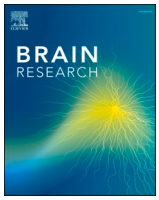
Brain Research 1869 (2025) 150000

Contents lists available at ScienceDirect

Brain Research

journal homepage: www.elsevier.com/locate/brainres



Apprehending relational events: The visual world paradigm and the interplay of event perception and language☆

Alon Hafri [a,*], John C. Trueswell [b,*]

[a] University of Delaware, United States
[b] University of Pennsylvania, United States

ARTICLE INFO

Keywords:
Visual world paradigm
Relation perception
Event perception
Eye movements during language use

ABSTRACT

When we observe the world, we appreciate not only the colors, shapes, and textures of people and objects, but also how they interact with one another (e.g., in events such as a girl pushing a boy). Although it might seem intuitive that extracting events and other relations would require active effort and multiple fixations, a growing body of vision research suggests that humans rapidly and automatically extract relational information—including the structure of events (i.e., who is acting on whom)—from a single fixation. These findings suggest that aspects of event structure can often be perceived without extensive visual interrogation. Yet despite this progress, much remains unknown about how visual events are perceived and represented—particularly for complex events (e.g., those involving roles beyond Agent and Patient, or events with multiple salient construals, such as *chase* vs. *flee*)—and how these emerging representations interact with language during interpretation and production of utterances about events. The visual world paradigm (VWP) offers a powerful tool to address these questions about the perception-language interface, by revealing which event representations are active when and by probing how language may guide event construal in real-time. We review eyetracking work in VWP studies of language comprehension and language production, as well as in related tasks, that provide initial insights into online event apprehension. This work suggests that (1) event apprehension and linguistic encoding are closely coordinated, interacting earlier and more continuously than previously recognized, and (2) fixations may serve to refine or disambiguate relational information extracted during initial processing— such as identifying event participants or clarifying their roles (e.g., as Instruments, Goals, or Recipients)—with language in some cases guiding attentional prioritization towards certain event components. More generally, this perspective offers a foundation for future VWP research exploring the dynamic relationship between seeing, listening, and speaking.

1. Introduction

What we appreciate from a visual scene goes far beyond retinal stimulation. Consider the image in Fig. 1. You surely notice that the individuals are similar in real-world size, despite their different angular extents; that the girl has two hands despite that one of them is occluded; and even that the two individuals on the left are both boys, despite their differences in posture and clothing. This ability reflects what is widely held to be a primary function of vision: to infer distal properties of objects, including their sizes, shapes, surface properties, locations, and perhaps even their categories (Marr, 1982). But beyond such properties, you likely also grasp other aspects of this scene: that the boys are *on* a skateboard, that the girl is *pushing* them, and that this activity is occurring *in* the street. These impressions capture relations: attributes that describe interactions or connections between two or more entities, extending beyond the individual properties of each entity to encompass their spatiotemporal and causal structure within their broader environment (see e.g., Hafri and Firestone, 2021; Zacks, 2020).

What kinds of mental processes give rise to relational representations such as these? Under the traditional perspective on visual perception mentioned above, this kind of process is one that requires active engagement and deliberative inference, over and above the usual "unconscious inferences" performed by visual processing. In other words, if what vision provides is primarily the lower-level properties and locations of objects, some kind of additional processes—active "visual routines" of sorts (Ullman, 1984)—would have to sequentially "stitch

☆ This article is part of a special issue entitled: '30 Years of Visual World Paradigm: The State of the Art' published in Brain Research.

* Corresponding authors.
E-mail addresses: alon@udel.edu (A. Hafri), trueswel@psych.upenn.edu (J.C. Trueswell).

https://doi.org/10.1016/j.brainres.2025.150000
Received 20 January 2025; Received in revised form 10 October 2025; Accepted 14 October 2025
Available online 17 October 2025

Fig. 1. An everyday scene featuring multiple events (such as *pushing, sitting,* and *rolling*) and other relations (such as *contact* and *support*). A growing body of literature suggests that we extract many such relations spontaneously in the course of visual processing, just as we extract more basic properties of the world, such as the colors and textures of the clothing, or the sizes, shapes, and locations of the visible objects (the boys, the girl, and the skateboard). Photo credit: Pressmaster, Shutterstock.

together" these more basic elements into a relational representation. This has often been assumed to require *successive shifts of attention* (Ullman, 1984, 1996; Yuan et al., 2016; Holcombe et al., 2011) that would typically manifest in corresponding successive shifts in eye position through saccades and fixations (e.g., Hoffman, 1998).

However, a growing body of research suggests that, for some kinds of physical and social relations (including PUSH, HIT, and even CHASE), the visual system can rapidly construct relational representations—including their abstract structure (e.g., who is acting on whom)—based on information gleaned from a single fixation, and that it does so spontaneously or even automatically (for a review, see Hafri and Firestone, 2021). Moreover, these representations appear to be in a format that is accessible to higher-level cognitive systems such as language and reasoning, aligning with recent proposals arguing that perception furnishes abstract, structured information that can be "readily consumed" by higher-level cognitive processes (Quilty-Dunn, 2020; see also Altmann and Kamide, 2007, 2009; Cavanagh, 2021; Hafri et al., 2023; Hafri and Papeo, 2025; for more detailed discussion of the content and format of these representations, see Section 2.3).

These and related findings provide an interesting opportunity to connect with the field of psycholinguistics and in particular one of its primary behavioral methods: the visual world paradigm (VWP). In the VWP, participants' eye movements are recorded while they are engaged in a task that involves the use of spoken language to interact with a visually co-present referent world. Gaze locations and durations are measured in order to make inferences about linguistic and cognitive processes on a moment-by-moment basis, including inferences about spoken word recognition (e.g., Allopenna et al., 1998), sentence comprehension (e.g., Spivey et al., 2002; Degen and Tanenhaus, 2016), sentence production (e.g., Griffin and Bock, 2000; Gleitman et al., 2007) and even joint communication among interlocutors (e.g., Brown-Schmidt et al., 2008; Heller et al., 2008). A good deal of this research happens to be about the language of events: the processing of verbs, tense, case markers, prepositions, and syntactic structure—all of which convey who-is-doing-what-to-whom, when, and where. Yet, as we discuss further below, only a small subset of this work has focused specifically on understanding how events are perceived and interact with linguistic processing.

In what follows, we first highlight key findings from the recent literature on event perception, demonstrating that visual processing rapidly and automatically extracts relational categories and structure from visual scenes (both foveally and extrafoveally).[1] While there are limits to the complexity and granularity of information gleaned in this way, such representations appear to be in a format that can be easily "read out" by higher-level cognitive systems, including language.

We then turn to how eyetracking has been used to study language processing (both comprehension and production) in the context of events, in the VWP and related paradigms. Our goal is to address two central questions that arise from the findings on relation perception:

1. What does the current VWP research tell us about the perception of events and how it interacts with language processing?
2. How can the VWP be used to better understand these processes?

Our review leads us to two main conclusions. First, event apprehension and linguistic encoding are tightly coordinated, interacting earlier and more continuously than is often assumed. Second, fixations in these tasks may function not only to index the allocation of overt attention, but also to refine or disambiguate relational information extracted during initial processing—for example, by identifying event participants or clarifying their roles (e.g., as Instruments, Goals, or Recipients)—with language sometimes guiding attention toward particular event components.

Along the way, we outline a roadmap for integrating event perception and VWP in future research. In particular, we suggest that systematically manipulating the timing of visual information relative to linguistic engagement will help us understand both the event perception process itself, and the role—and limits—of language in guiding event construal.

2. Rapid extraction of relation information

2.1. A new perspective on relation perception

The idea that visual perception represents certain events and other relations dates at least to Michotte's investigations of now classic 'launching' stimuli (Michotte, 1946/1963). When one disc approaches another, stops, and the second disc starts moving, people experience the impression of a causal interaction: of one disc *causing* another to move. The experience can be phenomenologically compelling and resistant to higher-level knowledge and beliefs (e.g., knowledge that the discs are simply shapes projected on a screen), and more modern tools of vision science have provided evidence that such interactions are genuinely *perceived* (e.g., they show retinotopic adaptation; Rolfs et al., 2013; Kominsky and Scholl, 2020).

More generally, a growing body of empirical literature has found that many types of relations, involving both social and physical events, are extracted rapidly and automatically in visual processing (for a review, see Hafri and Firestone, 2021). This is true even for naturalistic static snapshots of events, from which full information about what will or has transpired must typically be inferred (e.g., from an image of the moment of contact between a kicker and kickee; Dobel et al., 2007, 2010; Glanemann et al., 2016; Hafri et al., 2013; Hafri et al., 2018; Hafri et al., 2024; Peng et al., 2020). This work has used brief-display paradigms and tasks with time-constrained responses to provide evidence that certain aspects of relational encoding are indeed accomplished via rapid perceptual processes, including the extraction of abstract roles (Agent and Patient; Hafri et al., 2013, 2018; Vettori et al., 2024b) and relational categories (such as KICK, TAP, CONTAINMENT, or SUPPORT; Dobel et al., 2010; Glanemann et al., 2016; Hafri et al., 2013, 2024; Vettori et al., 2024a) (see Fig. 2).

[1] Out of a need to constrain the length of this paper, our focus will be on event relations and event structure. We do, however, situate our discussion within the broader topic of relations as a whole, including spatial relations, which are often intimately connected to events.

