



The psychophysics of compositionality: Relational scene perception occurs in a canonical order

Zekun Sun ^{a,b}, Chaz Firestone ^b, Alon Hafri ^{b,c} , ^{*}

^a Department of Psychology, Yale University, 100 College St., New Haven, CT, 06510, USA

^b Department of Psychological & Brain Sciences, Johns Hopkins University, 3400 N Charles St., Baltimore, MD 21218, USA

^c Department of Linguistics and Cognitive Science, University of Delaware, 105 The Green, Newark, DE 19716, USA

ARTICLE INFO

Dataset link: <https://osf.io/vzxdg>

Keywords:

Relations
Scene perception
Intuitive physics
Anchor objects
Scene grammar
Language of thought

ABSTRACT

We see not only objects and their features (e.g., glass vases or wooden tables) but also relations between them (e.g., a vase on a table). An emerging view accounts for such relational representations by positing that visual perception is compositional: Much like language, where words combine to form phrases and sentences, many visual representations contain discrete constituents that combine systematically. This perspective raises a fundamental question: What principles guide the composition of relational representations, and how are they built over time? Here, we tested the hypothesis that the mind constructs relational representations in a canonical order. Inspired by a distinction from cognitive linguistics, we predicted that ‘reference’ objects (typically large, stable, and able to physically control other objects; e.g., tables) take precedence over ‘figure’ objects (e.g., vases) during scene composition. In Experiment 1, participants who arranged items to match linguistic descriptions (e.g., “The vase is on the table”, “The table is supporting the vase”) consistently placed reference objects first (e.g., table, then vase). Experiments 2–5 extended these findings to visual recognition itself: participants were faster to verify scene descriptions when reference objects appeared before figure objects in a scene, rather than vice versa. This Reference-first advantage emerged rapidly (within 100 ms), persisted in a purely visual task, and reflected abstract principles (e.g., physical forces) beyond simple differences in size or shape. Our findings reveal psychophysical principles underlying compositionality in visual processing: the mind builds representations of object relations sequentially, guided by the objects’ roles in those relations.

1. Introduction

Look at the image in Fig. 1A. What do you see? Certainly you see colors, textures, edges, and countless other visual features—the deep green of a plant, the glossiness of a vase, the grain of a wooden table, and the wicker of a square basket. However, beyond these properties, you may also appreciate something about how the objects *relate* to one another: The plant is sitting *on* the table and the basket is resting *below* it (Fig. 1B). Relational representations are a core topic of study in many domains of higher-level cognition, such as analogical reasoning (Gattis, 2004; Goldwater & Gentner, 2015; Jamrozik & Gentner, 2015; Webb, Fu, Bihl, Holyoak, & Lu, 2023), linguistic reference (Johannes, Wilson, & Landau, 2016; Landau & Jackendoff, 1993; Levinson, 2003; Talmy, 1983; Webb, Holyoak, & Lu, 2023), and causal ascription (Gerstenberg, Peterson, Goodman, Lagnado, & Tenenbaum, 2017; Kominsky et al., 2017; Wolff & Song, 2003).

^{*} Corresponding author at: Department of Linguistics and Cognitive Science, University of Delaware, 105 The Green, Newark, DE 19716, USA.

E-mail addresses: zekun.sun@yale.edu (Z. Sun), chaz@jhu.edu (C. Firestone), alon@udel.edu (A. Hafri).

<https://doi.org/10.1016/j.cogpsych.2025.101765>

Received 22 April 2025; Received in revised form 9 September 2025; Accepted 18 September 2025

Available online 11 October 2025

0010-0285/© 2025 Elsevier Inc. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

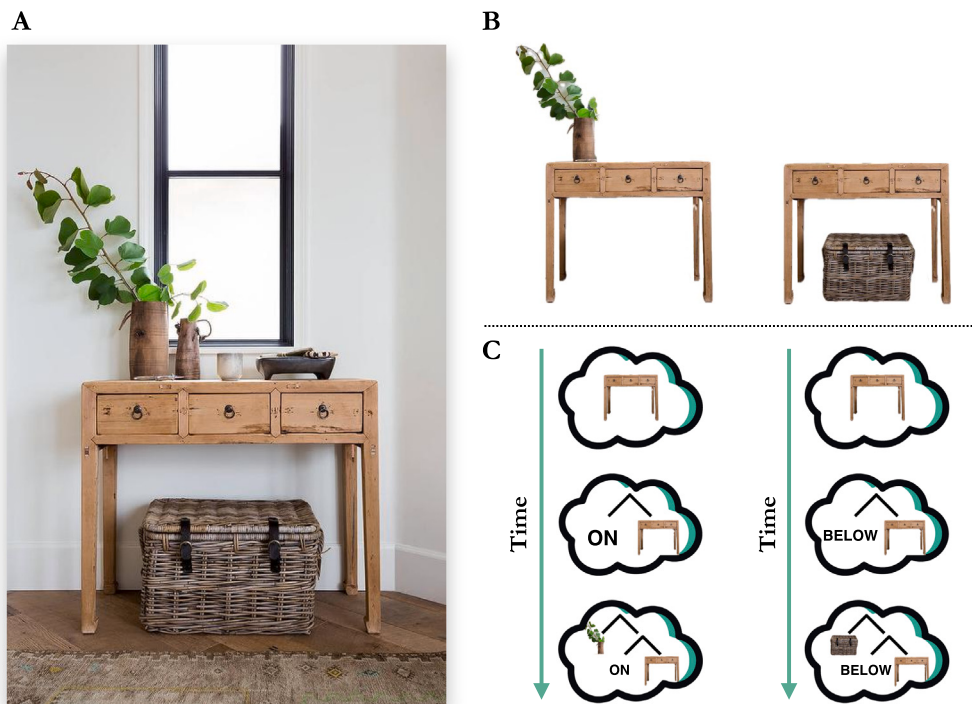


Fig. 1. The world contains not only objects and features, but also relations holding between them. (A) We see the plant and its greenness, the table and its size, and so on. But we also appreciate that the plant is sitting *on* the table, and the basket is resting *below* it. (B) A growing literature suggests that visual perception rapidly and spontaneously forms representations of such relations. How does it do so? (C) Despite the fact that the objects are presented to the eyes simultaneously, here we explore the possibility that the mind adopts a sequential order to combine discrete elements into relational representations, with reference objects (e.g., the table) serving as the primary element or ‘scaffold’.

However, a growing body of empirical evidence suggests that more basic processes of visual perception also encode sophisticated relations such as those depicted in Fig. 1B (Hafri, Bonner, Landau, & Firestone, 2024; Hafri & Firestone, 2021; Hafri, Trueswell, & Strickland, 2018; Lovett & Franconeri, 2017). Moreover, this evidence suggests that such relations are not merely detected by visual processing but also represented by the visual system in ways that are structured and systematic, as though such representations have parts that combine into wholes; in other words, they are represented *compositionally* (Hafri, Green, & Firestone, 2023). On this scheme, for example, when seeing a plant on a table, the mind represents the scene not as an undifferentiated collection of pixels or textures, but rather in terms of the discrete constituents *plant*, *table*, and even the relation *ON*.

This perspective raises a question: How does the mind join these parts together in forming sophisticated relational representations? Here, we explore the nature of this visual compositional process, using as a case study the sort of relational representations depicted in Fig. 1A.

1.1. Compositionality in visual perception

Compositionality refers to a form of representation in which complex representations are systematically constructed by combining their constituent parts (Fodor & Pylyshyn, 1988). In this sense, it is not particularly controversial to suggest that compositionality can be observed in some forms of basic visual representation. For example, recent theoretical work has examined how principles of compositionality apply to visual representations of the bounding contours of objects, suggesting that perceiving contours depends on different modes of composition, such as combining features into fragments and fragments into contours (Lande, 2023). Compositionality is also a fundamental principle in influential theories of object recognition. For example, the Recognition-by-Components theory proposes that objects are represented by freely combining a set of basic components (‘geons’) (Biederman, 1987) into more sophisticated structures, and more recent accounts describe shapes via their parts’ intrinsic axes and connections (their ‘skeletons’), often in a hierarchical tree format (Feldman & Singh, 2006). Ample empirical evidence supports the psychological reality of such compositional representations (Ayzenberg & Lourenco, 2022; Bonnen, Wagner, & Yamins, 2025; Firestone & Scholl, 2014; Lescroart & Biederman, 2013; Lewis & Frank, 2016; Lowet, Firestone, & Scholl, 2018; Sun & Firestone, 2021; Van Tonder, Lyons, & Ejima, 2002; Wilder, Feldman, & Singh, 2011).

While at first compositionality might seem to apply only to the contours and parts *within* objects, more recent work has made the case that this form of representation extends to relations that hold *between* objects. Cavanagh (2021) has proposed a ‘language of

vision,’ whereby visual processing separates images into language-like components—‘visual nouns’ (objects), ‘visual verbs’ (actions), and ‘visual prepositions’ (spatial relations)—and structurally combines them into ‘sentences’ (descriptions). In this framework, such ‘visual sentences’ embody the principle of compositionality in vision, capturing structured relations among discrete components.

A substantial body of recent empirical work supports this idea (for a review, see Hafri & Firestone, 2021). For example, observers performing rapid target-recognition tasks ‘confuse’ scenes that contain different objects but share the same relation (e.g., mistaking a phone in a basket for a knife in a cup), suggesting that certain relations are encoded abstractly, independent of the particular objects involved (Hafri et al., 2024; Vettori, Hochmann, & Papeo, 2024; see also Lovett & Franconeri, 2017). Moreover, visual processing binds entities to specific relational roles, such that changes in role assignment (e.g., a boy as Agent in one trial and Patient in the next) slow responses in a simple visual task (e.g., report the location of the target individual) even when these roles are irrelevant to the task (Hafri, Papafragou, & Trueswell, 2013; Hafri et al., 2018; Vettori, Odin, Hochmann, & Papeo, 2025).

The above work suggests that visual processing represents relations in ways that preserve the identities of both the entities themselves and the relations in which they participate. This representational scheme is often called *role-filler independence* (Quilty-Dunn, Porot, & Mandelbaum, 2023), and it is, in part, what makes such representations compositional. The existence of visual relational representations with this property suggests that some aspects of perception may exhibit core properties of a ‘Language of Thought (LoT)’ (Fodor, 1975; Quilty-Dunn et al., 2023), a format of representation that can readily accommodate compositionality in ways that other formats more traditionally associated with visual perception may not (i.e., iconic or ‘picture-like’ formats; Block, 2023; Burge, 2022; Carey, 2009; Kosslyn, Thompson, & Ganis, 2006).

1.2. What constraints govern the process of visual composition?

While compositionality is traditionally discussed in terms of representational *format*, the existence of LoT-like representations in visual perception raises an equally important question about the compositional *process*: How are such representations built by the mind from their constituent parts?¹ In other domains such as speech processing, structured representations are constructed incrementally, as the speech signal is dynamic and temporally extended, unfolding over time (Christiansen & Chater, 2016). By contrast, in visual processing, relational content (objects and their visual features) is in principle immediately available from an image. Despite this, might visual relational representations also be ‘built’ sequentially by the mind (Fig. 1C)?

Classic research in visual cognition offers clues about the dynamic nature of relational processing in vision. Ullman (1987) proposed the concept of *visual routines*—sequences of spatial operations executed to extract simple relations, such as whether one object is *inside*, *on*, or *collinear* with another. A well-known example is *curve-tracing*, in which the visual system systematically follows a curve’s contour in order to judge whether two points lie on the same curve (Jolicoeur, Ullman, & Mackay, 1986, 1991). These routines occur dynamically in both space and time, and typically require effort and intentional initiation (though see Wong & Scholl, 2024). However, this work primarily examines simple geometric features such as points, lines, and curves. It remains unclear how—or whether—such routines extend to more sophisticated relations between real-world objects (e.g., those shown in Fig. 1A). These relations often involve abstract roles or physical forces (e.g., Support, Containment), and thus cannot be traced or followed in any literal sense. Instead, they may require a different kind of sequential process—one that incrementally combines objects into a structured mental representation. If so, what principles govern this compositional process?

Insights from event cognition suggest one candidate principle: the *roles* of participants (often called ‘thematic roles’ in linguistics). Agents, relative to Patients, enjoy consistent processing advantages: they are prioritized in visual search, recognition, and attention (Segalowitz, 1982); allow for better predictions about upcoming events (Cohn & Paczynski, 2013); facilitate the processing of actions (Cohn & Paczynski, 2013); and elicit stronger neural responses (Cohn, Paczynski, & Kutas, 2017). The primacy of Agents is even more evident in psycholinguistics: sentences are processed more easily when the Agent appears first (e.g., *the dog bit the man* is easier than its passive counterpart, *the man was bitten by the dog*; Ferreira, 2003). On one hand, it might be that this processing asymmetry is specific to Agent-Patient events, which would not be surprising, as the Agent’s role naturally lends itself to priority in perception and language. On the other hand, the ease of processing Agent-first order might reflect a more general mechanism in relational processing that applies across eventive, spatial, and physical relations. How, then, are other types of visual relations processed?

Here, we extend research on Agent-Patient events into more fundamental forms of relational processing, by testing the hypothesis that the mind builds spatial and physical relations sequentially, according to the roles of the participating objects. In spatial or physical relations such as *ON* or *BELOW*, these roles are known as ‘Reference’ (sometimes called ‘Ground’) and ‘Figure’. Reference objects are typically large, stable, and/or physically ‘control’ other objects, while figure objects are typically small and/or mobile. According to a tradition known as cognitive linguistics (Miller & Johnson-Laird, 1976; Talmy, 1975), the reference object establishes the spatial or physical reference frame relative to which the figure object is located (see also Gleitman, Gleitman, Miller, & Ostrin, 1996; Landau & Jackendoff, 1993).

