AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General

ISSN: 0096-3445

2024, Vol. 153, No. 10, 2441–2453
https://doi.org/10.1037/xge0001636

"Visual Verbs": Dynamic Event Types Are Extracted Spontaneously During Visual Perception

Huichao Ji[1] and Brian J. Scholl[1, 2]
[1] Department of Psychology, Yale University
[2] Wu Tsai Institute, Yale University

During visual processing, input that is continuous in space and time is segmented, resulting in the representation of discrete tokens—objects or events. And there has been a great deal of research about how object representations are generalized into types—as when we see an object as an instance of a broader category (e.g., an animal or plant). There has been much less attention, however, to the possibility that vision represents dynamic information in terms of a small number of primitive *event types* (such as twisting or bouncing). (In models that posit a "language of vision," these would be the foundational visual verbs.) Here we ask whether such event types are extracted spontaneously during visual perception, even when entirely task irrelevant during passive viewing. We exploited the phenomenon of categorical perception—wherein differences are more readily noticed when they are represented in terms of different underlying categories. Observers were better at detecting changes to images or short videos when the changes involved switches in the underlying event type—even when the changes that maintained the same event type were objectively larger (in terms of both brute image metrics and higher level feature change). We observed this categorical "cross-event-type" advantage for visual working memory for *twisting* versus *rotating*, *scooping* versus *pouring*, and *rolling* versus *bouncing*. Moreover, additional control experiments confirmed that such effects could not be explained by appeal to lower-level non-categorical stimulus differences. This spontaneous perception of "visual verbs" might promote both generalization and prediction about how events are likely to unfold.

Public Significance Statement

Suppose that you see two photographs or short movies—each presented for only a moment, one after the other—and then you must indicate whether they were identical or not. How would you do this? What sorts of differences would your mind automatically latch onto? Your ability to notice changes might be controlled by the low-level properties of the stimuli—the precise placements of colored pixels and objects here and there. In contrast, this project shows that your ability to notice changes also depends on how your mind automatically and spontaneously *categorizes* them into distinct "event types"—such as *bouncing*, *rolling*, *pouring*, or *scooping*—with differences that involve changes in perceived event types being easier to notice (controlling for lower level image properties). This shows how perception involves representations of the world in terms of a foundational set of "visual verbs," which might be a core part of scene understanding.

Keywords: event types, event perception, categorical perception, change detection

Supplemental materials: https://doi.org/10.1037/xge0001636.supp

Karen Schloss served as action editor.

Brian J. Scholl https://orcid.org/0000-0003-3610-0890

This project was supported by the Office of Naval Research Global MURI Grant N00014-16-1-2007 awarded to Brian J. Scholl. This work was presented at the 2023 meeting of the *Vision Sciences Society*. For helpful conversation and/or comments on previous drafts, we thank Brent Strickland and the members of the Yale Perception and Cognition Laboratory.

Huichao Ji played a lead role in data curation, formal analysis, investigation, software, and writing–original draft, a supporting role in validation, and an equal role in conceptualization and writing–review and editing. Brian J. Scholl played a lead role in funding acquisition, project administration, resources, supervision, and validation and an equal role in conceptualization and writing–review and editing.

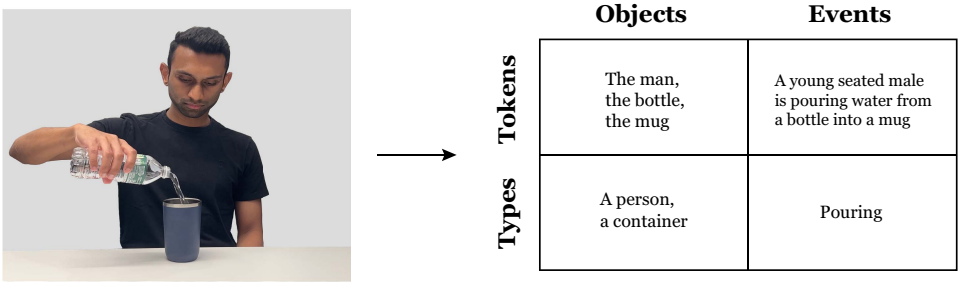
Correspondence concerning this article should be addressed to Huichao Ji or Brian J. Scholl, Department of Psychology, Yale University, Box 208047, New Haven, CT 06520-8047, United States. Email: huichao.ji@yale.edu or brian.scholl@yale.edu

Perception provides the external input to our mental lives, and perhaps the most salient feature of this input is that it is initially *continuous*—for example, when vision starts with an undivided array of activation across space and time on the retina. A central lesson of cognitive science in recent decades, however, has been that the underlying units of many perceptual and cognitive processes are intrinsically *discrete*.

A 2 × 2 Matrix: Objects and Events, Tokens and Types

Discrete perceptual representations seem to come in a sort of 2 × 2 matrix, as shown in Figure 1. One of the dimensions of this matrix stems from the initially continuous dimension itself: *objects* (segmented in space) or *events* (segmented in time). And the other dimension stems from the degree of specificity versus generalization: *tokens* (of particular entities at particular times and/or places)

Figure 1
Categorizing Discrete Representations



Note. Left panel: a person pouring water from a bottle into a mug. Right panel: a 2 × 2 matrix of different types of discrete representations that could be formed when viewing such a scene. See the online article for the color version of this figure.

and *types* (representing broader categories of possible tokens). Three of the cells of this matrix have received a great deal of study by vision scientists in recent years, but the fourth has received less attention.

Perhaps the most familiar cell of this 2 × 2 matrix involves *object tokens*. The way in which particular bits of spatial stimulation are packaged into individual objects has many important downstream consequences for other types of perceptual and cognitive processing. Attention, for example, appears to be irresistibly object based in many contexts, as when it is easier to select two probes when they lie on the same object, compared with different objects, even equating spatial distance (e.g., Duncan, 1984; Egly et al., 1994; for reviews see Cavanagh et al., 2023; Scholl, 2001). And memory, to take another example, may sometimes be limited not primarily in terms of the total number of features encoded but rather by the number of distinct objects into which those features are distributed (e.g., Luck & Vogel, 1997; Ngiam et al., 2024; Vogel & Machizawa, 2004; for a review of the active debate about such issues, see Suchow et al., 2014).

At the same time, there has also been a great deal of research on *object types*—as when we encode an entity as a member of a broader category—such as an animal or a plant. In this cell of the matrix, there has been much work both on how such generalization occurs and on the consequences for downstream processing. For example, many researchers have explored just how the visual system identifies particular entities as being *animals* (vs. inanimate objects, e.g., Chao et al., 1999; Downing et al., 2006; Wiggett et al., 2009), and others have explored how entities that are recognized as tokens of this "animate" type attract more attention (e.g., Calvillo & Hawkins, 2016; J. New et al., 2007, 2010) and are remembered better (e.g., Bonin et al., 2014; Loucks et al., 2020; van Buren & Scholl, 2017). And other work has also explored such questions in the context of many other object types—such as *indoor scenes* (e.g., Fuentemilla et al., 2010; Li et al., 2007), and *food* (e.g., Cunningham & Egeth, 2018; Khosla et al., 2022; Pennock et al., 2023).

Switching from space to time, there has also been a wealth of research in recent years on the nature and importance of *event tokens*. Continuous streams of visual input in time are automatically segmented into discrete events (for reviews see Kurby & Zacks, 2008; Richmond et al., 2017; Zacks, 2020), and these event tokens also go on to have many downstream consequences. Take, for example, the same two cognitive processes discussed above: attention and memory. Attention appears to automatically wax and wane as a function of event segmentation, such that targets are more difficult to detect (and the frequency of inattentional blindness is greater) at the boundaries between events (e.g., Huff et al., 2012; Levin & Varakin, 2004). In addition, memory is heavily influenced by whether information is presented within events or at their boundaries (e.g., Sargent et al., 2013; Swallow et al., 2009). Perhaps most famously, when working memory is assessed after a walk down a long hallway, performance is worse when people pass through a doorway (as a type of event boundary) during their walk, equating time and distance traveled (e.g., Ongchoco et al., 2023; Radvansky & Copeland, 2006; for reviews see Radvansky, 2012; Radvansky & Zacks, 2014).

That leaves just one cell left in the matrix from Figure 1: *event types*—dynamic sequences that are both temporally segmented and also encoded as members of a broader category. This cell of the matrix is rather curious. On one hand, we can certainly distinguish various event types, especially given their ubiquity in everyday language: *rolling*, *bouncing*, *sliding*, *pouring*, *twisting*, *rotating*, *stirring*, *scooping*, and so forth (and of course such types have certainly received a great deal of study in the context of language, e.g., Bach, 1986; Pustejovsky, 1991). On the other hand, there has been surprisingly little work on such event types in the context of visual perception and memory (see Cavanagh et al., 2001; Strickland & Scholl, 2015). The exceptions to this rule mostly involve extensive study of very particular putative types of events. For example, decades of research have explored the perception of *causal launching*, as in a billiard-ball collision, where the impact of one ball appears to cause the motion of the other (e.g., Michotte, 1963; for a review see Scholl & Tremoulet, 2000). (This may seem like a high-level inference, but recent work has demonstrated that this property is truly detected in lower level perception, as it involves retinotopically specific adaptation; Kominsky & Scholl, 2020; Rolfs et al., 2013). Another possible event type that has been extensively studied over many decades is *walking*, as in the perception of biological motion, where we readily perceive locomotion from surprisingly sparse displays of moving dots (in "point-light walkers"; for reviews see Blake & Shiffrar, 2007; Troje, 2013). And such particular types of events involving agents may be particularly salient when two agents are seen to be *facing* each other (Abassi & Papeo, 2020; Papeo & Abassi, 2019).

