TOPICS
TOPICS IN COGNITIVE SCIENCE

Topics in Cognitive Science 12 (2020) 311–351

ISSN: 1756-8765 online
DOI: 10.1111/tops.12455

This article is part of the topic "Visual Narrative Research: An Emerging Field in Cognitive Science," Neil Cohn and Joseph P. Magliano (Topic Editors). For a full listing of topic papers, see http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1756-8765/earlyview

The Scene Perception & Event Comprehension Theory (SPECT) Applied to Visual Narratives

Lester C. Loschky,[a] Adam M. Larson,[b] Tim J. Smith,[c] Joseph P. Magliano[d]

[a]*Department of Psychological Sciences, Kansas State University*
[b]*Department of Psychology, University of Findlay*
[c]*Department of Psychological Sciences, Birkbeck, University of London*
[d]*College of Education & Human Development, Georgia State University*

Received 26 November 2018; received in revised form 5 August 2019; accepted 5 August 2019

Abstract

Understanding how people comprehend visual narratives (including picture stories, comics, and film) requires the combination of traditionally separate theories that span the initial sensory and perceptual processing of complex visual scenes, the perception of events over time, and comprehension of narratives. Existing piecemeal approaches fail to capture the interplay between these levels of processing. Here, we propose the Scene Perception & Event Comprehension Theory (SPECT), as applied to visual narratives, which distinguishes between front-end and back-end cognitive processes. Front-end processes occur during single eye fixations and are comprised of attentional selection and information extraction. Back-end processes occur across multiple fixations and support the construction of event models, which reflect understanding of what is happening now in a narrative (stored in working memory) and over the course of the entire narrative (stored in long-term episodic memory). We describe relationships between front- and back-end processes, and medium-specific differences that likely produce variation in front-end and back-end processes across media (e.g., picture stories vs. film). We describe several novel research questions derived from SPECT that we have explored. By addressing these questions, we provide greater insight into

Correspondence should be sent to Lester C. Loschky, Department of Psychological Sciences, Kansas State University, 471 Bluemont Hall, 1114 Mid-Campus Dr. North, Manhattan, KS 66506, USA. E-mail: loschky@ksu.edu

how attention, information extraction, and event model processes are dynamically coordinated to perceive and understand complex naturalistic visual events in narratives and the real world.

Keywords: Scene perception; Event perception; Narrative comprehension; Visual narratives; Film; Comics; Eye movements; Attention

1. Introduction

How do viewers both *perceive* and *understand* visual narratives? This is a difficult and complex question that has not previously been addressed in any comprehensive theory (but see Cohn, 2013a; Cohn, 2019b). It involves coordinating perceptual and comprehension processes that operate over multiple images and produce a durable mental model of a narrative (Loschky, Hutson, Smith, Smith, & Magliano, 2018). Consider Fig. 1A, which shows three images from Mercer Mayer's (1967) visual narrative, *A Boy*, *a Dog*, and *a Frog*. In the first image, the viewer sees a boy carrying a net and a pail, running down a wooded hill with his dog, toward a frog on a lily pad, in a pond at the bottom of the hill. The viewer also sees a tree branch close to the ground about halfway down the hill. Several things are worth noting here. First, the viewer needs to recognize the boy, net, pail, dog, frog, lily pad, pond, hill, and tree branch as such. Likewise, the viewer needs to recognize that the boy and dog are running down the hill, and that the boy is carrying the net and pail. Research suggests that all of these things can be recognized very rapidly, with the boy, dog, tree branch, and hill potentially being recognized within the time frame of a single eye fixation (Fei-Fei, Iyer, Koch, & Perona, 2007). The fact that the boy and dog are running may also be recognized within the first fixation, or require a further fixation to extract the necessary visual detail (Glanemann, 2008; Larson, Hendry, & Loschky, 2012). The frog may also be too small to detect peripherally and require an additional fixation (Nelson & Loftus, 1980). Similarly, the fact that the boy is carrying a net and a pail will likely require one or two further fixations. Importantly, each additional fixation requires the viewer to attentionally select part of the image for further processing (Deubel & Schneider, 1996), though these selections are usually made preconsciously (Belopolsky, Kramer, & Theeuwes, 2008; Memmert, 2006). All of these processes can be thought of as basic perceptual building blocks of the scene.

But the viewer must also make sense of the how these agents, their actions, objects, and scene background elements depict *events* in a narrative. For example, to understand the narrative, viewers must infer the goal of the boy (Graesser & Clark, 1985; Long, Golding, & Graesser, 1992; Suh & Trabasso, 1993), which is to the catch the frog. Inferences of this sort can be generated very quickly in the context of reading texts (Long et al., 1992). When viewing such a picture story, viewers can generate an inference within two extra fixations on details of the scene that suggest the inference (e.g., the direction of the boy's eye gaze relative to the location of the frog, and the position of the boy's net, suggest his goal of catching the frog: Hutson et al., 2018).

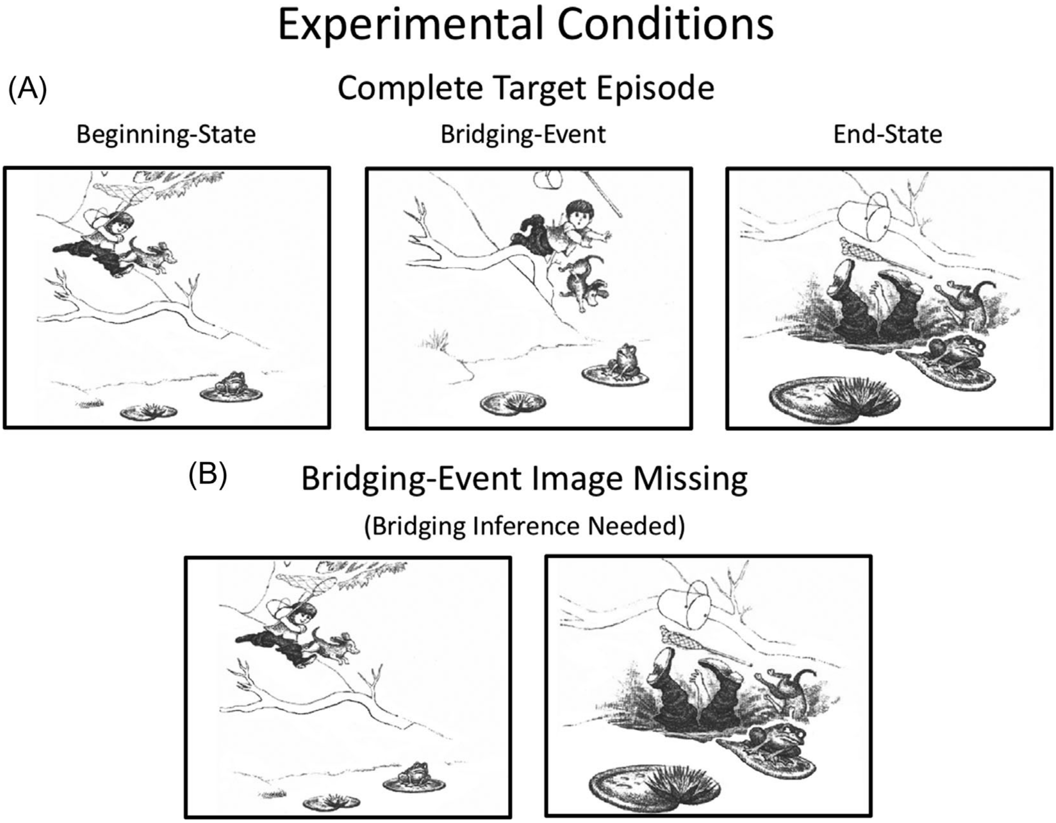

Fig. 1. Experimental conditions used to elicit bridging inferences while viewers read a picture story (Hutson, Magliano, & Loschky, 2018; Magliano, Larson, Higgs, & Loschky, 2016). (A) Complete target episode from "A Boy, a Dog, and a Frog" (Mayer, 1967), including beginning-state, bridging-event, and end-state images. (B) The target episode missing the bridging-event image, which requires the viewer to generate a bridging inference when viewing the end-state image to maintain coherence with the beginning-state image.

The second image shows the boy and dog tripping over the tree branch, with the boy having let go of his net and pail, and shows the frog noticing these events. Understanding this picture also requires the same processes of scene perception and attentional selection described for the first picture, but those processes should be supported by representations of the prior narrative context held in working memory and episodic memory (Graesser, Millis, & Zwaan, 1997). Importantly, the boy and dog tripping on the tree branch is inconsistent with their inferred goal of catching the frog from the first picture;thus, it reflects a *failure* of that goal (Trabasso, van den Broek, & Suh, 1989). Understanding how this picture fits into the narrative involves inferring the causal relationship between it and the prior narrative context (e.g., the boy tripped because he was running down the hill and there was a tree branch blocking his way; the boy failed to achieve his goal) (Trabasso & Suh, 1993). The final image shows the same scene from a slightly zoomed-in view. This image illustrates that viewers need to understand how the products of scene perception are related across images. The viewer sees boots sticking out of the water and must recognize that those are *the boy's* boots. Establishing this relationship has implications at the level of the narrative event model because it implies that the boy has fallen in the water, and that this is a result of his tripping on the tree branch (as depicted in the previous image).

Clearly, this illustrates the coordination of visual perception and narrative comprehension in the cognitive processing of visual narratives. Thus, it is no surprise that scholars who study these different levels of cognitive processes have become interested in how they are involved in visual narratives. Specifically, research on visual narrative processing has been rapidly expanding in the domains of visual scene perception (e.g., Hutson, Smith, Magliano, & Loschky, 2017), event perception/cognition (e.g., Zacks, Speer, & Reynolds, 2009), psycholinguistics (e.g., Cohn, 2013a), and narrative comprehension (e.g., Magliano, Kopp, McNerney, Radvansky, & Zacks, 2012). While these research areas are seemingly disparate, the above example illustrates that comprehensive theoretical frameworks are needed to explain how these processes are coordinated to support the perception and understanding of visual narratives. Additionally, such theoretical frameworks may be necessary to prevent fragmentation of the visual narrative research field, as occurred, for example, in reading research, where multiple models accounted for aspects of reading, such as word identification, syntactic parsing, discourse representations, and the roles of the reader's eye movements (Rayner & Reichle, 2010). Thus, the novel theoretical contribution of our theoretical framework lies in integrating processes from the scene perception literature and processes from the event perception and narrative comprehension literatures, raising interesting research questions concerning interactions between them. In this way, research on visual narratives within our framework is an example of complex cognition that can inform our broader understanding of naturalistic visual processing and transcend the currently compartmentalized research on visual narrative processing in the separate, minimally interacting research fields.

Below, we outline a theoretical framework, the Scene Perception & Event Comprehension Theory (SPECT: Loschky et al., 2018), that describes how perceptual processes and event model construction processes are coordinated during visual narrative processing. The novel theoretical contribution of SPECT lies in being an *integrative theoretical framework*, which identifies important interactions between perceptual and event model processes. SPECT allows researchers to identify the core perceptual and cognitive processes for perceiving and comprehending visual media. Critically, these core processes are also utilized in non-narrative contexts, such as real-world scenes. In formulating SPECT, we demonstrate how visual narratives are an example of complex cognition of broader interest to the cognitive sciences in general.

2. The SPECT framework

SPECT builds on decades of theoretical developments in general cognition and its subsystems (e.g. working memory, attentional control, etc.). Thus, SPECT is the application of general models of visual cognition to visual narratives, and many of SPECT's assumptions equally apply to real-world scene perception. SPECT bridges theories of scene perception (Henderson & Hollingworth, 1999; Irwin, 1996), event cognition (Radvansky & Zacks, 2011, 2014), and narrative comprehension (Gernsbacher, 1990; Zwaan & Radvansky, 1998). SPECT specifically pertains only to processing visual content; namely, it does *not* specify processes involved in processing either *language* narrowly defined, or non-linguistic *audio*. The basic architecture of SPECT distinguishes between *stimulus features* and *front-end* and *back-end* cognitive processes involved in visual event and narrative cognition, as illustrated in Fig. 2. Note that the front-end versus back-end distinction is *not* equivalent to bottom-up and top-down processes. We will clarify these distinctions below. We will briefly overview how these processes are conceptualized in SPECT before outlining each component in more detail with supporting evidence.

SPECT's starting point is the stimulus. All visual narratives are composed of either static (e.g., in Fig. 1) or dynamic (in the case of film, theater, or virtual reality) visual images of varying degrees of complexity and realism composed in sequence. Some stimulus properties constrain later processes within SPECT via *medium–agnosti*c *mechanisms* such as the *salience* of primitive visual features (e.g., luminance, contrast, or motion; Itti & Koch, 2001). For example, Fig. 3A shows the computed saliency of the "Beginning State" image from Fig. 1. For this saliency algorithm (AWS) (Garcia-Diaz, Fdez-Vidal, Pardo, & Dosil, 2009), the highest computed saliency regions (i.e., most likely to capture a viewer's attention) are for the Boy's head and the Frog's legs. This is based on analyzing the orientations of image elements at numerous size scales, and finding the local regions that are the most different from the rest of the image. Fig. 3B shows an actual fixation heat map, based on 39 viewers' fixations while viewing this image within the context of the entire visual narrative. The computed saliency is very close to the empirical fixation probabilities. Other stimulus properties are *medium-specific* such as the panels, layout, and action lines in comics, which are assumedly learned, rather than universal, in contrast to visual saliency (Cohn, 2013b). The three-panel layout of the images in Fig. 1A is familiar to comic readers, and it is meant to be read from left to right. In film, camera movements, cuts, and the predetermined pace of the moving images are similarly meant to guide viewers' attention (Bordwell & Thompson, 2003). Thus, the combination of medium-agnostic and medium-specific stimulus features shape what potential information is available to the viewer, and likely influence how front-end and back-end processes interact in processing this information.