Of note, the non-linguistic construal of such entities is systematically reflected in linguistic structure. Reference objects tend to occupy lower positions in syntax than figure objects (i.e., they are structurally beneath figure objects in syntax trees; Landau & Gleitman, 2015). For example, in ‘The bike is to the left of the garage’, the grammatical subject (bike) is Figure, and the grammatical

¹ A related line of work examines how knowledge of typical object-scene configurations (e.g., pots appear on stoves in kitchens; Kaiser, Quek, Cichy, & Peelen, 2019)—sometimes described as a kind of *scene grammar* (Draschlow & Vö, 2017; Vö, 2021)—supports visual processing. That work, however, concerns how prior knowledge shapes recognition within familiar contexts. Our focus is different: we aim to identify the more basic principles, or “syntax”, that govern the composition of relational representations itself, even in general, non-specific cases. See the General Discussion (Section 7.2) for a fuller contrast and implications.

object (garage) is Reference. Although figure objects often appear in subject position and thus (in English) first in a sentence, reference objects are *cognitively* primary, since they provide the spatial or physical frame against which figures are interpreted. This becomes clear when syntactic positions are reversed: “The garage is next to the bike” sounds odd, and may even encourage reinterpretation of the entities according to the roles canonically associated with their syntactic positions (as if, e.g., a small garage on wheels were circling around a giant bike statue; Gleitman et al., 1996). Thus syntactic structure not only reflects, but can, in some cases, actively shape relational interpretation. (As we show later, this linguistic influence may also play a role in scene composition when visual properties alone do not provide sufficient cues.)

Similar asymmetries appear in visual processing. Visual scene perception takes advantage of the unique role of ‘anchor’ objects to guide search and recognition (Vö, Boettcher, & Draschkow, 2019). For example, Boettcher, Draschkow, Dienhart, and Vö (2018) showed that in cluttered environments (e.g., a classroom), observers were slower to find a target object (e.g., an eraser) when the anchor object (e.g., a chalkboard) was replaced with an irrelevant one (e.g., a map). Studies with simpler geometric stimuli also suggest that figure and reference objects play different roles in the temporal construction of relations. Perceiving certain spatial relations requires serial processing (Holcombe, Linares, & Vaziri-Pashkam, 2011), and previewing the location of one object in a spatial relation can influence the speed with which people answer prompts such as “Is the red disc above the blue disc?” (Roth & Franconeri, 2012).

Developmental research points in the same direction. Huttenlocher, and Straus (1968) found that when children acted out a relational scene with real-world objects according to a statement (e.g., “The green block is on the top of the pink block”), they made fewer errors when they placed the figure object (green block) relative to a fixed reference object (pink block) than when the order was reversed. In similar work in which children reconstructed scenes according to statements such as “The red truck is pushing the blue truck” and “The blue truck is pulled by the red truck”, they positioned objects more quickly when placing the Agent (the red truck) with respect to the Patient (the blue truck), regardless of which entity served as the grammatical subject (Huttenlocher, Eisenberg, & Strauss, 1968).

Although the above work is suggestive, the hypothesis that relational perception follows a canonical order has not yet been tested in psychophysical experiments. It remains unclear whether the visual system builds relations in an order that respects the objects’ roles.

1.3. The present experiments: How to build a scene

Our work takes inspiration from the cognitive linguistics literature in hypothesizing that reference objects (e.g., tables, shelves, etc.) rather than figure objects (e.g., vases, laptops, etc.) serve as the scaffold for relational representations in visual perception (Gleitman et al., 1996; Landau & Jackendoff, 1993; Talmy, 1975). To test this hypothesis, we asked whether participants would construct relational scenes following a ‘Reference-first’ order, and whether participants’ visual processing of relational scenes would be facilitated when they have visual access to the reference object before the figure object (as opposed to vice versa).

We created relational scenes from various household objects (e.g., laptop, desk, lamp, nightstand), encompassing physical relations (e.g., desk supporting laptop) and spatial relations (e.g., laptop below desk). In Experiment 1, participants read linguistic descriptions and arranged objects on-screen to match the described relation, replicating previous findings of a Reference-first advantage in manual construction tasks (Huttenlocher, Eisenberg, & Strauss, 1968; Huttenlocher, & Straus, 1968). Experiments 2–5 used a recognition paradigm in which participants matched visual scenes to linguistic or pictorial probes viewed just beforehand. This paradigm builds on the classic sentence–picture verification task, which has traditionally been used to study how people determine whether a linguistic statement accurately describes a subsequently presented picture (Clark & Chase, 1972, 1974). Here we introduced a crucial modification to the task: instead of presenting the full relational scene at once (as is standard in this paradigm), we presented the objects asynchronously, either with the reference or figure object appearing first.

To foreshadow the key results, we found a Reference-object advantage across all five experiments: Participants employed a reference-first order in composing relational scenes, and they were faster to recognize a visual scene when the reference object appeared before the figure object rather than vice-versa. We further found that when visual differences between objects were eliminated, the linguistic structure of the probe sentence influenced the order of composition. Taken together, these results suggest that even though visual scene information is in principle available to the observer all at once, the mind composes relational representations sequentially, in ways that respect the role of each element in the relation.

2. Experiment 1 – Manual construction of relational scenes

How do people construct relational representations? Experiment 1 took a literal approach to this question by asking whether participants have a preference for the order in which they place constituent objects to compose relational scenes. Our approach was inspired by early studies of sentence comprehension conducted by Huttenlocher and colleagues, in which children placed one object with respect to a fixed object according to a statement (Huttenlocher, Eisenberg, & Strauss, 1968; Huttenlocher, & Straus, 1968). Children had an easier time doing so when the reference object was fixed in place first. A more recent object-placement study found a similar pattern in adults: when constructing rooms in a virtual environment (e.g., a kitchen or bathroom), participants tended to place “global” objects (reference objects such as a desk, bed, or stove) before “local” objects (figure objects such as a computer, pillow, or pot) (Draschkow & Vö, 2017).

The current study aimed to extend these previous findings by employing a similar paradigm, but with key differences: While the previous studies on sentence comprehension in children fixed an object in place and asked them to position another object

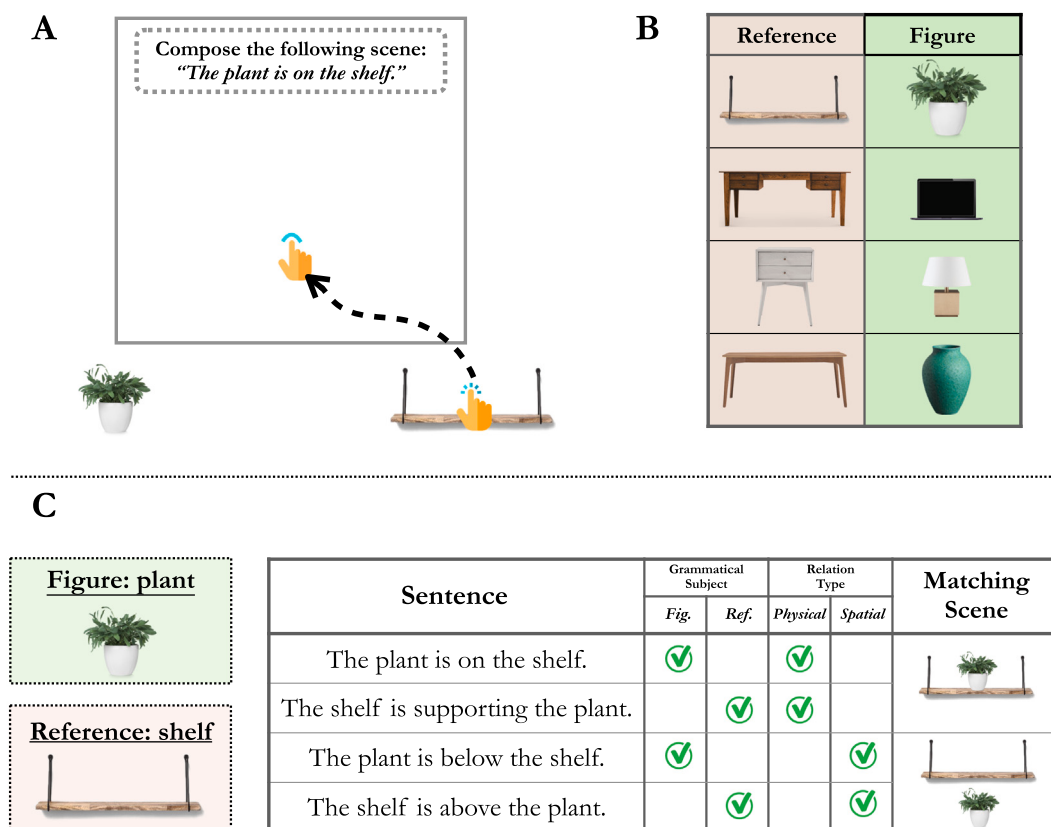


Fig. 2. (A) Illustration of the manual-composition task used in Experiment 1. On each trial, participants were asked to compose a scene by moving two objects into the workspace, according to a sentence that described a relational scene. (B) Stimuli used in Experiments 1–4. There were four pairs of objects, with one reference object and one figure object in each pair. (C) Descriptive sentences of visual scenes, using the plant/shelf pair as an example. The relational scenes were divided into two types: physical and spatial. Each scene was described in two ways: either the figure object or the reference object was the grammatical subject of the descriptive sentence, and thus was the first mentioned entity in the sentence. Fig. = figure, Ref. = reference.

relative to it (Huttenlocher, Eisenberg, & Strauss, 1968; Huttenlocher, & Straus, 1968), our task allowed adult participants to place objects in whatever order they wanted. On each trial, participants were asked to compose a scene to match a pre-specified linguistic description by dragging objects into a framed workspace. While they could do so in whatever order they liked, we predicted that they would move the reference object first, even without external pressure to do so.

2.1. Method

2.1.1. Open science practices

An archive of the data, code, stimuli, preregistrations, and other relevant materials is available at: <https://osf.io/vzxdg/>.

For each experiment, we preregistered the sample size, experimental design, and analyses (including exclusion criteria and some secondary analyses). Demos of the experiments can be viewed at <https://palresearch.org/buildingrelations>, so readers can experience them as participants did.

2.1.2. Participants

We recruited 40 participants for this experiment from the online platform Prolific (<https://www.prolific.com/>). (For a discussion of this participant pool's reliability, see Peer, Brandimarte, Samat, & Acquisti, 2017.) This sample size was determined based on a smaller pilot study. Participants were prescreened for a minimum approval rate of 85%, at least 50 prior submissions, normal or corrected-to-normal vision, native English proficiency (self-reported), and U.S. nationality. Sample sizes were preregistered for this and all other experiments. All studies were approved by the Johns Hopkins University Institutional Review Board.

2.1.3. Stimuli

Eight colored images were used in the experiment, grouped into four pairs: vase/table, laptop/desk, lamp/nightstand, and plant/shelf (Fig. 2B). Each pair consisted of a reference object that was relatively large and stable (i.e., table, desk, nightstand, or shelf) and a figure object that was relatively small and mobile (i.e., vase, laptop, lamp, or plant).

Participants were provided with sentences that described a relation between the two objects in each pair. These sentences varied along two dimensions: (1) whether the depicted relation was physical or (merely) spatial (e.g., “the plant is on the shelf” describes a physical relation, and “the plant is below the shelf” describes a spatial relation); and (2) whether the reference object or the figure object was the grammatical subject of the sentence or not (e.g., in “the plant is on the shelf”, the figure object, *plant*, is the subject, while in “the shelf is supporting the plant”, the reference object, *shelf*, is the subject). See Fig. 2C for a complete list of sentences for one of the object pairs.

Images ranged in size from 120×68 pixels to 417×186 pixels and were presented in the participant's Web browser. The workspace was presented at 600×600 pixels, with a white background. Because of the nature of online studies, we could not know the exact viewing distance, screen size, and luminance (etc.) of these stimuli as they appeared to participants. However, any distortions introduced by a given participant's viewing distance or monitor settings would have been equated across all stimuli and conditions.

2.1.4. Procedure

The experimental task is depicted in Fig. 2A. On each trial, participants first read a statement describing a relation between two objects, and then they clicked a button to indicate they were ready to compose the scene. Immediately after the click, the two mentioned objects appeared beneath the workspace (one on the right and the other on the left). Participants dragged each object into the workspace in whatever way they chose to compose a scene that correctly reflected the statement. Once both objects were inside the workspace, participants were able to click on a button to proceed to the next trial.

Overall, the experiment consisted of 32 trials. The four object pairs were combined with the four types of sentences, which made 16 unique trials. Each combination of object pair and sentence appeared twice: once with the figure object on the left side (below the workspace), once with the figure object on the right side. Trial order was randomized across participants.

2.1.5. Exclusions

As specified in the preregistration, we planned to exclude trials where the composed visual scenes did not accurately match the linguistic descriptions. To do so, we preregistered scene-specific boundaries within which each object should be placed in the particular scene in order to be considered accurate (with these boundaries detailed in the preregistration). We also planned to exclude any participant who had low overall accuracy ($<90\%$), lacked at least one trial in each combination of the key factors (Relation Type, Sentence Structure, and Object Side), or failed to provide a complete dataset.

2.2. Results

One participant was excluded for failing to submit a complete dataset. As expected, participants had little difficulty completing the task, with a mean accuracy of 98.7% in composing the visual scenes to match the linguistic descriptions given.

Crucially, participants overwhelmingly placed the reference object first (e.g., shelf before plant), doing so on 96.7% of trials across participants, $t(38) = 48.13$, $p < 2.2 \times 10^{-16}$ (98.1% for physical relations and 95.2% for spatial relations). This pattern held regardless of sentence order (i.e., whether the reference object was the grammatical subject or object of the sentence), object identity (all four pairs of images), and relation type (physical or spatial). Moreover, this Reference-first preference emerged at the very beginning of the task: Even in the very first trial of the experiment, a majority of participants (34 out of 39) moved the reference object before moving the figure object (binomial test, $p = 2.43 \times 10^{-16}$). These effects held even when including all trials (including those coded as incorrect, $t(38) = 48.66$, $p < 2.2 \times 10^{-16}$), and they generalized across the different object pairs (t -test across object-pair means, $t(38) = 35.90$, $p < 2.2 \times 10^{-16}$). We also conducted a repeated-measures ANOVA across participant means of Reference-first proportion, with the factors of interest as Relation, Sentence Structure, and Object Side, and found that these factors did not significantly modulate the Reference-first effect ($F(1, 38) < 3.65$, $p_s > .064$).