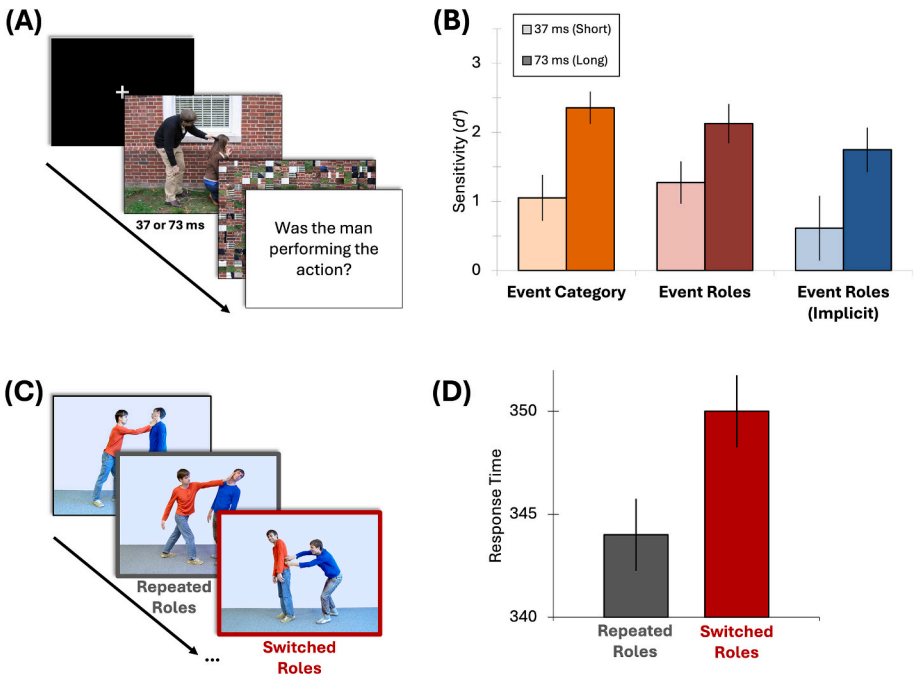


Fig. 2. (a) The task used in Hafri et al. (2013), in which participants reported event category or event roles after observing a briefly displayed and masked photograph of two-participant interactions. (b) Participants could reliably extract event category and event role, even when probed with an implicit prompt that did not directly reveal when event role information was being probed. (c) In Hafri et al. (2018), participants had to quickly perform a simple visual task (which side is the blue person on, left or right?) on a continuous sequence of photographs of two-participant interactions involving Agents and Patients. (d) When the role of the target individual switched from one trial to the next, participants were slower to respond than when roles repeated, even though event information was not directly probed, suggesting that participants were spontaneously encoding event-role information. Error bars in B and D reflect 95% within-participant confidence intervals. Figures adapted from Hafri et al. (2013, 2018).

For example, Hafri et al. (2013) found that after viewing a naturalistic two-participant event scene for a brief amount of time (37–73 ms) followed by a visual mask (Fig. 2A), observers could select which of two event categories they saw when prompted (e.g., "Did you see *kicking* or *tapping*?"; Fig. 2B). Similarly, observers could also make judgments of role assignment from these brief displays ("Did the *man* (or *woman*) do the kicking?"; Fig. 2B). Evidence suggests that observers make these judgments based on coarse-grain postural information, as disrupting the canonical postures or orientations of event participants (e.g., making the Patient more "Agent-like" by having them lean forward toward the Agent with outstretched arms) disrupts this ability (De Freitas and Hafri, 2024; Hafri et al., 2013; Vettori et al., 2024b). Relatedly, other work has demonstrated that observers can determine the coherence of a visual scene (i.e., whether the entities were interacting meaningfully or not) from a display of just 30 ms (Dobel et al., 2010; Glanemann et al., 2016).

Remarkably, the assignment of event roles in visual processing appears to be spontaneous, occurring without conscious deliberation about the particular event taking place (or indeed about events, more generally). For example, Hafri et al. (2018) asked observers to make simple, speeded judgments on event scenes involving asymmetric roles (i.e., Agent and Patient) while viewing these scenes in a continuous sequence (Fig. 2C). Observers reported the location (left or right) of a target individual (e.g., the blue-shirted man)—a task orthogonal to event information. Despite the ease of this task and the irrelevance of event role information to successfully completing it, changes in role assignment (e.g., the blue-shirted individual switching from Agent to Patient) interfered with participants' performance, resulting in a "role-switch cost" to their response times, even across changes in event type (e.g., tickling to biting) (Fig. 2D). Role information appears to be extracted somewhat independently from event category information (Hafri et al., 2013), as similar role-switch costs occur even in the absence of meaningful event category information, when Agent- and Patient-like postures are randomly paired with one another (Vettori et al., 2024b). Taken together, this work demonstrates that event structure is rapidly and automatically extracted and that the role information extracted at these timescales is abstract in nature, not tied to or dependent on specific knowledge about the precise kind of event taking place.

2.2. *Extra-foveal relational information: Extraction and its limits*

The work discussed above shows that in many cases, information about events and other relations can be extracted from a single fixation. These findings suggest a rethinking of approaches that advance a primary role for active, sequential routines for relational encoding (Ullman, 1984, 1996). However, it is important to note that much of the information available about events extends beyond the fovea. To experience this yourself, glance at the edge of the image in Fig. 1, and notice the difference between what event-relevant information is available at fixation and what you apprehend from the corner of your eye. This experience should make clear that the extraction of event information from a single fixation must rely at least in part on parafoveal and

peripheral visual input.[2] As we discuss below, there are limits to what information may be extracted extrafoveally, paving the way for investigations of eye movements—including in the visual world paradigm—in revealing the nature of event apprehension and how it makes contact with linguistic processes.

Work in scene perception has found that the rapid extraction of scene 'gist' (category, e.g., *kitchen*) can be supported by foveal, parafoveal, and peripheral processing simultaneously (Castelhano and Heaven, 2011; Castelhano and Henderson, 2007; Greene and Oliva, 2009a, 2009b; Larson and Loschky, 2009; Pereira and Castelhano, 2014; Torralba et al., 2006). Growing evidence suggests that this gist information includes not only category information, but also structural information about spatial layout and even the hierarchical and functional relations between objects in a scene (e.g., pots appear on stoves) (Josephs et al., 2016; Kadar and Ben-Shahar, 2012; D. Kaiser et al., 2019; Sun et al., 2016; Turini and Võ, 2022; Võ and Henderson, 2009)—a format described by some researchers as a kind of "scene grammar" (Võ et al., 2019; Vo, 2021). While earlier work has richly documented how co-occurrence associations between objects and scenes (e.g., pots appear in kitchens) facilitate object and scene recognition (Davenport and Potter, 2004; Hollingworth and Henderson, 1998; Hwang et al., 2011; Mack and Eckstein, 2011; Oliva and Torralba, 2007; Underwood et al., 2008), structural information captures finer-grained expectations about how objects are arranged and interact within a scene.

Others have extended investigations of scene gist to event perception, examining the kinds of representations that can be formed extrafoveally. For instance, Dobel et al. (2010) found that observers can report Agent and Patient role assignments and the coherence of event scenes presented briefly (<300 ms) and peripherally, but they required fixations on action-relevant areas (e.g., the Agent's hands) to identify many actions, apart from those inferrable from coarse postural information (e.g., *kicking*). Corroborating results have shown that observers can accurately identify the roles and (in some cases) category of action in scenes presented briefly (300 ms) in the periphery (Isasi-Isasmendi et al., 2023; Sauppe and Flecken, 2021). Moreover, a recent study found that observers show categorical-perception effects for event categories (e.g., *pouring* vs. *scooping*) even when scenes are presented peripherally for just 100 ms (Ji and Scholl, 2024).

Together, these studies reveal the types of information available parafoveally and peripherally: coarse postural and animacy cues that allow for role identification and, in some cases, event-category identification. Crucially, the difference here seems to be in the *availability* of information rather than a qualitative difference in *processing*, as similar results occur when foveal information is blurred to simulate the spatial resolution available in the parafovea or periphery (Dobel et al., 2010). Likewise, the rapidity of extraction appears comparable across foveal and extrafoveal input. In brief-exposure paradigms where scenes are initially presented peripherally (e.g., Gerwien and Flecken, 2016; Isasi-Isasmendi et al., 2023; Sauppe and Flecken, 2021), the first saccade toward the event region often occurs within about 200 ms of stimulus onset. Because programming a saccade typically takes about 150 ms, this suggests that the saccade planning itself began within about 50 ms of scene onset, which aligns with findings from brief-display paradigms using central presentation (e.g., the 37-ms displays in Hafri et al. (2013); see also Dobel et al., 2010; Glanemann et al., 2016).

The visual system's ability to rapidly extract relational information—including extrafoveally—does not imply that *all* aspects of causal, spatial, and event relations are fully resolved during initial processing. While relational roles and event categories (e.g., *kicking* or *tapping*) may be extracted from coarse postural information available in lower-resolution extrafoveal input, finer details about participating entities may be left unresolved. For example, peripheral vision may traffic in broad category distinctions such as animacy (animate vs. inanimate) or approximate size (large vs. small), rather than in more specific categories or identities (e.g., distinguishing a dog from a cat, or a doctor from a lawyer) (Freeman and Simoncelli, 2011; Long et al., 2018). Thus, determining the precise category or identity of participants may sometimes require targeted fixations.

Nevertheless, as brief-display paradigms demonstrate, extrafoveal cues may suffice to bind roles to individuals in particular spatial locations even without determining the category of the individual (e.g., recognizing the person on the left as the Agent, even without information about exactly who the Agent is). This dissociation between participant identity and event roles resembles the classic feature-binding problem (Treisman and Gelade, 1980). Supporting this distinction, Hafri et al. (2024) found that participants who were searching for a target image (e.g., knife-in-cup) in a rapid image sequence false-alarmed more to distractor images depicting different objects in the same relation (containment, e.g. phone-in-basket) rather than those in a different relation (e.g., phone-on-basket), suggesting a form of "role-filler independence" in visual processing (see also Vettori et al., 2024a).

There may also be limits on what kind of event information is available at a glance. In particular, thematic role information for more complex event structures (e.g., *transfer events*, such as *giving*) or roles (e.g., Recipient or Instrument) may take more deliberate inspection of an event, as we detail below in Section 3.2.1. Given these limits, overt attentional shifts (indexed by fixations) may be needed to refine or elaborate relational representations that were initially constructed via extrafoveally extracted information (whether through covert attentional mechanisms or otherwise).

