



The role of event understanding in guiding attentional selection in real-world scenes: The Scene Perception & Event Comprehension Theory (SPECT)

Lester C. Loschky¹ · Maverick E. Smith² · Prasanth Chandran¹ · John P. Hutson³ · Tim J. Smith⁴ · Joseph P. Magliano⁵

Received: 18 October 2024 / Accepted: 10 January 2026
© The Author(s) 2026

Abstract

Your understanding of what you see now surely influences what you will look at next. Yet this simple concept has only recently begun to be systematically studied and elaborated within theoretical frameworks. The Scene Perception & Event Comprehension Theory (SPECT) distinguishes between front-end and back-end processes that occur while viewers perceive and comprehend dynamic real-world events. Front-end processes occur during each eye fixation (information extraction, attentional selection) and back-end processes occur in memory (the current event model, the stored event model, prior knowledge, and executive processes). We begin with a selective review of the scene perception literature on bottom-up and top-down effects on attentional selection in scenes, and highlight unanswered questions regarding the impact of the viewer's event model—their understanding of what is happening now. Then, we outline the SPECT theoretical framework, and review empirical evidence about how the viewer's current event model influences attentional selection. This influence is contrasted with those of visual saliency (e.g., color, brightness, motion) and task-driven control (i.e., goal setting, attentional control, inhibition). From this review, we specify a hierarchy of factors affecting attentional selection, in the order of task-driven control, visual saliency, and event models. We then propose several mechanisms by which the viewer's event model influences attentional selection, and propose a systematic approach to investigating how that happens while watching dynamic scenes.

Keywords Scene perception · Event perception · Narrative comprehension · Film · Attentional selection · Eye movements

✉ Lester C. Loschky
loschky@ksu.edu
Maverick E. Smith
msmith@truman.edu
Prasanth Chandran
prasanthp@ksu.edu
John P. Hutson
jhutson@ets.org
Tim J. Smith
tim.smith@arts.ac.uk
Joseph P. Magliano
jmagliano@gsu.edu

¹ Department of Psychological Sciences, Kansas State University, Manhattan, KS, USA

² Present Address: Psychology and Counseling Department, Truman State University, Kirksville, MO, USA

³ Educational Testing Service, Princeton, NJ, USA

⁴ Creative Computing Institute, University of the Arts London, London, UK

⁵ Department of Learning Sciences, Georgia State University, Atlanta, GA, USA

Comprehension and attention in events

As you look around at the world, your understanding of what you see now influences what you will look at next. What you pay attention to greatly impacts your consciousness (Mack & Rock, 1998; Simons & Chabris, 1999), understanding (Just & Carpenter, 1987; Rouinfar et al., 2014), and memory (Tatler et al., 2005; C. C. Williams et al., 2005) for what you see. This raises the question, why do we attend to some things, but ignore others? We argue that our understanding of events influences what we attend to. Events are considered an important unit of cognition (Radvansky & Zacks, 2014). We intuitively view our experiences while engaging with the world as *events*. Events are typically defined as segments in time, at a location, perceived by a viewer to have a beginning and ending, with state changes in between (Zacks & Tversky, 2001). Importantly, events have components, such as agents, their behaviors, and cause and effect, within a spatiotemporal framework. Events are important to psychologists who study event cognition (Radvansky & Zacks, 2014),

psycholinguistics (Altmann & Ekves, 2019), and narrative comprehension (Gernsbacher, 1990; Graesser et al., 1994; Trabasso et al., 1989). Psychologists who study how attention affects scene perception *are studying event cognition*, even if they may not construe that they are doing so; the colloquial definition of a “scene” is where something *happens*. As such, we contend that the study of attention and scene perception should explicitly embrace the construct of events, be informed of how they are represented, and study how and when those representations affect attention as events unfold dynamically. However, scene perception researchers have typically studied single events, at single times, using single static scenes (for reviews, see Epstein & Baker, 2019; Henderson & Hollingworth, 1999).

To illustrate the importance of events for understanding attentional selection in scene perception, consider Fig. 1, which shows 17 frames from the opening scene of Orson Welles’ (Welles & Zugsmith, 1958) film *Touch of Evil*. In the first two events (shown with red borders), we see someone put a time bomb in a car, and soon after, the owners of that car unwittingly get in and drive off. In the third event (again with red borders), we see the car being stopped by a traffic cop on a busy street. The fourth event introduces another couple walking down the street. The walking couple then become the focus of the scene, while the bomb-laden car comes into and out of view over time. This creates heightened suspense in the viewer. Importantly, this suspense depends on the viewer’s understanding of the events



Fig. 1 Frames illustrating taken from the film *Touch of Evil* (Welles & Zugsmith, 1958). Red borders around select frames illustrate normative event boundaries. These moments were judged as event

boundaries by more than 23% of the participants. Gray borders indicate non-boundaries (judged as event boundaries by less than 4% of participants). (Color figure online)

they are seeing—their *event model*—because the bomb (a known entity) has the causal power to kill both couples (other known entities) at any moment. Does the viewer's event model prioritize looking at the car whenever it enters the screen, even when the car is in competition with the walking couple for the viewer's attention? Our answer to this question is a qualified yes, but as we will see, there are important constraints on how viewers' event models influence their attention. Furthermore, individual frames in this scene, which would be treated as separate pictures in a scene perception study, have importance because of their relations to the viewer's event model for the entire scene. Thus, we will use the *viewer's event model* to explain how this scene is perceived and *how attention is deployed* as the scene unfolds. Importantly, under the right conditions, we get a significant and meaningful impact of viewers' event models on their selective attention, as measured by where they send their eyes, including their first saccade on a scene (Hutson et al., 2022).

It is well-established that 1) what we selectively attend to greatly affects how we understand what we see; however, our example in the *Touch of Evil* scene suggests that 2) how we understand what we see affects what we selectively attend to. Concerning Point 1, there is ample evidence that attentional selection is essential to comprehension (for reviews, see Cohn & Foulsham, 2020; Foulsham et al., 2016; Just & Carpenter, 1987; Rayner, 1998; Rensink et al., 1997; Rouinfar et al., 2014). However, far fewer studies have investigated Point 2—how comprehension affects what we selectively attend to (Hutson et al., 2018, 2022; Loschky et al., 2015; Tanenhaus et al., 1995).

To capture the dynamic nature of real-world scenes, psychologists must study events that unfold over time. Much of our research examines the relationship between event comprehension and attentional selection by using narrative film stimuli (Hutson et al., 2017, 2022; Loschky et al., 2015). Visual narratives are an excellent type of stimuli to use for this purpose, because viewers process unscripted videos of everyday events and visual narratives very similarly (Magliano et al., 2020; Zacks et al., 2009). Put differently, movies work the way they do because they build upon the mechanisms that humans have developed for perceiving and understanding real-world scenes (Cutting, 2021; Zacks, 2015). Visual narratives, in virtual reality, film, slide shows, comics, or picture stories, capture the temporal evolution of events. People understand events using working memory and episodic memory representations called *event models* (e.g., Radvansky & Zacks, 2011). An event model in working memory reflects one's understanding of the events as they experience them in real time. We argue that psychologists need to explicitly connect *attention*, in the context of dynamically unfolding scenes, to *event comprehension*, as reflected in the *event model*. To this end, we have proposed and begun testing the Scene Perception &

Event Comprehension Theory (SPECT), which provides a theoretical framework for explaining how attentional selection and event comprehension are coordinated, as we experience events over time, in both unscripted dynamic real-world scenes and visual narratives (Loschky et al., 2018, 2020).

While Mary Peterson's research program has not explicitly explored the top-down effects of event models on attention, our interest in developing a theoretical account of how that happens is in the spirit of her work. A key contribution of Mary Peterson's work to the study of perception has been by looking for top-down effects on perception in an established research area where nobody thought to look for such effects. For example, even before her groundbreaking figure-ground work, Hochberg and Peterson (1987) showed that object recognition could unfold piecemeal over time, with partial perceptual information activating long-term memory representations that guide ongoing perceptual organization. This provided evidence against a strictly sequential model (first perception, then recognition) and instead supported a more interactive, real-time view of perception. Later, Peterson extended this line of inquiry to show that matching a proto-object to a long-term memory representation influenced the strongly assumed earlier process of figure-ground segmentation (e.g., M. A. Peterson, 1994; M. A. Peterson & Gibson, 1994). That was a theory-driven question, which called into question key assumptions about top-down effects on object processing. Our research on the role of event models on attention (e.g., Hutson et al., 2018, 2022) and perception (M. E. Smith & Loschky, 2019) is similarly one that few have explored. In this case, it is not that the influence of online event understanding on attentional selection seems counter-intuitive—on the contrary, it appears quite intuitive to most people. Instead, it seems more likely that attention researchers consider the top-down theoretical construct of the viewer's "event model" (i.e., their understanding of what they are seeing now) as too complex to test its influence on attentional selection. Research on visual attentional selection has long been largely limited to studies of visual search, often for very simple targets (e.g., conjunctions of red and green vertical and horizontal bars; or digits and letters created by removing segments of a figure-eight premask; e.g., Theeuwes et al., 1999; Wolfe, 2007). Only recently have attention researchers begun to study viewers' eye movements while they watch real-world video clips, usually when studying the impact of visual saliency (as measured by a computational algorithm), or a task (e.g., search for a specific object in a scene; e.g., Henderson et al., 2007; T. J. Smith & Mital, 2013). Thus, our theoretically driven studies of how one's event model influences attentional selection, may seem to "come out of left field" analogously to Mary Peterson's studies of how object recognition influences figure-ground segmentation.

The goal of this paper is to discuss how viewers' understanding of events, unfolding over time, influences their

attentional selection. We focus on eye movements, rather than other attentional measures, such as self-paced dwell time in slideshows (Hard et al., 2011), because they provide a continuous and ecologically valid measure of attentional selection, across both time and space, as events dynamically unfold. We will first discuss what has been learned about attentional selection and event comprehension, either directly or incidentally, from prior research on scene perception. We will highlight gaps in this literature that necessitate the more targeted inclusion of *event models*. We then review SPECT (Loschky et al., 2018, 2020). In this review we add a well-established mechanism to SPECT, biased competition, to explain *a newly hypothesized mechanism for influencing attentional selection, the event model*. We then discuss recent research which is consistent with the event model hypothesis. Finally, we will discuss conclusions we can draw from our review of the research, unanswered questions raised by SPECT, and future directions for studying the effects of the event model on attentional selection in real-world scenes.

What influences attention in real-world scene perception, and what important unanswered questions involve event models?

Prior work in scene perception research has investigated both bottom-up and top-down influences on attentional selection. Regarding top-down effects, this work has been limited to the effects of *tasks* and *prior knowledge*. However, to understand how viewers comprehend what they are seeing, from moment to moment, we need to invoke the theoretical construct of *event models*. The SPECT theoretical framework was created, in part, to explain how event models influence attentional selection. Below is a selective and condensed overview of studies of attentional selection in real-world scenes. In reviewing these studies, we make the case for *why events* are critical for understanding attentional selection in scene perception.

Effects of bottom-up visual saliency on attentional selection in scenes

Bottom-up visual stimulus features and top-down factors influence attentional selection. Features that predict where people look are termed *visually salient*. Computational models of visual saliency have been developed and studied for roughly 30 years (Bruce & Tsotsos, 2009; Itti & Koch, 2000; Kummerer et al., 2017; Linardos et al., 2021; Tsotsos et al., 1995). Such models generate a 2-D saliency map, indicating the predicted probability of a viewer's attentional selection in an image. Early models used low-level bottom-up visual feature contrast

(e.g., luminance, color, orientation, and motion) to predict saliency (e.g., something light among dark things, something red among green things, something vertical among horizontal things; Itti & Koch, 2000). The most visually salient features are things that suddenly appear (Theeuwes et al., 1999), or are moving (Carmi & Itti, 2006; Mital et al., 2010). More recent computational saliency models have included higher-level semantic object classes known to strongly attract human gaze, such as human faces (Fletcher-Watson et al., 2008), or text (Cerf et al., 2009), to predict viewers' attention (Kummerer et al., 2017; Linardos et al., 2021). Figure 2 shows an example of the predicted saliency of regions in a screenshot from the film, *Touch of Evil*. Computational models perform above chance at predicting attentional selection; however, even the best models cannot explain significant proportions of variance in where people look, because top-down factors also influence attentional selection, especially in static scenes (Foulsham & Underwood, 2007; Henderson et al., 2007; Pedziwiatr et al., 2021; Tatler et al., 2011). As shown in the top of Fig. 2, a saliency map based on high-level features (Linardos et al., 2021) predicts that the walking couple are the most salient region of that film frame, whereas the trunk of the car carrying the bomb is not. The bottom of Fig. 2 shows a heat map based on eye fixations, from viewers who knew of the bomb in the car trunk, and maintained that knowledge in working

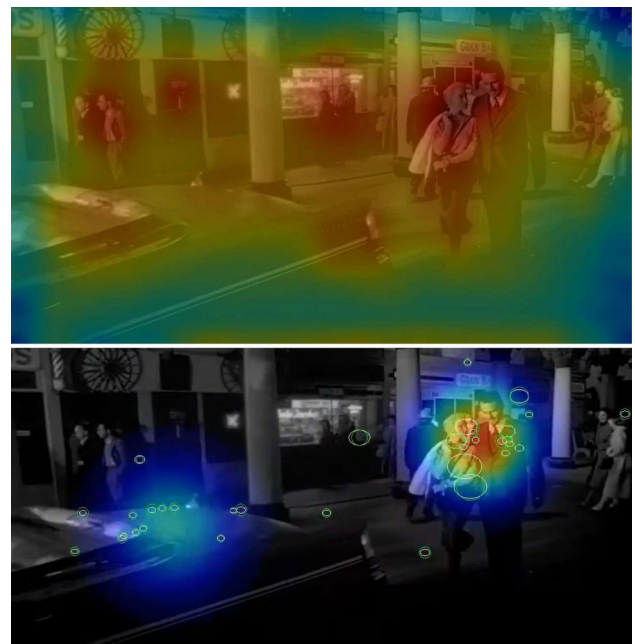


Fig. 2 Top: Saliency map (Linardos et al., 2021) of a frame from the opening scene of Orson Welles' *Touch of Evil* (Welles & Zugsmith, 1958). Bottom: Fixations and heat map on that frame from *Touch of Evil*. Note the low predicted saliency of the car trunk, compared with observed fixations on the car trunk, which these viewers know contains a bomb (data from Hutson et al., 2017). (Color figure online)

memory throughout the 3-min scene. Although viewers' eye fixations were consistent with the high-level saliency model's predicted *hot spot* on the walking couple, the saliency model did not predict the secondary hot spot on the car trunk. Conversely, the hot spot on the car trunk was both predicted, and shown, to be important for those viewers whose event models included the concealed bomb (Hutson et al., 2017, Exp. 2 A).

A key unanswered question is how bottom-up visual saliency and a top-down event model biases compete for attentional selection in ongoing scenes, such as visual narratives, and dynamic real-world scenes. The salience, the strength of bias from an event model, and the balance between the two are likely to change as a scene unfolds over time.

Effects of task-driven top-down attentional selection in scenes

We distinguish between two types of top-down influences on attentional selection, which we will call "*task-driven*" versus "*knowledge-based*" (see also Baluch & Itti, 2011). Task-driven top-down influences on attentional selection operate in everyday tasks like searching for your car in a parking lot. Computational models of visual search have become increasingly accurate at predicting where people will look while searching for targets (e.g., a red crossover SUV in a parking lot full of cars of various colors, types, and models; Wolfe, 2021; Zelinsky et al., 2020). Such visual search models use a 2-D priority map (like a saliency map, but including goals) to indicate the predicted probability that a viewer will fixate each image location. Importantly, a viewer's task is a much better predictor of what they will attend to in a scene than visual saliency (Einhäuser et al., 2008; Foulsham & Underwood, 2007; Henderson et al., 2007; Tatler et al., 2011) even in real-world scene videos (T. J. Smith & Mital, 2013), despite the fact that *motion* creates the strongest form of visual saliency (Mital et al., 2010).

A key unanswered question is the balance between *task-driven*, *knowledge-driven*, and specifically *event model-driven biases* in the competition for attentional selection. In contrast to task-driven situations, many everyday experiences, such as watching a movie or reading a graphic novel, unfold without an explicit external goal. In these cases, event models may guide attentional selection by default, organizing perception and memory in service of the internal goal of enjoyment (Vorderer et al., 2006). Such enjoyment is a rewarding state that emerges from the dynamic interaction between viewer motives and media features (e.g., narrative coherence, pacing)

Effects of knowledge-driven top-down attentional selection in scenes

Gist and schemas guide visual search in scenes Studies have investigated knowledge-driven top-down influences

on viewers' attentional selection in real-world scenes based on the viewer's relevant *schemas* (i.e., structured, generic, semantic long-term memory representations). Viewers can rapidly extract a holistic semantic representation of a scene within their first fixation (e.g., ≤ 330 ms); commonly known as its *scene gist* (Greene & Oliva, 2009; Larson et al., 2014; Loschky et al., 2007; Oliva & Torralba, 2006). The scene's gist activates its superordinate and basic-level category, the schema of the scene in semantic memory, and relevant knowledge about what objects tend to be where (Võ et al., 2019). It can also include the basic level category of one or more key animate entities in the scene (e.g., a man, a dog; Fei-Fei et al., 2007), the identity of an agent (i.e., the "doer" of an action; Hafri et al., 2013), and the scene's emotional valence (e.g., neutral, scary, happy; Calvo & Lang, 2005).

A key finding is that because a scene's gist rapidly activates a viewer's schemas in semantic long-term memory, it influences attentional selection of their first eye movement (saccade) on a scene. For example, Eckstein, et al. (2006) found that, when searching for a chimney, viewers' first saccade went to scene locations where chimneys were expected to be (i.e., the roof of a house). This was true, even when no chimney was there (Eckstein et al., 2006; Torralba et al., 2006). This is evidence that a schema (e.g., knowledge that houses have chimneys on their roofs) influences attentional selection (i.e., where to send a saccade). Note that this example combines task-driven top-down attentional selection (due to carrying out an explicit search task) and knowledge-based attentional selection (due to activating a relevant schema influencing saccade target selection). Note also that the speed of influencing attentional selection—namely, the first saccade on the image, was likely due to having time to process the search target identity (i.e., the word "chimney"), and activate its scene schema, before seeing the scene image.

More recently, Henderson and colleagues have extended this work by examining how the distribution of local scene meaning affects attentional selection (Hayes & Henderson, 2022b; Henderson & Hayes, 2018; Peacock et al., 2019). They obtained a 2D spatial distribution of local scene meaning comparable with a bottom-up visual salience map. For that, they generated meaning maps by crowdsourcing participants' ratings of the meaningfulness of many isolated patches taken from scenes. Henderson and colleagues have consistently demonstrated that meaning maps account for unique variance in the distribution of attention, even after statistically controlling for the correlation between meaning and bottom-up saliency. Such knowledge-based attentional guidance occurs because the scene's gist accesses stored knowledge about the scene category and a coarse representation of where informative regions in a scene are likely to be. However, the precise information that meaning maps capture

remains an open debate (Hayes & Henderson, 2022a; Pedziwiatr et al., 2021).

A key unanswered question is how the scene gist, acquired on the first fixation of a scene, is used to *lay the foundation* for the *current event model*, an *integrated dynamic meaning representation*, and how that, in turn, influences attentional selection. Gist activates schemas from semantic long term memory which contribute to the starting point (i.e., foundation) for the current event model, as described in detail later.

Effects of schema-inconsistent objects on attention in scenes What happens to viewers' attentional selection in a scene if an object violates the viewer's schema for that scene? For example, imagine a bathroom scene in which one sees a tube of toothpaste versus a small flashlight on a bathroom sink (Coco et al., 2020). Will the difference in the viewer's expectations for each object in the bathroom scene produce differences in attentional selection? Research has gone back and forth on this question, but it appears that the following statements are true: 1) viewers can detect a schema-inconsistent object using their parafoveal vision on *the last fixation before* they first fixate such an unexpected object (Coco et al., 2020); 2) once the viewer fixates such an unexpected object, their fixation durations are longer, and they may make extra fixations in an attempt to integrate the unexpected, schema-inconsistent object into their scene representation (which we would stipulate is their *event model* of the scene; Coco et al., 2020; Cornelissen & Vö, 2017; De Graef et al., 1990; Gareze & Findlay, 2007; Henderson et al., 1999; Vö & Henderson, 2011); 3) highly salient objects tend to be fixated quickly, regardless of their scene schema consistency, which can eliminate the top-down effect of schema-inconsistency (Underwood & Foulsham, 2006).

