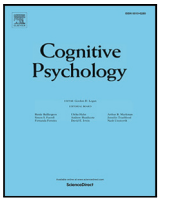
Cognitive Psychology 162 (2026) 101768

Contents lists available at ScienceDirect

Cognitive Psychology

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



Perceiving event structure in brief actions

Zekun Sun [a],*, Samuel D. McDougle [a,b]

[a] Department of Psychology, Yale University, 100 College St., New Haven, CT, 06510, USA
[b] Wu Tsai Institute, Yale University, 100 College St., New Haven, CT, 06510, USA

ARTICLE INFO

Dataset link: https://osf.io/j85fa

Keywords:
Event segmentation
Action perception
Schematic structure
Spatiotemporal dynamics
Motion

ABSTRACT

Event segmentation is a fundamental component of human perception and cognition. The field of event cognition studies how people decide where distinct events occur in incoming sensory data, how these "event boundaries" alter decision-making and memory processes, how events reveal themselves in neural activity, and how events may be represented within perception itself. The latter point is critical — the representation of events in the first place is filtered through perception. But what counts as a minimal event that is perceptible to humans? And to what extent is the perceptual representation of minimal events driven by physical properties within stimuli (e.g., sudden changes of a tennis ball's direction when one player strikes it) versus the semantic structure of events (e.g., "step one" versus "step two" of a tennis serve)? Here, across seven preregistered experiments, we explore the perceptual representation of event structure *within* single brief actions, and dissociate the roles of visual features and semantic structures in the perceptual segmentation of minimal events. First, participants produced boundary labels by segmenting videos of brief physical actions (e.g., kicking a ball). Then, separate groups of observers were asked to visually detect subtle disruptions in the video clips, unaware that the disruptions systematically occurred at boundary versus non-boundary timepoints. The results consistently showed an interfering effect of event boundaries on the detection of disruptions, suggesting a spontaneous perceptual representation of action structure even in very brief single actions. Moreover, boundary effects were strongest when stimuli were presented in recognizable forms versus distorted forms that only preserved lower-level features. Thus, automatic and rapid perceptual segmentation of single actions that only last several seconds may be driven by both sensory cues and our internal models of the world.

1. Introduction

The human mind tends to represent continuous experiences as discrete events, imposing "event boundaries" on incoming streams of sensory data. This phenomenon, known as event segmentation, not only shapes our perceptual experience but also underlies many cognitive processes (Zacks et al., 2007). Without it, we could not easily follow the arc of a story (Kumar et al., 2023; Ünal et al., 2021), recognize recurring phrases in music (Bregman, 1994), make useful predictions (Antony et al., 2021; Zacks et al., 2007), plan complex sequences of actions (Richmond & Zacks, 2017), or reason effectively about cause and effect (Shin & DuBrow, 2021; Zacks & Tversky, 2001). Developmental research suggests that the perceptual ability to parse dynamic, continuous actions into discrete units is a fundamental cognitive tool that emerges as early as 6–11 months of age (Baldwin et al., 2001; Hespos et al., 2009; Saylor et al., 2007).

While event segmentation occurs across modalities, it plays a particularly central role in vision. We naturally interpret streams of visual stimuli as meaningful, bounded episodes. In everyday life, a great amount of visual input involves observing other people's

* Corresponding author.
E-mail addresses: zekun.sun@yale.edu (Z. Sun), samuel.mcdougle@yale.edu (S.D. McDougle).

https://doi.org/10.1016/j.cogpsych.2025.101768
Received 27 May 2025; Received in revised form 9 October 2025; Accepted 10 October 2025
Available online 25 November 2025
0010-0285/© 2025 Elsevier Inc. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

actions — pouring a drink, opening a door, or tying a shoelace. Research shows that observers can reliably identify breakpoints, or boundaries, while observing everyday activities (e.g., making a bed, washing dishes, pitching a tent), suggesting that ongoing behavior is represented as a sequence of discrete actions with defined beginnings and endings (Buchsbaum et al., 2015; Hard et al., 2011; Newtson, 1973; Newtson et al., 1977; Zacks et al., 2001). Event segmentation can also occur at multiple levels of granularity (Baldassano et al., 2017; Kurby & Zacks, 2008; Zacks et al., 2009, 2007): For example, a long jump can be further segmented into sub-actions such as "running," "jumping," and "standing." (Fig. 1A). But what are the lower limits of single action events, or "minimal events"? Can perceptual event segmentation also occur even *within* single action events?

The present study investigates perceptual event structure within single, brief actions (e.g., a jump, a step, a golf swing; see Fig. 1B), with the goal of understanding event segmentation at this so-called "atomic" level. We approach this by examining the structure of action events at, to our knowledge, the shortest timescales studied to date. In addition to investigating event cognition in these minimal events, we further probe what sources of information drive the representation of action structure in visual perception.

1.1. Event structure within brief "atomic" actions

Although there is no clear definition, empirical work has shed light on the most basic units in event structures that unfold over time (see Yates et al., 2023 for a theoretical discussion on 'what counts as an event'). For example, seminal work in event perception described the types of brief, single basic actions that constitute the "atomic components of events" (Zacks & Tversky, 2001). In studies where observers were asked to identify the most fine-grained events in an ongoing sequence of actions, the minimal segmentable events appeared to be single actions with median lengths of 8 – 15 s; these included actions like grabbing a cup, closing a door, or opening a box (Zacks et al., 2009, 2001). Similar work using short clips of these atomic actions (e.g., drink, poke, blow) provided further evidence for boundary perception, not within, but *across* atomic actions (Baldwin et al., 2008; Buchsbaum et al., 2015; Hard et al., 2019). More recently, it has been proposed that a key feature of atomic actions is being "bounded", a concept analogous to objects in the spatial domain. That is, bounded events are defined as actions that are internally structured and have a well-defined endpoint (e.g., "stacking a deck of cards" or "folding a handkerchief"), in contrast to an unbounded action like "waving a handkerchief" (Lee et al., 2024).

Are such atomic actions the minimal, indivisible structure in event perception? Here we suggest that event structure can also be perceived *within* single, brief actions. To illustrate, consider simple motor skills, such as a golf swing, a tennis serve, or a swimming stroke. Although these actions can be brief (in the range of one to several seconds) and typically involve continuous, smooth body movements, they are often learned by breaking them into even smaller subroutines, e.g., a backswing and a downswing in golf (Fig. 1B). Do such "steps" within a brief action act as separable events in visual processing?

Research on motor skills and sports points to a defined event structure within even the briefest single actions. Some early work suggested that the mind encodes continuous body movements as a sequence of action units defined by a preparatory-completing structure — that is, a relatively stable motion followed by an unstable motion, such as a few quick steps followed by an arabesque in ballet. Evidence for this schematic representation was revealed using visual recognition tasks: After viewing movies of a ballet dancer, observers were more accurate in identifying preparatory-completing units than completing-preparatory units (Lasher, 1981). In a similar vein, researchers have shown that elite athletes can predict the outcome of others' actions based on early preparatory kinematic cues. For example, compared with novices, skilled basketball and volleyball players are more accurate anticipating the fate of the ball before it has even left the player's hand, relying on the perception of body kinematics during the preparatory phase of the action, prior to a putative event boundary (Smith, 2016; Urgesi et al., 2010, 2012). Physiological and neural data further point to the special status of preparatory phases in simple actions (Cohn et al., 2017; Urgesi et al., 2010).

Theoretical work in semantic analysis has likewise proposed a "preparation–head–coda" hierarchical structure to explain how ordinary actions are encoded and stored in memory. For example, the action sequence "making coffee" can be segmented into a preparation phase – "get out coffee," a head phase – "measure coffee," and a coda phase – "put away coffee"; and critically, each phase can be further segmented into minor prep-head–coda structure (Jackendoff, 2007). These ideas and findings suggest that a multi-phase causal structure may underlie our perception of even the briefest bounded actions. However, it is not yet known whether such granular action structure is automatically represented in visual processing. In the present work, we use psychophysical methods to examine how event structure unfolds within atomic actions, using a broad set of actions as stimuli.

1.2. Sensory cues and semantic content in perceptual segmentation

If atomic actions are indeed subdivided into discrete events, what information does the mind use to segment them? Both theoretical and empirical work suggests that event segmentation relies on a combination of sensory cues and semantic knowledge (Buchsbaum et al., 2015; Newtson et al., 1977; Zacks et al., 2007). Sensory cues for separable events are often salient, such as someone changing position in the kitchen as they move between the coffee maker and the sink. Motion cues are especially important, driving the perception of both fine-grained action events (e.g., 8–12 s, Zacks et al., 2009) and abstract events (e.g., simple animations with moving shapes, Pomp et al., 2024; Zacks, 2004). In one study, observers reported similar event boundaries in upright and inverted displays of action sequences, with these boundaries predicted by kinematic variables such as changes in direction, velocity, and acceleration of the relevant effectors (Hemeren & Thill, 2011).

Meanwhile, semantic event structure, e.g., "step one" versus "step two" of a tennis serve, could in theory be represented in perception, consistent with evidence that perception encodes abstract, structured, and sophisticated representations in other domains (Chen & Scholl, 2016; Cohn et al., 2017; Dasser et al., 1989; Hafri et al., 2018; Ji & Scholl, 2024; Loucks & Pechey, 2016; Scholl

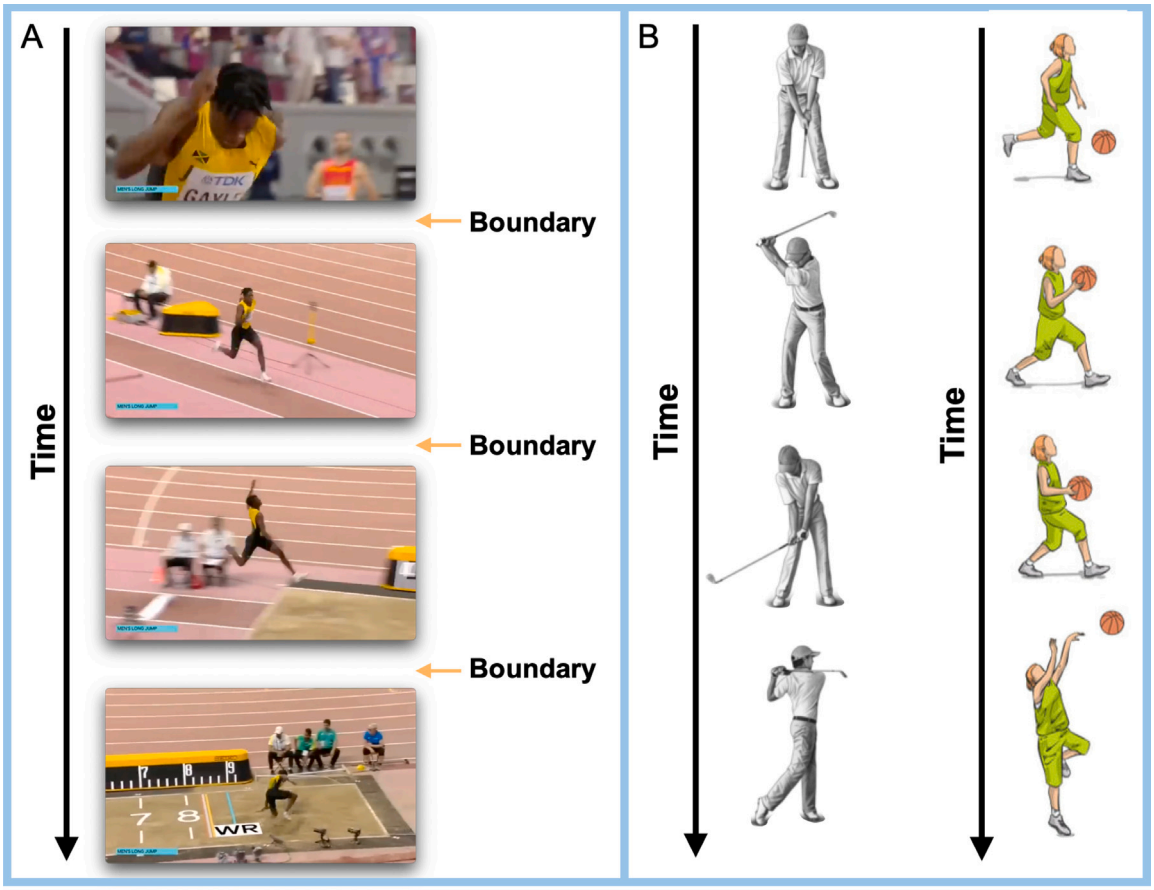


Fig. 1. (A) Mental representations of event boundaries emerge from ongoing perceptual processing of continuous sensory data. As shown here, a long jump video could be segmented at the shot cut and at the transition between run-to-jump and jump-to-stand — timepoints that coincide with abrupt changes in visual features and semantic information. How short can actions be to still retain event structure in perception? (B) We explore these possibilities for the case of single, brief "atomic" actions, such as a golf swing or a basketball shot. We ask whether and how the discrete 'steps' within a single brief action are represented in visual perception.