Front-end processes are involved in extracting content from the image, and back-end processes operate on their output to support the construction of an event model. Front-end processes *occur during single eye fixations*. These processes during a fixation extend from the earliest perceptual processes to activated semantic representations that are sent to working memory (WM). The front-end involves two key processes that occur during each fixation: *information extraction* and *attentional selection*. Information extraction is further subdivided between *broad* (the gist of the whole scene, e.g., woods) and *narrow* (detailed information from animate or inanimate *entities*, e.g., boy, dog, frog, net, pail, etc., in Fig. 1). Information extraction includes both entities and events, with event information extraction also producing both broad categorizations of what is seen (e.g., "trying

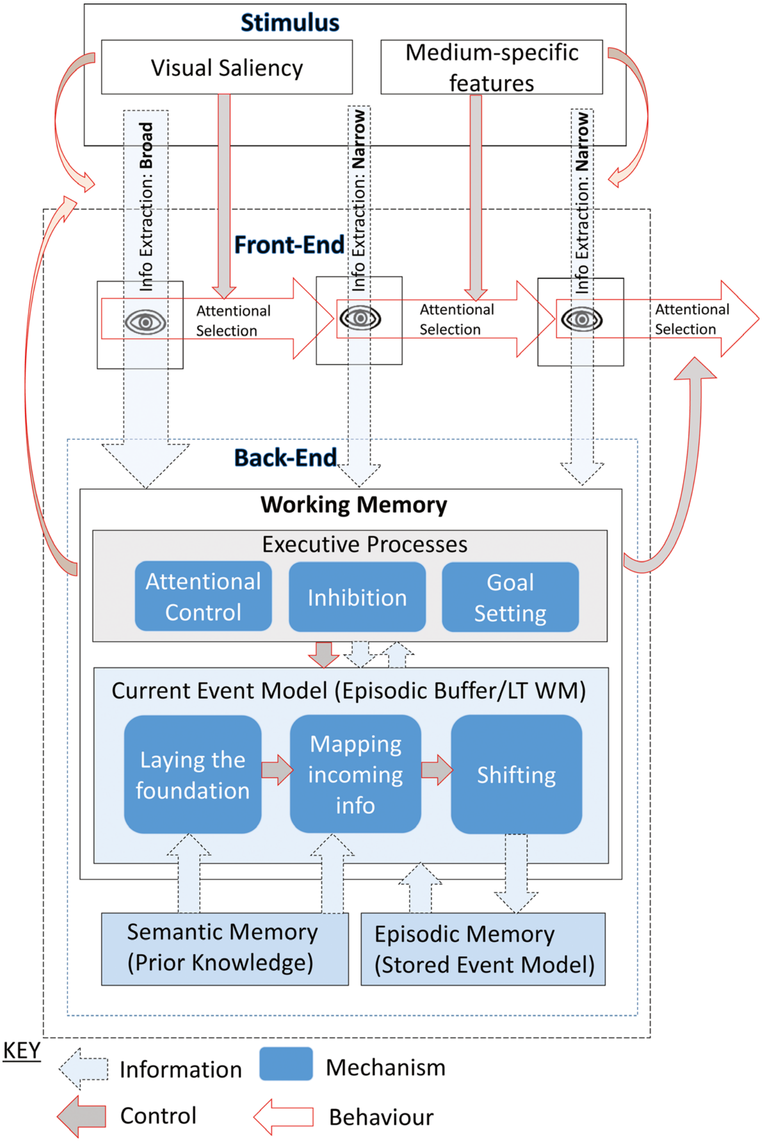

Fig. 2. Model of the Scene Perception & Event Comprehension Theory (SPECT) theoretical framework. The eye icon denotes the position of viewer gaze on the stimulus during a particular fixation. A further walk-through of the framework is provided in the text below.

to catch") and narrow categorizations (e.g., "running" in Fig. 1). The information extracted during each fixation is fed to the back-end. Attentional selection determines what information to process during single fixations, and where the eyes will be sent for the next fixation, and is influenced by both exogenous and endogenous factors. Note that the above definition of *front-end* processes is far more specific (i.e., occurring during

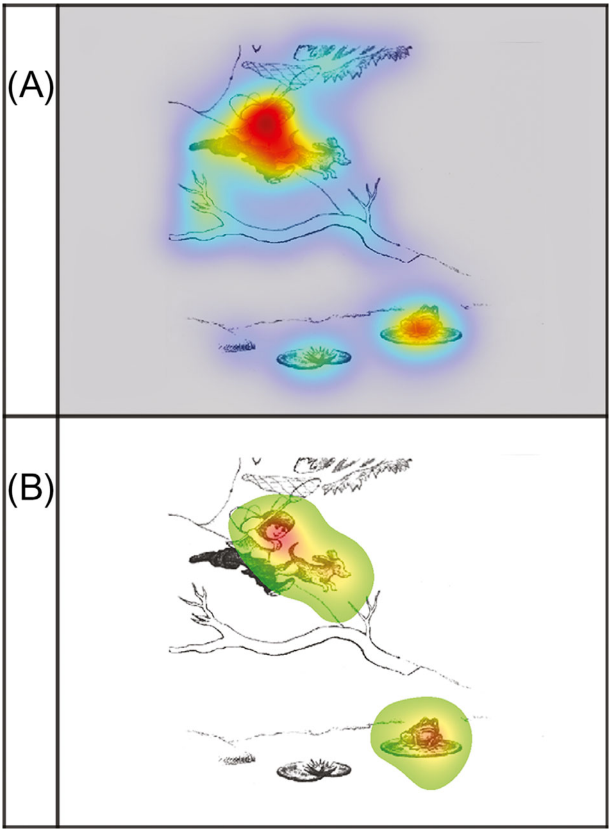

Fig. 3. (A) Example of computationally predicted visual saliency of regions in the Beginning-State image of Fig. 1, using the AWS saliency algorithm (Garcia-Diaz et al., 2009). (B) Fixation heat map from 39 viewers reading the wordless visual narrative. In both images, red = highest saliency/fixation probability. (Saliency and fixation heat map images courtesy of Maverick E. Smith.)

single fixations) and limited (i.e., information extraction and attentional selection) than the term *bottom-up* processes, and thus the two terms are not synonymous.

Back-end processes *occur in memory across multiple eye fixations*, specifically WM and long-term memory (LTM). The information represented in the back-end is accumulated over multiple eye fixations spanning durations extending from milliseconds to minutes. A key back-end process is the construction and maintenance of the *current event model* in WM, which represents what is happening now (e.g., a boy and his dog, trying to catch a frog in the woods, in Fig. 1). An event model is a particular type of *mental model* that captures a sequenced event. This representation is maintained until perceptual and conceptual content specifies that it is no longer relevant or valid (due to content changing over time–the boy's falling in the pond indicates the failure of his attempt to catch the frog, in Fig. 1). At that point, back-end processes encode the event model into episodic LTM, which we call a *stored event model* (e.g., the boy and dog tried to catch a frog in the woods, but fell in the pond, in Fig. 1). From these stored event models, more semantic event schemas can be derived in semantic LTM by averaging across multiple event model instances (e.g., Hintzman, 1988). The recently stored event models in episodic LTM will feed back to and influence the new current event model in WM (e.g., the expectation that the boy may make another attempt to catch the frog; see information arrows back from Episodic Memory to the Event Model in Fig. 2). The current event model is also influenced by *schemas* in semantic LTM (e.g., "little boys," "catching animals," etc.), and from *executive functions*, like goal setting, attention control, and inhibition. Note too that the above definition of *back-end* processes is also more specific (i.e., in memory across multiple fixations) and limited (to the event model building processes in WM, the stored event model in episodic LTM, stored knowledge in semantic LTM) than the term *bottom-up* processes.

An underlying assumption of SPECT is that front-end and back-end processes iteratively support the creation of the current event model in WM, and the management of stored event models in episodic LTM. Importantly, front-end attentional selection and information extraction guide the moment-to-moment knowledge retrieval from semantic LTM that supports the back-end processes of creating the current event model in WM (McKoon & Ratcliff, 1998; Myers & O'Brien, 1998). Thus, we cannot understand how knowledge is retrieved from LTM in the moment without understanding the role of these front-end processes. Similarly, a key theoretical issue raised by SPECT is whether and how back-end processes, including the current event model in WM, the stored event models in episodic LTM, and schemas and scripts in semantic LTM, influence the front-end information extraction and attentional selection processes. Thus, SPECT provides a theoretical framework to explore and explain the relationships between front- and back-end processes during visual narrative processing.

2.1. Theoretical foundations for front-end processes

When looking at real-world scenes, comics, or videos, visual information extraction only occurs during periods in which the eyes are stabilized relative to fixed points in space (fixations) or slowly moving objects (smooth pursuit or vestibulo–ocular reflex). This is because processing of visual detail is suppressed during the rapid shifts (saccadic eye movements) between locations (Matin, 1974; Ross, Morrone, Goldberg, & Burr, 2001). Thus, we can consider *eye fixations*[1] *to be the spatio-temporal input units of vision*. Furthermore, any extracted information maintained across multiple fixations is in short-term memory or WM (Irwin, 1996; Zelinsky & Loschky, 2005),[2] which is strongly constrained in terms of capacity (i.e., 3–4 items without rehearsal or chunking: Cowan, 2001), and encodable information (i.e., post-perceptual information: Hollingworth, 2009; Irwin, 1996). This key insight provides the rationale for distinguishing between front-end processes occurring during single fixations, and back-end processes occurring across multiple fixations occurring in memory. Furthermore, these constraints from eye movements necessarily shape how events in the environment, comics, or films are understood and become long-term episodic memories.

2.1.1. Information extraction

What types of information are extracted during a single eye fixation? SPECT distinguishes broad versus narrow information extraction (Loschky et al., 2018). Broad extraction is from all or most of an entire scene, producing holistic semantic information called scene gist (Oliva, 2005). This includes the basic level category of a scene (e.g., woods, a pond, in Fig. 1) (Fei-Fei et al., 2007; Greene & Oliva, 2009; Loschky & Larson, 2010), detecting animals or people (Fletcher-Watson, Findlay, Leekam, & Benson, 2008; Thorpe, Fize, & Marlot, 1996), the scene's emotional valence (Calvo, Nummenmaa, & Hyönä, 2007; Maljkovic & Martini, 2005), some rather rudimentary information about basic level actions (e.g., running vs. falling in Fig. 1) (Larson, 2012), and both the agent and patient of an action (e.g., the boy [agent] trying to catch the frog [patient], in Fig. 1) (Dobel, Gumnior, Bölte, & Zwitserlood, 2007; Hafri, Papafragou, & Trueswell, 2013). Narrow extraction operates on a particular entity (object, animal, or person) providing details such as colors, shapes, and sizes of object parts (Hollingworth, 2009; Pertzov, Avidan, & Zohary, 2009). Such broadly and narrowly extracted information in WM is used for comprehending events by back-end processes in the current event model. Importantly, despite the wide range of information extracted during a single eye fixation, the total amount of consciously available information from a single fixation remains limited, and thus increasingly detailed information from a scene or image must accrue in WM over multiple fixations (Hollingworth & Henderson, 2002; Pertzov et al., 2009) in the back-end. One slight caveat to this assumption of SPECT is that fixations are actually made up of micro movements (e.g. microsaccades, drift, etc), which may constitute phases of slight attentional shifts and changes in perceived information within a single fixation (Otero-Millan, Troncoso, Macknik, Serrano-Pedraza, & Martinez-Conde, 2008) and that the phases of attending to and processing a specific object could also be made up of multiple fixations dwelling within the object (Nuthmann & Henderson, 2010). Since both of these behaviors can still be considered fixations at different spatiotemporal scales, we will use the all-encompassing term *fixation* within SPECT.

2.1.2. Attentional selection

The other key front-end process during each fixation is *attentional selection*, which is *the gateway to WM, comprehension, and explicit LTM* for events. On each fixation, before moving the eyes, attention covertly shifts to the next to-be-fixated object (Deubel & Schneider, 1996; Hoffman & Subramaniam, 1995; Kowler, Anderson, Dosher, & Blaser, 1995). Attentional selection is affected by both exogenous, bottom-up, stimulus saliency, as described above (Borji & Itti, 2013; Wolfe & Horowitz, 2004), and endogenous, top-down, cognitive processes (DeAngelus & Pelz, 2009; Eckstein, Drescher, & Shimozaki, 2006; Findlay & Walker, 1999). Specifically, stimulus saliency is determined by visual *feature contrast* in terms of motion, brightness, color, orientation, and size (Mital, Smith, Hill, & Henderson, 2010; Peters, Iyer, Itti, & Koch, 2005). However, top-down, task-driven goals, such as searching for specific information, more strongly affect viewers' attention than saliency in pictures (Foulsham & Underwood, 2007; Henderson, Brockmole, Castelhano, & Mack, 2007), and some evidence of saliency-override by task has been demonstrated in film viewing, although this is believed to be more difficult (Hutson et al., 2017; Smith & Mital, 2013). More specifically, there are *volitional* (consciously controlled) versus *mandatory* (unconscious prior knowledge-based) top-down effects on attentional selection (Baluch & Itti, 2011). These can interact in tasks, such as visual search, in which the volitional top-down goal of finding a specific target (e.g., a chimney) is facilitated by mandatory top-down knowledge of likely target locations (e.g., at the top of a house: Eckstein et al., 2006; Torralba, Oliva, Castelhano, & Henderson, 2006). In SPECT, volitional top-down attentional control occurs in WM, using executive processes (Moss, Schunn, Schneider, McNamara, & VanLehn, 2011). Mandatory top-down processes can come from the event model or relevant world knowledge (i.e., schemas).

Attentional selection during single fixations can be narrowly focused, for example at the point of fixation, or broadly spread across a large portion of the visual field, also known as *attentional breadth*, or a person's *useful field of view* (Ball, Beard, Roenker, Miller, & Griggs, 1988; Eriksen & Yeh, 1985; Larson, Freeman, Ringer, & Loschky, 2014). Importantly, this can change dynamically based on the viewer's processing demands (Ringer, Throneburg, Johnson, Kramer, & Loschky, 2016; Williams, 1988). A viewer's breadth of attention also changes over the course of ~12–24 fixations in the first 4–6 s of viewing an image in the *ambient-to-focal shift* of eye movements (Pannasch, Helmert, Roth, Herbold, & Walter, 2008; Smith & Mital, 2013). Specifically, during the first 2 s of viewing an image, viewers tend to make long saccades, indicating broad attention, and short fixations, indicating shallow processing. Then, from 4 to 6 s of viewing, viewers shift to making short saccades, indicating narrowly focused attention, and long fixations, indicating deeper processing. Because this ambient-to-focal shift occurs across multiple fixations, back-end processes could possibly influence the front-end process of attentional selection.

While information extraction and attentional selection are considered independent within SPECT, and strong empirical and theoretical evidence supports their separation (Smith, Lamont, & Henderson, 2012; Triesch, Ballard, Hayhoe, & Sullivan, 2003), in active processing of scenes, these processes often operate in conjunction (Williams, Henderson, & Zacks, 2005). For example, when discussing how a viewer's fixation of an object influences his or her memory for it, we implicitly combine both attentional selection (i.e., choosing which object to send your eyes to) and information extraction (i.e., visual processing during the viewer's fixation on the object, as implicated by their later memory of it). As such, for parsimony we will sometimes refer to both processes together in later discussions.

2.2. Theoretical foundations for back-end processes

Back-end processes support the construction of a coherent current event model in WM, which later becomes a stored event model in episodic LTM (Magliano et al., 2012). A coherent event model contains information about the time and place in which the events unfold (the spatio-temporal framework), the entities in the event (people, animals, objects), the properties of those entities (e.g., colors, sizes, emotions, goals), the actions of the agents, the unintentional events that occur (e.g., acts of nature), and relational information (spatial, temporal, causal, ownership, kinship, social, etc.) (Magliano, Miller, & Zwaan, 2001; Zwaan, Magliano, & Graesser, 1995; Zwaan & Radvansky, 1998). As shown in Fig. 2, SPECT describes three key back-end processes involved in constructing the current event model: laying the foundation for a new event model, mapping incoming information to the current event model, and shifting to create a new event model (Gernsbacher, 1990).

2.2.1. Laying the foundation

Laying the foundation is the process of constructing the first nodes in an event model, where a node reflects a basic unit of representation (e.g., proposition, simple grounded simulation). These nodes then become memory structures to which subsequent information is connected or not (Gernsbacher, 1990, 1997). When a new event model is created, the viewer must lay the foundation for it. In the context of a visual narrative, the foundation will likely involve a representation of the spatial-temporal information that is extracted through gist processing, and any agents and actions recognized in the first fixation of the images.

As noted above, the information extraction process can gather some rudimentary information about basic level actions, including the agent and the patient, within a single eye fixation (Glanemann, 2008; Hafri et al., 2013). However, due to the limits of information processing within the time span of a single fixation (e.g., 330 ms), it takes at least two fixations to reach peak accuracy for identifying an action (Hafri et al., 2013; Larson, 2012). Thus, the information required to lay the foundation for the current event model, namely recognizing a basic action, requires integrating information across at least two fixations in WM.

2.2.2. Mapping incoming information

With each subsequent fixation, the viewer builds upon the foundation by mapping incoming information to WM, but only if it is coherent with the event model (Gernsbacher, 1990, 1997). This process involves monitoring continuities in the *event indices* of time, space, entities, causality, and goals (Gernsbacher, 1997; Zwaan & Radvansky, 1998). Specifically, situational information extracted by front-end processes serves as LTM retrieval cues, thus activating semantically related information in WM (Myers & O'Brien, 1998). Viewers assess the coherence of the event indices within the current event model and the newly activated information from LTM. Changes along any event index that are coherent with the current event model will lead viewers to incrementally update, or map, that change (Kurby & Zacks, 2012). In this way, the current event model becomes gradually elaborated as more information is extracted on each eye fixation.