Do such schemas play a role in guiding attention when the scene is dynamically changing, such as in movies, or in dynamic real-world scene videos? The assumption is that prior schema findings should apply to studies using dynamic scenes. However, by definition a schema can only influence our behavior if we believe it to be valid within a scene. It is currently unknown how schemas and scripts emerge or influence behavior in dynamically changing scenes. The concepts of schemas and scripts may be too limiting a theoretical construct to achieve that goal. *Event models*, which incorporate schemas, but are *specific to the given situation at that moment*, are needed to achieve greater breadth of explanatory power in terms of understanding what is expected, and how that affects attentional selection. For example, in Fig. 1, our gist of parking lots and street scenes activate those schemas, which include cars; however, as the scene continuously unfolds, *the particular car* driven by the first couple is expected to appear and reappear, because it is represented in the viewer's *current event model*. Moreover, the

car's significance to the scene is based on the viewer's event model *including the presence of the bomb in its trunk*, which is known, but not seen.

Additionally, the well-replicated finding of more fixations on scene-inconsistent objects, once they are found, is sometimes attributed to difficulty in "integrat[ion] with the scene representation" (Coco et al., 2020; Henderson et al., 1999, p. 226). However, little has been said about the nature of such *integration* processes. What is needed, then, is a theoretical framework that contains integration processes, including their inputs and outputs, what happens when integration processes fail, and how they affect attentional selection. In SPECT, such integration processes are handled by the *mapping* process of the current event model.

Studies exploring the relationship between event models and attentional selection Some prior research has considered events, either explicitly or implicitly when studying attentional selection. These studies were conducted to show how task instructions affect comprehension, but in the context of pictures and tasks that afford the creation of an event model (Borji & Itti, 2014; DeAngelus & Pelz, 2009; Kaakinen et al., 2011; Lahnakoski et al., 2014; Yarbus, 1967). However, it is important to note that the researchers did *not* describe their studies in terms of event models, thus we reinterpret them as such here.

Yarbus (1967) is recognized as a seminal study demonstrating "top-down" or "higher-level" influences on attentional selection. While this is certainly the case, we argue that it is also a seminal study demonstrating that *event models* affect attentional selection. Yarbus asked a participant several questions about a painting while their eyes were tracked. The painting showed a man entering a living room where several family members were watching him. Several of the questions required the participant to draw *inferences*, such as determining the relative wealth of the family members, estimating their ages, and what they had been doing before the visitor arrived. We argue that *inferences* evoked by the questions fostered the development of an *event model* for the picture because they emphasized situational information that could not be derived solely from the visual content. In Yarbus' demonstration, the viewer's eye movements differed considerably between questions, indicating that their attentional selection was affected by the inferential processes. Later studies have replicated and extended these results with more participants (DeAngelus & Pelz, 2009) and more stimuli, while controlling for visual saliency (Borji & Itti, 2014; but see Greene et al., 2012).

Other studies have investigated how viewers' event models influenced their attentional selection in scenes by manipulating viewers' *cognitive perspective*. Here the theoretical construct of the event model is more extensively

implicated, though not always explicitly discussed. Kaakinen et al. (2011) had participants view photographs of home interiors while imagining themselves as either a burglar or a home buyer (originally used in Anderson & Pichert, 1978). They found that adopting different perspectives guided viewers' attentional selection to objects relevant to each—for example, for the burglar perspective, a handbag or a painting, versus for the home buyer perspective, a discolored toilet seat or a wooden door. Similarly, Lahnakoski et al. (2014), manipulated viewers' imaginary perspective while they watched a video clip from an episode of *Desperate Housewives*. Participants alternatively adopted the perspectives of either a detective or an interior decorator. For the detective perspective, key information came from the two characters, usually near the center of the screen, while for the decorator perspective, key information came from the background, usually in the periphery of the screen. Viewers' fixation locations in the two perspectives differed qualitatively, but not significantly above chance. Whether the visual stimulus affords event model expression via viewer eye movements is an important factor. Static scenes impose fewer temporal demands on attention than dynamic or edited film sequences, which continually change. Because Lahnakoski et al. (2014) used edited TV clips composed by filmmakers to guide viewers' attention, this may have minimized the scope for viewers to use their event models to deviate from the filmmakers' intended viewing patterns. Likewise, although Kaakinen et al. (2011) used still images (which as noted above have less saliency-engendering features than movie clips), they found that the effect size of visual saliency was larger than that of cognitive perspective (which we construe as involving the viewers' event models).

As noted above, most of these studies investigated “knowledge-driven” top-down effects, but did so by giving participants an explicit task. That is good experimental practice, since otherwise, participants may not know what to do (Henderson et al., 2007). Thus, we assert that the effects shown are a combination of task-driven attentional selection, via the executive process of goal-setting, knowledge-driven attentional selection, based on activating schemas in semantic long-term memory, and using that activated prior knowledge to create event models.

A key unanswered question is the degree to which the top-down theoretical construct of the *viewer's event model* influences their attentional selection *in the absence of an explicit task goal*—understanding a visual narrative is something viewers spontaneously do by default. Future research should determine the extent to which visual saliency and the event model determine viewers' attentional selection, both in static events (e.g., slide shows, picture stories, or comics) and dynamic events (e.g., real-world videos, movies, virtual reality).

The Scene Perception & Event Comprehension Theory (SPECT)

SPECT provides a theoretical framework for explaining how a viewer's event model influences what they selectively attend to in real-world scenes and events. SPECT is unique in having such a comprehensive account of how high-level cognitive constructs, specifically within event models, affect something that occurs during every eye fixation, attentional selection. Before SPECT, theories of attentional selection have not included such higher-level cognitive constructs as the event model. Likewise, before SPECT, theories of real-world event comprehension have not explained how higher-level comprehension processes might influence lower-level constructs like attentional selection (Loschky et al., 2018, 2020). As shown in Fig. 3, SPECT distinguishes between stimulus features, front-end processes, and back-end processes. We describe each of these, and their sub-parts, below. We also note that Fig. 3 is a revised version of the previous SPECT box and arrow model (Loschky et al., 2018, 2020), which has primarily been edited for clarity. (The Appendix at the end of the paper describes the two substantive changes, changes for greater comprehensibility, and those changes for greater clarity.)

The stimulus

The starting point for SPECT is the visual stimulus. The stimulus can be composed of either static or dynamic real-world scenes, and can vary in complexity and realism. Stimulus properties constrain front- and back-end processes because they influence attentional selection. SPECT assumes that there are two aspects of the stimulus that affect attentional selection, saliency-engendering image features, and medium-specific features.

Saliency-engendering stimulus features

SPECT assumes that *most* saliency-engendering visual features operate similarly across different visual media. However, there is at least one medium-specific visual feature—namely, those changes across film frames that produce *motion* in film, video, and virtual reality—which is absent in still image media (i.e., comics, slideshows, or picture stories). Computed motion has a large impact on attentional selection (Carmi & Itti, 2006; Mital et al., 2010). Critically for SPECT, computing image salience requires no prior knowledge of the people in the scene, their interactions, or of what came before the current scene. Thus, in the saliency map shown in Fig. 2, from the film *Touch of Evil*, the saliency algorithm does not predict the car trunk to be

Scene Perception & Event Comprehension Theory (SPECT) V2

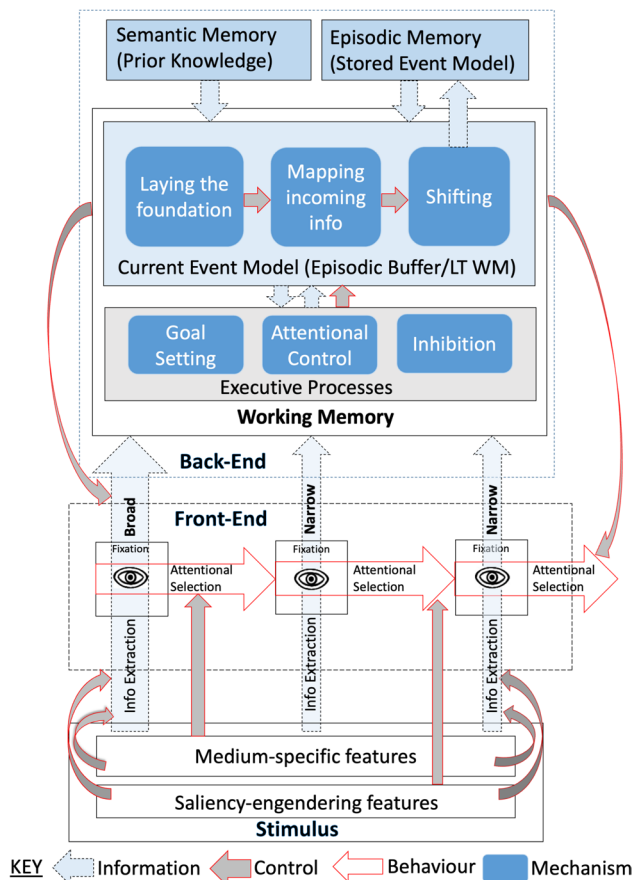


Fig. 3 The SPECT box and arrow model revised based upon the earlier version (Loschky et al., 2018, 2020) for clarity. (The Appendix at the end of the paper describes the two substantive changes, changes for greater comprehensibility, and changes for greater clarity.) (Color figure online)

salient. Conversely, if the viewer's *event model* contains the information that *there is a bomb in the car trunk*, that region becomes highly important and more likely to be attended, as shown in the bottom of Fig. 2. Note, however, that here we are only concerned with the saliency-engendering *features in the stimulus*—the actual computation of saliency is described as a sub-process of *attentional selection* in the *front-end* of SPECT.

Medium-specific features

Medium-specific features are aspects of visual design that can influence attention. These include individual pictures in wordless picture stories, the organization of panels and text in comics, and the shots, cinematography, and spatial arrangements of characters and objects in films (T. J. Smith, 2013). For example, narrative films follow conventions of

editing (Bordwell & Thompson, 2003) and other stylistic features (Hinde et al., 2018). Studies have demonstrated their impacts on eye movements (T. J. Smith & Martin-Portugues Santacreu, 2017), covert attention (Hinde et al., 2018), event perception (Cutting & Iricinschi, 2015), and affective responses (Lankhuizen et al., 2022). Such design features may guide viewers' attention to the most relevant visual information in a film shot or comic frame (Cohn, 2020; T. J. Smith, 2013). As illustrated in Fig. 1, conventions of panel structure inform the order in which panels are attended. Comics written in English, for example, are typically read in a "Z" pattern (i.e., left to right + top to bottom), though additional panel-specific constraints also affect reading order (Cohn, 2013). Some design features must be learned, such as the Z-pattern, whereas others do not. For example, after a *cut between shots* in a film (i.e., continuous camera run), viewers naturally tend to fixate near the center of the screen (T. J. Smith, 2013; Wang et al., 2012).

The front end

As shown in Fig. 3, the front end of SPECT has two components: 1) information extraction and 2) attentional selection. SPECT distinguishes between front-end processes that occur during every eye fixation, and back-end processes that occur in working memory and long-term memory. This distinction reflects the fact that only a limited portion of the visual field can be extracted during a single fixation, and minimal visual information is extracted during the saccades between fixations (Dorr & Bex, 2013; Ross et al., 2001). Once a fixation ends, the extracted information can only be processed further through passively activated long-term memories or more actively maintained short-term/working memories (Hollingworth, 2009; Irwin, 1996). Short-term/working memory is relatively abstract (Carlson-Radvansky & Irwin, 1995), and limited to only three to four objects or chunks (Irwin, 1996; Irwin & Gordon, 1998; Luck & Vogel, 1997). Thus, the distinction between the rich perceptual information available during fixations and the more limited preprocessed information used in memory is very important. Furthermore, attentional selection largely determines which elements of the available visual information are encoded and maintained in working memory (Irwin, 1996; Irwin & Gordon, 1998).

Information extraction

Figure 3 symbolizes the construct of information extraction in the front end with upward arrows from an eye icon to the back end, symbolizing passing extracted information to the event model in working memory. Information extraction encompasses everything that happens from the moment that light hits the retina until that activates the relevant semantic representations. Those include the identity of the scene, or

the person(s), object(s), or event(s) within it. Current theories of object, scene, and event identification are strongly influenced by the success of feed-forward convolutional neural networks (CNNs) at these tasks, which currently rival human accuracy (Kriegeskorte, 2015). Furthermore, studies have shown notable similarities to processing done by such neural networks and the activity of cells and regions of the ventral visual stream. Those include the visual cortical areas V1, V2, V4, inferior temporal cortex (IT), and lateral occipital cortex (LOC) for objects (Cichy et al., 2016; Yamins et al., 2014). They also include the parahippocampal place area (PPA) and occipital place area (OPA) for scenes (Cichy et al., 2017; for review see Wardle & Baker, 2020). Importantly, however, studies have also shown the importance of recurrent processing, including feedback connections, which is widespread in primate visual cortex, but not included in purely bottom-up/feed-forward CNN models (Kietzmann et al., 2019; Wardle & Baker, 2020). As we will point out later, top-down feedback connections are important for SPECT, which argues that the back-end event model influences front-end information extraction (Smith, 2021; Smith & Loschky, 2019).

Information extracted on each eye fixation is fed to the back end to support event model construction. Information extraction is either broad, encompassing much of an image, or narrow, focusing on only a small part of the scene, such as a single object (Ringer et al., 2016; Seiple et al., 2002; L. J. Williams, 1988). Broad information extraction occurs primarily in peripheral vision, which is the vast majority of the human visual field (i.e., $\pm 105^\circ$ eccentricity, horizontally from the point of fixation). Peripheral vision (from $>5^\circ$ eccentricity) is important for rapidly identifying the gist of a scene within a single eye fixation (Larson & Loschky, 2009; Loschky et al., 2019). Conversely, narrow information extraction occurs in central vision ($<5^\circ$ from fixation), which, by itself, is less important for acquiring scene gist (Larson & Loschky, 2009). On the other hand, there appears to be a bias for the first 100 ms of the first eye fixation on a scene to spread outward from central vision to peripheral vision (Larson et al., 2014).

Research also suggests that viewers transition from a broad mode of information extraction to a narrow mode as they make eye movements (i.e., saccades) in a scene (Panansch et al., 2008). Namely, the first 2–3 s of scene viewing are in the *ambient mode*, in which viewers' *saccades are longer*, going further into the visual periphery, but their *fixation durations are shorter*, indicating broad information extraction. The ambient mode allows viewers to rapidly locate key information in the scene. Then, from 3–6 s of processing of a scene, viewers shift to the *focal mode*, in which their *saccades are shorter*, staying within parafoveal or foveal vision, but their *fixation durations are longer*, together indicating more detailed, narrow information extraction. The

focal mode enables viewers to selectively extract detailed information from specific objects of interest. Viewers need approximately two fixations to extract narrower, more basic information, such as the details of a specific person or action in a scene (Larson, 2012). Increasingly detailed information is extracted on a fixation-by-fixation basis in the *front end*, and is accumulated across multiple fixations in the *back end*.

Importantly, front-end *information extraction* is influenced by the back-end *event model* processes. Specifically, having a spatiotemporally coherent event model in the back-end facilitates rapidly recognizing the gist of a scene in the front-end, as quickly as 100 ms after light hits the retina (M. E. Smith, 2021; M. E. Smith & Loschky, 2019). Conversely, there is a decrease in information extraction immediately before and after an event boundary (Crundall et al., 2002; Huff et al., 2012; Ji & Papafragou, 2022; Yates et al., 2024).

Attentional selection

Figure 3 illustrates the construct of attentional selection: rightward arrows between successive eye icons show that attentional selection mechanisms determine *what information* is processed during fixations, and *where to send the eyes* next. These two operations occur in parallel during each eye fixation (Findlay & Walker, 1999).

Here, as shown in Fig. 4 (Ptak, 2012), we update SPECT by incorporating fundamental concepts and assumptions from the *biased competition* approach to attentional selection. It is a widely supported approach, developed over the past 30 years (Adeli & Zelinsky, 2018; Desimone & Duncan, 1995; Ptak, 2012; Tsotsos et al., 1995). It draws on converging evidence from behavioral and neurophysiological research (Bichot et al., 2005; Buffalo et al., 2010; Desimone & Duncan, 1995), and computational modeling (Adeli & Zelinsky, 2018; Hamker, 2004; Itti & Koch, 2000; Tsotsos et al., 1995). Incorporating these important details to the SPECT theoretical framework opens avenues for future computational modeling efforts which integrate existing computational models of attentional selection. Critically, the biased competition approach captures how a viewer's task, prior knowledge, and—as we argue here—*their event model*, bias bottom-up competition for attentional selection.

As the name suggests, the approach includes bottom-up *competition* and top-down *bias*. Bottom-up competition is represented by a *saliency map*; top-down biases transform this saliency map into a *priority map*.

In describing the biased competition approach to attentional selection, we will refer to processing “units” within “layers,” consistent with computational models of object recognition discussed above. This is more abstract than “cells” within “brain areas.” But the abstraction is based on a relatively strong correspondence between *units in layers* and *cells in brain regions* as shown in computational

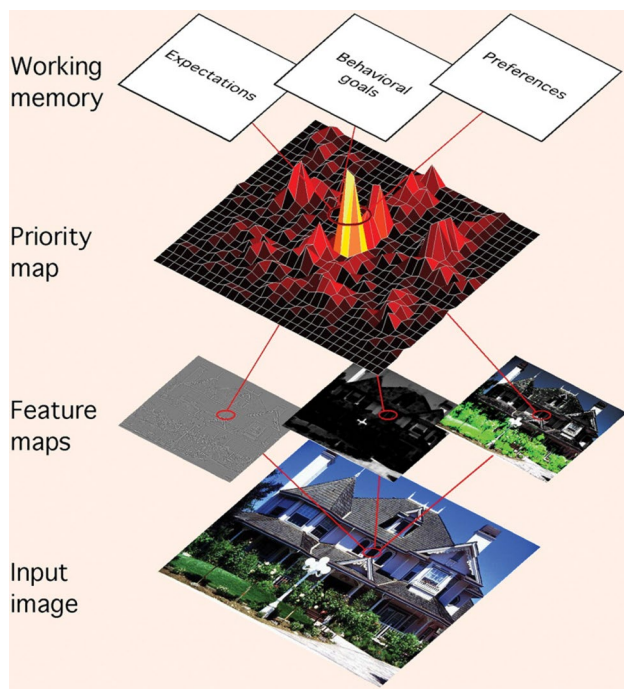


Fig. 4 Graphic representation of the biased competition approach to attentional selection (Ptak, 2012). (Color figure online)

neurophysiological studies (Nonaka et al., 2021; Yamins et al., 2014). In Fig. 4, each rectangular image represents a layer. In this architecture, bottom-up input ascends through multiple layers (e.g., hidden layers in a connectionist or deep neural network). Saliency and priority maps are implemented in higher layers, as shown in Fig. 4.

A key feature linking these maps is *retinotopic mapping*. This enables a spatial correspondence between regions in an input image, and in layers, as illustrated in Fig. 4. Retinotopic mapping enables selective attention to direct eye movements. The winning location in the priority map becomes the next fixation point.

SPECT captures the inherent strengths and limitations of peripheral vision by incorporating *cortical magnification of the fovea*. Namely, the fovea and parafovea (center of vision) cover only about 10° of the horizontal visual field, but have ~40% of the cells in area V1. This leaves the remaining 210° of the horizontal visual field¹ to be processed by only the remaining 60% of the cells (Curcio et al., 1990; Daniel & Whitteridge, 1961). Therefore, two small objects in the fovea may each be represented by their own separate cortical cells, but in the periphery the same two objects may share a single cell. Such cases produce a jumbled neural representation, and thus perceptual confusion, called “crowding” (Strasburger & Malania, 2013). Most saccades are short to

medium, with very few long ones, and this is explainable in terms of cortical magnification of the fovea (Raabe et al., 2023). Thus, models of attentional competition should incorporate cortical magnification, though only a few currently do (Adeli et al., 2017; Adeli & Zelinsky, 2018; Raabe et al., 2023).

SPECT also acknowledges the well-known divergence in visual processing between the “what” and the “where/how” pathways (for reviews, see Kravitz et al., 2011, 2013; Milner & Goodale, 2008). *Entities* (people, objects) are represented in the “what” pathway; whereas, the *locations* of those entities, and their *actions* are represented in the “where/how” pathway. Saliency and priority maps likely exist in both pathways (Fecteau & Munoz, 2006; Ptak, 2012; Zelinsky & Bisley, 2015).

Within each layer, *competition* selects the winning location for the next eye movement. The exact mechanisms of reducing the activity of most units, while amplifying ultimately one, vary between biased competition computational models (Adeli & Zelinsky, 2018; Hamker, 2004; Itti & Koch, 2000; Tsotsos et al., 1995). Regardless of the mechanism, the bottom-up/feed-forward sweep of activity combined with an inhibitory mechanism produces a “winner” within a retinotopic map. As shown in Fig. 4, such competition likely occurs in multiple feature maps, each for processing a different type of visual feature (e.g., color, luminance, orientation; Itti & Koch, 2000). Other high-level maps may also exist with winners being faces, or written text, both of which are known to strongly attract viewers’ gaze (Cerf et al., 2009). Winners from these maps compete in the final saliency map, where a single retinotopically mapped location is selected as most salient. This is shown in Fig. 4 as the central map having multiple peaks. But why is it labeled as a “priority map” instead of a “saliency map”?