& Tremoulet, 2000; Strickland & Scholl, 2015). Although both sources of information — sensory cues and semantic knowledge — can influence event segmentation, it remains unclear how they may function separately or jointly in perceptual segmentation at the shortest time scales. A key challenge is that these two sources of information are often correlated: semantically meaningful event boundaries frequently coincide with salient visual feature changes. This is evident in prior work that focused on perceptual effects driven by mental representations of event boundaries (Huff et al., 2012; Ji & Papafragou, 2022; Yates et al., 2024). For example, in a soccer game, an event boundary could correspond to one player passing the ball to another — or simply to a sudden change in motion cues in the scene (Huff et al., 2012). Dissociating these two sources requires removing semantic content from visual streams while preserving key physical stimulus properties. Building on previous efforts (Pomp et al., 2024; Zacks et al., 2009), our work here introduces simple new methods to do so.

1.3. The present study: Event structure within atomic actions

Here we ask whether event structure within brief actions is represented in visual perception, and to what extent spatiotemporal cues versus semantic content contribute to visual segmentation. We used simple animations of single actions originally created using motion capture technology. These action clips are not only much shorter than the stimuli used in previous studies of event segmentation, but they also occur in a stripped-down context that only shows a single agent's bodily motion. The simplicity of these stimuli allowed us to manipulate semantic content while preserving most lower-level visual features, such that semantic content and motion cues could be decorrelated. To foreshadow the key results, across seven experiments we find that perceptual boundary effects consistently emerged even within brief atomic actions, and that such effects were significantly stronger when the semantic structure of the actions was preserved. These findings suggest that the timescale of minimal events may be briefer than previously believed, and that both semantic knowledge and visual cues contribute to automatic perceptual segmentation even at these brief timescales.

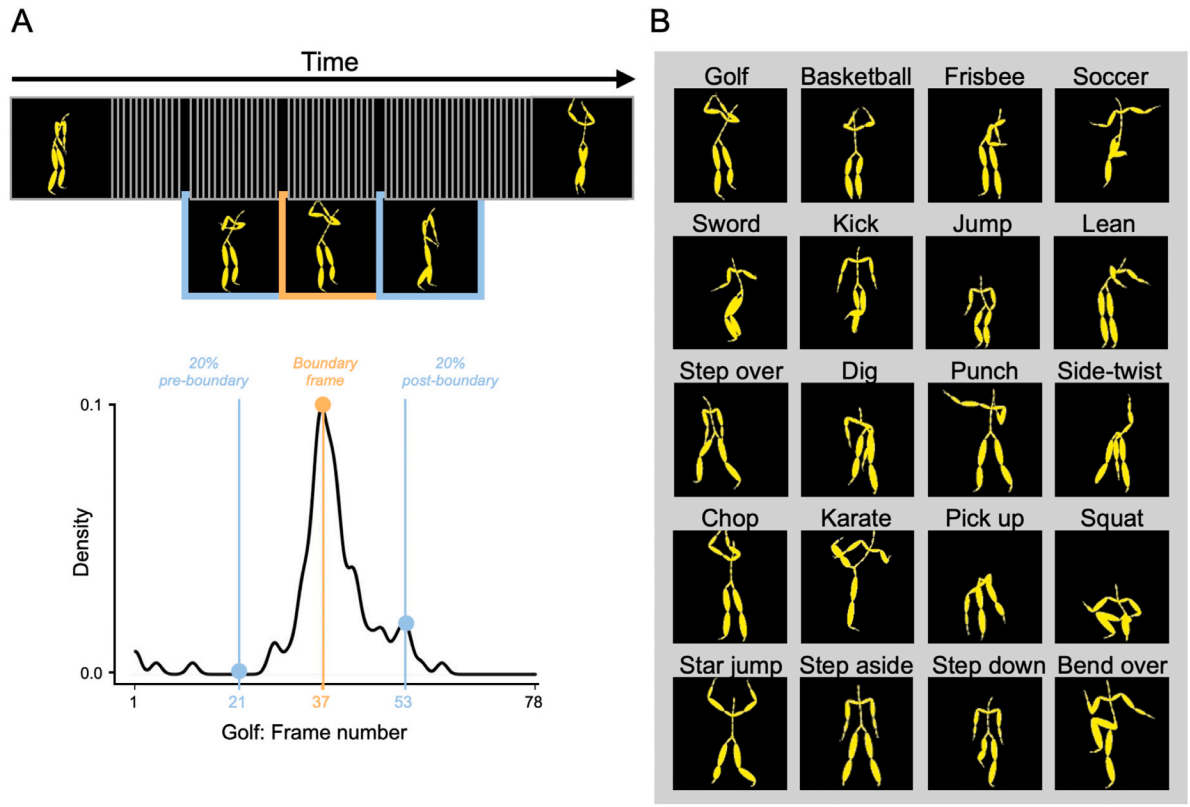


Fig. 2. (A) The method used to define the most characteristic event boundary for an action. Taking the golf swing video as an example, all event boundary responses from participants were combined and smoothed with a Gaussian kernel. The frame closest to the stimulus' global peak of the event boundary choices was identified as the boundary frame, and 20% frames of the video clip before the boundary as the pre-boundary frame and 20% after the boundary as the post-boundary frame. (B) The 20 actions used in this study. Each image represents the boundary frame of each action video, as determined by our analysis method in Experiment 1. (See action videos at: https://zk.actlabresearch.org/segmentation.)

1.4. Open science practice

This research received approval from Yale University's local ethics board. For all experiments, we preregistered the sample size, experimental design, and statistical analyses. We report all of the preregistered analyses; however, we also report several exploratory analyses suggested by reviewers. Where relevant, we indicate below which results reflect preregistered analyses and which are exploratory. The data, experiment code, stimuli, and experiment pre-registrations for all studies are available at: https://osf.io/j85fa/.

Demonstrations of all experiments and the full set of dynamic stimuli can be found at: https://zk.actlabresearch.org/segmentation/.

2. Experiment 1: Explicit segmentation

Experiment 1 aimed to identify the most characteristic boundary frame of the brief "atomic" actions in our stimulus set. We asked observers to deliberately segment a variety of action videos into two semantically distinct units. The simple MoCap animations used here appear as short, fluent and natural actions, stripped of contextual cues about scenes, agents, or objects. Given that the video clips used here could be very short, observers chose event boundaries offline, instead of during the video clip, such that our estimates of boundaries could be more accurate. We expected participants' boundary judgments to reflect how they explicitly represent the discrete steps of observed actions in a top-down manner (e.g., "step one" and "step two" of a tennis serve).

2.1. Method

2.1.1. Participants

As stated in our preregistration, we recruited 100 participants from the online platform Prolific (https://www.prolific.co/). Participants were pre-screened for age (15–35), a minimum approval rate of 99%, at least 50 prior submissions, normal or corrected-to-normal vision, fluency in English, and U.S. residence.

2.1.2. Stimuli

We compiled 20 animations depicting short natural actions, spanning sports, simple exercises, and everyday tasks, from the CMU *MoCap* database (http://mocap.cs.cmu.edu/). These videos involve an unidentified and skeletonized human figure performing a single, well-defined action (e.g., making a golf swing, picking up an object, taking a step aside) on a black background (Fig. 2B). Such simple, short actions have a salient basic structure, and are thus often referred to as "bounded events," which reflect events that lead to a salient start and endpoint that is naturally achieved unless there is a surprising interruption (Comrie, 1976; Ji & Papafragou, 2022; Mittwoch, 2013). The approximate duration of the 20 actions ranged from 1.0 s to 3.9 s (Mean = 2.2 s, SD = 0.7 s).

The action videos (200 × 200 pixels) were displayed in participants' web browser. The workspace covered 500 × 500 pixels with a black background. Because of the nature of online studies, we could not know the exact viewing distance, screen size, and luminance (etc.) of these stimuli as they appeared to participants. However, any distortions introduced by a given participant's viewing distance or particular monitor settings would have been equated across all stimuli and conditions.

2.1.3. Procedure

Each trial started with a "ready" cue appearing at the center of the workspace, which reminded participants to observe the full video of an action. After viewing the full video, they were given a slider they could toggle with their mouse, which allowed them to iterate through all the frames of the previously viewed action. Their job was to move the slider in order to find the most appropriate frame that divides each action into two units, such that each unit is "meaningful." The width of the slider was determined by the number of frames of the current action video, such that the amount of mouse movement required for seeing each frame was equal across all actions. The starting position of the slider was randomized on each trial. Each participant made these boundary judgments for each of the total 20 unique actions. No time pressure was applied to participants' responses. The serial order of actions was randomized across participants.

2.2. Results and discussion

Two participants were excluded for failing to submit a complete data set, leaving 98 participants with analyzable data. For each action, participants' choices of the boundary were combined and smoothed with a Gaussian kernel (bandwidth = 1 frame), which gave us the density of choices for each frame of the action video. The boundary of an action was defined as the frame corresponding to the highest peak of the fitted density function (see an example in Fig. 2A). Using this procedure, we obtained a single event boundary frame for all 20 actions (Fig. 2B). Observers showed high agreement in selecting boundaries: On average, 40.3% observers chose either the group-level peak frame or its neighboring frames as an action's event boundary.

Boundary selections were not limited to the videos' middle frames: The earliest boundary was 29.4% from the beginning across all videos, and the latest was 61.3%. Indeed, as predicted given our instructions, observers typically selected boundary frames that best described how to perform an action in discrete steps, e.g., kicking a soccer ball is divided into planting the supporting foot and then swinging the kicking foot, performing a simple jump requires bending the knees and then extending upward, etc. Fig. 2B depicts each boundary frame. As can be seen, the segmentation in some actions reflected the action schema that has been proposed in previous work, such as a preparing-completing structure in Lasher (1981), or prep-head–coda structure in Jackendoff (2007). For example, "shooting a basketball" was divided into a preparation phase and an execution phase, "picking up" appeared to have a "reaching" preparation phase and a "picking-up" phase, and "karate kick" appeared to reflect a head–coda structure (kicking, then withdrawing the foot). But other cases did not fit into obvious action schema, such as the exercise skills stimuli and the simple stepping stimuli. We note that the current work does not focus on any particular type of action schema, but aims to encompass a broad range of actions and potential event structures.

3. Experiment 2: Temporal change detection

Experiment 2 explored whether the action boundary was also spontaneously represented in visual perception, rather than just reflecting an explicit decision about where the boundary belonged. Previous empirical evidence has shown that perceptual processing of event transitions (i.e., boundaries) transiently impairs one's ability to detect subtle changes in continuous perceptual input (Huff et al., 2012; Ji & Papafragou, 2022; Repp, 1992, 1998; Reynolds et al., 2007; Yates et al., 2024; Zacks et al., 2007). If action boundaries are automatically encoded during visual processing, then observers' performance should be influenced by action structure even when the task does not demand attention to it. Here, we instructed observers to detect a transient slowdown in action videos, and tested whether their perceptual sensitivity differs when the slowdown occurs at boundary versus non-boundary frames.

