Journal of Experimental Psychology: Human Perception and Performance

Temporal Segmentation and "Look Ahead" Simulation: Physical Events Structure Visual Perception of Intuitive Physics

Tristan S. Yates, Shannon Yasuda, and Ilker Yildirim

Temporal Segmentation and "Look Ahead" Simulation: Physical Events Structure Visual Perception of Intuitive Physics

Tristan S. Yates[1], Shannon Yasuda[2], and Ilker Yildirim[3, 4]

[1] Department of Psychology, Columbia University
[2] Department of Psychology, New York University
[3] Department of Psychology, Yale University
[4] Department of Statistics and Data Science, Yale University

How we perceive the physical world is not only organized in terms of objects, but also structured in time as sequences of events. This is especially evident in intuitive physics, with temporally bounded dynamics such as falling, occlusion, and bouncing demarcating the continuous flow of sensory inputs. While the spatial structure and attentional consequences of physical objects have been well-studied, much less is known about the temporal structure and attentional consequences of physical events in visual perception. Previous work has recognized physical events as units in the mind, and used presegmented object interactions to explore physical representations. However, these studies did not address whether and how perception imposes the kind of temporal structure that carves these physical events to begin with, and the attentional consequences of such segmentation during intuitive physics. Here, we use performance-based tasks to address this gap. In Experiment 1, we find that perception not only spontaneously separates visual input in time into physical events, but also, this segmentation occurs in a nonlinear manner within a few hundred milliseconds at the moment of the event boundary. In Experiment 2, we find that event representations, once formed, use coarse "look ahead" simulations to selectively prioritize those objects that are predictively part of the unfolding dynamics. This rich temporal and predictive structure of physical event representations, formed during vision, should inform models of intuitive physics.

Public Significance Statement

Despite the continuous flow of sensory inputs, our perceptual experiences are deeply structured in space in terms of objects and in time in terms of events. In many ways, most research in visual perception has focused on half of this structure: objects. To help reveal the structure of the other half, we turn to intuitive physics—our ability to see and predict how scenes react to forces—where events are especially evident. In intuitive physics, we do not just see objects with certain physical properties occupying certain places, but we experience dynamic patterns of interactions, for example, falling, colliding, bouncing, and entering and exiting containers. In two experiments with human observers, we synthesize methods and theories across the studies of attention, music cognition, seeing, and thinking, and find that the visual system spontaneously marks physical events at subsecond timescales and prioritizes objects relevant to the ongoing event.

Against the undelimited influx of sensory inputs arriving at the eyes, our perceptual experiences are richly structured, with objects and events delimiting discrete spatial and temporal entities. Such structure is especially evident in the visual perception of intuitive physics—our ability to see, at a glance, and predict how objects move and react to forces. Yet, in many ways, most previous research on intuitive physics has focused on only half of this structure: representations of physical objects. Among others, these studies have illustrated the perception of physical stability of object configurations (Battaglia et al., 2013; Pramod et al., 2022), whether an object is heavier or harder than others (Guo et al., 2020; Hamrick et al., 2016; Schwettmann et al., 2019), and how intuitive physics may be recruited for object perception (Wong et al., 2023). Yet, physical scenes inherently unfold over time, expressing discrete temporal structure in demarcated episodes of object interactions, including collision, bouncing, toppling, falling, and entering and exiting containers. Despite their centrality to our perceptual experience and their importance in the broader cognitive science literature (Gibson, 1954, 1975; Michotte, 1963; Spelke, 1990), the structure of physical event representations, as they occur in the visual perception of intuitive physics, remains largely unaddressed.

Consider the simple dynamics of a blue marble rolling on a crowded surface, depicted in Figure 1 as two snapshots over time. When observing the marble, various scene elements ebb and flow into and out of our awareness: We attend to the planks under and across the marble in the earlier moment, and the narrow passage and the green container that the marble will enter in the later moment. It is only natural to describe in language a scene of this sort in terms of a sequence of interactions and the objects involved in those interactions (as we just did), but here we ask: Does the visual system spontaneously mark such physical event boundaries, even during passive viewing of dynamical scenes? And if so, what are the temporal characteristics of such real-time perceptual segmentation of unfolding dynamics? Moreover, once a segmentation occurs, that is, between a pair of physical events, what principle(s) guide which objects will be prioritized as part of the current event representation? Answers to these questions will reveal the basic structure of physical events in vision, and the goal of this article is to address them by introducing new performance-based psychophysical tasks.
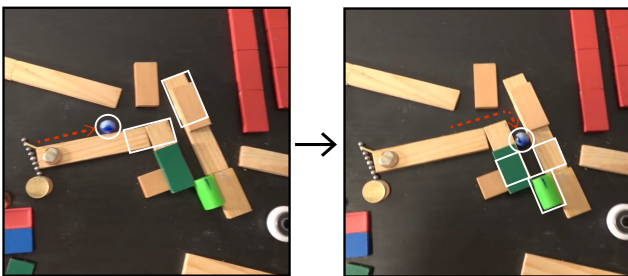
Physical Events as Units in the Mind

The study of physical events has a long history in cognitive science. Early work examined the perception of (real or apparent; rigid or nonrigid) object motion (Gibson, 1954), as well as how causality is perceived in so-called "launching" events (Michotte, 1963). These examples and others made clear that there is a distinction that can be made between physical events as they occur in the world and physical events as they occur in the mind—in other words, changes that are psychologically meaningful (e.g., the breaking of glass) may not necessarily be meaningful in classical physics (since matter is conserved; Gibson, 1975). People seem to have an intuitive understanding of how physical events unfold, even if their judgements do not always adhere to Newtonian physics (McCloskey, 1983). Landmark developmental research demonstrated that expectations about physical events are present early in infancy, suggesting that they may be part of a core system of knowledge (Spelke, 1990; Spelke & Kinzler, 2007).

