Received: 25 May 2023 | Revised: 1 December 2023 | Accepted: 7 December 2023

DOI: 10.1002/wcs.1675

ADVANCED REVIEW

WIREs COGNITIVE SCIENCE WILEY

Predicting attentional allocation in real-world environments: The need to investigate crossmodal semantic guidance

Kira Wegner-Clemens[1] | **George L. Malcolm**[2] | **Sarah Shomstein**[1]

[1]Psychological and Brain Sciences, George Washington University, Washington, DC, USA

[2]School of Psychology, University of East Anglia, Norwich, UK

Correspondence
Sarah Shomstein, Psychological and Brain Sciences, George Washington University, Washington, DC, USA.
Email: shom@gwu.edu

Funding information
National Institutes of Health, Grant/Award Number: F31EY034030; National Science Foundation (NSF), Grant/Award Number: BCS-1921415

Edited by: Wayne Wu, Editor

Abstract

Real-world environments are multisensory, meaningful, and highly complex. To parse these environments in a highly efficient manner, a subset of this information must be selected both within and across modalities. However, the bulk of attention research has been conducted within sensory modalities, with a particular focus on vision. Visual attention research has made great strides, with over a century of research methodically identifying the underlying mechanisms that allow us to select critical visual information. Spatial attention, attention to features, and object-based attention have all been studied extensively. More recently, research has established semantics (meaning) as a key component to allocating attention in real-world scenes, with the meaning of an item or environment affecting visual attentional selection. However, a full understanding of how semantic information modulates real-world attention requires studying more than vision in isolation. The world provides semantic information across all senses, but with this extra information comes greater complexity. Here, we summarize visual attention (including semantic-based visual attention), crossmodal attention, and argue for the importance of studying crossmodal semantic guidance of attention.

This article is categorized under:
Psychology > Attention
Psychology > Perception and Psychophysics

KEYWORDS
attention, attentional prioritization, crossmodal attention, semantics

1 | INTRODUCTION

Imagine walking through a city center. You see the blue sky above, pedestrians hurrying by, and flashing signs in shop windows. You hear cars honking in traffic, birds chirping in the trees, and eavesdrop on other pedestrians. You might stop by a coffee shop, enticed by the smell of fresh coffee and baked bread, or run into a neighbor walking their dog. If you turn the corner and encounter an unexpected event, like a parade, the colorful floats and cheering crowds integrate into your internal scene representation without any appreciable delay. Moment-to-moment, the sensory environment can change with new information flooding each sense.

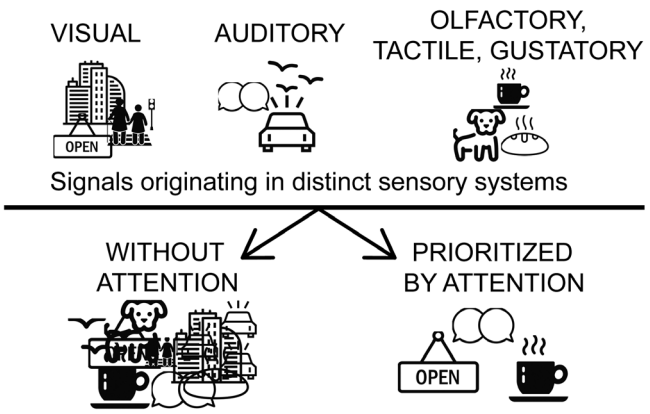

FIGURE 1 A "deconstructed" set of input signals impinging on various sensory systems: visual, auditory, tactile, olfactory, and gustatory information, each gathered by the corresponding sensory organ. Without attentional prioritization, this information competes simultaneously for the limited capacity of human cognition and is not readily interpretable. With attentional prioritization, a subset is selected for further processing that will facilitate our interactions with the world. This review focuses on how this prioritization happens in real-world multisensory environments.

Continuous information bombarding our senses would quickly overwhelm the limited capacity of human cognition without a selection process (as represented conceptually in Figure 1). To prevent important information from being lost, sensory signals are processed in a highly flexible manner determined by a number of factors that act to prioritize a subset information for further processing. In particular, various efforts have been undertaken to characterize how low-level visual features (e.g., location, color, size) and task demands shape attentional prioritization (Carrasco, 2011). While there have been giant leaps in our understanding of attentional mechanisms (Geng et al., 2019; see special issue), much of this work has been made solely within the visual modality. Additionally, most advances in our understanding of attentional prioritization have been made by elucidating how a particular unit of attention (e.g., space, features, objects, hysteresis) contributes to attentional selection in isolation, rather than how they interact to produce attentional prioritization (a more thorough discussion of these units follows in the next section). To understand how attention is guided in real-world complex environments, we need to better understand how various units of attention interact with one another and across sensory modalities. Here, we focus on crossmodal attentional orienting between audition and vision.

Let's get back to the example of walking in the city and the different representations that exist in the environment (or units) which are available for attentional selection and that can subsequently push and pull attentional prioritization. While enjoying your stroll through a city, your goal (or task) might be to attend to street signs to take the correct route. Your attention might be momentarily pulled away (captured) from this task by a signal (visual or from another sense) that is particularly salient, such as a movie theater's bright flashing sign or a honking car. General knowledge about environments, objects, and relationships between objects can also strongly influence attentional allocation. If someone holding a disposable coffee cup passes by, you may notice a nearby coffee shop because you know a container of that particular shape, size, and color, or the distinctive smell of fresh coffee, is generally bought at coffee shops. The meaningful information we have about one object (the coffee cup, the smell of coffee) leads to prioritization of another (the coffee shop). This example points out that not only simple features of the environment influence attentional allocation, but also that semantic relationships (such as that between a coffee cup and a coffee shop) have the potential to shape attention allocation in every real-world scene, both within and across modalities. These complex, semantically rich, multisensory representations of the objects and environments are key to how we allocate attention in real-world contexts, but their role and interactions have not been studied as extensively as other units of attention, be it low- or mid-level visual features. We argue here that the role of semantics and multisensory interactions must be studied in more depth to robustly understand attentional allocation in the real world.

In the sections that follow, we first briefly review some of the units of attentional orienting within the visual modality (note that it should not be taken as an exhaustive review). We then make a case for the importance of investigating how these identified visual units of attentional interact with similar information from other sensory modalities, with a particular focus on audition, and on higher order amodal semantic information. We conclude with suggestions for lines of future research inquiries.

2 | VISUAL ATTENTION: UNITS AND PROCESSES

Attention is, broadly speaking, the process of selection by which a subset of sensory information is prioritized over other sensory information. Selected information is then benefited by faster and more thorough processing (Carrasco, 2011). Most attention research conducted over the past 70 years has taken an approach of simplifying and controlling visual stimuli to isolate the mechanism of visual attentional selection. By breaking down complex visual scenes into well-controlled stimuli and paradigms, various "units" of visual information and attentional processes have been identified. These attention processes include, but are not limited to spatial attention, feature-based attention, object-based attention, task dependent attention, and guidance by semantics (meaning). Each is briefly summarized here to provide a broad understanding of major findings and theoretical framings from this approach.

2.1 | Spatial location

Visual information is intrinsically spatially coded, given the retinotopic organization of the visual system (Engel et al., 1997). Light is initially collected and processed on the retina in a spatially specific manner, with rods and cones relaying information related to the brightness and color of a particular location in the visual field. Cortical processing transforms this retinal input into recognizable shapes and objects while maintaining the spatial information from the retina, fundamentally tying visual information to spatial location (Golomb et al., 2008) even in object-specific ventral cortex (Arcaro et al., 2009; Ayzenberg & Behrmann, 2022; Uyar et al., 2016).