The answer is that a winner based on saliency can be overridden by top-down *biases*, thus transforming what was the saliency map into a *priority map*. As shown in Fig. 4, top-down bias spreads through feedback connections that begin at the top layer and go down sequentially through lower layers (Adeli & Zelinsky, 2018; Hamker, 2004; Tsotsos et al., 1995). Again, the exact mechanisms of excitation and inhibition vary among computational models. Whether the top-down bias is strong enough to overcome bottom-up saliency determines the winner of the biased competition for each new fixation.

The top of Fig. 4 illustrates several well-documented sources of top-down bias. One is the viewer’s task/behavioral goals (e.g., find X). Another is their expectations from prior knowledge/schemas (e.g., places where X are often found). A third is learned preferences based on reward values (e.g., X are good, or Y are dangerous). But, as noted in Fig. 4, all are contained in *working memory*, which is privileged for visual attentional selection (Han &

¹ We intentionally combine the left and right hemifields here, to capture the entire visual field, while recognizing that each hemisphere of the brain only processes half of that.

Kim, 2009; Olivers et al., 2006; Soto et al., 2005). Working memory can guide attention regardless of whether its contents are visual or verbal (Soto & Humphreys, 2007), and whether one or multiple objects are held in mind (Bahle et al., 2018).

Critically, SPECT proposes that the viewer's *current event model* residing in working memory can exert such a top-down *bias* on attention. As we will discuss in the section entitled, "The current event model," the event model in working memory tracks entities (e.g., people, or objects), actions, locations, time, goals, and causal relations. Thus, the event model representation in working memory could create a top-down bias focusing on a given person, location, time, action, or goal-object, which in turn, could amplify the activity of relevant units corresponding to those event indices in the *what* or *where/how* pathways.

The roles of task versus prior knowledge and event models in biased competition for attentional selection The biased competition model explains how task-driven goals can override visual stimulus features in determining what is attended (e.g., Hutson et al., 2017, Exp. 2B). However, a relatively unanswered question is how prior knowledge, and event models can enter this competition, and sometimes win. Imagine two different viewers of the 3-min scene illustrated in Fig. 1. The first viewer watches it from the beginning of the scene, so they know that the driving couple have a ticking *time bomb* in the *trunk of their car*. The second viewer does not start watching until approximately 1 min later in the scene, when the camera follows the walking couple as shown in sixth frame of Fig. 1, so they have no idea of either the time bomb or the driving couple. As shown in the bottom of Fig. 2, as the scene progresses, the car with the time bomb is highlighted in the event models for the first viewer. As we will discuss later in detail in the section entitled, "Stimulus features need to afford knowledge-driven or task-driven attentional control" (and Fig. 5), the car is just part of the scene background for the second viewer. SPECT assumes that competition for attentional selection will be biased differently between the two viewers, based on their different event models, and there is evidence supporting that (Hutson et al., 2017, 2022). However, SPECT also assumes that the influence of the event model is attenuated when viewers have a task-driven goal, or there is high visual saliency, competing with the event model to influence attentional selection (Hutson et al., 2017). Candidate brain areas for processing event models, which would likely enter this competition are those involved in the default mode network (DMN), a set of interconnected brain regions involved in internal thought, specifically for processing event models, such as the precuneus, the posterior cingulate cortex, and the medial prefrontal cortex (M. E. Smith et al., 2025).

The back end

As shown in Fig. 3, the back end involves processes that involve building the *current event model* in working memory, which in turn is important for building the *stored event model* in episodic long-term memory. Understanding the back-end is important for understanding attentional selection. This is because SPECT assumes that processes affecting construction of the current event model in working memory dynamically affect attentional selection as the event unfolds. Figure 3 shows that the *back end* of SPECT has four main components: 1) the current event model in working memory; 2) episodic long-term memory, which contains the stored event model; 3) semantic long-term memory, which contains prior knowledge (e.g., schemas and scripts); and 4) executive processes, such as goal-setting, attentional control, and inhibition, which affect what information is maintained in the current event model in working memory. The back end of SPECT is informed by theories of event comprehension that were first proposed and tested in the context of reading narrative texts (Gernsbacher, 1990; Zwaan & Radvansky, 1998) and more recently further developed in the context of real-world event cognition (Kurby & Zacks, 2008; Radvansky, 2012; Radvansky & Zacks, 2011).

The current event model

As shown in Fig. 3, the *current event model* is assumed to be dynamically constructed in working memory. From the perspective of Baddeley's (2000) working memory model, the current event model would use the episodic buffer. From Ericsson and Kintsch's (1995) perspective, the current event model may use long-term working memory. SPECT also includes the critical theoretical construct of *event indices*—these are the high priority pieces of information that viewers use to construct the current event model, and are likely to be recalled later from long-term episodic memory (Zwaan & Radvansky, 1998). Event indices include *events* (e.g., actions), *entities* (people, objects), *time*, *location*, *goal(s)* (of entities), other *causal relationships* (both social and physical), and entities' *emotions* (Gernsbacher et al., 1992). Figure 1 illustrates a number of events—for example, a time bomb being placed in the car trunk, the couple who own the car unknowingly driving off with it, and another couple walking nearby. These events typically involve hierarchically nested actions (e.g., opening the trunk, placing the bomb in it, closing the trunk), and a series of state changes (e.g., the bomb goes from outside the trunk to inside it), which unfold over time (e.g., a few seconds) and across locations (e.g., the parking lot, city streets). Monitoring these event models in working memory allows viewers to track entities, their

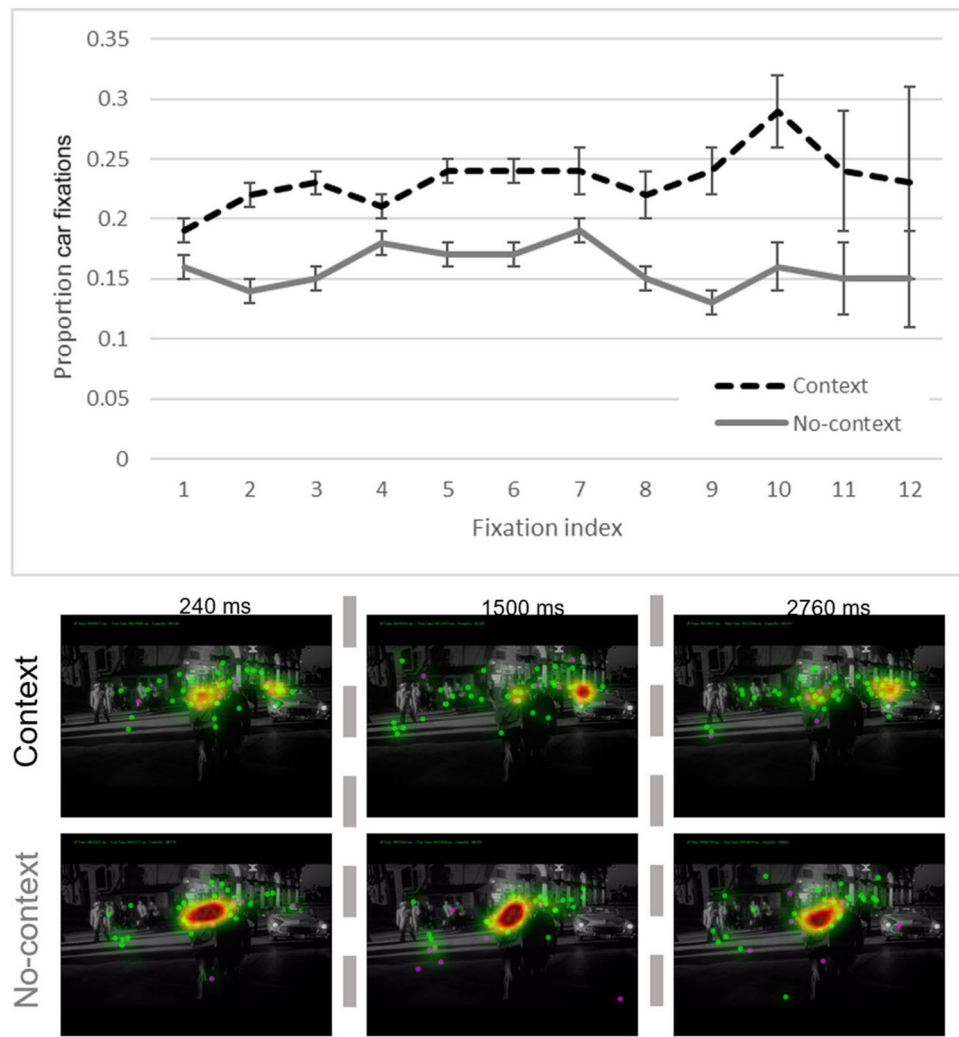


Fig. 5 Above: Proportion of participants who fixated the car with the bomb for each fixation on a given slide. Context participants are in dashed black; no-context in solid gray. Error bars are standard errors, which increased with fixation index due to lower N with increasing numbers of fixations during the 3-s viewing period (i.e., not many participants made 11 or 12 fixations on a single slide). Below: Example heatmap images for the context and no-context conditions at three

time points throughout the same slide. The first time point is 240 ms into the slide, the second is halfway through (1,500 ms), and the final is 240 ms from the end of the 3-s slide. Context condition participants can be seen to fixate the car more often than the no-context participants. Figure and caption recreated with permission from Hutson et al. (2022). (Color figure online)

goals, causal relationships, emotions, and the spatiotemporal framework of the unfolding narrative (Magliano et al., 2012; Zwaan & Radvansky, 1998).

SPECT draws upon the structure building framework (Gernsbacher, 1990) to describe how the current event model is constructed, and how it is later stored in episodic long-term memory as a *stored event model*. The stored event model is construed as a hierarchically organized representation of prior current event models (Zwaan & Radvansky, 1998), which are organized around the goal-structure of characters/agents (Trabasso et al., 1989). The structure building framework describes three phases of

event model construction: *laying the foundation*, *mapping*, and *shifting*.²

² Note that while SPECT has adopted the three phases from Gernsbacher's (1990) structure building framework, it has lightly modified certain constructs of it. First, Gernsbacher does not distinguish between long-term and working memory, whereas SPECT does. Second, Gernsbacher uses the term "structures" rather than the specific theoretical constructs of the *current event model*, or *stored event model*, used by SPECT. Third, Gernsbacher articulated these processing phases as operating on the coarse/global scene structure of narratives (e.g., goal-directed actions occurring in a specific time and place within a narrative world). However, SPECT assumes that the three phases operate at both *fine* and *coarse* grain sizes (Kurby & Zacks, 2011).

Laying the foundation As shown in Fig. 3, the first phase of constructing the current event model is called *laying the foundation*. The foundation of the event model contains a representation of its initial spatiotemporal context, key entities, and actions/events that are acquired within the viewer's first few fixations (Larson, 2012). Importantly, SPECT assumes that the *gist* of a scene is critical for laying the foundation (as described above for the *front end* in terms of *broad information extraction*). A concrete example of laying the foundation in Fig. 1 is the beginning of the first event of "the villain planting the bomb in the car." The beginning of the first event starts with a close-up of a man's hands setting the timer on a bomb. The viewer's first fixation would extract the information of "hands" (of a person), "holding" an inanimate entity, which may be recognized during a second fixation as a "time bomb." The person whose hands are seen will be perceived as an "agent" (the doer of an action), who is "holding" (an action) the "patient" (the bomb). Prior research has shown that the "agent" of an action is rapidly recognized within a single eye fixation, before recognizing the "patient" (Dobel et al., 2007; Hafri et al., 2013).

Thus, all of the above information could be extracted automatically within the time frame of one to two eye fixations (roughly 300 ms each; Dobel et al., 2007; Hafri et al., 2013; Larson, 2012). The viewer's first few fixations, which have extracted "man holding and setting a time bomb," may activate "terrorist" or "spy" scripts from semantic long-term memory. Thus, the person holding the bomb may be given a moral interpretation (e.g., "villain") by roughly 300–700 ms after fixating it (Ma et al., 2024).

Then, as the close-up on the villain's hands zooms out to show more of the scene, the foundation of the viewer's current event model will include the spatiotemporal scene gist (city, parking lot, evening; Fei-Fei et al., 2007; Greene & Oliva, 2009; Larson & Loschky, 2009; Loschky et al., 2007). The sum total of this beginning information could constitute the foundation for the viewer's event model (e.g., a villain [agent] has set the timer [action] on a time bomb [patient], in a city parking lot [place], at night [time]). Importantly, these key pieces of information are all event indices (Radvansky & Zacks, 2014; Zwaan & Radvansky, 1998).

Laying the foundation is an iterative process. As shown in Fig. 1, in the 3-min *Touch of Evil* scene, there are at least six distinct events that require laying the foundation. Those are: 1) the villain placing the bomb in the car, 2) the first couple unwittingly driving off in the bomb-laden car, 3) the second couple walking through the city streets, 4) both couples simultaneously going through the border checkpoint, 5) the walking couple kissing. The immediately following shot (*not* shown in Fig. 1), shows a sixth

event, the bomb exploding. Thus, the process of laying the foundation would occur in the first fixations at the start of each new current event model. When the viewer will lay the foundation is determined by the products of the mapping and shifting processes, described below. Finally, laying the foundation for the first event model of a narrative, which should occur during the first scene of a film, the first page of a picture story, or the first panel of a comic strip, typically produces longer processing times than subsequent ones (Foulsham et al., 2016; Gernsbacher, 1983). Similar results are well known for reading times of the first sentence in story compared with subsequent sentences (Haberlandt & Graesser, 1985). Thus, laying the first foundation represents a necessary cognitive investment that enables faster processing times later.

Mapping Figure 3 shows a rightward arrow from laying the foundation to mapping, because the former enables the latter. Viewers develop their current event models by mapping newly extracted information onto its foundation. However, whether incoming information is mapped onto the event model's foundation depends on how coherent and/or predictable that information is with the current event model. For example, as viewers process the events that make up the first *Touch of Evil* scene, they must judge whether the incoming subevents (e.g., villain setting a time bomb, putting the bomb in a car trunk, running away) belong together in the current event model. Two important processes support mapping.

The first process to support mapping is situational *coherence/continuity monitoring* (Gernsbacher, 1990; Radvansky, 2012; Zwaan & Radvansky, 1998). Specifically, as events unfold, people habitually assess the extent to which there is continuity in spatial locations, time, entities, causality, and the goals of agents (Huff et al., 2014; Zwaan et al., 1995). As a viewer watches the *Touch of Evil* opening scene, they will make eye movements and fixations on the villain, the car, and the owners of the car as they approach it, get in, and drive away. To the extent that the information from that series of fixations forms a coherent set of interrelated events for the viewer, the information will be mapped onto the viewers' evolving current event model.

The second process that supports mapping is *inference generation* (Graesser et al., 1994). Bridging inferences support *backward mapping*, which often involves inferring how current events are causally related to known prior events (Trabasso et al., 1989). For example, the first shot *after* the entire 3-min 20-s long shot depicted in Fig. 1 shows the car exploding. Viewers must map the event of the car exploding to the event of the bomb being placed in its trunk, and they

can do so by generating the bridging inference that the car exploded *because of* the bomb. (Viewers who have forgotten about the bomb, because it has not been visible for roughly 3 min, are surprised when the car explodes; Hutson et al., 2021).

The third process that supports mapping is *prediction generation*. The structure building framework did not emphasize *prediction* as a mapping process, but contemporary theories of event cognition assume that prediction is essential for event comprehension (Zacks et al., 2011). Prediction can be construed as *forward mapping* (using past events to predict future events). The perceived suspense of the *Touch of Evil* scene arguably requires the viewer to *predict* that the bomb will explode. The current event model in working memory is incrementally updated so long as new incoming event indices are perceived as continuously following from prior events (Kurby & Zacks, 2012) and that includes making bridging inferences to maintain coherence (M. E. Smith, Hutson, et al., 2024a, 2024b).

SPECT assumes that attentional selection is affected by the output of mapping processes in the current event model in two ways: 1) visual search and 2) predictive eye movements.

First, we consider search. When viewers experience a coherence gap during mapping in an event, they often engage in bridging inference generation to resolve it. But to bridge the gap, viewers may have to search the scene to find appropriate linking information (Hutson et al., 2018). Similarly, immediately following film cuts (i.e., transitions between separate shots), viewers make more eye movements, exploring the scene to establish event model continuity with the pre-cut shot. Specifically, viewers make more exploratory saccades after cuts to new scenes than to cuts continuing the same scene (T. J. Smith & Martin-Portugues Santacreu, 2017). A similar pattern of eye movements during event mapping is observed in the reading literature, which shows that regressive saccades search for information needed to repair breaks in coherence during text comprehension (Calvo et al., 2001; Ehrlich & Rayner, 1983; Poynor & Morris, 2003).

Second, consider predictive eye movements. These are found in both language comprehension, and action observation (Coco et al., 2016; Eisenberg et al., 2018; Flanagan et al., 2013). For example, in the *Touch of Evil* scene, consider the second event in which the first couple gets in the car and drives away. Some viewers will predict that the couple will get into the car when they start walking towards it, and so make a predictive saccade to the car seats (e.g., Eisenberg et al., 2018).

Shifting Figure 3 shows a rightward arrow from mapping to the third phase of constructing event models: shifting. This connection reflects that shifting occurs when mapping

processes fail, prompting the viewer to shift and create a new event model. In SPECT, shifting is conceptually aligned with *event segmentation*. When ongoing information no longer maps onto the current event model, the comprehension system shifts to build a new model, and this shift is experienced as an event boundary. A great deal of behavioral and neurophysiological evidence shows that the event comprehension system parses larger events into smaller ones (Zacks, 2020). The current event model is maintained by a network of brain regions, including those that make up a subsection of the brain's *default mode network*, known as the *midline default network core* (DN; Baldassano et al., 2017; M. E. Smith et al., 2025; Stawarczyk et al., 2021). Functional magnetic resonance imaging (fMRI) and electroencephalography (EEG) data show that neural activity patterns are stable as a viewer maps incoming information onto the current event model, but that rapid shifts in these patterns occur when shifts at event boundaries occur (Baldassano et al., 2017; Silva et al., 2019).

The *event segmentation task* is often used to identify when shifting occurs (Zacks et al., 2009). In this task, participants watch a movie, view a picture story, or read a text and indicate each time they perceive a new event has begun. We collected event segmentation data for the *Touch of Evil* scene. Figure 1 highlights the ten moments with the highest probabilities of perceiving event boundaries (depicted with red borders). Data from studies using this task indicate that event segmentation is hierarchically organized such that many fine, incremental events make up one course global event (Kurby & Zacks, 2011). Event segmentation occurs spontaneously, and viewers show high agreement in their segmentation behavior across the different levels of granularity (Zacks et al., 2010).

SPECT assumes that the processes of situational continuity monitoring, prediction, and bridging inference generation inform not only mapping, but also shifting (M. E. Smith, Hutson, et al., 2024a, 2024b). Shifting is probabilistically related to the perceived *number of changed event indices* (Magliano et al., 2001; Zwaan & Radvansky, 1998). In the *Touch of Evil* scene, shifting occurs when the villain moves away from the car, and when the man and woman get into the car. At this moment, there are changes in multiple event indices: the characters (one character leaves, then two others are introduced) and goals (one goal ends for the villain, and a new one begins for the man and woman). SPECT does not make specific claims regarding prioritization of specific situational indices in informing when shifting occurs. Shifting is also related to *prediction error*. The cognitive system is set up to support moment-to-moment predictions that events in the current event model will be spatiotemporally contiguous. When such a prediction fails, an event boundary is perceived (Kurby & Zacks, 2008; Zacks et al., 2011). A prediction

error occurs when the walking couple is introduced because no content of the current event model can anticipate seeing *those particular* people. Finally, shifting is also related to a *failure of inference processes* to resolve a coherence gap (Brich et al., 2024; M. E. Smith, et al., 2024a, 2024b). M. E. Smith, et al. (2024a, 2024b) showed that at coherence gaps, there was a negative correlation between the likelihood of making a causal bridging inference to resolve the gap (based on think aloud data) and the likelihood of perceiving an event boundary (based on event segmentation data).