Much like objects, physical events are only partially determined by the sensory data; rather, they emerge as units in the mind via perceptual processing, and provide a natural interface for attentional selection (Papafragou & Ji, 2023; Yousif & Scholl, 2019). Yet, previous work in intuitive physics either does not address physical events, or considers physical events as a given, as presegmented interactions in isolation: One object moves, one object falls, or one object is occluded. This is in contrast to how we normally encounter dynamic scenes, where an object in motion may participate in several different types of events, recruiting, at each event, a different causal context—varying subsets of all visible objects and surfaces relevant to ongoing dynamics. Even though using isolated physical events has led to key insights about physical representations (e.g., Gerstenberg et al., 2017; Karakose-Akbiyik et al., 2023; Little & Firestone, 2021), a full understanding of intuitive physics requires investigating how its temporal structure, that is, sequences of physical events, arises during the processing of continuous visual input, and how physical events impact attentional prioritization of the visible objects in a scene, particularly at the time scales relevant to fast perceptual processing. Therefore, to make progress in intuitive physics, it is essential to study the formation and contents of physical event representations during continuous visual processing.

Uncovering Physical Event Structure in Perception: The Current Study

In the broader event cognition literature, two main paradigms have been used to measure event segmentation: the unitization task (Newtson & Engquist, 1976; Radvansky & Zacks, 2017; Zacks, 2020; Zacks et al., 2007) and the dwell-time paradigm (Hard et al., 2011; Zheng et al., 2020). Both paradigms seek to uncover when human participants experience event boundaries during continuous, ongoing perception. In the unitization task, participants are simply asked to "press a button" or otherwise demarcate when they believe an event boundary occurs. Thus, as an explicit judgment task, the unitization task is not well-suited for revealing the spontaneous, automatic construction of event representations within visual perception. The dwell-time paradigm, on the other hand, is relatively more implicit: participants are simply asked to advance videos/stories at their own pace, and the more time that they "dwell" between frames or pages is seen as evidence of a boundary between events.

Figure 1

A Dynamical Scene With a Blue Marble Rolling on a Surface and Interacting With Other Objects, Shown in Two Snapshots

Note. The dashed red (dark grey) arrow shows the marble's previous trajectory, and white bounding boxes represent the objects relevant to the unfolding dynamics within each snapshot. Screenshot adapted from: https://www.youtube.com/watch?v=Hmb0Q0Q_7jo. See the online article for the color version of this figure.

However, this task may be insufficiently granular for fast timescale events like physical events, and may also interfere with relevant aspects of the stimuli (e.g., the velocity of objects). A different but related method comes from the domain of audition and music cognition. Repp (1992, 1998) showed that when participants listened to a musical piece where certain notes were extended in time relative to the original musical score, participants were less likely to detect these changes at the transitions (i.e., boundaries) between phrases. As we show in this article, this method suggests a paradigm for visual perception that is more objective via its performance-based manner, allowing us to determine if boundaries between physical events are spontaneously formed and segmented during seeing (see also Ji & Papafragou, 2022).

In perception, segmenting time is not the goal in itself, but a means to a more tractable setting for selective processing of otherwise complex, dynamic visual environments. For example, event representations guide which aspects of an object are prioritized: Subtle changes made to the relevant dimension of an object (height vs. width) with respect to the given event category (occlusion vs. containment) are differentially more detectable (Strickland & Scholl, 2015). Additionally, seeing an actor's reaction to invisible surfaces or objects implies their physical presence, influencing participants' responses to congruent or incongruent probes (Little & Firestone, 2021). How do events guide such selective processing in perception? One principle advanced and strongly supported in the event cognition literature is the idea of prediction (Reynolds et al., 2007; Zacks et al., 2007, 2011): Event models should compose of elements that will facilitate prediction during the epoch of that event, with boundaries between events signaled by a decreased ability to predict what will happen next, known as prediction errors. Applied to physical events, this principle suggests a selective processing mechanism that should take into account the simulated trajectories of objects (i.e., it should "look ahead" in time) to prioritize those objects that will impact dynamics within the locality of the current event.

In the current project, we investigate the basic structure of physical event representations within continuous perception by addressing (a) the time course of how a new event is formed—that is, when boundaries are imposed on continuous input—and (b) what determines the contents of an event representation. Experiment 1 generalizes the performance-based tasks developed in the domain of music cognition (Repp, 1992, 1998) to the domain of seeing dynamical scenes. Based on prior work, we predict that participants' detection accuracy for transient slow-downs that occur during videos of multiple physical events will decrease at event boundaries relative to nonboundary intervals (although it is also possible that detection accuracy may increase at boundaries; Baker & Levin, 2015). This task allows us to investigate the subsecond time course of event segmentation in terms of how detection accuracies rise and fall, and the symmetry of this pattern along the event boundary. In Experiment 2, we predict that participants will be better at detecting "simulation" probes—briefly presented letters on stationary objects in dynamical scenes—if they appear on objects that are along the simulated trajectory of the current event, relative to spatially controlled counterparts. This task allows us to test the hypothesis that physical event representations that are formed spontaneously during vision guide selective processing via prediction. Additionally, and consistent with event segmentation theory (Zacks, 2020), we predict that selective processing of relevant objects will be lower at event boundaries compared to event middles, since attention allocated to the transition between event models may impact

performance on a secondary task (see also Huff et al., 2012; Pradhan & Kumar, 2022). Critically, in both experiments, participants are not informed of physical events and are simply tasked with paying attention to either temporal (Experiment 1) or simulation (Experiment 2) probes. Thus, our study design should tap into automatic, visual processes and help reveal the structure of physical event representations as they are formed during vision.

Experiment 1

In Experiment 1, we examined how and when physical events are spontaneously segmented as part of visual processing. To reveal the time course of how physical event segmentation occurs at subsecond resolution, we showed participants videos of multiple physical events and asked them to indicate if and when during the video they noticed a temporary slowing of the video.