The relative lack of work on broad notions of visual event types stands in stark contrast to work in language and higher level cognition. These fields have extensively studied the representational "formats" of events in terms of their abstract properties—for example, being bounded or unbounded (Papafragou et al., 2008), involving both agents and patients (Mecklinger et al., 1995), and being oriented toward goals and away from sources (Lakusta et al., 2017). Work in this domain has shown how such abstract properties (e.g., who acted on whom) are highlighted even when task irrelevant for many different kinds of events (Hafri et al., 2018). And such properties may be similar to how many distinct categories of object representations are still represented in terms of the same underlying set of key properties. For example, shapes may be represented in terms of a key property of being *closed* versus *open* (e.g., Kovács & Julesz, 1993; Marino & Scholl, 2005), just as events may be represented in terms of being *bounded* versus *unbounded*—but in each case, these properties may be common across vastly different putative categorical *types* (e.g., *food* vs. *tools* in the case of objects or *bouncing* vs. *rolling* in the case of events).

Returning to visual perception, however, this "event types" cell of the matrix seems relatively barren in terms of previous research, in at least two ways. First, there has just been very little work on many (or most) of the other most obvious putative examples of event types in the context of vision research (as opposed to event representation in language and higher level cognition). (Consider, e.g., how relatively little work has focused on the visual perception of *scooping* or *stirring*.) Second, there has correspondingly been relatively little work on any possible downstream consequences of representing the world in terms of event types. The current article is the first step in trying to fill these gaps.

The Present Study: Categorical Perception of "Visual Verbs"?

Are there general representations of event types in perception? Despite the paucity of past work in this domain, this seems like an important question to explore, given the seemingly foundational nature of the 2 × 2 matrix of representations discussed above. Putting this point differently, we can consider for a moment that some adventurous models of visual processing liken perception to *language* (e.g., Cavanagh, 2021). Such models note that visual representations must end up interfacing with the rest of the mind, including representations in a "language of thought" (e.g., Quilty-Dunn et al., 2023). Accordingly, such models may distinguish visual "nouns" from "verbs" in a categorical manner—in which case it is the putative event types that would count as the visual verbs (Cavanagh, 2021). The underlying notion here is that different types of visual events are recognized and processed in a category-specific manner—perhaps similar to how visual objects such as food or tools are categorically distinguished (e.g., Almeida et al., 2008; Khosla et al., 2022).

In this sense, how could we study whether there are visual verbs in perception? The essence of this claim, we suggest, lies in the previously discussed notion of categorical *types*. Accordingly, the foundational notion of visual verbs is that dynamic visual information may be automatically and spontaneously represented *categorically*—with representational differences in dynamic visual information depending not on their low-level features but on their underlying categorical "verb." Such representations could also involve or interface with verbal labels but may nevertheless be "visual" insofar as they would be spontaneously activated during simple passive viewing of visual stimuli and would have downstream consequences for other types of visual processing (e.g., influencing which visual details are encoded into visual working memory).

To explore this possibility, the current project asks whether such dynamic event types are spontaneously extracted during visual processing using the phenomenon of categorical perception—wherein two stimuli are more readily discriminated when they are represented in terms of different underlying categories (e.g., Newell & Bülthoff, 2002; for reviews see Goldstone & Hendrickson, 2010; Harnad, 1987). For example, even when equidistant in color space, two stimuli belonging to the same linguistic category (e.g., light blue and dark blue) are discriminated more slowly and less accurately than those from different categories (e.g., blue and green; Bornstein & Korda, 1984; cf. Bae et al., 2015; He et al., 2014). Similarly, people are better able to discriminate faces when they have categorically different expressions, independent of how extreme the actual visual differences are (such that, e.g., distinguishing a happy face from a sad face is easier than distinguishing a slightly happy face from a maximally happy face—even when the latter difference is greater; Etcoff & Magee, 1992).

We adapted this logic to the study of visual event types. Observers viewed pairs of stimuli (either photos of dynamic events or actual animations) and simply had to determine whether they were identical or not. When there was a change, it could involve either a categorical difference (e.g., a switch from pouring liquid to scooping liquid) or a within-category difference (e.g., a switch from one depiction of pouring liquid to another depiction of pouring liquid). If the visual system does indeed represent such stimuli in terms of a foundational set of visual verbs, then changes should be more readily detected when they involve different event types.

This underlying experimental logic has three key features. First, unlike previous work focused on particular putative event types (such as biological motion), these displays always centrally involved the *contrast* between two different event types (such as pouring vs. scooping). We were inspired in this respect by past work that has directly contrasted the categories of *containment* versus *occlusion*, although this work never tested categorical perception, per se (e.g., Baillargeon & Wang, 2002; Hespos & Baillargeon, 2001; Strickland & Scholl, 2015). Second, our observers were never explicitly asked (or instructed) about event types at all. In contrast, some past work in neighboring domains has made such contrasts maximally explicit by asking subjective categorical questions (e.g., "Did you see 'punching'?"; e.g., Hafri et al., 2013). Here, the categorical manipulations were always entirely task irrelevant: Observers merely had to discriminate whether a pair of stimuli were different *in any way*, and we simply measured their objective performance at doing so. This allows us to ask whether any categorical effects are spontaneous, occurring without any instruction or directed attention. Third, by explicit design, the objective magnitudes of the visual changes were always numerically *smaller* in pairs that involved different event types (e.g., pouring vs. scooping) and *greater* in pairs that involved the same event type (e.g., two different depictions of scooping). This was always true in terms of both brute image metrics (number of pixels changed) and higher level visual properties (quantified in terms of the squared Euclidean distance in vectorized feature-activation maps from the penultimate layer [layer fc1] in a

convolutional neural network pretrained for image classification and detection [VGG16; Simonyan & Zisserman, 2015]).

We employed this design in a series of six experiments, exploring three different putative event-type contrasts: *twisting* versus *rotating* (Experiments 1a and 1b), *pouring* versus *scooping* (Experiments 2a and 2b), and *rolling* versus *bouncing* (Experiments 3a and 3b). In each case, we took care to ensure that any differences could not be explained by appeal to lower level noncategorical stimulus differences.

Experiment 1a: Twisting Versus Rotating

Observers saw two images displayed very quickly, one after the other, and simply had to indicate whether the two images were identical or not. When different, the two images came from one of two conditions (depicted in Figure 2a). On within-type trials, the images were different depictions of the same underlying event type—either two distinct images of towels being twisted (as in the left two images in Figure 2a) or two distinct images of towels being rotated (as in the right two images in Figure 2a). On across-type trials, the two images came from different event types (one twisting, one rotating, as in the middle two images in Figure 2a). Would observers be better at detecting across-type changes, even when the objective magnitudes of those changes were smaller than in within-type changes?

Method

Transparency and Openness

The methods and analyses for each experiment were preregistered and can be accessed at https://aspredicted.org/blind.php?x=YXX_WMZ (for Experiments 1a and 1b), https://aspredicted.org/blind.php?x=82X_8D7 (for Experiments 2a and 2b), and https://aspredicted.org/blind.php?x=8QS_N4M (for Experiments 3a and 3b). Data for all experiments are available as online Supplemental Material.

Participants

One hundred observers (48 women, 52 men, average age = 37.69, all fluent English speakers from either the United States or the United Kingdom) participated via the online platform Prolific (Palan & Schitter, 2018) for monetary compensation. This sample size was preregistered before data collection began and was fixed to be identical across all six experiments reported here, which are all of the preregistered experiments that we ran. Gender information was collected via a multiple-choice question, with the options of "male," "female," or "other." Observers were excluded with replacement according to the preregistered criteria if they self-reported an attention level below 80 in a postexperimental debriefing question (where they reported how well they paid attention throughout the experiment on a continuous scale, with 1 being very distracted, and 100 being very focused; $n = 18$) or if their mean response time across all trials was longer than 10 s ($n = 0$).[1]

Apparatus

Stimuli were presented, and data were collected via custom software written in HTML, CSS, JavaScript, PHP, and the JsPsych libraries (de Leeuw, 2015). Observers completed the experiment on either a laptop or desktop computer. Display size, viewing distance, and display resolutions could vary because the experiment was rendered on observers' web browsers, so we report stimulus dimensions below using pixel [px] values.

Stimuli

All text in instructions and postexperimental debriefing questions appeared in white (font size 15px), presented on a black display.

