from the back of the chair to the head of the horse), and other participants preferring another mapping (e.g., from the back of the chair to the back of the horse), though the same qualitative effects were still present even after accounting for these separate clusters (main effect of category consistency, $F_{1,40} = 151.53$, $p < 0.0001$; interaction effect, $F_{1,40} = 11.91$, $p < 0.005$.

Figure 6.6 shows some examples of mappings produced by visiPAM for point-cloud representations of the 3D objects used in the human experiment. To apply visiPAM to these problems, we obtained embeddings for each point using the DGCNN, then clustered these points, and assigned each cluster to a node, where the attributes of that node were defined based on the average embedding for all points in the cluster. Edge attributes were defined based on the 3D spatial relations between cluster centers. Figure 6.6 shows examples of successful mappings for all four problem types, though visiPAM also sometimes made errors (see Section 6.4.1 for detailed error analysis).

To systematically compare the model's behavior with human responses, we applied the model to all image pairs used in the experiment, and measured the distance from the marker location predicted by visiPAM to the mean locations of human placements for each pair. Figure 6.7 shows the results of this comparison. We found that visiPAM reproduced the qualitative pattern displayed by human mappings across conditions. As with the human participants, visiPAM's mappings showed greater deviation from the mean human placement when mapping between objects from different vs. same superordinate categories. The size of this difference was also larger when the target category was an animal vs. man-made object, capturing the significant interaction effect observed in the human data (see Figure 6.9 for an explanation of this interaction). Overall, marker locations of analogous parts predicted by visiPAM were an average of 25 pixels from mean locations of human placements, close to the average human distance to mean locations (20 pixels). Relative to the object sizes (average height of 213 pixels and width of 135 pixels), the model and human distances were very similar. The same qualitative results were present when taking the bimodal nature of human

responses into account (i.e., when measuring distance to the mean of the closest human cluster for trials with a bimodal response distribution), with both visiPAM and human participants showing lower variance when behavior was quantified in this manner (Figure 6.11).

We also calculated the item-level correlation across all analogy problems between average human distances from mean placement locations and distances of the model predictions from the same mean locations. VisiPAM reliably predicted human responses at the item level ($r = 0.70$). In addition, as with the results for analogies between 2D images, we found that visiPAM performed best when mapping involved both node and edge similarity (Supplementary Table B.3). The ability of visiPAM to predict human responses at the item level was impaired both when focusing exclusively on node similarity ($r = 0.61$ for $\alpha = 1$), and when focusing exclusively on edge similarity ($r = 0.60$ for $\alpha = 0$).

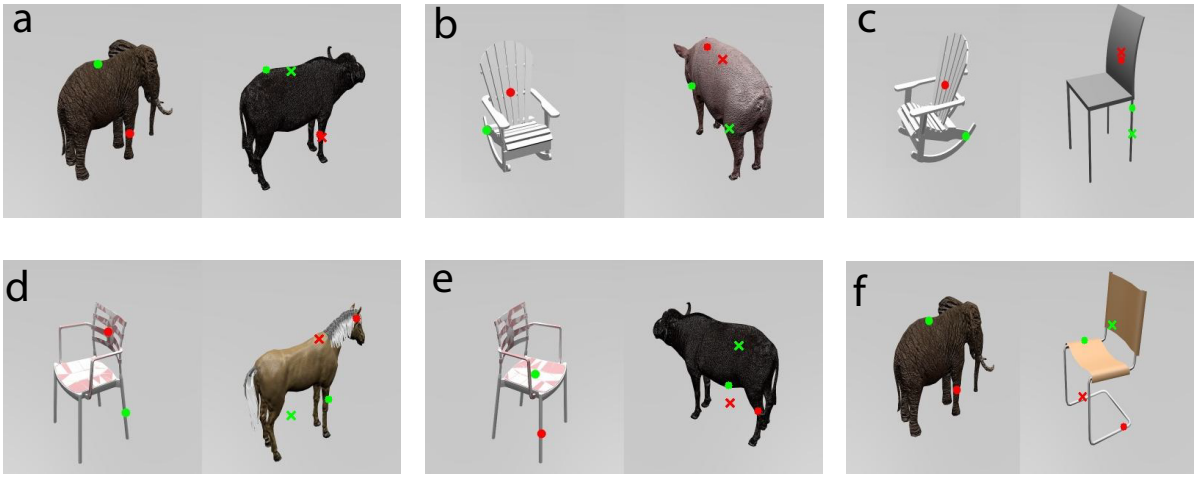


Figure 6.8: **Examples of errors made by visiPAM on 3D mapping task.** In each panel, images on the left panel are source images, and images on the right are target images that show both visiPAM predictions (circle) and human response centers (cross).

6.4.1 Error Analysis

We conducted a qualitative analysis of the errors generated by visiPAM in the context of the 3D marker mapping experiment. We identified two primary causes of errors in predicting human marker placements: (1) the model's over-reliance on local 3D geometric properties (e.g., curvature, corner, T-junctions); and (2) a lack of functional knowledge or semantic information pertaining to the parts. Figure 6.8 illustrates a few examples in which visiPAM gives undue attention to local geometric structure in its mapping decisions. Circles indicate visiPAM's predicted mapping locations, and crosses indicate the mean location of human responses. For example, for the image pairs in Figure 6.8a, visiPAM maps the green marker on the back of the elephant (source) to a position (green circle) located on the lower back of the ox (target). For this mapping, the model largely relies on local geometric commonalities due to their similar surface curvature. Humans, on the other hand, map the green dot to the middle of the ox's spine based on knowledge of overall body structure. In Figures 6.8b and 6.8c, the model maps the green marker using local geometric cues of the corner (L junction). In contrast, humans use functional knowledge to select the back leg of the target object (green cross), which provides support for the object. The mapping of chairs in Figure 6.8d shows a similar mapping error (red circles) based on shared vertical surfaces, whereas humans appear to rely on semantic knowledge to match the "back" of the chair to the "back" of the horse. Similarly, in Figure 6.8e, the model maps the green marker on the basis of shared flat surfaces (e.g., seating surface of a chair, and flat belly of an ox), whereas humans are sensitive to the functional support relations. Lastly, Figure 6.8f shows an interesting case in which the model maps the front leg of the elephant (red marker) to a horizontal bar of the chair. In this instance, the model matches the back legs of the elephant to the chair's vertical legs; accordingly, because the model favors one-to-one mapping, the two front legs of the elephant are mapped to the remaining narrow surfaces (despite the mismatch of orientations).

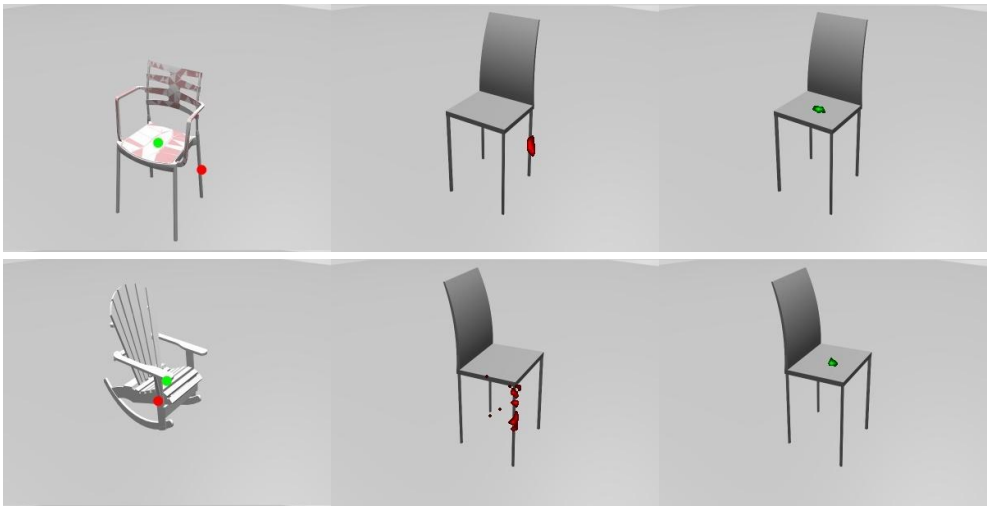


Figure 6.9: **Explanation of increased variability in chair-to-chair condition.** Both visiPAM and human participants displayed higher variability in the chair-to-chair condition than in the animal-to-animal condition. This difference was likely driven by the inherent diversity of chair shapes. To illustrate this, we have displayed two example problems. When the target and source chair had a similar shape (top panel), mappings had very low variability (heatmaps display distribution of human mapping judgments). When the target and source chair had very different shapes (bottom panel), mappings had much higher variability.

6.4.2 Consistency analysis

A noteworthy phenomenon is the consistent utilization of semantic, functional and visual information in human judgments when deciding the relative relation between the two marker locations. As illustrated in the right panel of Figure 6.5b, this tendency resulted in the formation of two distinct clusters for each marker in the between-category condition, with one cluster reflecting semantic knowledge (mapping the red dot on the back of the chair to the back of the horse) and the other based on spatial similarity (mapping the back of the chair to the neck of the horse). Depending on the mapping chosen for the red dot (the back of the chair corresponds to the back or neck of the horse), the green dot was placed on different legs of the horse.

To systematically analyze whether humans make consistent judgments for the two marker placements, we conducted the following analysis. First, we employed a dip test (with a threshold of $p < 0.05$) to determine whether marker locations reported by human participants

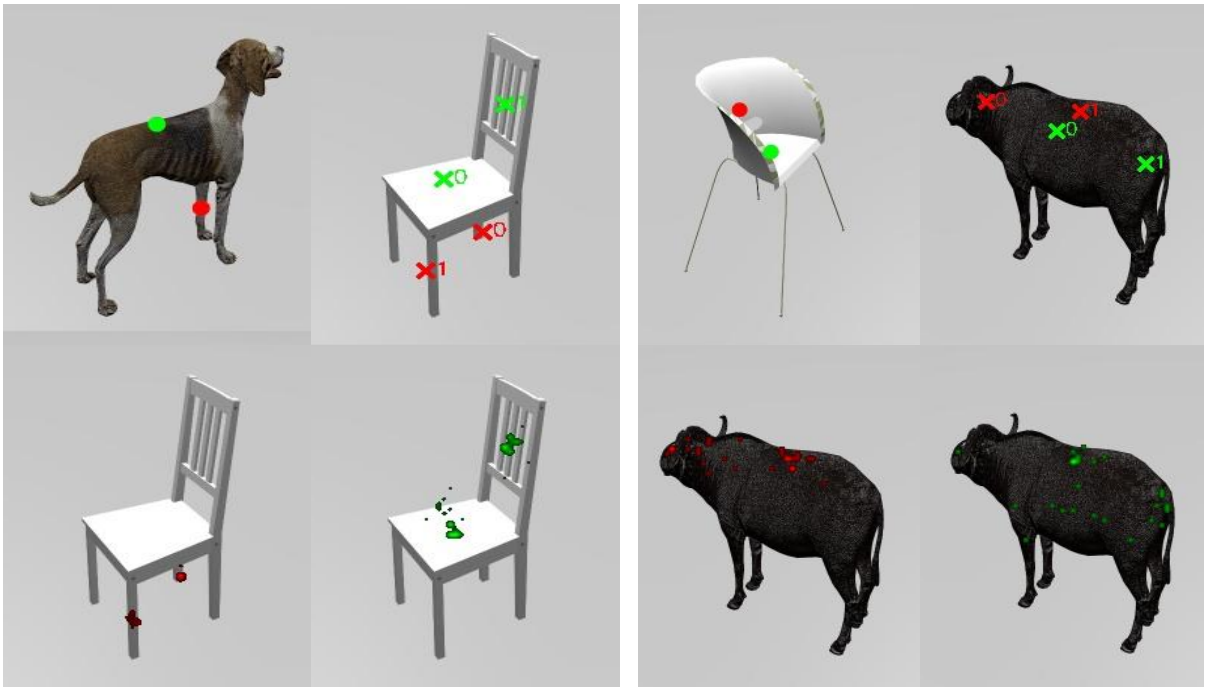


Figure 6.10: **Examples illustrating marker consistency.** Two example problems are shown. For each problem, human mapping judgments were bimodal, with two clusters of responses for each of the green and red markers. These clusters were generally consistent with a strategy that was based on either spatial relational information or semantic information. The clusters labeled with the index 1 follow a semantic strategy, e.g. mapping the back of the dog to the back of the chair (left panel), or mapping the back of the chair to the back of the buffalo (right panel). The clusters labeled with the index 0 follow a spatial strategy, e.g. mapping the back of the dog to the seat of the chair. Human responses were most often consistent, in the sense that both marker placements were governed *either* by a semantic strategy *or* a spatial strategy.

had a bimodal distribution. The dip test revealed 22 trials in which both red and green markers were bimodally distributed. For these trials, we used the KMeans++ algorithm [4] to segregate responses into two clusters. We proceeded to examine the consistency between the reported marker locations. In most of these 22 problems, we observed consistency in the type of information utilized by participants, with both markers being consistently matched based on either visual or semantic information.