3.1. Method

3.1.1. Participants

In line with our preregistration, 20 participants were recruited through Prolific. A smaller pilot suggested that this sample would have power above 95% to reveal perceptual boundary effects.

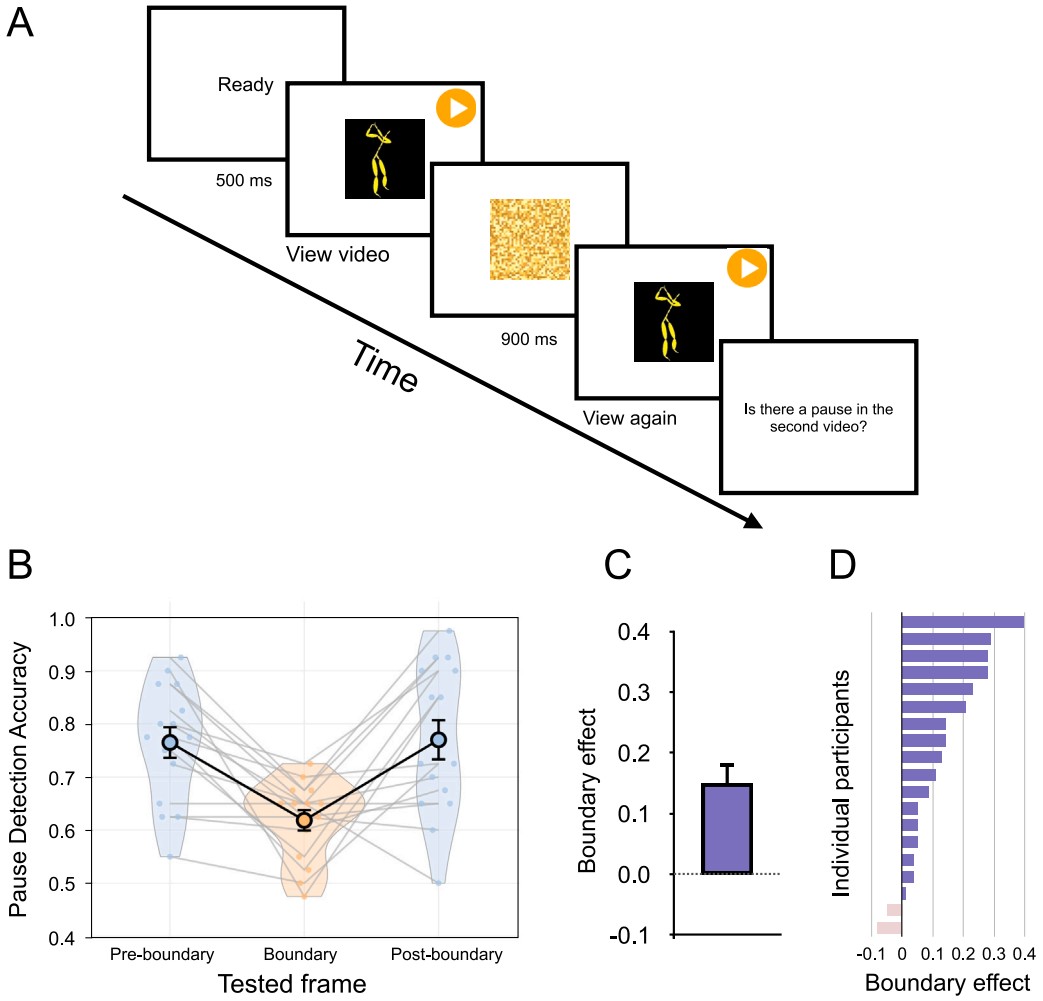


Fig. 3. Trial structure and results of Experiment 2. (A) In each trial, observers watched each action video twice. During the second viewing, a brief pause could happen at one of the three tested frames (pre-boundary, boundary, or post-boundary frame). Observers judged whether there was a pause or not in the second play. (B) Compared to the pre-boundary and post-boundary frames, observers were less likely to identify the pause occurred at the boundaries. (C) Mean accuracy reduced by 14.9% at the boundary of actions across participants. (D) A strong majority of participants were less sensitive to the disruptions at the boundary of actions. The boundary effect was calculated by subtracting the mean accuracy in the boundary condition from that in the non-boundary condition, where non-boundary accuracy was averaged across pre- and post-boundary trials. Error bars = 1 s.e.m.

3.1.2. Stimuli and procedure

Participants were instructed to detect transient slowdowns in the animated actions used in Experiment 1. On each trial, observers were first presented the word 'ready' for 500 ms, then watched the full video of an action, followed by a mask image (a yellow random noise pattern) for 900 ms, and then the video was played a second time. The second play was either identical to the first play or contained a transient slowdown (a 60- or 90-ms pause on two-thirds of trials). Observers pressed either F or J on their keyboard to indicate whether the second play did or did not contain a brief pause compared to the first play, with no time pressure to respond. Fig. 3A illustrates the trial sequence. Critically, the pause occurred at one of three specific frames determined by the results of Experiment 1: the pre-boundary frame (20% of frames earlier than the boundary), the boundary frame, or the post-boundary frame (20% of frames later than the boundary). The three frame conditions were crossed with the 2 pause time conditions, yielding 3 (pre-, on-, post-boundary) × 2 (60, 90 ms), or 6 unique trial types. All 20 actions appeared as each of these trial types, for 20 × 6 = 120 trials. Additionally, each action was shown three times without a pause, adding 60 control trials. Trial order was randomized across participants. Participants were also given 3 practice trials at the onset of the task (using an action video that did not appear elsewhere in the experiment).

3.2. Results and discussion

Four participants were excluded for low accuracy (< 60%; i.e., not significantly above chance according to a binomial test with $\alpha = .05$) and 1 for failing to provide complete data, leaving 15 participants with analyzable data. Across all participants, the mean accuracy was 75.3%. We only analyzed the trials with a pause (60 ms and 90 ms).

As shown in Fig. 3B, observers were less accurate in detecting pauses at boundary frames relative to non-boundary frames (averaging pre- and post-boundary frames), $t(14) = 4.52$, $p = .00049$, Cohen's $d = 1.17$, 95% $CI_{effect} = .15[.084, .21]$.[1] Thirteen out of 15 participants (87%) showed lower visual sensitivity to subtle disruptions at action boundaries. Observers were more accurate in detecting longer pauses, and their accuracy was reduced at boundary frames in both 60- and 90-ms conditions (60-ms: 57.7% for boundary vs. 70.7% for nonboundary; 90-ms: 66% for boundary vs. 83% for nonboundary).

We also conduced a repeated-measures one-way ANOVA to examine the effect of pause frame (pre-boundary, boundary, post-boundary) on accuracy. This exploratory analysis revealed a significant main effect of pause frame, $F(2, 28) = 12.18$, $p < .001$, $\eta^2_{partial} = .34$. Pairwise comparisons showed that accuracy was significantly lower in the boundary condition compared to both the pre-boundary (Tukey-adjusted $p < .001$) and post-boundary ($p_{adjusted} = .0071$) conditions, which did not differ significantly from each other ($p_{adjusted} = .98$).[2]

These results suggest that the transition of events within single atomic actions impedes observers' ability to detect the temporal disruptions in visual input. Note that the cover task here did not demand observers to intentionally pay attention to event structure or boundaries of video clips, as they only had to judge whether two videos were the same or different. Thus, the boundary effects here did not reflect observers' explicit knowledge (as in Experiment 1) but rather a spontaneous representation of action structure in visual perception. Considering the simple nature of the stimuli (e.g., a single continuous scene with no objects, perspective changes, or dramatic transitions), two possible sources of information could independently or jointly drive such perceptual effects: (a) subtle spatiotemporal changes and motion cues in these dynamic scenes, and (b) the semantic structure of the actions.

4. Experiment 3: Spatial change detection

The previous experiment addressed visual segmentation using dynamic stimuli (i.e., brief videos of actions). However, evidence exists showing that the mind can also form rich event representations from a series of discrete images, where event segmentation cannot rely on motion or dynamic signals (Baldwin et al., 2008; Cohn, Holcomb, et al., 2012; Ezzyat & Clements, 2024; Zheng et al., 2020). For example, people segment narratives while reading the static images of a comic book in a certain order (Cohn, Paczynski, et al., 2012). Here, we tested if perceptual boundary effects similar to those reported in Experiment 2 can emerge from such a scenario, where no motion signals can be used. In this experiment, we asked participants to detect subtle changes of spatial features instead of temporal disruptions as in the previous experiment. We predicted that the detection of spatial change would also be weakened at boundary frames if the transition between action steps is indeed represented in perception.

4.1. Method

4.1.1. Participants

As stated in our preregistration, 20 participants were recruited through Prolific.

4.1.2. Stimuli and procedure

This experiment asked participants to view a sequence of discrete images taken from the original action videos (maintaining the temporal order) and then judge whether the last image shown changed from the previous image. The images were selected from the frames of the action videos used in Experiments 1–2, one out of every three frames to create the sequence. We sequentially presented the selected images before the designated "key" frame, which was either the boundary, pre-boundary, or post-boundary frame obtained in Experiment 1 (e.g., in Fig. 4A, the orange bordered frame is the boundary frame of the action "tossing a frisbee"). Following the key frame was a testing frame, which was either the same as the key frame, or 2 frames forward in time from the key frame. Thus, from the viewpoint of the observer, on each trial, a series of images was sequentially displayed (each for 700 ms, with a 100 ms interstimulus interval), and they simply needed to determine whether the last frame changed from its previous frame or not (Fig. 4A). Since observers were not informed about how many images would be shown on each trial, they had to monitor every change of the images.

[1] The difference scores between boundary and non-boundary conditions were normally distributed, without violating the normality assumption of our preregistered t-tests (Shapiro–Wilk normality test: W = 0.97, p-value = 0.90). We note that for each t-test reported throughout, we confirmed the data were normally distributed (all $Ps > .1$).

[2] Compared to pre- and post-boundary frames, the boundary frame was temporally closer to the middle time point of the videos in many cases. Does decreased sensitivity at action boundaries simply reflect a kind of "middle frame" effect? To address this possibility, we ran a version of the task in which either the first or last frame of an action was extended by an additional 300–500 ms (9–15 frames; representing a 12%–50% increase in total duration, across 20 actions). This manipulation effectively shifted the original pre-, on-, and post-boundary timepoints away from the temporal center of the video, such that either the pre- or post-boundary frame was closer to the middle timepoint than the boundary frame. Despite this shift, the boundary effect was replicated: Accuracy was 17.8% lower at boundary frames compared to non-boundary frames, $t(11) = 4.14$, $p = .0016$, $d = 1.20$. This control study, together with Experiments 4–5, suggested that middle-frame effects did not account for our findings.