Mapping is supported by inference generation (Graesser, Singer, & Trabasso, 1994), particularly *bridging inferences* (Magliano, Zwaan, & Graesser, 1999). Bridging inferences connect two or more story events and are considered necessary for maintaining a coherent mental model (Graesser et al., 1994). Virtually all comprehension models consider bridging inferences important (McNamara & Magliano, 2009) because they are required when comprehenders perceive a gap in the narrative events (e.g., Magliano et al., 2016), or when two narrative events are causally related (e.g., Suh & Trabasso, 1993). For example, in Fig. 1B, the Bridging-Event image shown in Fig. 1A is missing. Thus, for viewers seeing only the Beginning-State and End-State images in Fig. 1B, they would need to generate a bridging inference to coherently map the information from the End-State image (boy and dog fell in the pond) onto the foundation of the event model created based on the information from the Beginning State image (boy and dog running down the hill to catch a frog).

2.2.3. Shifting

When mapping is no longer possible, the viewer shifts to create a new event model. This occurs when new incoming information produces a trigger signal, resulting in *event segmentation*, which parses this continuous activity into discrete events (Kurby & Zacks, 2008; Magliano et al., 2012). For example, when watching someone making breakfast, we recognize the discrete actions of taking a slice of bread out of a loaf, putting the slice in a toaster, toasting it, taking it out of the toaster, and putting it on a plate (Newtson, 1973; Newtson, Engquist, & Bois, 1977). Segmentation is critical for understanding and remembering complex events (Magliano et al., 2012; Radvansky & Zacks, 2011; Sargent et al., 2013).

Segmentation also occurs when we experience narratives, and triggers can be either perceptual or more conceptual in nature. For example, visual motion is strongly associated with event segmentation (Zacks, Swallow, Vettel, & McAvoy, 2006). Other important triggers are when viewers perceive shifts in situational continuities, such as shifts in time and space, causal discontinuities, the introduction of new characters, or changes in characters' goal-plans (Magliano et al., 2012; Zacks et al., 2009; Zwaan & Radvansky, 1998). If such changes are important enough, it indicates an *event boundary*, also known in older story grammar theories as a boundary between narrative *episodes* (Baggett, 1979; Gernsbacher, 1985; Thorndyke, 1977). For example, most readers of the visual narrative fragment in Fig. 1 will perceive an event boundary to have occurred on the End-State image, assumedly because the Boy's attempt to achieve his goal of catching the Frog has failed. Perceiving an event boundary means the current event has ended, which triggers a shift (Kurby & Zacks, 2012), and leads to storing the current event model in LTM as a global update to the previously stored event models in episodic LTM (Gernsbacher, 1985). Once this boundary has been perceived and the stored event model updated, information from the previous event model becomes less accessible (Gernsbacher, 1985; Swallow, Zacks, & Abrams, 2009). Once shifting is complete, the cycle begins again with laying the foundation for a new event model.

2.3. Executive processes

The back-end comprehension processes discussed above occur by default without the viewer's volition. Yet viewers can exert volitional control over their mental processes when they feel the need to do so. This likely happens when the viewer is given a task unrelated to understanding the story while viewing a visual narrative (Hutson et al., 2017; Lahnakoski et al., 2014). This seems relatively uncommon when people read comics or watch movies for pleasure, but it is very common when students are given educational tasks in school settings (Britt, Rouet, & Durik, 2018; McCrudden, Magliano, & Schraw, 2010). Such volitional strategic comprehension processes are more cognitively demanding (Kaakinen, Hyönä, & Keenan, 2003), and they engage frontal and prefrontal brain regions known to be involved in executive processes (Moss et al., 2011). This suggests that volitional control of comprehension processes involves executive processes, such as goal setting (i.e., deciding to carry out a specified task), attentional control (i.e., paying attention to task-relevant information), and inhibition (i.e., intentionally ignoring irrelevant information), as indicated in Fig. 2. For example, in Hutson et al. (2017, Exp 2B), prior to watching a film clip from *Touch of Evil*, viewers were told that after watching the clip, they would be asked to draw a map of all landmarks and their relative locations from memory. Assumedly, doing this task successfully would involve setting the goal of memorizing the landmarks and their locations, volitionally controlling one’s attention to meet this goal (e.g., by fixating background buildings, street signs, etc.), and inhibiting attending to the protagonists of the narrative (which would conflict with the spatial memorization goal). We assume that such executive processes are available to viewers of visual narratives, but they only use them when necessary, and only if they have the required WM resources, given their cognitive load, and the processing demands of the stimulus (e.g., rapidly edited film sequences may overload cognitive resources: Andreu-Sánchez, Martín-Pascual, Gruart, & Delgado-García, 2018; Lang, 2000).

3. Differences between static and dynamic media

There are potential differences in the complexities of processing narratives across media (Loschky et al., 2018; Magliano, Loschky, Clinton, & Larson, 2013). First, a growing literature indicates that fluency in processing visual narratives requires exposure and learning (Cohn & Kutas, 2017; Fussell & Haaland, 1978; Ildirar & Schwan, 2014; Liddell, 1997). This may explain why proficiency in comprehending text is weakly correlated with proficiency in comprehending visual narratives in children (Pezdek, Lehrer, & Simon, 1984), whereas they are robustly correlated in adults (Gernsbacher, Varner, & Faust, 1990). Second, there are non-trivial cultural differences in the structure of visual narratives across cultures, which in turn produce non-trivial differences in comprehension (Cohn & Kutas, 2017). Finally, the extent to which attentional selection can support event segmentation, inference generation, and model updating may be affected by whether consumption of the visual narrative is self-paced or externally controlled (Hutson et al., 2018; Magliano et al., 2013), a difference which SPECT is intended to describe.

Attentional selection and subsequent processing will also be shaped by medium-specific differences between reading comics or text and viewing films (Magliano et al., 2013). The static versus dynamic nature of visual narratives has a strong effect on the front-end process of attentional selection, as shown in Fig. 4, since viewers show greater *attentional* *synchrony* (i.e., looking at the same places at the same times) during a video clip in comparison to individual frames from the same video (Dorr, Martinetz, Gegenfurtner, & Barth, 2010; Smith & Mital, 2013). Viewers may tend to fixate similar locations in static scene perception (Mannan, Ruddock, & Wooding, 1997) *but not at the same time*, suggesting a larger influence of individual differences in both front- and back-end processes for static images (Hayes & Henderson, 2018; Le Meur, Le Callet, & Barba, 2007). Consistent with the finding of greater attentional synchrony during film viewing, motion is perhaps the most salient stimulus feature in guiding attention (Carmi & Itti, 2006; Le Meur et al., 2007; Mital et al., 2010). Another important medium-specific difference between comics and film, which will affect attentional selection, is the typical way each is viewed. When people read comics, their eyes move from panel to panel of a page layout, which stereotypically follows a "Z-path" of left-to-right and top-to-bottom, consistent with text (though the specific pattern varies by language, such as Hebrew and Japanese being read from right to left, and top to bottom). This is entirely different from film viewing, which instead shows a viewing pattern oriented toward the center of the screen due, in part, to the temporal presentation of visual information (Dorr et al., 2010; Le Meur et al., 2007; Mital et al., 2010).

Perhaps just as importantly, in reading both text and comics, comprehension differences are evident in the duration of fixations and the frequency of regressive saccades (Foulsham, Wybrow, & Cohn, 2016; Hutson et al., 2018; Laubrock, Hohenstein, & Kummerer, 2018; Rayner, 1998). Conversely, the predetermined pace of film does not provide viewers much time to look around and refixate things, when they have difficulty understanding what they saw (although this may occur more while watching digital video, where viewers can pause, stop, or rewind the video). Consider Fig. 5, which shows two different versions of the same scene from the graphic novel *Watchmen* (Moore & Gibbons, 1987) and the film adaptation of it (Snyder, 2009). The events depicted in both versions are identical. One would expect the general products of information extraction to be similar when processing either version. However, the graphic narrative affords more endogenous control of attentional selection because it is a static representation of the events. In contrast, the dynamic presentation of the film version has cinematic features that provide stronger exogenous influences on attentional selection such as dynamic framing through camera movements (e.g., the single zoom shot pulling back from the Comedian's badge in the film version versus the three frames depicting the same change in viewpoint in the comic; Fig. 5, middle row), lighting and focal depth changes, and choreography of actor motion within the frame (Hutson et al., 2017; Loschky, Larson, Magliano, & Smith, 2015; Smith, 2012a).

The scene depicted in Fig. 5 is predominantly action based, which means that the movie stills (essentially a storyboard) roughly convey the same content as the original comic version (as was intended for this particular film). The comic version may place more demands on back-end processes to guide attention and actively construct the event model than when viewing the movie version, for which the events are self-evident in the actions depicted. But in a scene like this it would seem reasonable to assume the resulting narrative event models would be medium-agnostic.

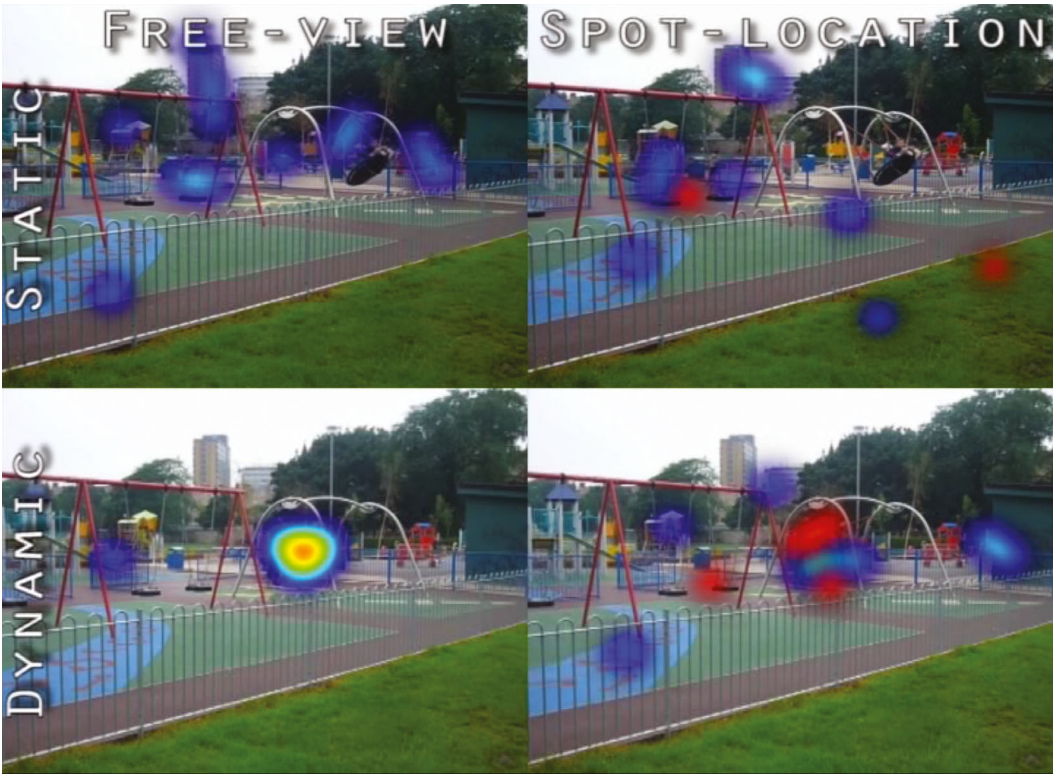

Fig. 4. The difference in gaze exploration of a scene (represented as a fixation heatmap of multiple viewers' gaze locations) across static (top row) and dynamic versions (bottom row), and free-viewing (left column) versus a spot-the-location task, which prioritizes background details (right column). Note that the most tightly clustered gaze is in the dynamic free-viewing condition, which contains motion. Note, also though, that this gaze clustering due to motion is somewhat reduced by giving viewers an explicit task (i.e., spot-the-location). (Reproduced with permission from Smith and Mital [2013].)

However, for scenes involving richer characterization and dialogue, the formal decisions comic artists and film directors make when composing their scenes may result in very different event models. For example, Fig. 6 depicts a later scene from *Watchmen* in which Rorschach meets with his old partner, Nite Owl. The comic version uses four panels each containing multiple visual centers of interest.[3] The reader likely must perform multiple fixations within each panel to extract information about the characters, their actions, and Nite Owl's emotional response to Rorschach (Laubrock et al., 2018). By comparison, the film version uses 12 shots varying widely in shot scale to convey the same information. By conveying each action serially (i.e., one per shot) this likely reduces the need for multiple fixations per shot, which raises interesting questions about whether back-end WM processes would differ depending on the number of front-end attentional shifts and fixations across media. Studies of medium differences in front-end attentional selection and information extraction on back-end processes process have been

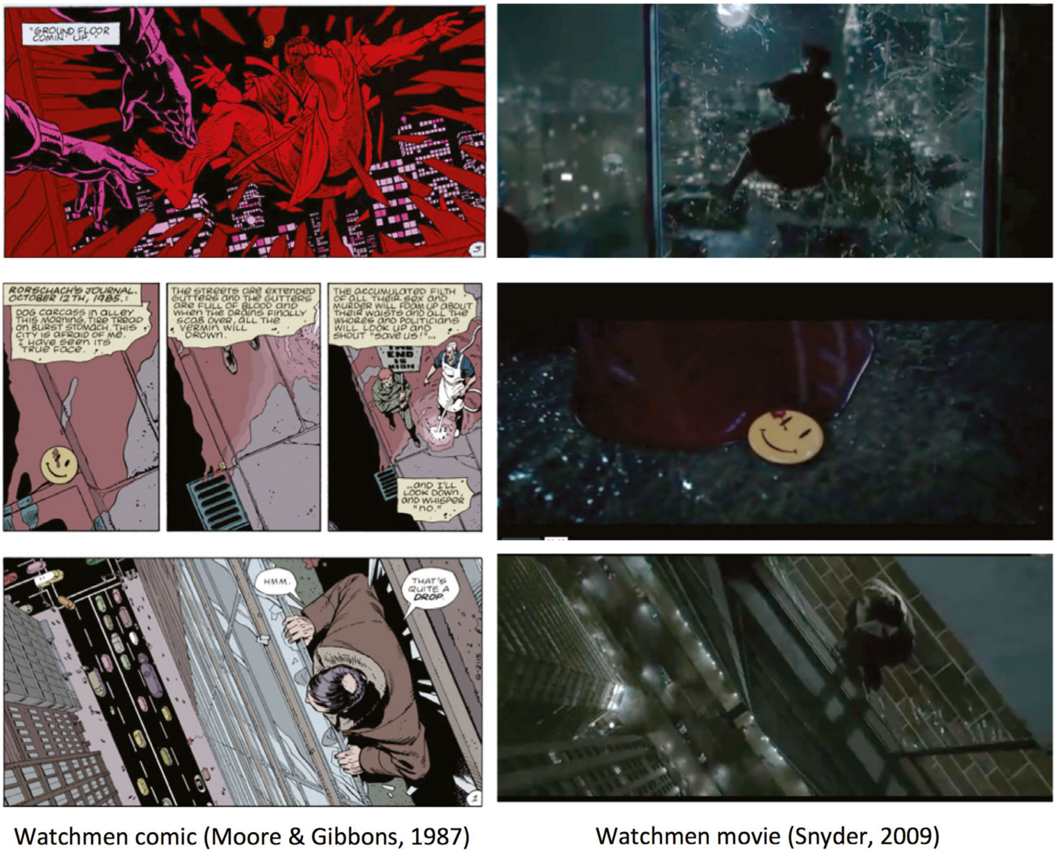
"GROUND FLOOR COMIN' UP."
3
RORSCHACH'S JOURNAL. OCTOBER 12TH, 1985.:
DOG CARCASS IN ALLEY THIS MORNING, TIRE TREAD ON BURST STOMACH. THIS CITY IS AFRAID OF ME. I HAVE SEEN ITS TRUE FACE.
THE STREETS ARE EXTENDED GUTTERS AND THE GUTTERS ARE FULL OF BLOOD AND WHEN THE DRAINS FINALLY SCAB OVER, ALL THE VERMIN WILL DROWN.
THE ACCUMULATED FILTH OF ALL THEIR SEX AND MURDER WILL FOAM UP ABOUT THEIR WAISTS AND ALL THE WHORES AND POLITICIANS WILL LOOK UP AND SHOUT "SAVE US!"...
THE END IS NIGH
...AND I'LL LOOK DOWN, AND WHISPER "NO."
HMM.
THAT'S QUITE A DROP.
1
Watchmen comic (Moore & Gibbons, 1987)
Watchmen movie (Snyder, 2009)

Fig. 5. Sample panels/stills from the original *Watchmen* (Moore & Gibbons, 1987) graphic novel and movie (Snyder, 2009) depicting the same actions. Note the images are not presented in their original sequence.

largely unexplored (Magliano, Clinton, O'Brien, & Rapp, 2018). However, the SPECT framework specifies the importance of exploring these issues.