Method

Participants

Per our preregistration, we recruited 110 participants from Prolific, such that there were 11 groups of 10 participants who saw the same set of videos. Data were collected in 2021. We excluded one participant for not having at least half of the trials usable. For our subject-wise analyses, one additional participant was excluded for not having enough event boundary trials usable (see below for preregistered trial exclusion criteria). This number of participants does not include data from an additional 29 participants who started the task but did not complete all of the test trials. Included participants ($N = 50$ female, $N = 58$ male, $N = 2$ unknown) were between 20 and 37 years of age ($M_{age} = 28.05$ years) and were majority white ($N = 80$ White, $N = 7$ Black, $N = 9$ Asian, $N = 9$ multiple ethnicities, $N = 2$ other, $N = 3$ unknown) from the United States or United Kingdom ($N = 32$ United States, $N = 76$ United Kingdom, $N = 2$ unknown). Results should therefore be interpreted with these sample demographics in mind.

Stimuli

We created multievent videos using a combination of Python (Van Rossum & Drake, 2009) and Blender (Community, 2018) software, inspired by recent work on causal graph structures in physics representations (Roussel et al., 2019). Specifically, we modified the code base created by Roussel and colleagues to compose videos where objects interacted with one another across multiple different physically plausible events. We focused our research on a handful of event types that were simple and easily identifiable by human participants: collision (two objects collide), occlusion (one object goes behind another or becomes visible again), containment (one object enters another object), falling (one object falls from a ramp), and toppling (one unstable object hits the ground; see Figure 2A). For the current experiment, we focused on the former four event types, as they could easily be strung together into every possible two-event combination (e.g., an object falls and then collides with another object; an object is occluded and then contained by another object, etc.). We created two unique examples for each two-event combination for a total of 32 "base" videos (mean video duration: 9.75 s; range: 6.23–14.3 s). Additionally, we added a 3-s countdown to

Figure 2
Design of the Two Experiments Presented in This Article
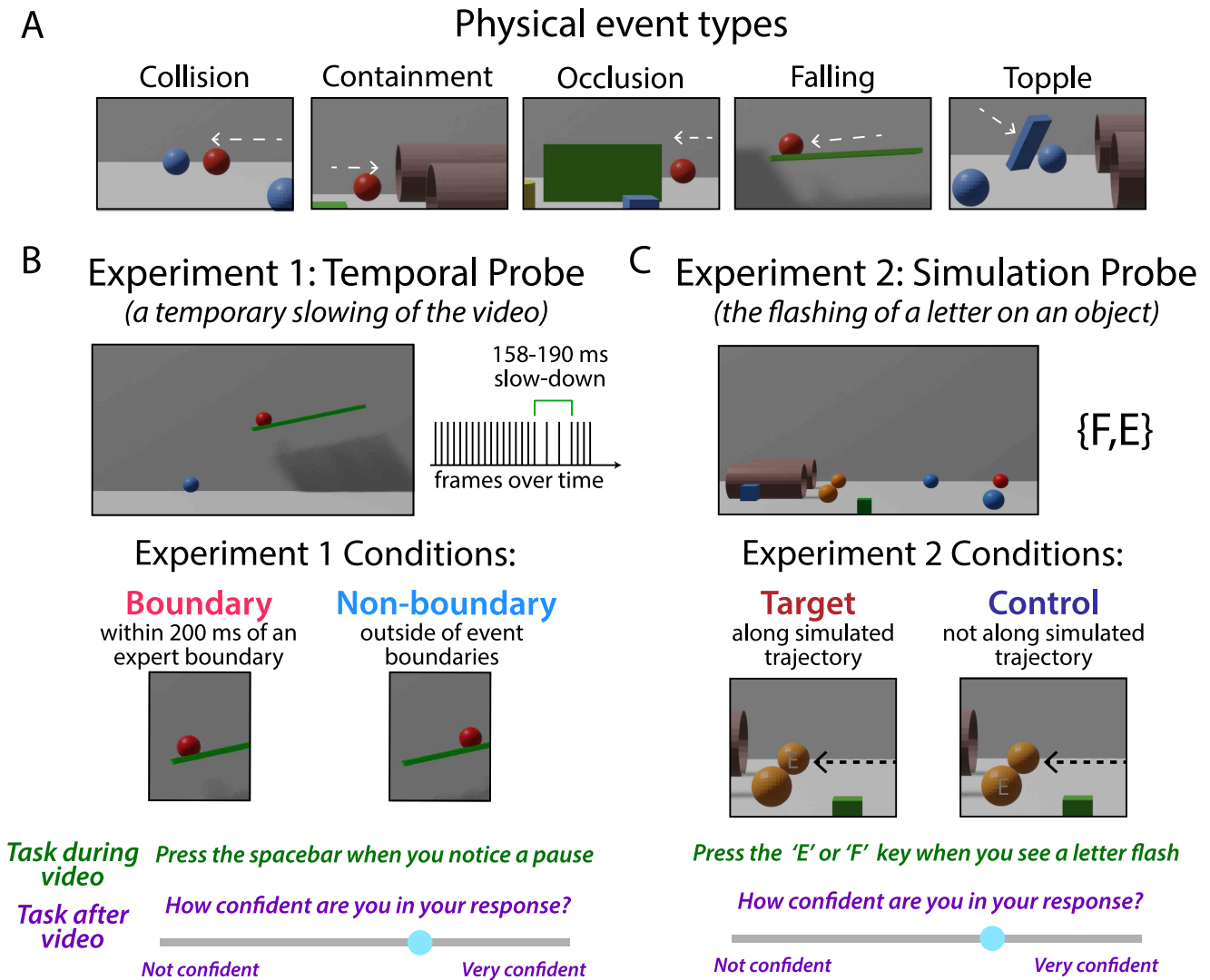
Note.   (A) In both experiments, participants saw videos that consisted of a sequence of two or more different physical event types: collision, containment, occlusion, falling, and topple. Dashed arrows indicate the motion of the target object—the object that is set in motion initially in each scene—and were not shown during the actual experiment. (B) In Experiment 1, participants watched videos with or without temporal probes (temporary slowing of the video frame rate for between 158 and 190 ms) that occurred either at expert-determined event boundaries (e.g., the onset of a ball falling off of a ramp) or at other points in the video. Participants were tasked with pressing the spacebar when they noticed a temporal probe in the video and reporting their confidence at the end of the video. (C) In Experiment 2, participants watched videos with or without simulation probes (briefly presented letters on stationary objects) that occurred either on objects that were on the simulated trajectory of the ongoing event (e.g., the orange ball that the red ball will hit; indicated with the dashed black arrows) or matched objects that were outside of the ongoing event. Participants were tasked with pressing either the "E" or "F" key when they saw the letter flash on the screen, and reporting their confidence at the end of the video. See the online article for the color version of this figure.

the start of each video so that participants could prepare for the video onset.