When a viewer *shifts* to create a new current event model, the *stored event model* in long-term episodic memory is updated to reflect the contents of the just previous *current event model* (Radvansky, 2012; Zwaan & Radvansky, 1998). This is shown by the fact that people are less accurate at retrieving information from previous events after perceiving an event boundary (Gernsbacher, 1985; Pettijohn & Radvansky, 2016; Swallow et al., 2009). Further support comes from studies measuring fMRI and EEG brain activity at event boundaries. Specifically, shifts in neural activity patterns at event boundaries across cortex are accompanied by increases in neural activity in the *hippocampus* (Baldassano et al., 2017; Ben-Yakov & Dudai, 2011). Assumedly, boundary-evoked hippocampal activity reflects early long-term memory consolidation processes, where the contents of the previous current event model are updated in episodic long-term memory as part of the stored event model. This raises the question of whether updating an event model in episodic long-term memory only occurs at coarse event boundaries, with transitions between fine event boundaries only occurring within the current event model in working memory. One of the few studies to address this question found that EEG correlations with previously encoded stimuli were greater at coarse boundaries than at fine boundaries (Sols et al., 2017). Furthermore, memory decreased after coarse event boundaries, but *not* after each new image of an event, or page in a picture story (Gernsbacher, 1985; Swallow et al., 2009). Finally, as noted earlier, there appears to be a brief time of limited information extraction for roughly 1 s before and after an event boundary (Crundall et al., 2002; Huff et al., 2012; Ji & Papafragou, 2022; Yates et al., 2024).

Coordination of front- and back-end processes

In Fig. 3, the coordination of front- and back-end processes is indicated by downward arrows from the back-end event model to the front-end attentional selection and information extraction processes, and by upward arrows from information extraction to the event model. Such coordination is further illuminated by considering the previously discussed front-end ambient-to-focal shift in eye movements (Pannasch et al., 2008) together with the three event model construction processes (laying the foundation, mapping, and shifting).

Information extracted during the first two seconds of viewing a static scene image, during the ambient mode, likely serves to lay the foundation for the current event model. Then, the focal mode occurs from roughly 4–6 s into viewing a scene, which may correspond to the mapping process. Once the back-end shifting process occurs, and laying the foundation restarts, and the front-end ambient mode will begin again (Eisenberg & Zacks, 2016). Thus, we can give additional motivation for these front-end changes to attentional selection, from ambient mode (broad) to focal mode (narrow) in terms of the three back-end event model processes of laying the foundation, mapping, and shifting.

Prior knowledge

Episodic long-term memory Episodic long-term memory is a key part of SPECT because it contains the *stored event model*. Figure 3 shows a downward arrow from episodic long-term memory to the three phases of structure building, and an upward arrow from shifting to episodic long term memory. The downward arrow reflects the fact that the contents of episodic memory can become activated in WM. This occurs to the extent that current event model information serves as a retrieval cue for the stored event model (Kintsch, 1988). Accessing the stored event model is important for laying the foundation for a new current event model because agents, entities, locations, and goals that are represented in the stored event are often reintroduced during this stage. The stored event model is also an important source of information used in mapping (Graesser et al., 1994; Kintsch, 1988) because accessing *the stored event model* supports *continuity monitoring* (Zwaan & Radvansky, 1998), and the generation of *bridging inferences* (Graesser et al., 1994) and *predictive inferences* (Loschky et al., 2015).

The Arrow going up from shifting to episodic memory reflects that the stored event model is updated when shifting occurs (Baldassano et al., 2017; Sols et al., 2017). Viewers shift to create a new event model (as indicated behaviorally by making an event segmentation response) roughly every 12 s in studies by Zacks and colleagues. Thus, when watching a feature length movie, a viewer's *stored event model* for the entire movie may be the concatenation of a hundred or more previous current event models. Research has studied the processes involved in structuring such stored event models in episodic long-term memory, and later retrieving them (Radvansky, 2012; Radvansky & Zacks, 2014).

Semantic long-term memory Semantic long-term memory refers to general knowledge of the structure of the world. Figure 3 shows a downward arrow from semantic memory to the three stages of structure building. Similar to episodic memory, the downward arrow reflects the fact that semantic

memory can become part of the *current event model*. This happens to the extent that information in the current event serves as a retrieval cue for *semantic memory* (Kintsch, 1998). Semantic memory can be accessed during all three stages of structure building. This assumption is grounded in theories and research on both scene perception (Eckstein et al., 2006; Palmer, 1975; M. E. Smith & Loschky, 2019) and text comprehension (Kintsch, 1988, 1998; Myers & O'Brien, 1998). Both areas assume that the activation and use of semantic knowledge is necessary for recognizing and comprehending events. SPECT assumes that semantic long-term memory for types of events is represented in *schemas*, also known as *scripts* or *frames* (Barsalou, 1992; Elman, 2009; Hare et al., 2009). Event schemas represent the key information that probabilistically co-occurs in the context of events, and reflects the specific contents of event indices: *time* (e.g., of day, or of year), *location*, *entities* (people, animals, objects), and *actions/events* (both intentional and unintentional; Magliano et al., 2008). The probability that information in an event schema becomes activated depends on *constraint satisfaction* via *spreading activation* from the products of information extraction (McNamara & Magliano, 2009).

Executive processes

As shown in Fig. 3, SPECT includes executive processes to account for non-automatic influences on attentional selection. These executive processes include cognitive flexibility via *goal setting* (i.e., identifying and voluntarily switching between tasks), *attentional control* (i.e., volitionally paying attention to task-relevant information), and *inhibition* (i.e., intentionally ignoring irrelevant information; Diamond, 2013). Executive processes are often considered to include working memory due to its *active* and *volitional* maintenance of information (Diamond, 2013). However, within SPECT we conceptualize working memory as a separate memory process *acted upon* by executive functions, namely volitional attentional selection to encode task-relevant information into working memory.

Executive processes are thought to be important in enabling task-driven influences on attentional selection and information extraction. Task-driven influences are volitional in nature (Baluch & Itti, 2011; Theeuwes, 2018), and thus appear to be qualitatively different from experience/knowledge-driven influences, which are typically automatic (e.g., schema activation). SPECT considers executive processes (i.e., the management of goal-directed behaviors) to be part of the back end, because they operate on time scales far longer than single fixations and may occur infrequently, only when necessary due to the effort involved. Executive processes are mediated by frontal and prefrontal brain areas, specifically the dorsolateral prefrontal cortex (DLPFC)

and the frontal eye fields (FEF). Both areas are involved in *attentional control*, a key executive process in SPECT. An extreme example of such attentional control is the *anti-saccade task*, in which, when a simple target appears on a screen, the viewer must make a saccade to the opposite direction (Hallett, 1978). Anti-saccades require *goal setting* by the viewer, to exert *attentional control* of their eyes, and *inhibiting* saccades to the distractor (Mitchell et al., 2002; Unsworth et al., 2004), all key executive processes in SPECT. Most people can do the anti-saccade task reasonably accurately (a sample of >2,000 healthy young men made 23% errors), but much more slowly than they make pro-saccades to the visual target (Evdokimidis et al., 2002). Importantly, making anti-saccades is more difficult when a person's executive working memory is taxed (e.g., by doing mental math, or other working memory tasks; Mitchell et al., 2002; Roberts et al., 1994; Unsworth et al., 2004).

A more ecologically valid example of deliberate and task-driven executive processes influencing viewers' attentional selection is visual search (Wolfe, 2021; Zelinsky et al., 2020). By definition, all cases of visual search involve the executive process of goal setting, because search is inherently goal directed. Nevertheless, in easy cases of visual search, such as the earlier example of searching for a chimney in a picture of a yard with a house (Eckstein et al., 2006), very little task-driven attentional control or inhibition are needed. However, the need for greater attentional control and inhibition increases as the difficulty of the visual search increases, because there is less guidance by prior knowledge, greater visual similarity between the target and nontargets, or greater saliency of the nontargets. The most famous ecologically valid examples are the "Where's Waldo" images, during which the goal and target template for Waldo must be actively maintained in working memory, and saccades to similar distractor characters inhibited, to successfully locate Waldo in complex scenes (Smirl et al., 2016; T. J. Smith & Henderson, 2011). Importantly, SPECT assumes that visual search can also be motivated by the nature of the event model. This assumption is influenced by what is known about visual search in the context of reading. When readers encounter comprehension problems (which in the context of narratives, involve the event model), they engage in regressive eye movements to correct those problems (Calvo et al., 2001; Foulsham et al., 2016; Poynor & Morris, 2003; Wiley & Rayner, 2000). We assume that when visual saliency is low in the context of a dynamic event, the event model can also affect attentional selection.

Hutson et al. (2018) showed that visual search produced additional eye fixations when a coherence gap caused difficulty in the mapping process, which could be resolved based on extracting information from the pictures. This is consistent with the assumption of SPECT that event models can trigger executive processes when there is a difficulty in

creating the current event model. However, an open question is whether the visual search for information to create a bridging inference in picture stories was volitional. We suspect that those fixations are not the result of effortful processing, based on prior research on text processing, which shows that many inferences needed to create a coherent event model result from passive, memory-based processes (e.g., Myers & O'Brien, 1998). Additionally, executive load when processing visual narratives that require generating bridging inferences attenuates inference processes, but does not eliminate them (Magliano et al., 2016). Clearly it is possible for comprehension problems in the event model to lead to deliberate, effortful executive control over attention to resolve them. However the visual salience of the narrative and the encoding conditions (e.g., self-paced processing) must afford that executive control (Magliano, 2024). Thus, a key hypothesis generated by SPECT, which we have tested, is that goal-driven executive attentional selection, involving visual search, occurs when there is difficulty in processing the current event model. However, such visual search will vary in the degree to which it is effortful.

Importantly, we have also shown that executive attentional control can volitionally override the impact of stimulus saliency of a dynamic film stimulus when the task is at odds with comprehending the narrative (Hutson et al., 2017; Simonson et al., 2021). In Simonson et al. (2021), participants performed one of two tasks, which were either congruent with comprehending the narrative (watching the clips in preparation to draw a four-panel comic of the narrative) or at odds with comprehending the narrative (watching the clips in preparation to draw a map of the key landmarks and their relative locations). The map task can be conceptualized as involving a high-level visual search for spatial landmarks. The results showed that participants looked more at the background of the scene when performing the map task than the comic task—thus, task-driven attentional selection overrode the tendency to look at the main characters (agents) in the center of the screen. To determine whether the map task required executive attentional resources, in another condition, participants had a secondary auditory 2-back task, while watching the silent film clips. The executively demanding 2-back dual task reduced the extent to which those in the map task looked at the background, but it did not reduce eye movements to the central part of the screen and the agents in the comic task. In sum, viewers can executively control their eye movements while watching narrative film—perhaps via selective inhibition of visually salient features around the screen center—and doing so can override all the stimulus factors in film used to guide their gaze, but it requires effortful executive control. Similar evidence of inhibitory mechanisms has been shown by Bezdek and colleagues (Bezdek & Gerrig, 2017; Bezdek et al., 2015) who showed narrowing of attentional focus and reduced activity

in the default mode network during suspenseful film scenes, suggesting inhibition of irrelevant sensory features.

Conversely, during typical movie viewing, the degree to which viewers engage executive processes is unclear. The highly designed nature of most film stimuli (e.g., shot composition, framing, editing) is believed to encourage stimulus-driven (i.e., *not* executively driven) attention to the most meaningful information for understanding the movie (T. J. Smith, 2013). Nevertheless, prioritization of the bottom-up features of film scenes may actually constitute an executively maintained attentional prior. Namely, studies attempting to artificially capture gaze with sudden onsets during film viewing have failed to do so (Hinde et al., 2017). Fixation durations immediately following distractor onsets are artificially elongated, suggesting distractors are seen and processed but capture is suppressed, possibly due to their irrelevance to the film viewing task (Hinde et al., 2017). Whether such inhibition of saccades occurs volitionally via executive processes, or more automatically once an overarching viewing goal is set at the start of viewing, is unknown.

Thus, while viewers can volitionally control where they look, it unlikely their first option, because it requires cognitively “expensive” executive resources, that are slow and effortful. Instead, in everyday viewing of visual narratives, task-driven control of attentional selection may be the exception rather than the rule (Theeuwes, 2018). This reasoning leads to a key testable assumption of SPECT—namely, while viewers watch visual narratives, knowledge-driven top-down influences on their attentional selection are more common than task-driven top-down influences, which, if at odds with following the narrative, will require more executive resources and be more effortful.

Using SPECT to study the role of event models in attentional selection in scenes

SPECT provides a theoretical framework for generating testable hypotheses about how moment-by-moment real-world scene understanding, formalized in terms of event model processes, guides attentional selection. SPECT assumes that the extent to which event models exert a knowledge-driven top-down influence on attentional selection depends on their competition with bottom-up features (visual salience, or medium-specific features). Specifically, top-down effects of the event model should be stronger when bottom-up visual features are weaker. We next review studies demonstrating this push-and-pull between the stimulus and the viewer's current event model to control of attentional selection. We show this during both real-world events, and visual narratives, including film clips, slide shows, and picture stories, which vary in their bottom-up features. Sequential visual narratives (picture stories or slide shows of films) are static,

whereas films are dynamic. The dynamic nature of film introduces *motion*, a potent *stimulus feature* that strongly competes for attentional selection. Accordingly, SPECT predicts that top-down influences of the event model on attentional selection will be more likely in *sequential* visual narratives (e.g., picture stories, or slideshows) than in filmed narratives. Also important to note in these studies, the need for volitional tasks beyond basic comprehension was minimal.

Stimulus features must afford knowledge-driven or task-driven attentional control

Films, by definition, move. In doing so, they exert the maximum potential to guide viewer attention via visual saliency (Carmi & Itti, 2006; Mital et al., 2010). Furthermore, filmmakers frequently make cuts between shots (i.e., periods of continuous recording of a single camera), immediately after which viewers look at the center of the screen (T. J. Smith, 2012a; Wang et al., 2012). Together, these two stimulus factors (the saliency-engendering feature of motion, and the medium-specific feature of cuts) minimize the viewer's opportunities to use their event models to guide their attention. Studies of gaze behavior during film viewing are consistent with SPECT's assumption that bottom-up features in filmed narratives (e.g., motion) decrease the likelihood that event models will exert a top-down influence on attentional selection (Hutson et al., 2017; Loschky et al., 2015). Loschky and colleagues (2015) manipulated the viewing context of participants who watched a short film clip from a James Bond film, by showing the preceding 3 min of the clip, in the context condition, or not, in the no-context condition. The manipulation created a strong difference in viewers' event models for the clip, as measured by viewers' think-aloud protocols, event segmentation, and predictions of what would happen next. But these big differences in event models produced only small (though statistically reliable) differences in eye movements. The only moment when there was a significant effect of viewers' event models on their attentional selection was in a shot with virtually no motion. The authors concluded that this strong attenuation of the effect of the event model on attentional selection, which they called *the tyranny of film*, was due to the nature of the film stimulus. Namely, the clip used Hollywood style filmmaking techniques to direct viewers' attention, including editing with many short (2 s) shots, showing only a single character, mostly at the center of the screen. This left little room for viewers' event models to influence their attentional selection. Furthermore, there was no key *target of attention* to differentiate the attentional selection of viewers having different event models.

Thus, a follow-up study by Hutson et al. (2017), showed participants the opening scene of Orson Welles' film *Touch*

of Evil (Welles & Zugsmith, 1958), which we have discussed at length above. A key medium-specific feature of this scene is that it consists of a single, continuous 3-min shot with no cuts. Thus, it was predicted that this movie clip would provide greater opportunity for viewers' event models to influence their attentional selection. Experiment 1 manipulated viewers' event models through context. Viewers in the context condition watched the entire clip, while those in the no-context condition started watching after the bomb was hidden in the car trunk (i.e., they did not know about it). As one would expect, participants in the context condition were much more likely to predict at the end of the scene that the bomb would explode—showing they had a very different event model for the scene than participants in the no-context condition (who predicted, say, that the two couples would have dinner together). However, there were no significant differences in any of the eye movement variables tested between the two conditions—fixation durations, saccade lengths, or fixations of the car. These results again supported the tyranny of film hypothesis, that bottom-up features of the film scene guided viewers' attention despite large differences in their event models.

Hutson et al. (2017) hypothesized that the *No-context* participants may have continued looking at the car (the *target of attention*) because they considered the driving couple as *protagonists*,³ or more formally, *agents* (Zwaan & Radvansky, 1998). Experiment 2 tested that hypothesis by having the no-context condition participants start watching the film about half way through the clip, when the *walking couple* were at the center of the screen, thus establishing them as the *agents* for those participants. Several seconds later, the walking couple passed by the car, which was parked near an intersection (as shown in Fig. 2). Here, it was hypothesized that the no-context participants, who had never seen the couple in the car before, would not consider them as agents, but instead part of the scene background. Consistent with that hypothesis, for the first 8 s that the car was on-screen, participants in the *new no-context condition* were significantly less likely to fixate the car than the context participants. Hutson et al. (2017) called this the *agent effect*. Namely, entities treated as protagonists (i.e., agents) had higher priority in their current event model, which influenced their attentional selection. Nevertheless, after those first 8 s, the couple in the car became part of the ongoing action, and thus were treated as agents by participants in both conditions, like the walking couple. Thus, the difference in attentional selection between the context and no-context conditions due to the agent effect was very short lived.

To test whether the broader lack of top-down guidance was limited to knowledge-driven top-down processes, or also extended to task-driven top-down processes, in Experiment

³ We thank Jonathan Frome for this insight, many years ago.

2B, Hutson et al. (2017) added a condition in which viewers were given the task of watching the film clip in preparation to draw a detailed map of all the landmarks and their spatial relationships afterwards. Note that one could predict that this task would guide viewers' attention to the *background* of the scene, rather than the main characters, similarly to Lahnakoski et al.'s (2014) "interior decorator" perspective discussed earlier. Indeed, the authors found a clear effect of top-down task-driven guidance of viewers' attention to the film's background elements. Importantly, the map task was at odds with understanding the narrative—namely, the map task required viewers to ignore the main characters at the center of the screen, and instead study the background. Consistent with this claim, participants in the map task condition, who saw the bomb placed in the car, were far less likely to make a prediction about the bomb exploding than those who simply watched for comprehension (i.e., 13% vs. 50%). This strongly suggests that the map task interfered with comprehending the narrative. A follow-up study (Simonson et al., 2020) replicated and extended this result to a set of eight film clips, and two viewing tasks: watching the clip in preparation to do 1) the map task versus 2) a comic drawing task (i.e., watch the clip in preparation to convey the film clip narrative in a four-panel comic). Attentional selection differed between the tasks, with only the map task producing fixation patterns away from the main characters (entities).

The above results raised the question: Why were the effects of the event model on viewers' attentional selection relatively small and short lived (Hutson et al., 2017; Loschky et al., 2015)? As discussed earlier, one of the strongest bottom-up predictors of attentional selection in film is motion (Carmi & Itti, 2006; Mital et al., 2010). This led to the prediction that motion may wash out effects of the viewer's event model on attentional selection while watching film. To test this motion hypothesis, Hutson et al. (2022) used the same *Touch of Evil* scene, but eliminated motion by transforming the film clip into a slideshow (i.e., one frame sampled from the film every 3 s was presented for 3 s to maintain the rough timing of events). All other aspects of the experiment and conditions were the same as Hutson et al. (2017, Exp. 1). When they removed motion from the visual narrative, they found the predicted effects of different event models on viewers' attentional selection (Hutson et al., 2022; see also Pedziwiatr et al., 2023). For example, as shown in Fig. 5, viewers in the context condition were more likely to fixate the car with the bomb, and this difference in attentional selection was fast, occurring *from viewers' first saccade on an image showing the car*. This speed was likely due to the fact that participants already had an event model that prominently included the bomb in the car, which was built earlier during the context manipulation.

Furthermore, when comparing viewer eye movements in the dynamic (film) and slideshow experiments, it was found

that in the context condition viewers' gaze covered a similar proportion of the screen, but for the no-context condition there was a difference (Fig. 6). Specifically, participants in the no-context condition in the slideshow experiment had less clustered gaze (i.e., their fixations covered a larger proportion of the screen) than participants in the film experiment. Thus, the attentional selection of viewers with a *weak* event model was more influenced by *bottom-up stimulus motion*.

It is also important to consider how the *temporal processing constraints* of event models may influence attentional control. Attention has to be optimally deployed by viewers to identify and encode the most important information at any particular moment. This is constrained by the pace of information flow, the viewer's speed of information extraction, and temporal constraints on overt attentional selection via eye movements. During dynamic scenes, the impetus is likely to be placed on the scene dynamics driving this process due to the inability for viewers to stop the flow of information. If viewers are given the opportunity to control the temporal flow of information in a dynamic scene, for instance by using a self-paced slideshow presentation, they have been shown to slow the presentation of frames at event boundaries (Hard et al., 2011). They have similarly been shown to slow the presentation during moments critical to spontaneous theory of mind reasoning, which would constitute a recursive event model—namely, the viewer's event model of a character's event model (Cabañas et al., 2026). For example, when watching silent film scenes in which a character is ignorant of an important piece of narrative information that viewers know, such as the fact that the character's fiancé is embracing her brother rather than a secret lover (known as *dramatic irony scenes*), the discrepancy between the viewer's knowledge and the viewer's representation of the characters' knowledge leads to a slowing down of self-paced film viewing at critical dramatic moments (Cabañas et al., 2026). The difference in the viewer's event models does not change processing overall, just at the moments when the recursive event model is relevant to the current scene. This context effect is further emphasized by changes in gaze behavior during these critical frames. Dwell time on characters and objects critical to the viewer's event model of the scenes' dramatic irony are longer for viewers who perceive the dramatic irony (the recursive event model). This evidence directly supports SPECT's predictions that the current event model can directly influence front-end processing but only if the stimulus affords such control (as the self-paced silent film sequences do), when such control is relevant to the event model.