4. Research questions raised by SPECT and their investigation

We have been using this framework to guide our program of research on the processing of visual narratives (both static and dynamic) for the past 8 years. Most of these studies have directly involved narratives, but a few have involved non-narrative content that mirrors important features of visual narratives. In those cases, we have adopted that approach because it afforded the experimental control needed to ask and answer the central questions raised by SPECT regarding processes involved in visual narratives. In addition, SPECT suggests that the distinction between static narratives and film may be important for attentional selection because the two media differ in terms of their degree

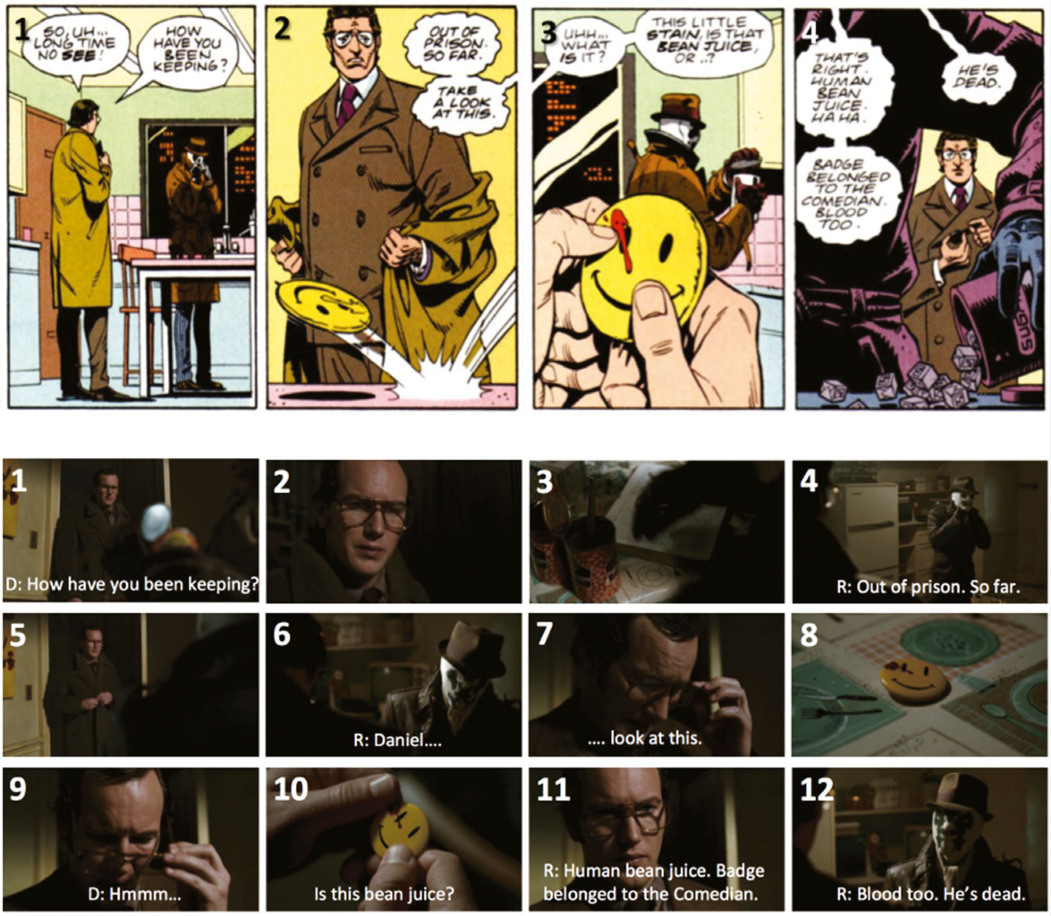
1
SO, UH... LONG TIME NO SEE!
HOW HAVE YOU BEEN KEEPING?
2
OUT OF PRISON. SO FAR.
TAKE A LOOK AT THIS.
3
UHH... WHAT IS IT?
THIS LITTLE STAIN, IS THAT BEAN JUICE, OR ..?
4
THAT'S RIGHT. HUMAN BEAN JUICE. HA HA.
HE'S DEAD.
BADGE BELONGED TO THE COMEDIAN. BLOOD TOO.
1
D: How have you been keeping?
2
3
4
R: Out of prison. So far.
5
6
R: Daniel....
7
.... look at this.
8
9
D: Hmmm...
10
Is this bean juice?
11
R: Human bean juice. Badge belonged to the Comedian.
12
R: Blood too. He's dead.

Fig. 6. Rorschach surprises Nite Owl in his kitchen and reveals the Comedian has died. Top row: panels taken from the original *Watchmen* (Moore & Gibbons, 1987) graphic novel. Bottom three rows: shots depicting the same action from the movie version (Snyder, 2009).

of visual salience (e.g., due to film having motion, but not comics). Nevertheless, to date, we have not made within-study comparisons between film and comics using the same content. With those caveats, in this section, we will illustrate SPECT's utility as a theoretical framework to guide research on the coordination of information extraction, attentional selection, basic event cognition, and event model construction in processing visual narratives.

To date, we have investigated two key lines of research, one regarding information extraction, and the other regarding attentional selection. Regarding information extraction, we have previously carried out studies on the discourse comprehension of visual narratives, which show how viewers monitor situation indices in their current event model, which then affects their event segmentation. However, those studies beg the question, *what information is extracted during single fixations in the front-end*? SPECT provides a framework for asking questions about how information extraction affects event model construction. Thus, we have investigated how extracting information on the viewer's first fixation on a new scene allows the viewer to lay the foundation of a new event model (Larson et al., 2012), and how that newly laid foundation primes extraction of further event indices on further fixations (Larson & Lee, 2015). We have also investigated how laying the foundation of the event model can, in turn, allow the viewer to predict what spatiotemporal context he or she will see next, which influences further information extraction on subsequent eye fixations (Smith & Loschky, in press). This can lead to further questions, such as how are subsequent event indices mapped onto an existing event model or signify an event model shift?

Regarding attentional selection, we have investigated how it is influenced by event model construction while viewing visual narratives, including both static picture stories and film (Hutson et al., 2017, 2018; Loschky et al., 2015). Specifically, we have studied how mapping incoming information to the current event model guides attentional selection in visual narratives with static images (picture stories) (Hutson et al., 2018). More specifically, how does the mapping process in the event model, and its subprocess of bridging inference generation, affect attentional selection, as measured by what viewers fixate on in a given picture in a visual narrative. We have also studied how the current and stored event models guide attentional selection in dynamic visual narratives (film clips) (Hutson et al., 2017; Loschky et al., 2015). More specifically, how does the mapping process in the event model, and its subprocess of predictive inference generation (forward mapping), affect attentional selection, as measured by what viewers look at from moment to moment while watching a narrative film?

Below, we discuss these studies, what they have shown that speaks to the SPECT framework, and a non-exhaustive sample of other relevant work that speaks to the same issues. We describe these studies in sections below, first on *The Relationship between Information Extraction and Event Model Construction*, and second on *The Relationship between Event Model Construction and Attentional Selection*.

4.1. The relationship between information extraction and event model construction

According to SPECT, the first stage of creating an event model is laying its foundation. This iteratively operates as one processes each picture (or frame) in a visual narrative, in a manner akin to the processes that support reading sentences in the context of narrative text (Magliano et al., 2013). For example, when viewing the first image of Fig. 1A (labeled "Beginning State"), the reader needs to quickly perceive who is doing what, when, and where. SPECT raises critical questions about how the process of information extraction, on each eye fixation enables the viewer to lay the foundation over the course of the first few eye fixations. Is there a temporal order in which the viewer recognizes that the scene takes place on a wooded hill, that there is a boy and a dog, and that they are both running? Perhaps the viewer recognizes the boy, the dog, and the wooded hill on the first fixation and stores that information as event indices in the foundation of the new event model.[4] And perhaps the viewer recognizes that both boy and dog are running down the hill on the second fixation and map that onto the foundation of the event model. If so, could this temporal order of information extraction imply that recognizing the spatiotemporal context ("wooded hill") on the first fixation facilitates recognizing the event ("running down the hill") on the second fixation? Alternatively, since comprehending a narrative requires recognizing the main character and his or her actions, perhaps the foundation of the event model requires recognizing this event information within the first eye fixation (Dobel et al., 2007; Hafri et al., 2013). Furthermore, attention is strongly biased to people in scene images within a single fixation (Fletcher-Watson et al., 2008; Humphrey & Underwood, 2010; Zwickel & Võ, 2010). These two points strongly suggest that people and their actions form the basis of an event model.

Larson (2012) explored these above issues within a non-narrative context, in order to gain an understanding of the processes involved when looking at the very first image in a narrative. Specifically, Larson (2012) examined the rapid categorization of locations and actions in static photographic scenes both within single eye fixations, and across multiple fixations. Larson found that viewers were able to rapidly categorize locations within a single fixation, but that actions required a second fixation. This suggests that laying the foundation consists of first recognizing the spatiotemporal framework, and then recognizing and mapping the actions that entities carry out within it. In a further study, Larson and Lee (2015) found that recognizing an action was facilitated by seeing it within the context of a recognizable scene. Importantly, however, this facilitation was only found after viewers had processed the image long enough to relatively accurately recognize the scene context (about 100 ms). Nevertheless, we have yet to explore if this hierarchy of recognizing the spatiotemporal framework first and then action occurs across pictures in a visual narrative.

SPECT raises further questions about the relationship between information extraction, laying the foundation, and mapping to the current event model, in the context of visual narrative sequences. For example, in Fig. 1A, according to SPECT, the boy's spatiotemporal context (i.e., "wooded hill") will be extracted while viewing the first picture, and stored as the foundation of the event model in working memory. A key question raised by SPECT is whether that foundation should facilitate information extraction of the spatiotemporal context on the subsequently viewed second and third pictures in Fig. 1A. Similar to anticipatory processes in language processing, we would expect to find priming of the upcoming spatiotemporal contexts, whether they remain the same, as in Fig. 1A (all showing a "wooded hill"), or they are different but spatiotemporally related (e.g., a transition from the wooded area into a field), but not if they clash with expectations (e.g., a transition from the wooded area into a bustling city street).

Smith and Loschky (in press) have investigated the above questions using simple first-person visual narratives of traveling from one location to another (e.g., going from an office to a parking lot) akin to the short narratives often used by discourse psychologists. As shown in Fig. 7, that study presented viewers with short narrative sequences of 0–9 scene priming images, each briefly flashed for enough time to both recognize and store them in working memory (about 300 ms), followed by a single target image that was briefly flashed and immediately masked to limit processing time (for 24 ms), after which the viewer was asked the categorize the target scene. The key manipulation was to present the image sequences in coherent versus randomized order. The results showed that viewers were much more accurate at categorizing the scenes shown in coherent sequences than in randomized sequences, showing clear priming of the current spatiotemporal context by the preceding context. Furthermore, the priming was greater when the priming was from the same category as the target (e.g., the second of two hallways in a row) than from a different but spatiotemporally related category to the target (e.g., a hallway seen immediately after one or more office images). This is consistent with the above hypothetical scenario for processing the image sequence in Fig. 1A, in which recognition of the second and third "wooded hill" images would be primed by recognizing the first such image. However, Smith and Loschky's (in press) results also showed that expectations about up-coming different spatiotemporal contexts also produced priming, but to a lesser degree.

Finally, Smith and Loschky (in press) investigated whether the spatiotemporal priming shown in coherent image sequences was simply due to response biases (i.e., guessing the scene category at the time of being tested), or was actually due to facilitation of perceptual sensitivity. To tease apart those possibilities, they showed participants the exact same coherent and randomized image sequences, but participants' task was changed from (1) identifying the scene category of the target to (2) visually discriminating whether the target was a real scene image or a noise image (with a 50/50 mix of both types of target images). Note that a viewer's ability to predict the category of the next scene should not bias him or her to respond either "real scene" or "noise image." Importantly, the results showed that participants were more perceptually sensitive to targets in the coherent than the randomized scene sequences, while their response bias was neutral (i.e., they equally responded "real scene" vs. "random noise") and did not differ between the coherent and randomized sequences. Thus, a viewer's expectations about the up-coming spatiotemporal context can facilitate their perception of that context.

4.1.1. Further research on the relationship between information extraction, attentional selection, and event model construction

A limitation of the above studies is that they either used only single images, or minimal visual narratives, rather than the more naturalistic ones found in comics, picture stories, and film. As previously discussed, the perceptual processing demands of static versus dynamic visual narratives differ greatly and these may alter the degree to which back-end processing influences front-end attentional selection and information extraction. For example, Smith and colleagues (Smith & Henderson, 2008; Smith & Martin-Portugues Santacreu, 2017) have demonstrated that continuity of a basic level action percept across a cut can obscure viewer awareness of the cut (i.e., *edit blindness*), which involved a global change in viewpoint of the spatiotemporal context. Object features and even actor identity can also change across cuts without viewers noticing (Levin & Simons, 1997). Diminished awareness of the shot change only occurs if sufficient action motion is present across the cut (hence the film technique name Match-On-Action: Smith & Martin-Portugues Santacreu, 2017), suggesting that viewers may often lack the capacity (e.g., attentional resources, working memory, or executive resources) to encode detailed surface