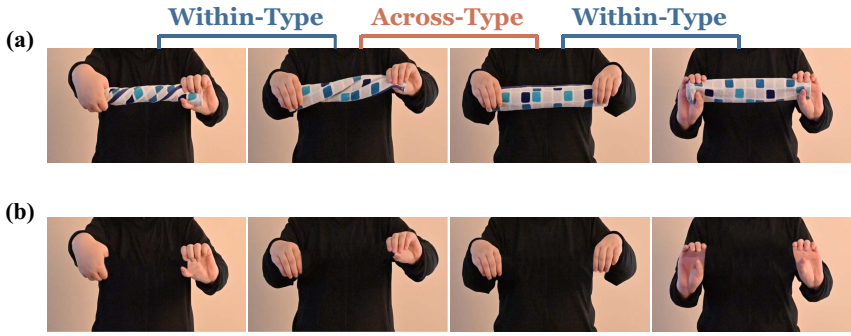
Videos of an individual rotating or twisting a towel were recorded, and we manually selected two frames from each video (cropped to display only a 1108 × 546 central region) as distinct images of the same underlying event type: two images of towels being *twisted* or *rotated* to a lesser or greater extent (Figure 2a). (While it may be possible to twist a towel from almost any starting point, that is surely not equally likely. From the starting point of the leftmost panel in Figure 2a, e.g., it would be nearly impossible to rotate your hands in place along the relevant axis because your right hand cannot go forward any more than it already is, and your left hand cannot go backward any more than it already is.) Image analyses then confirmed that the within-type image pairs (i.e., the leftmost or rightmost pairs of images in Figure 2a) involved a numerically greater degree of visual change than did the across-type image pair (i.e., the central two images in Figure 2a) in two ways. First, although roughly 99.7% of the pixels ended up changing across all images, more pixels changed in the within-type image pairs (average: 604,619; twisting: 604,613; rotating: 604,625) compared with the across-type image pair (twisting vs. rotating: 602,719). Second, the distance in vectorized feature-activation maps from the penultimate layer (fc1) in VGG16 (Simonyan & Zisserman, 2015) was numerically greater for within-type image pairs (average: 167.76; twisting: 143.26; rotating: 192.25) compared with the across-type image pair (143.09).

Procedure and Design

Each trial began with the 250-ms presentation of a white fixation cross (each bar 15px, with a stroke width of 2px) in the center of the display. After a 250-ms blank interval, two images were presented one after another (each for 100 ms, separated by a 1,500-ms blank interval). Each image (sized with a width of 15% of the display width) appeared in a distinct randomly selected quadrant of the display (with the image center jittered by 1/30 of the display width away from the display center). Observers then indicated (using a keypress, in response to a centered prompt 100px above the bottom of the display) whether the two images were identical. Observers completed 64 trials in a different randomized order for each observer: 2 matching possibilities (same vs. different) × 2 change types (within-type vs. across-type) × 16 repetitions. (Of course, the change-type variable did not apply to same-image trials, but we nevertheless include this variable when describing the overall design to ensure that there are an

[1] Following our preregistered analysis plan, we report the results of all experiments in this article after excluding observers based on these criteria. Because the self-reported attention levels ended up excluding a relatively large number of observers (as is common in such online studies), however, we also replicated all of the analyses on the full population of observers, without excluding any observers at all. And as detailed in our online data file, these supplementary analyses reproduced the relevant effects in every case (such that all significant effects with exclusions remained significant without exclusion and ditto for null effects).

Figure 2
Stimuli From Experiments 1a and 1b



Note. (a) In Experiment 1a, a towel is twisted or rotated. (b) In Experiment 1b, the same hand positions were preserved without the towel. See the online article for the color version of this figure.

equal number of same-image trials [32] vs. different-image trials [16 across-type trials and 16 within-type trials].) For the 32 no-change trials, one of the images was simply presented twice—with each of the four images used in eight different trials. The 16 across-type trials each used the central two images in Figure 2a (in a randomized order). Eight of the 16 within-type trials used the leftmost two images in Figure 2a (in a randomized order), and the other eight used the rightmost two images in Figure 2a (in a randomized order).

Results and Discussion

According to the preregistered criteria, we excluded individual trials whose response times were greater than 3 *SD* away from the mean response time for all observers (0.83/64 on average), or shorter than 200 ms (0.17/64 on average). As depicted in the top of Figure 3a, changes on across-type trials were detected more accurately than were changes on within-type trials, 80.93% versus 58.68%, $t(99) = 10.48$, $p < .001$, $d = 1.05$—and as depicted in the bottom of Figure 3a, this was true for the vast majority (79%) of individual observers. (Because half of the trials were no-change trials, we were able to focus on accuracy. These same patterns all hold with signal detection, measuring d'. The results of these analyses are included in the online Supplemental Material.) Examination of the response times confirmed that there was no speed–accuracy trade-off because the response times actually differed in the same direction, 925 ms versus 1,013 ms, $t(99) = 5.62$, $p < .001$, $d = 0.56$. These initial results are consistent with the spontaneous representation of visual event types: Observers were better able to distinguish depictions of twisting versus rotating, even when those changes were objectively smaller than changes within a given event type (two examples of twisting or two examples of rotating).

Experiment 1b: Twisting Versus Rotating Control

The contrast between twisting and rotating in Experiment 1a was implemented in an especially straightforward fashion, as might be frequently observed in real-world experience—simply by having a person twist or rotate a towel in their hands. This raises the possibility, though, that the results could simply reflect this difference in hand positions, independent of the dynamic event type being applied to the towel. To ensure that this could not explain our results, we replicated Experiment 1a with only the hands themselves depicted (as in Figure 2b)—now predicting that the difference observed in Experiment 1a should disappear.

Method

This experiment was identical to Experiment 1a except as noted. A new set of 100 observers (42 women, 57 men, 1 other, average age = 36.34) participated. Observers were excluded with replacement according to the same preregistered criteria ($n = 24$ and $n = 0$, respectively). Stimuli were identical to those used in Experiment 1a, except that the towels in the images were removed, as depicted in Figure 2b (implemented via the "Content-Aware Fill" tool in Photoshop, Version 23.5.2).

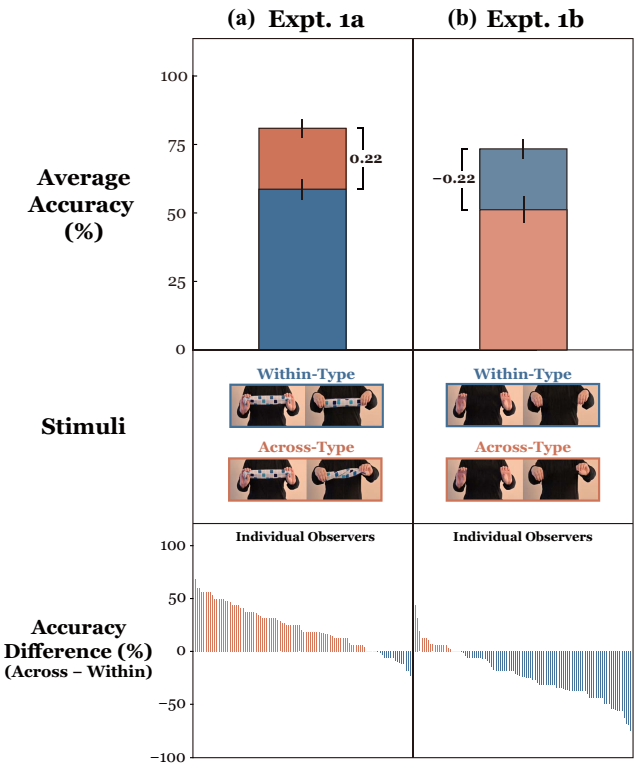
Results and Discussion

We again excluded individual trials whose response times were greater than 3 *SD* away from the mean response time for all observers (0.99/64) or shorter than 200 ms (0.69/64). As depicted in the top of Figure 3b, changes on across-type trials were detected less accurately than changes on within-type trials, 51.17% versus 73.43%, $t(99) = 9.70$, $p < .001$, $d = 0.97$—and as depicted in the bottom of Figure 3b, this was true for the vast majority (79%) of individual observers. Examination of the response times confirmed that there was no speed–accuracy trade-off, 1,051 ms versus 1,027 ms, $t(99) = 1.52$, $p = .132$, $d = 0.15$. Moreover, a cross-experiment comparison revealed that the accuracy difference (across-minus-within) in this experiment was significantly different from the effect in Experiment 1a, 22.26% smaller versus 22.25% larger, $t(198) = 14.24$, $p < .001$, $d = 2.01$. These results confirm that the cross-event-type advantage observed in Experiment 1a cannot be explained simply by appeal to the differing hand positions; indeed, these results unexpectedly suggest that any effect of hand position is in the opposite direction, suggesting that the results of Experiment 1 may even underestimate the true categorical effect of event types per se.

Experiment 2a: Scooping Versus Pouring

We interpreted the results of Experiments 1a and 1b in terms of a cross-event-type advantage rather than in terms of any specific properties of twisting versus rotating. To explore whether this effect generalizes, we next conducted the same type of change detection

Figure 3
Results From (a) Experiment 1a and (b) Experiment 1b



Note. The top panels depict the average accuracy for within-type versus across-type trials. The bottom panels depict the magnitude of the across-minus-within difference separately for each observer (with observers who performed better for across-type trials depicted in orange and those who performed better for within-type trials depicted in blue.) Error bars indicate 95% confidence intervals. See the online article for the color version of this figure.

study but now using depictions of *scooping* versus *pouring* (as depicted in Figure 4a). This new contrast has the possible added advantage of being less readily describable in terms of adjective-like states: The relevant manipulation from Experiments 1a and 1b could perhaps be described in terms of the "twistedness" of the towel (in part because you might easily pause mid-twist), but there is no simple analogue of "scoopedness" or "pouredness" (in part because one cannot freeze mid-pour in the same manner). Would observers also be better at detecting across-type changes between these event types (as in the middle two images in Figure 4a) compared with within-type changes (two distinct images of scooping, as in the left two images in Figure 4a; or two distinct images of pouring, as in the right two images in Figure 4a)? We tested this while again ensuring that the objective degree of change was always greater for within-type changes.