Overall, these results provide initial evidence that the mind applies a canonical routine in constructing relational scenes. They also raise the possibility that the mind adopts this Reference-first routine in representing visual relations more broadly, including in visual recognition. We explore this possibility in the remaining experiments of this paper.

3. Experiment 2 – Visual recognition of relational scenes

Experiment 1 revealed a predominant order in relational scene ‘production’: participants preferred to move the reference object into the scene first and position the figure object relative to it. Does this pattern extend beyond mere preferences and drive visual processing itself? Experiment 2 asked whether relational representations are built according to a similar compositional routine in visual recognition.

To do so, we used a variant of the ‘sentence-picture verification task’ initially developed by Clark and Chase (1972, 1974) to investigate how people decide whether or not a linguistic statement accurately describes a picture. In such a task, participants read a sentence (e.g., “The star is above the plus sign”) and are asked to verify whether it is a correct description of an image that

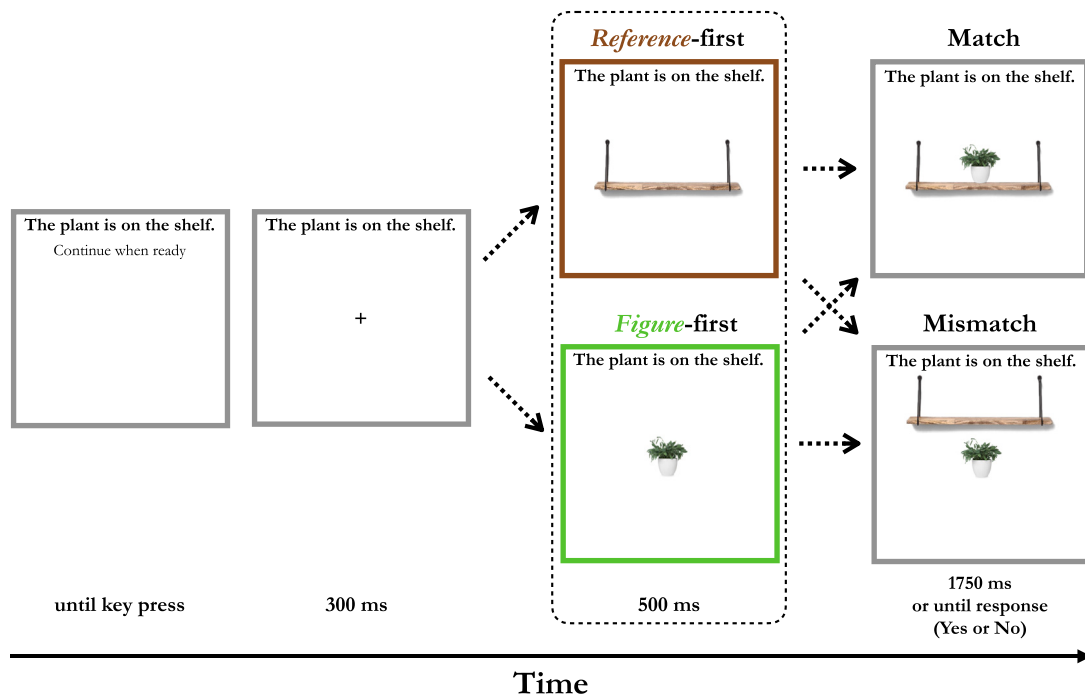


Fig. 3. Illustration of the visual recognition task. At the beginning of each trial, participants were given a sentence describing a relational scene. Once they pressed a key to proceed, a fixation cross appeared in the center of the frame for 300 ms, and then either the reference or figure object appeared at the center. After 500 ms, this was followed by the other object, which completed the scene. (On some trials, the objects appeared simultaneously.) Participants indicated whether the resultant scene matched or mismatched the sentence description as rapidly and accurately as possible, before the trial timed out. (For simplicity, the 500 ms blank display before fixation is omitted in this figure.).

appears soon after. We made one crucial change to the standard paradigm: Instead of showing the image all at once, sometimes the reference object (e.g., the table) appeared a half-second before the figure object (e.g., the laptop), or vice versa.

Manipulating this display order allowed us to ask whether there is a privileged order for visual recognition. We reasoned that if the visual system builds relational representations sequentially, following the Reference-first order pattern observed in Experiment 1, then participants would be faster to verify the target sentence when they saw the reference object right before the figure object, because this presentation order would ‘match’ the order in which the mind constructs such representations.

3.1. Method

3.1.1. Participants

As per our preregistration, 40 participants were recruited through Prolific. This sample size was determined by a power analysis on a small pilot study that produced similar results and indicated a 99% probability of detecting the effect of interest. Participants in this and subsequent experiments had to pass the same pre-screening criteria as in Experiment 1, and they could participate only if they had not previously completed a related experiment in this series.

3.1.2. Design and procedure

Participants were instructed that on each trial, they would read a sentence and then have to verify whether a subsequently presented visual scene matched the sentence. Fig. 3 illustrates the trial sequence. All sentences and scenes were presented on a white background within a 600 × 600 pixel frame. At the beginning of each trial, participants read a sentence that described a relational scene (e.g., “The laptop is on the table”), presented at the top of the frame. After reading, they pressed the space bar to indicate that they were ready to continue (but had to remain on the sentence screen for at least 500 ms). Once participants pressed the key, a blank screen appeared for 500 ms, followed by a fixation cross at the center of the scene for 300 ms. Then the to-be-verified scene was presented. The sentence remained at the top of the image frame throughout the trial.

Each visual scene was presented in one of three object-order conditions: (1) Reference-first, in which the reference object was presented first and then the figure object appeared 500 ms later; (2) Figure-first, in which the figure object was presented first, followed by the reference object 500 ms later; and (3) Simultaneous, in which both reference and figure objects appeared at the same time, right after the fixation cross disappeared. After the full scene was displayed, participants judged whether the complete scene matched the pre-specified sentential description as fast as possible without sacrificing accuracy, by pressing either Y for a match or N for a mismatch.

In mismatch trials, the visual scene involved the objects described in the linguistic probe, but the relational category was changed (e.g., if the sentence stated, “The plant is *on* the shelf”, the visual scene depicted the plant *below* the shelf). In other words, participants always saw the right pair of objects in the visual scene, but sometimes the relation was incorrect. Thus, they needed to evaluate the relation in both the sentence and the scene to make a correct response. Trials timed out if no response was given within 1750 ms. Of note, in the two 500 ms delay conditions (Reference-first and Figure-first), the first presented object always appeared at central fixation, so its location was not predictive of the correct response. (In the Simultaneous condition, the center object could be either Figure or Reference, as described below.)

All four pairs of objects and 16 sentences used in Experiment 1 (Fig. 2) were also used in this experiment. Several factors were fully crossed within participants: (a) Relation Type (spatial or physical), (b) Second Image Delay (0 ms or 500 ms), (c) Sentence Structure (Figure-as-subject or Reference-as-subject), (d) Trial Type (match or mismatch), and (e) Center Object Type (either the reference or figure object), yielding 128 test trials ($2 \times 2 \times 2 \times 2 \times 2 \times 4$ object pairs).

The key factor of interest was the presentation order of objects, determined by the combination of the factors Second Image Delay and Center Object. In particular, at the 0 ms second image delay (i.e., the Simultaneous condition), the object presented centrally in the visual scene was either Reference or Figure; likewise, at the 500 ms delay, the object presented centrally (and thus presented first) was either Reference or Figure. Six practice trials preceded the test trials and contained objects not used in the main study (a book and bookshelf), all in the Simultaneous object-order condition. This resulted in 134 trials in total. Test trial order was randomized for each participant.

3.1.3. Analysis

As stated in our preregistration, the dependent variable was the participant’s response time (RT) on each trial, measured from the onset of the second image. (For the Simultaneous condition, the onset of the second image was also the onset of the full scene.) Only Match trials in which the visual scene matched the sentence description were analyzed. We also excluded trials that timed out before a response was given, error trials, and trials in which RTs were extraordinarily fast (<200 ms). Additionally, we excluded participants who met any of the following criteria: (a) Low overall accuracy (<90%); (b) too many timeouts (>5% of trials); (c) too many extraordinarily fast RTs on test trials (>5% of RTs <200 ms); (d) after trial exclusion, not having at least one trial in a cell for each combination of the factors of interest (i.e., Relation Type, Object Order, and Sentence Structure); or (e) failing to contribute a complete dataset. However, none of the results reported here or in subsequent experiments were dependent on the particular exclusion criteria used for RT or accuracy; in other words, the effects were significant in the same direction regardless of whether the exclusion criteria were applied or not.

We conducted a repeated-measures analysis of variance (ANOVA) on participant means of inverse-transformed response times ($-1000/\text{RT}$, correctly answered Match trials only) to examine the main effects of Object Order, Relation Type, and Sentence Structure, as well as their interactions. Our primary question in this study concerned Object Order: In particular, we expected to observe shorter RTs when the reference object was presented right before the figure object in the scene, as compared to vice versa. We expected this to hold regardless of relation type (spatial or physical). Crucially, we predicted that this would also hold regardless of sentence structure (i.e., whether the figure or reference object appeared as the grammatical subject of the sentence).

3.2. Results

Three participants were excluded based on preregistered criteria, leaving 37 participants for further analysis. As expected, participants had little difficulty completing the task, with a mean accuracy of 97% and a mean response time (across all conditions) of 741 ms.

The ANOVA revealed a significant main effect of Object Order, $F(2, 72) = 77.98, p = 2 \times 10^{-16}, \eta_p^2 = 0.68$. As shown in Fig. 4A, the mean RT for the Reference-first condition was shorter than Figure-first and Simultaneous conditions, indicating that participants were faster to recognize the visual scene and verify that it matched the sentence when the reference object appeared before the figure object. Subsequent Holm-Bonferroni-corrected paired-samples t -tests revealed significant pairwise differences between the three Object-Order conditions (Reference-first vs. Figure-first, Reference-first vs. Simultaneous, and Figure-first vs. Simultaneous, all $t_s(36) > 5.44, p_s < .0001, d_s > 0.25$). We note that the Simultaneous condition was slower than both the other two conditions. While we did not have strong predictions with respect to this condition, one possible explanation is that sequential presentation, even in the Figure-first condition, helps pre-segment the objects, whereas in the Simultaneous condition, participants can only begin to perform this segmentation when the objects have appeared on the display.

Fig. 4B illustrates the effect of Object Order, split by Relation Type and Sentence Structure. As predicted, the ‘Reference-object advantage’ arose for both physical and spatial relational scenes, and it held no matter the order of elements mentioned in the sentence descriptions. Indeed, while there was a significant interaction of Object Order and Sentence Structure ($F(2, 72) = 7.0, p = 0.0017, \eta_p^2 = 0.16$), Holm-Bonferroni-corrected paired-samples t -tests confirmed significant pairwise differences between the three Object-Order conditions at each level of Sentence Structure (all $t_s(36) > 3.22, p_s < .005, d_s > 0.54$). In addition, Relation Type and Sentence Structure both emerged as significant main effects (Relation Type: $F(1, 36) = 129.80, p = 1.7 \times 10^{-13}, \eta_p^2 = 0.78$; Sentence Structure: $F(1, 36) = 11.89, p = .0015, \eta_p^2 = 0.25$). However, there was no significant three-way interaction among the three factors ($F(2, 72) = 1.40, p = .25, \eta_p^2 = 0.037$).²

² It is worth noting that one prompt version differed slightly from the others: the Reference-first physical-relation sentence probe used constructions such as “The table is supporting the vase”, which employs a verb (“support”) rather than a preposition (as in “on”, “above”, or “below”). However, no appreciable slowdown was observed in this condition compared to the others (see Fig. 4B). Moreover, the absence of a significant three-way interaction suggests that the object-order effect for this condition was largely consistent in both direction and magnitude with the other conditions.

Experiment 2

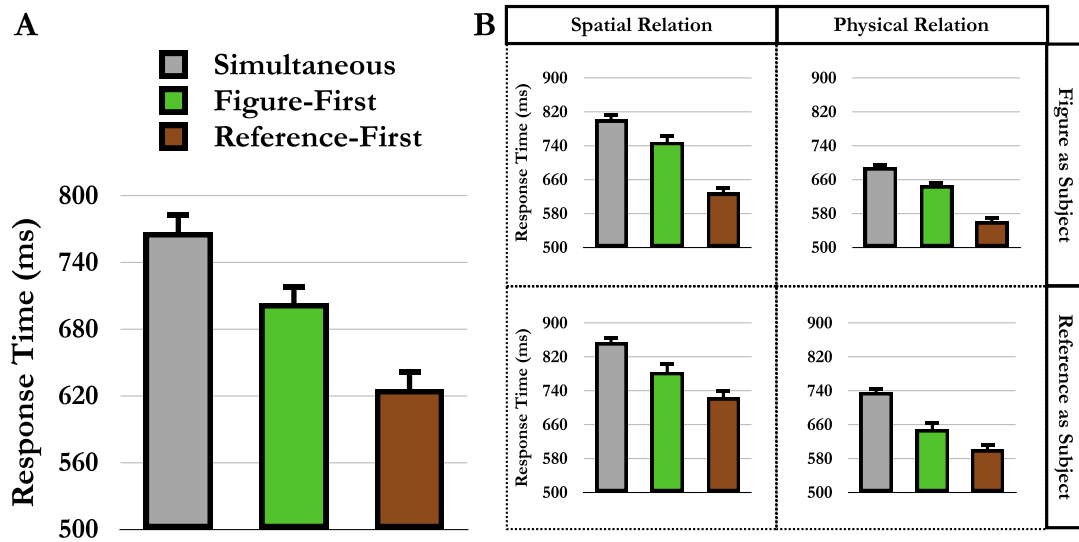


Fig. 4. Results of Experiment 2. (A) Participants were faster verifying the relational scene when the reference object appeared first and the figure object second, compared to when the order was reversed or when both objects appeared simultaneously. (B) This Reference-first advantage in relational recognition emerged for both spatial (“above”/“below”) and physical relations (“on”/“support”), regardless of which object was mentioned first in the sentence. Bars reflect mean response times across participants (computed from correct trials only), and error bars reflect within-participant 95% confidence intervals.