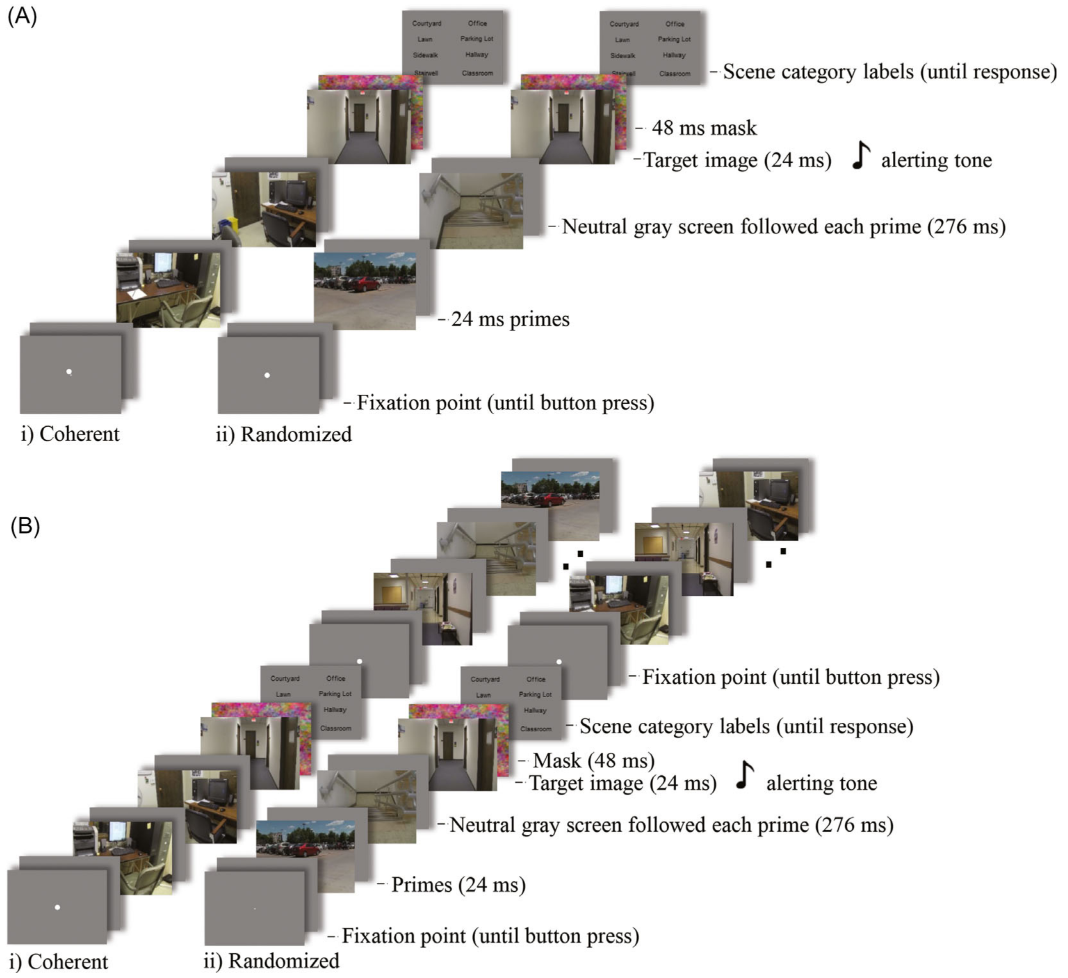
(A)
Courtyard
Office
Lawn
Parking Lot
Sidewalk
Hallway
Stairwell
Classroom
Scene category labels (until response)
48 ms mask
Target image (24 ms)
alerting tone
Neutral gray screen followed each prime (276 ms)
24 ms primes
Fixation point (until button press)
i) Coherent
ii) Randomized
(B)
Fixation point (until button press)
Scene category labels (until response)
Mask (48 ms)
Target image (24 ms)
alerting tone
Neutral gray screen followed each prime (276 ms)
Primes (24 ms)
Fixation point (until button press)
i) Coherent
ii) Randomized

Fig. 7. Experimental conditions used by Smith and Loschky (in press) to investigate priming of the current spatiotemporal context by the preceding context. Viewers saw either spatiotemporally coherent or randomized image sequences ending with a briefly flashed and visually masked target image. Participants then identified the scene category of the target from a list of all possible scene categories in that sequence. The coherent spatiotemporal sequence shows two office images followed by a hallway image, taken from a route from office to parking lot. The randomized sequence shows a parking lot, a stairwell, and then the target hallway image. Participants found the target images more predictable and were more accurate at identifying them, when presented in coherent sequences. (A) shows the beginnings of two sequences, including 2 primes, the target, and the response screen, in the (i) coherent, and (ii) randomized conditions. (B) shows a more complete representation of each sequence of 10 images, including those images that appeared after the participant's response.

information that does not change key event indices (Sampanes, Tseng, & Bridgeman, 2008) or that such information is obscured by the image motion (Smith, 2012a). Similar failures to notice differences between two different versions of the same static image are well known from "spot the difference" tasks. Assumedly, such changes are also missed in comics across pairs of adjacent busy panels sharing much of the same background. Within the Attentional Theory of Cinematic Continuity (AToCC) (Smith, 2012a, 2012b), these effects are explained as *postdiction*, the backwards inference of the details of the event after it has begun rather than predictive inference (Smith & Martin-Portugues Santacreu, 2017). Whether this absence of predictive inference is specific to the fast-paced sequences used in these studies is not currently known. In fact, such postdictive inferences are very similar to the bridging inferences that have shown to be commonly drawn in picture story studies (Hutson et al., 2018; Magliano et al., 2016). Predictive inferences do obviously occur in film, with many having been intentionally targeted by the filmmakers through filmmaking techniques (Magliano, Dijkstra, & Zwaan, 1996). Thus, it is possible that postdictive bridging inferences are more commonly generated during film viewing than predictive inferences, which appears to also be the case with narrative text (Graesser et al., 1994; Magliano et al., 1996).

However, not all cuts are missed, and their rate of detection is in proportion to the number of spatiotemporal and semantic features changed across the cut (Smith & Henderson, 2008; Smith & Martin-Portugues Santacreu, 2017) and object changes will be noticed if they change meaning, even if the changes are relatively small (Sampanes et al., 2008). In support of this, recent eye movement evidence indicates that low-level visual salience does not entirely account for gaze biases toward continued scene content across cuts—instead, memory-guided attention facilitates the deployment of attention but only if the viewer is actively tracking scene content (Valuch, König, & Ansorge, 2017). Whether such active tracking occurs automatically during visual narrative viewing is currently unknown. However, it is worth noting that drawn American visual narratives often circumvent this processing, by first introducing an environment early on in a sequence, and then leaving out the background entirely in later panels, though this intuition should be tested with corpus analyses.

Within the SPECT framework, we would suggest that important event indices are tracked by viewers across shots and cuts, interacting with visual salience to guide attention and gaze, and allowing changes to important semantic features of a scene to be detected (e.g., entities or actions that could change the goals of a protagonist in the visual narrative) but allowing unimportant features to pass unnoticed. Indeed, what constitutes an important event index has been the subject of much study in the event perception literature. Studies analyzing the likelihood of discontinuities in particular feature dimensions being perceived as event boundaries during film viewing have revealed that discontinuities of the goal of characters trump space and time (Magliano & Zacks, 2011). Exactly what information is used to construct and maintain a representation of *action,* or to detect changes to it, is currently unclear and will require further study in terms of the stages of processing outlined by SPECT.

4.1.2. Conclusions regarding information extraction and event model construction

Thus far, the studies by Larson (Larson, 2012; Larson & Lee, 2015) and Smith and Loschky (in press) have shown how rapid scene categorization processes, typically investigated by scene perception researchers, interact with higher-level event model processes, such as laying the foundation and mapping, typically studied by discourse comprehension researchers. These studies have shown evidence for a temporal order of processing event indices in which the spatiotemporal context is processed earlier than actions, with the former priming the latter (Larson, 2012; Larson & Lee, 2015). They have also shown that such spatiotemporal contexts can prime each other when encountered in sequential visual narratives (Smith & Loschky, in press). Further research is needed to investigate the temporal order of information extraction of the full range of key event indices across multiple fixations while viewing visual narratives. Other studies of change blindness and edit blindness while people watch films, however, raise questions about how much information viewers encode while viewing visual narratives (Levin & Simons, 1997; Smith & Henderson, 2008; Smith & Martin-Portugues Santacreu, 2017). A testable hypothesis consistent with SPECT is that viewers will detect those changes that change important event indices in the current event model or, to a lesser degree, recently stored event models (Sampanes et al., 2008).

4.2. The relationship between event model construction and attentional selection

SPECT assumes that not only do front-end information extraction and attentional selection processes affect back-end event model building, but also that back-end event model building processes affect front-end processes, such as attentional selection. We have conducted a series of studies that have been motivated by this general assumption and have explored whether and how event model building affects attentional selection. However, we have found evidence suggesting that the nature of this relation may vary as a function of whether narratives are static (comics, pictures stories) or dynamic (TV shows, videos, and films). Specifically, it seems that dynamic visual narratives, such as films, exert quite a bit of exogenous control over attentional selection, as measured by eye movements, and thus they may not afford much influence of the event model. This may be due to the fact that dynamic visual narratives (by definition) include motion, which is the single strongest stimulus feature for predicting eye movements and guiding attentional selection (Carmi & Itti, 2006; Mital et al., 2010). Conversely, because static narratives lack motion, and reading is self-paced, it seems that they may afford more endogenous influences on attentional selection via the back-end event model.

First consider static sequential picture stories. Magliano et al. (2016) had viewers read six wordless “Boy, Dog, Frog” stories. In each story, as illustrated in Fig. 1A, the authors identified three-image sequences that showed a beginning-state (e.g., Boy running down a hill), a bridging-event (e.g., Boy tripping over tree branch), and an end-state (e.g., Boy face first in the pond). As shown in Fig. 1A versus 1B, Magliano et al. (2016) manipulated whether the bridging-event image was present or not. When the bridging-event image was absent, viewers would need to generate a bridging inference in order to map event indices from the end-state picture onto their event model based on the beginning-state image. Magliano et al. (2016) found direct evidence of this in a pilot study in which they asked viewers to read the wordless picture stories on a computer screen, one image at a time, and do a think aloud after each end-state image. As predicted if an inferred bridging event was more highly activated in WM than an actually viewed bridging event, the authors found that participants were more likely to mention the bridging event in the absent condition than the present condition. In a follow-up study, the authors dropped the think-aloud task, and simply had viewers read the wordless picture stories at their own pace, while their viewing times were recorded. Consistent with the hypothesis that viewers were generating bridging inferences, the authors found that viewing times were longer when the bridging-event images were absent than when they were present.

Hutson et al. (2018) carried out a follow-up study that investigated more precisely why viewing times were longer in the bridging-state absent condition. They measured viewers' eye movements and asked whether viewing time differences were due to differences in either mean fixation durations or the mean number of fixations. They found that there were no differences in mean fixation durations, but there were approximately 20% additional fixations in the bridging-state absent condition relative to the bridging-event present condition. This suggested that, rather than the bridging event generation requiring further internal processing (during fixations), it may have required gathering additional information (in extra fixations). Thus, Hutson et al. (2018) empirically identified regions of the pictures that were informative for generating the bridging inference when the bridging-state picture was absent. Consistent with the hypothesis that viewers would preferentially fixate image regions that were more informative for generating the bridging inference, they found that the inferential-informativeness of image regions was more strongly correlated with the likelihood of eye fixations falling within them in the bridging-event absent condition. These data demonstrate that processes that support constructing the event model can influence attentional selection in scenes. Specifically, when visual narrative readers detect that they need to generate an inference to support the mapping process, their attentional selection system is engaged to support constructing that inference. Presumably, each fixation to support a bridging inference engages in information extraction, and that process continues until either (a) sufficient knowledge in semantic LTM is activated to support generating the inference, or (b) the viewer decides that the information is insufficient. The coordination of information extraction and attentional selection to support bridging inference generation warrants further investigation.

The story is much different in the context of film, likely because, as noted above, there are stronger exogenous features that attract attention. SPECT assumes that the event model will have less of an impact on attentional selection under such conditions. We have conducted a series of studies that have shown that the nature of the event model has a real but relatively small impact on attentional selection. Consider the film clip narrative sequence from James Bond *Moonraker* used in Loschky et al. (2015) illustrated in Fig. 8. This clip was chosen because Magliano et al. (1996) found that the use of cross cutting in shots 3–6 (alternating shots between two locations, in this case, a man in free fall and a circus tent) engendered a similar predictive inference, namely "the man will fall on the circus tent," across most viewers. Loschky et al. (2015) varied whether participants saw the prior 2 minutes of movie context leading up to this scene, and they found that participants in the "No-context" condition were less likely to generate the predictive inference than in the prior exposure ("Context") condition. Thus, this manipulation changed viewers' event models. However, when we measured viewers' eye movements as they watched the film clips, their gaze behavior indicated a high level of attentional synchrony both within and across the Context and No-context conditions. Only in a shot that had essentially no motion (Fig. 8, Shot 4), in which viewers were free to explore the shot of the circus tent, did we find gaze differences across the two context conditions. Thus, the nature of the event model appeared to have only a small effect on attentional selection, at least in the context of this film clip. We dubbed this phenomenon *the tyranny of film*, because despite large differences in viewers' understanding, there were small differences in attentional selection, assumedly due to the power of the film stimulus in guiding their attention.

Hutson et al. (2017) further investigated whether such tyranny of film on attentional selection, found in a highly edited film clip, would operate in a film clip with no editing. Given that editing practices are designed to influence attentional selection (Smith, 2012a), perhaps a lack of editing would minimize the tyranny of film. Hutson et al. (2017) explored this possibility by using the opening scene from *Touch of Evil* (Welles & Zugsmith, 1958), which consists of a single continuous long shot (i.e., no cuts), showing two couples navigating the streets of a Mexico/US border town. As shown in Fig. 9, the opening segment shows a man setting a time bomb and putting it in the trunk of a car. Soon after, a couple who owns the car unwittingly gets into the car and drives away. The couple in the car then passes a walking couple on the street. Hutson et al. reasoned that, since the bomb has tremendous causal power in the event models of viewers who know about it, if viewers had no knowledge of the bomb, they would be less likely to fixate on the car. Thus, in Experiment 1, Hutson et al. manipulated whether participants saw the bomb placed in the car trunk (Context condition) or not (No-context condition). Similarly to Loschky et al (2015), this context manipulation strongly affected participants' predictions of what would happen next at the end of the clip (e.g., either "the car will explode," or "the two couples will have dinner together"). This showed that the heavy-handed context manipulation indeed dramatically changed the nature of viewers' event models for the movie clip. Surprisingly, however, Hutson et al. (2017) found equal proportions of fixations on the car in both the Context (bomb-present) and No-context (bomb-absent) conditions. Thus, this showed that the tyranny of film was still operating even without film editing. Apparently, the structure of the long shot was such that the movement of the car exerted exogenous control of attentional selection.

In Experiment 2 of Hutson et al. (2017), the No-context condition began watching the clip when only the walking couple was on screen; thus, viewers would not consider the couple in the car as protagonists. When the walking couple passed the temporarily parked car, this was the first time viewers in the No-context condition saw it, and they were much less likely to fixate on it than those in the Context condition. Assumedly, this was because the No-context viewers perceived the car as background, whereas the Context condition viewers knew about the bomb, and also treated the couple in the car as protagonists/agents. Hutson et al. called this the *agent effect*. However, once the car began to move again, viewers in both context conditions fixated on the car equally, regardless of knowledge of the bomb. Thus, as in Loschky et al. (2015), the effect of the event model on attentional selection was real, but small.

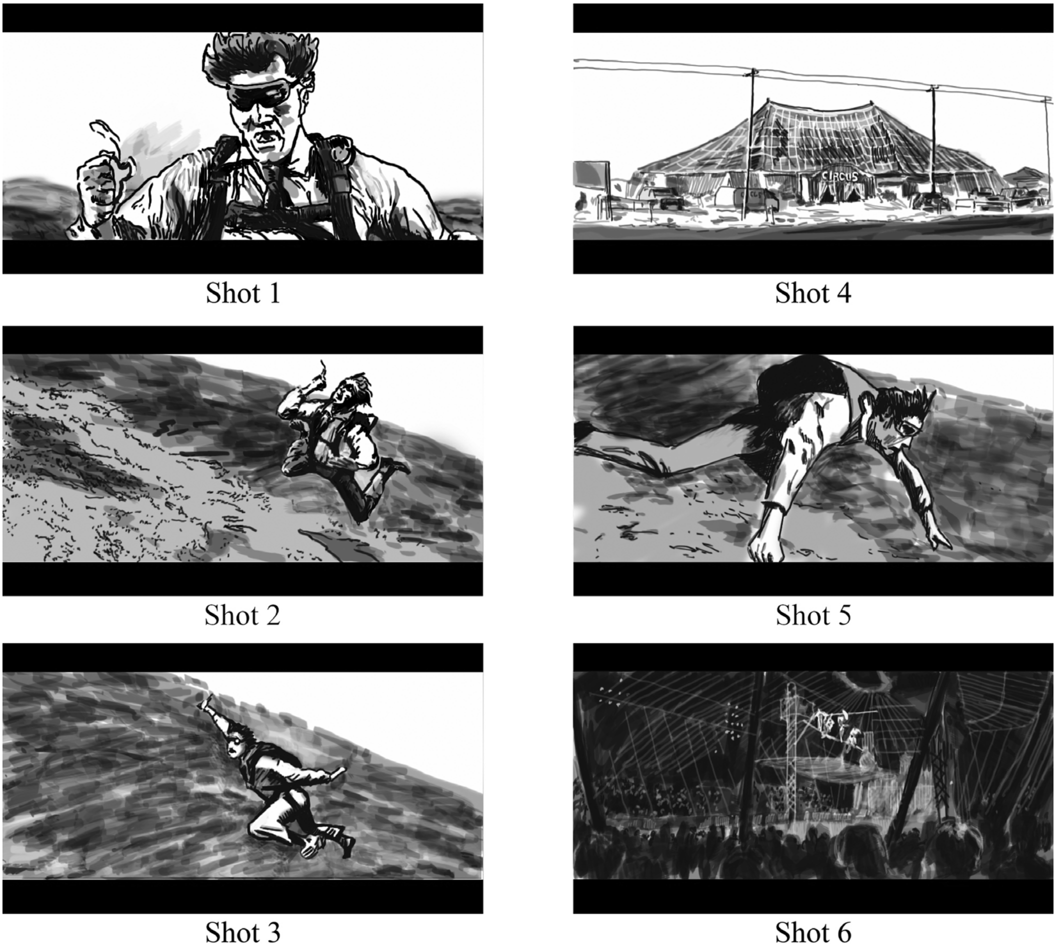

Fig. 8. Drawings of six frames from six sequential shots from James Bond *Moonraker* (Broccoli & Gilbert, 1979). (Reproduced with permission of Loschky et al., 2015.)