Physical event boundaries were defined by two expert coders (authors) who resolved disagreements by discussion. The criteria for determining physical event boundaries were subjective, but coders aimed to identify the onsets and offsets of every physical event shown in the scene. Each video was determined to have two or three event boundaries that separated different physical event types (e.g., the end of an occlusion event prior to a collision event).

These event boundaries were used to relate to temporal probe detection accuracy.

Apparatus

Because this experiment was run online, we had little control over the monitor setup (e.g., refresh rate, distance to the display). Nonetheless, we embedded procedures in our recruitment and experimental task to ensure that the experience was similar across participants.

First, we required that participants in our study participated on a desktop or laptop computer (and not a tablet or mobile device), using Prolific's inclusion criteria. Second, prior to starting the experiment, participants were told to adjust the size of an image onscreen to be the size of a credit card while keeping their arms comfortably outstretched. This allowed us to adjust our stimulus presentation size accordingly. Once participants completed this, they were told to maintain this arms-length distance from the monitor for the remaining duration of the experiment. Finally, after the experiment, all participants reported that the videos ran clearly and smoothly. In an open-ended response, two participants noted the difficulty of telling apart temporal probes from the refresh rate ("most of the distortions are pretty easy to spot apart from the few where it is hard to differentiate between framerate and distortion because the distortion is so minor" and "hard to tell sometimes if it was an intentional pause or a display glitch"). However, both of these participants showed response profiles that were in the hypothesized direction (i.e., better detection of nonboundary than boundary probes). Thus, although we had little control over task presentation by virtue of conducting an online study, participants did not seem to have difficulty viewing the stimuli or performing the probe detection task as intended.

Temporal Probe Task

For the temporal probe task, participants were asked to quickly and accurately respond to a slowdown of the video frame rate by pressing the spacebar. These slowdowns were created using FFmpeg (Tomar, 2006), such that for between 158 and 190 ms, the frame rate of the video was slowed by half (from 30 to 15 fps, for approximately four frames; see Figure S1 in the online supplement materials). Videos were shown centrally on the screen with a reminder at the top of the page: "Press the spacebar when a pause occurs." (Note that no real "pause" occurred in our videos—rather, on some trials, there was a temporary slowdown in the frame rate. Still, the participants seemed to understand the task, as evidenced by their performance, the low rate of exclusions, and debrief survey responses; see the online supplemental materials for a summary of our debrief survey.) The border around the video flashed red when it registered a spacebar response; otherwise, the border was black. After each video finished, participants rated their confidence on a sliding scale from *not very confident* to *very confident* (Figure 2B). Participants could then press the next button to start the next trial. Progress through the experiment was shown with a trial counter on the bottom of the screen.

Temporal probes could occur at any 200 ms time point from the start of a video, as long as there was motion on the screen (e.g., probes did not occur when a ball was occluded). Across the 32 base videos, there were many possible temporal probe locations ($M = 42$, range = 26–64), but we only sampled 22 of these probe locations for each video. To do this, we created 11 different sets of videos (each containing 96 videos), and collected data from 10 participants for each set. All participants completed a total of 96 trials, where 64 videos contained a temporal probe, and 32 did not. Some temporal probes (average of 4.85 per participant) occurred at event boundaries (e.g., as a ball was going to fall off of a ramp), while the majority of temporal probes (average of 52.84 per participant) occurred at timepoints within events (e.g., a ball further up the ramp; Figure 2B). The same base videos were shown 3 times throughout the experiment—twice with a temporal probe and once without it (as a filler). These filler trials were included to ensure that there were some trials in which a temporal probe was not shown at all, and that participants would be correct if they withheld their response. For our main analysis, we only analyze the trials in which a temporal probe was presented. Videos were shuffled and presented in a random order for each participant. Per our preregistration, individual trials were excluded if participants pressed the spacebar more than once during a video, or if participants responded outside of the acceptable response time window (within 300 and 2,500 ms of the probe).

At the start of the experiment, participants were asked to scale a bounding rectangle to the size of a credit card, to ensure that videos were approximately the same size across participants. They then were shown the experiment instructions and completed two practice trials. On the first practice trial, participants were shown a video of a ball colliding with another ball and then a third ball and told there would not be a temporal probe in this video. On the second practice trial, the same video was shown, but now there was a temporary slowdown of the video as the ball rolled towards the first ball it was going to collide with. After the practice, participants completed a two-question quiz on the instructions to test for comprehension. If either question was answered incorrectly, the instructions and practice trials were repeated, and participants only proceeded to the main experimental trials once they obtained 100% on the comprehension test. After completing the experimental trials, participants were asked a series of debriefing questions about task difficulty, engagement, and any feedback they had about the task. The code used to run this experiment was built using psiTurk, a platform for running behavioral experiments online (Eargle et al., 2020; Gureckis et al., 2016).