We also conducted secondary preregistered analyses to test whether the effect of Object Order generalized across the different object pairs. A repeated-measures ANOVA for mean RTs across object pairs showed a significant main effect of Object Order, $F(2, 6) = 55.8, p = 1.33 \times 10^{-14}, \eta_p^2 = 0.95$. Subsequent Holm-Bonferroni-corrected paired-samples t -tests revealed significant pairwise differences of the three Object-Order conditions, all $t_s(3) > 3.85, p_s < 0.033, d_s > 1.92$.

4. Experiment 3 – Timing of the compositional process

Experiment 2 revealed a Reference-first advantage in the recognition of relational visual scenes: Participants were faster to match the scene to the pre-specified linguistic description when the reference object appeared just a half-second before the figure object, rather than vice versa. This result raises a natural question: How quickly does this effect emerge? Here we probed the timing of the compositional process in detail. We ran the same study as before, but with one change: We systematically varied the onset of the second object in the scene, from very early (100 ms) to late (1000 ms). This manipulation also allowed us to distinguish between different possible underlying mental processes driving the Reference-first advantage. One possibility is that the effect arises from deliberate, cognitive expectations: Participants may form predictions about what will appear next based on the first object’s identity. In that case, the effect might take time to emerge and possibly strengthen as more time is available for reasoning between presentation of the two objects. Alternatively, visual processing itself might enforce a Reference-first order when constructing relational representations. In this case, we would expect the effect to emerge rapidly, even with minimal delays between the first and second object. By systematically varying the presentation timing, we aimed to distinguish between these possibilities.

4.1. Method

4.1.1. Participants

One-hundred fifty participants were recruited through Prolific. This sample size was chosen based on a power analysis of a small pilot study and was preregistered.

4.1.2. Procedure

Stimuli, procedures, and exclusion criteria were identical to those in Experiment 2, except that the onset delay of the second object varied across trials: 100 ms, 250 ms, 500 ms, and 1000 ms (Fig. 5A). There was no Simultaneous (i.e., 0 ms) condition in this experiment. Otherwise, the factors in the experiment were the same as in Experiment 2: 4 (Second Image Delay) $\times$ 2 (Object Order: Reference-first or Figure-first) $\times$ 2 (Relation Type: Physical or Spatial) $\times$ 2 (Trial Type: Match or Mismatch) $\times$ 2 (Sentence Structure: Figure-as-subject or Reference-as-subject), yielding 64 trials. The four pairs of objects were fully distributed among three primary

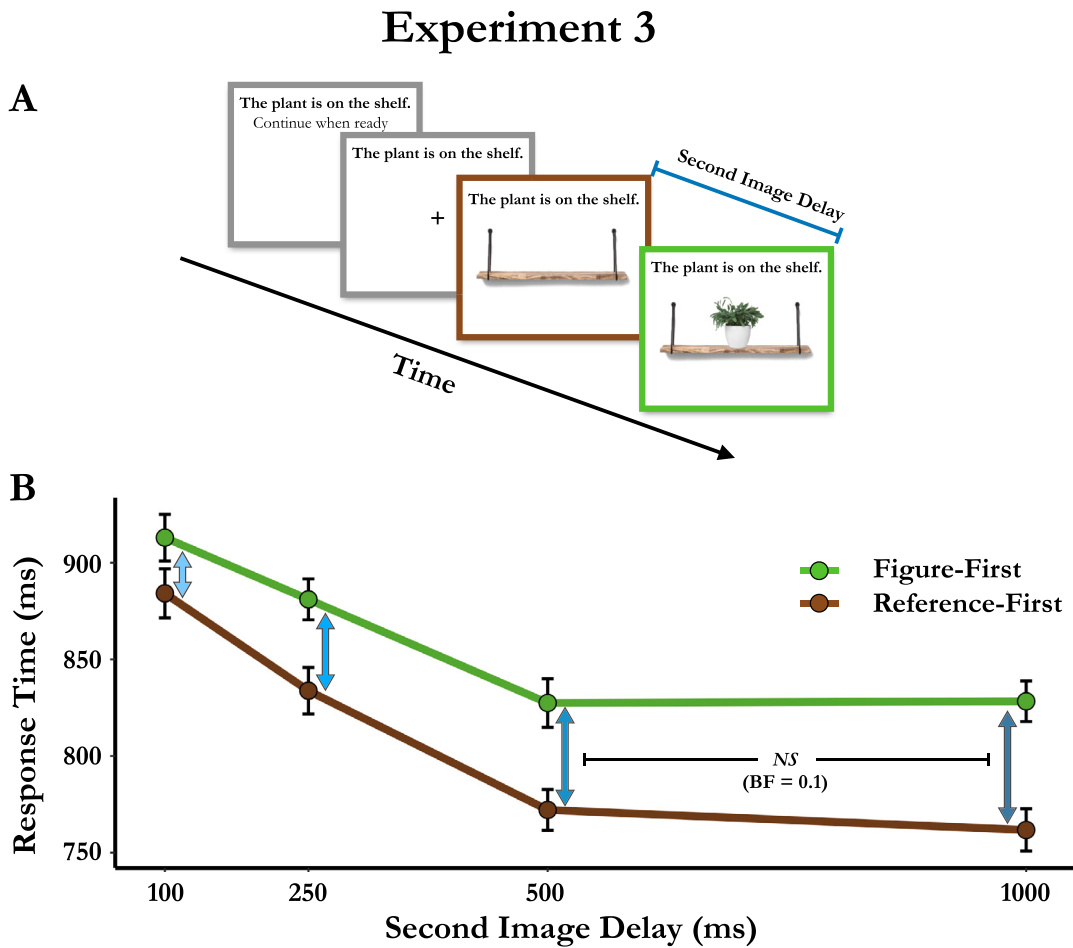


Fig. 5. Trial structure and results of Experiment 3. (A) The delay between figure object and reference object presentation was systematically varied from 100 ms to 1000 ms. (For simplicity, blank displays during the trial are omitted from this figure.) (B) The Reference-first advantage (i.e., faster RTs when the reference object appeared first) emerged as early as 100 ms, increased to a peak at 500 ms, and then plateaued. Filled circles reflect mean response times across participants (computed from correct trials only), and error bars reflect within-participant 95% confidence intervals.

factors—Second Image Delay, Object Order, and Relation Type—and randomly assigned among other factors (Trial Type [Match or Mismatch] and Sentence Structure [Reference-as-subject or Figure-as-subject]). The visual scene in mismatch trials presented the right pair of objects but in the wrong relational category, as in the previous study. Participants were given six practice trials at the beginning of the experiment to become familiar with the task (all with 0 ms delay between the first and second object).

4.2. Results

Twenty-four participants were excluded based on preregistered criteria, leaving 126 participants for further analysis. The remaining participants had little difficulty completing the study, with a mean accuracy of 97% and a mean response time (across all conditions) of 865 ms.

We first conducted a repeated-measures ANOVA on participant means of inverse-transformed response times ($-1000/\text{RT}$, correctly answered Match trials only) to examine the main effects of interest in this experiment—Second Image Delay, Object Order, Relation Type—as well as their interactions. Consistent with Experiment 2, the ANOVA revealed a significant main effect of Object Order, $F(1, 125) = 85.89, p = 7.02 \times 10^{-16}, \eta_p^2 = 0.41$, confirming an overall Reference-first RT advantage.

The ANOVA also revealed a significant interaction between Object Order and Second Image Delay, $F(3, 375) = 4.35, p = 0.0050, \eta_p^2 = 0.033$, suggesting that the magnitude of Reference-first advantage differed depending on the onset delay of the second object. To examine this further, we first computed the mean Reference-first RT advantage at each level of delay (collapsing over Relation Type) by subtracting Reference-first from Figure-first inverse RTs. One-sample t -tests showed a significant Reference-first advantage at all delay conditions (100 ms: $t(125) = 2.5, p = .013, d = 0.22$; 250 ms: $t(125) = 4.7, p = 1.29 \times 10^{-5}, d = 0.42$; 500 ms:

$t(125) = 5.8$, $p = 1.31 \times 10^{-7}$, $d = 0.52$; 1000 ms: $t(125) = 6.4$, $p = 9.60 \times 10^{-9}$, $d = 0.57$). That is, participants were faster to verify relational scenes when the reference object appeared right before the figure object, even with a minimal 100 ms delay (Fig. 5B).

To further explore how this effect unfolded over time, we conducted a series of paired-sample t -tests to compare the Reference-first advantage between each pair of delay conditions. The Reference-first advantage significantly increased from the 100 ms delay to the 500 ms delay ($t(125) = 3.14$, $p_{corrected} = 0.011$, $d = 0.28$), but then showed no further increase in magnitude from 500 ms to 1000 ms ($t(125) = 0.25$, $p_{corrected} = .80$). An exploratory Bayesian paired-sample t -test provided further evidence in favor of no difference between the 500 ms and 1000 ms conditions ($BF_{01} = 0.1$, using the default Cauchy prior with scale $\sqrt{2}/2$). In other words, the Reference-first RT advantage peaked at the 500 ms delay and then plateaued (See Fig. 5B).

Overall, the Reference-first advantage emerged even when the figure object appeared just a brief moment after the reference object (just 100 ms). This effect quickly increased and then plateaued after a half-second delay. These results suggest that the compositional process for building relational representations from visual scenes is rapid and does not rely on slow, deliberate reasoning processes, such as predicting what should happen next after seeing a given object.

5. Experiment 4 - Image-only composition

The previous experiments showed a Reference-first advantage in recognition. However, they involved both visual and linguistic components: participants compared a visual representation to a linguistic one (i.e., the previously presented sentence). Because natural language is inherently compositional, the linguistic probe could have influenced how the recognition process unfolded, rather than reflecting compositional process naturally occurring within vision.

The present experiment minimized that potential influence by running the original recognition study with one crucial change: we replaced the linguistic probe with a visual one. This allowed us to test whether evidence of compositional visual processing emerges independent of the format of the probe stimulus.

5.1. Method

5.1.1. Participants

Forty participants were recruited for this experiment through Prolific. This sample size was the same as Experiments 1 and 2.

5.1.2. Procedure

The same stimulus set from Experiments 2 and 3 was used in this recognition task. The paradigm was similar to that used in Experiments 2 and 3, with one key difference: instead of using linguistic descriptions as probes, each trial began with a pictorial probe showing a target relational scene. As shown in Fig. 6A, the probe image was presented for 350 ms, followed by a mask image for 350 ms. The mask was a box-scrambled version of all object images (chosen randomly with replacement from 16 masks, made up of 22×22 blocks), and was included to interrupt the formation of afterimages and iconic memory for the probe. After the mask, a blank screen of 100 ms and a fixation cross of 300 ms were displayed. Then, either the figure object or the reference object was displayed, followed by the second object 500 ms later, as in Experiment 2. At this point, participants responded whether the scene matched the probe image or did not. The experiment followed a 2 (Relation Type: Physical vs. Spatial) $\times$ 2 (Object Order: Figure-first or Reference-first) $\times$ 2 (Trial Type: Match or Mismatch) $\times$ 4 (Object Pairs) design, resulting in 32 unique trials. Each unique trial repeated twice, and there were also 6 practice trials at the beginning of the session, resulting in a total of 70 trials. The visual scene in mismatch trials presented the right pair of objects but in the wrong relational category, as in the previous two studies. Practice trials always featured a 500 ms delay between objects. The probe was presented centered on a frame with a gray background and dashed border, scaled to 75% of the size of the target scene in order to avoid overlap in image features between probe and target.

5.2. Results

Six participants were excluded based on our preregistered exclusion criteria, leaving 34 participants for further analysis. The remaining participants had little difficulty completing the study, with a mean accuracy of 96% and a mean response time (across all conditions) of 773 ms.

As shown in Fig. 6B, a Reference-first RT advantage was again observed: Participants were faster to match the visual scene with the pictorial probe when the reference object appeared right before the figure object in the scene, rather than vice versa (725 ms vs. 800 ms). A 2×2 repeated-measures ANOVA on participant means of inverse-transformed response times ($-1000/\text{RT}$, correctly answered Match trials only) that included Object Order and Relation Type as factors confirmed a significant main effect of Object Order ($F(1, 33) = 44.61$, $p = 1.33 \times 10^{-7}$, $\eta_p^2 = 0.57$). There was also a significant effect of Relation Type ($F(1, 33) = 112.4$, $p = 3.67 \times 10^{-12}$, $\eta_p^2 = 0.77$), with responses to physical relations being faster than those to spatial relations (690 ms vs. 836 ms). There was no significant interaction between Object Order and Relation Type ($F(1, 33) = 0.51$, $p = .48$, $\eta_p^2 = 0.015$).

These findings demonstrate a Reference-first advantage independent of the format of the probe stimulus. This suggests that the compositional process for building relational representations can arise in visual processing alone, rather than arising only from a cognitive comparison between representations formed from linguistic and visual modalities.