Studies of the effects of visual attentional selection have revealed behavioral facilitation in attended spatial locations, such that response time to detect or identify targets within the attended spatial locations are faster than at unattended locations. Since 1970, benefits of spatial selection have been consistently and robustly demonstrated, mostly relying on variants of a spatial cuing paradigm originally described by Posner (1980). In this paradigm (represented in Figure 2a) a spatial location is cued (here with an endogenous cue) and sensory information at that location is prioritized over other locations. The location can be cued either directly through a salient exogenous spatial sensory event, or

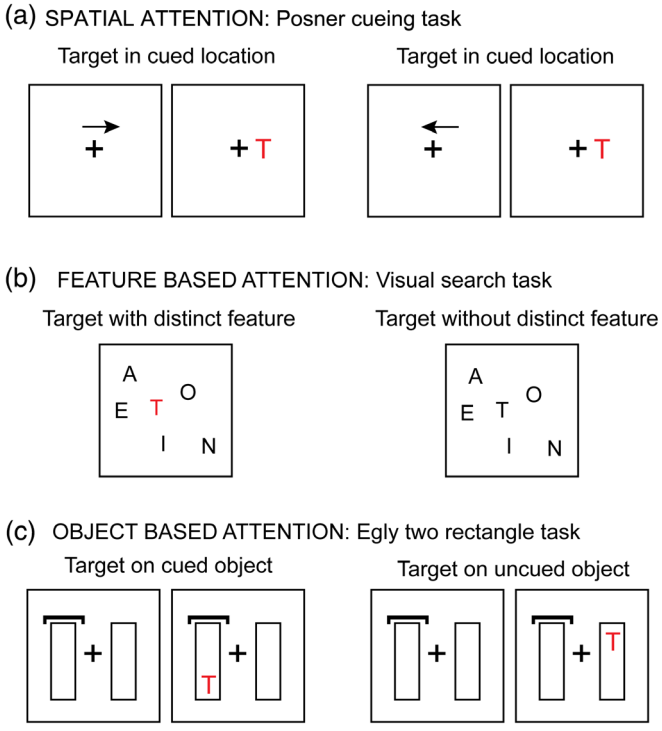

FIGURE 2 Schematics of classic visual attention, demonstrating (a) attention prioritization for cued spatial locations over uncued spatial locations, (b) attentional prioritization for distinct features over non-distinct features, and (c) attention prioritization for cued objects over uncued objects. In each panel, the target is represented by a "T" and the example trial condition on the left tends to prioritized over the sample trial condition on the right, as measured by response time and accuracy.

with an endogenous cue, such as the arrow in Figure 2a. Following the cue, visual targets are detected and identified more quickly if they appear in or near the cued location, relative to the non-cued locations, indicating that the cue has attracted attention toward its location (for review, see: Carrasco, 2011; Eriksen & Eriksen, 1974; Posner, 1980). Importantly, it is also demonstrated that unilateral damage to the posterior parietal lobe impairs the ability to orient attention, and particularly to shift attention from the unimpaired hemifield (the hemifield ipsilateral to the damage) into the impaired, contralateral hemifield (Mesulam, 1999).

Consistent with the role of the posterior parietal cortex (PPC) in spatial attention, studies employing functional magnetic resonance imaging (fMRI) found that corresponding frontal (frontal eye fields, FEF) and parietal areas in human inferior parietal sulcus (IPL), superior parietal lobule (SPL), and temporo-parietal junction (TPJ) contain topographic representations related to saccade planning and attention (Husain & Nachev, 2007; Molenberghs et al., 2007; Serences & Yantis, 2006; Sheremata & Silver, 2015; Silver & Kastner, 2009). For example, one robust finding is that when cues direct attention to specific visual field locations, activation is noted in superior frontal, inferior parietal, and superior temporal cortices (for review see Corbetta & Shulman, 2002; Shomstein, 2012). Voluntary deployments of spatial attention are associated with neural activity in regions of the dorsal parietal cortex (IPL, SPL) and frontal area FEF while involuntary spatial orienting, such as attentional capture by a perceptual singleton, is associated with ventral parietal cortex (TPJ) and ventral frontal cortex (VFC) (Downar et al., 2000; Serences et al., 2005; Shomstein et al., 2010).

2.2 | Feature

Spatial locations that are worth attending to are rarely empty. Rather, they contain objects that are made up of visual features (Buswell, 1935; Yarbus, 1967). For example, that cup of coffee that was carried by a passerby in our city walking example from above can be deconstructed into a basic geometric shape, as well as its colors. Shape and color are just some of the mid-level features that exist in the environment that extend beyond spatial properties. Studies showing that features guide attentional selection have capitalized on tasks that control for spatial attention, while varying what feature is selected from the visual stimulus. Particularly, it has been shown that visual features such as color, motion, and orientation can be prioritized regardless of where they are located within a visual scene (Liu, 2019; Maunsell & Treue, 2006; Wolfe & Horowitz, 2004). Visual search paradigms have been particularly useful for studying feature-based attention since participants can be instructed to search for a target with a given feature without providing any information about the spatial location (for reviews, see Carrasco, 2011; Nakayama & Martini, 2011). Feature-based attention is not strictly a matter of spatial search through an array looking for items that match for a feature, but rather can modulate attention even outside of the focus of spatial attention. Features are more likely to attract attention when they contrast with their surrounding features, like a red square in a visual array of green squares (Itti & Koch, 2000; Parkhurst et al., 2002) or as represented in Figure 2b, a red letter among black letters. Targets with a unique feature will "pop out" and rapidly capture attention, even in large displays (Neisser, 1963; Treisman, 1985; Treisman & Gelade, 1980).

Similarly to spatial attentional orienting, regions within the fronto-parietal network have been implicated in control of feature-based attention. In most paradigms, two superimposed features are presented, for example, motion and color (colored dots move in a particular orientation) or color and orientation (colored lines that vary in orientation) (Liu et al., 2011; Liu & Hou, 2013). For these particular paradigms, given that spatial attention is held constant, any observed neural difference would have to be attributed to feature-based selection. Signals specific to feature-based attention have been identified within the intraparietal sulcus and areas within the precentral sulcus (Jigo et al., 2018).

2.3 | Object

Features, described in the previous section, almost never exist in isolation. Features make up objects that exist in our environment. Until the early 1980s, it was widely assumed that attention is typically directed in a space-based or feature-based manner. Starting in the early 1980s, evidence began to accumulate that some tasks engage a selective mechanism that operates on an object-based, in addition to a location-based, representation (Duncan, 1984; Rock & Gutman, 1981).

A large body of evidence in support of object-based attentional orienting have been gleaned from many different studies employing superimposed objects (Duncan, 1984; O'Craven et al., 1999; Serences et al., 2004; Valdes-Sosa et al., 1998) and what is known as a two-rectangle paradigm, originally developed by (Egly et al., 1994). In the two-

rectangle paradigm (represented in Figure 2c), two adjacent rectangles are presented parallel to one another in such a way as to equate distance between each end of the rectangle and the fixation point. A participant's spatial attention is cued to one of the ends of the rectangle, and after a brief delay a target event appears either within the cued location or in one of the two equidistant ends of the rectangle. In the key finding from this paradigm, targets that are the same spatial distance from the cue are prioritized differently based on which of the two rectangles they appear on. That is, targets are detected faster and more accurately when they appear in an invalid (uncued) location on the same object than targets that appear in an invalid different-object location. The latter finding reflects the contribution of object-based attention to the quality of perception, indicating that other dimensions (e.g., spatial locations) of objects are facilitated by virtue of being part of the cued object. This paradigm has been extended in several subsequent studies investigating the role of object-based attention in visual perception, rendering the findings, for the most part, robust and replicable (Shomstein, 2012).

Several neuroimaging studies have directly demonstrated that object-based orienting engages attentional control regions within the inferior parietal lobule (IPL) (Arrington et al., 2000; Lee & Shomstein, 2013; Müller & Kleinschmidt, 2003; O'Craven et al., 1999; Serences et al., 2004; Shomstein & Behrmann, 2006; Valdes-Sosa et al., 1998), and that early sensory regions that correspond to spatial locations bound by the attended object reflect sensory enhancement consistent with object-based attentional selection. The neuroimaging results suggest that the neural mechanisms underlying object-based attention involve integration of space- and object-based representations within the left IPL and earlier sensory regions of the visual system. A dynamic circuit between the parietal and sensory visual regions may enable observers to preferentially focus on objects of interest that appear in complex visual scenes.

2.4 | Task

In addition to visual "units" (spatial location, feature, and object) that can be selectively prioritized, there are mechanisms by which a unit might be prioritized, such as task relevance. Task where a viewer will fixate in a scene (Yarbus, 1967). In a landmark study by Yarbus, participants were asked to view the same painting multiple times while their eye movements were tracked, but given different tasks, such that the elements of the scene that were relevant to the task changes (Yarbus, 1967). Viewers alternatively fixated different locations in the painting, despite the actual visual information available being the same. Viewers will even often attend to task-relevant items over highly salient scene regions (Foulsham & Underwood, 2007; Henderson et al., 2009), allowing for efficient interaction with the world.