Event Segmentation Task

To supplement the expert-determined boundaries and to test whether naive participants' explicit boundary judgments relate to detection accuracy on the temporal probe experiment, we additionally collected event segmentation data from a separate set of Prolific participants. We had 20 new participants complete this explicit segmentation task, which does not include six participants who did not finish the task. Demographics of these participants were similar to the temporal probe task: Included participants ($N = 14$ female, $N = 6$ male) were between 21 and 37 years of age ($M_{age} = 29.20$ years) and were majority white ($N = 16$ White, $N = 3$ Asian, $N = 1$ other) from the United States or United Kingdom ($N = 1$ United States, $N = 19$ United Kingdom). For each of the 32 base videos, participants were first shown the video in its entirety without the ability to pause or rewind. Then, participants were able to scroll through the video and press a key to indicate that an event boundary occurred at a given time window. Participants were allowed to denote multiple event boundaries and were free to change their answers by pressing the same key again when they landed on the previously determined event boundary frame. Once participants were satisfied, they could proceed to the next trial. Example aggregated responses on several typical base videos can be seen in Figure S2 in the online supplemental materials. As can be appreciated, while there was some variability in where participants decided an event boundary occurred, the peaks in responses aligned well with expert-determined boundaries.

Transparency and Openness

Deidentified data for Experiment 1, as well as the videos that were shown, are publicly accessible at https://osf.io/ma32p/. The code used

to run the tasks is publicly accessible on Github (Experiment 1 temporal probe task: https://github.com/CNCLgithub/phys-env-psiturk/tree/temporal_probe; Experiment 1 event segmentation task: https://github.com/CNCLgithub/phys-env-psiturk/tree/event_seg). The code used to create the videos for both experiments is also publicly accessible on Github (https://github.com/CNCLgithub/physical_event_primitives) as is the code for the analyses (https://github.com/CNCLgithub/physical_events_analysis). There is a preregistration for Experiment 1 here https://aspredicted.org/blind.php?x=5ZF_D5T.

Results

Detection Accuracy: People Are Worse at Detecting Probes That Occur at Event Boundaries

For our main analysis, we asked whether detection accuracy for temporal probes placed at expert-determined boundaries (i.e., between collision, containment, falling, and occlusion events) was lower than temporal probes placed at nonboundary timepoints. We focus our analyses on participants' average probe detection accuracy, rather than accuracy adjusted for participants' false alarm rates, since false alarms were extremely low in this sample (average false alarm rate of 1.3%, with only 3% of participants false alarming to more than one nonprobe trial). In this and all following analyses, we only consider the trials in which participants responded once within the detection window (average 1.26% of trials per participant dropped for multiple key presses; average 8.64% of trials per participant dropped for out-of-range response times). Probe detection accuracy was averaged across boundary and nonboundary probe trials within each participant, and a two-sided paired t test was used to determine significance. As depicted in Figure 3A, participants' detection accuracy for temporal probes that occurred at event boundaries ($M = 33.20\%$) was lower than that of nonevent boundaries ($M = 61.06\%$)—$t(107) = -11.36$, $p < .001$—a strikingly strong effect (Cohen's $d = 1.10$), that was evident in 96 of the 108 participants we analyzed (Figure 3B). This was also true in our preregistered analysis collapsing across participants (instead of trial types within participant): using Fisher's exact test, we found that temporal probes that occurred at event boundaries (36.32% of 479 boundary probes detected) were less likely to be detected compared to nonevent boundaries (62.67% of 5,760 nonboundary probes detected; two-sided test, $p < .001$). Thus, participants readily experience the boundaries between physical events, as evidenced by their difficulty in recognizing changes in the speed of a video at event boundaries.

As an exploratory follow-up, we investigated probe detection accuracy in each event type separately (Figure 3C). Because different participants saw different subsets of physical events, varying numbers of participants contributed data to this analysis. Across all event types, detection accuracy was substantially decreased relative to nonboundary probe detection accuracy—collision: $M = 33.82\%$, $t(67) = -5.34$, $p < .001$; containment: $M = 29.26\%$, $t(93) = -6.95$, $p < .001$; falling: $M = 28.60\%$, $t(73) = -7.40$, $p < .001$; occlusion: $M = 38.16\%$, $t(74) = -5.13$, $p < .001$. This was also true when we performed our preregistered analysis of collapsing across participants—collision: 42.86%, $p < .001$; containment: 37.13%, $p < .001$; falling: 26.77%, $p < .001$; occlusion: 40.86%, $p < .001$. This suggests that the overall lower detection accuracy we see is not being driven by a particular event type, but may be a shared property of physical events.

In addition to investigating probe detection accuracy at expert-determined boundaries, we asked whether probe detection accuracy tracks naive participants' agreement about the placement of event boundaries. To test this, we averaged probe detection accuracy across participants for each possible temporal probe (704 unique videos)—on average, this value represented detection accuracy aggregated from 8.9 participants per temporal probe. Then, we calculated for each of the 32 base videos the proportion of event segmentation participants who indicated an event boundary had occurred within 200 ms time windows. Finally, we correlated average probe detection accuracy from Experiment 1 with the proportion of participants who indicated an event boundary within the window of time of the corresponding temporal probe. When considering all trials (including both the boundary and nonboundary trial types), we found a significant negative relationship between probe detection accuracy and event boundary agreement ($r = -.41$, $p < .001$), indicating that the more people agreed that an event boundary occurred, the less likely a temporal probe would be detected at that time point. This relationship persists when we only consider nonboundary probe trials ($r = -.38$, $p < .001$), and is marginally significant when we only consider boundary probe trials ($r = -.27$, $p = .052$). These results illustrate the robustness of our categorical analysis and further establish that event boundaries are experienced during spontaneous visual processing.

Our results thus far have not taken into consideration an important potential confound: that is, at event boundaries, we may expect to see more changes when physical dynamics of the video change, which would be reflected in the low-level pixel values of the video. This means that we may not be picking up on event boundary representations per se, but instead the impact of image changes on temporal probe detection accuracy. To address this, we first calculated the amount of pixel change that occurred during temporal probes separately for boundary and nonboundary probe trials. Specifically, we used OpenCV (Bradski, 2000) to average the difference in pixel values between adjacent frames for each frame during the temporal probe. First, we found that the amount of pixel change that occurred for temporal probes at event boundaries ($M = 0.06$) was not different from that of nonboundaries ($M = 0.05$), $t(703) = 1.36$, $p = .173$. Second, we investigated whether boundary status still had a significant impact on temporal probe detection accuracy when we include pixel change as a predictor. In a linear mixed effects model with participant as a random effect, we found that pixel change did not predict temporal probe detection accuracy ($\beta = 0.02$, $p = .902$), while boundary status did ($\beta = -0.27$, $p < .001$; Table 1). Therefore, our results are inconsistent with a purely low-level account of pixel change.