Experiment 4: Image-only Composition

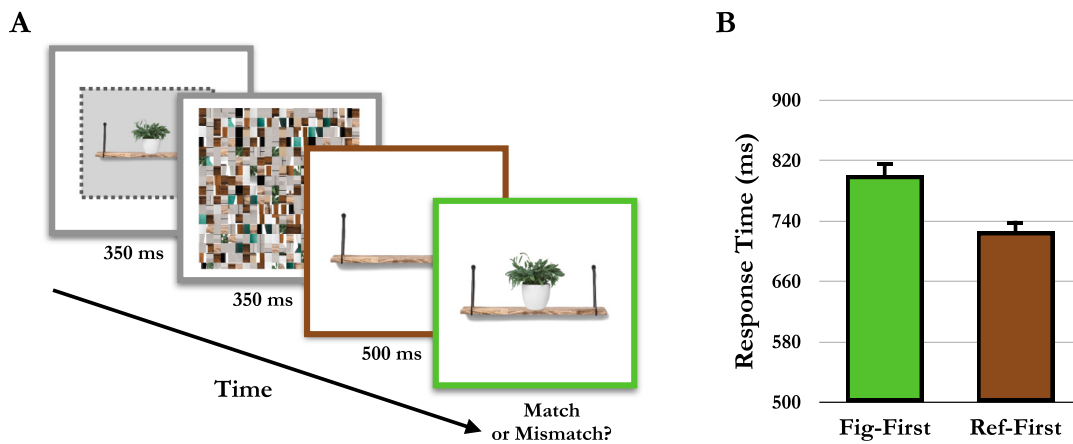


Fig. 6. Trial structure and results of Experiment 4. (A) Participants viewed a relational scene; after being replaced by a mask, its constituent objects reappeared either in a Reference-first or Figure-first sequence (e.g., the shelf appearing before the plant). Participants then judged whether the second scene matched the first. (For simplicity, the blank display and fixation cross during the trial are omitted from this figure.) (B) Responses were faster when the reference object was displayed first. Bars reflect mean response times across participants (computed from correct trials only), and error bars reflect within-participant 95% confidence intervals.

6. Experiment 5 - Identical-object composition

In our previous experiments, the reference objects (i.e., table, desk, nightstand, and shelf) were typically large, stable, and rectilinear, while the figure objects (i.e., vase, laptop, lamp, and plant) were smaller, more mobile, and had more rounded features (Fig. 2B). This raises a question: How are role assignment and compositional order determined when no such asymmetry exists between the objects in a relation?

If the Reference-first compositional order reflects a general principle of relation perception, then it should also emerge in scenes involving physically identical objects differing only in role, where lower-level visual differences are not present (Long, Konkle, Cohen, & Alvarez, 2016; Long, Yu, & Konkle, 2018). In such cases, we reasoned that role assignment—and the sequence in which objects assigned these roles are composed—would rely on other cues: intuitive physics (for physical relations) and linguistic structure (for spatial relations).

For physical relations such as support-from-below (ON), Reference-hood is determined by whether one object physically *controls* another (Landau & Gleitman, 2015; Talmy, 1975). These ‘hidden’ forces underlie many sophisticated relations between objects, transcending the lower-level visual properties of the objects themselves (Hafri & Firestone, 2021). Work on intuitive physics highlights how observers infer such forces, constraints, and causal interactions even when they are not directly visible, relying on structured internal models of the physical world (e.g., Battaglia, Hamrick, & Tenenbaum, 2013; Kubricht, Holyoak, & Lu, 2017; Ullman, Spelke, Battaglia, & Tenenbaum, 2017). For example, in Fig. 7A, the red book *supports* the blue book against gravity, establishing the red book as the reference object and the blue book as the figure object. We therefore hypothesized that the supporting object (red book) should be composed earlier in processing, producing a corresponding response-time advantage.

For spatial relations without physical control, such as ABOVE (e.g., “The blue book is above the red book”; Fig. 7B), the relevant cue to role assignment is linguistic in nature. Talmy and others (Gleitman et al., 1996; Landau & Jackendoff, 1993; Talmy, 1975) have noted that conceptual Figure and Reference roles parallel grammatical roles in language, with figure objects (e.g., a plant) often mapped to subject position and reference objects (e.g., a table) to complement position—a lower position in syntactic structure (see Fig. 7B). Notably, this mapping has bidirectional effects: not only can objects’ conceptual roles determine their positions in syntactic structure, but objects’ positions in syntactic structure can also shape how they are conceptually construed. For instance, the sentences *The bike is next to the garage* and *The garage is next to the bike* differ in their implied interpretations. While the latter sentence is unusual, the described scenario becomes plausible in a context where the garage is mobile and the bike is stationary—precisely the properties that figure and reference objects, respectively, often possess. Thus, syntax can guide relational interpretation in the absence of visual cues and even ‘imbue’ participating objects with the typical properties of their assigned roles.

Given this potential influence of linguistic structure, we hypothesized that for purely spatial relations, the object in complement position in the sentence probe would be composed earlier in scene processing and enjoy a response-time advantage, consistent with the Reference-first effect observed in our previous recognition experiments. In Fig. 7B, for example, the sentence *The blue book*

Experiment 5: Identical-Object Composition

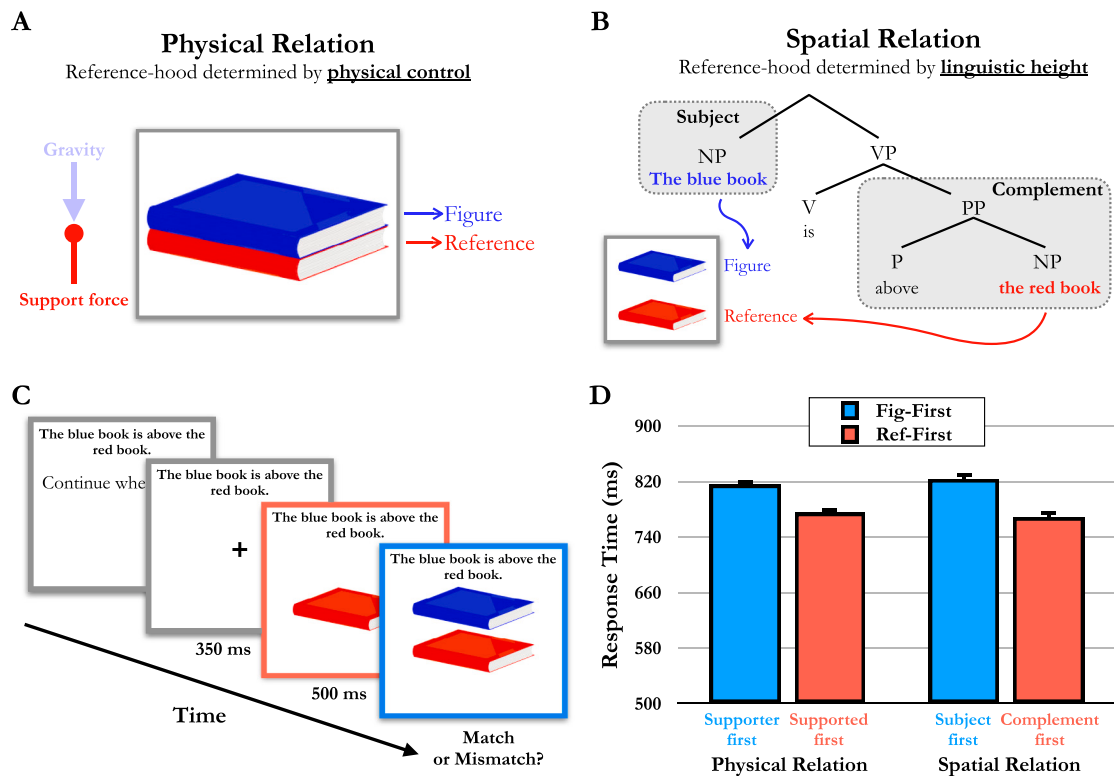


Fig. 7. Illustration of the identical-object task in Experiment 5. Since the two objects differed only in color, their roles were determined by either physical control or linguistic structure. In the experiment, the colors of figure and reference object were counterbalanced. (A) For physical relations, the reference object was the one that provided support against the force of gravity for the other object. (B) For spatial relations, the reference object was the entity embedded lower in the linguistic structure than the other object (syntax tree shown schematically to illustrate relative embedding rather than commit to a particular syntactic analysis). (C) In the task, participants read a sentence description at the beginning of each trial and then verified whether the subsequent scene matched the description. (The 500 ms blank between sentence and fixation is omitted from the figure for simplicity.) (D) Results showed that participants were faster in verifying scenes when the reference object appeared before the figure object, for both physical and spatial relations. Bars reflect mean response times across participants (computed from correct trials only), and error bars reflect within-participant 95% confidence intervals.

is above the red book has the blue book in subject position and the red book in complement position; we therefore predicted an advantage when the red book appeared before the blue book in the scene.³

To test these predictions, Experiment 5 reintroduced the sentence-picture verification task with identical-object stimuli (two books), enabling us to examine how the compositional process operates when visual asymmetries are absent and role assignment must rely on intuitive-physical or linguistic factors.

6.1. Method

6.1.1. Participants

In line with our preregistration, 100 participants were recruited through Prolific. This sample size was chosen based on a power analysis of a small pilot study.

³ We also piloted an image-only version of the task using the book stimuli, where participants viewed spatial relations (e.g., red book above blue book) without linguistic prompts. Preliminary data appeared to show a *bottom-first* RT advantage: Participants responded faster when the object positioned at the bottom of the scene (e.g., the blue book) appeared first. While we initially expected no preference for spatial relations, this result suggests that, in the absence of linguistic or physical-control cues (i.e., support-from-below), a bottom-first strategy may emerge as a default visual bias (see Langley & McBeath, 2023 for evidence that lower regions of scenes are perceptually salient). However, as will be shown, Experiment 5 revealed that the bottom-first bias for visual relations is flexible: Linguistic structure effectively overrides it for spatial relations like *ABOVE*, while physical control dominates role assignment for *ON* relations regardless of the syntactic structure of the linguistic probe.

6.1.2. Stimuli and procedure

The relational scenes in this experiment consisted of two identical objects differing only in color: a red book and blue book. As in previous experiments, we tested two types of relations: physical and spatial. For physical relations, Reference and Figure roles were determined by physical control, with the supporting object designated as the Reference (Fig. 7A). For spatial relations, however, Reference and Figure roles were determined by the grammatical structure of the sentence probe (Fig. 7B). For example, when the sentence “The red book is below the blue book” served as the linguistic probe, the grammatical complement (i.e., blue book) functioned as Reference, and the grammatical subject (i.e., red book) functioned as Figure.

As in previous experiments, there were several factors fully crossed within participant: (a) Relation Type (Physical or Spatial); (b) Book Order (Red or Blue book appearing first in the visual scene); (c) Sentence Structure (Red or Blue book as grammatical subject of the linguistic probe); and (d) Role Assignment (Red or Blue book serving as Reference in the visual scene, with role assignment determined by relation type, as described above). This resulted in 16 unique trials ($2 \times 2 \times 2 \times 2$), with each repeated 4 times, yielding 64 Match trials. In addition, there were 64 Mismatch trials, consisting of 3 types: (i) Role Mismatch (16 trials): The Reference and Figure roles were swapped (e.g., if the sentence stated, “The red book is on the blue book”, the visual scene depicted a *blue* book on a *red* book); (ii) Relation Mismatch (16 trials): The relational category was changed (e.g., if the sentence stated, “The red book is on the blue book”, the visual scene depicted a red book *below* a blue book); and (iii) Full Mismatch (32 trials): Both role assignment and relational category were inconsistent with the sentence. The full experiment consisted of 136 trials (8 practice trials, 64 Match trials, and 64 Mismatch trials). The 128 test trials were presented in a fully randomized order for each participant.

The experimental paradigm was similar to that used in Experiments 2 and 3: After participants proceeded past the sentence probe, one object appeared, followed by the second object 500 ms later, at which point participants could make their response (Fig. 7C).

For analyses, we created a new Object Order factor (with Reference-First and Figure-First conditions), where Reference and Figure were defined differently based on the Relation Type. For physical relations, the supporting object was Reference and the supported object was Figure. By contrast, for spatial relations, these were based on grammatical position in the sentence probe: the grammatical object served as Reference and the grammatical subject as Figure. See Figs. 7A and 7B for examples.

6.2. Results

In accordance with our preregistered exclusion criteria, 16 participants were excluded, leaving 84 participants for further analysis. The remaining participants had little difficulty completing the study, with a mean accuracy of 97% and a mean response time (across all conditions) of 846 ms.

As shown in Fig. 7D, even though the figure and reference objects were identical in size and shape, participants were still faster verifying the relational scene when the reference object appeared before the figure object than vice versa (772 ms vs. 820 ms). This Reference-first RT advantage was confirmed by a 2×2 repeated-measures ANOVA, which revealed a main effect of Object Order ($F(1, 83) = 57.40, p = 4.51 \times 10^{-11}, \eta_p^2 = 0.41$). There was also a marginally significant interaction between Relation Type and Object Order ($F(1, 83) = 3.77, p = .056, \eta_p^2 = 0.043$), but no significant main effect of Relation Type ($F(1, 83) = 0.12, p = .74, \eta_p^2 = 0.0014$).

To further examine the effect of Object Order within each Relation Type, we conducted Holm-Bonferroni-corrected paired-samples *t*-tests separately for physical relations and spatial relations. Both analyses confirmed the advantage for Reference-first order (Physical: $t(83) = 4.03, p_{corrected} = 1.24 \times 10^{-4}, d = 0.44$; Spatial: $t(83) = 7.70, p_{corrected} = 5.20 \times 10^{-11}, d = 0.84$). Finally, to determine whether the Object-Order effect was stronger for certain relational words in the sentence probes, we conducted two separate paired *t*-tests on the Reference-first RT advantage (computed as in Experiment 3): one for physical relations (“on” vs. “supporting”) and one for spatial relations (“below” vs. “above”). Neither test was significant ($ts(83) < 1.37$, corrected $ps > .34$).