Figure 6.10 presents two examples that illustrate marker consistency. Each marker in the figure has two cluster centers with the same color, and a number plotted next to the cluster center indexing the cluster. Out of 41 participants, 15 placed the green and red markers in cluster 0 in the left panel of the four images, whereas 16 placed the markers in cluster 1; 10 other participants did not follow a semantic or visual pattern. In this context, cluster 0 represents mapping primarily based on geometric properties in the object, indicating that participants relied on visual similarity between the source and target objects. In contrast, cluster 1 is a mapping based on semantic knowledge: mapping the back of the dog to the back of the chair despite their different local geometry properties. When participants mapped the back of the dog to the back of the chair, the front of the chair is on the left of the chair image. Hence, the front leg of the dog (red dot) would map to the cluster-1 location of the chair.

Mapping consistency was also observed in the mapping between a chair and a buffalo (right panel). Specifically, out of the total of 41 participants, 13 performed mapping based on visual similarity, placing both the green and red markers in cluster 0. In contrast, 24 participants relied on semantic similarity, placing the markers in cluster 1. The other 4 participants did not show any apparent pattern in their judgments. These analyses revealed a general pattern of consistent mapping among the participants, with a majority utilizing semantic information for mapping. Overall, across the 22 between-category mapping problems (for which 41 participants yielded a total of 902 trials), 568 trials (63%) showed a consistent mapping strategy for human judgments.

An examination was also conducted to determine if visiPAM exhibited a consistent mapping strategy. The results revealed that in 19 out of the 22 problems, VisiPAM consistently applied the mapping strategy based on visual similarity. Thus, similar to human participants, visiPAM was generally consistent in the mappings that it produced.

6.4.3 Cluster analysis

To account for the possibility that participants may have formed multiple response modes in certain mapping problems, we performed a cluster analysis based on the distribution of human responses. Each problem was classified as having either a unimodal or bimodal response pattern using Hartigan's dip test with a significance threshold of $p < 0.05$. For unimodal problems, responses were quantified using the distance to the overall mean (as in the primary analysis shown in Figure 6.7). For bimodal problems, responses were clustered into two groups, and the distance to the closest cluster mean was used as the measure of alignment—for both human and visiPAM responses. All problems in the same-superordinate-category condition were classified as unimodal, while 48.4% of problems in the different-superordinate-category condition were classified as bimodal. Even after accounting for bimodality, human and visiPAM responses remained more variable in the different-superordinate-category condition, as shown in Figure 6.11.

6.5 Discussion

We have presented a computational model that performs zero-shot analogical mapping based on rich, high-dimensional visual inputs. To accomplish this, visiPAM integrates representations derived using general-purpose algorithms for representation learning together with a similarity-based reasoning mechanism derived from theories of human cognition. Our experiments with analogical mapping of object parts in 2D images showed that visiPAM outperformed a state-of-the-art deep learning model, despite receiving no direct training on

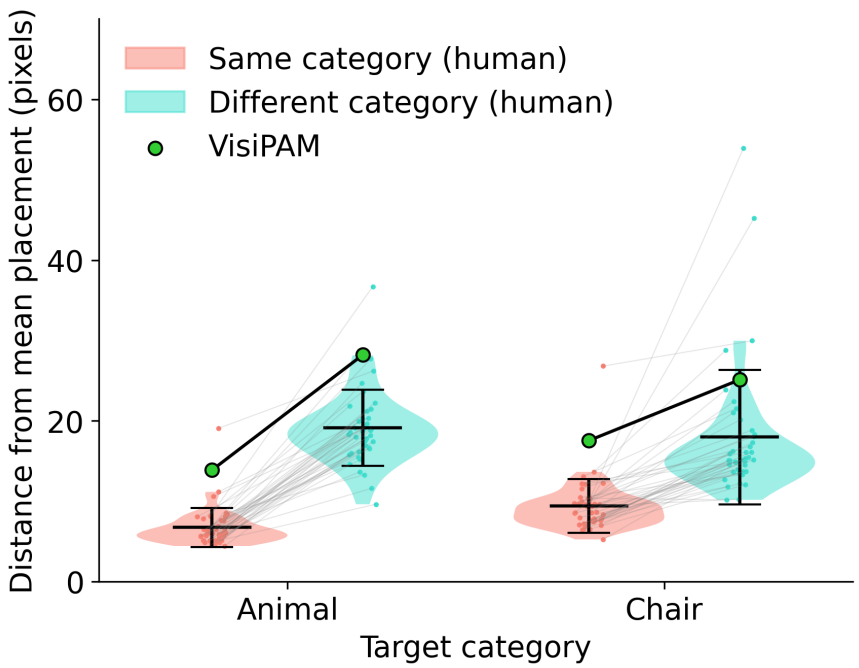


Figure 6.11: Comparison of visiPAM with human behavior after accounting for bimodal response patterns. Violin and strip plots show distances to the overall or cluster mean in the same- and different-superordinate-category conditions. Green dots represent visiPAM predictions; black lines and error bars indicate human means and standard deviations.

the analogy task. Our experiment with mapping of 3D object parts showed that, when armed with rich 3D visual representations, visiPAM matched the pattern of human mappings across conditions.

These results build on other recent work that has employed similarity-based reasoning over representations derived using deep learning [67, 102, 104], enabling analogy to emerge zero-shot. Relative to this previous work, visiPAM is unique in that it both operates over real-world visual inputs, and performs analogical *mapping* (rather than computing similarity of pre-mapped components). VisiPAM also has commonalities with some recent deep learning models. These include models of visual analogy that are applied to real-world visual inputs [31, 89, 119, 131], and models that incorporate graph-based representations similar to visiPAM's structured visual representations [10, 78, 132]. Unlike visiPAM, these deep learning models all involve end-to-end training on a large number of examples for a specific task, with limited generalization to new analogy tasks. Among deep learning

models, visiPAM is most closely related to models that incorporate a 'relational bottleneck', notably involving a central role for similarity-based mechanisms [3, 76, 155]. These models achieve better out-of-distribution generalization in relational tasks, and require significantly fewer training examples than standard deep learning models, but unlike visiPAM they are still centered around end-to-end training on specific tasks, and therefore are not capable of zero-shot reasoning on completely novel tasks.

A few recent studies have applied a more traditional cognitive model of analogical mapping, the Structure-Mapping Engine (SME), to visual analogy problems involving simple geometric forms [98, 99]. This work represents an impressive effort, showing that SME is able to solve the majority of problems in the Raven's Standard Progressive Matrices [127], a commonly used measure of fluid intelligence. However, a significant limitation of this approach lies in its traditional symbolic representations, which are derived from image-based inputs using hand-designed algorithms. This characteristic arguably limits the ability of such an approach to address more complex, real-world inputs such as natural images. Other so-called 'neurosymbolic' methods have attempted, with some success, to extract symbolic representations directly from pixel-level inputs, which can then be passed to a task-specific symbolic algorithm [161]. However, these approaches may not be able to capture the richness of real-world perceptual inputs, or the wide variety of tasks that human reasoners can perform over those inputs. Here, by contrast, we have specifically sought to employ representations that preserve as much of this perceptual richness as possible (by representing elements as vectors), while also incorporating the structured nature of symbolic representations (by maintaining explicit bindings between entities and relations). This was made possible by using a probabilistic reasoning algorithm capable of accommodating such high-dimensional, uncertain inputs.

Human analogical reasoning is thought to be driven by similarity both at the relational level and at the basic object level [80]. This latter influence has often been framed as a deficiency in human reasoning – an inability to achieve the pure abstraction that would

presumably be enabled by focusing exclusively on relations. An important finding from the present study is that visiPAM performed best when constrained both by edge (relational level) and node (object level) similarity. This result suggests that the influence of object-level similarity in human analogical reasoning may be driven in part by the fact that it is a useful constraint in the context of complex, real-world analogy-making [9].

One limitation of the present work is the lack of a method for deriving 3D representations (such as point-clouds) from the 2D inputs that human reasoners typically receive. We found the incorporation of topological information (e.g., whether two parts are connected) significantly improved visiPAM's mapping performance, but it would be desirable to derive this information directly from 2D inputs. Some recent advances may make it possible to take a step in that direction [126], but visiPAM will likely benefit from the rapid pace of innovation in representation learning algorithms. We are particularly excited about the continued development of general-purpose, self-supervised algorithms (such as the iBOT algorithm that we employed [173]), which we hope will provide increasingly rich representations over which reasoning algorithms such as PAM can operate. Similarly, it will be useful in future work to incorporate non-visual semantic and functional information into visiPAM's representations. We found that this kind of information was a significant contributing factor to human mapping judgments. Potential sources of such knowledge include word embeddings [110], multimodal embeddings [121], and representations of relations between verbal concepts [104].

Finally, an important next step will be to expand the capabilities of visiPAM to incorporate other basic cognitive elements of analogy. For example, human analogical reasoning often involves an interaction between bottom-up and top-down processes, in which the reasoning process can sometimes guide the reinterpretation of perceptual inputs [23, 61, 111]. VisiPAM in its current formulation is purely bottom-up – visual representations are constructed by the vision module, and then passed to PAM to perform mapping – but a promising avenue for future work will be to incorporate top-down feedback to the vision module. One possibility for doing so is to incorporate additional parameters (e.g., parameters for visual feature

attention) to visiPAM's optimization procedure, such that the visual representations are modulated based on the downstream graph-matching objective. Additionally, the present work focused on the identification of analogies between pairs of images, but human reasoners are also capable of inducing more general schematic representations that capture abstract similarity across multiple examples [48]. Future work should investigate how visiPAM's reasoning mechanisms can be extended to capture this capacity for schema induction. There are many exciting prospects for the continued synergy between rich visual representations and human-like reasoning mechanisms.

Chapter 7

Conclusion

This dissertation investigates a foundational question that bridges cognitive science and artificial intelligence: how do humans perceive and reason about visual relations, and to what extent can machine learning models replicate or approximate this relational reasoning? Relational understanding is central to human cognition, yet remains a persistent challenge for artificial systems. Through a series of five interrelated studies, this dissertation examines the cognitive and computational mechanisms that support relational reasoning, drawing on behavioral experiments, computational modeling, and systematic model evaluation.