Taken together, these results testing hypotheses generated by SPECT indicate that the top-down influence of the viewer's event model does guide attentional selection in scenes, but only when bottom-up saliency (e.g., from motion signal) is weak.

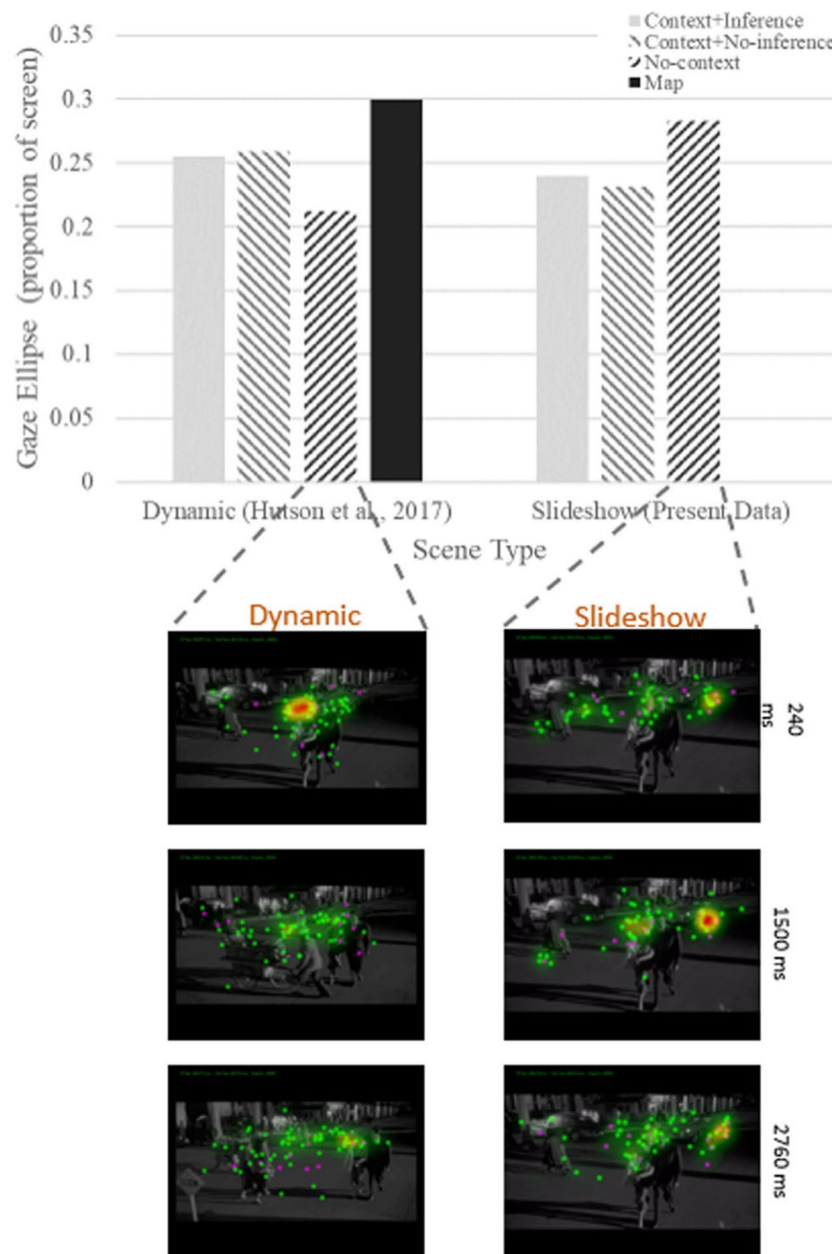


Fig. 6 Above: Gaze ellipse areas for the Hutson et al. (2017; dynamic [film]) versus the Hutson et al. (2022; slideshow). The ellipse area is the proportion of the screen the ellipse covers. Context + inference is solid light gray; context + no inference is a diagonal pattern slanted downwards from left to right; no-context is diagonal pattern slanted upwards from left to right; map task is solid black. Below: The heat-map images illustrate the gaze differences between the No-context

conditions of the dynamic (left) and slideshow (right) presentations. For the slideshow, the image is from 240 ms into the 3-s slide. For the dynamic, the first image is the same frame as the one in the slideshow. The heatmaps show that no-context viewers' gaze in the slideshow condition was more dispersed than in the dynamic (movie) condition. (Figure and caption recreated with permission from Hutson et al., 2022.) (Color figure online)

For the top-down factor of task, it must also be strong and at odds with following the narrative to manifest in differential attentional selection in film. Otherwise, attention to information important to comprehending the narrative will not differ between top-down driven and stimulus-driven elements.

Studies showing that event model processes of mapping and shifting affect attentional Selection

In this section we discuss studies that show that the back-end sub processes of mapping (here, bridging and predictive

inferences) and shifting (event segmentation) affect attentional selection. One common feature of these studies is that the impact of salient features on attentional selection is relatively lower than in studies involving commercially produced movies described above. These studies either used picture stories or videos of actors performing everyday activities as stimuli.

Bridging inferences

Hutson et al. (2018) investigated the role of viewer's event models on attentional selection in visual narratives. To do that, they tested how the *mapping process of bridging inference generation* influenced attentional selection in wordless sequential picture stories (e.g., Mayer, 1975). They identified picture sequences that showed a sequence of causally related events that consisted of a beginning state (Boy starts to run down a hill), a bridging event (Boy trips on a tree root), and end state (Boy falls in a pond). They created coherence gaps by manipulating the presence/absence of the bridging event, which when absent would require inferring that event (e.g., the boy fell because he tripped on something). They verified that participants could infer the missing bridging events using a think aloud task. They then collected eye movement data. Consistent with prior research that shows that the computational effort needed to compute inferences requires time (e.g., Magliano et al., 2016), they found longer viewing times on end-state pictures when the bridging events were absent than when they were present. Importantly, however, they found that this occurred because viewers made approximately two more fixations (11 vs. nine fixations) on the end-state pictures when the bridging-events were absent than when they were present. Moreover, they found that these extra fixations were more likely to go to regions of the end-state pictures that were empirically determined to be more informative for generating the bridging inferences than regions that were less informative. They interpreted these findings as suggesting that the perceived need to generate a bridging inference during the mapping process influences attentional selection processes. Specifically, viewers search for content in the pictures to support the needed bridging inference. Thus, mapping processes in the current event model influence attentional selection.

Predictive eye movements

As mentioned previously, viewers make predictive saccades to objects that they infer actors will interact with before the actor does so (see Gredebäck & Falck-Ytter, 2015, for a review). Wahlheim et al. (2022) tested this hypothesis by showing participants videos of an actor performing everyday activities on two fictional days. Some of the actions repeated on the second day, while others changed. For example, on

the first day viewing, the actor may have opened the closet to retrieve a *bath towel*, but later, on the second day, the actor opened the same closet to retrieve a *hand towel*. Replicating prior work (Eisenberg et al., 2018; Flanagan et al., 2013), Wahlheim et al. (2022) found that viewers made a predictive eye movement to objects before they were touched by the actor on both days. More importantly, the viewers who recalled the action from the first day made a predictive eye movement to the predicted Day 1 object on the second day. This would be based on their stored event model in episodic long-term memory for actions from the first day. As shown in Fig. 3, this is consistent with the bidirectional arrows between the stored event model in episodic long-term memory and the current event model in working memory. Thus, viewers who retrieved information from their stored event model (from Day 1) on the second day, generated predictions based on their stored event model, which influenced their attentional selection.

Event segmentation

There is also evidence that event segmentation affects attentional selection. Specifically, the likelihood of making a predictive eye movement decreases at event boundaries (Eisenberg et al., 2018), consistent with the idea that viewers shift when the next action becomes less predictable. Instead, viewers fixate more on actors' hands at fine event boundaries, to see what they will do (i.e., between stages within a larger goal such as reaching for a piece of clothing while doing laundry). Conversely, viewers fixate on actors' faces more at coarse boundaries (i.e., signifying the change from one goal to another, such as finishing doing laundry and sitting down to read). The latter suggests that viewers are trying to infer the actor's next goal from their facial expressions or gaze direction (T. J. Smith, 2012b). Preliminary research also suggests that event boundaries affect pupil dilation, gaze similarity, and the shift in attentional selection from the ambient to the focal mode. For instance, M. E. Smith, et al. (2024a, 2024b) found that pupil dilation and gaze similarity increases at event boundaries. This is in part because viewers shift into the ambient mode of processing at event boundaries, as they lay the foundation of the new current event model (Eisenberg & Zacks, 2016), like how viewers enter the ambient mode at the onset of a new picture (Pannasch et al., 2008). Viewers then return to the focal mode of processing within events, as it becomes easier to map incoming information onto the current event model (Eisenberg & Zacks, 2016). Together, these results are consistent with hypotheses generated from SPECT that failures in both 1) predictive processing and 2) coherently mapping incoming information onto the current event model cause viewers to shift and lay the foundation of a new current event model. At that moment, viewers cycle back to the ambient

mode of processing, to lay the foundation for the next event model. However, research on this topic is quite limited, and many questions remain about how the active construction and maintenance of an event model, such as that proposed within SPECT, may influence attentional selection at event boundaries. A key question is the degree to which this shift involves task-driven control, or relinquishing control to the visual saliency of the scene.

Developing a systematic approach to studying the effects of event models on attentional selection in film viewing

Most of the research testing SPECT has been in the context of narrative films, which, by definition, are created by highly skilled filmmakers to prioritize visual saliency in attention guidance over other factors, such as the viewer's current event model, which might lead to idiosyncratic viewing. In the context of these studies, we have found robust evidence for the tyranny of film in narrative movies (i.e., large differences in viewers' event models, but small differences in their attentional selection), suggesting that visual saliency has a greater priority than event models on attentional selection in film viewing. This is because in conventional film clips, there is a strong congruence between the high visual saliency of the target and its importance based on the event model. This makes attentional selection of the target automatic and effortless, which allows the viewer to easily continue following the story (T. J. Smith, 2012a; T. J. Smith et al., 2012). In such cases, the target may reasonably be said to "pop-out" for the film viewer, similarly to high saliency targets in visual search, which are found effortlessly (Braun & Julesz, 1998; Wolfe & Horowitz, 2004).

Thus, an important methodological development for testing hypotheses from SPECT has been to develop a systematic approach to investigating the effects of the event model on attentional selection in film clips. Therefore, to isolate the top-down effect of the viewer's event model on their attentional selection, we must find example film clips, or make them, which have a *dissociation between the event model and bottom-up saliency*. Taking this idea, we have developed two criteria for situations in which pre-existing film clips can be used to show the impact of the viewer's event model on their attentional selection while watching them (Chandran et al., 2022, 2024):

Criterion 1

The target of attention should be prioritized by a rich event model. The differences in event models that guide attention to the target need to be measurably distinct in terms of event indices, producing *differential elements*—namely, different event indices (Zwaan & Radvansky, 1998) and causal criteria

(van den Broek, 1990) between two distinct event models created in the *no-context* versus *context* conditions. Participants in the no-context condition view only what we call the *common viewing period* clip. Conversely, participants in the context condition first view a context manipulation clip before watching the common viewing period clip. The differential elements can be identified using a content analysis of viewers' written predictions of what will happen next at the end of the common viewing period clip. When multiple differential elements emphasize the importance of the target of attention, we predict that the viewer's event model can more strongly influence their attentional selection of that target.

To further test Criterion 1, we can ask the participants to show us what is the most important thing at each moment in the common viewing period clip, by having them hover their mouse over it while watching the clip. These mouse positions over time can then be used to generate an aggregated dynamic *importance map*, based on viewers' event models, for each moment in the common viewing period clip. The importance map is broadly similar to a meaning map (Hayes & Henderson, 2022a; Henderson & Hayes, 2017), but importantly differs in that it reflects the judged importance of scene regions *derived from the viewers' event models*, for a dynamically changing narrative or real-world event.

Criterion 1 may also be accompanied by necessary subconditions for meeting it:

- a. There is a *minimum number* of differential elements, or a *minimum subset of critical differential elements*, needed for the viewer's event model to significantly influence their attentional selection.
- b. There is a *minimum threshold of time* required for *laying the foundation* of the event model and for the *key differential elements* in the viewer's event model to significantly influence their attentional selection.

These subconditions also need to be tested.

Criterion 2

The target of attention should have *relatively low visual saliency* (as measured using a state-of-the-art model), so it will not reflexively pull the viewer's attention to the target, despite having a strong event model. The following are some, but not all, of the conditions that can reduce the saliency of the target of attention in a film clip:

- a. The target is not moving, nor does it suddenly appear, since motion and sudden onsets have high visual saliency (Carmi & Itti, 2006; Mital et al., 2010; Theeuwes et al., 1999). If the target is moving, other surrounding things should also be moving, so the target is similar to everything else (e.g., as shown in Fig. 7).



Fig. 7 Above: Importance map for a frame from the closing shot of *Cache* (Berger et al., 2006). Warmer colors correspond to a greater proportion of participants who hovered their mouse over that region. Below: Saliency of the different locations, with warmer colors corresponding to higher visual saliency in the frame from *Cache*, as estimated by the DeepGaze IIE saliency model (Linardos et al., 2021). Note that the target of attention (shown in the red area of interest), which should be deemed important in the target event model, is less salient than many other regions in the scene, particularly the cluster of boys near the exact center of the screen, etc.

- b. The target is not in the center of the screen, since the center bias (Dorr et al., 2010; Kummerer et al., 2017; Tatler, 2007) means that viewers will most frequently attend there (e.g., Fig. 7).
- c. The target is not a person or other animate entity, since people and animals capture attention (Fletcher-Watson et al., 2008; Humphrey & Underwood, 2010; e.g., the car trunk in Fig. 2). If the target is a person, there should be many people (e.g., a crowd), so the target is in competition for attention with the other people in the scene (e.g., Fig. 7).
- d. The target is not in camera focus, since blurred items are less salient (Enns & MacDonald, 2012; J. J. Peterson, 2018), but if it is in focus, *many other things should also be in focus*, so it is similar to everything else.

In sum, the saliency of a target of attention should be no greater than other regions of the scene in a saliency map (e.g., Fig. 7, bottom). Not all four of the above listed sub-conditions (a–d) need to be satisfied to achieve this second

criterion, as long as it satisfies enough of them. At this point, we do not know the relative saliency weights of each of the above four sub-conditions, though we have listed them in a predicted order of precedence for attentional selection based on our above review.

When both criteria are satisfied, the target of attention is important for the viewer's current and stored event models, but it is not visually salient enough to automatically attract the viewer's attention. This will create a dissociation between a behaviorally measured importance map and a computationally measured saliency map. The closing shot of the film *Cache* (Berger et al., 2006) meets both criteria as shown in Fig. 7 and is one film clip out of 10 used in a larger study (Chandran et al., 2024). The target of attention in Fig. 7 is indicated by a red rectangular area of interest in both the top and bottom images of the figure.

To evaluate whether the *Cache* clip meets the first criterion of having *quantifiably distinct event models* between the context and no-context conditions, we used the abovementioned context manipulation. For this clip, there are two targets of attention in the area of interest, two boys. The context condition introduced the two boys as characters (*differential entities*: an older teenage boy and a younger one). The context manipulation further indicated that there was a serious conflict between the boys' fathers, leading the older teenage boy to seek revenge on the younger boy's father (*differential goals*). There were several other differential elements as well (Gernsbacher et al., 1992; van den Broek, 1990; Zwaan et al., 1995).⁴ We then analyzed participants' predictions of what would happen next in terms of whether they included the differential elements, and found substantial differences in the expected direction (Chandran et al., 2022). To

⁴ Differential elements are differential event indices (Zwaan & Radvansky, 1998), causal criteria (van den Broek, 1990), and emotions of agents (Gernsbacher et al., 1992) that enable predictions unique to the participants who know the context of the scene. We have identified six primary differential elements based on the event indexing model, the causal inference maker, and the structure building framework:

1. Entities are agents/objects/highly delimited places of interest which are the designated targets of attention.
2. Goals are the purpose or motivations of an agent, typically the target.
3. Temporal operativity indicates whether the event that is leading up to the target prediction is still ongoing.
4. Spatial proximity indicates the closeness of entities in the scene to other entities that can lead to targeted predictions.
5. Enabling Features are any feature (other than spatial proximity and temporal operativity) in the story that enable targeted predictions.
6. Emotions indicate any emotion(s) of the agent(s) relevant for the target predictions.

provide converging evidence of the differential elements, translated to image locations over time, we used the abovementioned importance judgment task using mouse hovering (Chandran et al., 2024). This mouse data were used to create a dynamic importance map over the entire common viewing period, for both the context and no-context conditions. The top of Fig. 7 is an illustrative moment in the shot used in the study. It shows the importance map from the context condition for 30 participants. This clearly shows that for that moment in the clip, the targets of attention in the area of interest were judged the most important things in the scene, and this moment illustrates the general trend for the clip. In the no-context condition, no participants hovered their mouse over the dynamic area of interest illustrated in Fig. 7, but instead distributed their importance judgments across the image, with a strong bias to the center of the screen. All the other nine out of the 10 clips similarly showed substantial differences between context and no-context conditions (Chandran et al., 2024).

The bottom half of Fig. 7 shows a state-of-the-art saliency map of the same moment in the shot (Linardos et al., 2021). The two boys can be seen to have no greater saliency than the many other people in the image. Thus, this meets the second criterion for showing an effect of the viewer's event model on their attentional selection in this film clip.

Follow-up experiments using the 10 film clips have measured viewer's eye movements in the context versus no-context conditions (Chandran, 2026). This is a definitive existence proof test for whether different event models can produce different attentional selection while viewing commercially produced film clips. Having found such differences in attentional selection based on the context manipulation, it more clearly defines the conditions needed to overcome what we have called the tyranny of film.

Note, however, that the importance judgment task used to make the importance maps is metacognitive in nature—it asks participants to communicate a conscious judgment with their mouse. Such judgments are task driven and likely involve executive processes, such as inhibiting hovering over things deemed irrelevant to one's stored event model. Conversely, we did not expect viewers' eye movements to be so cleanly differentiated by the context manipulation, as viewers' only task was to comprehend the film clip in preparation to make a prediction of what will happen next. Furthermore, in the no-context condition, unlike mouse hovering in the importance judgment task, viewers' eye movements were more likely to be influenced by random fluctuations in their impoverished event model, and thus by random chance they might fixate on the targets of attention as they search for meaning. Nevertheless, the video clip from *Cache* presents a strong case of meeting the two criteria for showing the

effect of viewers' event models on their attentional selection while watching film. We preregistered a study to test the hypothesis that viewers' eye movements differ as a function of their event models when both criteria are met (Chandran et al., 2026), and we have found strong support for it (Chandran, 2026).

Discussion

In this paper we have discussed refinements to SPECT regarding how the current event model influences attentional selection. In this discussion, we specify the factors that influence attentional selection, underlying mechanisms by which the event model influences attentional selection, and discuss some future directions for the development and testing of SPECT.

The competition between visual saliency versus task- and knowledge-driven top-down processes for attentional selection in scenes

Our review of the relevant literature has pointed to a hierarchy of strength of influence on attentional selection in dynamic scenes and visual narratives. There are several factors that interact in determining this hierarchy. Those factors include 1) top-down and bottom-up influences, and the two subtypes of top-down influences (task-driven vs. knowledge-driven); 2) the type of stimuli (i.e., still images, vs. dynamic images, vs. film). In most cases, these influential factors are not in conflict with each other and hence could be contributing equally to attentional selection. For example, if a viewer were given a task to watch and understand a well-directed Hollywood-style film, the stimulus features are designed to drive attention to the same event model-driven area in the scene, which also matches their current goals of the task.