This experiment disentangled objects’ relational roles from relevant visual properties (e.g., size, shape), allowing us to draw several conclusions. First, the persistence of the Reference-first RT advantage suggests that the Reference-first advantage may be driven by objects’ abstract roles as Reference and Figure, rather than solely by their visual features. Second, when the asymmetry between objects is not given by their visual differences (e.g., size, shape) nor by their physical relationship (e.g., via physical control), the syntactic structure of a linguistic description can guide Figure and Reference assignments, shaping the order in which relational representations are mentally constructed.⁴

⁴ To further test whether the linguistic effects in Experiment 5 were driven by syntactic structure per se rather than surface word order, we piloted a version of the task using *it*-cleft sentences (e.g., *It is the red book that is below the blue book* vs. *It is the red book that the blue book is below*). These sentences reverse the order in which the objects are mentioned while preserving the deeper syntactic structure (i.e., which entity serves as subject vs. complement). Preliminary results trended in the predicted direction—consistent with syntactic structure driving the effect—but the task proved challenging for participants due to the processing complexity of *it*-cleft sentences, leading to longer and more variable response times. For this reason, we did not pursue a full version of this study, though the pilot findings support the conclusion that syntactic structure, rather than linear word order, determines relational composition in such cases.

7. General discussion

The present work investigated how everyday visual relations (such as one object supporting another) are represented in the mind, revealing key principles governing how they are composed in time. In a manual construction task (Experiment 1), participants assembled relational scenes by placing reference objects (e.g., tables, desks) first, followed by figure objects (e.g., vases, laptops). Likewise, across a series of visual recognition tasks (Experiments 2–5), participants demonstrated a Reference-first advantage, responding faster when the reference object appeared before the figure object. This Reference-first advantage emerged rapidly—within just 100 ms—and plateaued by 500 ms (Experiment 3), suggesting that it does not rely on deliberative expectations about object order. Notably, this effect persisted even without linguistic prompts, suggesting that the visual system independently employs a Reference-first compositional routine (Experiment 4). Finally, the effect extended to cases where the participating objects themselves provided no intrinsic asymmetries—i.e., two identical books (Experiment 5). Here, role assignment and the Reference-first advantage were determined instead by other cues: intuitive physics guided role assignment in physical relations, while linguistic structure did so for purely spatial relations. Taken together, our findings suggest that the mind automatically employs a sequential routine for composing relational representations, respecting each object's role in the relation: Reference first, Figure second.

7.1. Understanding perception as a compositional process

Our work aligns with a growing body of research exploring compositional representations in visual perception. Classic theories of object representation propose that objects are not processed as undifferentiated wholes, but rather as hierarchically decomposed structures consisting of parts and subparts (Biederman, 1987; Feldman & Singh, 2006; Marr & Nishihara, 1978). Such hierarchical structures not only define the relations among constituent parts but also establish a principle of primacy, whereby certain components carry greater weight than others in the representation. For example, the main skeletal axis of a shape is given greater representational priority than the 'ribs' that define the peripheral parts of a shape, reflecting an intrinsic ordering principle in object perception (El-Gaaly, Froyen, Elgammal, Feldman, & Singh, 2015; Feldman & Singh, 2006). This hierarchical precedence is even reflected in how object representations are 'grown' or generated in the mind: Mounting empirical evidence suggests that such representations are composed in a systematic order, where main axes emerge before subordinate branches, reinforcing the psychological reality of these processes (Ayzenberg & Lourenco, 2022; Destler, Singh, & Feldman, 2023; Sun & Firestone, 2022).

Our findings extend this principle to between-object relations: Just as certain object parts take precedence in forming object representations, certain relational constituents (here, reference objects) take precedence in forming relational representations. In doing so, our work contributes to growing scientific attention on visual relations as a fundamental unit of perception and cognition (for reviews and discussion, see Cavanagh, 2021; Hafri & Firestone, 2021; Hafri, Green, & Firestone, 2023; Hafri & Papeo, 2025; Hochmann & Papeo, 2021; Hummel & Holyoak, 2003; Kaiser et al., 2019; Miller & Johnson-Laird, 1976; Papeo, 2020; Peelen, Berlot, & de Lange, 2024; Quilty-Dunn et al., 2023; Vö et al., 2019).

The fact that compositional representations exist both within and between objects may suggest a general kind of compositional process that extends across diverse relational domains. One key example comes from social cognition, where event representations are sensitive to certain compositional orders for relational roles like Agent and Patient. In particular, Agents are prioritized in perception, recognition, and prediction of events, substantiating their primary role in relational encoding (Brocard, Wilson, Berton, Zuberbühler, & Bickel, 2024; Cohn & Paczynski, 2013; Cohn et al., 2017; Hafri et al., 2018; Sauppe & Flecken, 2021; Ünal, Wilson, Trueswell, & Papafragou, 2024; Wilson, Zuberbühler, & Bickel, 2022). Our findings extend these insights from social relations to physical and spatial ones, advancing beyond previous work in this area (Boettcher et al., 2018; Franconeri, Scimeca, Roth, Helseth, & Kahn, 2012; Holcombe et al., 2011). This is consistent with the idea that relational structure, rather than individual objects, form the backbone of visual scene perception writ large.

In this way, the present work also dovetails with a renewed interest in the 'Language-of-Thought' (LoT) hypothesis (Fodor, 1975), which has re-emerged in debates over the formats of mental representation (see Quilty-Dunn et al., 2023, and commentaries therein). The LoT framework proposes that certain types of mental representations are composed of discrete, combinable elements—a principle originally formulated to explain systematicity and productivity in language and high-level reasoning. Recent extensions of this framework suggest that LoT-like representations may exist across many cognitive systems, including in visual perception (Mandelbaum et al., 2022). In a recent article (Hafri, Green, & Firestone, 2023), we built on this idea, proposing that visual perception exhibits core LoT-like properties, where discrete perceptual units (e.g., object parts or whole objects) combine systematically. The current work extends this proposal not only by reinforcing the notion that visual perception involves compositional representations, but also by revealing the *processes* by which they are constructed—in other words, the 'psychophysics' of compositionality.⁵

⁵ At first glance, there might seem to be a tension between the sequentiality revealed in our experiments, which suggest a visual routine-like process, and other findings demonstrating the rapid, spontaneous extraction of relational representations—even from exposures as brief as 37 ms (e.g., Chen, 1982; Hafri et al., 2013). Yet, these observations are not contradictory. Visual routines need not be deliberate or effortful; they can unfold automatically and at great speed (Wong & Scholl, 2024). Moreover, the temporal intervals we used (typically 500 ms between objects) were chosen to elicit sequential effects, not to capture the natural time course of representation building. Indeed, the same ordering effect emerged even with interstimulus intervals as short as 100 ms (Experiment 3), consistent with the idea that such routines operate extremely rapidly.

7.2. Syntax in vision?

Our work explores compositional processes in perception with a focus on *syntax*—the structure of representations—rather than their content. Related research on *scene grammar* usefully distinguishes the ‘semantics’ of a scene—*what* objects belong (e.g., a computer mouse, rather than a pot of soup, should appear in an office)—from its ‘syntax’—*where* those objects belong (e.g., a mouse should appear next to a laptop instead of on top of its screen; Boettcher et al., 2018; Draschkow & Vö, 2017; Vö, 2021; see also Kaiser et al., 2019). This work even highlights the role of “anchor” objects in scaffolding scene representations (Draschkow & Vö, 2017; Vö, 2021).

Our approach complements this work by going beyond such statistical regularities about specific objects and their familiar contexts. Instead, we probe the structural principles by which visual elements combine into relational representations in the first place. This effort is informed by analogous investigations in linguistics that seek to identify the syntactic principles that determine how words combine to form phrases, clauses, and sentences (Chomsky, 1995), both at a theoretical level and in terms of the algorithmic mechanisms (e.g., ordering and timing) that generate them. In this sense, our framework treats vision as governed by a similar kind of syntax that posits abstract categories and principles determining how elements combine into structured representations (Cavanagh, 2021). This perspective opens new directions for identifying potential “grammatical” categories in visual cognition (e.g., analogs of nouns, verbs, or prepositions; Ji & Scholl, 2024) and for investigating the algorithmic processes that transform visual input into structured knowledge of the world.

7.3. The interface between vision and language

Our findings highlight a novel aspect of how visual and linguistic systems might interact during scene perception. On the one hand, the exact word order of the linguistic prompts we employed did not generally alter the order by which relational representations were constructed (indexed by RT differences between Reference- and Figure-first orders). On the other hand, Experiment 5 demonstrated that linguistic structure can guide the assignment of Figure and Reference roles when strong visual cues to asymmetry are absent, influencing the compositional process (e.g., in visually symmetric relations such as *above* or *below*; Gleitman et al., 1996; Talmy, 1975). However, we suspect that this influence of language does not directly alter visual representations themselves (e.g., by changing the objects’ appearance in the scene; Firestone & Scholl, 2016). Instead, it may function as a kind of ‘cognitive instruction,’ guiding how observers attend to and encode an upcoming scene (Knowlton et al., 2021)—albeit in a manner that may not reach explicit awareness. Crucially, whatever the precise mechanism by which language influences scene processing here, its effect is not determined merely by the probe’s surface word order (in which the figure object appears first in sequence, the reference object second), but by its syntactic structure—that is, by what occupies the hierarchically lower syntactic position.

This pattern joins classic and more recent literature detailing the interactions between linguistic and visual systems (Cavanagh, 2021; Jackendoff, 1987; Miller & Johnson-Laird, 1976; Strickland, 2017). For example, in recent work, we found that the mind is sensitive to correspondences between linguistic and visual notions of symmetry (Hafri, Gleitman, Landau, & Trueswell, John, 2023). Despite the striking differences between a butterfly’s appearance and a sentence like *Mary and Bill marry*, both share an abstract symmetry—an invariance to transformation. In images, this is evident in the bilateral symmetry of a butterfly; in language, it is reflected in flexible argument order for certain predicates like *marry* (e.g., *Mary marries Bill* vs. *Bill marries Mary*). In cross-modal matching tasks, we observed surprising correspondences across such stimuli, providing evidence for these intuitive psychological connections between symmetry in vision and language. Other studies demonstrate a tight link between visual event apprehension and language, whether in the formation of descriptions of observed events (e.g., Gleitman, January, Nappa, & Trueswell, 2007) or in the rapid lexical activation of event labels triggered by briefly masked action scenes (Zwitserslood, Bölte, Hofmann, Meier, & Dobel, 2018). These and other findings (e.g., De Freitas & Alvarez, 2018) suggest that the mind employs common formats and principles across cognitive systems—allowing perceptual representations of relations to be readily accessed by higher-level processes (Quilty-Dunn, 2020) and, in some cases, enabling linguistic representations to guide attentional patterns in scene perception (Isasi-Isasmendi et al., 2023; Sauppe & Flecken, 2021; Trueswell & Papafragou, 2010).

7.4. Open questions and future directions

These connections across cognitive systems raise a broader question: Why does perceptual compositional order appear to align so strongly with the hierarchical order of elements in linguistic syntax? While speculative, one possibility is that both may reflect a general cognitive principle in which structure is built incrementally from the most foundational elements (e.g., reference objects) upward. In syntax, complements—where reference objects often reside—occupy lower positions in the hierarchy than subjects (where figure objects often end up), and many frameworks in both syntax and semantics assume that structures are derived bottom-up through recursive composition operations (e.g., Chomsky, 1995; Pietroski, 2011). Likewise, psycholinguistic evidence shows that planning and parsing unfold incrementally—often anticipating syntactically lower material even when it appears earlier in the linear sequence (Momma & Phillips, 2018). This pattern mirrors our findings on structure-building in scene composition, suggesting a striking parallel between visual composition and linguistic derivation. Such parallels resonate with longstanding hypotheses in evolutionary and developmental cognitive science, which propose that core aspects of linguistic grammar may be rooted in more fundamental, potentially cross-domain perceptual and cognitive capacities (Christiansen & Chater, 2008; Jackendoff, 2002; Strickland, 2017)—capacities that may even predate the emergence of language itself (Hafri, 2024; Papeo et al., 2024; Wilson et al., 2022).

Our findings highlight an intriguing dissociation between event and non-event relational representations in the alignment of compositional order and grammatical order. In event representations, Agents—typically dynamic entities that move or initiate change—are psychologically primary and usually mapped to grammatical subject position. In contrast, for physical and spatial relations like those studied here, reference objects are primary yet are generally mapped to grammatical *complement* position, occupying lower levels in hierarchical syntactic structure. While speculative, we suggest that this difference in mapping may reflect underlying differences in the functional properties of Agents and reference objects. Dowty's (1991) proto-role theory aims to formalize how prototypically Agent-like properties (such as movement or initiating change) determine the mappings between event participants (arguments of verbs like *kick* or *fear*) and grammatical roles (subject and object); however, this framework does not straightforwardly extend to non-event relations. Reference objects illustrate this tension clearly: although capable of initiating change (an “Agent-like” property—e.g., a table moving can cause a vase atop it to move as well), they typically function as stable, stationary anchors (a more “Patient-like” property). Thus, a complementary theoretical framework may be needed to capture generalizations across non-event relations like those studied here.

Our work also opens important avenues for future research. One critical question concerns how the mind learns to identify reference objects as such, including their capacity to support other objects in physical or spatial relations. Experiments 1–4 showed that visual properties such as size, stability, and rectilinearity often determine Reference role assignment, consistent with findings on “anchor” objects in scene perception (Boettcher et al., 2018; Vö, 2021; Vö et al., 2019). However, the results of Experiment 5 reveal that Reference role assignment can also emerge from more abstract principles, such as intuitive physics (Battaglia et al., 2013; Firestone & Scholl, 2017; Kubricht et al., 2017; Little & Firestone, 2021; Ullman et al., 2017) or even linguistic structure (Gleitman et al., 1996; Talmy, 1975). Developmental research suggests that preverbal infants are sensitive to abstract relational concepts like support and containment (Baillargeon et al., 2012; Hespos & Spelke, 2004), raising the possibility that early interactions with physical forces inform the mind's sensitivity to Reference-hood. Nevertheless, how these early experiences integrate with objects' visual properties and with higher-level conceptual principles remains an open question.