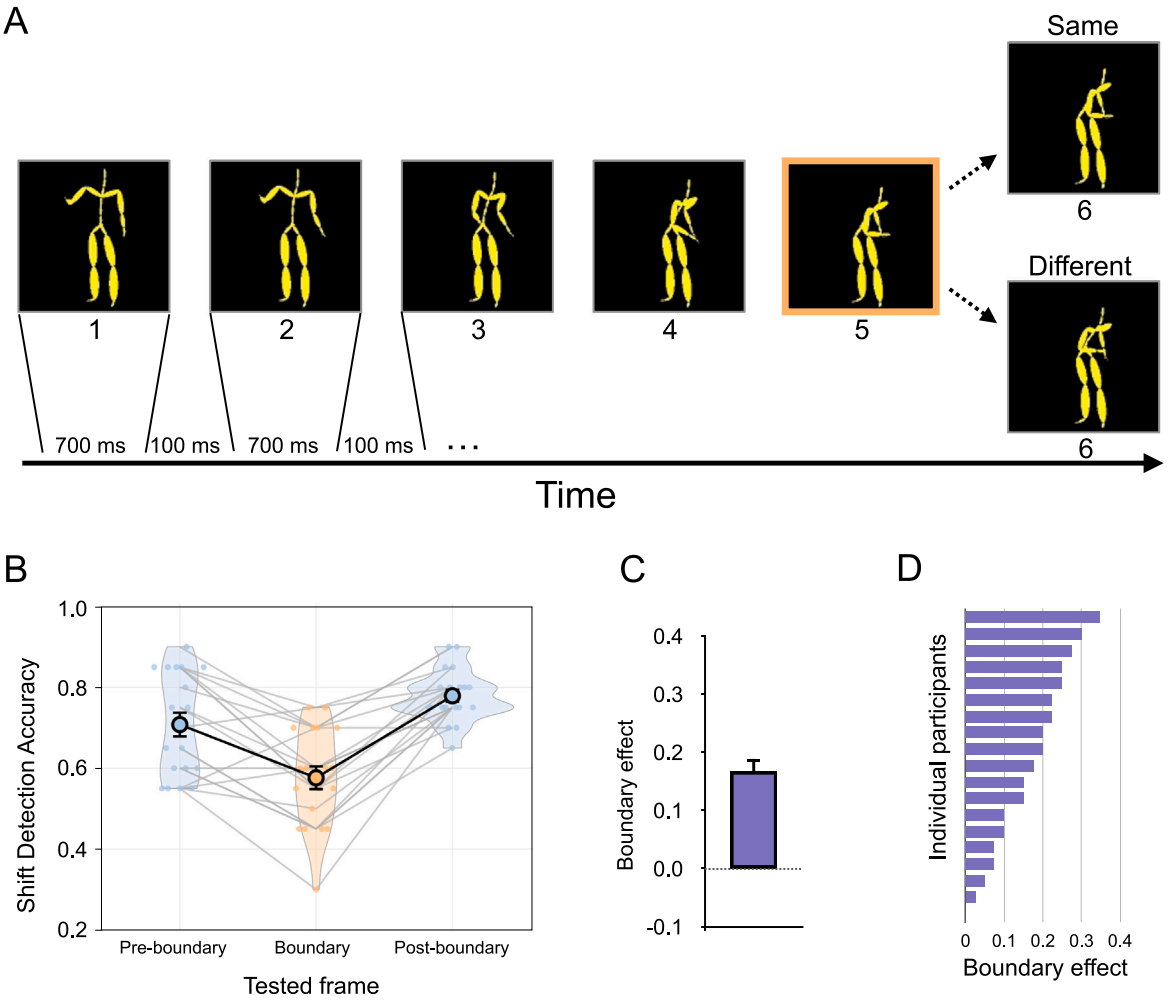


Fig. 4. Trial structure and results of Experiment 3. (A) In each trial, participants observed a sequence of images (i.e., individual frames of each action video) turning on and off, with 700 ms duration and a 100 ms interstimulus interval. Without knowing how many images would show up, participants were asked to judge if the last image shown in the sequence changed from the previous image. Critically, the second-to-last image on each trial (in this example, the fifth frame) could be the pre-boundary, boundary, or post-boundary frame of the action. (B) When a change occurred at the boundary frame, participants were less likely to identify it, similar to Experiment 2. (C) Mean accuracy reduced by 16.7% at boundary across participants. (D) All participants, except one, were less sensitive to changes of boundary frames compared to non-boundary frames. The boundary effect was calculated as the difference in accuracy between non-boundary and boundary conditions. Error bars = 1 s.e.m.

On each trial, participants pressed a key to observe the sequence of images depicting each action. After the last image was displayed, participants pressed either F or J on their keyboard to indicate whether the last image had shifted at all relative to the previous one. Two factors were fully crossed within-subject: 3 (pre-, on-, post-boundary) × 2 (shift, no shift) × 20 (actions) + 4 practice trials = 124 trials. Trial order was randomized across participants (except for practice trials).

4.2. Results and discussion

One participant was excluded for low accuracy (< 60%, not significantly above chance), leaving 19 participants with analyzable data, with a mean accuracy of 73.7%. Across 20 actions, observers' detection accuracy at pre-boundary, boundary, and post-boundary frames was 70.8%, 57.6%, and 77.9%, respectively (Fig. 4B). Disparate performance between boundary and non-boundary frames was confirmed by a paired t-test, $t(18) = 7.41$, $p = 7.14 \times 10^{-7}$, $d = 1.70$, 95% $CI_{effect} = .17[.12, .21]$ (Fig. 4C). All participants, except one, showed lower sensitivity to image changes at boundary compared to non-boundary frames (Fig. 4D).

A repeated-measures ANOVA was conducted to assess how accuracy varied across three temporal conditions: on-boundary, post-boundary, and pre-boundary. This exploratory analysis revealed a significant main effect of condition, $F(2, 36) = 28.17$, $p < .001$, $\eta_p^2 = .61$. Post-hoc comparisons indicated that accuracy was significantly lower during the on-boundary condition compared to the

post-boundary ($p_{adjusted} < .0001$) and pre-boundary conditions ($p_{adjusted} = .0006$). The difference between pre and post conditions was marginal and did not reach statistical significance ($p_{adjusted} = .058$).

Though we do not read too deeply into this latter result as the difference was marginal, the higher accuracy right after the action boundary (relative to pre-boundary) is consistent with the predictions of Event Segmentation Theory, where a transient increase in prediction error at event boundaries opens a perceptual gate, allowing the current event representation to be updated (Reynolds et al., 2007; Zacks, 2020). This effect might have been concealed in Experiment 2 where observers viewed the full action and then made offline judgments afterward. Here, observers viewed image sequences up to the testing frame, and thus their increased visual sensitivity to image features at the post-boundary frames may reflect how the action boundary functions to update an internal event model.

Overall, the results suggest that the perception of action boundaries interrupted the processing of subtle differences, even when viewing static images. Unlike Experiment 2, observers made their responses here in the absence of any image motion signals, and were never given a chance to preview the complete action videos. Yet, when the feature change happened around the boundary frame, they were less likely to detect it.

5. Experiment 4a: Low-level dynamic features

Experiments 2–3 revealed a perceptual effect wherein participants were less likely to detect subtle changes at action boundaries compared to non-boundaries. However, are such effects merely driven by salient visual differences between boundary and non-boundary frames? If frame-to-frame changes are more subtle around action boundaries (i.e., slower movement of pixels at boundaries), transient slowdowns at boundary frames would be less noticeable than those at non-boundary frames. In this experiment, we introduced a method to generate "spiralized" videos, whereby pixel-level dynamics are sufficiently preserved while high-level semantic information is removed from the stimuli, allowing us to ask whether boundary effects were fully explained by visual cues in the stimuli or not.

5.1. Method

5.1.1. Participants

In line with our preregistration, 35 participants were recruited through Prolific. A power analysis on the main results of a smaller pilot (i.e., the paired t-tests on the boundary effect) suggested that this sample would have power above 95% to reveal differences between normal and spiralized videos in terms of boundary effects (considering accidental loss of data). For Experiments 4a–b and 5b, which used the same number of conditions and trials, we preregistered the same sample size.

5.1.2. Stimuli and procedure

We used the same videos as earlier experiments but distorted all the frames of the actions with a spiral-shaped filter, generating a new set of videos that preserved most lower-level motion signals in the original videos while removing higher-level semantic information. The spiralized videos were created by taking each frame of the given standard action video clip and running it through a twirl filter with 600° using Adobe Photoshop (Fig. 5A). Readers can experience the similar dynamics between normal and spiralized videos at: https://zk.actlabresearch.org/segmentation/skeleton.html.

Indeed, the distorted videos maintained a nearly identical frame-to-frame profile in terms of the pixel changes across pairs of frames as the original actions. For example, in the golf swing stimulus, the obvious change in pixel movements between the backswing and the downswing is shown as the transition from clear outward to inward motion in its spiralized version, and the deceleration and acceleration during the swing are similarly salient in the spiral pattern. In fact, pixel changes across consecutive frames were highly correlated between original and spiralized videos (r = 0.97 across 20 actions), giving rise to highly similar impressions of motion and dynamics. In some cases, the motion signal was even stronger in the spiralized videos due to the stretching of the pixels (on average, the between-frame shift is 3824 pixels in normal videos versus 4350 pixels in spiralized videos).

We quantified pixel-level motion features at the pre-boundary, boundary, and post-boundary frames by calculating the number of pixels that changed relative to the preceding (–1) and following (+1) frames. Across the 20 actions, pixel change was numerically greater at the pre- and post-boundary frames than at the boundary frame. For normal videos, the average number of changed pixels was 3907 (pre-boundary), 3701 (boundary), and 3921 (post-boundary). For spiralized videos, the values were 4435 (pre-boundary), 4206 (boundary), and 4462 (post-boundary). However, the pixel-level difference between boundary and non-boundary frames were not statistically significant across 20 videos (all $Ps > 0.1$). The unique contribution of pixel-level features were further examined in an exploratory regression analysis reported below.

Observers attempted to detect transient pauses in the second play of each video. The experimental design was the same as Experiment 2, except that: (1) two types of videos were used — normal videos versus spiralized videos, and (2) for each participant, 10 randomly-selected actions used 60 ms for the pause trials and the other 10 used 90 ms. The Video Type condition was crossed with the other conditions as in the previous experiment, yielding 20 (actions) × 3 (pre-, on-, post-boundary) × 2 (with/without pause) × 2 (normal, spiralized video) + 4 practice trials = 244 trials. We imposed brief pauses either at the pre-boundary, boundary, or post-boundary frames in both the normal and spiralized video conditions, using the same frame locations for each condition. Trial order was randomized across participants (except for practice trials).