This understanding of what kind of event information can (and cannot) be extracted at a glance may help to resolve a puzzle about event perception more broadly: how do the eyes "know where to look" in order to optimally extract additional information about what is taking place? The answer is that the rapid perceptual processes we have been discussing here—including gist extraction for events and other relations—provide a structured scaffold for more targeted information extraction (Wolfe et al., 2011; Võ and Wolfe, 2013). Some kinds of information (e.g., relational categories or event roles) can be triggered by coarse-grain input and accessed without overt attention, while others (e.g., detailed object identity or fine-grained attributes) require high-resolution, localized input.[3] The key distinction, then, lies not in

[2] An important distinction in considering the link between eye movements and attention is that between overt attention—shifting gaze to fixate on a location or object—and covert attention, which allows selective processing of locations without moving the eyes (Posner, 1980). This distinction is related to, but not identical with, the difference between foveal and extrafoveal information extraction. Eye movements generally indicate overt attention, and, by extension, foveal information extraction. By contrast, finding evidence for extrafoveal information extraction is more subtle: some types may depend on covert attention, whereas others may not. While some models of saccadic control hold that covert attention typically entails saccade programming (even when execution is inhibited), we set aside these mechanistic debates about which kinds of information require attention, as our arguments do not hinge on them. For present purposes, saccades can be interpreted as reflecting overt attention—and thus targeted information extraction—toward an object or location, regardless of whether the extrafoveal information was covertly attended beforehand (see Rosenholtz, 2024, for discussion).

[3] This view is necessarily simplified. Because events unfold over time, nonselective "scene" processing—if it includes event information—is inherently dynamic, with representations that continuously update and incorporate memory for prior states (see Davis & Altmann, 2021). Indeed, many everyday interactions depend on such informational histories (e.g., knowing whether your cast-iron pan is hot or cold). For present purposes, we set aside these dynamic aspects to focus on how relational information is discussed in the perception and VWP literatures, which have primarily examined static visual scenes. However, we expect the kinds of rapid relational extraction discussed here to apply to dynamic, temporally unfolding events as well, though likely in a more complex and iterative manner.

the depth or complexity of the representation but in the amount and specificity of visual information required to trigger it.

2.3. The content and format of rapidly extracted relational information

These findings also have broader implications for the content and format of perceptual scene representations and how they interface with higher-level cognitive systems, including language. First, the *content* of such representations extends beyond what is immediately visible in a "snapshot" of an event. It includes inferred information about what has or will likely transpire on or between objects—that is, event-based object *histories* (for a review, see Altmann and Ekves, 2019). For example, in event gist-extraction studies, the "key moment" of an event (e.g., contact between the Agent's foot and the Patient) is sufficient to infer the category *kicking* (Dobel et al., 2010; Glanemann et al., 2016; Hafri et al., 2013). Such effects are so deep enough to alter episodic memory for dynamic events, provided that continuity cues are present (Boger and Strickland, 2025; Kominsky et al., 2021; Strickland and Keil, 2011). Observers even "play forward" state-change events in memory (e.g., melting ice), misremembering objects as more changed than they were (Hafri et al., 2022). Object histories can also be inferred rapidly and spontaneously in visual processing—for instance, whether a shape with a jagged mouth-like indentation must have been bitten (Y.C. Chen and Scholl, 2016) or which objects were placed first to yield a stable block tower (Wong et al., 2025). Thus, both perceptual and memory-based event representations contain content going beyond the co-present scene.

Moreover, the existence of structured representations for rapidly extracted scene content—and in particular the highly abstract content of events (including general roles like Agent and Patient)—suggests that their *format* is not simply imagistic or "picture-like" (in which the parts of the representation correspond to parts of the represented scene; Kosslyn et al., 2006). Instead, the format of these representations may be more language-like than previously assumed—symbolic and abstract—perhaps bearing similarities to the traditional notion of a "language-of-thought" (Fodor, 1975; Quilty-Dunn et al., 2023; Hafri et al., 2023). By virtue of this shared format, these representations are likely to interface seamlessly with higher-level systems for language and reasoning (see, e.g., Altmann and Kamide, 2007; Altmann and Kamide, 2009; Cavanagh, 2021; Hafri et al., 2023; Hafri and Papeo, 2025; Quilty-Dunn, 2020). In particular, perceptual representations may be readily translated into conceptual ones, supporting higher-level inference (e.g., about who did what to whom) and the updating of internal models of unfolding events. These conceptual representations (rather than perceptual ones) are likely those that language accesses and updates, and they may in turn interface with, or modify, spatial models of the visual world—updating where entities are located and how they relate within a scene.

Together, this work suggests that the scene representations that interface with language—and guide eye movements in response—are the product of rich inferences extending beyond what is visually present, and are encoded in a format well suited for efficient interaction with higher-level cognition.

3. Understanding the relationship between linguistic and visual representations of events using the VWP and related tasks

Given the theoretical perspective outlined above, the visual world paradigm (VWP) could provide a compelling window into event perception itself. In many VWP studies, participants are asked to inspect a visual scene and listen to or produce utterances that reference events, event roles, or spatial relations. Because fixations reflect overt attention, eye movements in the VWP could, in principle, provide a moment-to-moment window into how observers go from a coarse-grained event gist to a more fine-grained representation of that event.

However, much existing VWP work, especially in language comprehension, eschews this issue by introducing an extended visual "preview" period before language-relevant aspects of the task begin. This design permits researchers to assume that perceptual processing of the scene or event is complete, and that the subsequent language merely accesses or refers to the *product* of this process. Doing so allows for a simplifying linking assumption common among these studies: that eye movements in the VWP are largely driven by linguistic processes, rather than perceptual ones (for review and discussion, see Apfelbaum et al., 2021; Huettig et al., 2011; Magnuson, 2019; Salverda and Tanenhaus, 2017; Tanenhaus and Trueswell, 2006).

We suggest below, however, that parametric manipulation of preview time within the domain of events may give a clearer picture of how eye movements relate to the encoding of, and linguistic reference to, events. While this increases the complexity of the linking assumptions needed to relate eye movements to cognitive processes, it may be necessary to advance our understanding of how perceptual and linguistic representations interface. Specifically, how do rapidly formed event representations guide online language comprehension, and conversely, how does linguistic content—whether heard, anticipated, planned, or produced—guide apprehension? In the next section, we will review key findings from the VWP comprehension literature that address these issues. We then turn to VWP studies of language production.

3.1. Studies of language comprehension

There is ample evidence from the VWP that the auditory recognition of verbs, prepositions, and other relational terms triggers the rapid retrieval of combinatory linguistic information that pertains to event meaning. For example, hearing the start of a verb phrase (e.g., "The boy eats…") prompts anticipatory looks to semantically plausible objects that afford the described action (e.g., a cake; Altmann and Kamide, 1999), and this effect generalizes across languages with varying word orders, including the verb-initial Mayan language Tseltal (Garrido Rodriguez et al., 2023). Similarly, hearing the start of a prepositional phrase in an instruction (e.g., "Put the duck inside…") leads to anticipatory looks toward objects that could plausibly serve as the grammatical object of the preposition (e.g., containers such as a can; Chambers et al., 2004). For native speakers of case-marking languages, hearing a case-marked noun triggers anticipatory looks toward entities that could plausibly participate in the event denoted by the verb (e.g., E. Kaiser and Trueswell, 2004; Kamide et al., 2003; Özge et al., 2019, Özge et al., 2022). Hearing a locative noun phrase modifier (e.g., "Put the frog that's on/in…") guides eye movements to the object visually perceived as participating in the specified spatial relation (e.g., the one on something or in something; Novick et al., 2008). Finally, auditory verb recognition influences downstream parsing and reference resolution, altering eye movements during sentence processing (e.g., "Tap the frog with…" vs. "Choose the frog with…"; Snedeker and Trueswell, 2004; see also Tanenhaus et al., 1995; Spivey et al., 2002). These VWP studies of language comprehension demonstrate rapid anticipatory activation of relational information, showing how listeners rapidly connect linguistic input to conceptual knowledge of events and other relations made available by the visual displays.

Extending this work, other VWP studies have explored how comprehenders infer event-based information while spoken language unfolds over a viewed scene, anticipating which objects are likely to participate in an event based on their current states or inferred histories. For example, Altmann and Kamide (2007) found that a verb's tense (*will drink* vs. *has drunk*) modulated fixations to objects such as full versus empty glasses. Later work (Altmann and Kamide, 2009) showed that listeners can revise their expectations when language conveys updates about the scene: hearing that a bottle had been moved from the floor to a table enhanced anticipatory looks toward the updated location (the table) upon hearing "The woman will pour…", especially when the conflicting visual display (still showing the bottle on the floor) was

removed. These findings suggest that comprehension of event descriptions involves dynamic inferences about object states and spatial configurations, rather than simple word-to-visual-referent matching, and that such anticipatory eye movements can arise from internally generated event representations even without a co-present visual scene.

Nevertheless, these anticipatory effects do not reveal how this information is visually extracted and represented before the linguistic input arrives, or how quickly this visually apprehended information becomes available to interact with language comprehension in real-time. In other words, they do not reveal *when* event information is represented and *how* it is represented. This is because these studies typically involve scenarios in which participants have ample time to visually interrogate or "preview" the environment (typically for at least one second), during which the construction of relational representations could proceed slowly and incrementally through visual routines, rapidly and holistically through gist extraction, or through some combination of both. Thus, the dynamics of visual processing are not discernible in these cases.

Additionally, most of the VWP comprehension studies described above provided participants with visual environments that did not explicitly depict the linguistically relevant events. Instead, they relied on participants' ability to infer object-action affordances (i.e., the action possibilities of objects based on shape, size, and function) from their spatial arrangement—a process that itself can occur rapidly through spontaneous perception (Gibson, 1977, 1979; Guan and Firestone, 2020; Wong and Scholl, 2024) or more slowly through deliberative inference (Ye et al., 2009; Wagman et al., 2016, 2018; for detailed discussion as it pertains to the VWP, see Chambers, 2016). This is important because the nature of the display can shape what relational information is rapidly available. Simple arrays of isolated objects—often used in early VWP work—eliminate cues external to objects themselves but by the same token lack the structural and semantic constraints present in richer, more naturalistic or realistic scenes. Yet, as Henderson and Ferreira (2004) note, these constraints can actually facilitate rapid gist extraction, meaning that greater scene complexity does not necessarily increase processing demands and may, in fact, support relation extraction.