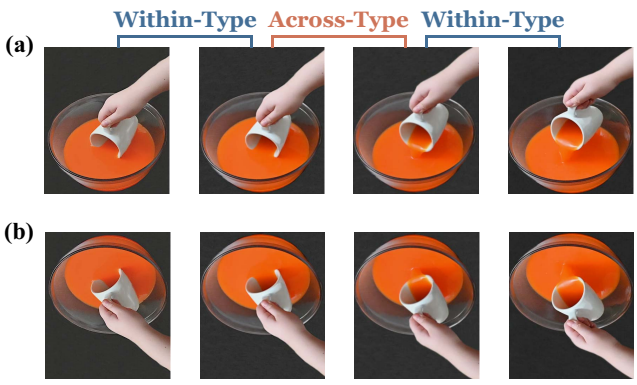
Method

This experiment was identical to Experiment 1a except as noted. A new set of 100 observers (52 women, 45 men, 3 other, average age = 38.91) participated. Observers were excluded with replacement according to the same preregistered criteria ($n = 25$ and $n = 0$, respectively). Videos of an orange liquid being *scooped* or *poured out* with a mug (from or into a glass bowl) were recorded, and we selected two frames from each video (cropped to display only an 816 × 1026 central region) as distinct images of the same underlying event type: two images of less or more liquid being scooped or poured out (Figure 4a). Image analyses also confirmed that the within-type image pairs involved a greater degree of visual change than did the across-type image pair. First, although roughly 99.5% of the pixels ended up changing across all images, more pixels changed in the within-type image pairs (average: 836,508; scooping: 836,015; pouring: 837,000) compared with the across-type image pair (scooping vs. pouring: 826,890). Second, the distance in vectorized feature-activation maps from the penultimate layer (fc1) in VGG16 was greater for within-type image pairs (average: 118.63; scooping: 112.11; pouring: 125.15) compared with the across-type image pair (111.67). The across-type pair (the third and then the second images in Figure 4a) was repeated in 16 trials, which were in a different randomized order for each observer. The images were presented (sized with a width of 25% of the display width) with the image center jittered by 1/25 of the display width away from the display center for 400 ms.

Results and Discussion

We excluded individual trials whose response times were greater than 3 *SD* away from the mean response time for all observers (0.07/64) or shorter than 200 ms (3.18/64). As depicted in the top of Figure 5a, changes on across-type trials were detected more accurately than changes on within-type trials, 77.06% versus 53.66%, $t(99) = 10.27$, $p < .001$, $d = 1.03$—and as depicted in the bottom of Figure 5a, this was true for the vast majority (86%) of individual observers. Examination of the response times confirmed that there was no speed–accuracy trade-off because the response times actually differed in the same direction, 763 ms versus 866 ms, $t(99) = 3.88$, $p < .001$, $d = 0.39$. These results confirm that the cross-event-type advantage observed in Experiment 1a can be generalized to at least one other event type.

Figure 4
Stimuli From Experiments 2a and 2b



Note. (a) In Experiment 2a, a colorful liquid is being scooped or poured out, using a mug. (b) In Experiment 2b, these same images were inverted. See the online article for the color version of this figure.

Experiment 2b: Scooping Versus Pouring Control

The images of pouring versus scooping used in Experiment 2a differed in terms of their categorical event type, but of course they also necessarily differed in terms of various lower level properties—for example, the heights or precise angles of the mugs. To ensure that the results reflect the event categories, per se, we employed a familiar manipulation from face-perception research (e.g., Valentine, 1988), and we simply repeated the experiment with the same images *inverted* (as in Figure 4b). This obviously maintains all of the lower level properties of the images while seeming to obscure categorical differences. (In terms of brute phenomenology, we note informally that the images in Figure 4a seem immediately and spontaneously recognizable as instances of pouring vs. scooping, whereas the images in Figure 4b all seem to blend together—with any pouring/scooping differences being apparent only after careful scrutiny, if at all.)

Method

This experiment was identical to Experiment 2a except as noted. A new set of 100 observers (39 women, 61 men, average age = 40.37) participated. Observers were excluded with replacement according to the same preregistered criteria ($n = 30$ and $n = 0$, respectively). Stimuli were identical to those used in Experiment 2a, except that all the images were inverted.

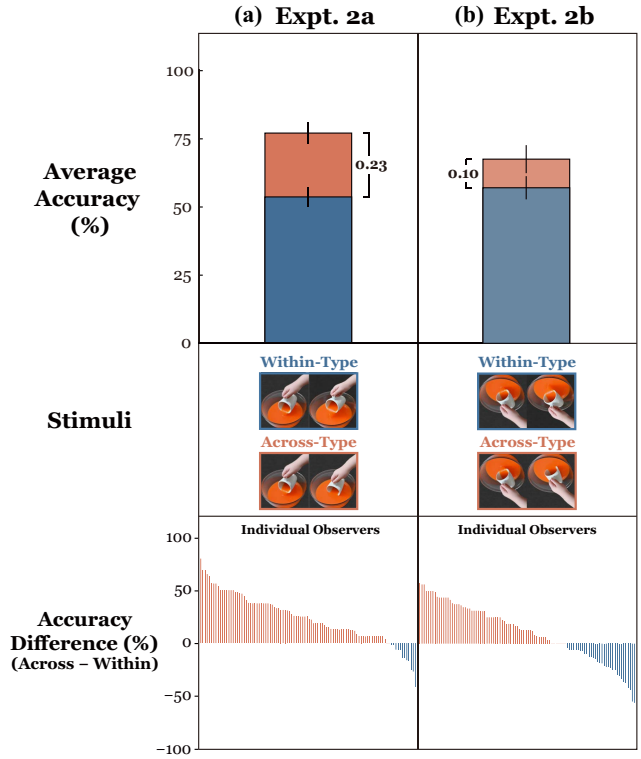
Results and Discussion

We again excluded individual trials whose response times were greater than 3 *SD* away from the mean response time for all observers (0.75/64) or shorter than 200 ms (2.92/64). As depicted in the top of Figure 5b, changes on across-type trials were again detected more accurately than changes on within-type trials, 67.48% versus 57.00%, $t(99) = 3.87$, $p < .001$, $d = 0.39$—and as depicted in the bottom of Figure 5b, this was again true for the majority (61%) of individual observers. Examination of the response times confirmed that there was no speed–accuracy trade-off, 873 ms versus 899 ms, $t(99) = 1.24$, $p = .218$, $d = 0.12$. A comparison between the top panels of Figure 5a and 5b, however, suggests that the effect was considerably greater in the upright images, and a cross-experiment analysis confirmed this, 23.4% in Experiment 2a versus 10.48% in Experiment 2b, $t(198) = 3.66$, $p < .001$, $d = 0.52$. These results confirm that the large and robust cross-event-type advantage observed in Experiment 2a cannot be explained simply by appeal to lower level features.

Experiment 3a: Rolling Versus Bouncing

Throughout this article we have been considering the nature of dynamic event types, but the stimuli in Experiments 1 and 2 were always static depictions of such dynamic events. To explore whether the cross-event-type advantage also occurs with dynamic animations, we next conducted the same type of change detection study but now using dynamic animations of *rolling* versus *bouncing* (as depicted in Figure 6a). Would observers also be better at detecting across-type changes between rolling and bouncing (as illustrated in the middle two images of Figure 6a) than within-type changes (two distinct videos of rolling, as illustrated in the left two images from Figure 6a; or two distinct videos of bouncing, as illustrated in the right two images of Figure 6a)?

Figure 5
Results From (a) Experiment 2a and (b) Experiment 2b

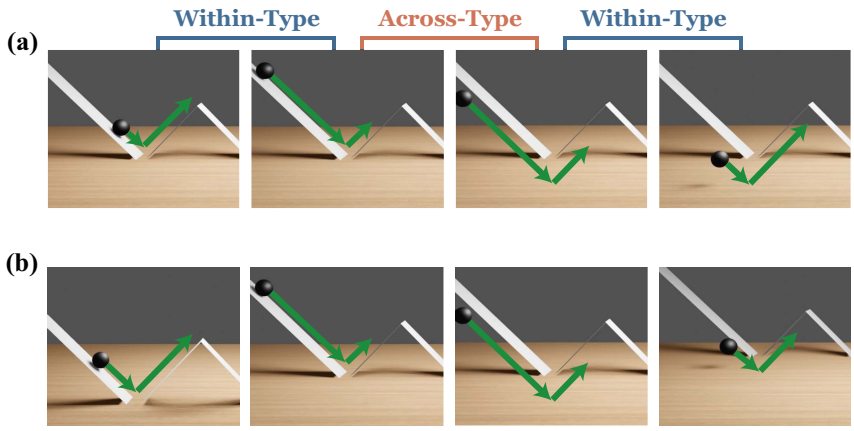


Note. The top panels depict the average accuracy for within-type versus across-type trials. The bottom panels depict the magnitude of the across-minus-within difference separately for each observer (with observers who performed better for across-type trials depicted in orange and those who performed better for within-type trials depicted in blue.) Error bars indicate 95% confidence intervals. See the online article for the color version of this figure.