However, it is also possible that boundary status and pixel change interact to predict temporal probe detection accuracy. Thus, we ran a second linear mixed effects model with participant as a random effect, this time including an interaction term as an additional predictor. We found that there was in fact a significant interaction between boundary status and pixel change ($\beta = -1.24$, $p < .001$). To investigate the direction of this effect, we calculated the average temporal probe detection accuracy for each probe and then correlated this with the amount of pixel change, separately for boundary probe and nonboundary probe trials. We found that pixel change was not significantly related to temporal probe detection accuracy for nonboundary probes ($r = .06$, $p = .109$), but was significantly related to boundary probes ($r = -.29$, $p = .033$; Figure S3 in the online supplemental materials), such that boundary probes with higher amounts of pixel change

Figure 3
Temporal Probe Detection Accuracy From Experiment 1, Averaging the Boundary and Nonboundary Responses Within Participants
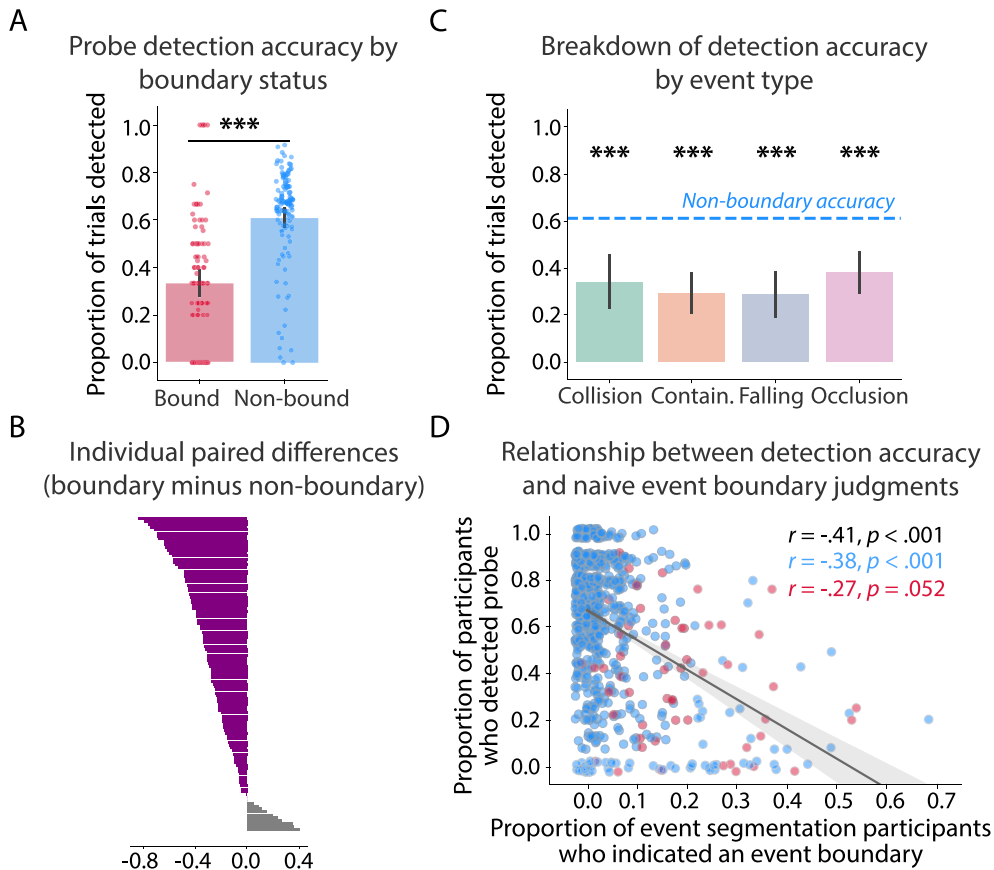
Note. (A) Temporal probes were significantly less likely to be detected when they occurred at expert-determined event boundaries compared to nonboundaries. Dots represent individual participant averages, and error bars reflect the 95% confidence interval derived from bootstrap resampling of the mean. (B) Differences at the individual participant level, sorted by the size of the effect. The expected difference (boundary accuracy less than nonboundary accuracy) was found in 96 out of the 108 participants included in this analysis. (C) Exploratory breakdown of temporal probe detection accuracy for the four different event types. Average participant accuracy on nonboundary trials is marked with a dashed line. Because the number of event type boundaries that were tested varied by participant group, subsets of participants are included in this analysis per event type (collision: $N = 68$; containment: $N = 94$; falling: $N = 74$; occlusion: $N = 75$). (D) The proportion of participants who detected a temporal probe was significantly negatively related to the proportion of event segmentation participants who indicated an event boundary in the 200-ms window around that time point when considering all possible probes, and when specifically looking at nonboundary probe trials (blue dots), but only marginally so when considering boundary probe trials (red dots). Dots represent averages across participants for different probe trials. The shaded area reflects the 95% confidence interval derived from bootstrap resampling of the correlation. Contain. = containment. See the online article for the color version of this figure.
*** $p < .001$.

showed lower average detection accuracy. Together, these results suggest that although pixel changes can influence temporal probe detection, boundary status remains an important factor influencing temporal probe detection accuracy.

The Time Course of Event Segmentation: Temporal Symmetry in Probe Detection Relative to Event Boundaries

So far, the results support our hypothesis that participants segment the continuous flow of sensory inputs according to physical events, as evidenced by their decreased ability to detect temporal probes at

expert-determined event boundaries as well as boundaries empirically determined by naive participants.