Notably, some VWP studies *have* examined the comprehension of spoken sentences that referred to explicitly depicted events (e.g., Divjak et al., 2020; Knoeferle et al., 2005; Hoover and Richardson, 2008; Nappa et al., 2009; Mitsugi, 2017; Soroli, 2024), showing that visually apprehended event representations can support real-time sentence processing—but often after substantial scene preview. For example, Knoeferle et al. (2005) recorded eye movements as participants inspected Agent-Patient events (e.g., *washing*, *painting*) and heard verbal descriptions in German that contained temporary syntactic ambiguities related to thematic role assignment. The visual event scenes aided listeners in resolving these initial ambiguities, demonstrating that representations of the perceived events made contact with ongoing linguistic processing. However, because participants were given ample time for scene preview, and because eye movements during preview were not analyzed relative to those that arose during sentence processing, this work cannot reveal how quickly event relations were apprehended or how they interacted with linguistic processing in real time.

Other VWP comprehension studies have examined how linguistic input can bias the ultimate interpretation or construal of depicted events, although again typically measured after extended scene preview (e.g., Divjak et al., 2020; Nappa et al., 2009). Nappa et al. investigated so-called perspective predicates—verbs whose use depends on the perspective adopted by the speaker (e.g., *chase* vs. *flee*, *buy* vs. *sell*, *win* vs. *lose*). Children (ages 3–5) were shown static scenes depicting events such as a bunny chasing an elephant while a speaker on-screen described the event using a novel verb ("He's gorping him!") while gazing from one participant or the other. Older children used these gaze cues to infer the verb's meaning (e.g., a gaze to the bunny indicating a "chase" construal). When syntactic positional cues were also provided ("The bunny is gorping the elephant!" vs. "The elephant is gorping the bunny!"), children overwhelmingly relied on syntax to infer the intended meaning. In short, the linguistic cues of a co-occurring utterance shaped how children construed the same visual event. Notably, however, the utterance began at least 1.5 s after event scene onset, leaving open whether language would alter construal in the absence of such preview time (for related VWP work that also used extensive scene preview, see Divjak et al., 2020). Supporting this possibility, studies of spatial relations using variations of the classic sentence-picture verification task (Clark and Chase, 1972)—where a sentence *precedes* a visual display to be verified—have found that the perspective expressed in the sentence (e.g., *red above blue* vs. *blue below red*) influences attentional patterns and response latencies (Roth and Franconeri, 2012; Yuan et al., 2016; Sun et al., 2025).

These considerations are not merely methodological details to be resolved for improving VWP research; they are both practically and theoretically important. Practically, in everyday conversation, a person's visual environment is constantly changing—objects appear, move, and disappear, and attention itself shifts with locomotion or task demands. Sometimes language use occurs in familiar, predictable contexts (e.g., explaining a televised football game to a friend), and other times in novel ones (e.g., arriving at an unfamiliar party and asking someone to introduce you to the host). Language taps into these dynamic representations, and situating a given VWP preview manipulation along this continuum can clarify its relevance to real-world language use. Theoretically, extended preview time may mask how language itself—and differences across languages—guides event construal and how observers interrogate different components of visual events. The VWP, with its fine-grained temporal sensitivity, is well-suited to probing such influences.

3.1.1. Manipulating preview time to understand how visual information modulates comprehension

Only relatively recently have the temporal dynamics of visual apprehension during VWP language-comprehension tasks become a more central topic of interest (e.g., Apfelbaum et al., 2021; Q. Chen and Mirman, 2015; Ferreira et al., 2013; Huettig et al., 2011; Hintz et al., 2017, 2020; Yee et al., 2011). Key debates center on how much, and what kinds of, information are computed prior to hearing an utterance. Specifically, this work asks whether previewing the visual context allows participants not only to generate an accurate perceptual and conceptual encoding of the environment, but also to preemptively generate expected linguistic codes—ranging from high-level messages to phonemic forms.

Although this line of research has not, to our knowledge, specifically investigated *depicted* events, it has examined how manipulating preview time affects the activation of information about objects and their affordances during comprehension (e.g., Q. Chen and Mirman, 2015; Hintz, Meyer, and Huettig, 2020; Yee et al., 2011). For example, Hintz, Meyer, and Huettig (2020) found that longer preview times restricted anticipatory eye movements to objects with appropriate affordances, excluding objects that were merely semantically associated or visually similar. Hearing "The man will peel…" elicited anticipatory looks to a banana (an affordance-matching object) but not to a monkey (a semantically associated object) or a canoe (a visually similar object). In contrast, shorter visual preview times resulted in anticipatory looks to all three object types compared to unrelated objects.

These findings support two interpretations. One is that extended visual preview leads to the retrieval of linguistic information, such as word-object mappings or even phonological representations (p. 465, Hintz et al., 2017). The other, consistent with the extrafoveal visual-processing work reviewed in Section 2.2 above, is that longer preview times might simply allow for more accurate and detailed visual, spatial, and semantic representations, which linguistic input subsequently accesses. For example, greater certainty that a monkey is not holding a banana could suppress looks to that region when hearing "The man will peel…". By contrast, shorter previews leave perceptual and conceptual

representations more uncertain—perhaps there is uncertainty about whether the monkey is really holding a banana, or perhaps a canoe in the periphery might be mistaken for a banana—increasing attention to semantically or visually related objects. Apfelbaum et al. (2021) provide converging evidence for this interpretation within the domain of comprehending nouns.

Regardless of the specific interpretation, these studies make clear that the amount of time that is available to process visual input prior to hearing speech influences both language comprehension and attentional guidance in real-time.

3.1.2. Extending preview-time manipulations to event apprehension

These preview-time effects suggest a way forward for understanding how event apprehension interacts with language. The perceptual gist-extraction work reviewed in Section 2.1 shows that basic relational information (e.g., who is acting on whom) is rapidly and automatically available even with little or no preview. Furthermore, because this process is automatic, it should remain largely unaffected by the additional cognitive demands of comprehending speech (Endress and Potter, 2012), allowing it to occur concurrently with language processing. It is therefore plausible that within the VWP, ongoing language-comprehension processes could access basic event-gist information even under conditions of little or no preview.

Empirical support for this idea comes from Zwitserlood et al. (2018), who found that briefly flashed (50–150 ms) action scenes that were then masked primed the naming of subsequent actions, including activation of their associated word forms. Crucially, they also found that linguistic primes produced a similar facilitation. Such results suggest that event scenes viewed for just a brief glance provide sufficiently detailed conceptual information to facilitate both scene apprehension and linguistic encoding. Likewise, preliminary VWP evidence that event gist is immediately available to ongoing linguistic processes comes from J. Chen and Trueswell (2025). When asked to select among two depicted events based on spoken linguistic input, observers showed above-chance eyegaze to the target event image at verb offset ("The red person is kicking…") even in the absence of visual preview, i.e., when the depicted events were not displayed until the onset of the verb itself. Given these findings, it is possible that event gist information available from minimal preview could affect real-time language processing, such as aiding in the resolution of syntactic disambiguities (e.g., Knoeferle et al., 2005).

At the same time, important questions remain about the *flexibility* of the initial event apprehension process itself, particularly in contexts of language comprehension (Ferreira et al., 2013). For example, does knowing that there will be language produced by some other co-present observer lead to 'message-level' predictions about their upcoming utterance, as proposed by the Thinking-for-Speaking / Thinking-for-Listening hypothesis (Slobin, 1996, 2003; see also Huettig et al., 2011)? If so, might such expectations influence the apprehension process, i.e., what is extracted from a scene or how it is extracted?

Relatedly, might cross-linguistic differences in how events are typically encoded in utterances influence this process even during visual preview—and perhaps even in situations that do not explicitly involve interpreting or anticipating linguistic input? Recently, Soroli (2024) explored these questions experimentally by comparing French and English speakers' visual interrogation of movies depicting motion events (e.g., *riding, crawling*, etc.) under tasks that either did or did not involve native-language input. While the pattern of findings is complex, they do point to some influences of native language on event interrogation, especially when linguistic input is part of the task, in line with Thinking-for-Listening/Speaking.

These findings point to the importance of considering the limits of rapid event apprehension (as outlined in Section 2.2). While the early extraction of event gist may support language comprehension in many contexts, it does not always yield fully specified representations—especially regarding details such as participant identity or fine-grained role assignment (as discussed in Section 3 below). In such cases, event construal may proceed more incrementally, leaving room for language to guide attention or bias interpretation while details are still being determined (e.g., Nappa et al., 2009). Critically, systematically manipulating preview time—from no preview, to brief glimpses, to extended viewing—provides a means to test whether these language-driven effects emerge only after an initial, relatively "language-neutral" gist has formed or whether they influence event apprehension from the very start.

Such methods could also address broader theoretical debates: Are language-specific effects on event processing deep and enduring, or do they instead arise only transiently in specific online processing contexts? (For extensive discussion of these issues, see Gleitman and Papafragou, 2013). Although preview-time manipulations have proven valuable in object-based comprehension studies, they have yet to be fully leveraged in cross-linguistic research on events, making this a promising direction for future research.