Overall, task-driven top-down influences seem to be at the top of the hierarchy; whether in still images (Einhauser et al., 2008; Henderson et al., 2007; Underwood & Foulsham, 2006) or in filmed events, if their composition provides the time and space for scene exploration (Hutson et al., 2017; Simonson et al., 2021; T. J. Smith & Mital, 2013). Dynamic scenes that are composed to focus attention on key features and limit individual exploration, such as movies or video clips, have been shown to allow less expression of the task influences (Taya et al., 2012) and the event model (Hutson et al., 2017; Loschky et al., 2015). Visual saliency seems to be next in the hierarchy, being more influential than knowledge-driven top-down processes, particularly in filmed events (Hutson et al., 2017; Loschky et al., 2015), and with still images in some

cases (Kaakinen et al., 2011; Underwood & Foulsham, 2006), but not in others (Hutson et al., 2022). Knowledge-driven top-down influences, such as schemas or the viewer's event model, seem to be last in the hierarchy, but can win the competition for attentional selection if 1) the visual saliency is low enough, such as in still images (Hutson et al., 2018, 2022) or 2) for brief moments in dynamic images when motion is essentially absent (Hutson et al., 2017, Exp. 2 A, 2022; Loschky et al., 2015), or 3) if the event model is specifically relevant (Cabañas et al., 2026). As such, the literature review clearly indicates that the extent to which factors lower in the hierarchy can affect attentional selection is contingent upon how salient visual features vary across and within visual media.

There are various ways one can talk about the hierarchy. It can be the proportion of variance in fixation locations that are explained by each factor, which is the basis of our comparisons above. It can also be the time course of attentional selection. Saliency and knowledge-driven factors, including the event model, are fast, while task-driven executive control is slow (for a review, see Theeuwes, 2018). Thus, in the absence of a task, both saliency (Carmi & Itti, 2006; Mital et al., 2010) and knowledge-driven selection (Hutson et al., 2022; T. J. Smith & Martin-Portugues Santacreu, 2017; Valuch et al., 2017) can rapidly influence the first saccade in a scene, with semantics having more of an influence after roughly 2,000 ms (Carmi & Itti, 2006). But in the long run, task-driven attentional selection tends to overcome the influences of saliency (Einhauser et al., 2008; Henderson et al., 2007; Hutson et al., 2017; Simonson et al., 2021; T. J. Smith & Mital, 2013; Underwood & Foulsham, 2006).

While our review points to ranking the influences that impact attentional selection in this way, future research is needed to quantify the relative strengths of influence of each of these factors. The studies that tested SPECT and informed this proposed ranking thus far have characterized the influences of the factors affecting attentional selection hierarchy in binary terms (i.e., a rich event model, or motion, are present vs. absent). The work we have done with visual narratives has until recently not used saliency models to measure visual saliency in order to quantify its strength, relative to that of the knowledge-driven top-down influence of event models. Instead, those studies have simply held the visual stimulus constant while varying the strength of the event model, and determined whether the event model produced a significant effect on viewers' gaze. Thus, future studies of the effects of the knowledge-driven influence of event models on viewers' gaze, versus that of visual saliency, need to quantify both the strength of the event model (e.g., varied in terms of event indices, and measured in terms of importance judgments) and visual saliency (i.e., measured using saliency models) in order to quantitatively compare the relative strengths of each (e.g., see Chandran, 2026).

Mechanisms for the current event model to influence attentional selection in scenes

In this article, we introduce three new mechanisms within SPECT by which the current event model can influence attentional selection: 1) the ambient-to-focal shift as event models transition from laying the foundation to mapping incoming information; 2) the working memory contents of the current event model during both laying the foundation and mapping; and 3) visual search-like attention deployments during mapping or shifting to create a new event model. Below, we discuss key topics that need to be investigated to further develop and test these mechanisms in SPECT.

A first mechanism through which the event model influences attentional selection is via the routine of the ambient-to-focal shift of eye movements during laying the foundation and mapping for the current event model (Eisenberg & Zacks, 2016; Pannasch, 2014; T. J. Smith, 2012b). This routine seems to be triggered by shifting to create a new current event model (Eisenberg & Zacks, 2016; Pannasch, 2014; T. J. Smith, 2012b). Shifting to the ambient mode of processing at event boundaries may facilitate laying the foundation because key global information is quickly extracted at the start of a new event. Reverting to the focal model of processing within events may facilitate mapping detailed information onto the event model as it is extracted. An important question, however, is the degree to which the ambient-to-focal shift is uniquely tailored to each new event. Is each saccade sent to the ideal location for extracting the information needed (Najemnik & Geisler, 2005) to fill out the new current event model? If so, this would implicate the event model's influence from its nascent beginning, perhaps based on the stored event model, and associated schemas or scripts. Alternatively, is the ambient-to-focal shift more of a standard routine that can be followed automatically, much like demonstrations of automated eye movements in the reading literature (e.g., "Z-reading"; Vitu et al., 1995)? Such reading-like eye movements have also been successfully simulated by a computational model of eye movements based on the architecture of the midbrain structure, the superior colliculus (Vitu et al., 2021). Furthermore, the same model, when provided with a priority map indicating a task-driven goal, also successfully simulated eye movements during categorical visual search (e.g., search for a "clock" in a picture; Adeli & Zelinsky, 2018). Further research could tease apart these two possibilities by assessing the degree to which the ambient-to-focal shift is differentially affected by the specifics of the viewer's current and stored event models (e.g., their goal structure).

A second mechanism through which the current event model can influence attentional selection is through the fact that it is created, processed, and stored in working memory.

We previously mentioned the well-attested finding that the contents of working memory (e.g., objects or object features) are privileged for visual attention (Bahle et al., 2018; Olivers et al., 2006; Soto et al., 2005). Critically, the contents of working memory are maintained in a heightened state of activation (Cowan, 1993), and event indices in the event model in working memory are highly prioritized in comprehension and memory (Radvansky & Zacks, 2014; Zwaan & Radvansky, 1998). The biased competition model proposes that working memory retention requires that the perceptual representation of the remembered information be maintained in a sustained level of activation. Subsequent processing that shares the preactivated features is facilitated, and this primes spatial selection toward features that are similar to the information held in working memory. Such a mechanism is consistent with the previously discussed agent effect from Hutson et al. (2017). Namely, viewers are more likely to fixate entities that have been introduced earlier in an event (i.e., agents) than entities that have not been previously introduced. This is because viewers have a representation of the agents as part of the ongoing event model in working memory. The agent effect is one example of how situational indices of event models may affect attentional selection. This raises a key unanswered question: Is this influence of event indices on attentional selection largely passive in nature, or does it require effort? There has been a related debate concerning event models for narrative texts as to the relative roles of passive, memory-based processes (e.g., spreading activation) and effortful, constructive processes (e.g., explanation; Graesser et al., 1994; Long & Lea, 2005; Myers & O'Brien, 1998). The contemporary perspective is that much of the heavy lifting in constructing event models in working memory involves passive mechanisms (Kintsch, 1998), but that effortful constructive mechanisms are important when there are difficulties in comprehension (Chi et al., 1989). Further targeted research on this topic needs to be conducted.

A third mechanism by which the event model influences attentional selection is by engaging in visual search during mapping and shifting to create a new event model (Hutson et al., 2018). This should typically occur when repair of the event model is needed—specifically, difficulty in mapping incoming information to the current event model. This comes from the aforementioned evidence that viewers search for inferentially relevant information during the mapping process, when they need to generate a bridging inference to maintain coherence of the in-coming information with their current event model. This would be a case of an effortful constructive process as described above. Note also that such a search implies goal setting, an executive process, but how often such search rises to the level of conscious awareness is an open question. We speculate that any time the viewer's current event model requires interrogation of a scene for

specific sorts of information, either when needing to lay the foundation for a new event model, or when mapping information to the current event model, categorical visual search processes may be engaged (Zelinsky et al., 2020). Such searches may even produce anticipatory/predictive saccades when the viewer's event model leads to a prediction of what an entity will do next (Eisenberg et al., 2018; Wahlheim et al., 2022). An interesting theoretical question in this case is why the viewer is better off making an anticipatory saccade to the target of the entity's inferred goal, rather than waiting to find out what the entity does first. One possibility is that predictions, while computationally expensive to generate, function to stabilize the current event model, facilitating information extraction (M. E. Smith & Loschky, 2019; Zacks et al., 2007). Actions in our environment are not random, such that the probability of one action often depends on another. An event model that can capitalize on the regularities of the environment by generating predictions for what will happen next may ease the computational effort associated with information extraction (M. E. Smith & Loschky, 2019). Note that this would seem to be the most active influence of the event model on attentional selection.

These proposed mechanisms point to fertile ground for further work to refine SPECT and elaborate the time course of event model influences on front-end processes. The above mechanisms also provide new hypotheses about how the current event model influences attentional selection, and we have suggested some new ways to test them.

Conclusion

A key question in vision research is, what influences viewers' moment-to-moment attentional selection? We highlight a previously largely unexplored top-down influence on attentional selection—namely, the viewer's event model—their understanding of what is happening now. Borrowing from established fields of literature spanning scene perception, event cognition, and text and discourse comprehension, we describe the event model's mechanisms for influencing viewers' attentional selection within scenes. This synthesizes scene perception theories (in the front-end) with theories of viewers' event models (in the back end). SPECT generates novel testable hypotheses concerning how viewers' event models influence their attentional selection in dynamic real-world scenes. Importantly, in this update of SPECT, we incorporate the biased competition mechanism of attentional selection, which takes both bottom-up and top-down influences. We also incorporate the ambient-to-focal shift, the fact that the event model resides in working memory (which is well-known to bias attentional selection), and visual search as mechanisms to explain how salient features and the event model compete for selection. We introduce criteria necessary to empirically

test the event model hypothesis in the context of narrative film. As such, we believe that SPECT provides a testable and quantifiable theoretical framework for exploring attentional selection in the context of dynamic events. Furthermore, we encourage scene perception researchers to use SPECT, or to propose competing theories, to identify when viewers' event models affect their attentional selection.

Importantly with respect to this special issue, we see the contributions of SPECT to understanding the event model's influences on attentional selection as being broadly akin to those made by Mary Peterson regarding top-down effects on object perception. Among other things, she identified how object recognition processes produce top-down influences on figure-ground segmentation, which were previously ignored in the vision sciences because they were thought to be impossible. We see SPECT as similarly providing a theoretical framework that provides vision sciences with a basis for how and when event models affect attentional selection. This theoretical space has largely been ignored till now, because the constituent research areas encompassed by SPECT, namely scene perception, event cognition, and text comprehension, have previously remained siloed. Just as Mary Peterson has shown that "impossible" influences in object perception do in fact occur, we think event models can influence attentional selection in dynamic scenes in ways that can be studied and understood.

Appendix: Changes to the SPECT box & arrow model V2 figure from the previous version (Loschky et al., 2018, 2020):

1. Substantive Changes:

- a. In "Stimulus":
- i. "Saliency" has been changed to "Saliency-engendering features." This is to reflect the fact that saliency is not inherently a stimulus property. Rather, stimulus features engender the perceptual process of determining saliency.
- b. In "Back-End"
- i. Downward arrows from "Semantic Memory (Prior Knowledge)" and "Episodic Memory (Stored Event Model)" only touch the "Current Event Model (Episodic Buffer/LT WM)" rather than connecting to specific processes in the "Current Event Model". This avoids making strong claims about how Semantic Memory and Episodic Memory influence the sub-processes of "Laying the Foundation," "Mapping incoming info," and "Shifting."

2. Changes to ease cognitive processing:

- a. The entire figure has been rotated 180 degrees so that "top-down" (feedback) influences are at the top,

and "bottom-up" (feed-forward) influences are at the bottom. This is to make the figure conceptually easier to process.

3. Changes only to add clarity, without changing substance:

- a. In "Stimulus":
 - i. Arrows from "Saliency-engendering features" and "Medium-specific features" to both "broad" and "narrow" information extraction ("Info extraction") are explicitly shown. This adds clarity to the previous version of the figure.
 - ii. Arrows from "Saliency-engendering features" and "Medium-specific features" to "Attentional Selection" are explicitly shown. This adds greater clarity to the previous version of the figure.
- b. In "Front-End":
 - i. The Front-End now has a separate dashed line surrounding it, similar to that for the Back-End, and for the Stimulus. Previously, there was a dashed line around both the Front-End and Back-End, indicating that they are separate from the Stimulus. That dashed line has been removed. These changes simply clarify that the Stimulus, Front-End, and Back-End are each considered separate.
 - ii. Boxes around the eye icon, which encompass both "Info extraction" and "Attentional Selection" are now labeled "Fixation" to clarify that both processes are occurring on each fixation. This adds clarity to the previous version of the figure.
- c. In "Back-End":
 - i. Downward arrows from the Back-End "Current Event Model (Episodic Buffer/LT WM)" to "Front-End" "Info Extraction" and "Attentional Selection" touch those processes. This is to add visual clarity.

1. However, the lack of such a downward arrow from the "Current Event Model..." to "Info Extraction Narrow" does not imply that this relationship does not exist. The lack of this arrow is only to avoid visual clutter. It is assumed that the "Current Event Model..." equally influences both "Broad" and "Narrow" information extraction.

Acknowledgments We sincerely thank Neil Cohn for encouraging us to develop SPECT. We are grateful to Gregory Zelinsky and Hossein Adeli for helpful discussions of the biased competition model. We also thank Jonathan Frome for pointing us to the agent effect. And we thank Tom Foulsham for helpful discussion of task effects that could overcome the influence of visual saliency while watching film.

Author Contributions Writing initial draft: L.L., J.M.; writing contributions on subsequent drafts: M.S., P.C., J.H.; conceptual development of theoretical framework: L.L., J.M., M.S., P.C., T.S.; feedback on subsequent drafts: L.L., J.M., M.S., P.C., T.S.

Funding Not applicable

Data availability Not applicable

Code Availability Not applicable

Declarations

Ethics Approval Not applicable

Consent to Participate Not applicable

Consent for Publication All authors consent to the publication of this article.

Conflicts of Interest/Competing Interests The authors declare that they have no conflicts of interest relevant to the content of this manuscript.

Open Practices Statement All data discussed in this paper is based on summaries of existing published results and analyses.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Adeli, H., Vitu, F., & Zelinsky, G. J. (2017). A model of the superior colliculus predicts fixation locations during scene viewing and visual search. *Journal of Neuroscience*, 37(6), 1453–1467. <https://doi.org/10.1523/jneurosci.0825-16.2016>
- Adeli, H., & Zelinsky, G. J. (2018). *Deep-BCN: Deep networks meet biased competition to create a brain-inspired model of attention control*. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT.
- Altmann, G. T. M., & Ekves, Z. (2019). Events as intersecting object histories: A new theory of event representation. *Psychological Review*, 126(6), 817–840. <https://doi.org/10.1037/rev0000154>
- Anderson, R. C., & Pichert, J. W. (1978). Recall of previously unrecalled information following a shift in perspective. *Journal of Verbal Learning and Verbal Behavior*, 17(1), 1–12. [https://doi.org/10.1016/s0022-5371\(78\)90485-1](https://doi.org/10.1016/s0022-5371(78)90485-1)
- Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, 4(11), 417–423. [https://doi.org/10.1016/S1364-6613\(00\)01538-2](https://doi.org/10.1016/S1364-6613(00)01538-2)
- Bahle, B., Beck, V. M., & Hollingworth, A. (2018). The architecture of interaction between visual working memory and visual attention. *Journal of Experimental Psychology: Human Perception and Performance*, 44(7), 992–1011. <https://doi.org/10.1037/xhp0000509>
- Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., & Norman, K. A. (2017). Discovering event structure in continuous narrative perception and memory. *Neuron*, 95(3), 709–721.e705. <https://doi.org/10.1016/j.neuron.2017.06.041>
- Baluch, F., & Itti, L. (2011). Mechanisms of top-down attention. *Trends in Neurosciences*, 34(4), 210–224. <https://doi.org/10.1016/j.tins.2011.02.003>
- Barsalou, L. W. (1992). Frames, concepts, and conceptual fields. In A. Lehrer & E. F. Kittay (Eds.), *Frames, fields, and contrasts: New essays in semantic and lexical organization* (pp. 21–74). Erlbaum.
- Ben-Yakov, A., & Dudai, Y. (2011). Constructing realistic engrams: Poststimulus activity of hippocampus and dorsal striatum predicts subsequent episodic memory. *Journal of Neuroscience*, 31(24), 9032–9042.
- Berger, C., Haneke, M., Menegoz, M., Heiduschka, V., Auteuil, D., Binoche, J., & Benichou, M. (2006). *Cache* [Motion picture]. Sony Pictures Home Entertainment.
- Bezdek, M. A., & Gerrig, R. J. (2017). When narrative transportation narrows attention: Changes in attentional focus during suspenseful film viewing. *Media Psychology*, 20(1), 60–89. <https://doi.org/10.1080/15213269.2015.1121830>
- Bezdek, M. A., Gerrig, R. J., Wenzel, W. G., Shin, J., Revill, K. P., & Schumacher, E. H. (2015). Neural evidence that suspense narrows attentional focus. *Neuroscience*, 303, 338–345.
- Bichot, N. P., Rossi, A. F., & Desimone, R. (2005). Parallel and serial neural mechanisms for visual search in macaque area V4. *Science*, 308(5721), 529–534. <https://doi.org/10.1126/science.1109676>
- Bordwell, D., & Thompson, K. (2003). *Film Art: An Introduction*. McGraw-Hill.
- Borji, A., & Itti, L. (2014). Defending Yarbus: Eye movements reveal observers' task. *Journal of Vision*, 14(3), 29–29.
- Braun, J., & Julesz, B. (1998). Withdrawing attention at little or no cost: Detection and discrimination tasks. *Perception & Psychophysics*, 60(1), 1–23. <https://doi.org/10.3758/BF03211915>
- Brich, I. R., Papenmeier, F., Huff, M., & Merkt, M. (2024). Construction or updating? Event model processes during visual narrative comprehension. *Psychonomic Bulletin & Review*. <https://doi.org/10.3758/s13423-023-02424-w>
- Bruce, N. D. B., & Tsotsos, J. K. (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision*, 9(3), 5–5. <https://doi.org/10.1167/9.3.5>
- Buffalo, E. A., Fries, P., Landman, R., Liang, H., & Desimone, R. (2010). A backward progression of attentional effects in the ventral stream. *Proceedings of the National Academy of Sciences*, 107(1), 361–365.
- Cabañas, C., Senju, A., & Smith, T. J. (2026). *Seeing other's minds in your eyes: Viewers actively extract and maintain theory of mind representations in dramatic irony scenes*. Manuscript under review.
- Calvo, M. G., & Lang, P. J. (2005). Parafoveal semantic processing of emotional visual scenes. *Journal of Experimental Psychology: Human Perception and Performance*, 31(3), 502–519.
- Calvo, M. G., Meseguer, E., & Carreiras, M. (2001). Inferences about predictable events: Eye movements during reading.