Critically, beyond the basic contrast of boundary versus nonboundary trials, our experimental design allows us to investigate the time course of how a new event is segmented in visual perception. We analyzed how probe detection accuracy changes as a function of time—i.e., when during the video the probe occurred relative to the moment of the nearest event boundary. To do so, we first collapsed the detection accuracy across trials, resulting in between 189 and 479 responses over the time interval of −1,000 to 1,000 ms centered at the event boundary at increments of 200 ms.

Table 1
Results From a Linear Mixed Effects Model

Temporal probe detection accuracy by pixel change + boundary condition				
Predictor	Beta estimate	*SE*	*t* statistic	*p*
Intercept	.613	0.020	30.695	<.001
Condition (boundary)	−.270	0.022	−12.491	<.001
Pixel change	.015	0.121	0.123	.902
Participant random effect	.036	0.012	—	—

Note. Probe detection (1 for *detected*, 0 for *not detected*) was predicted from the amount of pixel change at the probe and the boundary condition (1 for *boundary*, 0 for *nonboundary*), using data from all trials across participants (with participant as a random effect). The equation used in the model is shown as the table heading. The beta parameter, standard error, *t* statistic, and *p* values are given for each of the predictors.

Second, we calculated confidence intervals for detection accuracy at each temporal distance to the event boundary by resampling the corresponding trials 1,000 times with replacement. Finally, we characterized the rate of change in detection accuracy using a sigmoid function, instead of a linear fit, based on our qualitative observations that the way detection accuracy changed as a function of time was often nonlinear. We separately fit sigmoid functions preevent boundary (between −1,000 and 0 ms) and postevent boundary (between 0 and 1,000 ms) using SciPy curve fitting functions (Virtanen et al., 2020). Specifically, we conducted 1,000 bootstrap resamples of the average detection accuracy for each window of time, fitting a sigmoid function and saving the parameter values on each iteration. We then performed a two-tailed direct bootstrap hypothesis test on the resulting parameter estimates for the logistic growth/decay rate to assess whether the rates of change were symmetric or significantly different pre- and postboundary.

As shown in Figure 4A, there was a dropoff in average temporal probe detection accuracy when approaching the event boundary, followed by a return to baseline accuracy following the event boundary. Roughly centered at the event boundary, there is approximately one full second period (from −400 to 400 ms relative to the boundary) in which accuracy drops by a large amount and recovers that back. Indeed, this was confirmed in our analysis of the resampled sigmoid slope values: The rate of logistic decay preboundary—$\log(\beta) = 2.86$—was not significantly steeper than the rate of logistic growth postboundary—$\log(\beta) = 1.23$, $p = .168$ (Figure 4B). However, a bimodal distribution for preboundary rates of logistic decay was evident, with some values being in the same range of the postboundary rate of logistic growth, and some seemingly higher. In an exploratory analysis, we next investigated whether this bimodal distribution may be due to differences across different types of event boundaries. Indeed, we found that the rates of logistic growth and decay in detection accuracy somewhat varied by event type (Figure 4C). Specifically, consistent with the overall trend, preboundary rates of change were not significantly different from postboundary rates of change for containment—$\log(\beta) = 1.50$ versus $\log(\beta) = 0.77$, $p = .576$—or falling events—$\log(\beta) = 3.66$ versus $\log(\beta) = 3.32$, $p = .640$. However, for the remaining event types, it is evident that there is a quick dropoff in detection accuracy preboundary, followed by a more gradual pickup postboundary; for collision, preboundary rates of change were significantly higher/steeper than postboundary rates of change—$\log(\beta) = 3.63$ versus $\log(\beta) = 0.78$, $p = .048$—and the same was true for

occlusion—$\log(\beta) = 3.34$ versus $\log(\beta) = 0.53$, $p = .014$. These results indicate that while there is an overall decrease in temporal probe detection accuracy at event boundaries, how soon this detection accuracy decreases and then returns to baseline accuracy level seems to depend on the particular event types. These exploratory results indicate that there may be differences in how event models come online during perception.

Confidence and Response Time Also Reflect Event Boundaries

The focus of our research was on temporal probe detection accuracy, but we can also ask whether participants' confidence and response times were impacted by whether temporal probes were placed at boundary versus nonboundary timepoints. First, we investigated probe detection confidence averaged across boundary and nonboundary probe trials within each participant using a two-sided paired *t* test. As depicted in Figure S4A in the online supplemental materials, participants' confidence in their performance at detecting temporal probes at event boundaries ($M = 69.46$) was lower than their confidence for detecting temporal probes at nonevent boundaries ($M = 75.44$), $t(108) = −5.52$, $p < .001$. This is consistent with our main detection accuracy findings, which suggest that detecting temporal probes at event boundaries is more difficult. Importantly, although the amount of pixel change that occurred during a probe was related to participants' confidence, there was still a significant effect of event boundary status (Table S1 in the online supplemental materials). There was also a significant relationship between confidence and event boundary agreement across our event segmentation participants for all probe trials ($r = −.20$, $p < .001$) and when only considering nonboundary probe trials ($r = −.14$, $p < .001$), but not when only considering boundary probe trials ($r = −.13$, $p = .340$; Figure S4B).

Finally, we asked whether participants were slower to respond to temporal probes when they occurred at event boundaries. Similar to our confidence results, we found that event boundary status predicted response time such that even if people detected a probe, they were in fact slower to respond when that probe occurred at an event boundary. This result is based on a linear mixed effects model that took into account pixel change (Table S2 in the online supplemental materials). We find broadly consistent (though weaker) results in participant-level and continuous analyses of response time data (Figures S4C and S4D in the online supplemental materials).