3.2. Studies of language production

Studies of language production (using the VWP and other methods) have been more centrally concerned with the relationship between event apprehension and linguistic encoding (see Ünal et al., 2024, for a review of relevant empirical and theoretical work). Rather than relegating scene apprehension (perception) to a preview stage, even the very first VWP production study (Griffin and Bock, 2000) tackled the issue of the temporal relationship and potential temporal overlap between these two processes. In a series of experiments, participants' eye movements were recorded as they viewed line drawings of two-participant actions (e.g., a mailman chasing a dog) and performed one of several different tasks. In a "Patient-search" task in which participants had to identify the Patient in a scene, eye movements diverged between Agents and Patients after 300 ms, signaling a rapid understanding of the event's relational structure. Patterns in a scene description task corroborated these findings: while eye movements during the first 300 ms did not predict the speaker's choice of a Subject, fixations occurring about one second before speech onset did. Taken together, these findings led Griffin and Bock to suggest that event description in language production unfolds in two stages: an initial stage of holistic event apprehension (akin to rapid gist extraction), followed by a second stage where remaining fixations are dedicated to planning the utterance. This view is highly compatible with (and prescient of) the results sketched above in Section 2 regarding rapid extraction of event gist.

However, subsequent work suggests that this temporal separation between apprehension and linguistic planning may be too strict. In a study similar to Griffin and Bock, Gleitman et al. (2007) found, contrary to the two-stage model, that initial fixation positions on characters in a scene partially predicted the later order-of-mention in participants' descriptions, indicating that perceptual and linguistic processes may interact in a cascading fashion. For instance, participants who first fixated on a man in an image of a dog chasing a man were more likely to produce the (dispreferred) description "A man is running from a dog" than those who first looked at the dog. This pattern held across various perspective-predicate pairs (e.g., *win/lose, give/get, buy/sell*), which describe the same event from different event participants' perspectives. A plausible interpretation that still aligns with Griffin and Bock's two-stage model is that initial fixation positions alter figure-ground assignment (i.e., which event participant is most prominent in the non-linguistic event representation), which subsequently impacts linguistic output. Although we endorse this explanation, Gleitman et al. also found that initial fixation influenced order-of-mention in cases where figure-ground relationships were held constant (e.g., in reciprocal interactions, such as a dog and a cat growling at each other, "a dog and a cat…" versus "a cat and a dog…"). This suggests that early fixation on a character not only affects event apprehension but also directly facilitates linguistic encoding of that character, leading to earlier mention in a

cascading sequence. This cascading model is further supported by studies showing that even parafoveally presented objects can activate lexical representations (e.g., Malpass and Meyer, 2010).

3.2.1. *Sequential attention in encoding complex event structure*

It is also unlikely that event apprehension is always entirely completed in an initial gist extraction stage, especially for more complex or ambiguous events. Studies demonstrating rapid and automatic extraction of event information have largely focused on simple and relatively unambiguous Agent-Patient or symmetrical social interactions, such as those discussed in Section 2.2 (Dobel et al., 2010; Glanemann et al., 2016; Hafri et al., 2013, 2018; Isik et al., 2020; Vettori et al., 2024b). In contrast, for more complex or ambiguous events, fixations—guided by this initial gist—may be needed to refine and update event representations. This refinement may be necessary not only to determine participant identity (as discussed in Section 2.2), but also to resolve aspects of the event structure itself. Evidence for this staged refinement, while still emerging, is supported by several lines of research. For instance, Yin et al. (2022) found asymmetries in how object-transfer events like *giving* and *taking* are encoded in working memory. Participants had more difficulty detecting changes to the participant who lost an object to a taker than to the one who received it from a giver, revealing biases in how events like giving and taking are represented. Similarly, studies of motion events show that Goals are encoded with greater fidelity than Sources, particularly in events involving animate Figures (Lakusta and Landau, 2012; Lakusta et al., 2007; cf. Y. Chen et al. (2024)). Additionally, encoding certain spatial relations, such as one object being above or to the left of another, appears to require deliberate initiation of a visual routine in the form of sequential attentional shifts from the Figure to the Ground object (Franconeri et al., 2012; Yuan et al., 2016).

A more recent role-identification study involving eye-tracking during apprehension further supports the view that roles for certain types of complex events require more effortful (or simply longer) scene analysis to extract. Ünal et al. (2024), inspired by the Patient-search task of Griffin and Bock (2000), examined the time course of identifying roles in static images of caused-motion events, with different participants performing Agent, Patient, Goal, and Instrument searches (see Fig. 3A). Eye movement data showed rapid extraction of each role type, but with systematic delays in line with the Thematic Hierarchy (Baker, 1997; Jackendoff, 1990), where roles differ in prominence: Agents, followed by Patients, Goals, and finally Instruments (see Fig. 3B). This temporal ordering held even when statistically controlling for factors such as size and distance from central fixation. Interestingly, role identification often appeared to depend on overtly attending to other event participants first. For example, participants searching for the Patient (always inanimate) showed evidence of attending to the Agent; those searching for the (inanimate) Goal attended to both Agent and Patient; and those searching for the Instrument attended to the Agent and Patient but not the Goal (see Fig. 3C). While alternative explanations remain to be

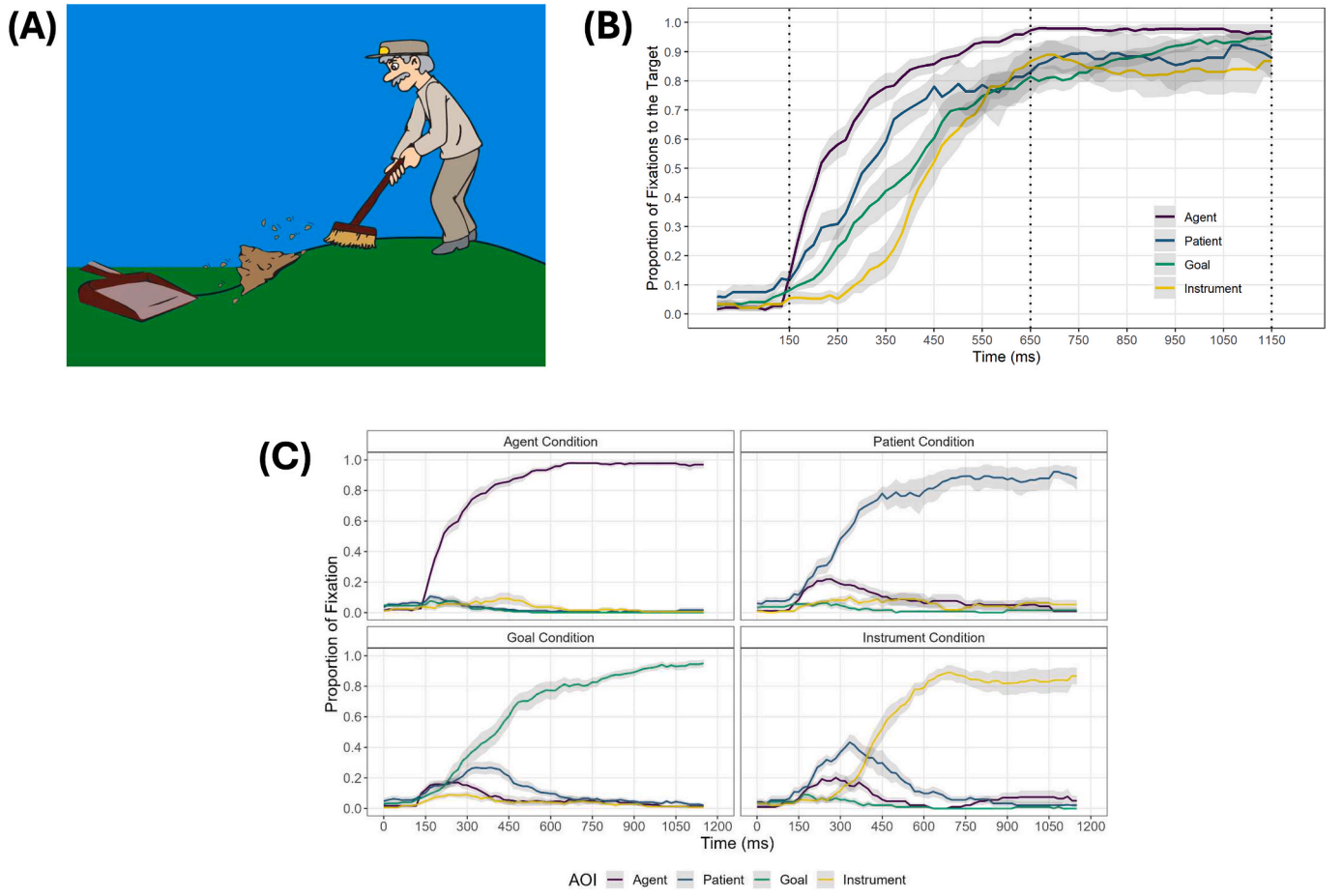


Fig. 3. A. Example stimulus image from Ünal et al. (2024): a caused-motion event involving a man (Agent) pushing dirt (Patient) into a dustpan (Goal) using a broom (Instrument). B. Eye movements (Exp. 2 of Ünal et al.) for four different groups of participants tasked with searching for either the Agent, Patient, Goal, or Instrument. Correct trials only. C. The same eye movement data, including looks to competing (task-incorrect) event roles. Searching for Agents shows little consideration of other roles. Searching for Patients shows consideration of Agents but not other roles. Searching for Goals shows consideration of both Agents and Patients but not Instruments. Searching for Instruments shows consideration of Agents and Patients but not Goals. Shaded areas in B and C indicate standard error of participant means. Figure adapted from Ünal et al. (Copyright 2024).

tested (including the degree to which animacy differences led to certain inferences about roles), these findings suggest that forming a complete event representation that includes certain event components (e.g., Goals) requires targeted, overt attention to other, more prominent components (e.g., Agents and Patients) (see also Wilson et al., 2011; Wilson et al., 2014).

Together, this set of work reveals a more nuanced picture of event apprehension, indicating that apprehension of more complex events (involving roles beyond just Agents and Patients) may proceed through a sequence of interdependent, attention-guided processes. What makes additional attention useful (or perhaps even necessary) for encoding such roles needs to be a topic of further research. It may be that to overcome initial biases in how certain events or other relations are encoded structurally, more effortful visual routines must be initiated, akin to the basic routines for mid-level vision tasks like detecting collinearity or containment in geometric scenes (Ullman, 1984; Ullman, 1996; Jolicoeur et al., 1991; McCormick and Jolicoeur, 1992).