Method

This experiment was identical to Experiment 1a except as noted. A new set of 100 observers (56 women, 42 men, 2 other, average age = 39.09) participated. Observers were excluded with replacement according to the same preregistered criteria ($n = 24$ and $n = 0$, respectively). A black ball (20 kg) was created in Blender (Version 3.3.0; https://www.blender.org/), either rolling along white ramps (with a friction value of 0.02 and a bounciness value of 1.80) or bouncing in front of the ramps along with a similar trajectory (controlled by a Bezier curve). Four videos were rendered and cropped to display only a 1138 × 720 central region (Figure 6a). Image analyses of all the frames in each video confirmed that the within-type video pairs involved a numerically greater degree of average visual change than did the across-type video pair. First, although roughly 36.4% of the pixels ended up changing across all video pairs, more pixels changed in the within-type video pairs (average: 335,499; rolling: 335,740; bouncing: 335,258) compared with the across-type video pair (rolling vs. bouncing: 335,038). Second, the distance in vectorized feature-activation maps from the penultimate layer (fc1) in VGG16 (Simonyan & Zisserman, 2015) was greater for within-type video pairs (average: 115.89; rolling:

Figure 6
Stimuli From Experiments 3a and 3b



Note. (a) In Experiment 3a, a ball is rolling along or bouncing in front of the ramps. (b) In Experiment 3b, the depth difference between within-type trials was larger. Because the actual stimuli were dynamic animations, the trajectories of balls are illustrated here by the green arrows (which were not shown in the actual experiments). See the online article for the color version of this figure.

97.89; bouncing: 133.88) compared with the across-type video pair (96.81). Each video was presented for 300 ms. Animations of these conditions are available online at https://perception.yale.edu/visual-verbs/.

Results and Discussion

We excluded individual trials whose response times were greater than 3 *SD* away from the mean response time for all observers (0.63/64) or shorter than 200 ms (1.07/64). As depicted in the top of Figure 7a, changes on across-type trials were detected more accurately than changes on within-type trials, 74.45% versus 56.23%, $t(99) = 5.55$, $p < .001$, $d = 0.56$—and as depicted in the bottom of Figure 7a, this was true for the majority (63%) of individual observers. Examination of the response times confirmed that there was no speed–accuracy trade-off, 1,014 ms versus 1,042 ms, $t(99) = 1.15$, $p = .252$, $d = 0.12$. These results confirm that the cross-event-type advantage observed in former experiments also applies to the perception of dynamic animations.

Experiment 3b: Rolling Versus Bouncing Control

Among the lower level image differences in the stimuli used in Experiment 3a, one basic visual property is especially salient: relative *depth*. In particular, on across-type trials, the ball was rolling on the ramps, but bouncing *in front of* the (similarly positioned) ramps, closer to the observer. To ensure that this could not explain our results, we simply replicated Experiment 3a by increasing the depth difference on within-type trials (as in Figure 6b)—so that the two animations of rolling (or bouncing) themselves occurred at an even greater difference in relative depth. (In Figure 6b, this can be appreciated by noticing that the entities in the first and third panels are closer to the observer than are the entities in the second and fourth panels, respectively.)

Method

This experiment was identical to Experiment 3a except as noted. A new set of 100 observers (45 women, 53 men, 2 other, average age = 37.62) participated. Observers were excluded with replacement according to the same preregistered criteria ($n = 22$ and $n = 1$, respectively). Stimuli were identical to those used in Experiment 3a, except that the balls and ramps were moved closer (as in the first panel in Figure 6b) or further away (as in the fourth panel in Figure 6b) from the observer.

Results and Discussion

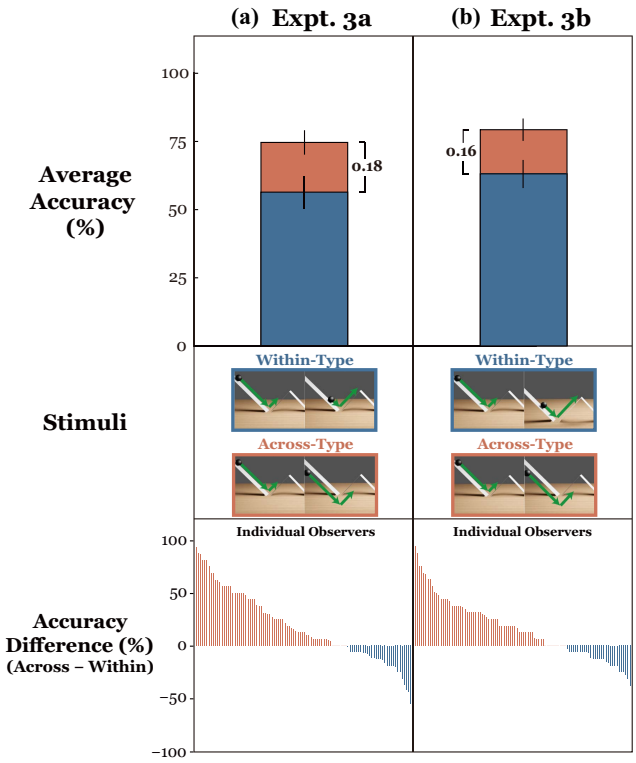
We excluded individual trials whose response times were greater than 3 *SD* away from the mean response time for all observers (0.82/64) or shorter than 200 ms (0.86/64). As depicted in the top of Figure 7b, changes on across-type trials were again detected more accurately than changes on within-type trials, 79.08% versus 62.93%, $t(99) = 5.87$, $p < .001$, $d = 0.59$—and as depicted in the bottom of Figure 7b, this was true for the majority (62%) of individual observers. Examination of the response times confirmed that there was no speed–accuracy trade-off, 973 ms versus 1,018 ms, $t(99) = 2.31$, $p = .023$, $d = 0.23$. Moreover, a cross-experiment comparison revealed that the accuracy difference (across-minus-within) in this experiment did not differ from the effect in Experiment 3a, 16.15% versus 18.22%, $t(198) = 0.48$, $p = .629$, $d = 0.07$. These results confirm that the cross-event-type advantage observed in Experiment 3a cannot be explained simply by appeal to the differing relative depth.

General Discussion

The central empirical contribution of this project was the discovery of a cross-event-type advantage in visual perception and memory, as demonstrated in patterns of objective performance (rather than subjective judgments). Across six experiments, people

Figure 7

Results From (a) Experiment 3a and (b) Experiment 3b



Note. The top panels depict the average accuracy for within-type versus across-type trials. The bottom panels depict the magnitude of the across-minus-within difference separately for each observer (with observers who performed better for across-type trials depicted in orange and those who performed better for within-type trials depicted in blue.) Error bars indicate 95% confidence intervals. See the online article for the color version of this figure.

were better at detecting changes to images and animations when the differences involved a change in the perceived event types (e.g., from rolling to bouncing) compared with when the differences involved only a single event type (e.g., from one animation of bouncing to a different animation of bouncing). This sort of categorical perception for event types occurred during spontaneous perception, insofar as events and categories were never mentioned: Observers had to detect any changes to the images and animations, and the manipulation of different categorical types was always entirely task irrelevant. Moreover, this effect seemed tied to the categorical types themselves because (a) the stimuli were carefully constructed so that the objective magnitudes of changes were always numerically *smaller* in cross-event-type displays compared with within-event-type displays, and (b) careful controls ruled out explanations based on lower level properties. Per the 2 × 2 matrix from Figure 1, these results suggest that visual perception involves the spontaneous representation not only of object tokens, object types, and event tokens but also event types—a foundational kind of visual representation that may be likened to visual verbs.

The cross-event-type advantage observed here seemed especially robust in at least five related ways. First, the relevant differences themselves were relatively enormous: The degree to which observers were better at detecting across-type changes compared with (the objectively greater) within-type changes was always at least 15% and averaged more than 20% across the four such demonstrations reported here (in Experiments 1a, 2a, 3a, and 3b). Second, these effects were highly statistically robust, with the relevant p values nowhere near the traditional significance boundaries: In each case, following standard convention, we simply report that the critical (across-minus-within) p values were $<.001$, but in fact they averaged $p < 7.332 \times 10^{-8}$. Third, the effect sizes of the cross-event-type advantages were accordingly always very large—always greater than $d = 0.5$, and even greater than $d = 1.0$ for two of the key tests (in Experiments 1a and 2a). Fourth, these results generalized across both static images of dynamic events (in Experiments 1 and 2) and actual dynamic animations (in Experiment 3). Fifth, the cross-event-type advantage observed here generalized across several different case studies: People were better at detecting changes across *twisting* versus *rotating* (in a way that could not be explained by hand positions; Experiments 1a and 1b), *scooping* versus *pouring* (in a way that could not be explained by any low-level image features; Experiments 2a and 2b), and *rolling* versus *bouncing* (in a way that could not be explained by differences in relative depth; Experiments 3a and 3b). These case studies were designed to span examples that centrally involved an agent (e.g., *scooping*) as well as those that did not (e.g., *bouncing*).

These effects could have been driven by differences in the underlying visual memories themselves, or they could have been driven by the (possibly even unconscious) activation of verbal labels (such as "bouncing") while viewing the images or animations. This remains an open question, and we intend our conclusions to be equally consistent with (and novel with respect to) both possibilities: In either case, the extracted event categories seemed "visual" insofar as (a) the effects were spontaneous and occurred during passive viewing of visual stimuli, even when such distinctions were entirely task irrelevant, and (b) such representations were activated in a way that directly influenced which visual details were encoded into visual working memory. In any case, we suspect that any activation of verbal labels would have to have been relatively subtle because (a) such labels would not be helpful in the underlying task (which involved the detection of subtle visual details), (b) such labels on their own would make the detection of within-category changes impossible by definition, and (c) in response to debriefing questions (e.g., "Did you find yourself using any particular strategy to complete the experiment?"), none of the 600 observers mentioned any form of verbal encoding.

Relation to Previous Work

The present study builds on a large and diverse foundation of previous work on event categorization in higher level cognition, which shows how both adults and children can represent events in terms of abstract categories. Some of this work aims to show simply that this is possible in the first place for various categories (such as *give* vs. *take* or even *give* vs. *hug*), especially in infancy (e.g., Gordon, 2004; Tatone et al., 2015). But much or even most of the extant work in this domain has focused instead on the various ways in which many kinds of events have a similar structure, for example, along the following dimensions.