In a further control experiment, Hutson et al. (2017) found they could reduce the tyranny of film by using a task that directed viewers' volitional attention away from the narrative events in the shot, namely asking viewers to prepare to draw a map from memory of the spatial environment in the film clip. As noted earlier, SPECT assumes that this requires the use of effortful volitional executive processes (see Fig. 2). Additionally, Hutson et al. (2017) compared the levels of attentional synchrony found in the highly edited shot sequence of James Bond *Moonraker* used in Loschky et al. (2015) versus the continuous long shot from *Touch of Evil*. As predicted, the levels of attentional synchrony were less in the continuous long shot than in the highly edited sequence. This analysis suggests that there may be differences in the extent to which features of dynamic visual narratives affect the relationship

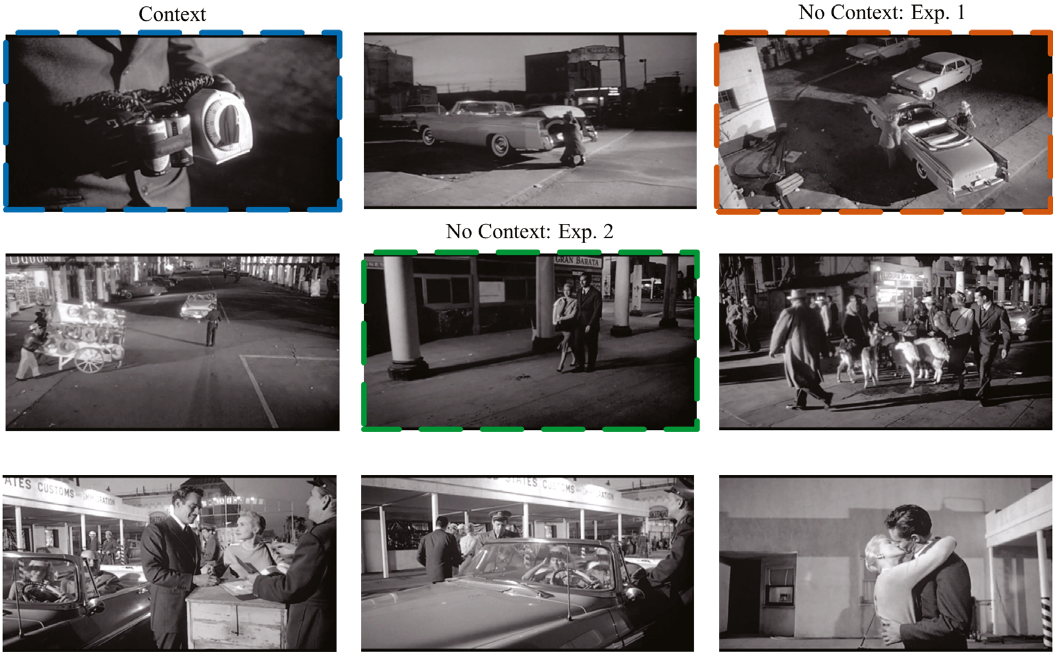

Fig. 9. Nine frames from the opening long shot of *Touch of Evil* (Welles & Zugsmith, 1958) and the experimental conditions used in Hutson et al. (2017). The blue dashed outline indicates the video starting point in the Context Condition (Experiments 1 and 2); the orange outline shows the starting point for the No-context condition (Experiment 1); the green outline shows the starting point for the No-context condition (Experiment 2). (Published with permission of Hutson et al. [2017]).

between the event model and attentional selection, and more research is warranted to address this issue.

4.2.1. Conclusions regarding the relationship between event model construction and attentional selection

The studies described in this section have shown effects of event model building processes on attentional selection in visual narratives, including both static picture stories (Hutson et al., 2018) and movie clips (Hutson et al., 2017; Loschky et al., 2015). However, these effects appeared stronger in the static picture stories than in the film clips. This has led us to modify the assumption of SPECT that back-end and front-end processes have bi-directional influences. Specifically, we have added the further assumption that the influence of event model building processes on attentional selection is moderated by whether a visual narrative is static or dynamic. Nevertheless, this conjecture is in need of more direct tests. More generally, the implications of differences between media in terms of affording control over attentional selection need to be carefully explored.

Related to the above, there are likely trade-offs involved in the tyranny of film. Filmmakers can utilize the properties of film to direct viewers' attention to specific portions of the screen, which should affect the process of information extraction, which then affects passive knowledge activation, which in turn affects back-end event model building processes (e.g., Kintsch, 1988). However, the lack of opportunities for regressive eye movements in film that might support comprehension repair is a price that is paid for the tyranny and the lack of self-paced control. SPECT provides a motivation for research that addresses these important issues.

5. Discussion

The intent of SPECT is to explain how visual narratives are processed and understood from early perceptual processes to relatively late processes that support event model building. In doing so, SPECT integrates previously separate research domains for visual narrative perception and comprehension, which has rarely occurred in research on text comprehension (for exceptions, see the computational models of reading, e.g., SWIFT: Engbert, Nuthmann, Richter, & Kliegl, 2005; EZ-Reader: Reichle, Rayner, & Pollatsek, 1999). Our intent in this article was to inspire future research on the processing and comprehension of visual narratives that explores the interplay between multiple levels of front- and back-end cognitive processing. We have made a case that the framework has been invaluable in guiding our program of research on visual narrative processing, and we contend that new research questions are afforded by it.

5.1. Future research questions

As noted above, we have been using this framework to guide a program of research. However, that program is by no means exhaustive in addressing the important research questions that can be derived by SPECT. In this section, we discuss pressing questions that we believe should be addressed in order to further illustrate the utility SPECT as a theoretical framework.

An important unanswered question raised by SPECT regarding front-end information extraction is the temporal order of information extraction for event indices (e.g., spatiotemporal framework, agents, objects, actions, goals of agents) across multiple eye movements. As noted earlier, Larson (Larson, 2012; Larson & Lee, 2015) has begun to answer this question by showing that the spatiotemporal event index is extracted prior to the action underlying an event. However, further, more detailed investigations are needed to determine when entities, their inferred goals, and inferred causal relationships are extracted across multiple eye fixations. It seems likely that not only the spatiotemporal context, and actions, but also the entities of agents and patients are among the first event indices to be extracted in a new event model. Furthemore, given that identifying goals and causal relationships require more inferential processes, these event indices are likely extracted and generated in the event model later. However, tests of these hypotheses are needed. Doing so would elucidate the role of front-end information extraction during single fixations in event model building across multiple fixations in WM.

As noted above, a key research question suggested by SPECT is whether static versus dynamic visual narratives differ in the degree to which the event model influences attentional selection. To answer this question will require at least two things: (a) visual narratives in which manipulations of viewers' event models influence attentional selection, and (b) versions of those narratives that primarily differ in terms of the static versus dynamic distinction. Meeting both criteria is non-trivial. However, answering this question will more broadly help answer the question of the conditions under which the viewer's event model influences attentional selection in visual narratives.

A further key unanswered research question suggested by SPECT is whether and how the back-end process of shifting to build a new event model affects the front-end process of attentional selection (but see Huff, Papenmeier, & Zacks, 2012). Research has shown better memory for event boundaries than middles (Huff, Meitz, & Papenmeier, 2014; Swallow et al., 2009), suggesting that attentional selection is affected by shifting. Interestingly, it is possible that attention is heightened at event boundaries (Huff et al., 2014; Swallow et al., 2009) or, conversely, that it is diminished (Huff et al., 2012). This apparent contradiction may be resolved by other results showing that gaze patterns change just before and after event boundaries (Eisenberg & Zacks, 2016; Smith, Whitwell, & Lee, 2006), consistent with the ambient-to-focal eye movement shift—namely, attention may expand and contract over time near event boundaries (Ringer 2016; Ringer, 2018). This warrants further research to clarify these relationships.

We invite the reader to identify questions that have not yet been pursued. Such efforts are essential for revisions to SPECT that would allow it to become a formalized and implementable model. Moreover, in the pursuit of such research, we acknowledge that alternative and perhaps contradictory frameworks could emerge. We see that possibility as healthy and indicative of the study of visual narratives as being a vibrant and growing area of research.

5.2. *Future computational and neurophysiological tests of SPECT*

SPECT is not yet a formally complete cognitive model of visual narrative processing, primarily because many assumptions of the model remain to be empirically validated, as laid out above. However, while it has not yet been computationally implemented, our goal is to refine the model such that it eventually can be. Thus, future studies should develop and test computational approximations of key elements of SPECT. For front-end mechanisms, there are already deep neural networks that can extract the event indices needed for laying the foundation of an event model (e.g., locations, people, animals, objects, and actions) from video (Du, El-Khamy, Lee, & Davis, 2017; Hoai, Lan, & De la Torre, 2011; LeCun, Bengio, & Hinton, 2015; Manohar, Sharath Kumar, Kumar, & Rani, 2019; Zhou, Lapedriza, Khosla, Oliva, & Torralba, 2018). There are also neural networks for attentional selection (Adeli & Zelinsky, 2018; Huang, Shen, Boix, & Zhao, 2015). For the back-end, formal ontologies use techniques such as description logics to represent events (Baader & Nutt, 2003; Neumann & Möller, 2008). Inferential processes based on event representations can be modeled in terms of Bayesian weights for likely inferences (Bateman & Wildfeuer, 2014; Grosz & Gordon, 1999).[5] A key challenge is to link the front-end event index outputs in ways that are usable by the back-end ontologies.

The neurophysiological foundations of SPECT are based on numerous related but non-visual-narrative-based studies. The distinction between the front-end processes of information extraction and attentional selection is strongly supported by their implementation within different functional brain networks, and having differentiable time courses. Front-end information extraction of foundation event indices (i.e., locations, people, animals, actions, and objects) is extremely rapid, with perceptual decisions occurring within 150–225 ms post-stimulus, as shown by EEG and MEG studies (Cichy, Khosla, Pantazis, Torralba, & Oliva, 2016; Greene & Hansen, 2018; Ramkumar, Hansen, Pannasch, & Loschky, 2016; VanRullen & Thorpe, 2001). Such event indices can be decoded from fMRI brain activity within functionally defined areas for locations (Walther, Caddigan, Fei-Fei, & Beck, 2009), objects (Majaj, Hong, Solomon, & DiCarlo, 2015), and actions (Gallivan & Culham, 2015). Front-end attentional selection is also extremely fast, with ventral stream neurons activating roughly 100 ms before the eyes fixate an object of interest (Sheinberg & Logothetis, 2001). Very fast stimulus saliency effects on attentional selection are controlled by the superior colliculus (Boehnke & Munoz, 2008), and slower back-end influences are likely controlled by the fronto-parietal and fronto-temporal networks (Baldauf & Desimone, 2014). Knowing these basic facts can guide research to test front-end hypotheses of SPECT.

The functional distinction of back-end from front-end processes is also strongly supported by their having different time courses and involving different brain networks. EEG research has shown that mapping processes, such as when the same people and/or locations recur across multiple panels in a visual narrative, elicit decreased N400 amplitude (roughly 300–500 ms post-stimulus) (Cohn, Paczynski, Jackendoff, Holcomb, & Kuperberg, 2012). Importantly, consistent with SPECT, this time course is later than the 150–225 ms needed to extract an event index (e.g., the person, the location) and is operating over multiple items in WM. Other EEG studies have shown that mapping processes such as updating the event model with new event indices and bridging inference generation occur even later, eliciting the P600 (roughly 400–900 ms post-stimulus) (Cohn & Kutas, 2015). Likewise, consistent with SPECT's assumption that mapping and shifting are separate processes, fMRI studies have shown that they involve separate brain regions (Ezzyat & Davachi, 2011). SPECT further argues that shifting at event boundaries leads the event model in WM to be stored in LTM. Consistent with this claim, fMRI studies have shown event boundaries lead to activity in parietal and posterior medial cortex being temporarily synchronized (Baldassano et al., 2017; Ezzyat & Davachi, 2011). Such interactions between brain areas involved in front-end and back-end processes are critical predictions of SPECT, but the bidirectionality of these predicted interactions requires considerable further neuroimaging support.

Additionally, SPECT assumes that event segmentation is similar across different representational formats, which has been supported by fMRI studies showing the same posterior-medial network (Inhoff & Ranganath, 2017) being engaged when reading written narratives (Baldassano et al., 2017; Speer, Zacks, & Reynolds, 2007) and watching visual narratives (Baldassano et al., 2017; Kurby & Zacks, 2018; Zacks et al., 2001). This further supports the relevance of research from outside of the context of visual narratives for establishing the neurophysiological bases of SPECT. Indeed, processing visual narratives likely involves a complex coordination of neurophysiological systems that are both domain-specific (Cohn & Maher, 2015) and domain-general (Cohn, 2019a, 2019b).

Future studies should use behavioral and computational modeling methods together with neuroimaging methods to test hypotheses of SPECT in terms of their time course or functional differentiation. Furthermore, while research using non-visual-narrative materials is valuable for understanding the neurophysiological bases of SPECT, there is behavioral evidence, and some neurophysiological evidence, that visual narratives require processing unique to each medium (Cohn & Ehly, 2016; Cohn & Maher, 2015; Smith, 2012a; Smith, Levin, & Cutting, 2012) and may require specialized literacy skills (Cohn, 2019b; Cohn & Magliano, in press; Schwan & Ildirar, 2010). Thus, future tests of the neurophysiological bases of SPECT should prioritize visual narratives.

5.3. Limitations of SPECT

What is missing from SPECT? One obvious limitation of SPECT is that it does not specify how prior world knowledge supports the comprehension of visual narratives. In contrast, theories of text comprehension focus on how semantic knowledge is activated and integrated into a mental model for a text (McNamara & Magliano, 2009). The complexities of exploring how front-end process support mental model construct are such that, at this juncture, we deem this to be a necessary omission.

As described here, SPECT principally applies to traditional non-interactive media (though reading comics and picture stories allows self-pacing). SPECT does not account for visual narrative experience in which the viewer is also an active participant, such as video game and virtual reality experiences. Given that first-person experiences are processed in similar fashion to narrative experiences (Magliano, Radvansky, Forsythe, & Copeland, 2014), SPECT should be able to accommodate these experiences. However, the fact that one is an active agent in many of these contexts will obviously have implications for attentional selection. SPECT neglects the rich and important social aspects (e.g., communal viewing at a cinema, or a parent reading a picture book to their child) and emotional aspects (i.e., the affective profile of joy and despair so important to narrative arcs) of visual narratives. This is a systemic issue with many theories of comprehension, but it does not imply that these processes are unimportant for comprehension.