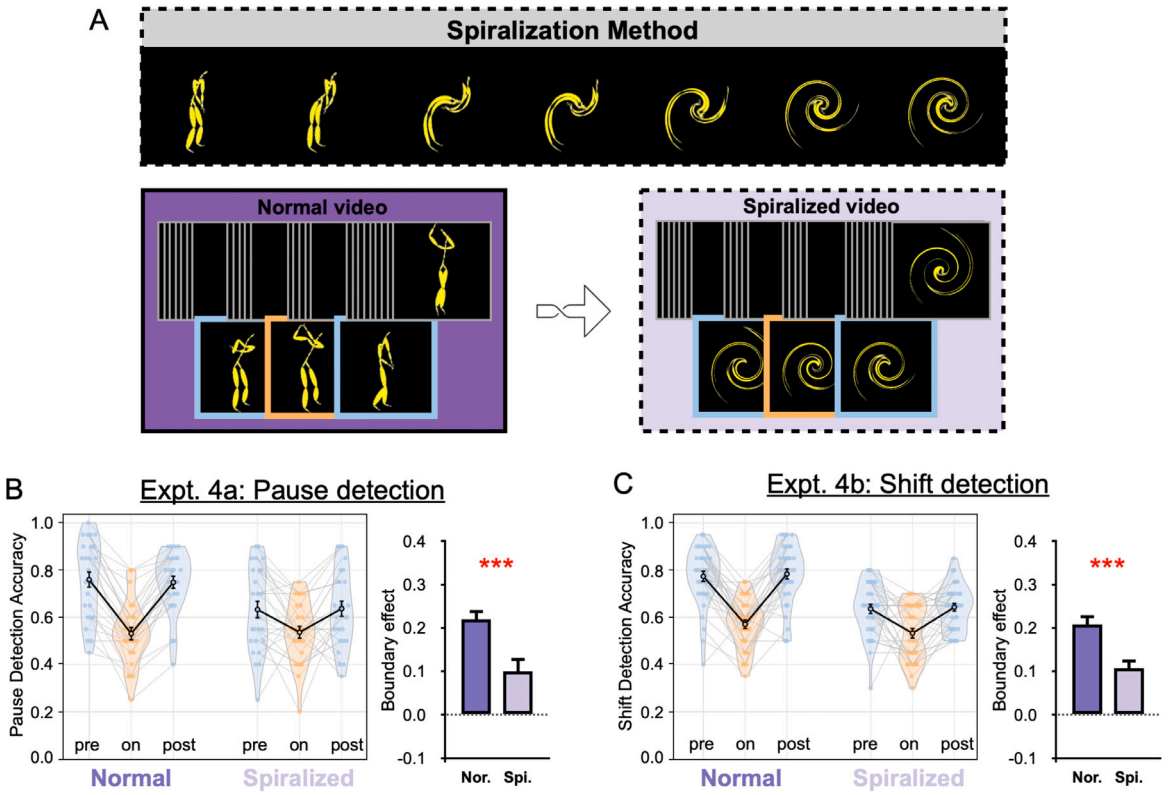


Fig. 5. (A) Method used to generate the spiralized video stimuli. For each action video, we radially twisted all its frames using a spiral-shaped filter, rendering an abstract, meaningless pattern that maintained many of the lower-level spatiotemporal dynamics of the video clips. Experiments 4a and 4b used the same design as Experiment 2 and 3 respectively, except that observers saw both the normal and spiralized versions of videos and images. (B) In Experiment 4a, the boundary effect (i.e., the reduced detectability of pauses at boundaries relative to non-boundaries) was significantly weakened in the altered videos. (C) In Experiment 4b, the boundary effect decreased for the spiralized images. The boundary effect was calculated as the difference in accuracy between non-boundary and boundary conditions. Error bars = 1 s.e.m. Asterisks indicate significant differences between means (***$p < .001$).

5.2. Results

Eight participants were excluded for low accuracy (< 57%, not significantly above chance according to a binomial test with $\alpha = .05$) and 1 for failing to provide complete data, leaving 26 participants with analyzable data, with a mean accuracy of 75.3%. Significant boundary effects were found in both types of videos: Observers were less accurate in detecting brief pauses at boundary frames relative to non-boundary frames (normal video: $t(25) = 11.05$, $p = 4.12\times10^{-11}$, $d = 2.17$, 95% $CI_{effect} = .22[.18, .26]$; spiralized video: $t(25) = 3.16$, $p = .004$, $d = 0.62$, 95% $CI_{effect} = .10[.041, .16]$). However, the boundary effect (i.e., the decreased detection accuracy at boundary relative to non-boundary frames) was significantly weakened for the spiralized stimuli ($t(25) = 4.19$, $p = .00030$, $d = 0.82$, 95% $CI_{difference} = .12[.059, .18]$; Fig. 5B).

Two follow-up exploratory analyses were used to further examine the data. First, we conducted a 2 (normal, spiralized) × 3 (pre-boundary, boundary, post-boundary) repeated-measures analysis of variance (ANOVA), which revealed significant main effects of Testing Frame, $F(2, 50) = 35.21$, $p < .001$, $\eta_p^2 = .58$, and Video Type, $F(1, 25) = 18.05$, $p < .001$, $\eta_p^2 = .42$, as well as a significant Testing Frame × Video Type interaction, $F(2, 50) = 11.52$, $p < .001$, $\eta_p^2 = .32$. Follow-up pairwise comparisons showed that in the normal condition, accuracy was significantly lower at the boundary compared to both the pre-boundary (Tukey-corrected $p < .0001$) and post-boundary time points ($p_{adjusted} < .0001$). In the spiralized condition, accuracy at the boundary was numerically lower than both pre- and post-boundary, but these comparisons did not reach significance thresholds ($p_{adjusted} = .082$ and $p_{adjusted} = .053$, respectively). Finally, accuracy was significantly higher in the normal condition than the spiralized condition at both the pre-boundary ($p_{adjusted} = .0015$) and post-boundary ($p_{adjusted} = .0006$) frames (Fig. 5B).

Second, a multiple linear regression model was used to examine the extent to which Video Type (normal vs. spiralized, coded as 1 and 0), Testing Frame (pre-, on-, post-boundary, coded as 1, −1, 1) and Pixel Change (see 5.1.2 for details about how this metric is computed) explained variance in response accuracy. The overall model explained approximately 24.7% of the variance in accuracy ($R^2 = .247$). All three predictors were statistically significant. Accuracy was significantly lower for twisted videos compared to normal ones ($\beta = -0.13$, $p < .001$), and higher at non-boundary frames ($\beta = 0.08$, $p < .001$) compared to boundaries. Pixel change

also showed a modest but significantly positive relationship with response accuracy (β = 0.00006, p = .022). Relative importance analysis revealed that Testing Frame uniquely accounted for the largest proportion of explained variance (58.3%), followed by Video Type (29.6%) and Pixel Change (12.1%).

These results showed that action boundaries impaired performance more reliably in normal videos compared to spiralized videos where semantic content was removed. This suggests that the perception of event boundaries within single actions may not solely rely on pixel-level, spatiotemporal changes in stimuli, but also on internal representations of semantic structure; in this case, the structure of familiar bodily actions.

6. Experiment 4b: Low-level static features

To examine the effect of pixel-level change involved in the image shift paradigm, we again used spiralized images in a task similar to Experiment 3, allowing us to compare the boundary effect in normal images versus their spiralized counterparts.

6.1. Method

6.1.1. Participants

As stated in our preregistration, 35 participants were recruited through Prolific.

6.1.2. Stimuli and procedure

The experimental design was the same as Experiment 3, except that two types of images were used: normal versus spiralized. We quantified pixel-level change at the pre-boundary, boundary, and post-boundary frames by calculating the number of changed pixels between the last two images in the sequence (2 frames apart in the original video). Across the 20 actions, pixel change was numerically greater at the pre- and post-boundary frames compared to the boundary frame. For normal videos, the average number of changed pixels was 4231 (pre-boundary), 4018 (boundary), and 4188 (post-boundary). For spiralized videos, the corresponding values were 4816 (pre-boundary), 4570 (boundary), and 4778 (post-boundary). However, the pixel-level difference between boundary and non-boundary frames were not statistically significant (all $Ps \geq 0.1$). The unique contribution of pixel-level features were further examined in the exploratory analysis, as reported below.

The Image Type condition was crossed with other conditions as in previous experiments, yielding 20 (actions) × 3 (pre-, on-, post-boundary) × 2 (shift, no shift) × 2 (normal, spiralized image) + 4 practice trials = 244 trials. Participants were instructed to detect whether the last image in the sequence was identical to the one immediately preceding it. Trial order was randomized across participants (except for practice trials).

6.2. Results and discussion

One participant was excluded for failing to submit complete data, leaving 34 participants with analyzable data, with a mean accuracy of 72.4%. Thirty-four observers attempted to detect image changes across both conditions.

We again observed a significant boundary effect in the normal images ($t(33) = 10.31$, $p = 7.52 \times 10^{-12}$, $d = 1.77$, 95% $CI_{effects} = .21[.17, .25]$), and also observed this effect in the spiralized images ($t(33) = 5.04$, $p = 1.63 \times 10^{-5}$, $d = 0.86$, 95% $CI_{effects} = .11[.065, .15]$); however, the effect was about half as strong in the spiralized images versus the normal images ($t(33) = 3.59$, $p = .0011$, $d = 0.62$, 95% $CI_{difference} = .10[.045, .16]$; Fig. 5C), echoing the results of Experiment 4a.

We then conducted two follow-up exploratory analyses. First, a 2 (normal, spiralized) × 3 (pre-boundary, boundary, post-boundary) repeated-measures ANOVA revealed significant main effects of Testing Frame, $F(2, 66) = 50.29$, $p < .001$, $\eta_p^2 = .60$, and Image Type, $F(1, 33) = 40.30$, $p < .001$, $\eta_p^2 = .55$. Importantly, there was also a significant Testing Frame × Image Type interaction, $F(2, 66) = 7.73$, $p < .001$, $\eta_p^2 = .19$, indicating that the effect of frame position varied across image types. Post hoc comparisons showed that in the normal condition, accuracy was significantly lower at the boundary compared to both pre-boundary ($p_{adjusted} < .0001$) and post-boundary ($p_{adjusted} < .0001$) frames. Similarly, in the spiralized condition, boundary accuracy was significantly lower than both pre-boundary ($p_{adjusted} = .0062$) and post-boundary ($p_{adjusted} = .0002$), but the differences were smaller in magnitude. As can be seen in Fig. 5C, observers were significantly more accurate in the normal condition than the spiralized condition at both the pre-boundary ($p_{adjusted} < .0001$) and post-boundary ($p_{adjusted} < .0001$).

Second, a multiple linear regression was conducted to examine the extent to which Image Type, Testing Frame, and Pixel Change accounted for variance in response accuracy. The overall model was significant, $F(3, 116) = 8.14$, $p < .001$, $R^2 = .174$. All three predictors made significant contributions: accuracy was lower for twisted videos compared to normal ones ($\beta = -0.11$, $p = .010$), higher at non-boundary frames ($\beta = 0.073$, $p = .001$) compared to boundaries, and increased slightly with greater pixel change ($\beta = 0.00005$, $p = .041$). Relative importance analysis indicated that frame position accounted for the largest share of explained variance (50.5%), followed by video type (27.8%) and pixel change (21.7%). The results revealed a robust unique contribution of image type to the boundary effect.

Boundary effects were present under both conditions but were more pronounced in normal action videos. The difference of boundary effects between normal versus spiralized stimuli mostly came from observers' lower accuracy in detecting changes of spiralized non-boundary frames (Fig. 5), suggesting that observers did better predicting the flow of actions when semantic information was present, and thus were more sensitive to subtle changes within action steps. Taken together, these results suggest that in addition to pixel-level changes and motion cues, mental representations of internal structures in atomic actions may be involved in rapid, spontaneous visual segmentation.

7. Experiment 5a: Point-light displays

Our last two experiments aimed to replicate and extend our findings to the case of point-light displays. Although the twisting procedure used in Experiment 4a and 4b maintained large-scale pixel-level and spatiotemporal changes in the action videos, it inevitably altered the percept of biological motion and other visual features, like motion direction and figure rigidity. In this experiment, we re-created the 20 actions as "point-light walkers" (Johansson, 1973; Van Boxtel & Lu, 2013) and attempted to replicate the previous effects. Each action video now consisted of a number of coordinated moving points that represent the joints of the human figures. In this experiment, we presented PLWs either upright or upside-down. Inverted PLWs serve as an interesting case wherein the semantic information of actions is only partially preserved and can impede recognition (Shipley, 2003; Sumi, 1984). We asked whether the boundary effect emerged in visually minimal PLW stimuli, and whether the effect differed between upright versus inverted stimuli.

7.1. Method

7.1.1. Participants

Consistent with our preregistration, 80 participants were recruited through Prolific. We increased the sample size for this experiment based on the preliminary result of a small pilot study.

7.1.2. Stimuli and procedure

This experiment transformed the 20 original action videos into "point-light walker" (PLW) stimuli. This was implemented using the BioMotion MATLAB toolbox (Van Boxtel & Lu, 2013). The number of joints used to define actions ranged from 20 to 30. This new set of videos depicted the same 20 actions as the ones used in previous experiments, yet were slightly different in the number of frames as a higher sampling rate was used. The boundary frame was hand-selected as the frame that had the matching posture as the boundary frame used in previous experiments, and the pre- and post-boundary frames were selected as before.