3.2.2. Task-dependent effects of language on visual event apprehension

Beyond these perceptual findings, there is growing evidence that language itself can modulate how events are visually apprehended in certain contexts—a form of "Looking-for-Speaking." Isasi-Isasmendi et al. (2023) compared Basque and Spanish speakers, two populations differing in how their languages mark Agents: Basque overtly case-marks Agent roles (with ergative case), whereas Spanish typically does not. In a scene description task, Basque speakers showed more frequent fixations to Agents than Spanish speakers—a difference that persisted even in a nonlinguistic memory task. Although some effects varied with task order, overall these findings suggest that long-term experience with a particular linguistic encoding can shape overt visual attention in certain linguistic and non-linguistic tasks (see also Gerwien and Flecken, 2016). Relatedly, Sauppe and Flecken (2021) demonstrated that the sentence structure that participants were instructed to use when describing an event (e.g., active vs. passive voice) altered their initial fixations to Agents and Patients when viewing a separate, briefly displayed peripheral image. These results extend the classic "Thinking-for-Speaking" view, showing that even abstract features of linguistic planning (e.g., which role will be mentioned first) can bias where overt attention is allocated during event apprehension.

Other studies converge on a similar conclusion but suggest that such effects may depend on task demands. For example, cross-linguistic work comparing English and Greek speakers has shown differences in how motion events are encoded—English tends to lexicalize manner of motion (e.g., *skip, run*), whereas Greek more often encodes path (e.g., *ascend, exit*)—with corresponding differences in eye movements when memory demands are high or when overt verbalization is required (Papafragou et al., 2008; Trueswell and Papafragou, 2010). Together, these findings suggest that while language experience can bias event construal and attention, enduring differences in early event apprehension appear limited, with the most robust effects emerging when language is actively engaged or when the non-linguistic task is particularly demanding.

In sum, our theoretical position advocates for a cascading model in which event apprehension and linguistic encoding are dynamically interconnected, rather than separated into discrete stages. This view suggests that eye movements in language production tasks serve multiple overlapping functions: they not only reflect the allocation of overt attention for event apprehension but also directly influence the sequence of linguistic encoding, including figure-ground assignment and role order within descriptions. The fact that linguistic factors such as sentence structure or case marking can sometimes influence early fixations—at least when the task involves use of language—suggests that event apprehension is not fully insulated from language. An intriguing implication is that, if linguistic structure is made relevant before or during event construal, similar language-specific effects could also emerge in comprehension, shaping how events are visually apprehended in real time.

4. Conclusions

Here we have offered an examination of the role of eye movements in event apprehension and linguistic processing within the framework of the visual world paradigm (VWP). Across studies (both non-VWP and VWP), a central finding is that relational information, such as who does what to whom, is often extracted rapidly, spontaneously, and sometimes even peripherally. This evidence demonstrates the ability of the visual system to generate structured (and perhaps symbolic) representations of relational content that are readily accessible to higher-level cognitive processes, including language. Furthermore, the content of these representations often includes inferred information about what has transpired or what might soon after (e.g., that a kicking event observed for a brief moment will be carried out in full; or that an ice cube in the process of melting will continue to melt).

At the same time, event apprehension is not always complete at a glance or from extrafoveal input. While coarse relational structure can be established rapidly (e.g., with relational roles such as Agent or Patient bound to entities in certain spatial locations), details such as the precise identities of individual event participants, fine-grained role assignments, or perspective-dependent construals (e.g., *chase* vs. *flee*) may require additional fixations. These open-ended aspects of early event apprehension may create opportunities for language—and cross-linguistic differences in how events are encoded—to guide attention and influence how events are construed, particularly in the context of language comprehension.

Production studies further reveal that event apprehension and linguistic encoding unfold in a cascading, integrated fashion rather than as strictly separate stages. Even early fixations appear to directly influence linguistic planning, including figure–ground assignment and order of mention (e.g., *dog chases man* vs. *man runs from dog*), while later fixations refine initially established role assignments, especially for events involving less prominent roles (e.g., Instruments, Goals, or Recipients). Moreover, language itself can sometimes bias how events are visually inspected, particularly when observers are engaged in demanding tasks and language is available as a tool for encoding observed events in memory.

Together, these findings demonstrate how the VWP can track, in real time, how visual event structure is extracted and selectively refined to meet the demands of comprehension and production. Eye movements provide a direct window into how relational information supports and constrains linguistic interpretation and utterance formulation and offer a way to probe how language may guide attention as visual event representations are being constructed. Preview-time manipulations, especially when combined with cross-linguistic designs, offer a promising path for identifying whether and when language-driven influences emerge and for addressing broader questions about how perception and language work together to shape event understanding.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used OpenAI's ChatGPT and Anthropic's Claude in order to assist in proofreading and improving drafted text. After using this service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

CRediT authorship contribution statement

Alon Hafri: Writing – review & editing, Writing – original draft, Conceptualization. **John C. Trueswell:** Writing – review & editing, Writing – original draft, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Allopenna, P.D., Magnuson, J.S., Tanenhaus, M.K., 1998. Tracking the time course of spoken word recognition using eye movements: evidence for continuous mapping models. J. Mem. Lang. 38 (4), 419–439.

Altmann, G.T.M., Ekves, Z., 2019. Events as intersecting object histories: a new theory of event representation. Psychol. Rev. 126 (6), 817–840. https://doi.org/10.1037/rev0000154.

Altmann, G.T., Kamide, Y., 1999. Incremental interpretation at verbs: restricting the domain of subsequent reference. Cognition 73 (3), 247–264.

Altmann, G.T., Kamide, Y., 2007. The real-time mediation of visual attention by language and world knowledge: linking anticipatory (and other) eye movements to linguistic processing. J. Mem. Lang. 57 (4), 502–518.

Altmann, G.T., Kamide, Y., 2009. Discourse-mediation of the mapping between language and the visual world: eye movements and mental representation. Cognition 111 (1), 55–71.

Apfelbaum, K.S., Klein-Packard, J., McMurray, B., 2021. The pictures who shall not be named: Empirical support for benefits of preview in the Visual World Paradigm. J. Mem. Lang. 121, 104279.

Baker, M.C., 1997. Thematic roles and syntactic structure. In: Haegeman, L. (Ed.), Elements of Grammar: Handbook in Generative Syntax. Springer, Netherlands Dordrecht, pp. 73–137.

Boger, T., Strickland, B., 2025. Object persistence explains event completion. Cognition 259, 106110.

Brown-Schmidt, S., Gunlogson, C., Tanenhaus, M.K., 2008. Addressees distinguish shared from private information when interpreting questions during interactive conversation. Cognition 107 (3), 1122–1134.

Castelhano, M.S., Heaven, C., 2011. Scene context influences without scene gist: Eye movements guided by spatial associations in visual search. Psychon. Bull. Rev. 18, 890–896.

Castelhano, M.S., Henderson, J.M., 2007. Initial scene representations facilitate eye movement guidance in visual search. J. Exp. Psychol. Hum. Percept. Perform. 33 (4), 753.

Cavanagh, P., 2021. The language of vision. Perception 50, 195–215.

Chambers, C., 2016. The role of affordances in visually situated language comprehension. Visually Situated Language Comprehension 12, 205–226.

Chambers, C.G., Tanenhaus, M.K., Magnuson, J.S., 2004. Actions and affordances in syntactic ambiguity resolution. J. Exp. Psychol. Learn. Mem. Cogn. 30 (3), 687.

Chen, Q., Mirman, D., 2015. Interaction between phonological and semantic representations: time matters. Cognit. Sci. 39 (3), 538–558.

Chen, Y.-C., Scholl, B.J., 2016. The perception of history: seeing causal history in static shapes induces illusory motion perception. Psychol. Sci. 27 (6), 923–930. https://doi.org/10.1177/0956797616628525.

Chen, J., & Trueswell, J. (2025). Searching for Events: Rapid visual extraction of language-compatible event representations. In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society.*.

Chen, Y., Trueswell, J., Papafragou, A., 2024. Sources and goals in memory and language: fragility and robustness in event representation. J. Mem. Lang. 135, 104475.

Clark, H.H., Chase, W.G., 1972. On the process of comparing sentences against pictures. Cogn. Psychol. 3 (3), 472–517.

Kaiser, D., Quek, G.L., Cichy, R.M., Peelen, M.V., 2019. Object vision in a structured world. Trends Cogn. Sci. 23 (8), 672–685.

Davenport, J.L., Potter, M.C., 2004. Scene consistency in object and background perception. Psychol. Sci. 15 (8), 559–564. https://doi.org/10.1111/j.0956-7976.2004.00719.x.

Davis, F., Altmann, G.T., 2021. Finding event structure in time: What recurrent neural networks can tell us about event structure in mind. Cognition 213, 104651.

De Freitas, J., Hafri, A., 2024. Moral thin-slicing: forming moral impressions from a brief glance. J. Exp. Soc. Psychol. 112, 104588.

Degen, J., Tanenhaus, M.K., 2016. Availability of alternatives and the processing of scalar implicatures: a visual world eye-tracking study. Cognit. Sci. 40 (1), 172–201.

Divjak, D., Milin, P., Medimorec, S., 2020. Construal in language: a visual-world approach to the effects of linguistic alternations on event perception and conception. Cognit. Linguist. 31 (1), 37–72.

Dobel, C., Gumnior, H., Bölte, J., Zwitserlood, P., 2007. Describing scenes hardly seen. Acta Psychol. 125 (2), 129–143.

Dobel, C., Glanemann, R., Kreysa, H., Zwitserlood, P., & Eisenbeiß, S. (2010). Visual encoding of coherent and non-coherent scenes. In J. Bohnemeyer & E. Pederson (Eds.), *Event Representation in Language and Cognition* (pp. 189–215). Cambridge University Press. https://doi.org/10.1017/CBO9780511782039.009.