Tasks flexibly change which features and units of information are prioritized on a moment-to-moment basis. When looking for a traffic light before crossing a road, the immediate task might trigger feature-based attention for the physical properties of the traffic light (rectangular with red, yellow, and green lights) and spatial attention to its usual location (generally found on street corners, high up off the ground) (Eckstein et al., 2006; Malcolm & Henderson, 2010; Neider & Zelinsky, 2006). When the viewer's task updates, e.g., cross the road, the new goal (navigate through the environment) similarly drives attentional deployment to new feature and spatial properties (white stripes outlining the crosswalk on the ground, gaps between approaching pedestrians, etc.). In this way, it is the features and spatial arrangement of the world interacting with the viewer's current task which determines where we visually attend to in a scene.

3 | VISUAL ATTENTION: BEYOND LOW- AND MID-LEVEL UNITS OF ATTENTION

3.1 | Semantic relatedness

The world we observe is not solely made up of low- and mid-level sensory information (as in spatial locations, features, and object shapes), but is also rich with semantic meaning, which includes information related to an object's identity, how an object relates to or interacts with other objects in the scene, and in what contexts an object is likely to appear. Vision research, in particular, has shown the importance of semantic information in guiding attention. Both the global meaning of a scene (Eckstein et al., 2006; Torralba et al., 2006) as well as local meaning such as the identity and spatial relationship of objects (Peacock et al., 2019; Wu et al., 2014) can direct eye movements to relevant spatial locations. In particular, expected spatial relationships, or *spatial-licenses*, shape expectations about the location of objects in a visual scene (Hollingworth & Henderson, 2002; Malcolm, Groen, & Baker, 2016; Torralba et al., 2006). These spatial licenses

can be physical: objects generally are supported by surfaces rather than floating in space (Demiral et al., 2012; Võ et al., 2019). Alternatively, spatial-licenses can be semantic: how an object is used provides probabilistic information about its likely location. For instance, when given a description of the function of an invented object, participants were able to find it more quickly if it was in a location consistent with its function, despite no prior visual experience with the object in a scene (Castelhano & Witherspoon, 2016).

Attention can also be guided by more abstract high-level semantic relationships (Belke et al., 2008; Hwang et al., 2011; Mack & Eckstein, 2011; Moores et al., 2003; Neider & Zelinsky, 2006; Shomstein et al., 2019). For instance, if you see a small dog, you expect to look up and see an owner nearby. This type of object-based semantic bias of attention has been suggested to intrude even when it is irrelevant to your goal (Cornelissen & Võ, 2017; Malcolm, Rattinger, & Shomstein, 2016; Nah et al., 2019, 2021; Peacock et al., 2019) suggesting that once an object is in working memory—either from a cue or direct fixation—it activates a semantic network that biases attention within a scene. In this manner, certain items will be prioritized without intentional deployment of attention.

Recent neuropsychological and neuroimaging literature provides evidence that semantic information is derived by a broadly distributed neural network, lateralized toward the left hemisphere (Binder et al., 2009). Specifically, the left inferior frontal gyrus (IFG) has been demonstrated to be crucial in the control of semantic information, including the retrieval and evaluation of meaning (Fiez, 1997; Gabrieli et al., 1996; Thompson-Schill et al., 1997; A. D. Wagner et al., 2001; R. K. Wagner et al., 1997) and as a key region that computes semantic similarity (Carota et al., 2017). Additionally, several recent studies point to the pervasiveness of task-relevant semantic information within the perceptual system (Bar & Aminoff, 2003; Binder et al., 2009; Brady & Oliva, 2008; Greene & Fei-Fei, 2014; Livne & Bar, 2016; Lupyan & Spivey, 2010; Thompson-Schill et al., 1997) with direct evidence of semantic representations decoded throughout the ventral visual cortex, from the occipital to temporal pole (Çukur et al., 2013). Importantly, Nah et al., 2021, in a recent study, fleshed out the neural network that is responsible for semantic influence on attentional selection which is independent of the task. It was demonstrated that the left IFG shows sensitivity to object semantic relatedness, with activity in IFG directly predicting the degree of behavioral benefit of faster response times for targets that appear on task-irrelevant semantically related objects. It was also shown that semantic relatedness biases spatial attention maps in the intraparietal sulcus, subsequently modulating early visual cortex activity.

4 | SENSORY INFORMATION IS SELECTIVELY PRIORITIZED BOTH WITHIN AND BETWEEN MODALITIES

Unisensory studies of attention often rely on a core assumption: findings about attentional processes identified in a study of one sensory modality generalize to attention as a whole. The visual attention findings described above robustly describe how attended visual information is selected over non-attended information at other visual spatial locations, with other visual features, or associated with other visual objects. The mechanisms identified through this vision-only research are no doubt important in real-world settings. However, if attentional resources are distributed in a dynamic fashion across sensory systems, studies of attention in one sensory modality in isolation cannot fully capture the processing underlying prioritization in real-world scenes. A cue or a distractor may differentially influence attention to a target, depending on whether it is in the same or a different modality than the target. For example, a bright light (visual distractor) and equally intense loud music (an auditory distractor) may differentially impact someone's ability to attend to the book they're reading (a visual task), even if at the same signal intensity or spatial location (Figure 3). To understand how signal intensity modulates distraction from a visual task, it is necessary to understand both how same and different modality distractors influence attention to that task. Above, we provide a brief summary of the *visual* units and processes that shape visual attention. Below, we briefly summarize how *auditory* units and processes have been shown to shape visual attention (Driver & Spence, 1998; Hillyard et al., 2016; Koelewijn et al., 2010; Noppeney, 2021; Spence, 2020; Talsma, 2015; Talsma et al., 2010), which provides an example of how two sensory modalities may interact to produce attentional prioritization.

4.1 | Crossmodal interactions

Multisensory environments fundamentally consist of information from different sensory modalities competing for attentional prioritization (Driver & Spence, 1998; Talsma et al., 2010). This information can be integrated into a

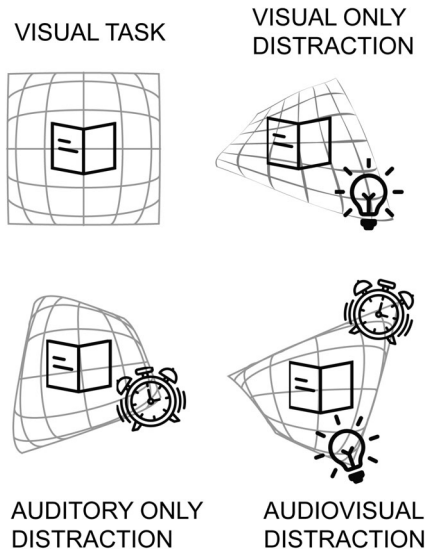

FIGURE 3 A conceptual schematic of how an attentional priority map during a visual task could be differentially changed by a visual only distractor, auditory only distractor, or audiovisual distractor – necessitating direct study of each condition.

coherent whole (multisensory integration), but direct integration is only one of a broader class of crossmodal interactions that include multisensory integration, crossmodal correspondences between stimuli with shared features, as well as contexts where a stimuli from one modality influences memory or attentional prioritization of another. Each sensory modality fundamentally captures a distinct but overlapping subset of information about the environment and is coded according to different properties of the stimulus within the corresponding sensory cortices. Some information is represented through multiple senses, while other information is unique to a particular sense. For example, information about an object's texture is processed by the visual and somatosensory systems, while information about an object's color can only be processed by vision. It then follows that sensory information is selected for attention either through factors that are unique to a particular modality or factors that extend across multiple modalities. For example, color information is specific to vision, so attentional prioritization for a particular color can be explained entirely through prioritization of visual information. If you are looking through a room for something red and see a red alarm clock by the bedside, the visual information associated with the alarm clock (e.g., round, on the nightstand) was prioritized due to a visual feature (redness). However, that same visual information could have been prioritized on the basis of an auditory feature. If you hear an alarm ringing, you can use interaural differences in the auditory signal to identify the ringing's spatial location and selectively prioritize visual information at the corresponding location. In this way, information that is specific to audition (the alarm sound) could selectively prioritize information specific to vision (the alarm clock's face, showing the time).