Chapters 2 and 3 highlight a persistent and substantial gap between human relational reasoning and AI performance. Chapter 2 presents a behavioral benchmark of visual relation detection under rapid exposure, establishing human accuracy across various relational types and using these data to evaluate the performance of visual and multimodal models. Chapter 3 extends this finding, showing that even the most advanced multimodal generation systems struggle on out-of-distribution and compositional visual scenes. These results collectively confirm that modern AI systems, despite their successes, still fall far short of the relational inference flexibility exhibited by humans, especially when task demands diverge from training distributions

To address this, Chapters 4–6 investigate how integrating cognitive principles can effectively

narrow this gap. Chapter 4 explored the role of hierarchical abstraction in 3D object recognition. It demonstrates that humans exhibit robust recognition performance even under degraded visual input, and that models incorporating hierarchical downsampling are better aligned with human behavior, suggesting the importance of coarse-to-fine shape processing. Chapter 5 shows that human participants spontaneously align part-whole relations across objects, and that a compositional model based on learned object-part representations can closely mirror this analogical reasoning. Chapter 6 introduces a vision-based probabilistic analogical mapping framework, VisiPAM, which integrates learned visual embeddings with structural inference. Notably, VisiPAM requires no task-specific training yet approximates human performance in analogy tasks, highlighting the power of domain-general relational reasoning.

In sum, these chapters reveal two critical insights. First, the gap between human and machine relational reasoning remains considerable, particularly when tasks require abstraction, compositional inference, or generalization beyond the training distribution. Second, the findings show that incorporating cognitive principles like hierarchical abstraction, compositional representation, and structural inference into model architectures significantly enhances relational performance and alignment with human behavior. By integrating insights from human cognition into computational modeling, this dissertation advances toward a more principled and interpretable framework for artificial intelligence—one that moves beyond pattern matching to reasoning with structured, meaningful relations.

Future Directions and Open Questions

This dissertation lays a foundation for understanding relation perception and reasoning in both humans and machines, but it also opens several compelling directions for future research aimed at building more generalizable and cognitively aligned AI systems.

Chapter 4 explored how the integration of hierarchical abstraction mechanisms can

enhance the robustness of deep learning models in 3D object recognition tasks. While this approach improved classification performance, it revealed persistent limitations in how current architectures represent the compositional structure of objects. Despite the incorporation of hierarchical features, many deep learning models continue to operate as "black boxes," lacking explicit encodings of how object parts relate to wholes. This stands in contrast to human cognitive processes, where part-whole hierarchies are fundamental to recognition and reasoning. For example, Biederman's Recognition-by-Components (RBC) theory argues that human vision parses objects into primitive volumetric parts ("geons") and their spatial configurations, enabling robust recognition across occlusions and viewpoints [13]. Complementary work by Hummel and Stankiewicz has shown that humans encode not only parts, but also their precise relational arrangements, through mechanisms such as dynamic binding [64, 66].

The absence of such explicit compositional representations in deep models hampers both interpretability and generalization. Without internal structures that mirror part-whole relationships, models struggle to transfer knowledge to novel object configurations or explain their decisions in transparent terms. Visualizing and interpreting the learned features of these models remains an open challenge, further complicating efforts to diagnose or trust model predictions. Addressing these issues will require developing architectures that embed dynamic part-based bindings (more broadly variable-binding) and support structured inference over visual inputs. Hence, aligning model representations with the compositional strategies evident in human vision is an important direction for future research to impose as a foundamental learning objective, rather than a byproduct of learning with massive amount of data.

In Chapter 6, I introduced the VisiPAM model as a step toward human-like analogical reasoning by combining visual embeddings from machine learning models with probabilistic analogical mapping. VisiPAM successfully captured structural correspondences across images and 3D objects without any task-specific training. However, this model still provides ample space for enhancement.

One prominent future direction involves leveraging multimodal large language model

(MLLM) embeddings—such as Flamingo [2], InternVL [30], or Qwen [148]—as rich, general-purpose feature representations. These models are trained on extensive image–text corpora and have been shown to capture nuanced semantic, functional, and relational knowledge within their latent spaces. Instead of relying on supervised fine-tuning, future systems should apply supervision-free reasoning algorithms on top of these embeddings. Methods such as graph-based relational mapping, self-supervised relational reasoning heads, and unsupervised graph neural networks can extract structured relational patterns without human labeling [118, 159, 164]. By combining MLLM embeddings with these unsupervised or self-supervised reasoning methods, models could perform more flexible relational inference and support better generalization while avoiding the need for large annotated datasets.

A second direction expands VisiPAM's relational abstraction. Currently, VisiPAM depends on raw spatial coordinates to define relations between objects, which works for mapping low-level layouts but fails to support higher-order relational generalization. In contrast, the Bayesian Analogy with Relational Transformations (BART) model has demonstrated how structured representations can emerge from unstructured feature vectors through probabilistic inference over learned relational predicates [101]. BART's success underscores the importance of incorporating explicit relational representations—such as comparative and role-based structures—into analogical models for relation reasoning. Inspired by BART, future enhancements could introduce abstract relational tokens into VisiPAM's architecture. NLP transformers like BERT and GPT-3 have demonstrated how contextual embeddings can capture relational semantics implicitly [150]. Likewise, MLLMs that integrate visual and textual modalities offer fertile ground: by anchoring visual representations to descriptive language, these models can endow embeddings with semantic and functional relational structure. The key challenge will be extracting explicit relational abstractions, symbolic or predicate-like representations, from these massive, transformer-based models and integrating them into mapping frameworks like VisiPAM.

Nonetheless, as shown in Chapter 3, current MLLMs are still limited in their capacity

to understand and generalize over compositional scenes and textual inputs. Although these models perform well on many downstream tasks, they often fail when required to compose multiple relational concepts or reason about novel combinations of familiar entities via relations. Such failures reflect a broader challenge: while semantic integration is a strength of MLLMs, compositional inference remains underdeveloped. Overcoming this bottleneck will require advances in both model architecture and training strategies. Architecturally, models should incorporate explicit representations of relations—drawing inspiration from cognitive frameworks like schema induction and analogical mapping—to support more systematic and transferable inference. For instance, integrating symbolic-like relational structures or graph-based inductive biases could help models generalize more reliably across relational compositions. On the training side, methods that prioritize compositional generalization—such as in-context learning with structural prompting, graph-guided self-training, or multi-task curricula emphasizing relational recombination—may instill stronger relational abstraction even in large, pretrained models

Together, these directions reflect a broader research agenda aimed at bridging the gap between human and machine reasoning with relations. By examining visual relations, we can broaden the scope of human research on relation reasoning using realistic images rather than solely relying on text inputs for verbal analogy or synthesic images in Raven's progress matrix test. Meanwhile, by grounding AI models in cognitive principles such as part-whole hierarchies, relation understanding, and analogical mappings, we can work toward systems that not only perform well but also reason in ways that are interpretable, generalizable, and aligned with human cognition.

Appendices

Appendix A

Appendix for Chapter 3

A.1 Code and Data Availability

All experimental materials, code, and data are available at:

https://github.com/andrewjlee0/evaluating_compositionality_VLMs

A.2 Prompts for Relational Image Generation Experiments

To test relational image generation in text-to-image models, we used a set of 75 relational prompts developed by [34]. Due to policy violations identified by DALL-E 3, we made the following modifications to these prompts:

- 'kicking' → 'hugging'
- 'hitting' → 'chasing'
- 'knife' → 'spoon'
- 'a teacup near a cylinder' → 'a teacup near a box'

- ‘a monkey touching an iguana’ → ‘a monkey touching a teacup’
- ‘a cylinder near a bowl’ → ‘a teacup near a bowl’
- ‘a man hindering a man’ → ‘a man hindering a robot’
- ‘a woman pulling an iguana’ → ‘a monkey pulling an iguana’
- ‘a child pulling a box’ → ‘a robot pulling a car’
- ‘a monkey hindering a monkey’ → ‘a monkey hindering an iguana’
- ‘a man helping a man’ → ‘a man helping a monkey’
- ‘a robot helping a robot’ → ’a robot helping a man’

This list describes all modifications for both Experiments 1 and 2. In Experiment 1, we sought to re-test the original list of 75 prompts but changed 25 due to policy violations. In Experiment 2, we modified 15 of these 75 (7 of the already 25 modified ones in Experiment 1 and 8 new ones, resulting in 33 total prompts with at least one modification) to deal with reversed prompts that yielded policy violations, or prompts for which the reversed version was identical (e.g., ‘a man helping a man’). From this set of 75, we then selected 30 prompts and their corresponding reversed versions for Experiment 2. Specifically, we selected the three physical and agentic relations for which DALL-E 3 displayed the highest level of agreement when generating images based on basic prompts. The selected physical relations were ‘near’, ‘on’, and ‘in’. The selected agentic relations were ‘touching’, ‘helping’, and ‘kicking’. Because prompts with the relation ‘kicking’ frequently resulted in policy violations, we replaced ‘kicking’ with ‘chasing’. For each of these 6 relations, we used 5 basic prompts, and 5 reversed prompts in which the subject and object were swapped, resulting in a total of 60 prompts (30 basic and 30 reversed). For each prompt, we used DALL-E 3 to generate 10 images, yielding 600 images in total.

A.3 Details for Human Behavioral Evaluation of Generated Images

We performed three human behavioral experiments to assess the extent to which the images generated by DALL-E 3 matched the prompts. To evaluate images generated from basic relational prompts (Section 3.2.1), we recruited 34 participants (26 female, average age = 20.18 years). Four of these participants were excluded from analysis because they indicated a lack of seriousness in the post-experiment survey. To evaluate images generated from reversed relational prompts (Section 3.2.2), we recruited 43 participants (38 female, average age = 19.6 years). One participant was excluded for taking too long to complete the experiment (> 3000 minutes), and four of these participants were excluded because they indicated a lack of seriousness in the post-experiment survey. To evaluate images generated from compositional relational prompts (Section 3.2.3), we recruited 42 participants (34 female, average age = 20.9 years). Five participants were excluded because they indicated a lack of seriousness in the post-experiment survey. All participants were recruited through the [redacted for anonymity] Psychology subject pool, and were compensated with course credit. All behavioral experiments were approved by the [redacted for anonymity] Institutional Review Board.

All experiments were conducted remotely via an online platform (Figure A.1). On each trial, participants were presented with a prompt, along with a collection of 10 images generated by DALL-E 3, and asked to select all of the images that matched the prompt (Figure A.1). Although there was no time limit, participants were encouraged to respond as quickly and accurately as possible. There were 75 trials for experiment 1 (basic relational prompts), 60 trials for experiment 2 (reversed prompts), and 60 trials for experiment 3 (compositional prompts). Trials were presented in a random order for each participant.

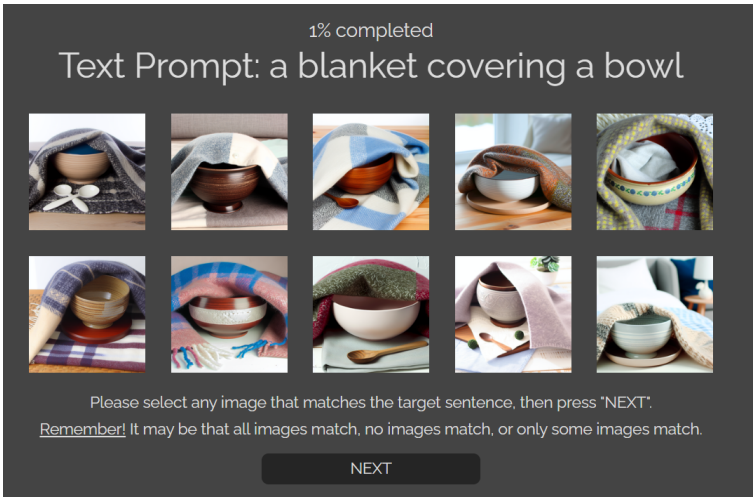


Figure A.1: An example of the display presented to participants in the behavioral experiment.