Endress, A.D., Potter, M.C., 2012. Early conceptual and linguistic processes operate in independent channels. Psychol. Sci. 23 (3), 235–245. https://doi.org/10.1177/0956797611421485.

Ferreira, F., Foucart, A., Engelhardt, P.E., 2013. Language processing in the visual world: effects of preview, visual complexity, and prediction. J. Mem. Lang. 69 (3), 165–182.

Fodor, J.A., 1975. The language of thought. Harvard University Press.

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.

Freeman, J., Simoncelli, E.P., 2011. Metamers of the ventral stream. Nat. Neurosci. 14 (9), 1195–1201.

Garrido Rodriguez, G., Norcliffe, E., Brown, P., Huettig, F., Levinson, S.C., 2023. Anticipatory processing in a verb-initial mayan language: eye-tracking evidence during sentence comprehension in tseltal. Cognit. Sci. 47 (1), e13292.

Gerwien, J., Flecken, M., 2016. In: First Things First? Top-down Influences on Event Apprehension. Cognitive Science Society, Austin, TX, pp. 2633–2638.

Gibson, J.J., 1977. The theory of affordances. In: Shaw, R., Bransford, J. (Eds.), *Perceiving, Acting, and Knowing*. Lawrence Erlbaum, pp. 67–82.

Gibson, J.J., 1979. The Ecological Approach to Visual Perception. Houghton Mifflin.

Glanemann, R., Zwitserlood, P., Bölte, J., Dobel, C., 2016. Rapid apprehension of the coherence of action scenes. Psychon. Bull. Rev. 23, 1566–1575.

Gleitman, L., Papafragou, A., 2013. Relations between language and thought. In: Reisberg, D. (Ed.), The Oxford Handbook of Cognitive Psychology. Oxford University Press, pp. 504–523. https://doi.org/10.1093/oxfordhb/9780195376746.013.0032.

Gleitman, L.R., January, D., Nappa, R., Trueswell, J.C., 2007. On the give and take between event apprehension and utterance formulation. J. Mem. Lang. 57 (4), 544–569.

Greene, M.R., Oliva, A., 2009a. Recognition of natural scenes from global properties: seeing the forest without representing the trees. Cogn. Psychol. 58 (2), 137–176. https://doi.org/10.1016/j.cogpsych.2008.06.001.

Greene, M.R., Oliva, A., 2009b. The briefest of glances: the time course of natural scene understanding. Psychol. Sci. 20 (4), 464–472. https://doi.org/10.1111/j.1467-9280.2009.02316.x.

Griffin, Z.M., Bock, K., 2000. What the eyes say about speaking. Psychol. Sci. 11 (4), 274–279.

Guan, C., Firestone, C., 2020. Seeing what's possible: disconnected visual parts are confused for their potential wholes. J. Exp. Psychol. Gen. 149 (3), 590.

Hafri, A., Firestone, C., 2021. The perception of relations. Trends Cogn. Sci. 25 (6), 475–492.

Hafri, A., Papeo, L., 2025. The past, present, and future of relation perception. J. Exp. Psychol. Hum. Percept. Perform. 51 (5), 543–546.

Hafri, A., Papafragou, A., Trueswell, J.C., 2013. Getting the gist of events: recognition of two-participant actions from brief displays. J. Exp. Psychol. Gen. 142 (3), 880–905.

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, 336–352.

Hafri, A., Boger, T., Firestone, C., 2022. Melting ice with your mind: representational momentum for physical states. Psychol. Sci. 33 (5), 725–735.

Hafri, A., Green, E.J., Firestone, C., 2023. Compositionality in visual perception [commentary]. Behav. Brain Sci. 46, E277.

Hafri, A., Landau, B., Bonner, M.F., 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.

Heller, D., Grodner, D., Tanenhaus, M.K., 2008. The role of perspective in identifying domains of reference. Cognition 108 (3), 831–836.

Henderson, J.M., Ferreira, F., 2004. Scene perception for psycholinguists. In: Henderson, J.M., Ferreira, F. (Eds.), The Interface of Language, Vision, and Action: Eye Movements and the Visual World. Psychology Press, New York.

Hintz, F., Meyer, A.S., Huettig, F., 2017. Predictors of verb-mediated anticipatory eye movements in the visual world. J. Exp. Psychol. Learn. Mem. Cogn. 43 (9), 1352.

Hintz, F., Meyer, A.S., Huettig, F., 2020. Visual context constrains language-mediated anticipatory eye movements. Q. J. Exp. Psychol. 73 (3), 458–467.

Hoffman, J.E., 1998. Visual attention and eye movements. In: Pashler, H. (Ed.), Attention. Psychology Press, Hove, UK, pp. 119–153.

Holcombe, A.O., Linares, D., Vaziri-Pashkam, M., 2011. Perceiving spatial relations via attentional tracking and shifting. Curr. Biol. 21 (13), 1135–1139.

Hollingworth, A., Henderson, J.M., 1998. Does consistent scene context facilitate object perception? J. Exp. Psychol. Gen. 127 (4), 398.

Hoover, M.A., Richardson, D.C., 2008. When facts go down the rabbit hole: contrasting features and objecthood as indexes to memory. Cognition 108, 533–542.

Huettig, F., Rommers, J., Meyer, A.S., 2011. Using the visual world paradigm to study language processing: a review and critical evaluation. Acta Psychol. 137 (2), 151–171.

Hwang, A.D., Wang, H.C., Pomplun, M., 2011. Semantic guidance of eye movements in real-world scenes. Vision Res. 51 (10), 1192–1205.

Isasi-Isasmendi, A., Andrews, C., Flecken, M., Laka, I., Daum, M.M., Meyer, M., Bickel, B., Sauppe, S., 2023. The agent preference in visual event apprehension. Open Mind 7, 240–282.

Isik, L., Mynick, A., Pantazis, D., Kanwisher, N., 2020. The speed of human social interaction perception. Neuroimage 215, 116844.

Jackendoff, R., 1990. On Larson's treatment of the double object construction. Ling. Inq. 21 (3), 427–456.

Ji, H., Scholl, B.J., 2024. "Visual verbs": dynamic event types are extracted spontaneously during visual perception. J. Exp. Psychol. Gen. 153 (10), 2441.

Jolicoeur, P., Ullman, S., Mackay, M., 1991. Visual curve tracing properties. J. Exp. Psychol. Hum. Percept. Perform. 17 (4), 997–1022. https://doi.org/10.1037/0096-1523.17.4.997.

Josephs, E.L., Draschkow, D., Wolfe, J.M., Võ, M.L.H., 2016. Gist in time: scene semantics and structure enhance recall of searched objects. Acta Psychol. 169, 100–108.

Kadar, I., Ben-Shahar, O., 2012. A perceptual paradigm and psychophysical evidence for hierarchy in scene gist processing. J. Vis. 12 (13), 16.

Kaiser, E., Trueswell, J.C., 2004. The role of discourse context in the processing of a flexible word-order language. Cognition 94 (2), 113–147.