Carefully controlled studies of multisensory perception have begun to establish the underlying mechanisms and principles of how attention might transfer from one sensory modality to another. Auditory spatial cues improve perception of visual information (McDonald et al., 2000) and correspond to increased processing for that location (Busse et al., 2005; McDonald et al., 2013, 2003). These effects are not simply spatial in nature, but extend to learned associations. In a study where participants were taught associations between sounds and visual images, those learned associations between crossmodal features result in interference in a crossmodal flanker spatial attention task (Jensen et al., 2020). Spatial coincidence or learned associations can lead to multisensory attention effects, even though the perceiver is not necessarily combining them into a unified integrated object.

In the real world, simultaneous multisensory signals often originate from the same source and could be prioritized as an integrated whole. In certain situations, information is part of some overall integrated multisensory object or event, for example, an alarm clock ringing, in contrast to information originating from disparate sources, for example, a security alarm going off while looking at a clock on the wall. With the alarm clock, the auditory and visual signals are not only occurring at a similar time and place, but are caused by the same event and are perceived as a coherent whole through multisensory integration. Multisensory integration is a complementary process that combines subsets of signals across modalities, ensuring that information from the same event is grouped together and information from different events are kept apart. Multisensory events (e.g., passing a barking dog on the sidewalk) are collections of unisensory

signals (e.g., seeing and hearing the dog) bound together and kept separate from other overlapping multisensory events (e.g., an ambulance passing by).

Multisensory integration can occur pre-attentively, with audiovisual integration effects observed in patients with no awareness of the visual information (Bertelson et al., 2000). When multisensory information is integrated pre-attentively, some evidence shows that multisensory events are particularly effective at capturing attention because the redundant information from multiple sensory systems renders an item more salient. That is, "multisensory" may itself be a feature that leads to additional prioritization. In a multisensory event, the sensory systems have captured redundant information through different sensory systems. The additional redundant information may allow these objects to compete for attention more effectively: a cup of coffee might preferentially capture your attention because you can smell and see it at the same time and there is more sensory information in total.

Simultaneous information from multiple sensory systems has been shown to improve search even when the information itself is not relevant to an ongoing task. In the classic redundant signals paradigm, faster response times are observed for audiovisual over visual only or auditory only signals (Hershenson, 1962; Miller, 1982). This effect was extended to search through the "pip and pop effect," where a sound improved visual search performance even though it did not provide helpful information for the search task. Participants were presented an array of line segments and asked to identify either the vertical or horizontal line with trials either having no sound or a task-irrelevant tone. Despite being strictly visual task and the tone not providing any task-relevant information, participants responded faster to visual cues presented with a tone (van der Burg et al., 2008). Similarly, simultaneous auditory tones were shown to increase the magnitude of spatial cuing effects in visual search (Matusz & Eimer, 2011). These effects suggest there is some general attention benefit to having information available in multiple sensory modalities.

In a more direct comparison of uni- and multisensory stimuli, Santangelo and Spence (2007) had participants report the location of a visual target following either auditory, visual, or audiovisual cues. The multisensory cues captured attention even under high attentional load of a secondary letter identification task, while the unisensory auditory and visual cues did not capture attention when the load was high. Further research has shown that objects encountered through multiple sensory modalities are remembered to a greater extent than unisensory (Duarte et al., 2022; Thelen et al., 2014), potentially suggesting improved encoding due to greater attentional allocation. These findings support the suggestion that multisensory stimuli are particularly salient, since the effects are maintained even when other factors are competing for attention.

However, other findings show that the additional salience afforded to multisensory events is highly sensitive to the overall context, rather than an inherent feature of multisensory events due to the increased saliency of redundant information. These findings are consistent with studies showing that attention sometimes precedes multisensory integration, such that information is only integrated when both modalities are attended (Talsma et al., 2007). In a recent study, multisensory targets were found more quickly than unisensory targets in a search task, but when the same stimuli were used as distractors in a visual search task, the multisensory stimuli were not any more distracting than the unisensory stimuli (Lunn, et al 2019). A further study showed that multisensory distractors only interfered with task performance when they were within the attentional set as either a valid cue or a potential target (Spence, 2020). Similarly, crossmodal attention effects are not necessarily bidirectional. Auditory distractors for visual tasks and visual distractors for auditory tasks will not necessarily have an equivalent effect, which would be expected if the attention benefit is derived from there simply being more information available (Ward et al., 2000). The mixed results suggest that the influence of multisensory events is more complex than that events with more information are prioritized to a greater extent. Instead, multisensory objects interact with other factors (e.g., task relevance) in attention. Further research is necessary to unravel how multisensory signals are prioritized among the other ongoing factors impinging on attentional allocation.

4.2 | Neural mechanisms underlying audiovisual and visuotactile attention

Visual information inherently occupies a spatial location, so any attentional modulation ultimately results in a change in attention at a given spatial location (whether within or across modalities). The parietal cortex has been previously identified as a potential location for a spatial map of attentional prioritization (Behrmann et al., 2004; Bisley & Goldberg, 2010; Grefkes et al., 2002; Shomstein et al., 2022). These identified priority maps have been shown to be multisensory (Macaluso et al., 2003), but each sensory system processes spatial information in a manner specific to the sensory organ's biological structure (e.g., visual spatial information is retinotopic, while somatosensory spatial information

depends on receptor location). It is not fully understood how this information is coordinated and collapsed into a spatial priority map that would allow for the behavioral effects in crossmodal attention to be observed. If attentional selection occurs primarily through differential regulation of signals in the parietal maps, this information must all be remapped into a shared coordinate system that is continuously updated. Attention to a visual feature of an object can result in coactivation of auditory features of that object, and vice versa with auditory object and a visual feature (Busse et al., 2005; Molholm et al., 2007; Talsma et al., 2007). The parietal cortex does receive direct inputs from visual, auditory, and somatosensory cortices and represents multisensory information (Behrmann et al., 2004; Kravitz et al., 2011) and is therefore a likely candidate for this integration. Evidence has also shown that remapping does occur: visual input can be remapped to a coordinate system consistent with somatosensory input to underlie reaching behavior (Sereno & Huang, 2014) and auditory information has been shown to be converted to retinotopic coordinates, with sound information even remapped during saccades in parietal and intraparietal areas (Schut et al., 2018; Szinte et al., 2020). This evidence for remapping in parietal cortex is consistent with studies identifying locations for computations in the multisensory processing hierarchy, with sensory fusion processing linked to parietal–temporal regions (Cao et al., 2019) and intraparietal sulcus (Rohe & Noppeney, 2015). In addition to spatial coordinate remapping, crossmodal information can be lined up through temporal processing, which may explain the crossmodal effects in sensory modalities with lower spatial precision than vision. Oscillatory attention signaling has also been proposed as a mechanism by which sensory information is coordinated (for review, see: Senkowski et al., 2008; van Atteveldt et al., 2014). This entrainment may allow for better temporal prediction. These various mechanisms of multisensory processing must dynamically coordinate with other higher level, such as prior expectations and semantic processing, which remains an open question.

5 | HOW DOES SEMANTIC KNOWLEDGE GUIDE ATTENTION IN AUDIOVISUAL CONTEXTS?

Investigating the role of crossmodal information can deepen our understanding spatial, feature, object based, and task-relevant attention. However, it is particularly crucial to understand the role of multisensory information in studies of semantic guidance of attention. As described briefly above, semantics are known to strongly guide attentional prioritization. Semantic information is derived from our prior experience, taken from any available sensory information. Your past experiences with a kitchen were multisensory in nature (the look of a fridge, the whistle of the kettle, the smell of spices cooking). These multisensory experiences resulted in semantic information that is either an amodal (not specific to any sensory modality) representation or crossmodal due to semantic associations across sensory modalities. To fully understand real-world attention in meaningful multisensory environments, the role of semantic information in guiding attention must be studied in greater depth. Semantic guidance in audiovisual attention has often been studied by displaying pictures of objects with "characteristic" sounds, that is sounds that are strongly associated and matched to the visual information such as a picture of a cat with a meow sound. These characteristic sounds—even when they do not provide spatial information (they come from central speakers)—can direct spatial visual attention (Iordanescu et al., 2010, 2008).