We also conducted an additional behavioral experiment in which participants assessed images using a Likert scale (from 1 to 7) rather than making a binary judgment. We recruited 45 participants for this study (37 female, average age = 19.84 years). Three participants were excluded from the analysis because they indicated a lack of seriousness in the post-experiment survey. Each participant completed a total of 750 trials, each of which included a single image generated from a basic relational prompt. The results of this experiment were qualitatively similar to the experiment that used a binary decision (Figure A.2), so we used the binary decision paradigm for the experiments with reversed and compositional prompts.

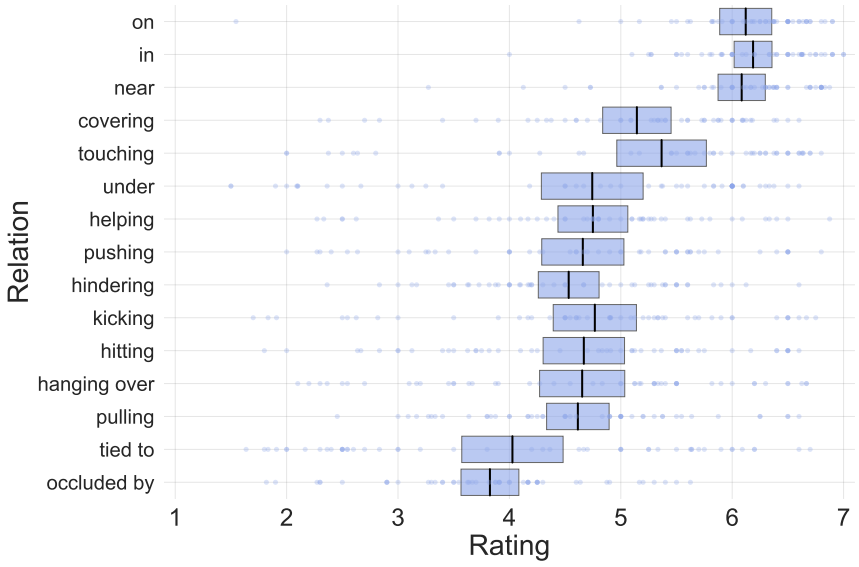


Figure A.2: Results for basic relational prompts when participants responded using a Likert scale from 1 to 7 instead of making a binary judgment. Relations are sorted along the Y axis based on the rank order of agreement in the binary decision paradigm (in Figure 3.2). These results illustrate that the overall qualitative pattern (i.e., the rank ordering of the relations) is very similar between the two paradigms. Each point reflects the average rating for an individual image. Vertical lines indicate the mean rating for each relation, and boxes indicate 95% confidence intervals.

A.4 Additional Examples of Images Generated by DALL-E 3

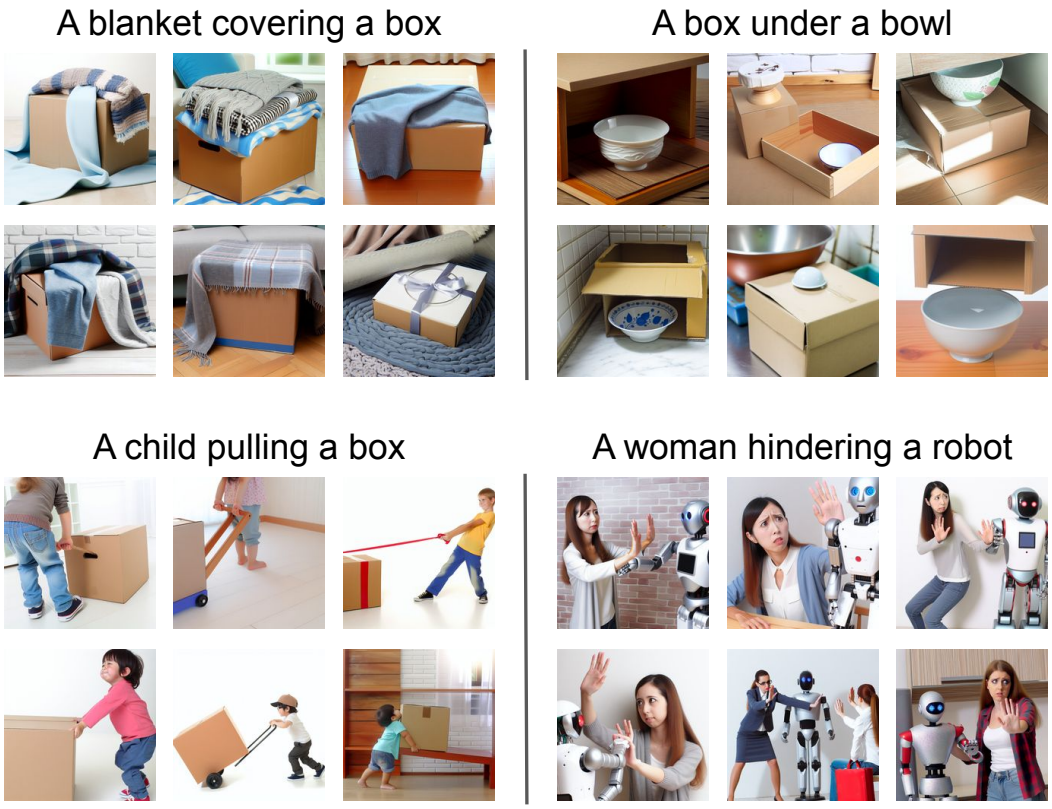


Figure A.3: Additional examples of images generated by DALL-E 3 for basic relational prompts using the 'natural' style.

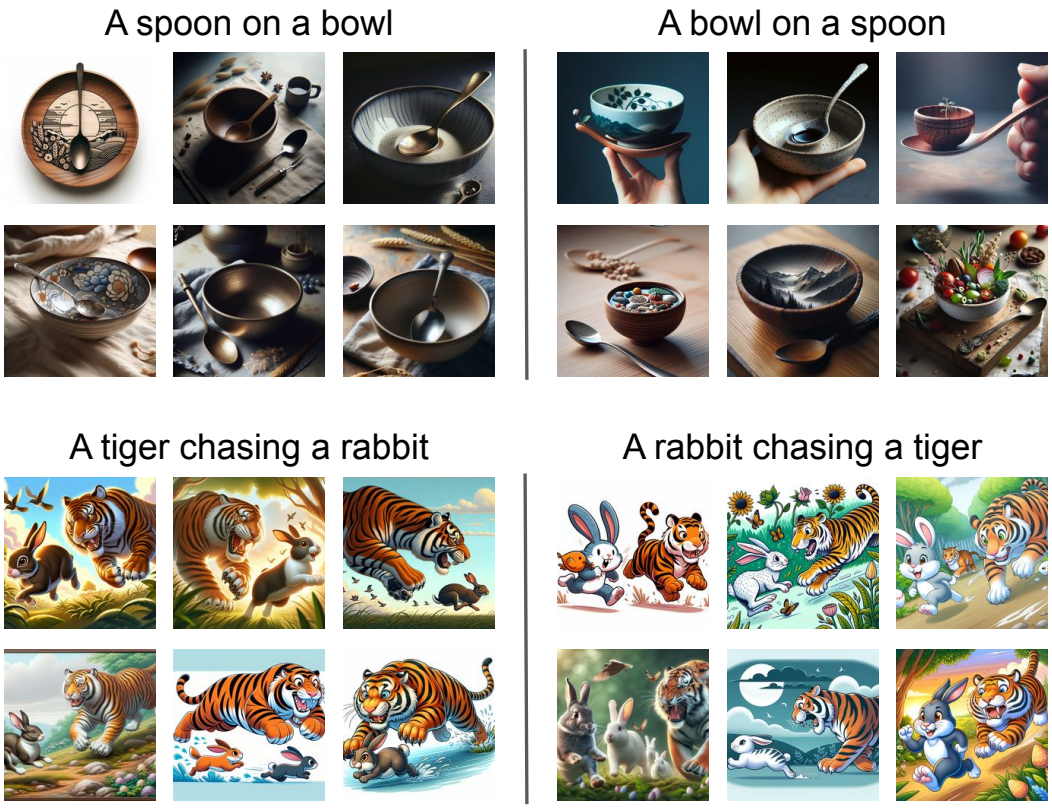


Figure A.4: Additional examples of images generated by DALL-E 3 for basic relational prompts and their corresponding reversed prompts using the 'vivid' style.

Figure A.5: Additional examples of images generated by DALL-E 3 for compositional relational prompts using the 'vivid' style.

A.5 Analysis of Errors in Images Generated by DALL-E 3

We performed an analysis to better understand the types of errors found in the images generated by DALL-E 3. This analysis focused on the images generated for reversed relational prompts. We inspected each of the images with an average agreement of less than 0.2, and categorized them according to the following types of errors: wrong subject, wrong object, wrong relation, and relation binding error. For example, given the prompt 'a rabbit chasing a tiger', a wrong subject error would indicate that the image does not contain a rabbit, a wrong subject error would indicate that the image does not contain a tiger, a wrong relation error would indicate that the image does not contain a 'chasing' relation, and a relation binding error would indicate that the roles are reversed, i.e. the image depicts a tiger chasing

a rabbit. The results (Figure A.6) showed that the overwhelming majority of errors involved either a wrong relation or a relation binding error.

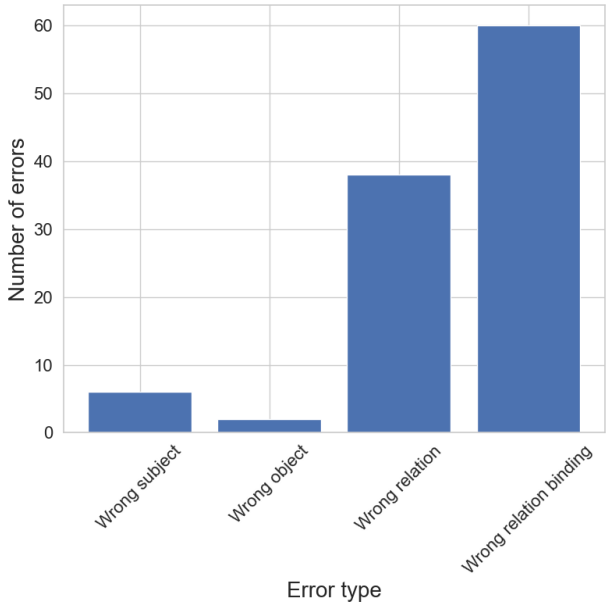


Figure A.6: Error analysis for images generated by DALL-E 3.

A.6 Using Multimodal Language Models to Evaluate Images Generated by DALL-E 3

We performed an analysis to determine whether multimodal language models could be used to automatically evaluate the images generated by DALL-E 3, focusing on the generated images for basic relational prompts. To do so, we prompted each model to rate whether a generated image matched the original prompt on a scale from 0 to 10. We then performed a correlation analysis between these scores and the average score for each image from the human behavioral experiment. As a baseline, we estimated the inter-subject reliability of the human ratings. This was accomplished by randomly splitting the human participants into two groups and computing the correlation between these groups. We performed this simulation 100 times and computed the average of the correlation scores obtained in each

run.