The PLW stimuli were presented either upright or upside-down. To reduce the number of repetitions of each action video (which could lead to recognition of the inverted stimuli), we asked participants to view each action only once in each trial and detect if the video was briefly frozen at any point. A 120-ms pause was applied to pre-, on-, and post-boundary frames of each action video. All 20 actions appeared for each of these trial types, for 20 (actions) × 3 (pre-, on-, post-boundary) × 2 (with, without pause) × 2 (upright, inverted video) + 4 practice trials = 244 trials. Crucially, we note that because the experiments were crowd-sourced, we could not monitor participants for any behavior that would change the perceived orientation of the video clips (e.g., head-tilting).

7.2. Results and discussion

18 participants were excluded for low accuracy (< 57%) and 1 participant for failing to submit a complete data set, leaving 61 participants with analyzable data, with a mean accuracy of 66.1%. The boundary effects arose similarly in both the upright videos ($t(60) = 13.34$, $p = 1.81 \times 10^{-19}$, $d = 1.72$, 95% $CI_{effects} = .24[.21, .28]$) and inverted videos ($t(60) = 11.35$, $p = 1.82 \times 10^{-16}$, $d = 1.47$, 95% $CI_{effects} = .21[.17, .25]$). As predicted, the boundary effect was numerically stronger in upright videos versus inverted videos (24.4% versus 21.1%), however, the difference was only marginally reliable given our preregistered criteria ($t(60) = 1.75$, $p = .085$, $d = 0.22$, 95% $CI_{difference} = .033[-.004, .07]$; Fig. 6A). An exploratory analysis using Bayesian paired-sample t-test showed weak evidence for the null hypothesis; $BF_{10} = 0.59$, using JZS prior with scale $\sqrt{2}/2$.

A 2 (upright, inverted) × 3 (pre-boundary, boundary, post-boundary) repeated-measures ANOVA revealed a robust main effect of Testing Frame, $F(2, 120) = 120.44$, $p < .001$, $\eta_p^2 = .67$. There was no main effect of Video Type, $F(1, 60) = 0.12, p = .73$, $\eta_p^2 = .002$. However, there was a significant Testing Frame × Video Type interaction, $F(2, 120) = 3.82$, $p = .025$, $\eta_p^2 = .06$, suggesting that the effect of condition was indeed modulated by orientation. Post-hoc comparisons revealed that in the upright condition, accuracy at the boundary was significantly lower than both the pre-boundary ($p_{adjusted} < .0001$) and post-boundary frames ($p_{adjusted} < .0001$). In the inverted condition, a similar pattern was observed: accuracy at the boundary was significantly lower than both pre-boundary ($p_{adjusted} < .0001$) and post-boundary ($p_{adjusted} < .0001$). This result implies that spatiotemporal information significantly contributes to perceptual segmentation in this set of stimuli. It may be that partially preserved semantic information in inverted actions — particularly when it is potentially recovered through repeated exposure to the stimuli — as well as the lack of control for head tilting movements may help explain why the boundary effect remained comparably robust in the inverted condition.

8. Experiment 5b: Scrambled PLWs

Upside-down PLW actions are less recognizable yet still preserve partial semantic information (e.g., people may see still actions but misperceive some properties; see Barclay et al., 1978; Sumi, 1984). Is there a more effective way to eliminate semantic information in PLW stimuli that avoids these limitations? Here, we shuffled the initial positions of the "joints" such that the configuration of the body shape no longer resembled the human figure, while the local joint motion and rigidity was maintained. This method fully removes the semantic content related to human actions while preserving the unique motion patterns of the biological movements (producing a bizarre action made by some kind of "creatures", see Chen et al., 2022). In this case, observers lack internal knowledge for the event structure but can still rely on biological motion cues. Thus, from the observer's perspective, the spatiotemporal motion perceived in the intact point-light displays is preserved in the scrambled videos, but the stimuli lose semantic meaning.

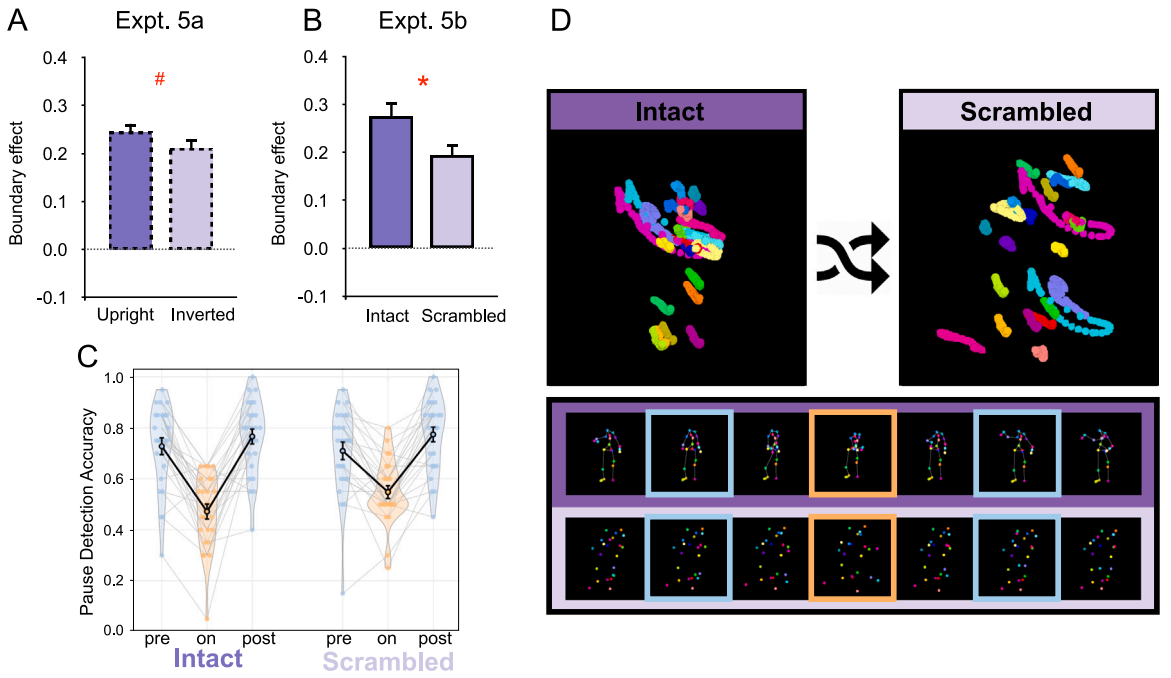


Fig. 6. (A) Results of Experiment 5a. The reduction of boundary effect in upside-down point-light walker stimuli was marginally significant relative to upright stimuli. (B&C) Results of Experiment 5b. The reduction of accuracy occurred at boundary for both intact and scrambled videos, but the size of reduction differed between two conditions. The boundary effect was weaker when the global configuration of the observed agent was disrupted. (D) Example stimuli used in Experiment 5b. For a given action, the movements of all the individual points (joints) were identical between its intact form and scrambled form. Here, each color represents a unique joint, and the exact movement trajectory of each colored joint across all frames was identical between two types of stimuli. Similar to Experiment 2, a brief pause could occur at the pre-boundary, boundary, or post-boundary frame. (Note that all the visual points of the actual stimuli were colored yellow as in the original skeletonized stimuli, and the underlying skeleton, indicated by the thin lines in the above figure, was not visible to observers.) The boundary effect was calculated as the difference in accuracy between non-boundary and boundary conditions. Error bars = 1 s.e.m. Asterisks indicate significant differences between means for two-tailed t-test ($*p < .05$; $\#p < .1$).

8.1. Method

8.1.1. Participants

Consistent with our preregistration, 35 participants were recruited through Prolific.

8.1.2. Stimuli and procedure

The point-light videos used in Experiment 5a were made into "scrambled" versions, where the initial positions of the joints were randomly selected — separately by x, y and z — from the ranges of width, height, and depth of the original videos (Fig. 6A). Readers can experience the intact and scrambled stimuli at https://zk.actlabresearch.org/segmentation/plw.html.

The experimental design was the same as Experiment 4a, except that: (1) point-light stimuli (normal and scrambled videos) were used, and (2) the pause duration was always 75 ms (to counter the overall increased difficulty of the task). Similar to Experiment 4a, all 20 actions appeared for each of these trial types, for 20 (actions) × 3 (pre-, on-, post-boundary) × 2 (with, without pause) × 2 (normal, scrambled video) + 4 practice trials = 244 trials. Trial order was randomized across participants (except for practice trials).

8.2. Results

Ten participants were excluded for low accuracy (< 57%), leaving 25 participants with analyzable data, with a mean accuracy of 76.6%. We again observed robust boundary effects in both the intact ($t(24) = 9.02$, $p = 3.53 \times 10^{-9}$, $d = 1.80$, 95% $CI_{effects} = .28[.22, .33]$) and scrambled videos ($t(24) = 8.11$, $p = 2.45 \times 10^{-8}$, $d = 1.62$, 95% $CI_{effects} = .19[.15, .24]$). Critically, these boundary effects were reduced in the scrambled videos, where semantic action structure was not perceivable ($t(24) = 2.22$, $p = .036$, $d = 0.44$, 95% $CI_{difference} = .081[.01, .15]$; Fig. 6B).

We also conducted a 2 (intact, scrambled) × 3 (pre-boundary, boundary, post-boundary) repeated-measures ANOVA, as a secondary analysis.[3] This revealed a strong main effect of Testing Frame, $F(2, 48) = 55.20$, $p < .001$, $\eta_p^2 = .70$, and a marginal Testing Frame × Video Type interaction, $F(2, 48) = 2.83$, $p = .069$, $\eta_p^2 = .11$. The main effect of video type was not statistically significant, $F(1, 24) = 2.46$, $p = .13$, $\eta_p^2 = 0.09$. Follow-up comparisons revealed that, for intact videos, accuracy at the boundary was significantly lower than both the pre- ($p_{adjusted} < .0001$) and post-boundary moments ($p_{adjusted} < .0001$). For scrambled videos, accuracy at the boundary was also significantly lower than both the pre- ($p_{adjusted} = .0001$) and post-boundary ($p_{adjusted} < .0001$). The numerically larger boundary drop for intact videos (e.g., accuracy at boundary between intact and scrambled conditions, $p_{adjusted} = .085$) is consistent with the observed interaction effect (Fig. 6C).

Though biological motion is thought to be processed as a special category (Blakemore & Decety, 2001), it cannot easily explain the stronger boundary effect observed in the real actions relative to the scrambled biological movements. Thus, while we observed a significant contribution of spatiotemporal dynamics and biological motion to the visual segmentation (i.e., significant boundary effects in the scrambled condition), our results demonstrate that semantic knowledge may also play a significant role in the perception of action event structure.

9. Discussion

In this work, we asked if the event structure *within* single, brief "atomic" actions is represented in visual perception, and if this representation involves semantic knowledge. Using motion-captured video stimuli, static action images, and biological motion stimuli, we consistently observed that the transition of an action's salient internal sub-routines impaired observers' ability to visually detect subtle changes in the stimuli, suggesting that the event structure within these brief actions is indeed represented in the visual system. Control studies suggested that this boundary representation is not solely driven by basic visual features of the stimulus that happen to change at the boundary (e.g., spatiotemporal features and motion cues), but also by internal models of the actions themselves. That is, semantic information of human actions, even very brief actions, plays a key role in driving perceptual effects at event boundaries, and thus may support automatic segmentation in visual processing.

9.1. Breaking down "minimal events"

In their seminal work, Zacks and Tversky (2001) raised two fundamental questions about event perception: What are the basic units of actions, and what are events? They proposed that the atomic components of events are *basic actions*, such as raising one's hand or moving one's head. Later empirical studies supported this view, showing that atomic actions tend to be single actions with well-defined starts and ends, similar to those used here. When observers segment streams of behavior into coarse and finer segments, fine-grained boundaries often arise between single actions like "walks into the room", "open door", "take out food", etc., as in Zacks et al. (2001). More recently, it has been proposed that bounded events (e.g., "scoop up yogurt", "fold a handkerchief", "roll up a towel") constitute the basic units or "mental individuals" in the temporal domain (Lee et al., 2024, see a related work about "objecthood" of auditory events in Goh et al., 2025). This idea, that events can be considered objects in time, dates back to earlier philosophical analyses (Quine, 1985).