In a series of visual search paradigms, participants were shown an array of objects and asked to indicate the location of a target, while a sound played that was either related to the target, another item in the array, or an object not in the display. For example, when the target was a dog, hearing a bark facilitated visual search speed. However, when the sound related to a distractor item (e.g., the sound of flushing when a toilet was also in the search array), this facilitation effect disappeared, showing that it was specific to the relationship between sound and object and not a general benefit of sound. The effect has been further replicated in more naturalistic conditions, with characteristic sounds facilitating search for objects in videos of real-world scenes (Kvasova et al., 2019). Characteristic sounds have further been shown to improve memory for objects in complex scenes (Almadori et al., 2021; Heikkilä et al., 2015). This effect additionally seems to be rooted into the semantic relationship specifically between the auditory and visual signals, rather than a purely conceptual relationship between items that does not rely on the sensory information at all. Rather than a sound leading to prioritization of a picture, it is possible that both are mapped to an abstract concept, just like the word "dog" would be, meaning semantic guidance of attention does not need to be studied with information from a particular modality. If this was the case, it would not be important to study the role of semantic guidance of attention in multisensory contexts because the effects can be fully studied with words. However, when Iordanescu et al. (2008) replaced the object array with a word array (the word dog, rather than a picture of one), the congruent facilitation disappeared suggesting that congruous sounds facilitate processing of the target's visual features (here, the representation

of dog-like visual features) but the processing does not occur on a purely semantic level. In other words, external information provided in one sense (sound of a bark) facilitated the processing of external information in another sense (sight of a dog), but did not facilitate the processing of more abstract representation (a word).

It is still unclear whether crossmodal semantic guidance is an automatic process, with relevant studies showing how semantic guidance interacts with overall attention load. In a paradigm using characteristic sounds, Mastroberardino et al. (2015) showed participants a dog and cat, and played either a bark or meow. The dog and cat then disappeared, and participants had to perform an orientation discrimination task on a Gabor patch where the objects had been. When the Gabor was in the same location as the sound congruent picture, participants better identified the orientation but only when the orientation discrimination task was difficult, suggesting that semantic guidance of audiovisual attention is modulated by attentional load. The results suggest that semantic guidance might not be an automatic process, but instead provides additional prioritization only in contexts where it can be helpful. However, future work will be needed to determine how semantic information modulates prioritization along with each of the other possible factors guiding attention in multisensory contexts.

6 | REMAINING QUESTIONS

Despite several decades of progress in studying single sensory processing mechanisms, many open questions remain about the role of multisensory attention in perception of real-world environments. The factors and mechanisms that shape attentional prioritization within each sensory modality need to be tested in multisensory paradigms with naturalistic, semantically rich stimuli. This research is likely to become the 'new frontier' within attention research.

6.1 | What types of semantic relationships guide attention in multisensory contexts?

Few of the semantic relationships that individuals encounter have been examined and mechanistically understood, particularly for relationships across sensory systems. Even within the same "unit" of attention (e.g., object-to-object) sensory signals can potentially have numerous semantic associations: two sensory signals might belong the same source (e.g., the sight of a dog and barking), might share the same overarching category (e.g., an image of an apple and an orange both show fruit), or might be likely to occur in the same context (e.g., an umbrella and a rain coat are both used in a rainstorm) (Figure 4, left box). In addition, each of these factors do not influence attention in isolation. Carrots may be classed as a piece of produce (e.g., related to apples and oranges), as an item found in kitchens (e.g., related to oven mitts and kettles whistling), and as an ingredient in a stew (e.g., related to a spoon and a pot) (Figure 4, right

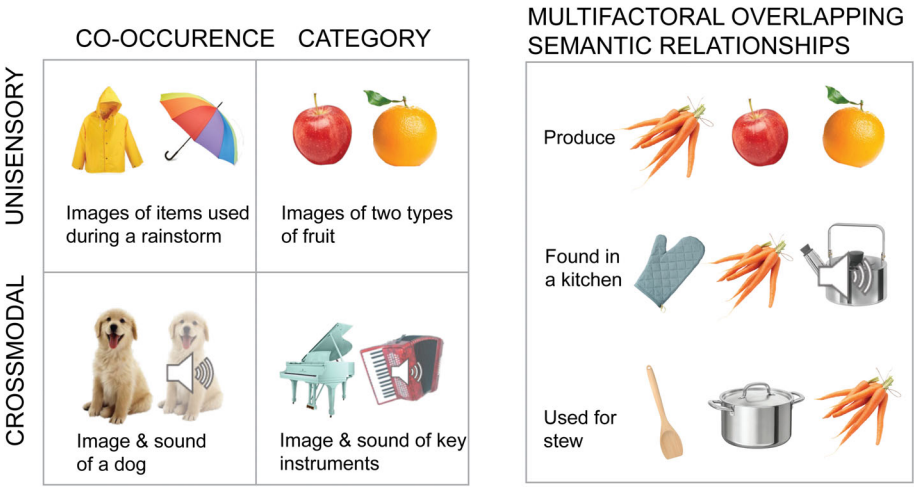

FIGURE 4 An illustration of the possible and overlapping ways that sensory information might be semantically related, both within and between sensory modalities. Any given item may have multiple semantic relationships with various items in a scene, as is shown with the carrots on the right box.

box). Some of these relationships overlap or are hierarchical, like the fact that making a stew usually happens in a kitchen. This complexity exists for each object in a scene: a pot could be associated with carrots as part of making a stew, or alternatively with a drum because of a similar percussive sound, an oven mitt could be associated with other kitchen objects like the carrots or with winter gloves because both protect the hands. To complicate matters further, these individual objects are embedded within a larger scene context that holds additional semantic information. The source of the signal, an object, therefore, can either be semantically related to the scene it is embedded in (e.g., a kettle boiling in a kitchen) or it can be semantically unrelated (e.g., a kettle boiling in a dog park). Any of numerous highly complex, multimodal, and hierarchical relationships guide attention in real world environments. Human judgments of similarity, such as those in the Sight Sound Semantics Database (Wegner-Clemens et al., 2022), can account for some of this complexity by basing it in participants' overall perception of similarity, rather than the experimenter selecting a particular definition of semantic relatedness (e.g., co-occurrence, category membership) and using it as a proxy for semantic relatedness as a whole.

Despite the diversity of semantic associations possible, existing semantic guidance studies using audiovisual contexts (e.g., Iordanescu et al., 2008; Kvasova et al., 2019) have focused on the semantic relationship between two signals that share a source (e.g., an image of a kettle and the sound of a kettle). It is not yet known whether the demonstrated effect of shared-source semantic relationships on attention extends to other types of semantic relationships (category, co-localization, scenes) or is derived from an overarching semantic process. It is possible that shared-source relationships are unique among semantic relationships because these are signals that can, and are likely to, be integrated into a coherent multisensory object. Multisensory integration is highly sensitive to source information and prior expectations (Cao et al., 2019; Gau & Noppeney, 2016; Körding et al., 2007; Noppeney, 2021). The neural mechanisms underlying causal inference could explain the attentional advantage for signals with a shared source semantic relationship. If the underlying mechanism is linked to causal inference, the attentional benefit is not necessarily derived from the semantic relationship and may be unique to the shared-source semantic relationship. Without testing other types of semantic relationships, it is still unknown whether crossmodal guidance of visual attention by sound is truly semantic guidance or guidance through a causal inference mechanism. Characterizing the attention effect with different distinct semantic relationships has the potential to provide a much more robust understanding of the exact underlying mechanisms and processes of crossmodal guidance of attention.

6.2 | How does semantic information interact with other attentional processes?