Finally, another important open question is what additional visual routines underlie relational composition beyond those documented for simple geometric relations (Jolicoeur et al., 1986, 1991; Ullman, 1987; Wong & Scholl, 2024). While our work uncovered a Reference-first order for scenes involving just two objects, additional routines may be involved for complex everyday scenes, which are often cluttered with a multitude of objects (e.g., a table, lamp, bookshelf, and books, to name a few) and thus may initially require selection processes of some sort. Moreover, questions remain about how more complex relations are represented compositionally, beyond the simpler dyadic cases studied here. For instance, transfer events (e.g., *giving*) or caused motion (e.g., hitting a ball with a racket into the net) involve at least three roles (Tatone & Csibra, 2024; Ünal et al., 2024). Likewise, relational structures may be embedded. Consider a cat on a mat that is in a box. Is this situation represented as two independent relations (cat-on-mat and mat-in-box) or as a single embedded relation (cat-on-mat in box)? A further intriguing case arises when the relation itself is symmetric (e.g., “The table and refrigerator are touching”), leaving no clear cue for assigning Figure versus Reference status, whether in language or in vision. Such cases may be among the rare contexts in which no consistent asymmetry in compositional order emerges (Hafri, Gleitman, et al., 2023). Our approach, using a sentence–picture verification task with delayed presentations, may prove fruitful in addressing these questions.

7.5. Conclusions

In sum, our work reveals a fundamental principle of relational composition: the mind constructs relational representations sequentially, operating over abstract relational roles. By uncovering a Reference-first compositional routine, we show that visual perception employs a structured process akin to compositional principles observed in event cognition and language. More broadly, these findings provide new insights into the nature of visual representation, the underlying construction algorithms, and the deep connections between perception, language, and intuitive physics.

CRedit authorship contribution statement

Zekun Sun: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Chaz Firestone:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Conceptualization. **Alon Hafri:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Software, Methodology, Investigation, Formal analysis, Conceptualization.

Acknowledgments

For helpful discussion and/or comments on drafts of this article, the authors thank Rebecca Tollan and members of the University of Delaware Perception & Language Laboratory. This work was supported by National Science Foundation Grant No. BCS-2021053 awarded to C. Firestone and National Science Foundation Directorate for Social, Behavioral and Economic Sciences (SBE) Postdoctoral Research Fellowship No. SMA-2105228 awarded to A. Hafri.

Data availability

An archive of the data, code, stimuli, preregistrations, and other relevant materials is available at: <https://osf.io/vzxdg>.

References

- Ayzenberg, V., & Lourenco, S. (2022). Perception of an object's global shape is best described by a model of skeletal structure in human infants. *Elife*, 11, Article e74943. <http://dx.doi.org/10.7554/eLife.74943>.
- Baillargeon, R., Stavans, M., Wu, D., Gertner, Y., Setoh, P., Kittredge, A. K., et al. (2012). Object individuation and physical reasoning in infancy: An integrative account. *Language Learning and Development*, 8(1), 4–46. <http://dx.doi.org/10.1080/15475441.2012.630610>.
- Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 110(45), 18327–18332. <http://dx.doi.org/10.1073/pnas.1306572110>.
- Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94(2), 115. <http://dx.doi.org/10.1037//0033-295x.94.2.115>.
- Block, N. (2023). *The border between seeing and thinking*. Oxford University Press, <http://dx.doi.org/10.1093/oso/9780197622223.001.0001>.
- Boettcher, S. E., Draschkow, D., Dienhart, E., & Vö, M. L.-H. (2018). Anchoring visual search in scenes: Assessing the role of anchor objects on eye movements during visual search. *Journal of Vision*, 18(13), 11. <http://dx.doi.org/10.1167/18.13.11>.
- Bonnen, T., Wagner, A. D., & Yamins, D. L. K. (2025). Medial temporal cortex supports object perception by integrating over visuospatial sequences. *Cognition*, 262(106135), Article 106135. <http://dx.doi.org/10.1016/j.cognition.2025.106135>.
- Brocard, S., Wilson, V. A., Berton, C., Zuberbühler, K., & Bickel, B. (2024). A universal preference for animate agents in hominids. *iScience*, 27(6), <http://dx.doi.org/10.1016/j.isci.2024.109996>.
- Burge, T. (2022). *Perception: First form of mind*. Oxford University Press, <http://dx.doi.org/10.1093/oso/9780198871002.001.0001>.
- Carey, S. (2009). *The origin of concepts*. Oxford University Press, <http://dx.doi.org/10.1093/acprof:oso/9780195367638.001.0001>.
- Cavanagh, P. (2021). The language of vision. *Perception*, 50(3), 195–215. <http://dx.doi.org/10.1177/0301006621991491>.
- Chen, L. (1982). Topological structure in visual perception. *Science*, 218(4573), 699–700. <http://dx.doi.org/10.1126/science.7134969>.
- Chomsky, N. (1995). *The MIT press, The minimalist program*. London, England: MIT Press, <http://dx.doi.org/10.7551/mitpress/9780262527347.001.0001>.
- Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, 31(5), 489–509. <http://dx.doi.org/10.1017/S0140525X08004998>.
- Christiansen, M. H., & Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, 39, Article e62. <http://dx.doi.org/10.1017/S0140525X1500031x>.
- Clark, H. H., & Chase, W. G. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology*, 3(3), 472–517. [http://dx.doi.org/10.1016/0010-0285\(72\)90019-9](http://dx.doi.org/10.1016/0010-0285(72)90019-9).
- Clark, H. H., & Chase, W. G. (1974). Perceptual coding strategies in the formation and verification of descriptions. *Memory & Cognition*, 2(1), 101–111. <http://dx.doi.org/10.3758/bf03197499>.
- Cohn, N., & Paczynski, M. (2013). Prediction, events, and the advantage of agents: The processing of semantic roles in visual narrative. *Cognitive Psychology*, 67(3), 73–97. <http://dx.doi.org/10.1016/j.cogpsych.2013.07.002>.
- Cohn, N., Paczynski, M., & Kutas, M. (2017). Not so secret agents: Event-related potentials to semantic roles in visual event comprehension. *Brain and Cognition*, 119, 1–9. <http://dx.doi.org/10.1016/j.bandc.2017.09.001>.
- De Freitas, J., & Alvarez, G. A. (2018). Your visual system provides all the information you need to make moral judgments about generic visual events. *Cognition*, 178, 133–146. <http://dx.doi.org/10.1016/j.cognition.2018.05.017>.
- Destler, N., Singh, M., & Feldman, J. (2023). Skeleton-based shape similarity. *Psychological Review*, 130(6), 1653–1671. <http://dx.doi.org/10.1037/rev0000412>.
- Dowty, D. (1991). Thematic proto-roles and argument selection. *Language*, 67(3), 547. <http://dx.doi.org/10.2307/415037>.
- Draschkow, D., & Vö, M. L. H. (2017). Scene grammar shapes the way we interact with objects, strengthens memories, and speeds search. *Scientific Reports*, 7, 16471. <http://dx.doi.org/10.1038/s41598-017-16739-x>.
- El-Gaaly, T., Froyen, V., Elgammal, A., Feldman, J., & Singh, M. (2015). A Bayesian approach to perceptual 3D object-part decomposition using skeleton-based representations. In *Proceedings of the AAAI conference on artificial intelligence: Vol. 29*, <http://dx.doi.org/10.1609/aaai.v29i1.9793>.
- Feldman, J., & Singh, M. (2006). Bayesian estimation of the shape skeleton. *Proceedings of the National Academy of Sciences*, 103(47), 18014–18019. <http://dx.doi.org/10.1073/pnas.0608811103>.
- Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive Psychology*, 47(2), 164–203. [http://dx.doi.org/10.1016/S0010-0285\(03\)00005-7](http://dx.doi.org/10.1016/S0010-0285(03)00005-7).
- Firestone, C., & Scholl, B. J. (2014). “Please tap the shape, anywhere you like”: Shape skeletons in human vision revealed by an exceedingly simple measure. *Psychological Science*, 25(2), 377–386. <http://dx.doi.org/10.1177/0956797613507584>.
- Firestone, C., & Scholl, B. J. (2016). Cognition does not affect perception: Evaluating the evidence for “top-down” effects. *Behavioral and Brain Sciences*, 39, Article e229. <http://dx.doi.org/10.1017/S0140525X15000965>.
- Firestone, C., & Scholl, B. (2017). Seeing physics in the blink of an eye. *Journal of Vision*, 17(10), 203. <http://dx.doi.org/10.1167/17.10.203>.
- Fodor, J. A. (1975). *The language of thought*. Harvard University Press.
- Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1–2), 3–71. <http://dx.doi.org/10.7551/mitpress/2103.003.0002>.
- Franconeri, S. L., Scimeca, J. M., Roth, J. C., Helseth, S. A., & Kahn, L. E. (2012). Flexible visual processing of spatial relationships. *Cognition*, 122(2), 210–227. <http://dx.doi.org/10.1167/11.11.268>.
- Gattis, M. (2004). Mapping relational structure in spatial reasoning. *Cognitive Science*, 28(4), 589–610. <http://dx.doi.org/10.1016/j.cogsci.2004.02.001>.
- Gerstenberg, T., Peterson, M. F., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2017). Eye-tracking causality. *Psychological Science*, 28(12), 1731–1744. <http://dx.doi.org/10.1177/0956797617713053>.
- Gleitman, L. R., Gleitman, H., Miller, C., & Ostrin, R. (1996). Similar, and similar concepts. *Cognition*, 58(3), 321–376. [http://dx.doi.org/10.1016/0010-0277\(95\)00686-9](http://dx.doi.org/10.1016/0010-0277(95)00686-9).
- Gleitman, L. R., January, D., Nappa, R., & Trueswell, J. C. (2007). On the give and take between event apprehension and utterance formulation. *Journal of Memory and Language*, 57(4), 544–569. <http://dx.doi.org/10.1016/j.jml.2007.01.007>.
- Goldwater, M. B., & Gentner, D. (2015). On the acquisition of abstract knowledge: Structural alignment and explication in learning causal system categories. *Cognition*, 137, 137–153. <http://dx.doi.org/10.1016/j.cognition.2014.12.001>.
- Hafri, A. (2024). Cognitive development: The origins of structured thought in the mind. *Current Biology*, 34(18), R856–R859. <http://dx.doi.org/10.1016/j.cub.2024.07.096>.
- Hafri, A., Bonner, M. F., Landau, B., & Firestone, C. (2024). A phone in a basket looks like a knife in a cup: Role-filler independence in visual processing. *Open Mind*, 8, 766–794. <http://dx.doi.org/10.31234/osf.io/jx4yg>.
- Hafri, A., & Firestone, C. (2021). The perception of relations. *Trends in Cognitive Sciences*, 25(6), 475–492. <http://dx.doi.org/10.1016/j.tics.2021.01.006>.
- Hafri, A., Gleitman, L. R., Landau, B., & Trueswell, J. C. (2023). Where word and world meet: Language and vision share an abstract representation of symmetry. *Journal of Experimental Psychology: General*, 152(2), 509–527. <http://dx.doi.org/10.1037/xge0001283>.
- Hafri, A., Green, E., & Firestone, C. (2023). Compositionality in visual perception. *Behavioral and Brain Sciences*, 46(E277), <http://dx.doi.org/10.1017/S0140525X23001838>.