In contrast to a large body of literature that has explored event segmentation *between* distinct actions (Baldwin et al., 2008; Buchsbaum et al., 2015; Franklin et al., 2020; Lea et al., 2016; Newtson, 1973; Newtson et al., 1977; Pomp et al., 2024; Swallow et al., 2009; Wang et al., 2013; Zacks et al., 2009), the present work broke down the minimal units in event perception and focused on event boundaries that occur *within* short, bounded (atomic) actions (Ji & Papafragou, 2022; Vendler, 1957). Through this approach, we demonstrate that atomic actions are not indivisible events; but instead, they are represented as having their own internal event structure. Consistent with prior work that conceptualized objects and events as individuated regions in space and time, respectively, our results might point to an even deeper parallel between objects and events: As the visual system represents a compositional spatial structure within single objects (Biederman, 1987; Feldman & Singh, 2006; Sun & Firestone, 2021), it may also represent components and internal structure within single actions in the time domain.

Our finding arguably fits with the recent resurgence of interest in ideas like the Language-of-Thought approach to vision, which proposes multiple types of compositional and structured representations within perception (Hafri et al., 2024; Papeo et al., 2024; Quilty-Dunn et al., 2023). While most evidence for LoT-like representation in perception has focused on the spatial domain (Hafri & Firestone, 2021; Lovett & Franconeri, 2017; Strickland & Scholl, 2015; Sun et al., 2025), the current study presents preliminary evidence that visual perception spontaneously represents basic episodes of experiences as discrete, inherently structured events unfolding over time. By extending the idea that visual perception may exhibit LoT-like properties to the domain of temporal structure, this work points to a broader presence of compositional and structured representations in perception.

[3] In addition to the main analysis, we reported repeated-measures ANOVA across participant means for detection accuracy for Experiment 4a-b and 5a-b. Note that we determined sample size only using power analyses on the main results (i.e., the paired t-tests on boundary effects) using pilot studies. Thus, the exploratory ANOVAs in some studies likely do not have sufficient power with the current sample size.

9.2. Perceptual effects in event segmentation

Our study builds on a number of studies that have discussed the spontaneous nature of representing event boundaries in perception. A number of experiments have shown that the representation of the transition between events interferes with the perceptual detection of subtle disruptions, resulting in a lower detection accuracy in detecting disruptions at boundary relative to non-boundary timepoints, seen in both visual perception (Huff et al., 2012; Ji & Papafragou, 2022; Yates et al., 2024) and auditory perception (Repp, 1992, 1998). For example, observers are less sensitive to visual interruptions at the end points of events (e.g., a girl folding her handkerchief) compared to the middle points (Ji & Papafragou, 2022; but also see Hard et al., 2019; Ongchoco, Chun, & Bainbridge, 2023; Swallow et al., 2009 and Footnote #2 for middle frame effects that may interact with boundary effects). This perceptual effect also is found in physical events: The boundary between simple physical events (such as collision, containment, or falling) has been found to impede the processing of distracting information (Yates et al., 2022). Such results have been explained by the rise of prediction error at the boundary and subsequent enhanced attention to event information (Antony et al., 2021; Pradhan & Kumar, 2022; Reynolds et al., 2007; Zacks et al., 2011, 2007).

However, in all of the above studies, feature, motion, and movement information was inevitably correlated with the event structure in ongoing visual input. Our study thus contributes to and builds on this line of work by attempting to dissociate two possible mechanisms that may account for the perceptual consequences of event segmentation: spatiotemporal dynamics and motion cues within the stimulus, and internal representations of semantic event structure. Our findings suggest that predictive processes that define discrete events during visual perception may employ both visual features and semantic content to generate predictions.

While our study built on the logic of previous studies which focused on the perception of disruptions or distractions at event boundaries (Huff et al., 2012; Repp, 1992; Yates et al., 2024), event structure and boundary can influence perception in other ways. For example, an early study found that observers were more accurate at detecting frame deletion when the missing frames occurred at "breakpoints" in complex action sequences (e.g., a person setting out tools), compared to non-breakpoints (Newtson, 1973). Enhanced perceptual encoding at an event boundary has also been shown through improved recognition performance in later memory tasks (Swallow et al., 2009). The interpretation of these effects as reflections of attentional or predictive processes could be better fleshed out; future work could look for other perceptual effects at boundaries in observed actions, such as increased processing of certain features versus others (Baker & Levin, 2015; Yates et al., 2024), and the creation of temporal distortions (Goh et al., 2025; Ongchoco, Yates, & Scholl, 2023).

9.3. Semantic structure in action perception

Though semantic knowledge is the key source of information underlying event cognition (e.g., "Sally was sitting, now she is standing"), it was not a foregone conclusion that the automatic perceptual boundaries in our atomic action stimuli would also be shaped by semantics. First, the action segmentation in our study was even more rapid and finer-grained than those described as "fine" units in previous work (Hard et al., 2011; Yates et al., 2024; Zacks et al., 2009; Zacks & Tversky, 2001). Fine boundaries are thought to be more perceptually determined and driven primarily by external changes of pixel-level movement and motion (Hard et al., 2006; Papeo et al., 2024; Zacks et al., 2009). Second, our work used simple, abstract animations of actions in a stripped-down context, rather than the complex, semantically-rich scenes in naturalistic stimuli. Prior studies that used more abstract, simple stimuli like ours in segmentation tasks (e.g., videos of moving geometric shapes) have revealed a stronger relation between boundary judgments and visual features within the stimuli (such as movement, motion features and dynamics), accompanied by increased activity in motion-related brain regions at event boundaries (Hard et al., 2006; Pomp et al., 2024; Zacks et al., 2009, 2006). Indeed, our results using multiple forms of control stimuli consistently pointed to rather robust effects of pixel-level and motion features. However, we also found that action schema appeared to contribute to visual event segmentation.

Our work sheds new light on our understanding of the mental representation of motor action structure. Some actions in our study exhibited a "preparation" stage followed by an "execution" phase, for example, the "focus" and "shoot" phases in a basketball shot. Empirical studies have revealed this preparation-execution structure and its function in facilitating action recognition and anticipation in ordinary actions (Cohn et al., 2017; Urgesi et al., 2010) and in athletic skills (Aglioti et al., 2008; Lasher, 1981; Smith, 2016; Urgesi et al., 2012). Yet, for most actions we examined, segmentation generally reflected discrete yet structured steps to complete an action. For example, the first step of actions such as a "star jump", a golf swing, and a punch typically belongs to the execution phase, occurring after the preparation phase highlighted in previous research (e.g., Lasher, 1981; Urgesi et al., 2012). Theoretical work in semantic analysis has applied a more richly-structured framework of action steps that may better explain our results. For instance, Jackendoff (2007) suggested that simple, routine actions (e.g., shaking hands) are mentally stored in a semantic hierarchy typically composed of three components: preparation (reaching and grasping), head (shaking), and coda (releasing and withdrawing). Speculatively, our findings suggest that such semantic structures exist not only as conceptual schema but perhaps also within perceptual systems.

Our findings leave open important questions about action segmentation for future research. First is the connection between segmentation and prediction in the context of human actions. Given evidence that early kinematic cues can predict the outcome of actions (e.g., Aglioti et al., 2008; Cohn et al., 2017), it is worth asking whether and how different ways of segmenting actions might alter an agent's ability to generate accurate predictions about how an action will unfold. Related questions may include whether finer segmentation ability precedes improved performance in anticipating future actions, and whether skilled athletes are also more efficient in segmenting others' actions. Second, our study employed salient, simple, familiar actions; this was by design, to make the semantic structure overdetermined. In future work, we could ask how more novel or unfamiliar actions are segmented, or how

learning about new actions may affect the perception of event boundaries. Finally, how our findings connect to motor learning of actions themselves could provide insights into how and why people divide simple motor routines into discrete steps. For example, in learning a new motor skill like a tennis serve, coaches and teachers tend to divide an otherwise smooth, rapid movement into multiple discrete "chunks" based on an internal structure (Fitts, 1964; Sakai et al., 2003). How do we decide where the chunk boundary should be?

9.4. Conclusion

In sum, this work explored the perception of event structure within single atomic actions: our results suggest that the mind represents rapid, continuous actions as containing discrete, inherently structured steps. These structured representations appear to emerge automatically over time during visual processing, and are driven by both visual motion cues and the semantic structure of actions.

CRediT authorship contribution statement

Zekun Sun: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Samuel D. McDougle:** Writing – review & editing, Supervision, Methodology, Investigation, Funding acquisition, Conceptualization.

Acknowledgments

For helpful discussions and comments on this work, the authors wish to thank Wenyan Bi, Ilker Yildirim, Brian Scholl, and the members of the Action, Computation and Thinking Laboratory at Yale. We also thank Jeffrey M. Zacks, Joan Danielle K. Ongchoco, and one anonymous reviewer for their constructive and thoughtful comments on the manuscript. This work was supported by grant R01 NS134754 (S.D.M.) from the National Institutes of Health.

Data availability

An archive of the data, code, stimuli, preregistrations, and other relevant materials is available at: https://osf.io/j85fa.

References

Aglioti, S. M., Cesari, P., Romani, M., & Urgesi, C. (2008). Action anticipation and motor resonance in elite basketball players. *Nature Neuroscience*, *11*(9), 1109–1116.
Antony, J. W., Hartshorne, T. H., Pomeroy, K., Gureckis, T. M., Hasson, U., McDougle, S. D., & Norman, K. A. (2021). Behavioral, physiological, and neural signatures of surprise during naturalistic sports viewing. *Neuron*, *109*(2), 377–390.
Baker, L. J., & Levin, D. T. (2015). The role of relational triggers in event perception. *Cognition*, *136*, 14–29.
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.
Baldwin, D., Andersson, A., Saffran, J., & Meyer, M. (2008). Segmenting dynamic human action via statistical structure. *Cognition*, *106*(3), 1382–1407.
Baldwin, D. A., Baird, J. A., Saylor, M. M., & Clark, M. A. (2001). Infants parse dynamic action. *Child Development*, *72*(3), 708–717.
Barclay, C. D., Cutting, J. E., & Kozlowski, L. T. (1978). Temporal and spatial factors in gait perception that influence gender recognition. *Perception & Psychophysics*, *23*, 145–152.
Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychological Review*, *94*(2), 115.
Blakemore, S.-J., & Decety, J. (2001). From the perception of action to the understanding of intention. *Nature Reviews. Neuroscience*, *2*(8), 561–567.
Bregman, A. S. (1994). *Auditory scene analysis: The perceptual organization of sound*. MIT Press.
Buchsbaum, D., Griffiths, T. L., Plunkett, D., Gopnik, A., & Baldwin, D. (2015). Inferring action structure and causal relationships in continuous sequences of human action. *Cognitive Psychology*, *76*, 30–77.
Chen, Y.-C., Pollick, F., & Lu, H. (2022). Aesthetic preferences for causality in biological movements arise from visual processes. *Psychonomic Bulletin & Review*, *29*(5), 1803–1811.
Chen, Y.-C., & Scholl, B. J. (2016). The perception of history: Seeing causal history in static shapes induces illusory motion perception. *Psychological Science*, *27*(6), 923–930.
Cohn, N., Holcomb, P., Jackendoff, R., & Kuperberg, G. (2012). Segmenting visual narratives: evidence for constituent structure in comics. In *Proceedings of the annual meeting of the cognitive science society*: *vol. 34*, (34).
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.
Cohn, N., Paczynski, M., & Kutas, M. (2017). Not so secret agents: Event-related potentials to semantic roles in visual event comprehension. *Brain and Cognition*, *119*, 1–9.
Comrie, B. (1976). Aspect: An introduction to the study of verbal aspect and related problems. *Cambridge UP*.
Dasser, V., Ulbaek, I., & Premack, D. (1989). The perception of intention. *Science*, *243*(4889), 365–367.
Ezzyat, Y., & Clements, A. (2024). Neural activity differentiates novel and learned event boundaries. *Journal of Neuroscience*, *44*(38).
Feldman, J., & Singh, M. (2006). Bayesian estimation of the shape skeleton. *Proceedings of the National Academy of Sciences*, *103*(47), 18014–18019.
Fitts, P. M. (1964). Perceptual-motor skill learning. In *Categories of human learning* (pp. 243–285). Elsevier.
Franklin, N. T., Norman, K. A., Ranganath, C., Zacks, J. M., & Gershman, S. J. (2020). Structured event memory: A neuro-symbolic model of event cognition. *Psychological Review*, *127*(3), 327.
Goh, R. Z., Zhou, H., Firestone, C., & Phillips, I. (2025). Event-based warping: A relative distortion of time within events. *Journal of Experimental Psychology: General*.