Semantic information may influence attention differently depending on non-semantic factors within the scene. For example, if a scene strongly drives attention to a particular object-based or low-level feature (e.g., the only red object among gray objects), semantic modulation may have a much smaller effect than it otherwise would in a situation with weaker task demands. As described above, some research has shown that multisensory information is ignored when not relevant to the goal (Mastroberardino et al., 2015), but also, conversely, that task irrelevant information from another modality can guide attention (Iordanescu et al., 2008). These findings appear contradictory, but likely reflect complex interactions between the various attentional processes both within and between sensory modalities. Different attentional processes variably compete or combine to set the overall prioritization map for every bit of information available in the environment. Interaction with other cognitive mechanisms and individual differences (such as variation across the lifespan or clinical groups) could also further modulate how attention prioritization operates in semantically rich multisensory contexts. These processes may also operate differently for different pairs of sensory modalities, such that an audiovisual relationship between two signals, as primarily discussed here, does not have the same influence on attention as a visuotactile relationship. Further work will need to determine how competing attentional processes prioritizing different information in different contexts, across individual differences, and across different pairs of sensory modalities are resolved.

7 | CONCLUSION

Traditionally, experimental psychology has approached the complexity introduced by real-world environments by separating attention from multisensory processing, sensory modalities from each other, and specific features of stimuli from the larger whole. This ground-up approach has been highly informative, producing a foundation of knowledge about

each component cognitive process, sensory system, and a number of factors that can modulate attention. Research has now begun to move to investigate real-world scenes, where multisensory interactions and semantic information is particularly relevant. The time, therefore, is ripe to focus on semantics and multisensory interactions to fill in the gap in knowledge and move toward understanding attention in real-world environments.

AUTHOR CONTRIBUTIONS

Kira Wegner-Clemens: Conceptualization (equal); writing – original draft (equal); writing – review and editing (equal). **George L. Malcolm:** Conceptualization (equal); writing – original draft (equal); writing – review and editing (equal). **Sarah Shomstein:** Conceptualization (equal); writing – original draft (equal); writing – review and editing (equal).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available at https://osf.io/v9rgy/.

DATA AVAILABILITY STATEMENT

This is a review piece and therefore no data are reported.

ORCID

Sarah Shomstein https://orcid.org/0000-0001-8278-6630

RELATED WIREs ARTICLE

Attention and platypuses

FURTHER READING

Castelhano, M. S., Mack, M. L., & Henderson, J. M. (2009). Viewing task influences eye movement control during active scene perception. *Journal of Vision*, *9*(3), 6.1–6.15.

Land, M. F., & Hayhoe, M. (2001). In what ways do eye movements contribute to everyday activities? *Vision Research*, *41*(25–26), 3559–3565.

REFERENCES

Almadori, E., Mastroberardino, S., Botta, F., Brunetti, R., Lupiáñez, J., Spence, C., & Santangelo, V. (2021). Crossmodal semantic congruence interacts with object contextual consistency in complex visual scenes to enhance short-term memory performance. *Brain Sciences*, *11*(9), 1206.

Arcaro, M. J., McMains, S. A., Singer, B. D., & Kastner, S. (2009). Retinotopic organization of human ventral visual cortex. *Journal of Neuroscience*, *29*(34), 10638–10652.

Arrington, C. M., Carr, T. H., Mayer, A. R., & Rao, S. M. (2000). Neural mechanisms of visual attention: Object-based selection of a region in space. *Journal of Cognitive Neuroscience*, *12*(Suppl 2), 106–117.

Ayzenberg, V., & Behrmann, M. (2022). Does the brain's ventral visual pathway compute object shape? *Trends in Cognitive Sciences*, *26*(12), 1119–1132.

Bar, M., & Aminoff, E. (2003). Cortical analysis of visual context. *Neuron*, *38*(2), 347–358.

Behrmann, M., Geng, J. J., & Shomstein, S. (2004). Parietal cortex and attention. *Current Opinion in Neurobiology*, *14*(2), 212–217.

Belke, E., Humphreys, G. W., Watson, D. G., Meyer, A. S., & Telling, A. L. (2008). Top-down effects of semantic knowledge in visual search are modulated by cognitive but not perceptual load. *Perception & Psychophysics*, *70*(8), 1444–1458.

Bertelson, P., Pavani, F., Ladavas, E., Vroomen, J., & de Gelder, B. (2000). Ventriloquism in patients with unilateral visual neglect. *Neuropsychologia*, *38*(12), 1634–1642.

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, *19*(12), 2767–2796.

Bisley, J. W., & Goldberg, M. E. (2010). Attention, intention, and priority in the parietal lobe. *Annual Review of Neuroscience*, *33*, 1–21.

Brady, T. F., & Oliva, A. (2008). Statistical learning using real-world scenes: Extracting categorical regularities without conscious intent. *Psychological Science*, *19*(7), 678–685.

Busse, L., Roberts, K. C., Crist, R. E., Weissman, D. H., & Woldorff, M. G. (2005). The spread of attention across modalities and space in a multisensory object. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(51), 18751–18756.

Buswell, G. T. (1935). How people look at pictures: a study of the psychology and perception in art. 198.

Cao, Y., Summerfield, C., Park, H., Giordano, B. L., & Kayser, C. (2019). Causal inference in the multisensory brain. *Neuron*, *102*(5), 1076–1087.e8.

Carota, F., Kriegeskorte, N., Nili, H., & Pulvermüller, F. (2017). Representational similarity mapping of distributional semantics in left inferior frontal, middle temporal, and motor cortex. *Cerebral Cortex*, *27*(1), 294–309.

Carrasco, M. (2011). Visual attention: The past 25 years. *Vision Research*, *51*(13), 1484–1525.

Castelhano, M. S., & Witherspoon, R. L. (2016). How you use it matters: Object function guides attention during visual search in scenes. *Psychological Science*, *27*(5), 606–621.

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, *3*(3), 201–215.

Cornelissen, T. H. W., & Võ, M. L.-H. (2017). Stuck on semantics: Processing of irrelevant object-scene inconsistencies modulates ongoing gaze behavior. *Attention, Perception & Psychophysics*, *79*(1), 154–168.

Çukur, T., Nishimoto, S., Huth, A. G., & Gallant, J. L. (2013). Attention during natural vision warps semantic representation across the human brain. *Nature Neuroscience*, *16*(6), 763–770.

Demiral, S. B., Malcolm, G. L., & Henderson, J. M. (2012). ERP correlates of spatially incongruent object identification during scene viewing: Contextual expectancy versus simultaneous processing. *Neuropsychologia*, *50*(7), 1271–1285.

Downar, J., Crawley, A. P., Mikulis, D. J., & Davis, K. D. (2000). A multimodal cortical network for the detection of changes in the sensory environment. *Nature Neuroscience*, *3*(3), 277–283.

Driver, J., & Spence, C. (1998). Attention and the crossmodal construction of space. *Trends in Cognitive Sciences*, *2*(7), 254–262.

Duarte, S. E., Ghetti, S., & Geng, J. J. (2022). Object memory is multisensory: Task-irrelevant sounds improve recollection. *Psychonomic Bulletin & Review*, *30*, 652–665.

Duncan, J. (1984). Selective attention and the organization of visual information. *Journal of Experimental Psychology. General*, *113*(4), 501–517.

Eckstein, M. P., Drescher, B. A., & Shimozaki, S. S. (2006). Attentional cues in real scenes, saccadic targeting, and Bayesian priors. *Psychological Science*, *17*(11), 973–980.

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.

Engel, S. A., Glover, G. H., & Wandell, B. A. (1997). Retinotopic organization in human visual cortex and the spatial precision of functional MRI. *Cerebral Cortex*, *7*(2), 181–192.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, *16*(1), 143–149.

Fiez, J. A. (1997). Phonology, semantics, and the role of the left inferior prefrontal cortex. *Human Brain Mapping*, *5*(2), 79–83.

Foulsham, T., & Underwood, G. (2007). How does the purpose of inspection influence the potency of visual salience in scene perception? *Perception*, *36*(8), 1123–1138.

Gabrieli, J. D. E., Desmond, J. E., Demb, J. B., Wagner, A. D., Stone, M. V., Vaidya, C. J., & Glover, G. H. (1996). Functional magnetic resonance imaging of semantic memory processes in the frontal lobes. *Psychological Science*, *7*(5), 278–283.

Gau, R., & Noppeney, U. (2016). How prior expectations shape multisensory perception. *NeuroImage*, *124*(Pt A), 876–886.

Geng, J. J., Leber, A. B., & Shomstein, S. (2019). Attention and perception: 40 reviews, 40 views. *Current Opinion Psychology*, *29,* v–viii.