Kamide, Y., Scheepers, C., Altmann, G.T., 2003. Integration of syntactic and semantic information in predictive processing: cross-linguistic evidence from German and English. J. Psycholinguist. Res. 32 (1), 37–55.
Knoeferle, P., Crocker, M.W., Scheepers, C., Pickering, M.J., 2005. The influence of the immediate visual context on incremental thematic role-assignment: evidence from eye-movements in depicted events. Cognition 95 (1), 95–127.
Kominsky, J.F., Scholl, B.J., 2020. Retinotopic adaptation reveals distinct categories of causal perception. Cognition 203, 104339.
Kominsky, J.F., Baker, L., Keil, F.C., Strickland, B., 2021. Causality and continuity close the gaps in event representations. Mem. Cognit. 49 (3), 518–531. https://doi.org/10.3758/s13421-020-01102-9.
Kosslyn, S.M., Thompson, W.L., Ganis, G., 2006. The Case for Mental Imagery. Oxford University Press.
Lakusta, L., Landau, B., 2012. Language and memory for motion events: origins of the asymmetry between source and goal paths. Cognit. Sci. 36 (3), 517–544. https://doi.org/10.1111/j.1551-6709.2011.01220.x.
Lakusta, L., Wagner, L., O'Hearn, K., Landau, B., 2007. Conceptual foundations of spatial language: evidence for a goal bias in infants. Lang. Learn. Dev. 3 (3), 179–197. https://doi.org/10.1080/15475440701360168.
Larson, A.M., Loschky, L.C., 2009. The contributions of central versus peripheral vision to scene gist recognition. J. Vis. 9 (10), 6.
Long, B., Yu, C.P., Konkle, T., 2018. Mid-level visual features explain the high-level categorical organization of the ventral stream. Proc. Natl. Acad. Sci. 115 (38), E9015–E9024.
Mack, S.C., Eckstein, M.P., 2011. Object co-occurrence serves as a contextual cue to guide and facilitate visual search in a natural viewing environment. J. Vis. 11 (9), 9.
Magnuson, J.S., 2019. Fixations in the visual world paradigm: where, when, why? J. Cult. Cognit. Sci. 3 (2), 113–139.
Malpass, D., Meyer, A.S., 2010. The time course of name retrieval during multiple-object naming: evidence from extrafoveal-on-foveal effects. J. Exp. Psychol. Learn. Mem. Cogn. 36 (2), 523.
Marr, D., 1982. Vision: A Computational Investigation Into The Human Representation And Processing of Visual Information. W.H. Freeman, San Francisco, CA.
McCormick, P.A., Jolicoeur, P., 1992. Capturing visual attention and the curve tracing operation. J. Exp. Psychol. Hum. Percept. Perform. 18 (1), 72–89. https://doi.org/10.1037/0096-1523.18.1.72.
Michotte, A. (1946/1963). *The Perception of Causality* (T. R. Miles & E. Miles, Trans.). New York: Basic Books. (Original work published 1946).
Mitsugi, S., 2017. Incremental comprehension of Japanese passives: evidence from the visual-world paradigm. Appl. Psychol. 38 (4), 953–983.
Nappa, R., Wessel, A., McEldoon, K.L., Gleitman, L.R., Trueswell, J.C., 2009. Use of Speaker's gaze and syntax in verb learning. Lang. Learn. Dev. 5 (4), 203–234. https://doi.org/10.1080/15475440903167528.
Novick, J.M., Thompson-Schill, S.L., Trueswell, J.C., 2008. Putting lexical constraints in context into the visual-world paradigm. Cognition 107 (3), 850–903.
Oliva, A., Torralba, A., 2007. The role of context in object recognition. Trends Cogn. Sci. 11 (12), 520–527.
Özge, D., Küntay, A., Snedeker, J., 2019. Why wait for the verb? Turkish speaking children use case markers for incremental language comprehension. Cognition 183, 152–180.
Özge, D., Kornfilt, J., Maquate, K., Küntay, A.C., Snedeker, J., 2022. German-speaking children use sentence-initial case marking for predictive language processing at age four. Cognition 221, 104988.
Papafragou, A., Hulbert, J., Trueswell, J., 2008. Does language guide event perception? Evidence from Eye Movements. Cognition 108 (1), 155–184.
Peng, Y., Ichien, N., Lu, H., 2020. Causal actions enhance perception of continuous body movements. Cognition 194, 104060.
Pereira, E.J., Castelhano, M.S., 2014. Peripheral guidance in scenes: the interaction of scene context and object content. J. Exp. Psychol. Hum. Percept. Perform. 40 (5), 2056.
Posner, M.I., 1980. Orienting of attention. Q. J. Exp. Psychol. 32 (1), 3–25.
Quilty-Dunn, J., 2020. Attention and encapsulation. Mind Lang. 35 (3), 335–349.
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. Behav. Brain Sci. 46, e261.
Rolfs, M., Dambacher, M., Cavanagh, P., 2013. Visual adaptation of the perception of causality. Curr. Biol. 23 (3), 250–254.
Rosenholtz, R., 2024. Visual attention in crisis. Behav. Brain Sci. 1–32. https://doi.org/10.1017/S0140525X24000323.
Roth, J.C., Franconeri, S.L., 2012. Asymmetric coding of categorical spatial relations in both language and vision. Front. Psychol. 3. https://doi.org/10.3389/fpsyg.2012.00464.
Salverda, A. P., & Tanenhaus, M. K. (2017). The visual world paradigm. *Research Methods in Psycholinguistics and the Neurobiology of Language: A practical guide*, 89-110.
Sauppe, S., Flecken, M., 2021. Speaking for seeing: Sentence structure guides visual event apprehension. Cognition 206, 104516.
Slobin, D., 1996. From 'thought and language' to 'thinking for speaking'. In: Gumperz, J., Levinson, S. (Eds.), Rethinking Linguistic Relativity. Cambridge University Press, New York, pp. 70–96.
Slobin, D.I., 2003. Language and thought online: Cognitive consequences of linguistic relativity. In: Gentner, D., Goldin-Meadow, S. (Eds.), Language in Mind: Advances in the Study of Language and Thought. MIT Press, Cambridge, MA, pp. 157–191.
Snedeker, J., Trueswell, J.C., 2004. The developing constraints on parsing decisions: the role of lexical-biases and referential scenes in child and adult sentence processing. Cogn. Psychol. 49 (3), 238–299.
Soroli, E., 2024. How language influences spatial thinking, categorization of motion events, and gaze behavior: a cross-linguistic comparison. Lang. Cogn. 16 (4), 924–968.
Spivey, M.J., Tanenhaus, M.K., Eberhard, K.M., Sedivy, J.C., 2002. Eye movements and spoken language comprehension: effects of visual context on syntactic ambiguity resolution. Cogn. Psychol. 45 (4), 447–481.
Strickland, B., Keil, F., 2011. Event completion: event based inferences distort memory in a matter of seconds. Cognition 121 (3), 409–415. https://doi.org/10.1016/j.cognition.2011.04.007.
Sun, Q., Ren, Y., Zheng, Y., Sun, M., Zheng, Y., 2016. Superordinate level processing has priority over basic-level processing in scene gist recognition. i-Perception 7 (6), 2041669516681307.
Sun, Z., Firestone, C., Hafri, A., 2025. The psychophysics of compositionality: Relational scene perception occurs in a canonical order. PsyArXiv. https://doi.org/10.31234/osf.io/97z4n_v1.
Tanenhaus, M.K., Trueswell, J.C., 2006. Eye movements and spoken language comprehension. In: Handbook of Psycholinguistics. Academic Press, pp. 863–900.
Tanenhaus, M.K., Spivey-Knowlton, M.J., Eberhard, K.M., Sedivy, J.C., 1995. Integration of visual and linguistic information in spoken language comprehension. Science 268 (5217), 1632–1634.
Torralba, A., Oliva, A., Castelhano, M.S., Henderson, J.M., 2006. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. Psychol. Rev. 113 (4), 766.
Treisman, A.M., Gelade, G., 1980. A feature-integration theory of attention. Cogn. Psychol. 12 (1), 97–136. https://doi.org/10.1016/0010-0285(80)90005-5.
Trueswell, J.C., Papafragou, A., 2010. Perceiving and remembering events cross-linguistically: evidence from dual-task paradigms. J. Mem. Lang. 63 (1), 64–82. https://doi.org/10.1016/j.jml.2010.02.006.
Turini, J., Võ, M.L.H., 2022. Hierarchical organization of objects in scenes is reflected in mental representations of objects. Sci. Rep. 12 (1), 20068.
Ullman, S., 1984. Visual routines. Cognition 18 (1–3), 97–159.
Ullman, S., 1996. Visual cognition and visual routines. In: High-Level Vision: Object Recognition and Visual Cognition. MIT Press, Cambridge, MA, pp. 263–315.
Ünal, E., Wilson, F., Trueswell, J., Papafragou, A., 2024. Asymmetries in encoding event roles: evidence from language and cognition. Cognition 250, 105868.
Underwood, G., Templeman, E., Lamming, L., Foulsham, T., 2008. Is attention necessary for object identification? Evidence from eye movements during the inspection of real-world scenes. Conscious. Cogn. 17 (1), 159–170.
Vettori, S., Hochmann, J.R., Papeo, L., 2024a. Fast and automatic processing of relations: the case of containment and support. J. Vis. 24 (10), 840. https://doi.org/10.1167/jov.24.10.840.
Vettori, S., Odin, C., Hochmann, J.-R., Papeo, L., 2024b. A perceptual cue-based mechanism for automatic assignment of thematic agent and patient roles. J. Exp. Psychol. https://doi.org/10.1037/xge0001657.
Vo, M.L.H., 2021. The meaning and structure of scenes. Vision Res. 181, 10–20.
Võ, M.L.H., Henderson, J.M., 2009. Does gravity matter? Effects of semantic and syntactic inconsistencies on the allocation of attention during scene perception. J. Vis. 9 (3), 24.
Võ, M.L.H., Wolfe, J.M., 2013. The interplay of episodic and semantic memory in guiding repeated search in scenes. Cognition 126 (2), 198–212.
Võ, M.L.H., Boettcher, S.E., Draschkow, D., 2019. Reading scenes: how scene grammar guides attention and aids perception in real-world environments. Curr. Opin. Psychol. 29, 205–210.
Wagman, J.B., Caputo, S.E., Stoffregen, T.A., 2016. Hierarchical nesting of affordances in a tool use task. J. Exp. Psychol. Hum. Percept. Perform. 42 (10), 1627.
Wagman, J.B., Stoffregen, T.A., Bai, J., Schloesser, D.S., 2018. Perceiving nested affordances for another person's actions. Q. J. Exp. Psychol. 71 (3), 790–799.
Wilson, F., Papafragou, A., Bunger, A., Trueswell, J., 2011. Rapid extraction of event participants in caused motion events. In: Carlson, L., Hölscher, C., Shipley, T.F. (Eds.), *Proceedings from the 33rd Annual Meeting of the Cognitive Science Society*. Erlbaum, Hillsdale, NJ, pp. 1206–1211.
Wilson, F., Ünal, E., Trueswell, J., & Papafragou, A. (2014). Homologies between language and event cognition: Evidence from event role prominence. Paper presented at the 39th Annual Boston University Conference on Language Development, Boston, MA.
Wolfe, J.M., Võ, M.L.H., Evans, K.K., Greene, M.R., 2011. Visual search in scenes involves selective and nonselective pathways. Trends Cognit. Sci. 15 (2), 77–84.
Wong, K.W., Scholl, B.J., 2024. Spontaneous path tracing in task-irrelevant mazes: Spatial affordances trigger dynamic visual routines. J. Exp. Psychol. Gen. 153 (9), 2230–2238.
Wong, K.W., Shah, A.D., Scholl, B., 2025. Seeing from the ground up: Spontaneous perception of 'causal history' due to intuitive physics. J. Vis. 25 (9), 2006.
Yee, E., Huffstetler, S., Thompson-Schill, S.L., 2011. Function follows form: activation of shape and function features during object identification. J. Exp. Psychol. Gen. 140 (3), 348.
Ye, L., Cardwell, W., Mark, L.S., 2009. Perceiving multiple affordances for objects. Ecol. Psychol. 21 (3), 185–217.
Yin, J., Csibra, G., Tatone, D., 2022. Structural asymmetries in the representation of giving and taking events. Cognition 229, 105248.
Yuan, L., Uttal, D., Franconeri, S., 2016. Are categorical spatial relations encoded by shifting visual attention between objects? PLoS One 11 (10), e0163141.
Zacks, J.M., 2020. Event perception and memory. Annu. Rev. Psychol. 71, 165–191. https://doi.org/10.1146/annurev-psych-010419-051101.
Zwitserlood, P., Bölte, J., Hofmann, R., Meier, C.C., Dobel, C., 2018. Seeing for speaking: Semantic and lexical information provided by briefly presented, naturalistic action scenes. PLoS One 13 (4), e0194762.