Hafri, A., Bonner, M. F., Landau, B., & Firestone, C. (2024). A phone in a basket looks like a knife in a cup: Role-filler independence in visual processing. *Open Mind*, *8*, 766–794.
Hafri, A., & Firestone, C. (2021). The perception of relations. *Trends in Cognitive Sciences*, *25*(6), 475–492.
Hafri, A., Trueswell, J. C., & Strickland, B. (2018). Encoding of event roles from visual scenes is rapid, spontaneous, and interacts with higher-level visual processing. *Cognition*, *175*, 36–52.
Hard, B. M., Meyer, M., & Baldwin, D. (2019). Attention reorganizes as structure is detected in dynamic action. *Memory & Cognition*, *47*, 17–32.
Hard, B. M., Recchia, G., & Tversky, B. (2011). The shape of action. *Journal of Experimental Psychology: General*, *140*(4), 586.
Hard, B. M., Tversky, B., & Lang, D. S. (2006). Making sense of abstract events: Building event schemas. *Memory & Cognition*, *34*(6), 1221–1235.
Hemeren, P. E., & Thill, S. (2011). Deriving motor primitives through action segmentation. *Frontiers in Psychology*, *1*, 243.
Hespos, S. J., Saylor, M. M., & Grossman, S. R. (2009). Infants' ability to parse continuous actions. *Developmental Psychology*, *45*(2), 575.
Huff, M., Papenmeier, F., & Zacks, J. M. (2012). Visual target detection is impaired at event boundaries. *Visual Cognition*, *20*(7), 848–864.
Jackendoff, R. (2007). Shaking hands and making coffee. In *The structure of complex actions* (pp. 111–144). MIT Press Cambridge, MA,
Ji, Y., & Papafragou, A. (2022). Boundedness in event cognition: Viewers spontaneously represent the temporal texture of events. *Journal of Memory and Language*, *127*, Article 104353.
Ji, H., & Scholl, B. J. (2024). "Visual verbs": Dynamic event types are extracted spontaneously during visual perception. *Journal of Experimental Psychology: General*, *153*(10), 2441.
Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, *14*, 201–211.
Kumar, M., Goldstein, A., Michelmann, S., Zacks, J. M., Hasson, U., & Norman, K. A. (2023). Bayesian surprise predicts human event segmentation in story listening. *Cognitive Science*, *47*(10), Article e13343.
Kurby, C. A., & Zacks, J. M. (2008). Segmentation in the perception and memory of events. *Trends in Cognitive Sciences*, *12*(2), 72–79.
Lasher, M. D. (1981). The cognitive representation of an event involving human motion. *Cognitive Psychology*, *13*(3), 391–406.
Lea, C., Reiter, A., Vidal, R., & Hager, G. D. (2016). Segmental spatiotemporal cnns for fine-grained action segmentation. In *Computer vision–ECCV 2016: 14th European conference, amsterdam, the netherlands, October 11-14, 2016, proceedings, part III 14* (pp. 36–52). Springer.
Lee, S. H., Ji, Y., & Papafragou, A. (2024). Signatures of individuation across objects and events. *Journal of Experimental Psychology: General*, *153*(8), 1997.
Loucks, J., & Pechey, M. (2016). Human action perception is consistent, flexible, and orientation dependent. *Perception*, *45*(11), 1222–1239.
Lovett, A., & Franconeri, S. L. (2017). Topological relations between objects are categorically coded. *Psychological Science*, *28*(10), 1408–1418.
Mittwoch, A. (2013). On the criteria for distinguishing accomplishments from activities, and two types of aspectual misfits. In *Studies in the composition and decomposition of event predicates* (pp. 27–48). Springer.
Newtson, D. (1973). Attribution and the unit of perception of ongoing behavior. *Journal of Personality and Social Psychology*, *28*(1), 28.
Newtson, D., Engquist, G. A., & Bois, J. (1977). The objective basis of behavior units. *Journal of Personality and Social Psychology*, *35*(12), 847.
Ongchoco, J. D. K., Chun, M. M., & Bainbridge, W. A. (2023). What moves us? The intrinsic memorability of dance. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *49*(6), 889.
Ongchoco, J. D. K., Yates, T. S., & Scholl, B. J. (2023). Event segmentation structures temporal experience: Simultaneous dilation and contraction in rhythmic reproductions. *Journal of Experimental Psychology: General*.
Papeo, L., Vettori, S., Serraille, E., Odin, C., Rostami, F., & Hochmann, J.-R. (2024). Abstract thematic roles in infants' representation of social events. *Current Biology*, *34*(18), 4294–4300.
Pomp, J., Garlichs, A., Kulvicius, T., Tamosiunaite, M., Wurm, M. F., Zahedi, A., Wörgötter, F., & Schubotz, R. I. (2024). Action segmentation in the brain: The role of object–action associations. *Journal of Cognitive Neuroscience*, *36*(9), 1784–1806.
Pradhan, R., & Kumar, D. (2022). Event segmentation and event boundary advantage: Role of attention and postencoding processing. *Journal of Experimental Psychology: General*, *151*(7), 1542.
Quilty-Dunn, J., Porot, N., & Mandelbaum, E. (2023). The best game in town: The reemergence of the language-of-thought hypothesis across the cognitive sciences. *Behavioral and Brain Sciences*, *46*, Article e261.
Quine, W. V. O. (1985). Events and reification. *Actions and Events: Perspectives on the Philosophy of Donald Davidson*, 162–171.
Repp, B. H. (1992). Probing the cognitive representation of musical time: Structural constraints on the perception of timing perturbations. *Cognition*, *44*(3), 241–281.
Repp, B. H. (1998). Variations on a theme by chopin: Relations between perception and production of timing in music. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(3), 791.
Reynolds, J. R., Zacks, J. M., & Braver, T. S. (2007). A computational model of event segmentation from perceptual prediction. *Cognitive Science*, *31*(4), 613–643.
Richmond, L. L., & Zacks, J. M. (2017). Constructing experience: Event models from perception to action. *Trends in Cognitive Sciences*, *21*(12), 962–980.
Sakai, K., Kitaguchi, K., & Hikosaka, O. (2003). Chunking during human visuomotor sequence learning. *Experimental Brain Research*, *152*, 229–242.
Saylor, M. M., Baldwin, D. A., Baird, J. A., & LaBounty, J. (2007). Infants' on-line segmentation of dynamic human action. *Journal of Cognition and Development*, *8*(1), 113–128.
Scholl, B. J., & Tremoulet, P. D. (2000). Perceptual causality and animacy. *Trends in Cognitive Sciences*, *4*(8), 299–309.
Shin, Y. S., & DuBrow, S. (2021). Structuring memory through inference-based event segmentation. *Topics in Cognitive Science*, *13*(1), 106–127.
Shipley, T. F. (2003). The effect of object and event orientation on perception of biological motion. *Psychological Science*, *14*(4), 377–380.
Smith, D. M. (2016). Neurophysiology of action anticipation in athletes: A systematic review. *Neuroscience & Biobehavioral Reviews*, *60*, 115–120.
Strickland, B., & Scholl, B. J. (2015). Visual perception involves event-type representations: The case of containment versus occlusion. *Journal of Experimental Psychology: General*, *144*(3), 570.
Sumi, S. (1984). Upside-down presentation of the johansson moving light-spot pattern. *Perception*, *13*(3), 283–286.
Sun, Z., & Firestone, C. (2021). Curious objects: How visual complexity guides attention and engagement. *Cognitive Science*, *45*(4), Article e12933.
Sun, Z., Firestone, C., & Hafri, A. (2025). The psychophysics of compositionality: Relational scene perception occurs in a canonical order. *Cognitive Psychology*, [ISSN: 0010-0285] *161*, Article 101765.
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.
Ünal, E., Ji, Y., & Papafragou, A. (2021). From event representation to linguistic meaning. *Topics in Cognitive Science*, *13*(1), 224–242.
Urgesi, C., Maieron, M., Avenanti, A., Tidoni, E., Fabbro, F., & Aglioti, S. M. (2010). Simulating the future of actions in the human corticospinal system. *Cerebral Cortex*, *20*(11), 2511–2521.
Urgesi, C., Savonitto, M. M., Fabbro, F., & Aglioti, S. M. (2012). Long-and short-term plastic modeling of action prediction abilities in volleyball. *Psychological Research*, *76*, 542–560.
Van Boxtel, J. J., & Lu, H. (2013). A biological motion toolbox for reading, displaying, and manipulating motion capture data in research settings. *Journal of Vision*, *13*(12), 1–16.
Vendler, Z. (1957). Verbs and times. *The Philosophical Review*, *66*(2), 143–160.
Wang, H., Kläser, A., Schmid, C., & Liu, C.-L. (2013). Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision*, *103*, 60–79.

Yates, T. S., Sherman, B. E., & Yousif, S. R. (2023). More than a moment: What does it mean to call something an 'event'? *Psychonomic Bulletin & Review*, *30*(6), 2067–2082.
Yates, T. S., Skalaban, L. J., Ellis, C. T., Bracher, A. J., Baldassano, C., & Turk-Browne, N. B. (2022). Neural event segmentation of continuous experience in human infants. *Proceedings of the National Academy of Sciences*, *119*(43), Article e2200257119.
Yates, T. S., Yasuda, S., & Yildirim, I. (2024). Temporal segmentation and "look ahead" simulation: Physical events structure visual perception of intuitive physics. *Journal of Experimental Psychology: Human Perception and Performance*, *50*(8), 859.
Zacks, J. M. (2004). Using movement and intentions to understand simple events. *Cognitive Science*, *28*(6), 979–1008.
Zacks, J. M. (2020). Event perception and memory. *Annual Review of Psychology*, *71*(1), 165–191.
Zacks, J. M., Kumar, S., Abrams, R. A., & Mehta, R. (2009). Using movement and intentions to understand human activity. *Cognition*, *112*(2), 201–216.
Zacks, J. M., Kurby, C. A., Eisenberg, M. L., & Haroutunian, N. (2011). Prediction error associated with the perceptual segmentation of naturalistic events. *Journal of Cognitive Neuroscience*, *23*(12), 4057–4066.
Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., & Reynolds, J. R. (2007). Event perception: a mind-brain perspective. *Psychological Bulletin*, *133*(2), 273.
Zacks, J. M., Swallow, K. M., Vettel, J. M., & McAvoy, M. P. (2006). Visual motion and the neural correlates of event perception. *Brain Research*, *1076*(1), 150–162.
Zacks, J. M., & Tversky, B. (2001). Event structure in perception and conception. *Psychological Bulletin*, *127*(1), 3.
Zacks, J. M., Tversky, B., & Iyer, G. (2001). Perceiving, remembering, and communicating structure in events. *Journal of Experimental Psychology: General*, *130*(1), 29.
Zheng, Y., Zacks, J. M., & Markson, L. (2020). The development of event perception and memory. *Cognitive Development*, *54*, Article 100848.