Golomb, J. D., Chun, M. M., & Mazer, J. A. (2008). The native coordinate system of spatial attention is retinotopic. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *28*(42), 10654–10662.

Greene, M. R., & Fei-Fei, L. (2014). Visual categorization is automatic and obligatory: Evidence from stroop-like paradigm. *Journal of Vision*, *14*(1), 1–11.

Grefkes, C., Weiss, P. H., Zilles, K., & Fink, G. R. (2002). Crossmodal processing of object features in human anterior intraparietal cortex: An fMRI study implies equivalencies between humans and monkeys. *Neuron*, *35*(1), 173–184.

Heikkilä, J., Alho, K., Hyvönen, H., & Tiippana, K. (2015). Audiovisual semantic congruency during encoding enhances memory performance. *Experimental Psychology*, *62*(2), 123–130.

Henderson, J. M., Malcolm, G. L., & Schandl, C. (2009). Searching in the dark: Cognitive relevance drives attention in real-world scenes. *Psychonomic Bulletin & Review*, *16*(5), 850–856.

Hershenson, M. (1962). Reaction time as a measure of intersensory facilitation. *Journal of Experimental Psychology*, *63*(3), 289–293.

Hillyard, S. A., Störmer, V. S., Feng, W., Martinez, A., & McDonald, J. J. (2016). Cross-modal orienting of visual attention. *Neuropsychologia*, *83*, 170–178.

Hollingworth, A., & Henderson, J. M. (2002). Accurate visual memory for previously attended objects in natural scenes. *Journal of Experimental Psychology. Human Perception and Performance*, *28*(1), 113–136.

Husain, M., & Nachev, P. (2007). Space and the parietal cortex. *Trends in Cognitive Sciences*, *11*(1), 30–36.

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

Iordanescu, L., Grabowecky, M., Franconeri, S., Theeuwes, J., & Suzuki, S. (2010). Characteristic sounds make you look at target objects more quickly. *Attention, Perception & Psychophysics*, *72*(7), 1736–1741.

Iordanescu, L., Guzman-Martinez, E., Grabowecky, M., & Suzuki, S. (2008). Characteristic sounds facilitate visual search. *Psychonomic Bulletin & Review*, *15*(3), 548–554.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, *40*(10–12), 1489–1506.

Jensen, A., Merz, S., Spence, C., & Frings, C. (2020). Perception it is: Processing level in multisensory selection. *Attention, Perception & Psychophysics*, *82*(3), 1391–1406.

Jigo, M., Gong, M., & Liu, T. (2018). Neural determinants of task performance during feature-based attention in human cortex. *ENeuro*, *5*(1), 1–14.

Koelewijn, T., Bronkhorst, A., & Theeuwes, J. (2010). Attention and the multiple stages of multisensory integration: A review of audiovisual studies. *Acta Psychologica*, *134*(3), 372–384.

Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., & Shams, L. (2007). Causal inference in multisensory perception. *PLoS One*, *2*(9), e943.

Kravitz, D. J., Saleem, K. S., Baker, C. I., & Mishkin, M. (2011). A new neural framework for visuospatial processing. *Nature Reviews. Neuroscience*, *12*(4), 217–230.

Kvasova, D., Garcia-Vernet, L., & Soto-Faraco, S. (2019). Characteristic sounds facilitate object search in real-life scenes. *Frontiers in Psychology*, *10*, 2511.

Lee, J., & Shomstein, S. (2013). The differential effects of reward on space- and object-based attentional allocation. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *33*(26), 10625–10633.

Liu, T. (2019). Feature-based attention: Effects and control. *Current Opinion in Psychology*, *29*, 187–192.

Liu, T., Hospadaruk, L., Zhu, D. C., & Gardner, J. L. (2011). Feature-specific attentional priority signals in human cortex. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *31*(12), 4484–4495.

Liu, T., & Hou, Y. (2013). A hierarchy of attentional priority signals in human frontoparietal cortex. *The Journal of Neuroscience*, *33*(42), 16606–16616.

Livne, T., & Bar, M. (2016). Cortical integration of contextual information across objects. *Journal of Cognitive Neuroscience*, *28*(7), 948–958.

Lupyan, G., & Spivey, M. J. (2010). Redundant spoken labels facilitate perception of multiple items. *Attention, Perception & Psychophysics*, *72*(8), 2236–2253.

Macaluso, E., Driver, J., & Frith, C. D. (2003). Multimodal spatial representations engaged in human parietal cortex during both saccadic and manual spatial orienting. *Current Biology*, *13*(12), 990–999.

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. *Journal of Vision*, *11*(9), 1–16.

Malcolm, G. L., Groen, I. I. A., & Baker, C. I. (2016). Making sense of real-world scenes. *Trends in Cognitive Sciences*, *20*(11), 843–856.

Malcolm, G. L., & Henderson, J. M. (2010). Combining top-down processes to guide eye movements during real-world scene search. *Journal of Vision*, *10*(2), 4.1–4.11.

Malcolm, G. L., Rattinger, M., & Shomstein, S. (2016). Intrusive effects of semantic information on visual selective attention. *Attention, Perception & Psychophysics*, *78*(7), 2066–2078.

Mastroberardino, S., Santangelo, V., & Macaluso, E. (2015). Crossmodal semantic congruence can affect visuo-spatial processing and activity of the fronto-parietal attention networks. *Frontiers in Integrative Neuroscience*, *9*, 45.

Matusz, P. J., & Eimer, M. (2011). Multisensory enhancement of attentional capture in visual search. *Psychonomic Bulletin & Review*, *18*(5), 904–909.

Maunsell, J. H. R., & Treue, S. (2006). Feature-based attention in visual cortex. *Trends in Neurosciences*, *29*(6), 317–322.

McDonald, J. J., Störmer, V. S., Martinez, A., Feng, W., & Hillyard, S. A. (2013). Salient sounds activate human visual cortex automatically. *The Journal of Neuroscience*, *33*(21), 9194–9201.

McDonald, J. J., Teder-Sälejärvi, W. A., Di Russo, F., & Hillyard, S. A. (2003). Neural substrates of perceptual enhancement by cross-modal spatial attention. *Journal of Cognitive Neuroscience*, *15*(1), 10–19.

McDonald, J. J., Teder-Sälejärvi, W. A., & Hillyard, S. A. (2000). Involuntary orienting to sound improves visual perception. *Nature*, *407*(6806), 906–908.

Mesulam, M. M. (1999). Spatial attention and neglect: Parietal, frontal and cingulate contributions to the mental representation and attentional targeting of salient extrapersonal events. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *354*(1387), 1325–1346.

Miller, J. (1982). Divided attention: Evidence for coactivation with redundant signals. *Cognitive Psychology*, *14*(2), 247–279.

Molenberghs, P., Mesulam, M. M., Peeters, R., & Vandenberghe, R. R. C. (2007). Remapping attentional priorities: Differential contribution of superior parietal lobule and intraparietal sulcus. *Cerebral Cortex*, *17*(11), 2703–2712.

Molholm, S., Martinez, A., Shpaner, M., & Foxe, J. J. (2007). Object-based attention is multisensory: co-activation of an object's representations in ignored sensory modalities. *The European Journal of Neuroscience*, *26*(2), 499–509.

Moores, E., Laiti, L., & Chelazzi, L. (2003). Associative knowledge controls deployment of visual selective attention. *Nature Neuroscience*, *6*(2), 182–189.

Müller, N. G., & Kleinschmidt, A. (2003). Dynamic interaction of object- and space-based attention in retinotopic visual areas. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *23*(30), 9812–9816.

Nah, J. C., Malcolm, G. L., & Shomstein, S. (2019). Semantic associations between scenes and objects bias attention even when task-irrelevant. *Journal of Vision*, *19*(10), 46a.

Nah, J. C., Malcolm, G. L., & Shomstein, S. (2021). Task-irrelevant semantic properties of objects impinge on sensory representations within the early visual cortex. *Cerebral Cortex Communications*, *2*(3), tgab049.

Nakayama, K., & Martini, P. (2011). Situating visual search. *Vision Research*, *51*(13), 1526–1537.

Neider, M. B., & Zelinsky, G. J. (2006). Scene context guides eye movements during visual search. *Vision Research*, *46*(5), 614–621.