- Psychological Research*, 65(3), 158–169. <https://doi.org/10.1007/s004260000050>
- Carlson-Radvansky, L. A., & Irwin, D. E. (1995). Memory for structural information across eye movements. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 21(6), 1441–1458.
- Carmi, R., & Itti, L. (2006). Visual causes versus correlates of attentional selection in dynamic scenes. *Vision Research*, 46(26), 4333–4345. <https://doi.org/10.1016/j.visres.2006.08.019>
- Cerf, M., Frady, E. P., & Koch, C. (2009). Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of Vision*, 9(12), 1–15. <https://doi.org/10.1167/9.12.10>
- Chandran, P. (2026). *Investigating the conditions under which movie viewers' event models influence their attention* [Doctoral dissertation, Kansas State University]. Manuscript in preparation.
- Chandran, P., Fitzgibbons, C., Hubbell, I., Nuwamanya, J., Pagen, J., Schultz, S., . . . Loschky, L. C. (2022). *Creating a generalizable process for manipulating participants' understanding of film clips*. Society for Cognitive Studies of the Moving Image (SCSMI) 2022, Gandia, Spain.
- Chandran, P., Magliano, J. P., & Loschky, L. C. (2026, January 13). *Investigating the conditions under which movie viewers' event models influence their attention*. osf.io/3vqtm
- Chandran, P., Smith, A., Lofing, A., Wilson, L., Wallace, B., Mass, W., . . . Loschky, L. C. (2024, May). *Preparations to break the Tyranny of Film: Designing stimuli that change comprehension and moment-to-moment content importance in video* The 24th Annual Meeting of the Vision Sciences Society, St. Pete Beach, FL, USA.
- Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, 13(2), 145–182. https://doi.org/10.1207/s15516709cog1302_1
- Cichy, R. M., Khosla, A., Pantazis, D., & Oliva, A. (2017). Dynamics of scene representations in the human brain revealed by magnetoencephalography and deep neural networks. *NeuroImage*, 153, 346–358.
- Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, 6, Article 27755. <https://doi.org/10.1038/srep27755>
- Coco, M. I., Keller, F., & Malcolm, G. L. (2016). Anticipation in real-world scenes: The role of visual context and visual memory. *Cognitive Science*, 40(8), 1995–2024.
- Coco, M. I., Nuthmann, A., & Dimigen, O. (2020). Fixation-related brain potentials during semantic integration of object–scene information. *Journal of Cognitive Neuroscience*, 32(4), 571–589. <https://doi.org/10.1162/jocn>
- Cohn, N. (2013). *The visual language of comics: Introduction to the structure and cognition of sequential images*. Bloomsbury Academic.
- Cohn, N. (2020). *Who understands comics?: Questioning the universality of visual language comprehension*. Bloomsbury Publishing.
- Cohn, N., & Foulsham, T. (2020). Zooming in on the cognitive neuroscience of visual narrative. *Brain and Cognition*, 146, Article 105634. <https://doi.org/10.1016/j.bandc.2020.105634>
- Cornelissen, T. H. W., & Vö, M. L. H. (2017). Stuck on semantics: Processing of irrelevant object–scene inconsistencies modulates ongoing gaze behavior. *Attention, Perception, & Psychophysics*, 79(1), 154–168.
- Cowan, N. (1993). Activation, attention, and short term memory. *Memory & Cognition*, 21(2), 162–167.
- Crundall, D. E., Underwood, G., & Chapman, P. R. (2002). Attending to the peripheral world while driving. *Applied Cognitive Psychology*, 16(4), 459–475. <https://doi.org/10.1002/acp.806>
- Curcio, C. A., Sloan, K. R., Kalina, R. E., & Hendrickson, A. E. (1990). Human photoreceptor topography. *Journal of Comparative Neurology*, 292(4), 497–523. <https://doi.org/10.1002/cne.902920402>
- Cutting, J. E. (2021). *Movies on our minds: The evolution of cinematic engagement*. Oxford University Press.
- Cutting, J. E., & Iricinschi, C. (2015). Re-presentations of space in Hollywood movies: An event-indexing analysis. *Cognitive Science*, 39(2), 434–456. <https://doi.org/10.1111/cogs.12151>
- Daniel, P., & Whitteridge, D. (1961). The representation of the visual field on the cerebral cortex in monkeys. *The Journal of Physiology*, 159(2), 203–221.
- De Graef, P., Christiaens, D., & D'Ydewalle, G. (1990). Perceptual effects of scene context on object identification. *Psychological Research*, 52(4), 317–329. <https://doi.org/10.1007/BF00868064>
- DeAngelus, M., & Pelz, J. (2009). Top-down control of eye movements: Yarbus revisited. *Visual Cognition*, 17(6), 790–811. <https://doi.org/10.1080/13506280902793843>
- Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, 18(1), 193.
- Diamond, A. (2013). Executive functions. *Annual Review of Psychology*, 64, 135–168. <https://doi.org/10.1146/annurev-psych-113011-143750>
- Dobel, C., Gummior, H., Bölte, J., & Zwitserlood, P. (2007). Describing scenes hardly seen. *Acta Psychologica*, 125(2), 129–143.
- Dorr, M., & Bex, P. J. (2013). Peri-saccadic natural vision. *Journal of Neuroscience*, 33(3), 1211–1217. <https://doi.org/10.1523/jneurosci.4344-12.2013>
- Dorr, M., Martinetz, T., Gegenfurtner, K. R., & Barth, E. (2010). Variability of eye movements when viewing dynamic natural scenes. *Journal of Vision*, 10(10), Article 28. <https://doi.org/10.1167/10.10.28>
- Eckstein, M. P., Drescher, B. A., & Shimozaaki, S. S. (2006). Attentional cues in real scenes, saccadic targeting, and Bayesian priors. *Psychological Science*, 17(11), 973–980. <https://doi.org/10.1111/j.1467-9280.2006.01815.x>
- Ehrlich, K., & Rayner, K. (1983). Pronoun assignment and semantic integration during reading: Eye movements and immediacy of processing. *Journal of Verbal Learning and Verbal Behavior*, 22(1), 75–87.
- Einhauser, W., Rutishauser, U., & Koch, C. (2008). Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. *Journal of Vision*, 8(2), 2–2. <https://doi.org/10.1167/8.2.2>
- Eisenberg, M. L., & Zacks, J. M. (2016). Ambient and focal visual processing of naturalistic activity. *Journal of Vision*, 16(2), Article 5. <https://doi.org/10.1167/16.2.5>
- Eisenberg, M. L., Zacks, J. M., & Flores, S. (2018). Dynamic prediction during perception of everyday events. *Cognitive Research: Principles and Implications*, 3(1), Article 53. <https://doi.org/10.1186/s41235-018-0146-z>
- Elman, J. L. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive Science*, 33(4), 547–582. <https://doi.org/10.1111/j.1551-6709.2009.01023.x>
- Enns, J. T., & MacDonald, S. C. (2012). The role of clarity and blur in guiding visual attention in photographs. *Journal of Experimental Psychology: Human Perception and Performance*, 39(2), 568–578. <https://doi.org/10.1037/a0029877>
- Epstein, R. A., & Baker, C. I. (2019). Scene perception in the human brain. *Annual Review of Vision Science*, 5(1), 373–397. <https://doi.org/10.1146/annurev-vision-091718-014809>
- Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review*, 102(2), Apr 1995, 1211–1245.

- Evdokimidis, I., Smyrnis, N., Constantinidis, T., Stefanis, N., Avramopoulos, D., Paximadis, C., & Stefanis, C. (2002). The antisaccade task in a sample of 2,006 young men. *Experimental Brain Research*, 147(1), 45–52. <https://doi.org/10.1007/s00221-002-1208-4>
- Fecteau, J. H., & Munoz, D. P. (2006). Saliency, relevance, and firing: A priority map for target selection. *Trends in Cognitive Sciences*, 10(8), 382–390. <https://doi.org/10.1016/j.tics.2006.06.011>
- Fei-Fei, L., Iyer, A., Koch, C., & Perona, P. (2007). What do we perceive in a glance of a real-world scene? *Journal of Vision*, 7(1), Article 10. <https://doi.org/10.1167/7.1.10>
- Findlay, J., & Walker, R. (1999). A model of saccade generation based on parallel processing and competitive inhibition. *Behavioral and Brain Sciences*, 22(4), 661–721.
- Flanagan, J. R., Rotman, G., Reichelt, A. F., & Johansson, R. S. (2013). The role of observers' gaze behaviour when watching object manipulation tasks: Predicting and evaluating the consequences of action. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1628), Article 20130063. <https://doi.org/10.1098/rstb.2013.0063>
- Fletcher-Watson, S., Findlay, J. M., Leekam, S. R., & Benson, V. (2008). Rapid detection of person information in a naturalistic scene. *Perception*, 37(4), 571–583.
- Foulsham, T., & Underwood, G. (2007). How does the purpose of inspection influence the potency of visual saliency in scene perception? *Perception*, 36(8), 1123–1138.
- Foulsham, T., Wybrow, D., & Cohn, N. (2016). Reading without words: Eye movements in the comprehension of comic strips. *Applied Cognitive Psychology*, 30(4), 566–579. <https://doi.org/10.1002/acp.3229>
- Gareze, L., & Findlay, J. M. (2007). Absence of scene context effects in object detection and eye gaze capture. In R. P. G. van Gompel, M. H. Fischer, W. S. Murray, & R. L. Hill (Eds.), *Eye movements: A window on mind and brain* (pp. 537–562). Elsevier. <https://doi.org/10.1016/B978-008044980-7/50031-8>
- Gernsbacher, M. A. (1983). *Memory for surface information in non-verbal stories: Parallels and insights to language processes* [Doctoral dissertation, University of Texas at Austin].
- Gernsbacher, M. A. (1985). Surface information loss in comprehension. *Cognitive Psychology*, 17(3), 324–363.
- Gernsbacher, M. A. (1990). *Language comprehension as structure building* (Vol. 11). Erlbaum.
- Gernsbacher, M. A., Goldsmith, H. H., & Robertson, R. R. (1992). Do readers mentally represent characters' emotional states? *Cognition & Emotion*, 6(2), 89–111.
- Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, 101(3), 371–395.
- Gredebäck, G., & Falck-Ytter, T. (2015). Eye movements during action observation. *Perspectives on Psychological Science*, 10(5), 591–598. <https://doi.org/10.1177/1745691615589103>
- Greene, M. R., Liu, T., & Wolfe, J. M. (2012). Reconsidering Yarbus: A failure to predict observers' task from eye movement patterns. *Vision Research*, 62(0), 1–8. <https://doi.org/10.1016/j.visres.2012.03.019>
- Greene, M. R., & Oliva, A. (2009). The briefest of glances: The time course of natural scene understanding. *Psychological Science*, 20(4), 464–472.
- Haberlandt, K. F., & Graesser, A. C. (1985). Component processes in text comprehension and some of their interactions. *Journal of Experimental Psychology: General*, 114(3), 357–374.
- Hafri, A., Papafragou, A., & Trueswell, J. C. (2013). Getting the gist of events: Recognition of two-participant actions from brief displays. *Journal of Experimental Psychology: General*, 142(3), 880–905.
- Hallett, P. E. (1978). Primary and secondary saccades to goals defined by instructions. *Vision Research*, 18(10), 1279–1296. [https://doi.org/10.1016/0042-6989\(78\)90218-3](https://doi.org/10.1016/0042-6989(78)90218-3)
- Hamker, F. H. (2004). The reentry hypothesis: The putative interaction of the frontal eye field, ventrolateral prefrontal cortex, and areas V4, IT for attention and eye movement. *Cerebral Cortex*, 15(4), 431–447. <https://doi.org/10.1093/cercor/bbh146>
- Han, S. W., & Kim, M.-S. (2009). Do the contents of working memory capture attention? Yes, but cognitive control matters. *Journal of Experimental Psychology: Human Perception and Performance*, 35(5), Article 1292.
- Hard, B., Recchia, G., & Tversky, B. (2011). The shape of action. *Journal of Experimental Psychology: General*, 140(4), Article 586.
- Hare, M., Jones, M., Thomson, C., Kelly, S., & McRae, K. (2009). Activating event knowledge. *Cognition*, 111(2), 151–167. <https://doi.org/10.1016/j.cognition.2009.01.009>
- Hayes, T. R., & Henderson, J. M. (2022). Meaning maps detect the removal of local semantic scene content but deep saliency models do not. *Attention, Perception & Psychophysics*, 84(3), 647–654. <https://doi.org/10.3758/s13414-021-02395-x>
- Hayes, T. R., & Henderson, J. M. (2022). Scene inversion reveals distinct patterns of attention to semantically interpreted and uninterpreted features. *Cognition*, 229, Article 105231. <https://doi.org/10.1016/j.cognition.2022.105231>
- Henderson, J. M., Brockmole, J. R., Castelano, M. S., & Mack, M. L. (2007). Visual saliency does not account for eye movements during visual search in real-world scenes. In R. v. Gompel, M. Fischer, W. Murray, & R. W. Hill (Eds.), *Eye movements: A window on mind and brain* (pp. 537–562). Elsevier.
- Henderson, J. M., & Hayes, T. R. (2017). Meaning-based guidance of attention in scenes as revealed by meaning maps. *Nature Human Behaviour*, 1(10), 743–747.
- Henderson, J. M., & Hayes, T. R. (2018). Meaning guides attention in real-world scene images: Evidence from eye movements and meaning maps. *Journal of Vision*, 18(6), 10–10. <https://doi.org/10.1167/18.6.10>
- Henderson, J. M., & Hollingworth, A. (1999). High-level scene perception. *Annual Review of Psychology*, 50, 243–271. <https://doi.org/10.1146/annurev.psych.50.1.243>
- Henderson, J. M., Weeks, P. A., & Hollingworth, A. (1999). The effects of semantic consistency on eye movements during complex scene viewing. *Journal of Experimental Psychology: Human Perception & Performance*, 25(1), 210–228. <https://doi.org/10.1037/0096-1523.25.1.210>
- Hinde, S. J., Smith, T. J., & Gilchrist, I. D. (2017). In search of oculomotor capture during film viewing: Implications for the balance of top-down and bottom-up control in the saccadic system. *Vision Research*, 134, 7–17. <https://doi.org/10.1016/j.visres.2017.01.007>
- Hinde, S. J., Smith, T. J., & Gilchrist, I. D. (2018). Does narrative drive dynamic attention to a prolonged stimulus? *Cognitive Research: Principles and Implications*, 3(1), Article 45. <https://doi.org/10.1186/s41235-018-0140-5>
- Hochberg, J., & Peterson, M. A. (1987). Piecemeal organization and cognitive components in object perception: Perceptually coupled responses to moving objects. *Journal of Experimental Psychology: General*, 116(4), Article 370.
- Hollingworth, A. (2009). Memory for real-world scenes. In J. R. Brockmole (Ed.), *The visual world in memory* (pp. 89–119). Psychology Press.
- Huff, M., Meitz, T. G., & Papenmeier, F. (2014). Changes in situation models modulate processes of event perception in audiovisual narratives. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(5), 1377–1388.

- Huff, M., Papenmeier, F., & Zacks, J. M. (2012). Visual target detection is impaired at event boundaries. *Visual Cognition*, 20(7), 848–864. <https://doi.org/10.1080/13506285.2012.705359>
- Humphrey, K., & Underwood, G. (2010). The potency of people in pictures: Evidence from sequences of eye fixations. *Journal of Vision*, 10(10), Article 10.1167/10.10.19.
- Hutson, J. P., Chandran, P., Magliano, J. P., Smith, T. J., & Loschky, L. C. (2022). Narrative comprehension guides eye movements in the absence of motion. *Cognitive Science*, 46(5), Article e13131.
- Hutson, J. P., Magliano, J. P., & Loschky, L. C. (2018). Understanding moment-to-moment processing of visual narratives [journal article]. *Cognitive Science*, 42, 2999–3033. <https://doi.org/10.1111/cogs.12699>
- Hutson, J. P., Magliano, J. P., Smith, T. J., & Loschky, L. C. (2021). This ticking noise in my head?: How sound design, dialogue, event structure, and viewer working memory interact in the comprehension of *Touch of Evil* (1958). *Projections*, 15(1), 1–27. <https://doi.org/10.3167/proj.2021.150102>
- Hutson, J. P., Smith, T. J., Magliano, J. P., & Loschky, L. C. (2017). What is the role of the film viewer? The effects of narrative comprehension and viewing task on gaze control in film [journal article]. *Cognitive Research: Principles and Implications*, 2(1), Article 46. <https://doi.org/10.1186/s41235-017-0080-5>
- Irwin, D. E. (1996). Integrating information across saccadic eye movements. *Current Directions in Psychological Science*, 5(3), 94–100.
- Irwin, D. E., & Gordon, R. D. (1998). Eye movements, attention and trans-saccadic memory. *Visual Cognition*, 5(1/2), 127–155.
- 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.
- Ji, Y., & Papafragou, A. (2022). Boundedness in event cognition: Viewers spontaneously represent the temporal texture of events. *Journal of Memory and Language*, 127, Article 104353. <https://doi.org/10.1016/j.jml.2022.104353>
- Just, M. A., & Carpenter, P. A. (1987). *The psychology of reading and language comprehension*. Allyn & Bacon.
- Kaakinen, J. K., Hyönä, J., & Viljanen, M. (2011). Influence of a psychological perspective on scene viewing and memory for scenes. *Quarterly Journal of Experimental Psychology*, 64(7), 1372–1387. <https://doi.org/10.1080/17470218.2010.548872>
- Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K. A., Cichy, R. M., Hauk, O., & Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*, 116(43), 21854–21863. <https://doi.org/10.1073/pnas.1905544116>
- Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, 95(2), 163–182.
- Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge University Press.
- 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.
- Kravitz, D. J., Saleem, K. S., Baker, C. I., Ungerleider, L. G., & Mishkin, M. (2013). The ventral visual pathway: An expanded neural framework for the processing of object quality. *Trends in Cognitive Sciences*, 17(1), 26–49. <https://doi.org/10.1016/j.tics.2012.10.011>
- Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, 1, 417–446. <https://doi.org/10.1146/annurev-vision-082114-035447>
- Kummerer, M., Wallis, T. S., Gatys, L. A., & Bethge, M. (2017). Understanding low-and high-level contributions to fixation prediction. *Proceedings of the 2017 IEEE International Conference On Computer Vision*. <https://ieeexplore.ieee.org/document/8237755>
- Kurby, C. A., & Zacks, J. M. (2008). Segmentation in the perception and memory of events. *Trends in Cognitive Sciences*, 12(2), Article 72.
- Kurby, C. A., & Zacks, J. M. (2011). Age differences in the perception of hierarchical structure in events. *Memory & Cognition*, 39, 75–91.
- Kurby, C. A., & Zacks, J. M. (2012). Starting from scratch and building brick by brick in comprehension. *Memory & Cognition*, 40(5), 812–826. <https://doi.org/10.3758/s13421-011-0179-8>
- Lahnakoski, J. M., Glerean, E., Jääskeläinen, I. P., Hyönä, J., Hari, R., Sams, M., & Nummenmaa, L. (2014). Synchronous brain activity across individuals underlies shared psychological perspectives. *NeuroImage*, 100, 316–324. <https://doi.org/10.1016/j.neuroimage.2014.06.022>
- Lankhuizen, T., Bálint, K. E., Savardi, M., Konijn, E. A., Bartsch, A., & Benini, S. (2022). Shaping film: A quantitative formal analysis of contemporary empathy-eliciting Hollywood cinema. *Psychology of Aesthetics, Creativity, and the Arts*, 16(4), 704–718. <https://doi.org/10.1037/aca0000356>
- Larson, A. M. (2012). *Recognizing the setting before reporting the action: Investigating how visual events are mentally constructed from scene images* [Doctoral dissertation, Kansas State University].
- Larson, A. M., Freeman, T. E., Ringer, R. V., & Loschky, L. C. (2014). The spatiotemporal dynamics of scene gist recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 40(2), 471–487. <https://doi.org/10.1037/a0034986>
- Larson, A. M., & Loschky, L. C. (2009). The contributions of central versus peripheral vision to scene gist recognition. *Journal of Vision*, 9(10), 1–16.
- Linardos, A., Kümmerer, M., Press, O., & Bethge, M. (2021). Deep-Gaze IIE: Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling. *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision*.
- Long, D. L., & Lea, R. B. (2005). Have we been searching for meaning in all the wrong places? Defining the “search after meaning” principle in comprehension. *Discourse Processes*, 39(2/3), 279–298.
- Loschky, L. C., Hutson, J. P., Smith, M. E., Smith, T. J., & Magliano, J. P. (2018). Viewing static visual narratives through the lens of the Scene Perception and Event Comprehension Theory (SPECT). In J. Laubrock, J. Wildfeuer, & A. Dunst (Eds.), *Empirical comics research: Digital, multimodal, and cognitive methods* (pp. 217–238). Routledge.
- Loschky, L. C., Larson, A. M., Magliano, J. P., & Smith, T. J. (2015). What would Jaws do? The tyranny of film and the relationship between gaze and higher-level narrative film comprehension. *PLoS One*, 10(11), 1–23. <https://doi.org/10.1371/journal.pone.0142474>
- Loschky, L. C., Larson, A. M., Smith, T. J., & Magliano, J. P. (2020). The scene perception & event comprehension theory (SPECT) applied to visual narratives. *Topics in Cognitive Science*, 12(1), 311–351. <https://doi.org/10.1111/tops.12455>
- Loschky, L. C., Sethi, A., Simons, D. J., Pydimari, T., Ochs, D., & Corbeille, J. (2007). The importance of information localization in scene gist recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 33(6), 1431–1450.
- Loschky, L. C., Szaffarczyk, S., Beugnet, C., Young, M. E., & Boucart, M. (2019). The contributions of central and peripheral vision to scene gist recognition with a 180° visual field. *Journal of Vision*, 19(5), 1–21. <https://doi.org/10.1167/19.5.15>
- Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, 390(6657), 279–281.