This analysis revealed that the human participant ratings were highly correlated with each other (human-to-human correlation: $r = 0.91$), whereas the models showed much lower correlations with the human ratings (GPT-4v: $r = 0.49$; GPT-4o: $r = 0.51$; Claude 3.5 Sonnet: $r = 0.43$; Qwen2-VL-72B: $r = 0.56$; InternVL2.5-38B: $r = 0.48$). These results indicate that multimodal language models cannot reliably be employed to automate the assessment of images generated by text-to-image models. This is consistent with the results of our evaluation of multimodal language models, which showed that they suffer from limited understanding of compositional images comparable to the limited ability of text-to-image models to generate those images.

A.7 Comparison of Multimodal Language Models with Human Behavioral Data on SVRT

We compared a set of multimodal language models (including GPT-4o, GPT-4-vision-preview, Claude 3.5 Sonnet, QWEN2-VL-72B, and InternVL2.5-38B) with human performance on SVRT for each number of few-shot examples. In the behavioral experiment from [86], episodes were terminated after participants made 7 correct classifications in a row. As a result, in some episodes, participants received fewer than 9 examples before the episode was terminated. For the purposes of comparison, we assume that participants in these episodes would have made correct classifications for subsequent examples. For instance, if a participant made 7 correct classifications for the first 7 examples (thus terminating the episode), we assume that they also would have made a correct classification for an 8th and 9th example. This assumption was necessary to avoid biasing the distribution for 8 and 9 examples toward lower performing participants, since the best performing participants were more likely to terminate episodes early by classifying all examples correctly. In Section A.8.1, we present an experiment in which GPT-4 was evaluated using the same paradigm as human participants, allowing for a

closer comparison.

A.8 Evaluating GPT-4 with Chain-of-Thought Prompting

In the main text, we evaluated GPT-4 on the SVRT and Bongard-HOI by presenting a set of labeled in-context examples followed by a target image for classification. Here, we also present results for experiments that used more extensive 'chain-of-thought' prompting strategies [156], in which language model output is used to perform intermediate reasoning steps. We investigated four separate Chain-of-Thought conditions, involving different amounts of intermediate computation:

- **Chain-of-Thought 0:** No Chain-of-Thought is performed (this corresponds to the default results presented in the main text).
- **Chain-of-Thought 1:** In-context examples were presented one at a time, each coupled with a revised message instructing GPT-4 to describe the contents of the image. These descriptions were then appended to the context window. The final target/query message was the same as the Chain-of-Thought 0 condition.
- **Chain-of-Thought 2:** Descriptions were not solicited for the in-context examples, but the target/query prompt was expanded to encourage compositional thinking across all images. Specifically, GPT-4 was instructed to think carefully about the objects and relations in the series of labeled examples, about the pattern of relations that may determine category membership, and to apply the patterns to the target.
- **Chain-of-Thought 3:** The combination of prompts for Chain-of-Thought 1 and 2.

The results of these experiments are shown below. On Bongard-HOI, the Chain-of-Thought 0 condition performed as well or better than any of the other conditions (Figure A.7). On SVRT, there were no discernible differences between the different Chain-of-Thought

conditions (Figure A.8). Taken together, these results suggest that performance on these tasks cannot be improved via Chain-of-Thought prompting. This is consistent with the interpretation that performance in these models is constrained by a basic difficulty with multi-object perception, rather than downstream reasoning processes.

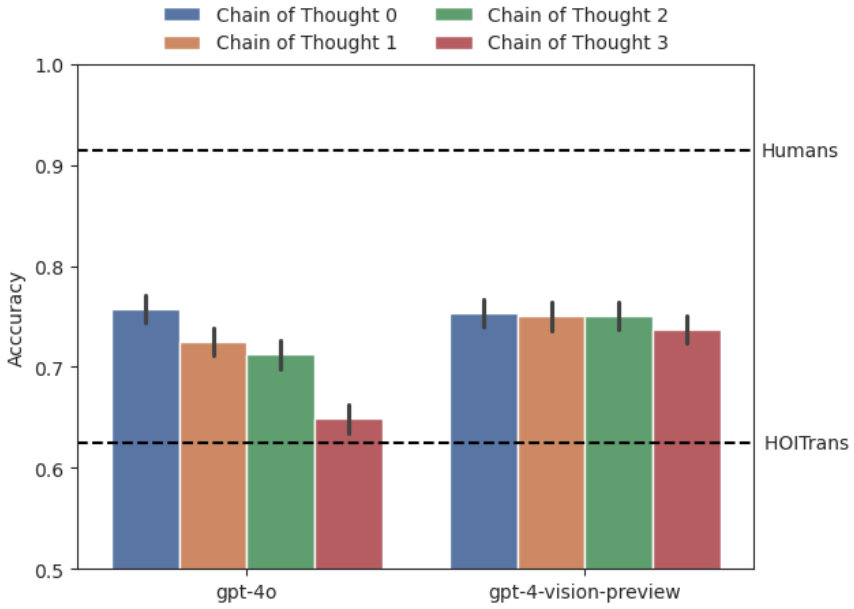


Figure A.7: Bongard-HOI Chain-of-Thought results. Dashed lines represent accuracies for humans and the HOITrans baseline [70]. Error bars reflect binomial standard errors.

A.8.1 Interactive Testing of GPT-4 on SVRT

We also performed an alternative evaluation of GPT-4 on the SVRT involving an active learning paradigm, similar to the paradigm that is typically used with human participants for this task. In that paradigm, GPT-4 was prompted to categorize each in-context example one at a time, receiving immediate feedback after each classification. The model maintained a context window of the previous 9 classifications, which allowed it to learn from past errors. On each trial, GPT-4 was presented with an image and asked to categorize it as either positive or negative. After each classification, the model received feedback ('Correct!' or 'Incorrect!') and acknowledged the feedback before moving on to the next image. As with human participants, this process was continued until there were 7 correct classifications in a

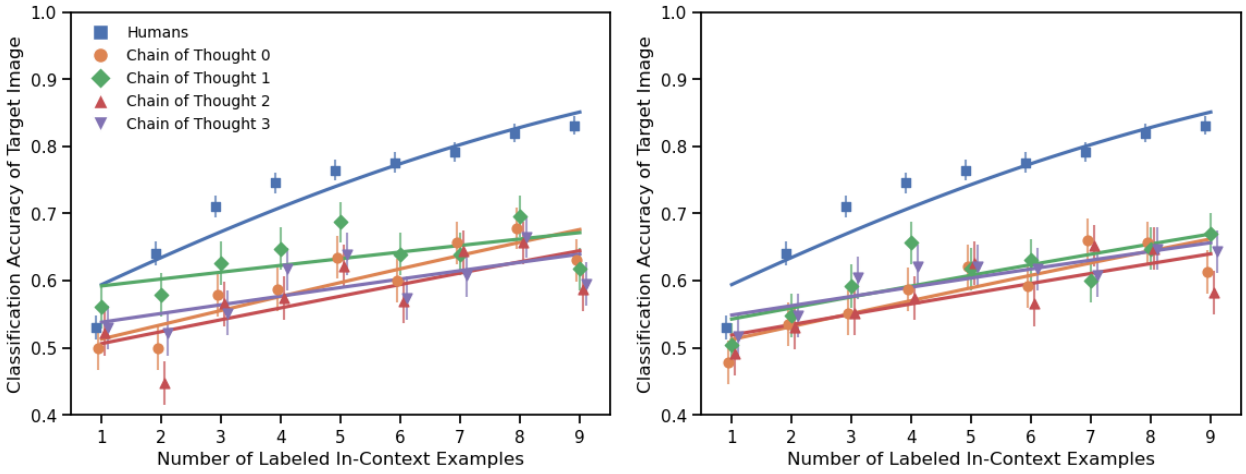


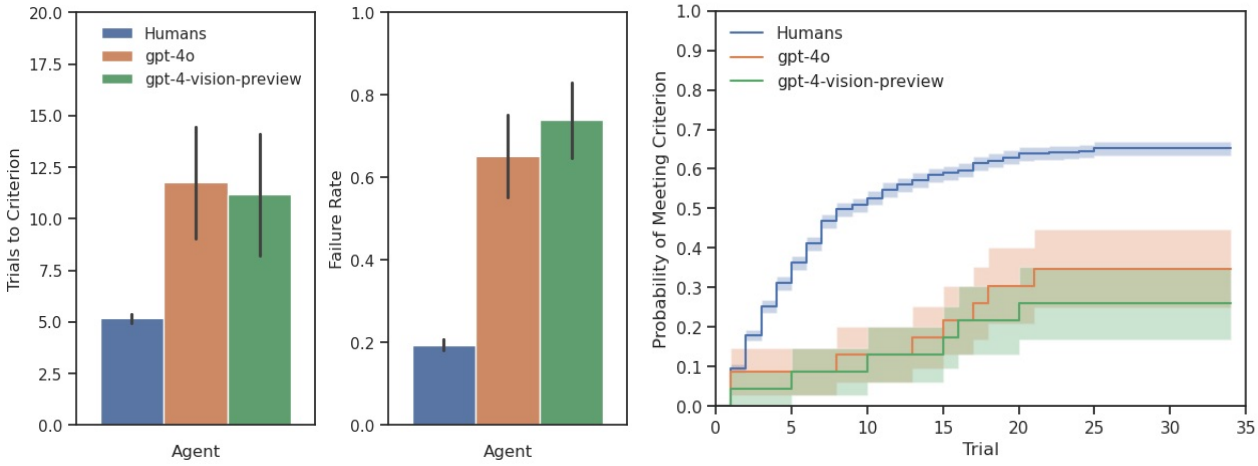
Figure A.8: SVRT Chain-of-Thought results for humans, (a) gpt-4o, and (b) gpt-4-vision-preview. Points are averages across all SVRT concepts and attempts per concept, coupled with binomial standard error bars. Curves are separate logistic regression fits to each chain-of-thought condition.

row or the maximum of 34 attempts was reached.

Two dependent variables were measured per SVRT concept: concept failure (whether criterion was met within the maximum 34 attempts) and trials-to-criterion (no data if concept failure; or total number of classification attempts before 7 correct in a row if criterion met). Figure A.9a displays average trials-to-criterion over non-failed concepts and failure rates over all concepts.

We fit a linear and logistic mixed-effects models to failure and trials-to-criterion data respectively, with participant type (human vs. gpt-4o vs. gpt-4-vision-preview) as a fixed effect and problem ID as a random effect. Pairwise contrasts with Tukey-adjusted p-values of 3 estimates reveal that humans failed less frequently (gpt-4o: $z = 5.14, p < 0.0001$; gpt-4-vision-preview: $z = 5.67, p < 0.0001$) and met criterion faster than than both variants (gpt-4o: $t(625) = 3.96, p = 0.0003$; gpt-4-vision-preview: $t(624) = 3.31, p = 0.0028$).

Because trials-to-criterion data are missing for concepts where criterion was not met (i.e., concept failure), average trials-to-criterion underestimate true rates of learning. However, precise estimates of learning speeds can be established using survival analysis, a method that takes into account censored data (e.g., the frequency of concept failures due to a ceiling of 34



Trials-to-criterion and failure rates

Aalen-Johansen cumulative incidence functions

Figure A.9: Active learning and survival analysis. (a) Trials-to-criterion across all non-failed SVRT concepts and failure rates across all concepts with standard error of the mean and binomial standard error, respectively. (b) Probabilities of meeting criterion by trial, or cumulative incidence functions, estimated by Aalen-Johansen survival analysis, with binomial standard error bands.

attempts) to estimate trial-by-trial probabilities of meeting criterion. A standard approach to survival analysis is to use a Kaplan-Meier estimator for such trial-by-trial probabilities. However, Kaplan-Meier assumes that censorship (i.e., concept failure) is independent of the probability of meeting criterion. Since our experiment violates this assumption by design, we used an Aalen-Johansen estimator with trials-to-criterion as the event of interest and concept failure (censorship) as a *competing risk*.