Probably the most important current omission is that SPECT specifically describes the relationship between visual processing and event model construction, but it does not describe how written or auditory information (linguistic and non-linguistic) contributes to the understanding of visual narratives. Cohn similarly does not specify how linguistic information is processed in his theory of visual narrative processing (Cohn, 2013a), though he acknowledges the importance of understanding the relationship between text and images (Cohn, 2016; Manfredi, Cohn, & Kutas, 2017) and sounds and images (Manfredi, Cohn, De Araújo Andreoli, & Boggio, 2018) in sequential visual narratives in conveying meaning. Auditory information is vital to the practices of storytelling in filmmaking (Batten & Smith, 2018; Bordwell, 1985) and representations of speech, thought, narration, and sound effects are vital to storytelling in comics (Cohn, 2013a). Moreover, when comic panels contain a large amount of text, readers allocate considerable attentional resources to process the text, and there is some suggestion that image content may be processed in parafoveal vision (Laubrock et al., 2018). Furthermore, in film, auditory and linguistic content support inference processes (Magliano et al., 1996). However, given the complexities of understanding the relationship between visual perception and event cognition, we argue that this is a necessary omission at this juncture.

5.4. Conclusion

With SPECT, we have taken the first steps toward outlining a comprehensive cognitive framework for visual narrative processing which extends from momentary attentional selection and information extraction from visual images to the longer-scale creation and maintenance of event models in WM and LTM. This theoretical framework incorporates contemporary theories of all of these stages of visual scene perception, event perception, and narrative comprehension, but by applying SPECT to complex visual narratives, a number of important ruptures, inconsistencies, and gaps in our understanding have emerged. Most important, as previously stated in relation to film (Smith, Levin, et al., 2012), by theorizing about and studying how we process visual narratives, we learn more about how we perceive and make sense of the real world.

Notes

1. Here “fixation” refers to all periods of low-velocity gaze stability relative to elements within a scene, whether the element is static, moving (e.g., a driving car), or moving on the retina due to head/body movement as we move our head in front of an image (e.g., the page of a comic). For further discussion of “fixation” definitions, see Hessels et al. (2018).
2. Note that much of this work was done on the topic of transsaccadic memory, namely memory across a saccade. That work eventually determined that the contents of transsaccadic memory are in short-term memory or WM (Irwin, 1996).
3. The regions of interest include speech balloons, though language, narrowly defined, is beyond the scope of SPECT, so we will not discuss the speech balloons or dialog here.
4. Research has shown comic readers’ first fixations within panels were more likely sent to characters than to background elements (Laubrock et al., 2018), which might suggest that characters are processed before backgrounds. This is consistent with the *person bias*, namely if an image of a person is present in a photograph, viewers’ first fixation usually goes to that person (Fletcher-Watson et al., 2008; Humphrey & Underwood, 2010). However, Laubrock et al. (2018, p. 249) point

out that comic readers can likely recognize the gist of the background within their first fixation of a comic panel using their peripheral vision (i.e., without fixating it). Specifically, studies have shown that, within a single fixation on a photograph, viewers can accurately categorize the scene background (e.g., beach vs. mountain vs. street vs. bedroom) using only their peripheral vision (Boucart, Moroni, Thibaut, Szaffarczyk, & Greene, 2013; Larson & Loschky, 2009; Loschky, Szaffarczyk, Beugnet, Young, & Boucart, 2019). Thus, it is currently unclear whether the category of a background or of a character is processed earlier while viewing visual narratives.

5. We thank John A. Bateman and William H. Hsu for help in conceptualizing computational modeling for these back-end processes and providing relevant references.

References

Adeli, H., & Zelinsky, G. J. (2018). Deep-BCN: Deep networks meet biased competition to create a brain-inspired model of attention control. Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Salt Lake City, Utah.

Andreu-Sánchez, C., Martín-Pascual, M. Á., Gruart, A., & Delgado-García, J. M. (2018). Chaotic and fast audiovisuals increase attentional scope but decrease conscious processing. *Neuroscience*, *394*, 83–97. https://doi.org/10.1016/j.neuroscience.2018.10.025

Baader, F., & Nutt, W. (2003). Basic description logics. Paper presented at the Description logic handbook.

Baggett, P. (1979). Structurally equivalent stories in movie and text and the effect of the medium on recall. *Journal of Verbal Learning & Verbal Behavior*, *18*(3), 333–356.

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. https://doi.org/10.1016/j.neuron.2017.06.041

Baldauf, D., & Desimone, R. (2014). Neural mechanisms of object-based attention. *Science*, *344*(6182), 424–427. https://doi.org/10.1126/science.1247003

Ball, K. K., Beard, B. L., Roenker, D. L., Miller, R. L., & Griggs, D. S. (1988). Age and visual search: Expanding the useful field of view. *Journal of the Optical Society of America*, *5*(12), 2210–2219. https://doi.org/10.1364/josaa.5.002210

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

Bateman, J. A., & Wildfeuer, J. (2014). A multimodal discourse theory of visual narrative. *Journal of Pragmatics*, *74*, 180–208.

Batten, J. P., & Smith, T. J. (2018). Looking at sound: Sound design and the audiovisual influences on gaze. In T. Dwyer, C. Perkins, S. Redmond, & J. Sita (Eds.), *Seeing into screens: Eye tracking the moving image* (pp. 85–102). London: Bloomsbury.

Belopolsky, A. V., Kramer, A. F., & Theeuwes, J. (2008). The role of awareness in processing of oculomotor capture: Evidence from event-related potentials. *Journal of Cognitive Neuroscience*, *20*(12), 2285–2297. https://doi.org/10.1162/jocn.2008.20161

Boehnke, S. E., & Munoz, D. P. (2008). On the importance of the transient visual response in the superior colliculus. *Current Opinion in Neurobiology*, *18*(6), 544–551. https://doi.org/10.1016/j.conb.2008.11.004

Bordwell, D. (1985). *Narration in the fiction film*. Madison: University of Wisconsin Press.

Bordwell, D., & Thompson, K. (2003). *Film art: An introduction*. New York: McGraw-Hill.

Borji, A., & Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(1), 185–207. https://doi.org/10.1109/tpami.2012.89

Boucart, M., Moroni, C., Thibaut, M., Szaffarczyk, S., & Greene, M. (2013). Scene categorization at large visual eccentricities. *Vision Research*, *86*, 35–42. https://doi.org/10.1016/j.visres.2013.04.006

Britt, M. A., Rouet, J.-F., & Durik, A. M. (2018). *Literacy beyond text comprehension: A theory of purposeful reading*. New York: Routledge/Taylor & Francis.

Broccoli, A. R. P., & Gilbert, L. D.(Writers). (1979). *Moonraker* [Film]. USA: CBS/Fox Video.

Calvo, M. G., Nummenmaa, L., & Hyönä, J. (2007). Emotional and neutral scenes in competition: Orienting, efficiency, and identification. *The Quarterly Journal of Experimental Psychology*, *60*(12), 1585–1593. https://doi.org/10.1080/17470210701515868

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

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*, 27755. https://doi.org/10.1038/srep27755

Cohn, N. (2013a). *The visual language of comics: Introduction to the structure and cognition of sequential images*. London: Bloomsbury.

Cohn, N. (2013b). Visual narrative structure. *Cognitive Science*, *37*(3), 413–452. https://doi.org/10.1111/cogs.12016

Cohn, N. (2016). A multimodal parallel architecture: A cognitive framework for multimodal interactions. *Cognition*, *146*, 304–323.

Cohn, N. (2019a). Visual narratives and the mind: Comprehension, cognition, and learning. In K. D. Federmeier& D. M. Beck (Eds.), *Psychology of learning and motivation* (Vol. 70, pp. 97–127). Cambridge, MA: Academic Press.

Cohn, N. (2019b). Your brain on comics: A cognitive model of visual narrative comprehension. *Topics in Cognitive Science*, https://doi.org/10.1111/tops.12421

Cohn, N., & Ehly, S. (2016). The vocabulary of manga: Visual morphology in dialects of Japanese Visual Language. *Journal of Pragmatics*, *92*, 17–29. https://doi.org/10.1016/j.pragma.2015.11.008

Cohn, N., & Kutas, M. (2015). Getting a cue before getting a clue: Event-related potentials to inference in visual narrative comprehension. *Neuropsychologia*, *77*, 267–278.

Cohn, N., & Kutas, M. (2017). What's your neural function, visual narrative conjunction? Grammar, meaning, and fluency in sequential image processing. *Cognitive research: principles and implications*, *2* (1), 27.

Cohn, N., & Magliano, J. P. (in press). Visual narrative research: An emerging field in cognitive science. *Topics in Cognitive Science*.

Cohn, N., & Maher, S. (2015). The notion of the motion: The neurocognition of motion lines in visual narratives. *Brain Research*, *1601*, 73–84. https://doi.org/10.1016/j.brainres.2015.01.018

Cohn, N., Paczynski, M., Jackendoff, R., Holcomb, P. J., & Kuperberg, G. R. (2012). (Pea)nuts and bolts of visual narrative: Structure and meaning in sequential image comprehension. *Cognitive Psychology*, *65*(1), 1–38. https://doi.org/10.1016/j.cogpsych.2012.01.003

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87–185.

DeAngelus, M., & Pelz, J. (2009). Top-down control of eye movements: Yarbus revisited. *Visual Cognition*, *17*(6), 790–811.

Deubel, H., & Schneider, W. X. (1996). Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research*, *36*(12), 1827–1837. https://doi.org/10.1016/0042-6989(95)00294-4

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

Dorr, M., Martinetz, T., Gegenfurtner, K. R., & Barth, E. (2010). Variability of eye movements when viewing dynamic natural scenes. *Journal of Vision*, *10*(10), 1–28. https://doi.org/10.1167/10.10.28

Du, X., El-Khamy, M., Lee, J., & Davis, L. (2017). Fused DNN: A deep neural network fusion approach to fast and robust pedestrian detection. Paper presented at the IEEE Winter Conference on Applications of Computer Vision (WACV), March 24–31, 2017.

Eckstein, M. P., Drescher, B. A., & Shimozaki, S. S. (2006). Attentional cues in real scenes, saccadic targeting, and Bayesian priors. *Psychological Science*, *17*(11), 973–980. https://doi.org/10.1111/j.1467-9280.2006.01815.x

Eisenberg, M. L., & Zacks, J. M. (2016). Ambient and focal visual processing of naturalistic activity. *Journal of Vision*, *16*(2), 1–12. https://doi.org/10.1167/16.2.5

Engbert, R., Nuthmann, A., Richter, E. M., & Kliegl, R. (2005). SWIFT: A dynamical model of saccade generation during reading. *Psychological Review*, *112*(4), 777–813.

Eriksen, C. W., & Yeh, Y. Y. (1985). Allocation of attention in the visual field. *Journal of Experimental Psychology: Human Perception & Performance*, *11*(5), 583–597. https://doi.org/10.1037/0096-1523.11.5.583

Ezzyat, Y., & Davachi, L. (2011). What constitutes an episode in episodic memory? *Psychological Science*, *22*(2), 243–252.

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*(1:10), 1–29. 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.

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 salience 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

Fussell, D., & Haaland, A. (1978). Communicating with pictures in Nepal: Results of practical study used in visual education. *Journal of Educational Broadcasting International*, *11*(1), 25–31.

Gallivan, J. P., & Culham, J. C. (2015). Neural coding within human brain areas involved in actions. *Current Opinion in Neurobiology*, *33*, 141–149.

Garcia-Diaz, A., Fdez-Vidal, X. R., Pardo, X. M., & Dosil, R. (2009). *Decorrelation and distinctiveness provide with human-like saliency*. Berlin: Springer-Verlag.

Gernsbacher, M. A. (1985). Surface information loss in comprehension. *Cognitive Psychology*, *17*(3), 324–363. https://doi.org/10.1016/0010-0285(85)90012-x

Gernsbacher, M. A. (1990). *Language comprehension as structure building* (Vol. xi). Hillsdale, NJ: Erlbaum.

Gernsbacher, M. A. (1997). Coherence cues mapping during comprehension. In J. Costermans & M. Fayol (Eds.), *Processing interclausal relationships: Studies in the production and comprehension of text* (Vol. x, pp. 3–21). Mahwah, NJ: Erlbaum.

Gernsbacher, M. A., Varner, K. R., & Faust, M. E. (1990). Investigating differences in general comprehension skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(3), 430–445.

Glanemann, R. (2008). To see or not to see–Action scenes out of the corner of the eye. PhD Dissertation, University of Münster, Münster.

Graesser, A. C., & Clark, L. F. (1985). *Structures and procedures of implicit knowledge: Advances in discourse processes*. Santa Barbara, CA: Praeger.

Graesser, A. C., Millis, K. K., & Zwaan, R. A. (1997). Discourse comprehension. *Annual Review of Psychology*, *48*(1), 163–189. https://doi.org/10.1146/annurev.psych.48.1.163

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, *101*(3), 371–395.

Greene, M. R., & Hansen, B. C. (2018). Shared spatiotemporal category representations in biological and artificial deep neural networks. *PLOS Computational Biology*, *14*(7), e1006327. https://doi.org/10.1371/journal.pcbi.1006327

Greene, M. R., & Oliva, A. (2009). The briefest of glances: The time course of natural scene understanding. *Psychological Science*, *20*(4), 464–472.

Grosz, B. J., & Gordon, P. (1999). Conceptions of limited attention and discourse focus. *Computational Linguistics*, *25*(4), 617–624.

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.

Hayes, T. R., & Henderson, J. M. (2018). Scan patterns during scene viewing predict individual differences in clinical traits in a normative sample. *PLoS ONE*, *13*(5), e0196654. https://doi.org/10.1371/journal.pone.0196654

Henderson, J. M., Brockmole, J. R., Castelhano, 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 (Ed.), *Eye movements: A window on mind and brain* (pp. 537–562). Amsterdam: Elsevier.

Hessels, R. S., Niehorster, D. C., Nyström, M., Andersson, R., & Hooge, I. T. C. (2018). Is the eyemovement field confused about fixations and saccades? A survey among 124 researchers. *Royal Society Open Science*, *5*(8), 180502. https://doi.org/10.1098/rsos.180502.

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.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, *95*(4), 528–551.

Hoai, M., Lan, Z. Z., & De la Torre, F. (2011). Joint segmentation and classification of human actions in video. 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 3265–3272).

Hoffman, J. E., & Subramaniam, B. (1995). The role of visual attention in saccadic eye movements. *Perception & Psychophysics*, *57*(6), 787–795. https://doi.org/10.3758/BF03206794

Hollingworth, A. (2009). Memory for real-world scenes. In J. R. Brockmole (Ed.), *The visual world in memory* (pp. 89–119). Hove, UK: Psychology Press.

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

Huang, X., Shen, C., Boix, X., & Zhao, Q. (2015). Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. Paper presented at the Proceedings of the IEEE International Conference on Computer Vision.

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), 1–19. https://doi.org/10.1167/10.10.19

Hutson, J. P., Magliano, J. P., & Loschky, L. C. (2018). Understanding moment-to-moment processing of visual narratives. *Cognitive Science*, *42*, 2999–3033. https://doi.org/10.1111/cogs.12699

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. *Cognitive Research: Principles and Implications*, *2*(1), 46. https://doi.org/10.1186/s41235-017-0080-5

Ildirar, S., & Schwan, S. (2014). First-time viewers' comprehension of films: Bridging shot transitions. *British Journal of Psychology*, *106*, 133–151.

Inhoff, M. C., & Ranganath, C. (2017). Dynamic cortico-hippocampal networks underlying memory and cognition: The PMAT framework. In D. E. Hannula & M. C. Duff (Eds.), *The hippocampus from cells to* *systems: Structure, connectivity, and functional contributions to memory and flexible cognition* (pp. 559–589). Cham: Springer.

Irwin, D. E. (1996). Integrating information across saccadic eye movements. *Current Directions in Psychological Science*, *5*(3), 94–100.

Itti, L., & Koch, C. (2001). Computational modeling of visual attention. *Nature Reviews Neuroscience*, *2*(3), 194–203.

Kaakinen, J. K., Hyönä, J., & Keenan, J. M. (2003). How prior knowledge, WMC, and relevance of information affect eye fixations in expository text. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(3), 447.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, *95*(2), 163–182.