Neisser, U. (1963). Decision-time without reaction-time: Experiments in visual scanning. *The American Journal of Psychology*, *76*(3), 376–385.

Noppeney, U. (2021). Perceptual inference, learning, and attention in a multisensory world. *Annual Review of Neuroscience*, *44*, 449–473.

O'Craven, K. M., Downing, P. E., & Kanwisher, N. (1999). fMRI evidence for objects as the units of attentional selection. *Nature*, *401*(6753), 584–587.

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, *42*(1), 107–123.

Peacock, C. E., Hayes, T. R., & Henderson, J. M. (2019). Meaning guides attention during scene viewing, even when it is irrelevant. *Attention, Perception & Psychophysics*, *81*(1), 20–34.

Posner, M. I. (1980). Orienting of attention. *The Quarterly Journal of Experimental Psychology*, *32*(1), 3–25.

Rock, I., & Gutman, D. (1981). The effect of inattention on form perception. *Journal of Experimental Psychology Human Perception and Performance*, *7*(2), 275–285.

Rohe, T., & Noppeney, U. (2015). Cortical hierarchies perform Bayesian causal inference in multisensory perception. *PLoS Biology*, *13*(2), e1002073.

Santangelo, V., & Spence, C. (2007). Multisensory cues capture spatial attention regardless of perceptual load. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(6), 1311–1321.

Schut, M. J., van der Stoep, N., & van der Stigchel, S. (2018). Auditory spatial attention is encoded in a retinotopic reference frame across eye-movements. *PLoS One*, *13*(8), e0202414.

Senkowski, D., Schneider, T. R., Foxe, J. J., & Engel, A. K. (2008). Crossmodal binding through neural coherence: Implications for multisensory processing. *Trends in Neurosciences*, *31*(8), 401–409.

Serences, J. T., Schwarzbach, J., Courtney, S. M., Golay, X., & Yantis, S. (2004). Control of object-based attention in human cortex. *Cerebral Cortex*, *14*(12), 1346–1357.

Serences, J. T., Shomstein, S., Leber, A. B., Golay, X., Egeth, H. E., & Yantis, S. (2005). Coordination of voluntary and stimulus-driven attentional control in human cortex. *Psychological Science*, *16*(2), 114–122.

Serences, J. T., & Yantis, S. (2006). Spatially selective representations of voluntary and stimulus-driven attentional priority in human occipital, parietal, and frontal cortex. *Cerebral Cortex*, *17*(2), 284–293.

Sereno, M. I., & Huang, R.-S. (2014). Multisensory maps in parietal cortex. *Current Opinion in Neurobiology*, *24*(1), 39–46.

Sheremata, S. L., & Silver, M. A. (2015). Hemisphere-dependent attentional modulation of human parietal visual field representations. *The Journal of Neuroscience*, *35*(2), 508–517.

Shomstein, S. (2012). Object-based attention: Strategy versus automaticity. *Wiley Interdisciplinary Reviews Cognitive Science*, *3*(2), 163–169.

Shomstein, S., & Behrmann, M. (2006). Cortical systems mediating visual attention to both objects and spatial locations. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(30), 11387–11392.

Shomstein, S., Lee, J., & Behrmann, M. (2010). Top-down and bottom-up attentional guidance: Investigating the role of the dorsal and ventral parietal cortices. *Experimental Brain Research*, *206*(2), 197–208.

Shomstein, S., Malcolm, G. L., & Nah, J. C. (2019). Intrusive effects of task-irrelevant information on visual selective attention: Semantics and size. *Current Opinion in Psychology*, *29*, 153–159.

Shomstein, S., Zhang, X., & Dubbelde, D. (2022). Attention and platypuses. *Wiley Interdisciplinary Reviews Cognitive Science*, *14*(1), 1–10.

Silver, M. A., & Kastner, S. (2009). Topographic maps in human frontal and parietal cortex. *Trends in Cognitive Sciences*, *13*(11), 488–495.

Spence, C. (2020). In C. W. Eriksen (Ed.), Special Issue *Extending the study of visual attention to a multisensory world*. Attention, Perception & Psychophysics.

Szinte, M., Aagten-Murphy, D., Jonikaitis, D., Wollenberg, L., & Deubel, H. (2020). Sounds are remapped across saccades. *Scientific Reports*, *10*(1), 21332.

Talsma, D. (2015). Predictive coding and multisensory integration: An attentional account of the multisensory mind. *Frontiers in Integrative Neuroscience*, *9*, 19.

Talsma, D., Doty, T. J., & Woldorff, M. G. (2007). Selective attention and audiovisual integration: Is attending to both modalities a prerequisite for early integration? *Cerebral Cortex*, *17*(3), 679–690.

Talsma, D., Senkowski, D., Soto-Faraco, S., & Woldorff, M. G. (2010). The multifaceted interplay between attention and multisensory integration. *Trends in Cognitive Sciences*, *14*(9), 400–410.

Thelen, A., Matusz, P. J., & Murray, M. M. (2014). Multisensory context portends object memory. *Current Biology*, *24*(16), R734–R735.

Thompson-Schill, S. L., D'Esposito, M., Aguirre, G. K., & Farah, M. J. (1997). Role of left inferior prefrontal cortex in retrieval of semantic knowledge: A reevaluation. *Proceedings of the National Academy of Sciences of the United States of America*, *94*(26), 14792–14797.

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. *Psychological Review*, *113*(4), 766–786.

Treisman, A. (1985). Preattentive processing in vision. *Computer Vision, Graphics, and Image Processing*, *31*(2), 156–177.

Treisman, A., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136.

Uyar, F., Shomstein, S., Greenberg, A. S., & Behrmann, M. (2016). Retinotopic information interacts with category selectivity in human ventral cortex. *Neuropsychologia*, *92*, 90–106.

Valdes-Sosa, M., Bobes, M. A., Rodriguez, V., & Pinilla, T. (1998). Switching attention without shifting the spotlight object-based attentional modulation of brain potentials. *Journal of Cognitive Neuroscience*, *10*(1), 137–151.

van Atteveldt, N., Murray, M. M., Thut, G., & Schroeder, C. E. (2014). Multisensory integration: Flexible use of general operations. *Neuron*, *81*(6), 1240–1253.

van der Burg, E., Olivers, C. N. L., Bronkhorst, A. W., & Theeuwes, J. (2008). Pip and pop: Nonspatial auditory signals improve spatial visual search. *Journal of Experimental Psychology. Human Perception and Performance*, *34*(5), 1053–1065.

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.

Wagner, A. D., Paré-Blagoev, E. J., Clark, J., & Poldrack, R. A. (2001). Recovering meaning: Left prefrontal cortex guides controlled semantic retrieval. *Neuron*, *31*(2), 329–338.

Wagner, R. K., Torgesen, J. K., Rashotte, C. A., Hecht, S. A., Barker, T. A., Burgess, S. R., Donahue, J., & Garon, T. (1997). Changing relations between phonological processing abilities and word-level reading as children develop from beginning to skilled readers: A 5-year longitudinal study. *Developmental Psychology*, *33*(3), 468–479.

Ward, L. M., McDonald, J. J., & Lin, D. (2000). On asymmetries in cross-modal spatial attention orienting. *Perception & Psychophysics*, *62*(6), 1258–1264.

Wegner-Clemens, K., Malcolm, G. L., & Shomstein, S. (2022). How much is a cow like a meow? A novel database of human judgements of audiovisual semantic relatedness. *Attention, Perception & Psychophysics*, *84*(4), 1317–1327.

Wolfe, J. M., & Horowitz, T. S. (2004). What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience*, *5*(6), 495–501.

Wu, C.-C., Wang, H.-C., & Pomplun, M. (2014). The roles of scene gist and spatial dependency among objects in the semantic guidance of attention in real-world scenes. *Vision Research*, *105*, 10–20.

Yarbus, A. L. (1967). In B. Haigh (Ed.), Trans. *Eye movements and vision*. Plenum Press.

How to cite this article: Wegner-Clemens, K., Malcolm, G. L., & Shomstein, S. (2024). Predicting attentional allocation in real-world environments: The need to investigate crossmodal semantic guidance. *WIREs Cognitive Science*, e1675. https://doi.org/10.1002/wcs.1675