Figure A.9b displays Aalen-Johansen probability curves. At least half of human participants meet criterion by the 9th trial, whereas gpt-4o and gpt-4-vision-preview have a 13% and 8.7% probability of meeting criterion by the same time. At trial 25, 6.5 people out of 10 meet criterion while neither variant reaches 50% of meeting criterion. To compare probability curves, we fit a Fine-Gray competing risks regression with *agent* as a fixed effect and *SVRT concept ID* as a stratified predictor. The Fine-Gray regression with humans as the reference level reveals that humans reach criterion faster than either GPT-4 variant (gpt-4o: $SHR = 0.25, 95\%CI : 0.13 - 0.47, p < 0.0001$; gpt-4-vision-preview: $SHR = 0.18, 95\%CI : 0.087 - 0.38, p < 0.0001$).

A.9 Testing with Swapped Labels on SVRT

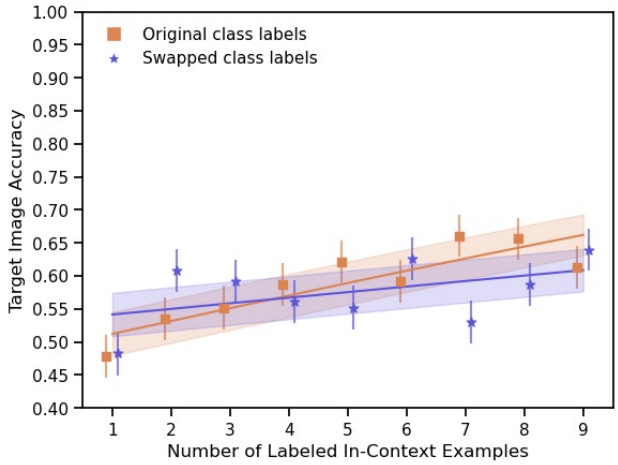


Figure A.10: SVRT few-shot classification accuracy for gpt-4-vision-preview with and without swapped class labels.

Finally, as with Bongard-HOI, we tested whether GPT-4's performance can be explained by memorization of the SVRT dataset. Although all SVRT images were generated from scratch for our experiments, we confirmed that memorization could not explain our results by re-testing gpt-4-vision-preview with swapped class labels. There was no significant difference between performance with original vs. swapped labels (Figure A.10; logistic regression: $z = 0.87, p = 0.38$).

Appendix B

Appendix for Chapter 6

B.1 Ablation analyses

B.1.1 2D image mapping

VisiPAM's edge embeddings involved two different types of spatial relations, one based on angular distance ($\mathbf{r}_\theta$), and one based on vector difference ($\mathbf{r}_\delta$). We performed ablation experiments in which visiPAM's edge embeddings only contained one of these spatial relation types. The results demonstrated that both components contributed significantly to visiPAM's performance (Table B.1).

	Within-category		Between-category
	Animals	Vehicles	Animals
VisiPAM	**63.2% (59.1%)**	**69.5% (58.8%)**	**67.9% (59.9%)**
VisiPAM ($\mathbf{r}_\theta$ only)	56.7% (51.8%)	66.7% (55%)	67.5% (59.4%)
VisiPAM ($\mathbf{r}_\delta$ only)	56.1% (51.2%)	65.4% (53.2%)	52.5% (40.6%)
Random	10%	26%	20%

Table B.1: Analogical mapping with 2D images – edge embedding ablation analyses. Mapping accuracy on part-matching task. VisiPAM performed worse when edge embeddings were based only on angular distance ($\mathbf{r}_\theta$) or vector difference ($\mathbf{r}_\delta$). 'Random' denotes chance performance (determined by average number of part comparisons). Values in parentheses reflect chance-normalized performance (percentage of the range between chance performance and 100% accuracy).

We also carried out an additional experiment to determine whether visiPAM's performance can be further improved by incorporating additional sources of relational information. Specifically, we augmented visiPAM's edge embeddings with information about whether two parts are *connected*. This is an important aspect of relations between object parts that may not be entirely captured by simple spatial relations, especially for non-rigid body objects such as animals. For instance, the spatial relation between two body parts (e.g., the hand and the shoulder) can change dramatically depending on posture, but the topological relationship between body parts (e.g., which body parts are connected) remains unchanged. To test whether this information might further improve visiPAM's performance, we tested a model in which edges were modified based on whether the corresponding object parts were connected (using ground-truth knowledge). For parts that were not connected, the values for the spatial relation components ($\mathbf{r}_\theta$ and $\mathbf{r}_\delta$) were set to 0, while spatial relation embeddings were left unchanged for connected parts. Additionally, an extra dimension was added to the edge embeddings with a value of 1 for connected parts, and a value of -1 for non-connected parts. We found that this version of visiPAM performed even better than the version that included spatial relations only, particularly for mapping problems involving animals (Table B.2). Though this preliminary experiment relied on ground-truth knowledge about the object parts, future work might explore methods for extracting this information directly from images.

Within-category		Between-category
Animals	Vehicles	Animals
72.4% (69.3%)	70.3% (59.9%)	77% (71.3%)

Table B.2: VisiPAM's performance is improved by incorporating topological relations. VisiPAM's performance was improved when edges were augmented with information about whether two parts were connected. Values in parentheses reflect chance-normalized performance (percentage of the range between chance performance and 100% accuracy).

B.1.2 3D object mapping

	Same superordinate category		Different superordinate category		
	Animal-to-animal	Chair-to chair	Animal-to-chair	Chair-to-animal	r
VisiPAM	**11.51**	**17.03**	**34.37**	**36.63**	**0.70**
VisiPAM (nodes only)	13.77	19.96	35.83	37.98	0.61
VisiPAM (edges only)	12.61	19.32	34.86	40.10	0.60

Table B.3: Ablation results for comparison of visiPAM with human behavior. VisiPAM performed better than ablated models based on either node or edge similarity only. Results in the middle four columns reflect average distance between visiPAM and human marker placements (smaller values indicate better fit to human responses) in each of four conditions, defined by the target category (animal vs. chair), and whether or not the source and target objects were from the same (e.g., animal-to-animal) or different (e.g., animal-to-chair) superordinate categories. Results in the rightmost column reflect the item-level correlation between human and visiPAM distances to the human mean placement (higher values indicate a better fit to human responses).

Bibliography

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. *ArXiv*, abs/2204.14198, 2022.

[3] Awni Altabaa, Taylor Webb, Jonathan Cohen, and John Lafferty. Abstractors: Transformer modules for symbolic message passing and relational reasoning. *arXiv preprint arXiv:2304.00195*, 2023.

[4] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. Technical report, Stanford, 2006.

[5] Renée Baillargeon. Physical reasoning in infancy. In Michael S. Gazzaniga, editor, *The Cognitive Neurosciences*, pages 181–204. MIT Press, Cambridge, MA, 1995.

[6] Nicholas Baker, Hongjing Lu, Gennady Erlikhman, and Philip J Kellman. Local features and global shape information in object classification by deep convolutional neural networks. *Vision Research*, 159:60–72, 2019.

[7] Victor Bapst, Alvaro Sanchez-Gonzalez, Carl Doersch, Kimberly L Stachenfeld, Pushmeet Kohli, Peter W Battaglia, and Jessica B Hamrick. Structured agents for physical construction. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 464–474. PMLR, 2019.

[8] David G.T. Barrett, Felix Hill, Adam Santoro, Ari S. Morcos, and Timothy Lillicrap. Measuring abstract reasoning in neural networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 511–520. PMLR, 2018.

[9] Miriam Bassok, Valerie M Chase, and Shirley A Martin. Adding apples and oranges: Alignment of semantic and formal knowledge. *Cognitive Psychology*, 35(2):99–134, 1998.

[10] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

[11] Marlene D. Berke, Robert Walter-Terrill, Julian Jara-Ettinger, and Brian J. Scholl. Flexible goals require that inflexible perceptual systems produce veridical representations: Implications for realism as revealed by evolutionary simulations. *Cognitive Science*, 46(10):e13195, 2022.

[12] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

[13] Irving Biederman. Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94(2):115–147, 1987.

[14] Irving Biederman, Robert J. Mezzanotte, and Jan C. Rabinowitz. Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, 14(2):143–177, 1982.

[15] Mikhail Moiseevich Bongard. *Pattern Recognition*. Spartan Books, New York, 1970. Edited by Joseph K. Hawkins. Translated from Russian by Theodore Cheron. Originally published as *Problema uznavaniia*.

[16] Jane Bromley, J.W. Bentz, Leon Bottou, Isabelle Guyon, Yann LeCun, C. Moore, Eduard Säckinger, and Roopak Shah. Signature verification using a "siamese" time delay neural network. In *Advances in Neural Information Processing Systems*, volume 6, pages 737–744. Morgan Kaufmann, 1993.

[17] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.

[18] Christopher P Burgess, Loic Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matt Botvinick, and Alexander Lerchner. Monet: Unsupervised scene decomposition and representation. *arXiv preprint arXiv:1901.11390*, 2019.

[19] Declan Campbell, Sunayana Rane, Tyler Giallanza, Nicolò De Sabbata, Kia Ghods, Amogh Joshi, Alexander Ku, Steven M Frankland, Thomas L Griffiths, Jonathan D Cohen, et al. Understanding the limits of vision language models through the lens of the binding problem. *arXiv preprint arXiv:2411.00238*, 2024.

[20] Patricia A. Carpenter, Marcel Adam Just, and Patricia Shell. What one intelligence test measures: A theoretical account of the processing in the raven's progressive matrices test. *Psychological Review*, 97(3):404–431, 1990.

[21] Raymond B Cattell. *Abilities: Their structure, growth, and action*. Houghton Mifflin, 1971.

[22] Patrick Cavanagh. The language of vision. *Perception*, 50(3):195–215, 2021.

[23] David J. Chalmers, Robert M. French, and Douglas R. Hofstadter. High-level perception, representation, and analogy: A critique of artificial intelligence methodology. *Journal of Experimental & Theoretical Artificial Intelligence*, 4(3):185–211, 1992.

[24] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. Shapenet: An information-rich 3d model repository. Technical report, arXiv preprint arXiv:1512.03012, 2015.

[25] Agneet Chatterjee, Yiran Luo, Tejas Gokhale, Yezhou Yang, and Chitta Baral. Revision: Rendering tools enable spatial fidelity in vision-language models. In *Computer Vision – ECCV 2024*, volume 15088 of *Lecture Notes in Computer Science*, pages 339–357. Springer, 2025.

[26] Kezhen Chen, Irina Rabkina, Matthew D. McLure, and Kenneth D. Forbus. Human-like sketch object recognition via analogical learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1336–1343, 2019.

[27] 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)*, pages 833–851. Springer, 2018.

[28] Lin Chen. Topological structure in visual perception. *Science*, 218(4573):699–700, 1982.

[29] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1979–1986, 2014.

[30] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

[31] Jonghyun Choi, Jayant Krishnamurthy, Aniruddha Kembhavi, and Ali Farhadi. Structured set matching networks for one-shot part labeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3627–3636, 2018.

[32] François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

[33] Colin Conwell, Rupert Tawiah-Quashie, and Tomer Ullman. Relations, negations, and numbers: Looking for logic in generative text-to-image models. *arXiv preprint arXiv:2411.17066*, 2024.

[34] Colin Conwell and Tomer Ullman. Testing relational understanding in text-guided image generation. *arXiv preprint arXiv:2208.00005*, 2022.

[35] David Ding, Felix Hill, Adam Santoro, Malcolm Reynolds, and Matt Botvinick. Attention over learned object embeddings enables complex visual reasoning. *Advances in neural information processing systems*, 34:9112–9124, 2021.