Kowler, E., Anderson, E., Dosher, B., & Blaser, E. (1995). The role of attention in the programming of saccades. *Vision Research*, *35*(13), 1897–1916. https://doi.org/10.1016/0042-6989(94)00279-u

Kurby, C. A., & Zacks, J. M. (2008). Segmentation in the perception and memory of events. *Trends in Cognitive Sciences*, *12*(2), 72.

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

Kurby, C. A., & Zacks, J. M. (2018). Preserved neural event segmentation in healthy older adults. *Psychology and Aging*, *33*(2), 232–245.

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.

Lang, A. (2000). The limited capacity model of mediated message processing. *Journal of Communication*, *50* (1), 46–70. https://doi.org/10.1111/j.1460-2466.2000.tb02833.x

Larson, A. M. (2012). Recognizing the setting before reporting the action: Investigating how visual events are mentally constructed from scene images. Ph.D. 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., Hendry, J., & Loschky, L. C. (2012). Scene gist meets event perception: The time course of scene gist and event recognition. *Journal of Vision*, *12*(9), 1077. https://doi.org/10.1167/12.9.1077

Larson, A. M., & Lee, M. (2015). When does scene categorization inform action recognition? *Journal of Vision*, *15*(12), 118–118. https://doi.org/10.1167/15.12.118

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.

Laubrock, J., Hohenstein, S., & Kummerer, M. (2018). Attention to comics: Cognitive processing during the reading of graphic literature. In A. Dunst, J. Laubrock, & J. Wildfeuer (Eds.), *Empirical comics research: Digital, multimodal, and cognitive methods* (pp. 239–263). New York: Routledge.

Le Meur, O., Le Callet, P., & Barba, D. (2007). Predicting visual fixations on video based on low-level visual features. *Vision Research*, *47*(19), 2483–2498. https://doi.org/10.1016/j.visres.2007.06.015

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444. https://doi.org/10.1038/nature14539

Levin, D. T., & Simons, D. J. (1997). Failure to detect changes to attended objects in motion pictures. *Psychonomic Bulletin & Review*, *4*(4), 501–506.

Liddell, C. (1997). Every picture tells a story—Or does it?: Young South African children interpreting pictures. *Journal of Cross-Cultural Psychology*, *28*(3), 266–283. https://doi.org/10.1177/0022022197283004

Long, D. L., Golding, J. M., & Graesser, A. C. (1992). A test of the on-line status of goal-related inferences. *Journal of Memory and Language*, *31*(5), 634–647. https://doi.org/10.1016/0749-596X(92)90032-S

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 (Ed.), *Empirical comics research: Digital, multimodal, and cognitive methods* (pp. 217–238). New York: Routledge.

Loschky, L. C., & Larson, A. M. (2010). The natural/man-made distinction is made prior to basic-level distinctions in scene gist processing. *Visual Cognition*, *18*(4), 513–536.

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., 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* (5:15), 1–21. https://doi.org/10.1167/19.5.15

Magliano, J. P., Clinton, J. A., O'Brien, E. J., & Rapp, D. N. (2018). Detecting differences between adapted narratives. In J. Laubrock, J. Wildfeuer, & A. Dunst (Eds.), *Empirical comics research: Digital, multimodal, and cognitive methods* (pp. 284–304). New York: Routledge.

Magliano, J. P., Dijkstra, K., & Zwaan, R. A. (1996). Generating predictive inferences while viewing a movie. *Discourse Processes*, *22*(3), 199–224.

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., 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., Loschky, L. C., Clinton, J. A., & Larson, A. M. (2013). Is reading the same as viewing? An exploration of the similarities and differences between processing text- and visually based narratives. In B. Miller, L. Cutting, & P. McCardle (Eds.), *Unraveling the behavioral, neurobiological, and genetic components of reading comprehension* (pp. 78–90). Baltimore, MD: Brookes.

Magliano, J. P., Miller, J., & Zwaan, R. A. (2001). Indexing space and time in film understanding. *Applied Cognitive Psychology*, *15*(5), 533–545.

Magliano, J. P., Radvansky, G. A., Forsythe, J., & Copeland, D. (2014). Event segmentation during first-person continuous events. *Journal of Cognitive Psychology*, *26*, 649–661.

Magliano, J. P., & Zacks, J. M. (2011). The impact of continuity editing in narrative film on event segmentation. *Cognitive Science*, *35*(8), 1489–1517. https://doi.org/10.1111/j.1551-6709.2011.01202.x

Magliano, J. P., Zwaan, R. A., & Graesser, A. C. (1999). The role of situational continuity in narrative understanding. In S. R. Goldman & H. van Oostendorp (Eds.), *The construction of mental representations during reading* (pp. 219–245). Mahwah, NJ: Erlbaum.

Majaj, N. J., Hong, H., Solomon, E. A., & DiCarlo, J. J. (2015). Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance. *Journal of Neuroscience*, *35*(39), 13402–13418. https://doi.org/10.1523/JNEUROSCI.5181-14.2015

Maljkovic, V., & Martini, P. (2005). Short-term memory for scenes with affective content. *Journal of Vision*, *5*(3), 6. https://doi.org/10.1167/5.3.6

Manfredi, M., Cohn, N., De Araújo Andreoli, M., & Boggio, P. S. (2018). Listening beyond seeing: Event-related potentials to audiovisual processing in visual narrative. *Brain and Language*, *185*, 1–8. https://doi.org/10.1016/j.bandl.2018.06.008

Manfredi, M., Cohn, N., & Kutas, M. (2017). When a hit sounds like a kiss: An electrophysiological exploration of semantic processing in visual narrative. *Brain and Language*, *169*, 28–38. https://doi.org/10.1016/j.bandl.2017.02.001

Mannan, S. K., Ruddock, K. H., & Wooding, D. S. (1997). Fixation sequences made during visual examination of briefly presented 2D images. *Spatial Vision*, *11*(2), 157–178.

Manohar, N., Sharath Kumar, Y. H., Kumar, G. H., & Rani, R. (2019). Deep learning approach for classification of animal videos. Singapore: Springer.

Matin, E. (1974). Saccadic suppression: A review and an analysis. *Psychological Bulletin*, *81*(12), 899–917.

Mayer, M. (1967). *A boy, a dog, and a frog*. New York: Dial Books for Young Readers.

McCrudden, M. T., Magliano, J. P., & Schraw, G. (2010). Exploring how relevance instructions affect personal reading intentions, reading goals and text processing: A mixed methods study. *Contemporary Educational Psychology*, *35*(4), 229–241. https://doi.org/10.1016/j.cedpsych.2009.12.001

McKoon, G., & Ratcliff, R. (1998). Memory-based language processing: Psycholinguistic research in the 1990s. *Annual Review of Psychology*, *49*(1), 25–42.

McNamara, D. S., & Magliano, J. P. (2009). Toward a comprehensive model of comprehension. In B. H. Ross (Ed.), *Psychology of learning and motivation* (Vol. 51, pp. 297–384). New York: Elsevier Science.

Memmert, D. (2006). The effects of eye movements, age, and expertise on inattentional blindness. *Consciousness and Cognition*, *15*(3), 620–627. https://doi.org/10.1016/j.concog.2006.01.001

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.

Moore, A., & Gibbons, D. (1987). *Watchmen*. New York: DC Comics.

Moss, J., Schunn, C. D., Schneider, W., McNamara, D. S., & VanLehn, K. (2011). The neural correlates of strategic reading comprehension: Cognitive control and discourse comprehension. *NeuroImage*, *58*(2), 675–686.

Myers, J. L., & O'Brien, E. J. (1998). Accessing the discourse representation during reading. *Discourse Processes*, *26*(2–3), 131–157.

Nelson, W. W., & Loftus, G. R. (1980). The functional visual field during picture viewing. *Journal of Experimental Psychology: Human Learning & Memory*, *6*(4), 391–399. https://doi.org/10.1037/0278-7393.6.4.391

Neumann, B., & Möller, R. (2008). On scene interpretation with description logics. *Image and Vision Computing*, *26*(1), 82–101.

Newtson, D. (1973). Attribution and the unit of perception of ongoing behavior. *Journal of Personality and Social Psychology*, *28*(1), 28–38.

Newtson, D., Engquist, G. A., & Bois, J. (1977). The objective basis of behavior units. *Journal of Personality and Social Psychology*, *35*(12), 847–862.

Nuthmann, A., & Henderson, J. M. (2010). Object-based attentional selection in scene viewing. *Journal of Vision*, *10*(8), 1–19. https://doi.org/10.1167/10.8.20

Oliva, A. (2005). Gist of a scene. In L. Itti, G. Rees, & J. K. Tsotsos (Eds.), *Neurobiology of attention* (pp. 251–256). Burlington, MA: Elsevier Academic Press.

Otero-Millan, J., Troncoso, X. G., Macknik, S. L., Serrano-Pedraza, I., & Martinez-Conde, S. (2008). Saccades and microsaccades during visual fixation, exploration, and search: Foundations for a common saccadic generator. *Journal of Vision*, *8*(14), 21–21. https://doi.org/10.1167/8.14.21

Pannasch, S., Helmert, 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), 1–19. https://doi.org/10.16910/jemr.2.2.4

Pertzov, Y., Avidan, G., & Zohary, E. (2009). Accumulation of visual information across multiple fixations. *Journal of Vision*, *9*(10:2), 1–12.

Peters, R. J., Iyer, A., Itti, L., & Koch, C. (2005). Components of bottom-up gaze allocation in natural images. *Vision Research*, *45*(18), 2397–2416.

Pezdek, K., Lehrer, A., & Simon, S. (1984). The relationship between reading and cognitive processing of television and radio. *Child Development*, *55*(6), 2072–2082. https://doi.org/10.2307/1129780

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*. New York: Oxford University Press.

Ramkumar, P., Hansen, B. C., Pannasch, S., & Loschky, L. C. (2016). Visual information representation and rapid-scene categorization are simultaneous across cortex: An MEG study. *NeuroImage*, *134*, 295–304. https://doi.org/10.1016/j.neuroimage.2016.03.027

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

Rayner, K., & Reichle, E. D. (2010). Models of the reading process. *WIREs Cognitive Science*, *1*(6), 787–799. https://doi.org/10.1002/wcs.68

Reichle, E. D., Rayner, K., & Pollatsek, A. (1999). Eye movement control in reading: Accounting for initial fixation locations and refixations within the E-Z Reader model. *Vision Research*, *39*(26), 4403–4411.

Ringer, R. V. (2018). The spatiotemporal dynamics of visual attention during real-world event perception. Doctor of Philosophy dissertation. Kansas State University, Manhattan, KS, USA.

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. https://doi.org/10.1167/16.2.7

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.

Sampanes, A. C., Tseng, P., & Bridgeman, B. (2008). The role of gist in scene recognition. *Vision Research*, *48*(21), 2275–2283. https://doi.org/10.1016/j.visres.2008.07.011

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

Schwan, S., & Ildirar, S. (2010). Watching film for the first time: How adult viewers interpret perceptual discontinuities in film. *Psychological Science*, *21*(7), 970–976.

Sheinberg, D. L., & Logothetis, N. K. (2001). Noticing familiar objects in real world scenes: The role of temporal cortical neurons in natural vision. *Journal of Neuroscience*, *21*(4), 1340–1350.

Smith, M. E., & Loschky, L. C. (in press). The role of sequential expectations in perceiving scene gist. *Journal of Vision*.

Smith, T. J. (2012a). The attentional theory of cinematic continuity. *Projections*, *6*(1), 1–27. https://doi.org/10.3167/proj.2012.060102

Smith, T. J. (2012b). Extending AToCC: A reply. *Projections*, *6*(1), 71–78.

Smith, T. J., & Henderson, J. M. (2008). Edit blindness: The relationship between attention and global change blindness in dynamic scenes. *Journal of Eye Movement Research*, *2*(2:6), 1–17.

Smith, T. J., Lamont, P., & Henderson, J. M. (2012). The penny drops: Change blindness at fixation. *Perception*, *41*(4), 489–492.

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. https://doi.org/10.1177/0963721412437407

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. https://doi.org/10.1080/15213269.2016.1160789

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), 1–24. https://doi.org/10.1167/13.8.16

Smith, T. J., Whitwell, M., & Lee, L. (2006). Eye movements and pupil dilation during event perception. Paper presented at the 2006 Symposium on Eye Tracking Research & Applications.

Snyder, Z. (Writer). (2009). *Watchmen* [Motion Picture]. H. Gains & T. Tull (Producers). United States: Warner Bros.

Speer, N. K., Zacks, J. M., & Reynolds, J. R. (2007). Human brain activity time-locked to narrative event boundaries. *Psychological Science*, *18*(5), 449–455. https://doi.org/10.1111/j.1467-9280.2007.01920.x

Suh, S., & Trabasso, T. (1993). Inferences during reading: Converging evidence from discourse analysis, talk-aloud protocols, and recognition priming. *Journal of Memory and Language*, *32*(3), 297.

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

Thorndyke, P. W. (1977). Cognitive structures in comprehension and memory of narrative discourse. *Cognitive Psychology*, *9*, 77–110.

Thorpe, S. J., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, *381* (6582), 520–522.

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., & Suh, S. (1993). Understanding text: Achieving explanatory coherence through on-line inferences and mental operations in working memory. Special Issue: Inference generation during text comprehension. *Discourse Processes*, *16*(1–2), 3–34.

Trabasso, T., van den Broek, P., & Suh, S. (1989). Logical necessity and transitivity of causal relations in stories. *Discourse Processes*, *12*(1), 1–25.

Triesch, J., Ballard, D. H., Hayhoe, M. M., & Sullivan, B. T. (2003). What you see is what you need. *Journal of Vision*, *3*(1), 9.

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

VanRullen, R., & Thorpe, S. J. (2001). The time course of visual processing: From early perception to decision-making. *Journal of Cognitive Neuroscience*, *13*(4), 454–461.

Walther, D. B., Caddigan, E., Fei-Fei, L., & Beck, D. M. (2009). Natural scene categories revealed in distributed patterns of activity in the human brain. *The Journal of Neuroscience*, *29*(34), 10573–10581. https://doi.org/10.1523/jneurosci.0559-09.2009

Welles, O., & Zugsmith, A. (Writers). (1958). *Touch of Evil*[Film]. A. Zugsmith (Producer). USA: Universal Pictures.

Williams, C. C., Henderson, J. M., & Zacks, R. T. (2005). Incidental visual memory for targets and distractors in visual search. *Percept Psychophys*, *67*(5), 816–827.

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., & 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.

Zacks, J. M., Braver, T. S., Sheridan, M. A., Donaldson, D. I., Snyder, A. Z., Ollinger, J. M., Buckner, R. L., & Raichle, M. E. (2001). Human brain activity time-locked to perceptual event boundaries. *Nature Neuroscience*, *4*(6), 651–655. https://doi.org/10.1038/88486

Zacks, J., Speer, N., & Reynolds, J. (2009). Segmentation in reading and film comprehension. *Journal of Experimental Psychology: General*, *138*(2), 307–327. https://doi.org/10.1037/a0015305

Zacks, J., Swallow, K., Vettel, J., & McAvoy, M. (2006). Visual motion and the neural correlates of event perception. *Brain Research*, *1076*(1), 150–162. https://doi.org/10.1016/j.brainres.2005.12.122

Zelinsky, G. J., & Loschky, L. C. (2005). Eye movements serialize memory for objects in scenes. *Perception & Psychophysics*, *67*(4), 676–690.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2018). Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(6), 1452–1464. https://doi.org/10.1109/TPAMI.2017.2723009

Zwaan, R. A., Magliano, J. P., & Graesser, A. C. (1995). Dimensions of situation model construction in narrative comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(2), 386–397.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, *123*(2), 162–185.

Zwickel, J., & Võ, M. L.-H. (2010). How the presence of persons biases eye movements. *Psychonomic Bulletin & Review*, *17*(2), 257–262.