Agents Versus Patients. Many researchers have argued that a central distinction in many event categories involves agents versus patients, often corresponding to subjects versus objects (e.g., Gärdenfors, 2014; for a review, see Rissman & Majid, 2019). People are able to reliably draw such distinctions (Braine & Wells, 1978) even with very sparse input (Hafri et al., 2013) and even just when hearing certain kinds of verbs in simple phrases (e.g., Mauner & Koenig, 2000). And they may also find it especially salient when such roles are swapped among actors (e.g., Leslie & Keeble, 1987).

Sources Versus Goals. People also seem especially sensitive to the ways in which many events are oriented *toward* goals (and correspondingly *from* sources). In the first place, even infants seem especially sensitive to such differences in simple visual depictions (e.g., when viewing a duck moving into a bowl [the goal] vs. moving out of a bowl [the source]; Lakusta et al., 2017). Moreover, when viewing videos depicting a goal-directed event (e.g., a person walking from a table [source] to a ladder [goal]), both adults and children are better at detecting goal changes compared with source changes (Lakusta & Landau, 2012). More generally, event roles defined by goals (such as being a *chaser*) also seem to be prioritized in many ways, including word learning (Yin & Csibra, 2015).

Boundedness Versus Unboundedness. As a final example, several recent studies have explored how people draw a fundamental distinction between events that are intrinsically bounded (having a clear endpoint, such as "blowing a bubble") from those that are unbounded (such as "blowing bubbles"). People are able to readily draw this distinction, even when linguistic coding is disrupted (Ji & Papafragou, 2020a), and they will naturally use this distinction when attempting to group different kinds of events (Ji & Papafragou, 2020b).

This foundational work has collectively emphasized the nature and importance of abstract event categories, which in turn directly inspired the current project. At the same time, however, we note that the experiments reported here differ from this previous work along several related dimensions. First, and perhaps most obviously, whereas most of this past work has explored various dimensions of event structure, which may apply across broad ranges of events (such as boundedness or agent/patient relationships), the current studies aimed instead to *contrast* potentially distinct event types (such as bouncing vs. rolling). Second, and perhaps most importantly, most of this past work has shown how people *can* draw such categorical distinctions when they are actively invited to do so in various ways that make them maximally task relevant (e.g., subjects might be directly asked a question such as "Did you see 'punching'?"; Hafri et al., 2013). In contrast, the central question explored here is whether such categorical distinctions are made in adult visual perception *spontaneously* (simply during passive viewing), even when they are entirely task irrelevant. As such, none of this previous literature would necessarily entail (or predict) the current pattern of results. Third, and in a more operationalized vein, the current experiments used a categorical perception measure (along with careful novel stimulus controls to ensure that across-type changes were actually *less* extreme than within-type changes)—where to our knowledge neither this task nor such controls have been previously employed in this domain.

Why Visual Verbs? Generalization and Prediction

Why might the visual system spontaneously encode the world in terms of event-type representations? The likely answers here might simply involve event-based analogues to the usual explanations of the more familiar representation of object types (see also Strickland & Scholl, 2015). In the first place, representations of types afford a kind of *generalization*. If you can recognize some part of a scene as an instance of the category *animal* or *agent*, for example—say, as opposed to an instance of *plant* or *mineral*—then you might be able to infer many other things about it (e.g., its likely spatiotemporal stability) simply from that categorization alone, even without yet having processed other details of the scene. Similarly, for event types, if you can recognize some part of a (perhaps dynamic) scene as an instance of the category *pouring*, for example—say, as opposed to an instance of *scooping*—then you can infer other things about it (e.g., which parts of the scene are likely to increase vs. decrease in size) simply from that categorization alone. Of course this sort of generalization may also help to promote the appreciation of similarity across tokens, as two scenes with radically different superficial features might still be both recognized as instances of *agents*, or *pouring*.

This same point can also be put in terms of *prediction*. Many models have stressed how perception involves active online predictions about what sensory stimulation is likely to arrive next, with prioritization of signals that fail to match these predictions (e.g., Huang & Rao, 2011; Rao & Ballard, 1999; Walsh et al., 2020). Indeed, some colleagues even suggest that such dynamics make it sensible to characterize "perceiving as predicting" (Clark, 2014). It seems natural how and why event-type representations would have advantages in this context by allowing for greater or improved predictions. If you can recognize some part of a dynamic scene as *bouncing* (vs. *rolling*), for example, that might readily allow you to predict the immediate future trajectory of the relevant object, even without processing any of its other properties (e.g., an object that is *bouncing* is more likely to vary its immediate vertical position in the coming moments more than is a *rolling* object, in the absence of changes to the ground plane). We speculate that such considerations may make the representation of visual verbs adaptive.

Constraints on Generality

The cross-event-type advantage reported here was observed during change detection with both images and short animations, with agents (in the form of visible hands) both present and absent, and with multiple distinct event types. However, there remain several open questions about the degree to which these results might generalize in other ways.

Of course future work could explore such effects in different subject populations. Perhaps most obviously, it remains unclear how or whether the current effects may relate to the ways in which event categories are lexicalized in various languages. We intentionally avoided choosing event types for these experiments that are already known to vary across languages (such as *breaking* vs. *cutting*; Kemmerer, 2019)—but it could be especially interesting to test event-type contrasts in languages that do not lexicalize such distinctions, if in fact such languages exist. If the present results disappeared in languages that did not lexicalize such distinctions, that would of course provide strong evidence that verbal coding helps to mediate such effects. But that possibility is by no means certain, for at least

two reasons. First, several other studies have uncovered cross-linguistically universal ways of categorizing events, which may reflect natural perceptual distinctions (Majid et al., 2008; Malt et al., 2008; Rissman et al., 2023). Second, it is possible that spontaneous event categorization during perception employs a standard "library" of dynamic primitives, only some of which are highlighted in some languages. This might end up being similar to how certain core contrasts are highlighted in infancy regardless of language—where, for example, all infants tend to highlight contrasts such as tight versus loose containment (e.g., McDonough et al., 2003), even when that distinction only ends up being lexicalized in some languages (famously, in Korean, but not English).

In a similar vein, it may also be especially important for future work to explore how these effects may or may not generalize to other event types. In particular, note that the "library" of event types in the present study—*twisting*, *rotating*, *scooping*, *pouring*, *rolling*, and *bouncing*—all stem from simple types of physical interactions. This was not an accident. Indeed, we chose these particular candidate event types precisely because they seemed so tightly bound with various forms of intuitive physics and are thus likely to be salient across vast swaths of space and time. (Events such as *twisting* and *rolling* seem likely to have been a part of human experience for as long as there have been humans.) But we do not yet know whether such effects would also obtain with putative event types that rely more on social or cultural conventions (such as *waving*, *nodding*, or *handshaking*).

Conclusion: From Language to Perception

Vision and language have always seemed like two of the most appealing candidates for possible modular/encapsulated systems in the human mind (e.g., Garfield, 1987; Pylyshyn, 1999), suggesting that they may interact only in relatively constrained ways. At the same time, however, a growing body of work suggests that they may often operate in similar ways (e.g., Quilty-Dunn et al., 2023). Examples include similar representations of symmetry in language and vision (Hafri et al., 2023), similar distinctions between agents and patients in terms of their event roles (Hafri et al., 2018), and perhaps even a notion of "grammar" as applied to typical scene layouts (for reviews see Võ, 2021; Võ et al., 2019).

The current results may contribute to this research program, insofar as the spontaneous visual representation of event types was inspired directly by the notion of a "language of vision" (Cavanagh, 2021). We leave future research to determine whether and how an individual's specific language and/or culture may influence such effects while noting that such visual verbs—at least in the context of the foundational event types studied here (all of which stem from universal properties of physical interactions)—may constitute a sort of library of dynamic primitives (Cavanagh et al., 2001), which could in turn constitute of another sort of core knowledge of the world (Scholl, 2024; Spelke, 2000, 2022).