[36] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold,

Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[37] Brian Falkenhainer, Kenneth D Forbus, and Dedre Gentner. The structure-mapping engine: Algorithm and examples. *Artificial intelligence*, 41(1):1–63, 1989.

[38] Chaz Firestone and Brian Scholl. Seeing stability: Intuitive physics automatically guides selective attention. *Journal of Vision*, 16:689, 09 2016.

[39] François Fleuret, Tao Li, Charles Dubout, Emma K Wampler, Steven Yantis, and Donald Geman. Comparing machines and humans on a visual categorization test. *Proceedings of the National Academy of Sciences*, 108(43):17621–17625, 2011.

[40] Kenneth D Forbus, Dedre Gentner, Arthur B Markman, and Ronald W Ferguson. Analogy just looks like high-level perception: Why a domain-general approach to analogical mapping is right. *Journal of Experimental & Theoretical Artificial Intelligence*, 10(2):231–257, 1998.

[41] Steven M. Frankland, Taylor W. Webb, and Jonathan D. Cohen. No coincidence, george: Capacity-limits as the curse of compositionality. *PsyArXiv*, 2021.

[42] Shuhao Fu, Keith J Holyoak, and Hongjing Lu. From vision to reasoning: Probabilistic analogical mapping between 3d objects. In *Proceedings of the 44th Annual Meeting of the Cognitive Science Society*, 2022.

[43] Shuhao Fu, Andrew Jun Lee, Anna Wang, Ida Momennejad, Trevor Bihl, Hongjing Lu, and Taylor W. Webb. Evaluating compositional scene understanding in multimodal generative models. *Transactions on Machine Learning Research*, Apr 2025. in press.

[44] Andrew Geiger, Alexandra Carstensen, Michael C Frank, and Christopher Potts. Relational reasoning and generalization using nonsymbolic neural networks. *Psychological Review*, 2022.

[45] Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Partial success in closing the gap between human and machine vision. *Advances in Neural Information Processing Systems*, 34:23885–23899, 2021.

[46] Dedre Gentner. Children's performance on a spatial analogies task. *Child development*, pages 1034–1039, 1977.

[47] Dedre Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive science*, 7(2):155–170, 1983.

[48] Mary L Gick and Keith J Holyoak. Schema induction and analogical transfer. *Cognitive psychology*, 15(1):1–38, 1983.

[49] Steven Gold and Anand Rangarajan. A graduated assignment algorithm for graph matching. *IEEE Transactions on pattern analysis and machine intelligence*, 18(4):377–388, 1996.

[50] Robert L Goldstone. Similarity, interactive activation, and mapping. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20(1):3, 1994.

[51] Michelle R. Greene and Aude Oliva. The briefest of glances: the time course of natural scene understanding. *Psychological Science*, 20(4):464–472, 2009.

[52] Klaus Greff, Sjoerd Van Steenkiste, and Jürgen Schmidhuber. On the binding problem in artificial neural networks. *arXiv preprint arXiv:2012.05208*, 2020.

[53] Yulan Guo, Hanyun Wang, Qingyong Hu, Hao Liu, Li Liu, and Mohammed Bennamoun. Deep learning for 3d point clouds: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12):4338–4364, 2021.

[54] Alon Hafri, Michael F. Bonner, Barbara Landau, and Chaz Firestone. A phone in a basket looks like a knife in a cup: Role-filler independence in visual processing. *Open Mind*, 8:766–794, 06 2024.

[55] Alon Hafri and Chaz Firestone. The perception of relations. *Trends in Cognitive Sciences*, 25(6):475–492, 2021.

[56] Alon Hafri, Anna Papafragou, and John C. Trueswell. Getting the gist of events: recognition of two-participant actions from brief displays. *Journal of experimental psychology. General*, 142 3:880–905, 2013.

[57] Alon Hafri, John C. Trueswell, and Brent Strickland. Encoding of event roles from visual scenes is rapid, spontaneous, and interacts with higher-level visual processing. *Cognition*, 175:36–52, 2018.

[58] 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*, pages 770–778, 2016.

[59] Will Douglas Heaven. This avocado armchair could be the future of ai. *MIT Technology Review*, January 2021. Accessed: 2025-05-29.

[60] Felix Hill, Adam Santoro, David G T Barrett, Ari S Morcos, and Timothy P Lillicrap. Learning to make analogies by contrasting abstract relational structure. In *International Conference on Learning Representations*, 2019.

[61] Douglas R Hofstadter, Melanie Mitchell, et al. The copycat project: A model of mental fluidity and analogy-making. *Advances in connectionist and neural computation theory*, 2:205–267, 1995.

[62] Keith J Holyoak and Paul Thagard. Analogical mapping by constraint satisfaction. *Cognitive science*, 13(3):295–355, 1989.

[63] Jennifer Hu and Roger Levy. Prompting is not a substitute for probability measurements in large language models. *arXiv preprint arXiv:2305.13264*, 2023.

[64] John E. Hummel. Object recognition. In Daniel Reisberg, editor, *The Oxford Handbook of Cognitive Psychology*, pages 32–46. Oxford University Press, Oxford, England, 2013.

[65] John E Hummel and Keith J Holyoak. Distributed representations of structure: A theory of analogical access and mapping. *Psychological review*, 104(3):427, 1997.

[66] John E. Hummel and Brian J. Stankiewicz. An architecture for rapid, hierarchical structural description. In Toshio Inui and James L. McClelland, editors, *Attention and Performance XVI: Information Integration in Perception and Communication*, pages 93–121. MIT Press, Cambridge, MA, 1996.

[67] Nicholas Ichien, Qing Liu, Shuhao Fu, Keith J. Holyoak, Alan L. Yuille, and Hongjing Lu. Visual analogy: Deep learning versus compositional models. In *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*, pages 1799–1805, Austin, TX, 2021. Cognitive Science Society.

[68] Nicholas Ichien, Qing Liu, Shuhao Fu, Keith J. Holyoak, Alan L. Yuille, and Hongjing Lu. Two computational approaches to visual analogy: Task-specific models versus domain-general mapping. *Cognitive Science*, 47:e13347, 2023.

[69] Leyla Isik, Anna Mynick, Dimitrios Pantazis, and Nancy Kanwisher. The speed of human social interaction perception. *NeuroImage*, 215:116844, 2020.

[70] Huaizu Jiang, Xiaojian Ma, Weili Nie, Zhiding Yu, Yuke Zhu, and Anima Anandkumar. Bongard-hoi: Benchmarking few-shot visual reasoning for human-object interactions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19056–19065, 2022.

[71] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017.

[72] Daniel Kaiser, Genevieve L. Quek, Radoslaw M. Cichy, and Marius V. Peelen. Object vision in a structured world. *Trends in Cognitive Sciences*, 23(8):672–685, 2019.

[73] Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9161–9175, Singapore, December 2023. Association for Computational Linguistics.

[74] Carina Kauf, Emmanuele Chersoni, Alessandro Lenci, Evelina Fedorenko, and Anna A Ivanova. Comparing plausibility estimates in base and instruction-tuned large language models. *arXiv preprint arXiv:2403.14859*, 2024.

[75] Philip J Kellman. Perception of three-dimensional form by human infants. *Perception & Psychophysics*, 36:353–358, 1984.

[76] Giancarlo Kerg, Sarthak Mittal, David Rolnick, Yoshua Bengio, Blake Richards, and Guillaume Lajoie. On neural architecture inductive biases for relational tasks. *arXiv preprint arXiv:2206.05056*, 2022.

[77] Junkyung Kim, Matthew Ricci, and Thomas Serre. Not-so-clevr: learning same–different relations strains feedforward neural networks. *Interface focus*, 8(4):20180011, 2018.

[78] Thomas N. Kipf, Elise van der Pol, and Max Welling. Contrastive learning of structured world models. In *8th International Conference on Learning Representations, ICLR*, 2020.

[79] Talia Konkle and Aude Oliva. A real-world size organization of object responses in occipitotemporal cortex. *Neuron*, 74(6):1114–1124, 2012.

[80] Daniel C Krawczyk, Keith J Holyoak, and John E Hummel. Structural constraints and object similarity in analogical mapping and inference. *Thinking & reasoning*, 10(1):85–104, 2004.

[81] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

[82] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, pages 1097–1105, 2012.

[83] Brenden M Lake and Steven T Piantadosi. People infer recursive visual concepts from just a few examples. *Computational Brain & Behavior*, 3(1):54–65, 2020.

[84] Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253, 2017.

[85] Andrew K. Lampinen and James L. McClelland. Transforming task representations to perform novel tasks. *Proceedings of the National Academy of Sciences*, 117(52):32970–32981, 2020.

[86] Andrew Jun Lee, Hongjing Lu, and Keith J. Holyoak. Human relational concept learning on the synthetic visual reasoning test. In Micah Goldwater, Florencia K. Anggoro, Brett K. Hayes, and Desmond C. Ong, editors, *Proceedings of the 45th Annual Meeting of the Cognitive Science Society*, pages 799–805. Cognitive Science Society, 2023.

[87] Martha Lewis and Melanie Mitchell. Using counterfactual tasks to evaluate the generality of analogical reasoning in large language models. *arXiv preprint arXiv:2402.08955*, 2024.

[88] Martha Lewis, Nihal V Nayak, Peilin Yu, Qinan Yu, Jack Merullo, Stephen H Bach, and Ellie Pavlick. Does clip bind concepts? probing compositionality in large image models. *arXiv preprint arXiv:2212.10537*, 2022.

[89] Jing Liao, Yuan Yao, Lu Yuan, Gang Hua, and Sing Bing Kang. Visual attribute transfer through deep image analogy. *arXiv preprint arXiv:1705.01088*, 2017.

[90] Xin Lin, Changxing Ding, Yibing Zhan, Zijian Li, and Dacheng Tao. Hl-net: Heterophily learning network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19476–19485, 2022.

[91] Dingning Liu, Xiaomeng Dong, Renrui Zhang, Xu Luo, Peng Gao, Xiaoshui Huang, Yongshun Gong, and Zhihui Wang. 3daxiesprompts: Unleashing the 3d spatial task capabilities of gpt-4v. *arXiv preprint arXiv:2312.09738*, 2023.

[92] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. LLaVA-1.5: MLP connector + academic VQA data, achieves SOTA on 11 benchmarks.

[93] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors,

Advances in Neural Information Processing Systems, volume 36, pages 34892–34916. Curran Associates, Inc., 2023.

[94] Qing Liu, Adam Kortylewski, Zhishuai Zhang, Zizhang Li, Mengqi Guo, Qihao Liu, Xiaoding Yuan, Jiteng Mu, Weichao Qiu, and Alan Yuille. Learning part segmentation through unsupervised domain adaptation from synthetic vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3955–3964, 2022.

[95] Zili Liu. Ideal observers of visual object recognition. In *Theoretical Aspects of Neural Computation*, pages 145–154. Springer, 1998.

[96] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *Advances in neural information processing systems*, 33:11525–11538, 2020.

[97] Gordon D. Logan and Daniel D. Sadler. A computational analysis of the apprehension of spatial relations. In Paul Bloom, Mary A. Peterson, Lynn Nadel, and Merrill F. Garrett, editors, *Language and Space*, pages 493–529. The MIT Press, 1996.

[98] Andrew Lovett and Kenneth Forbus. Modeling visual problem solving as analogical reasoning. *Psychological Review*, 124(1):60–90, 2017.

[99] Andrew Lovett, Emmett Tomai, Kenneth Forbus, and Jeffrey Usher. Solving geometric analogy problems through two-stage analogical mapping. *Cognitive science*, 33(7):1192–1231, 2009.