- Ma, C., Jin, Y., & Lauwereyns, J. (2024). Speed is associated with polarization during subjective evaluation: No tradeoff, but an effect of the ease of processing. *Cognitive Neurodynamics*, 18(6), 3691–3714. <https://doi.org/10.1007/s11571-024-10151-8>
- Mack, A., & Rock, I. (1998). *Inattention blindness* (Vol. 6). MIT Press.
- Magliano, J. P. (2024). Exploring inferences across media: An application of Kintsch's framework of inference processes. *Discourse Processes*. <https://doi.org/10.1080/0163853X.2024.2336416>
- Magliano, J. P., Kopp, K., McNerney, M. W., Radvansky, G. A., & Zacks, J. M. (2012). Aging and perceived event structure as a function of modality. *Aging, Neuropsychology, and Cognition*, 19(1/2), 264–282. <https://doi.org/10.1080/13825585.2011.633159>
- Magliano, J. P., Kurby, C. A., Ackerman, T., Garlitch, S. M., & Stewart, J. M. (2020). Lights, camera, action: The role of editing and framing on the processing of filmed events. *Journal of Cognitive Psychology*, 32(5/6), 506–525. <https://doi.org/10.1080/20445911.2020.1796685>
- Magliano, J. P., Larson, A. M., Higgs, K., & Loschky, L. C. (2016). The relative roles of visuospatial and linguistic working memory systems in generating inferences during visual narrative comprehension. *Memory & Cognition*, 44(2), 207–219. <https://doi.org/10.3758/s13421-015-0558-7>
- Magliano, J. P., Miller, J., & Zwaan, R. A. (2001). Indexing space and time in film understanding. *Applied Cognitive Psychology*, 15(5), 533–545. <https://doi.org/10.1002/acp.724>
- Magliano, J. P., Skowronski, J. J., Britt, M. A., Güss, C. D., & Forsythe, C. (2008). What do you want? How perceivers use cues to make goal inferences about others. *Cognition*, 106(2), 594–632. <https://doi.org/10.1016/j.cognition.2007.03.010>
- Mayer, M. (1975). *One frog too many*. Dial Press.
- McNamara, D. S., & Magliano, J. P. (2009). Toward a comprehensive model of comprehension. In B. H. Ross (Ed.), *Psychology of Learning and Motivation* (Vol. Volume 51, pp. 297–384). Elsevier Science. [https://doi.org/10.1016/S0079-7421\(09\)51009-2](https://doi.org/10.1016/S0079-7421(09)51009-2)
- Milner, A. D., & Goodale, M. A. (2008). Two visual systems reviewed. *Neuropsychologia*, 46(3), 774–785. <https://doi.org/10.1016/j.neuropsychologia.2007.10.005>
- Mital, P. K., Smith, T. J., Hill, R. L., & Henderson, J. M. (2010). Clustering of gaze during dynamic scene viewing is predicted by motion. *Cognitive Computation*, 3(1), 5–24. <https://doi.org/10.1007/s12559-010-9074-z>
- Mitchell, J. P., Macrae, C. N., & Gilchrist, I. D. (2002). Working memory and the suppression of reflexive saccades. *Journal of Cognitive Neuroscience*, 14(1), 95–103. <https://doi.org/10.1162/089892902317205357>
- Myers, J. L., & O'Brien, E. J. (1998). Accessing the discourse representation during reading. *Discourse Processes*, 26(2/3), 131–157. <https://doi.org/10.1080/01638539809545042>
- Najemnik, J., & Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature*, 434(7031), 387–391. <https://doi.org/10.1038/nature03390>
- Nonaka, S., Majima, K., Aoki, S. C., & Kamitani, Y. (2021). Brain hierarchy score: Which deep neural networks are hierarchically brain-like? *iScience*. <https://doi.org/10.1016/j.isci.2021.103013>
- Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research, Special Issue on Visual Perception*, 155, 23–36.
- Olivers, C. N., Meijer, F., & Theeuwes, J. (2006). Feature-based memory-driven attentional capture: Visual working memory content affects visual attention. *Journal of Experimental Psychology. Human Perception and Performance*, 32(5), Article 1243.
- Palmer, S. E. (1975). Visual perception and world knowledge: Notes on a model of sensory-cognitive interaction. In D. A. Norman, D. E. Rumelhart, & t. L. R. Group (Eds.), *Explorations in cognition* (pp. 279–307). Freeman.
- Pannasch, S. (2014). Characteristics of ambient and focal processing during the visual exploration of dynamic stimuli. *Journal of Vision*, 14(10), 1208–1208.
- Pannasch, S., Helmer, J. R., Roth, K., Herbold, A. K., & Walter, H. (2008). Visual fixation durations and saccade amplitudes: Shifting relationship in a variety of conditions. *Journal of Eye Movement Research*, 2(2):4, 1–19. <https://doi.org/10.16910/jemr.2.2.4>
- 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.
- Pedziwiatr, M. A., Heer, S., Coutrot, A., Bex, P., & Mareschal, I. (2023). Prior knowledge about events depicted in scenes decreases oculomotor exploration. *Cognition*, 238, Article 105544. <https://doi.org/10.1016/j.cognition.2023.105544>
- Pedziwiatr, M. A., Kümmerer, M., Wallis, T. S. A., Bethge, M., & Teufel, C. (2021). Meaning maps and saliency models based on deep convolutional neural networks are insensitive to image meaning when predicting human fixations. *Cognition*, 206, Article 104465. <https://doi.org/10.1016/j.cognition.2020.104465>
- Peterson, J. J. (2018). *The interaction between visual resolution and task-relevance in guiding visual selective attention*. Kansas State University.
- Peterson, M. A. (1994). Object recognition processes can and do operate before figure-ground organization. *Current Directions in Psychological Science*, 3(4), 105–111.
- Peterson, M. A., & Gibson, B. S. (1994). Must figure-ground organization precede object recognition? An assumption in peril. *Psychological Science*, 5(5), 253–259.
- Pettijohn, K. A., & Radvansky, G. A. (2016). Walking through doorways causes forgetting: Environmental effects. *Journal of Cognitive Psychology*, 28(3), 329–340.
- Poynor, D. V., & Morris, R. K. (2003). Inferred goals in narratives: Evidence from self-paced reading, recall, and eye movements. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 29(1), Article 3.
- Ptak, R. (2012). The frontoparietal attention network of the human brain: Action, saliency, and a priority map of the environment. *Neuroscientist*, 18(5), 502–515. <https://doi.org/10.1177/1073858411409051>
- Raabe, M. C., López, F. M., Yu, Z., Caplan, S., Yu, C., Shi, B. E., & Triesch, J. (November 2023). Saccade amplitude statistics are explained by cortical magnification. *2023 IEEE International Conference on Development and Learning (ICDL)*.
- Radvansky, G. A. (2012). Across the event horizon. *Current Directions in Psychological Science*, 21(4), 269–272.
- Radvansky, G. A., & Zacks, J. M. (2011). Event perception. *Wiley Interdisciplinary Reviews. Cognitive Science*, 2(6), 608–220.
- Radvansky, G. A., & Zacks, J. M. (2014). *Event cognition*. Oxford University Press.
- Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3), 372–422. <https://doi.org/10.1037/0033-2909.124.3.372>
- Rensink, R. A., O'Regan, J. K., & Clark, J. J. (1997). To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science*, 8(5), 368–373.
- Ringer, R. V., Throneburg, Z., Johnson, A. P., Kramer, A. F., & Loschky, L. C. (2016). Impairing the useful field of view in natural scenes: Tunnel vision versus general interference. *Journal of Vision*, 16(2), 1–25.
- Roberts, R. J., Hager, L. D., & Heron, C. (1994). Prefrontal cognitive processes: Working memory and inhibition in the antisaccade task. *Journal of Experimental Psychology: General*, 123(4), 1374–1393.

- Ross, J., Morrone, M. C., Goldberg, M. E., & Burr, D. C. (2001). Changes in visual perception at the time of saccades. *Trends in Neurosciences*, 24(2), 113–121.
- Rouinfar, A., Agra, E., Larson, A. M., Rebello, N. S., & Loschky, L. C. (2014). Linking attentional processes and conceptual problem solving: Visual cues facilitate the automaticity of extracting relevant information from diagrams. *Frontiers in Psychology*, 5. <https://doi.org/10.3389/fpsyg.2014.01094>
- Seiple, W., Clemens, C., Greenstein, V. C., Holopigian, K., & Zhang, X. (2002). The spatial distribution of selective attention assessed using the multifocal visual evoked potential. *Vision Research*, 42(12), 1513–1521.
- Silva, M., Baldassano, C., & Fuentemilla, L. (2019). Rapid memory reactivation at movie event boundaries promotes episodic encoding. *bioRxiv*. <https://doi.org/10.1523/JNEUROSCI.0360-19.2019>
- Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattention blindness for dynamic events. *Perception*, 28(9), 1059–1074.
- Simonson, T. L., Hutson, J. P., Yu, Y., Kumakiri, S., Ueda, Y., Saiki, J., & Loschky, L. C. (2021). The price of breaking the tyranny of film: The cognitive demand of top-down processes. *Journal of Vision*, 21(9), 2610–2610. <https://doi.org/10.1167/jov.21.9.2610>
- Simonson, T. L., Kumakiri, S., Yu, Y., Ueda, Y., Saiki, J., & Loschky, L. C. (2020, June). *What besides the film guides viewers' attention while watching films? The roles of culture and task*. The Annual Meeting of the Society for Cognitive Studies of the Moving Image, Virtual Conference.
- Smirl, J. D., Wright, A. D., Bryk, K., & van Donkelaar, P. (2016). Where's Waldo? The utility of a complicated visual search paradigm for transcranial Doppler-based assessments of neurovascular coupling. *Journal of Neuroscience Methods*, 270, 92–101. <https://doi.org/10.1016/j.jneumeth.2016.06.007>
- Smith, M. E. (2021). *Neural mechanisms underlying the influence of sequential predictions on scene gist recognition*. Kansas State University.
- Smith, M. E., Hutson, J. P., Newell, M. K., Wing-Paul, D., McCarthy, K. S., Loschky, L. C., & Magliano, J. P. (2024). Bridging a gap in coherence: The coordination of comprehension processes when viewing visual narratives. *Vision*, 8(3), Article 50.
- Smith, M. E., & Loschky, L. C. (2019). The influence of sequential predictions on scene-gist recognition. *Journal of Vision*, 19(12), 1–24. <https://doi.org/10.1167/19.12.14>
- Smith, M. E., Loschky, L. C., & Bailey, H. R. (2024). Eye movements and event segmentation: Eye movements reveal age-related differences in event model updating. *Psychology and Aging*, 39(2), 180–187. <https://doi.org/10.1037/pag0000773>
- Smith, M. E., Zacks, J. M., & Reagh, Z. M. (2025). Events in the stream of behavior. *Current Opinion in Behavioral Sciences*, 65, Article 101581. <https://doi.org/10.1016/j.cobeha.2025.101581>
- Smith, T. J. (2012). The attentional theory of cinematic continuity. *Projections*, 6(1), 1–27. <https://doi.org/10.3167/proj.2012.060102>
- Smith, T. J. (2012). The relationship between overt attention and event perception during dynamic social scenes. *Journal of Vision*, 12(9), Article 407. <https://doi.org/10.1167/12.9.407>
- Smith, T. J. (2013). Watching you watch movies: Using eye tracking to inform cognitive film theory. In A. P. Shimamura (Ed.), *Psychocinematics: Exploring cognition at the movies*. Oxford University Press.
- Smith, T. J., & Henderson, J. M. (2011). Looking back at Waldo: Oculomotor inhibition of return does not prevent return fixations. *Journal of Vision*, 11(1), 3–3. <https://doi.org/10.1167/11.1.3>
- Smith, T. J., Levin, D. T., & Cutting, J. E. (2012). A window on reality: Perceiving edited moving images. *Current Directions in Psychological Science*, 21(2), 107–113.
- Smith, T. J., & Martin-Portugues Santacreu, J. Y. (2017). Match-action: The role of motion and audio in creating global change blindness in film. *Media Psychology*, 20(2), 317–348.
- Smith, T. J., & Mital, P. K. (2013). Attentional synchrony and the influence of viewing task on gaze behaviour in static and dynamic scenes. *Journal of Vision*, 13(8):16, 1–24. <https://doi.org/10.1167/13.8.16>
- Sols, I., DuBrow, S., Davachi, L., & Fuentemilla, L. (2017). Event boundaries trigger rapid memory reinstatement of the prior events to promote their representation in long-term memory. *Current Biology*, 27(22), 3499–3504. <https://doi.org/10.1016/j.cub.2017.09.057>
- Soto, D., Heinke, D., Humphreys, G. W., & Blanco, M. J. (2005). Early, involuntary top-down guidance of attention from working memory. *Journal of Experimental Psychology. Human Perception and Performance*, 31(2), 248.
- Soto, D., & Humphreys, G. W. (2007). Automatic guidance of visual attention from verbal working memory. *Journal of Experimental Psychology. Human Perception and Performance*, 33(3), 730.
- Stawarczyk, D., Bezdek, M. A., & Zacks, J. M. (2021). Event representations and predictive processing: The role of the midline default network core. *Topics in Cognitive Science*, 13(1), 164–186. <https://doi.org/10.1111/tops.12450>
- Strasburger, H., & Malania, M. (2013). Source confusion is a major cause of crowding. *Journal of Vision*. <https://doi.org/10.1167/13.1.24>
- Swallow, K. M., Zacks, J. M., & Abrams, R. A. (2009). Event boundaries in perception affect memory encoding and updating. *Journal of Experimental Psychology. General*, 138(2), 236–257. <https://doi.org/10.1037/a0015631>
- 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. <https://doi.org/10.1126/science.7777863>
- Tatler, B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, 7(14), 4: 1–17.
- Tatler, B. W., Gilchrist, I. D., & Land, M. F. (2005). Visual memory for objects in natural scenes: From fixations to object files. *The Quarterly Journal of Experimental Psychology Section A*, 58(5), 931–960. <https://doi.org/10.1080/02724980443000430>
- Tatler, B. W., Hayhoe, M. M., Land, M. F., & Ballard, D. H. (2011). Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision*. <https://doi.org/10.1167/11.5.5>
- Taya, S., Windridge, D., & Osman, M. (2012). Looking to score: The dissociation of goal influence on eye movement and meta-attentional allocation in a complex dynamic natural scene. *PLoS One*, 7(6), 1–9.
- Theeuwes, J. (2018). Visual selection: Usually fast and automatic; Seldom slow and volitional. *Journal of Cognition*, 1(1), Article 29. <https://doi.org/10.5334/joc.13>
- Theeuwes, J., Kramer, A. F., Hahn, S., Irwin, D. E., & Zelinsky, G. J. (1999). Influence of attentional capture on oculomotor control. *Journal of Experimental Psychology. Human Perception and Performance*, 25(6), 1595–1608. <https://doi.org/10.1037/0096-1523.25.6.1595>
- 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.
- Trabasso, T., van den Broek, P., & Suh, S. (1989). Logical necessity and transitivity of causal relations in stories. *Discourse Processes*, 12(1), 1–25.
- Tsotsos, J. K., Culhane, S. M., Kei Wai, W. Y., Lai, Y., Davis, N., & Nufflo, F. (1995). Modeling visual attention via selective tuning.

- Artificial Intelligence, 78(1), 507–545. [https://doi.org/10.1016/0004-3702\(95\)00025-9](https://doi.org/10.1016/0004-3702(95)00025-9)
- Underwood, G., & Foulsham, T. (2006). Visual saliency and semantic incongruity influence eye movements when inspecting pictures. *Quarterly Journal of Experimental Psychology*, 59(11), 1931–1949. <https://doi.org/10.1080/17470210500416342>
- Unsworth, N., Schrock, J. C., & Engle, R. (2004). Working memory capacity and the antisaccade task: Individual differences in voluntary saccade control. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(6), Article 1302.
- Valuch, C., König, P., & Ansorge, U. (2017). Memory-guided attention during active viewing of edited dynamic scenes. *Journal of Vision*, 17(1), 1–31. <https://doi.org/10.1167/17.1.12>
- van den Broek, P. (1990). The causal inference maker: Towards a process model of inference generation in text comprehension. In D. A. Balota, G. B. Flores d'Arcais, & K. Rayner (Eds.), *Comprehension processes in reading* (pp. 423–445). Erlbaum.
- Vitu, F., Adeli, H., & Zelinsky, G. J. (2021). Readers move their eyes mindlessly using midbrain visuo-motor principles. *bioRxiv*
- Vitu, F., O'Regan, J. K., Inhoff, A. W., & Topolski, R. (1995). Mindless reading: Eye-movement characteristics are similar in scanning letter strings and reading texts. *Perception & Psychophysics*, 57(3), 352–364. <https://doi.org/10.3758/bf03213060>
- Vö, M.-H., Boettcher, S. E. P., & Draschkow, D. (2019). Reading scenes: How scene grammar guides attention and aids perception in real-world environments. *Current Opinion in Psychology*, 29, 205–210. <https://doi.org/10.1016/j.copsyc.2019.03.009>
- Vö, M. L. H., & Henderson, J. M. (2011). Object–scene inconsistencies do not capture gaze: Evidence from the flash-preview moving-window paradigm. *Attention, Perception & Psychophysics*, 73(6), 1742–1753. <https://doi.org/10.3758/s13414-011-0150-6>
- Vorderer, P., Klimmt, C., & Ritterfeld, U. (2006). Enjoyment: At the heart of media entertainment. *Communication Theory*, 14(4), 388–408. <https://doi.org/10.1111/j.1468-2885.2004.tb00321.x>
- Wahlheim, C. N., Eisenberg, M. L., Stawarczyk, D., & Zacks, J. M. (2022). Understanding everyday events: Predictive-looking errors drive memory updating. *Psychological Science*, 33(5), 765–781. <https://doi.org/10.1177/09567976211053596>
- Wang, H. X., Freeman, J., Merriam, E. P., Hasson, U., & Heeger, D. J. (2012). Temporal eye movement strategies during naturalistic viewing. *Journal of Vision*. <https://doi.org/10.1167/12.1.16>
- Wardle, S. G., & Baker, C. I. (2020). Recent advances in understanding object recognition in the human brain: deep neural networks, temporal dynamics, and context. *F1000Research*, 9(590). <https://doi.org/10.12688/f1000research.22296.1>
- Welles, O., & Zugsmith, A. (1958). *Touch of evil* [Motion picture]. Universal Pictures.
- Wiley, J., & Rayner, K. (2000). Effects of titles on the processing of text and lexically ambiguous words: Evidence from eye movements. *Memory & Cognition*, 28(6), 1011–1021. <https://doi.org/10.3758/bf03209349>
- Williams, C. C., Henderson, J. M., & Zacks, R. T. (2005). Incidental visual memory for targets and distractors in visual search. *Perception & Psychophysics*, 67(5), 816–827. <https://doi.org/10.3758/BF03193535>
- Williams, L. J. (1988). Tunnel vision or general interference? Cognitive load and attentional bias are both important. *American Journal of Psychology*, 101, 171–191.
- Wolfe, J. M. (2007). Guided search 4.0: Current progress with a model of visual search. In W. Gray (Ed.), *Integrated models of cognitive systems* (pp. 99–119). Oxford.
- Wolfe, J. M. (2021). Guided search 6.0: An updated model of visual search. *Psychonomic Bulletin & Review*, 28(4), 1060–1092. <https://doi.org/10.3758/s13423-020-01859-9>
- 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.
- Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23), 8619–8624. <https://doi.org/10.1073/pnas.1403112111>
- Yarbus, A. L. (1967). *Eye movements and vision* (B. Haigh, Trans.). Plenum.
- Yates, T. S., Yasuda, S., & Yildirim, I. (2024). Temporal segmentation and “look ahead” simulation: Physical events structure visual perception of intuitive physics. *Journal of Experimental Psychology: Human Perception and Performance*, 50(8), 859–874. <https://doi.org/10.1037/xhp0001218>
- Zacks, J., Speer, N., Swallow, K., & Maley, C. (2010). The brain's cutting-room floor: Segmentation of narrative cinema [Original Research]. *Frontiers in Human Neuroscience*, 4. <https://doi.org/10.3389/fnhum.2010.00168>
- Zacks, J., & Tversky, B. (2001). Event structure in perception and conception. *Psychological Bulletin*, 127(1), 3–21.
- Zacks, J. M. (2015). *Flicker: Your brain on movies*. Oxford University Press.
- Zacks, J. M. (2020). Event perception and memory. *Annual Review of Psychology*, 71(1), 165–191. <https://doi.org/10.1146/annurev-psych-010419-051101>
- Zacks, J. M., Kurby, C., Eisenberg, M. L., & Haroutunian, N. (2011). Prediction error associated with the perceptual segmentation of naturalistic events. *Journal of Cognitive Neuroscience*, 23(12), 4057–4066. https://doi.org/10.1162/jocn_a_00078
- Zacks, J. M., Speer, N., & Reynolds, J. (2009). Segmentation in reading and film comprehension. *Journal of Experimental Psychology: General*, 138(2), 307–327. 2009-05547-010
- Zacks, J. M., Speer, N., Swallow, K., Braver, T., & Reynolds, J. (2007). Event perception: A mind-brain perspective. *Psychological Bulletin*, 133(2), 273–293. <https://doi.org/10.1037/0033-2909.133.2.273>
- Zelinsky, G. J., & Bisley, J. W. (2015). The what, where, and why of priority maps and their interactions with visual working memory. *Annals of the New York Academy of Sciences*, 1339(1), 154–164.
- Zelinsky, G. J., Chen, Y., Ahn, S., & Adeli, H. (2020). Changing perspectives on goal-directed attention control: The past, present, and future of modeling fixations during visual search. In K. D. Federmeier & E. R. Schotter (Eds.), *Psychology of Learning and Motivation* (Vol. 73, pp. 231–286). Academic Press. <https://doi.org/10.1016/bs.plm.2020.08.001>
- Zwaan, R. A., Langston, M. C., & Graesser, A. C. (1995). The construction of situation models in narrative comprehension: An event-indexing model. *Psychological Science*, 6(5), 292–297. <https://doi.org/10.1111/j.1467-9280.1995.tb00513.x>
- Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, 123(2), 162–185.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.