References

Abassi, E., & Papeo, L. (2020). The representation of two-body shapes in the human visual cortex. *The Journal of Neuroscience*, *40*(4), 852–863. https://doi.org/10.1523/JNEUROSCI.1378-19.2019

Almeida, J., Mahon, B. Z., Nakayama, K., & Caramazza, A. (2008). Unconscious processing dissociates along categorical lines. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(39), 15214–15218. https://doi.org/10.1073/pnas.0805867105

Bach, E. (1986). The algebra of events. *Linguistics and Philosophy*, *9*(1), 5–16. https://doi.org/10.1007/BF00627432

Bae, G. Y., Olkkonen, M., Allred, S. R., & Flombaum, J. I. (2015). Why some colors appear more memorable than others: A model combining categories and particulars in color working memory. *Journal of Experimental Psychology: General*, *144*(4), 744–763. https://doi.org/10.1037/xge0000076

Baillargeon, R., & Wang, S. H. (2002). Event categorization in infancy. *Trends in Cognitive Sciences*, *6*(2), 85–93. https://doi.org/10.1016/S1364-6613(00)01836-2

Blake, R., & Shiffrar, M. (2007). Perception of human motion. *Annual Review of Psychology*, *58*(1), 47–73. https://doi.org/10.1146/annurev.psych.57.102904.190152

Bonin, P., Gelin, M., & Bugaiska, A. (2014). Animates are better remembered than inanimates: Further evidence from word and picture stimuli. *Memory & Cognition*, *42*(3), 370–382. https://doi.org/10.3758/s13421-013-0368-8

Bornstein, M. H., & Korda, N. O. (1984). Discrimination and matching within and between hues measured by reaction times: Some implications for categorical perception and levels of information processing. *Psychological Research*, *46*(3), 207–222. https://doi.org/10.1007/BF00308884

Braine, M. D. S., & Wells, R. S. (1978). Case-like categories in children: The actor and some related categories. *Cognitive Psychology*, *10*(1), 100–122. https://doi.org/10.1016/0010-0285(78)90020-8

Calvillo, D. P., & Hawkins, W. C. (2016). Animate objects are detected more frequently than inanimate objects in inattentional blindness tasks independently of threat. *Journal of General Psychology*, *143*(2), 101–115. https://doi.org/10.1080/00221309.2016.1163249

Cavanagh, P. (2021). The language of vision. *Perception*, *50*(3), 195–215. https://doi.org/10.1177/0301006621991491

Cavanagh, P., Caplovitz, G. P., Lytchenko, T. K., Maechler, M. R., Tse, P. U., & Sheinberg, D. L. (2023). The architecture of object-based attention. *Psychonomic Bulletin & Review*, *30*(5), 1643–1667. https://doi.org/10.3758/s13423-023-02281-7

Cavanagh, P., Labianca, A. T., & Thornton, I. M. (2001). Attention-based visual routines: Sprites. *Cognition*, *80*(1–2), 47–60. https://doi.org/10.1016/S0010-0277(00)00153-0

Chao, L. L., Haxby, J. V., & Martin, A. (1999). Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. *Nature Neuroscience*, *2*(10), 913–919. https://doi.org/10.1038/13217

Clark, A. (2014). Perceiving as predicting. In D. Stokes, M. Matthen, & S. Biggs (Eds.), *Perception and its modalities* (pp. 23–43). Oxford University Press. https://doi.org/10.1093/acprof:Oso/9780199832798.003.0002

Cunningham, C. A., & Egeth, H. E. (2018). The capture of attention by entirely irrelevant pictures of calorie-dense foods. *Psychonomic Bulletin & Review*, *25*(2), 586–595. https://doi.org/10.3758/s13423-017-1375-8

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

Downing, P. E., Chan, A. W., Peelen, M. V., Dodds, C. M., & Kanwisher, N. (2006). Domain specificity in visual cortex. *Cerebral Cortex*, *16*(10), 1453–1461. https://doi.org/10.1093/cercor/bhj086

Duncan, J. (1984). Selective attention and the organization of visual information. *Journal of Experimental Psychology: General*, *113*(4), 501–517. https://doi.org/10.1037/0096-3445.113.4.501

Egly, R., Driver, J., & Rafal, R. D. (1994). Shifting visual attention between objects and locations: Evidence from normal and parietal lesion subjects. *Journal of Experimental Psychology: General*, *123*(2), 161–177. https://doi.org/10.1037/0096-3445.123.2.161

Etcoff, N. L., & Magee, J. J. (1992). Categorical perception of facial expressions. *Cognition*, *44*(3), 227–240. https://doi.org/10.1016/0010-0277(92)90002-Y

Fuentemilla, L., Penny, W. D., Cashdollar, N., Bunzeck, N., & Düzel, E. (2010). Theta-coupled periodic replay in working memory. *Current Biology*, *20*(7), 606–612. https://doi.org/10.1016/j.cub.2010.01.057

Gärdenfors, P. (2014). *The geometry of meaning: Semantics based on conceptual spaces*. MIT Press. https://doi.org/10.7551/mitpress/9629.001.0001

Garfield, J. L. (1987). *Modularity in knowledge representation and natural-language understanding*. MIT Press. https://doi.org/10.7551/mitpress/4735.001.0001

Goldstone, R. L., & Hendrickson, A. T. (2010). Categorical perception. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(1), 69–78. https://doi.org/10.1002/wcs.26

Gordon, P. (2004). The origin of argument structure in infant event representations. *Proceedings of the Annual Boston University Conference on Language Development*, *28*, 189–198.

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. https://doi.org/10.1037/xge0001283

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), 880–905. https://doi.org/10.1037/a0030045

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. https://doi.org/10.1016/j.cognition.2018.02.011

Harnad, S. (1987). *Categorical perception: The groundwork of cognition*. Cambridge University Press.

He, X., Witzel, C., Forder, L., Clifford, A., & Franklin, A. (2014). Color categories only affect post-perceptual processes when same- and different-category colors are equally discriminable. *Journal of the Optical Society of America. A, Optics, Image Science, and Vision*, *31*(4), A322–A331. https://doi.org/10.1364/JOSAA.31.00A322

Hespos, S. J., & Baillargeon, R. (2001). Infants' knowledge about occlusion and containment events: A surprising discrepancy. *Psychological Science*, *12*(2), 141–147. https://doi.org/10.1111/1467-9280.00324

Huang, Y., & Rao, R. P. N. (2011). Predictive coding. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(5), 580–593. https://doi.org/10.1002/wcs.142

Huff, M., Papenmeier, F., & Zacks, J. M. (2012). Visual target detection is impaired at event boundaries. *Visual Cognition*, *20*(7), 848–864. https://doi.org/10.1080/13506285.2012.705359

Ji, Y., & Papafragou, A. (2020a). Is there an end in sight? Viewers' sensitivity to abstract event structure. *Cognition*, *197*, Article 104197. https://doi.org/10.1016/j.cognition.2020.104197

Ji, Y., & Papafragou, A. (2020b). Midpoints, endpoints and the cognitive structure of events. *Language, Cognition and Neuroscience*, *35*(10), 1465–1479. https://doi.org/10.1080/23273798.2020.1797839

Kemmerer, D. (2019). *Concepts in the brain: The view from cross-linguistic diversity*. Oxford University Press. https://doi.org/10.1093/oso/9780190682620.001.0001

Khosla, M., Ratan Murty, N. A., & Kanwisher, N. (2022). A highly selective response to food in human visual cortex revealed by hypothesis-free voxel decomposition. *Current Biology*, *32*(19), 4159–4171.e9. https://doi.org/10.1016/j.cub.2022.08.009

Kominsky, J. F., & Scholl, B. J. (2020). Retinotopic adaptation reveals distinct categories of causal perception. *Cognition*, *203*, Article 104339. https://doi.org/10.1016/j.cognition.2020.104339

Kovács, I., & Julesz, B. (1993). A closed curve is much more than an incomplete one: Effect of closure in figure-ground segmentation. *Proceedings of the National Academy of Sciences of the United States of America*, *90*(16), 7495–7497. https://doi.org/10.1073/pnas.90.16.7495

Kurby, C. A., & Zacks, J. M. (2008). Segmentation in the perception and memory of events. *Trends in Cognitive Sciences*, *12*(2), 72–79. https://doi.org/10.1016/j.tics.2007.11.004

Lakusta, L., & Landau, B. (2012). Language and memory for motion events: Origins of the asymmetry between source and goal paths. *Cognitive Science*, *36*(3), 517–544. https://doi.org/10.1111/j.1551-6709.2011.01220.x

Lakusta, L., Spinelli, D., & Garcia, K. (2017). The relationship between preverbal event representations and semantic structures: The case of goal and source paths. *Cognition*, *164*, 174–187. https://doi.org/10.1016/j.cognition.2017.04.003

Leslie, A. M., & Keeble, S. (1987). Do six-month-old infants perceive causality? *Cognition*, *25*(3), 265–288. https://doi.org/10.1016/S0010-0277(87)80006-9

Levin, D. T., & Varakin, D. A. (2004). No pause for a brief disruption: Failures of visual awareness during ongoing events. *Consciousness and Cognition*, *13*(2), 363–372. https://doi.org/10.1016/j.concog.2003.12.001

Li, F.-F., Iyer, A., Koch, C., & Perona, P. (2007). What do we perceive in a glance of a real-world scene? *Journal of Vision*, *7*(1), Article 10. https://doi.org/10.1167/7.1.10

Loucks, J., Verrett, K., & Reise, B. (2020). Animates engender robust memory representations in adults and young children. *Cognition*, *201*, Article 104284. https://doi.org/10.1016/j.cognition.2020.104284

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279–281. https://doi.org/10.1038/36846

Majid, A., Boster, J. S., & Bowerman, M. (2008). The cross-linguistic categorization of everyday events: A study of cutting and breaking. *Cognition*, *109*(2), 235–250. https://doi.org/10.1016/j.cognition.2008.08.009

Malt, B. C., Gennari, S., Imai, M., Ameel, E., Tsuda, N., & Majid, A. (2008). Talking about walking: Biomechanics and the language of locomotion. *Psychological Science*, *19*(3), 232–240. https://doi.org/10.1111/j.1467-9280.2008.02074.x

Marino, A. C., & Scholl, B. J. (2005). The role of closure in defining the "objects" of object-based attention. *Perception & Psychophysics*, *67*(7), 1140–1149. https://doi.org/10.3758/BF03193547

Mauner, G., & Koenig, J.-P. (2000). Linguistic vs. conceptual sources of implicit agents in sentence comprehension. *Journal of Memory and Language*, *43*(1), 110–134. https://doi.org/10.1006/jmla.1999.2703

McDonough, L., Choi, S., & Mandler, J. M. (2003). Understanding spatial relations: Flexible infants, lexical adults. *Cognitive Psychology*, *46*(3), 229–259. https://doi.org/10.1016/S0010-0285(02)00514-5