[100] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European Conference on Computer Vision*, 2016.

[101] Hongjing Lu, Dawn Chen, and Keith J. Holyoak. Bayesian analogy with relational transformations. *Psychological Review*, 119(3):617–648, 2012.

[102] Hongjing Lu, Nicholas Ichien, and Keith J Holyoak. Probabilistic analogical mapping with semantic relation networks. *Psychological Review*, 2022.

[103] Hongjing Lu, Qi Liu, Nicholas Ichien, Alan L. Yuille, and Keith J. Holyoak. Seeing the meaning: Vision meets semantics in solving pictorial analogy problems. In *Proceedings of the 41st Annual Meeting of the Cognitive Science Society*, pages 2201–2207, 2019.

[104] Hongjing Lu, Ying Nian Wu, and Keith J. Holyoak. Emergence of analogy from relation learning. *Proceedings of the National Academy of Sciences*, 116(10):4176–4181, 2019.

[105] Gary Marcus, Ernest Davis, and Scott Aaronson. A very preliminary analysis of dall-e 2. *arXiv preprint arXiv:2204.13807*, 2022.

[106] David Marr. *Vision: A computational investigation into the human representation and processing of visual information.* MIT Press, 1982.

[107] Bryan J. Matlen, Dedre Gentner, and Steven L. Franconeri. Spatial alignment facilitates visual comparison. *Journal of Experimental Psychology: Human Perception and Performance*, 46(5):443–457, 2020.

[108] Nicola Messina, Giuseppe Amato, Fabio Carrara, Claudio Gennaro, and Fabrizio Falchi. Recurrent vision transformer for solving visual reasoning problems. *arXiv preprint arXiv:2111.14576*, 2021.

[109] Nicola Messina, Giuseppe Amato, Fabio Carrara, Claudio Gennaro, and Fabrizio Falchi. Solving the same-different task with convolutional neural networks. *Pattern Recognition Letters*, 143:75–80, 2021.

[110] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

[111] Melanie Mitchell. *Analogy-making as perception: A computer model*. Mit Press, 1993.

[112] Melanie Mitchell. Abstraction and analogy-making in artificial intelligence. *Annals of the New York Academy of Sciences*, 1505(1):79–101, 2021.

[113] Melanie Mitchell, Alessandro B Palmarini, and Arseny Moskvichev. Comparing humans, gpt-4, and gpt-4v on abstraction and reasoning tasks. *arXiv preprint arXiv:2311.09247*, 2023.

[114] Shanka Subhra Mondal, Jonathan D Cohen, and Taylor W Webb. Slot abstractors: Toward scalable abstract visual reasoning. *arXiv preprint arXiv:2403.03458*, 2024.

[115] Shanka Subhra Mondal, Taylor Webb, and Jonathan D Cohen. Learning to reason over visual objects. *arXiv preprint arXiv:2303.02260*, 2023.

[116] Scott O Murray, Bruno A Olshausen, and David L Woods. Processing shape, motion and three-dimensional shape-from-motion in the human cortex. *Cerebral Cortex*, 13(5):508–516, 2003.

[117] Sam Musker, Alex Duchnowski, Raphaël Millière, and Ellie Pavlick. Semantic structure-mapping in llm and human analogical reasoning. *arXiv preprint arXiv:2406.13803*, 2024.

[118] Massimiliano Patacchiola and Amos Storkey. Self-supervised relational reasoning for representation learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[119] Julia Peyre, Ivan Laptev, Cordelia Schmid, and Josef Sivic. Detecting unseen visual relations using analogies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1981–1990, 2019.

[120] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 652–660, 2017.

[121] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[122] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

[123] Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind. *arXiv preprint arXiv:2407.06581*, 2024.

[124] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[125] Sunayana Rane, Alexander Ku, Jason Baldridge, Ian Tenney, Tom Griffiths, and Been Kim. Can generative multimodal models count to ten? In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46, 2024.

[126] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12179–12188, 2021.

[127] J. C. Raven. *Progressive Matrices: A Perceptual Test of Intelligence*. H. K. Lewis, London, 1938.

[128] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics.

[129] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[130] Gaetano Rossiello, Alfio Gliozzo, Robert Farrell, Nicolas R. Fauceglia, and Michael R. Glass. Learning relational representations by analogy using hierarchical siamese networks. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3235–3245, 2019.

[131] Fereshteh Sadeghi, C. Lawrence Zitnick, and Ali Farhadi. Visalogy: Answering visual analogy questions. In *Advances in Neural Information Processing Systems*, volume 28, pages 1882–1890, 2015.

[132] Adam Santoro, David Raposo, David G. T. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. In *Advances in Neural Information Processing Systems*, volume 30, pages 4967–4976, 2017.

[133] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[134] Richard E. Snow, Patrick C. Kyllonen, and Barbara Marshalek. The topography of ability and learning correlations. In Robert J. Sternberg, editor, *Advances in the Psychology of Human Intelligence*, volume 2, pages 47–103. Erlbaum, 1984.

[135] Charles Spearman. *The nature of "intelligence" and the principles of cognition*. Macmillan, 1923.

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

[137] Yihong Sun, Adam Kortylewski, and Alan L. Yuille. Amodal segmentation through out-of-task and out-of-distribution generalization with a bayesian model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10287–10296, 2020.

[138] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H. S. Torr, and Timothy M. Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1199–1208, 2018.

[139] Anne M Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980.

[140] Stefan Treue, Richard A. Andersen, Hiroshi Ando, and Ellen C. Hildreth. Structure-from-motion: Perceptual evidence for surface interpolation. *Vision Research*, 35(1):139–148, 1995.

[141] Stefan Treue, Masud Husain, and Richard A. Andersen. Human perception of structure from motion. *Vision Research*, 31(1):59–75, 1991.

[142] Barbara Tversky and Kenneth Hemenway. Objects, parts, and categories. *Journal of Experimental Psychology: General*, 113(2):169–193, 1984.

[143] Mohit Vaishnav and Thomas Serre. Gamr: A guided attention model for (visual) reasoning. *arXiv preprint arXiv:2206.04928*, 2022.

[144] Johan Wagemans, James H. Elder, Michael Kubovy, Stephen E. Palmer, Mary A. Peterson, Manish Singh, and Rüdiger von der Heydt. A century of gestalt psychology in visual perception: I. perceptual grouping and figure-ground organization. *Psychological Bulletin*, 138(6):1172–1217, 2012.

[145] Hans Wallach and Daniel N. O'Connell. The kinetic depth effect. *Journal of Experimental Psychology*, 45(4):205, 1953.

[146] Jianyu Wang, Zhishuai Zhang, Cihang Xie, Yuyin Zhou, Vittal Premachandran, Jun Zhu, Lingxi Xie, and Alan Yuille. Visual concepts and compositional voting. *Annals of Mathematical Sciences and Applications*, 3:151–188, 2018.

[147] Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Yixuan Li, and Neel Joshi. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. In *Advances in Neural Information Processing Systems (NeurIPS 2024)*, 2024.

[148] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[149] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (TOG)*, 38(5):1–12, 2019.

[150] Taylor Webb, Keith J Holyoak, and Hongjing Lu. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541, 2023.

[151] Taylor Webb, Keith J Holyoak, and Hongjing Lu. Evidence from counterfactual tasks supports emergent analogical reasoning in large language models. *arXiv preprint arXiv:2404.13070*, 2024.

[152] Taylor Webb, Shanka Subhra Mondal, and Jonathan D Cohen. Systematic visual reasoning through object-centric relational abstraction. *Advances in Neural Information Processing Systems*, 36, 2024.

[153] Taylor W Webb, Zachary Dulberg, Steven Frankland, Alexander Petrov, Randall O'Reilly, and Jonathan D Cohen. Learning representations that support extrapolation. In *International conference on machine learning*, pages 10136–10146. PMLR, 2020.

[154] Taylor W. Webb, Shuhao Fu, Trevor Bihl, Keith J. Holyoak, and Hongjing Lu. Zero-shot visual reasoning through probabilistic analogical mapping. *Nature Communications*, 14(1):5144, 2023.

[155] Taylor Whittington Webb, Ishan Sinha, and Jonathan D. Cohen. Emergent symbols through binding in external memory. In *9th International Conference on Learning Representations, ICLR*, 2021.

[156] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[157] Yang Wu, Shilong Wang, Hao Yang, Tian Zheng, Hongbo Zhang, Yanyan Zhao, and Bing Qin. An early evaluation of gpt-4v (ision). *arXiv preprint arXiv:2310.16534*, 2023.

[158] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1912–1920, 2015.

[159] Zhilin Yang, Jake Zhao, Bhuwan Dhingra, Kaiming He, William W. Cohen, Ruslan Salakhutdinov, and Yann LeCun. Glomo: Unsupervisedly learned relational graphs as transferable representations. *arXiv preprint arXiv:1806.05662*, 2018. NeurIPS workshop paper; also Code: kimiyoung/glomo.

[160] Meagan Yee, Susan S Jones, and Linda B Smith. Changes in visual object recognition precede the shape bias in early noun learning. *Frontiers in Psychology*, 3:533, 2012.

[161] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *Advances in neural information processing systems*, 31, 2018.

[162] Eunice Yiu, Maan Qraitem, Charlie Wong, Anisa Noor Majhi, Yutong Bai, Shiry Ginosar, Alison Gopnik, and Kate Saenko. Kiva: Kid-inspired visual analogies for testing large multimodal models. *arXiv preprint arXiv:2407.17773*, 2024.

[163] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Conference on Computer Vision and Pattern Recognition*, 2018.

[164] Daniel Zeng, Tailin Wu, and Jure Leskovec. Virel: Unsupervised visual relations discovery with graph-level analogy. In *Beyond Bayes: Paths Towards Universal Reasoning Systems Workshop, ICML 2022*, 2022. arXiv preprint arXiv:2207.00590.

[165] Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. Raven: A dataset for relational and analogical visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5317–5327, 2019.

[166] Chi Zhang, Baoxiong Jia, Feng Gao, Yixin Zhu, Hongjing Lu, and Song-Chun Zhu. Learning perceptual inference by contrasting. In *Advances in Neural Information Processing Systems*, volume 32, pages 1075–1087, 2019.

[167] Ji Zhang, Kevin J. Shih, Ahmed Elgammal, Andrew Tao, and Bryan Catanzaro. Graphical contrastive losses for scene graph parsing. In *CVPR*, 2019.

[168] Yong Zhang, Yingwei Pan, Ting Yao, Rui Huang, Tao Mei, and Chang-Wen Chen. Learning to generate language-supervised and open-vocabulary scene graph using pre-trained visual-semantic space. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2915–2924, 2023.

[169] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip H S Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16259–16268, 2021.

[170] Li Zhaoping and Nathalie Guyader. Interference with bottom-up feature detection by higher-level object recognition. *Current Biology*, 17(1):26–31, 2007.

[171] Yinyuan Zheng, Bryan J. Matlen, and Dedre Gentner. Spatial alignment facilitates visual comparison in children. *Cognitive Science*, 46(8):e13182, 2022.

[172] Alisa Zhila, Wen-tau Yih, Christopher Meek, Geoffrey Zweig, and Tomas Mikolov. Combining heterogeneous models for measuring relational similarity. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1000–1009, 2013.

[173] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan L. Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. In *10th International Conference on Learning Representations, ICLR*, 2022.

[174] Hongru Zhu, Peng Tang, Jeongho Park, Soojin Park, and Alan Yuille. Robustness of object recognition under extreme occlusion in humans and computational models. In *Proceedings of the 41st Annual Meeting of the Cognitive Science Society*, pages 3213–3219, 2019.