- Hafri, A., Papafragou, A., & Trueswell, J. C. (2013). Getting the gist of events: Recognition of two-participant actions from brief displays. *Journal of Experimental Psychology: General*, 142(3), 880. <http://dx.doi.org/10.1037/a0030045>.
- Hafri, A., & Papeo, L. (2025). The past, present, and future of relation perception. *Journal of Experimental Psychology: Human Perception and Performance*, 51(5), 543. <http://dx.doi.org/10.1037/xhp0001310>.
- Hafri, A., Trueswell, J. C., & Strickland, B. (2018). Encoding of event roles from visual scenes is rapid, spontaneous, and interacts with higher-level visual processing. *Cognition*, 175, 36–52. <http://dx.doi.org/10.1016/j.cognition.2018.02.011>.
- Hespos, S. J., & Spelke, E. S. (2004). Conceptual precursors to language. *Nature*, 430(6998), 453–456. <http://dx.doi.org/10.1038/nature02634>.
- Hochmann, J. R., & Papeo, L. (2021). How can it be both abstract and perceptual? Comment on Hafri, A., & Firestone, C. (2021) the perception of relations, trends in cognitive sciences. PsyArXiv, URL https://osf.io/preprints/psyarxiv/hm49p_v1.
- Holcombe, A. O., Linares, D., & Vaziri-Pashkam, M. (2011). Perceiving spatial relations via attentional tracking and shifting. *Current Biology*, 21(13), 1135–1139. <http://dx.doi.org/10.1016/j.cub.2011.05.031>.
- Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review*, 110(2), 220. <http://dx.doi.org/10.1037/0033-295x.110.2.220>.
- Huttenlocher, J., Eisenberg, K., & Strauss, S. (1968). Comprehension: Relation between perceived actor and logical subject. *Journal of Memory and Language*, 7(2), 527. [http://dx.doi.org/10.1016/s0022-5371\(68\)80044-1](http://dx.doi.org/10.1016/s0022-5371(68)80044-1).
- Huttenlocher, J., & Straus, S. (1968). Comprehension and a statement's relation to the situation it describes. *Journal of Memory and Language*, 7(2), 300. [http://dx.doi.org/10.1016/s0022-5371\(68\)80005-2](http://dx.doi.org/10.1016/s0022-5371(68)80005-2).
- Isasi-Isasmendi, A., Andrews, C., Flecken, M., Laka, I., Daum, M. M., Meyer, M., et al. (2023). The agent preference in visual event apprehension. *Open Mind*, 7, 240–282. http://dx.doi.org/10.1162/opmi_a.00083.
- Jackendoff, R. (1987). On beyond zebra: The relation of linguistic and visual information. *Cognition*, 26(2), 89–114. [http://dx.doi.org/10.1016/0010-0277\(87\)90026-6](http://dx.doi.org/10.1016/0010-0277(87)90026-6), URL <http://www.sciencedirect.com/science/article/pii/0010027787900266>.
- Jackendoff, R. (2002). *Foundations of language: Brain, meaning, grammar, evolution*. Oxford University Press, <http://dx.doi.org/10.1093/acprof:oso/9780198270126.001.0001>.
- Jamrozik, A., & Gentner, D. (2015). Well-hidden regularities: Abstract uses of in and on retain an aspect of their spatial meaning. *Cognitive Science*, 39(8), 1881–1911. <http://dx.doi.org/10.1111/cogs.12218>.
- Ji, H., & Scholl, B. J. (2024). “Visual verbs”: Dynamic event types are extracted spontaneously during visual perception.. *Journal of Experimental Psychology: General*, 153(10), 2441. <http://dx.doi.org/10.1037/xge0001636>.
- Johannes, K., Wilson, C., & Landau, B. (2016). The importance of lexical verbs in the acquisition of spatial prepositions: The case of in and on. *Cognition*, 157, 174–189. <http://dx.doi.org/10.1016/j.cognition.2016.08.022>.
- Jolicoeur, P., Ullman, S., & Mackay, M. (1986). Curve tracing: A possible basic operation in the perception of spatial relations. *Memory & Cognition*, 14, 129–140. <http://dx.doi.org/10.3758/bf03198373>.
- Jolicoeur, P., Ullman, S., & Mackay, M. (1991). Visual curve tracing properties. *Journal of Experimental Psychology: Human Perception and Performance*, 17(4), 997–1022. <http://dx.doi.org/10.1037/0096-1523.17.4.997>.
- Kaiser, D., Quek, G. L., Cichy, R. M., & Peelen, M. V. (2019). Object vision in a structured world. *Trends in Cognitive Sciences*, 23(8), 672–685. <http://dx.doi.org/10.1016/j.tics.2019.04.013>, URL <https://linkinghub.elsevier.com/retrieve/pii/S1364661319301056>.
- Knowlton, T., Hunter, T., Odic, D., Wellwood, A., Halberda, J., Pietroski, P., et al. (2021). Linguistic meanings as cognitive instructions. *Annals of the New York Academy of Sciences*, 1500(1), 134–144. <http://dx.doi.org/10.1111/nyas.14618>.
- Kominsky, J. F., Strickland, B., Wertz, A. E., Elsner, C., Wynn, K., & Keil, F. C. (2017). Categories and constraints in causal perception. *Psychological Science*, 28(11), 1649–1662. <http://dx.doi.org/10.1177/0956797617719930>.
- Kosslyn, S. M., Thompson, W. L., & Ganis, G. (2006). *The case for mental imagery*. Oxford University Press, <http://dx.doi.org/10.1093/acprof:oso/9780195179088.003.0001>.
- Kubricht, J. R., Holyoak, K. J., & Lu, H. (2017). Intuitive physics: Current research and controversies. *Trends in Cognitive Sciences*, 21(10), 749–759. <http://dx.doi.org/10.1016/j.tics.2017.06.002>.
- Landau, B., & Gleitman, L. R. (2015). Height matters. In *Structures in the mind: Essays on language, music, and cognition in honor of ray jackendoff*. MIT Press, <http://dx.doi.org/10.7551/mitpress/10044.003.0014>.
- Landau, B., & Jackendoff, R. (1993). “What” and “where” in spatial language and spatial cognition. *Behavioral and Brain Sciences*, 16(2), 217–238. <http://dx.doi.org/10.1017/s0140525x00029733>.
- Lande, K. (2023). Contours of vision: Towards a compositional semantics of perception. *British Journal for the Philosophy of Science*, <http://dx.doi.org/10.1086/725094>.
- Langley, M. D., & McBeath, M. K. (2023). Vertical attention bias for tops of objects and bottoms of scenes. *Journal of Experimental Psychology: Human Perception and Performance*, 49(10), 1281. <http://dx.doi.org/10.1037/xhp0001117>.
- Lescroart, M. D., & Biederman, I. (2013). Cortical representation of medial axis structure. *Cerebral Cortex*, 23(3), 629–637. <http://dx.doi.org/10.1093/cercor/bhs046>.
- Levinson, S. C. (2003). *Space in language and cognition: Explorations in cognitive diversity: Vol. 5*, Cambridge University Press, <http://dx.doi.org/10.1017/cbo9780511613609>.
- Lewis, M. L., & Frank, M. C. (2016). The length of words reflects their conceptual complexity. *Cognition*, 153, 182–195. <http://dx.doi.org/10.1016/j.cognition.2016.04.003>.
- Little, P. C., & Firestone, C. (2021). Physically implied surfaces. *Psychological Science*, 32(5), 799–808. <http://dx.doi.org/10.1177/0956797620939942>.
- Long, B., Konkle, T., Cohen, M. A., & Alvarez, G. A. (2016). Mid-level perceptual features distinguish objects of different real-world sizes. *Journal of Experimental Psychology: General*, 145(1), 95. <http://dx.doi.org/10.1037/xge0000130>.
- Long, B., Yu, C. P., & Konkle, T. (2018). Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proceedings of the National Academy of Sciences*, 115(38), E9015–E9024. <http://dx.doi.org/10.1073/pnas.1719616115>.
- Lovett, A., & Franconeri, S. L. (2017). Topological relations between objects are categorically coded. *Psychological Science*, 28(10), 1408–1418. <http://dx.doi.org/10.1177/095679761709814>.
- Lowet, A. S., Firestone, C., & Scholl, B. J. (2018). Seeing structure: Shape skeletons modulate perceived similarity. *Attention, Perception, & Psychophysics*, 80, 1278–1289. <http://dx.doi.org/10.3758/s13414-017-1457-8>.
- Mandelbaum, E., Dunham, Y., Feiman, R., Firestone, C., Green, E., Harris, D., et al. (2022). Problems and mysteries of the many languages of thought. *Cognitive Science*, 46(12), Article e13225. <http://dx.doi.org/10.1111/cogs.13225>.
- Marr, D., & Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 200(1140), 269–294. <http://dx.doi.org/10.1098/rspb.1978.0020>.
- Miller, G. A., & Johnson-Laird, P. N. (1976). *Language and perception*. Harvard University Press, <http://dx.doi.org/10.1037/11135-003>.
- Momma, S., & Phillips, C. (2018). The relationship between parsing and generation. *Annual Review of Linguistics*, 4(1), 233–254. <http://dx.doi.org/10.1146/annurev-linguistics-011817-045719>.
- Papeo, L. (2020). Twos in human visual perception. *Cortex*, 132, 473–478. <http://dx.doi.org/10.31219/osf.io/yuxza>.

- Papeo, L., Vettori, S., Serraille, E., Odin, C., Rostami, F., & Hochmann, J.-R. (2024). Abstract thematic roles in infants' representation of social events. *Current Biology*, <http://dx.doi.org/10.1016/j.cub.2024.07.081>.
- Peelen, M. V., Berlot, E., & de Lange, F. P. (2024). Predictive processing of scenes and objects. *Nature Reviews Psychology*, 3(1), 13–26. <http://dx.doi.org/10.1038/s44159-023-00254-0>.
- Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, 70, 153–163. <http://dx.doi.org/10.1016/j.jesp.2017.01.006>.
- Pietroski, P. M. (2011). Minimal semantic instructions. In B. Cedric (Ed.), *Oxford handbook of linguistic minimalism* (pp. 472–498). Oxford University Press, <http://dx.doi.org/10.1093/oso/9780198812722.003.0008>.
- Quilty-Dunn, J. (2020). Concepts and predication from perception to cognition. *Philosophical Issues*, 30(1), 273–292. <http://dx.doi.org/10.1111/phis.12185>.
- Quilty-Dunn, J., Porot, N., & Mandelbaum, E. (2023). The best game in town: The reemergence of the language-of-thought hypothesis across the cognitive sciences. *Behavioral and Brain Sciences*, 46, Article e261. <http://dx.doi.org/10.1017/s0140525x22002849>.
- Roth, J., & Franconeri, S. (2012). Asymmetric coding of categorical spatial relations in both language and vision. *Frontiers in Psychology*, 3, 464. <http://dx.doi.org/10.3389/fpsyg.2012.00464>.
- Sauppe, S., & Flecken, M. (2021). Speaking for seeing: Sentence structure guides visual event apprehension. *Cognition*, 206, Article 104516. <http://dx.doi.org/10.1016/j.cognition.2020.104516>.
- Segalowitz, N. S. (1982). The perception of semantic relations in pictures. *Memory & Cognition*, 10(4), 381–388. <http://dx.doi.org/10.3758/bf03202430>.
- Strickland, B. (2017). Language reflects “core” cognition: A new theory about the origin of cross-linguistic regularities. *Cognitive Science*, 41(1), 70–101. <http://dx.doi.org/10.1111/cogs.12332>, URL <http://doi.wiley.com/10.1111/cogs.12332>.
- Sun, Z., & Firestone, C. (2021). Curious objects: How visual complexity guides attention and engagement. *Cognitive Science*, 45(4), Article e12933. <http://dx.doi.org/10.1111/cogs.12933>.
- Sun, Z., & Firestone, C. (2022). Beautiful on the inside: Aesthetic preferences and the skeletal complexity of shapes. *Perception*, 51(12), 904–918. <http://dx.doi.org/10.1177/03010066221124872>.
- Talmy, L. (1975). Figure and ground in complex sentences. In *Annual meeting of the Berkeley Linguistics Society* (pp. 419–430). <http://dx.doi.org/10.3765/bls.v1i0.2322>.
- Talmy, L. (1983). How language structures space. In *Spatial orientation: Theory, research, and application* (pp. 225–282). Springer, http://dx.doi.org/10.1007/978-1-4615-9325-6_11.
- Tatone, D., & Csibra, G. (2024). The representation of giving actions: Event construction in the service of monitoring social relationships. *Current Directions in Psychological Science*, 33(3), 159–165. <http://dx.doi.org/10.1177/09637214241242460>, URL <https://journals.sagepub.com/doi/10.1177/09637214241242460>.
- Trueswell, J. C., & Papafragou, A. (2010). Perceiving and remembering events cross-linguistically: Evidence from dual-task paradigms. *Journal of Memory and Language*, 63(1), 64–82. <http://dx.doi.org/10.1016/j.jml.2010.02.006>.
- Ullman, S. (1987). Visual routines. In *Readings in computer vision* (pp. 298–328). Elsevier, <http://dx.doi.org/10.1364/oam.1986.thc4>.
- Ullman, T. D., Spelke, E., Battaglia, P., & Tenenbaum, J. B. (2017). Mind games: Game engines as an architecture for intuitive physics. *Trends in Cognitive Sciences*, 21(9), 649–665. <http://dx.doi.org/10.1016/j.tics.2017.05.012>.
- Ünal, E., Wilson, F., Trueswell, J., & Papafragou, A. (2024). Asymmetries in encoding event roles: Evidence from language and cognition. *Cognition*, 250, Article 105868. <http://dx.doi.org/10.1016/j.cognition.2024.105868>.
- Van Tonder, G. J., Lyons, M. J., & Ejima, Y. (2002). Visual structure of a Japanese zen garden. *Nature*, 419(6905), 359–360. <http://dx.doi.org/10.1038/419359a>.
- Vettori, S., Hochmann, J. R., & Papeo, L. (2024). Fast and automatic processing of relations: The case of containment and support. *Journal of Vision*, 24(10), 840. <http://dx.doi.org/10.1167/jov.24.10.840>.
- Vettori, S., Odin, C., Hochmann, J.-R., & Papeo, L. (2025). A perceptual cue-based mechanism for automatic assignment of thematic agent and patient roles. *Journal of Experimental Psychology: General*, 154(3), 787–798. <http://dx.doi.org/10.1037/xge0001657>.
- Vö, M. L. H. (2021). The meaning and structure of scenes. *Vision Research*, 181, 10–20. <http://dx.doi.org/10.1016/j.visres.2020.11.003>.
- Vö, M. L. H., Boettcher, S. E., & Draschkow, D. (2019). Reading scenes: How scene grammar guides attention and aids perception in real-world environments. *Current Opinion in Psychology*, 29, 205–210. <http://dx.doi.org/10.1016/j.copsyc.2019.03.009>.
- Webb, T., Fu, S., Bihl, T., Holyoak, K. J., & Lu, H. (2023). Zero-shot visual reasoning through probabilistic analogical mapping. *Nature Communications*, 14(1), 5144. <http://dx.doi.org/10.1038/s41467-023-40804-x>.
- Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9), 1526–1541. <http://dx.doi.org/10.1038/s41562-023-01659-w>.
- Wilder, J., Feldman, J., & Singh, M. (2011). Superordinate shape classification using natural shape statistics. *Cognition*, 119(3), 325–340. <http://dx.doi.org/10.1016/j.cognition.2011.01.009>.
- Wilson, V. A., Zuberbühler, K., & Bickel, B. (2022). The evolutionary origins of syntax: Event cognition in nonhuman primates. *Science Advances*, 8(25), eabn8464. <http://dx.doi.org/10.1126/sciadv.abn8464>.
- Wolff, P., & Song, G. (2003). Models of causation and the semantics of causal verbs. *Cognitive Psychology*, 47(3), 276–332. [http://dx.doi.org/10.1016/s0010-0285\(03\)00036-7](http://dx.doi.org/10.1016/s0010-0285(03)00036-7).
- Wong, K. W., & Scholl, B. J. (2024). Spontaneous path tracing in task-irrelevant mazes: Spatial affordances trigger dynamic visual routines. *Journal of Experimental Psychology: General*, 153(9), 2230. <http://dx.doi.org/10.1037/xge0001618>.
- Zwitserslood, P., Bölte, J., Hofmann, R., Meier, C. C., & Döbel, C. (2018). Seeing for speaking: Semantic and lexical information provided by briefly presented, naturalistic action scenes. *PLoS One*, 13(4), Article e0194762. <http://dx.doi.org/10.1371/journal.pone.0194762>.