Mecklinger, A., Schriefers, H., Steinhauer, K., & Friederici, A. D. (1995). Processing relative clauses varying on syntactic and semantic dimensions: An analysis with event-related potentials. *Memory & Cognition*, *23*(4), 477–494. https://doi.org/10.3758/BF03197249

Michotte, A. (1963). *The perception of causality* (T. Miles & E. Miles, Trans.). Basic Books. (Original work published 1946)

New, J., Cosmides, L., & Tooby, J. (2007). Category-specific attention for animals reflects ancestral priorities, not expertise. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(42), 16598–16603. https://doi.org/10.1073/pnas.0703913104

New, J. J., Schultz, R. T., Wolf, J., Niehaus, J. L., Klin, A., German, T. C., & Scholl, B. J. (2010). The scope of social attention deficits in autism: Prioritized orienting to people and animals in static natural scenes. *Neuropsychologia*, *48*(1), 51–59. https://doi.org/10.1016/j.neuropsychologia.2009.08.008

Newell, F. N., & Bülthoff, H. H. (2002). Categorical perception of familiar objects. *Cognition*, *85*(2), 113–143. https://doi.org/10.1016/S0010-0277(02)00104-X

Ngiam, W. X. Q., Loetscher, K. B., & Awh, E. (2024). Object-based encoding constrains storage in visual working memory. *Journal of Experimental Psychology: General*, *153*(1), 86–101. https://doi.org/10.1037/xge0001479

Ongchoco, J. D. K., Walter-Terrill, R., & Scholl, B. J. (2023). Visual event boundaries restrict anchoring effects in decision-making. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(44), Article e2303883120. https://doi.org/10.1073/pnas.2303883120

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27. https://doi.org/10.1016/j.jbef.2017.12.004

Papafragou, A., Hulbert, J., & Trueswell, J. (2008). Does language guide event perception? Evidence from eye movements. *Cognition*, *108*(1), 155–184. https://doi.org/10.1016/j.cognition.2008.02.007

Papeo, L., & Abassi, E. (2019). Seeing social events: The visual specialization for dyadic human-human interactions. *Journal of Experimental Psychology: Human Perception and Performance*, *45*(7), 877–888. https://doi.org/10.1037/xhp0000646

Pennock, I. M. L., Racey, C., Allen, E. J., Wu, Y., Naselaris, T., Kay, K. N., Franklin, A., & Bosten, J. M. (2023). Color-biased regions in the ventral visual pathway are food selective. *Current Biology*, *33*(1), 134–146.e4. https://doi.org/10.1016/j.cub.2022.11.063

Pustejovsky, J. (1991). The syntax of event structure. *Cognition*, *41*(1–3), 47–81. https://doi.org/10.1016/0010-0277(91)90032-Y

Pylyshyn, Z. (1999). Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. *Behavioral and Brain Sciences*, *22*(3), 341–365. https://doi.org/10.1017/S0140525X99002022

Quilty-Dunn, J., Porot, N., & Mandelbaum, E. (2023). The best game in town: The re-emergence of the language of thought hypothesis across the cognitive sciences. *Behavioral and Brain Sciences*, *46*, Article e261. https://doi.org/10.1017/S0140525X22002849

Radvansky, G. A. (2012). Across the event horizon. *Current Directions in Psychological Science*, *21*(4), 269–272. https://doi.org/10.1177/0963721412451274

Radvansky, G. A., & Copeland, D. E. (2006). Walking through doorways causes forgetting: Situation models and experienced space. *Memory & Cognition*, *34*(5), 1150–1156. https://doi.org/10.3758/BF03193261

Radvansky, G. A., & Zacks, J. M. (2014). *Event cognition*. Oxford University Press. https://doi.org/10.1093/acprof:Oso/9780199898138.001.0001

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87. https://doi.org/10.1038/4580

Richmond, L. L., Gold, D. A., & Zacks, J. M. (2017). Event perception: Translations and applications. *Journal of Applied Research in Memory and Cognition*, *6*(2), 111–120. https://doi.org/10.1016/j.jarmac.2016.11.002

Rissman, L., Horton, L., & Goldin-Meadow, S. (2023). Universal constraints on linguistic event categories: A cross-cultural study of child homesign. *Psychological Science*, *34*(3), 298–312. https://doi.org/10.1177/09567976221140328

Rissman, L., & Majid, A. (2019). Thematic roles: Core knowledge or linguistic construct? *Psychonomic Bulletin & Review*, *26*(6), 1850–1869. https://doi.org/10.3758/s13423-019-01634-5

Rolfs, M., Dambacher, M., & Cavanagh, P. (2013). Visual adaptation of the perception of causality. *Current Biology*, *23*(3), 250–254. https://doi.org/10.1016/j.cub.2012.12.017

Sargent, J. Q., Zacks, J. M., Hambrick, D. Z., Zacks, R. T., Kurby, C. A., Bailey, H. R., Eisenberg, M. L., & Beck, T. M. (2013). Event segmentation ability uniquely predicts event memory. *Cognition*, *129*(2), 241–255. https://doi.org/10.1016/j.cognition.2013.07.002

Scholl, B. J. (2001). Objects and attention: The state of the art. *Cognition*, *80*(1–2), 1–46. https://doi.org/10.1016/S0010-0277(00)00152-9

Scholl, B. J. (2024). Perceptual (roots of) core knowledge. *Behavioral and Brain Sciences*, *47*, Article e140. https://doi.org/10.1017/S0140525X23003023

Scholl, B. J., & Tremoulet, P. D. (2000). Perceptual causality and animacy. *Trends in Cognitive Sciences*, *4*(8), 299–309. https://doi.org/10.1016/S1364-6613(00)01506-0

Simonyan, K., & Zisserman, A. (2015). *Very deep convolutional networks for large-scale image recognition* [Conference session]. The 3rd International Conference on Learning Representations (ICLR 2015), San Diego, CA, United States. https://doi.org/10.48550/arXiv.1409.1556

Spelke, E. S. (2000). Core knowledge. *American Psychologist*, *55*(11), 1233–1243. https://doi.org/10.1037/0003-066X.55.11.1233

Spelke, E. S. (2022). *What babies know: Core knowledge and composition*. Oxford University Press. https://doi.org/10.1093/oso/9780190618247.001.0001

Strickland, B., & Scholl, B. J. (2015). Visual perception involves event-type representations: The case of containment versus occlusion. *Journal of Experimental Psychology: General*, *144*(3), 570–580. https://doi.org/10.1037/a0037750

Suchow, J. W., Fougnie, D., Brady, T. F., & Alvarez, G. A. (2014). Terms of the debate on the format and structure of visual memory. *Attention, Perception & Psychophysics*, *76*(7), 2071–2079. https://doi.org/10.3758/s13414-014-0690-7

Swallow, K. M., Zacks, J. M., & Abrams, R. A. (2009). Event boundaries in perception affect memory encoding and updating. *Journal of Experimental Psychology: General*, *138*(2), 236–257. https://doi.org/10.1037/a0015631

Tatone, D., Geraci, A., & Csibra, G. (2015). Giving and taking: Representational building blocks of active resource-transfer events in human infants. *Cognition*, *137*, 47–62. https://doi.org/10.1016/j.cognition.2014.12.007

Troje, N. F. (2013). What is biological motion? Definition, stimuli, and paradigms. In M. D. Rutherford & V. A. Kuhlmeier (Eds.), *Social perception: Detection and interpretation of animacy, agency, and intention* (pp. 13–36). MIT Press. https://doi.org/10.7551/mitpress/9780262019279.003.0002

Valentine, T. (1988). Upside-down faces: A review of the effect of inversion upon face recognition. *British Journal of Psychology*, *79*(4), 471–491. https://doi.org/10.1111/j.2044-8295.1988.tb02747.x

van Buren, B., & Scholl, B. J. (2017). Minds in motion in memory: Enhanced spatial memory driven by the perceived animacy of simple shapes. *Cognition*, *163*, 87–92. https://doi.org/10.1016/j.cognition.2017.02.006

Võ, M. L. H. (2021). The meaning and structure of scenes. *Vision Research*, *181*, 10–20. https://doi.org/10.1016/j.visres.2020.11.003

Võ, 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. https://doi.org/10.1016/j.copsyc.2019.03.009

Vogel, E. K., & Machizawa, M. G. (2004). Neural activity predicts individual differences in visual working memory capacity. *Nature*, *428*(6984), 748–751. https://doi.org/10.1038/nature02447

Walsh, K. S., McGovern, D. P., Clark, A., & O'Connell, R. G. (2020). Evaluating the neurophysiological evidence for predictive processing as a model of perception. *Annals of the New York Academy of Sciences*, *1464*(1), 242–268. https://doi.org/10.1111/nyas.14321

Wiggett, A. J., Pritchard, I. C., & Downing, P. E. (2009). Animate and inanimate objects in human visual cortex: Evidence for task-independent category effects. *Neuropsychologia*, *47*(14), 3111–3117. https://doi.org/10.1016/j.neuropsychologia.2009.07.008

Yin, J., & Csibra, G. (2015). Concept-based word learning in human infants. *Psychological Science*, *26*(8), 1316–1324. https://doi.org/10.1177/0956797615588753

Zacks, J. M. (2020). Event perception and memory. *Annual Review of Psychology*, *71*(1), 165–191. https://doi.org/10.1146/annurev-psych-010419-051101

Received August 2, 2023
Revision received June 6, 2024
Accepted June 12, 2024 ■