Experiment 2

Experiment 1 demonstrated how and when visual input is demarcated into discrete events during perception. However, this is only important insofar as it helps guide attention and other downstream processing. In Experiment 2, we tested whether a form of "look ahead" simulations guide which objects are prioritized as part of a physical event representation. Furthermore, we ask whether object prioritization depends in part on when during an event it is probed—with the idea that attention towards objects relevant to the scene should be higher during event middles than event boundaries (Zacks, 2020). We showed participants videos consisting of multiple physical events in a sequence and asked them to indicate if and when during the video they noticed the flashing of a letter on one of the objects in the scene.

Figure 4

Rates of Change in the Average Probe Detection Accuracy Approaching and Receding From Expert-Determined Event Boundaries
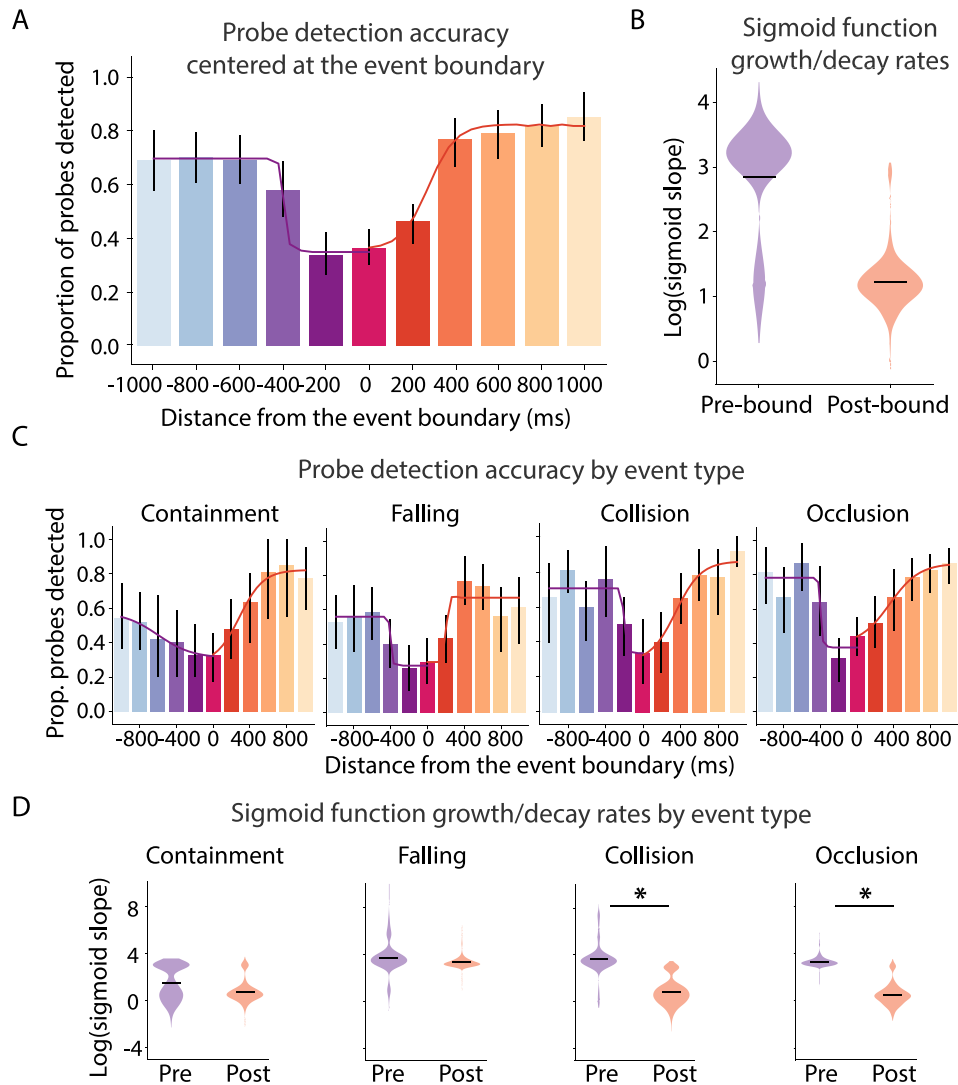
Note. (A) Detection accuracy rapidly declines prior to event boundaries, with the lowest detection accuracy occurring 200 ms before the boundary, and then increases following the event boundary. Sigmoid functions quantify such change in detection accuracy, and the best fits to the average data are displayed as lines pre- (purple) and postboundary (orange). (B) Sigmoid model slopes of the relationship between detection accuracy and temporal distance from the event boundary, calculated separately for preboundary and postboundary. Log-transformed slopes are shown based on 1,000 bootstrap resamples. The average sigmoid model slope was not significantly different pre- and postboundary. (C) Detection accuracy and distance plotted separately for each of the four event types. (D) The sigmoid model slopes after log transformation for 1,000 bootstrap resamples plotted separately for each of the four event types. Prop. = Proportion; Pre = preboundary; Post = postboundary. See the online article for the color version of this figure. * $p < .05$.

Method

Participants

We recruited undergraduate students from the psychology subject pool to complete an online version of this task. Data were collected in 2022. Our goal was to recruit 20 total participants in two groups of 10 participants each, with participants in the same group seeing the same combination of videos. However, due to the asynchronous nature of online recruitment, we ended up collecting 28 total participants (13 saw one set of videos, and 15 saw the other set). All participants were included in the study for having more than half of the trials usable (described more below). Demographic information of the participants is unavailable,

but psychology subject pool participants tend to be college-aged (18–22 years old) and balanced in terms of gender.

Stimuli

We created new multievent videos again using a combination of Python and Blender software, as in Experiment 1. This time, we used all five event types (collision, occlusion, containment, falling, and toppling) and strung them together into three-event videos. We also added numerous distracting objects to each scene in order to test for attentional allocation. We created a total of 32 "base" videos that were between 3 and 6 s long (mean video duration: 4.39 s).

From these base videos, we created probe videos using Adobe Premiere Pro software (Adobe Inc., 2019). Specifically, for each video, we first marked the transitions between physical events, and then determined whether in the time shortly before that event boundary there was (a) a target object present that would be on the simulated trajectory of the ongoing physical event (e.g., a ball that is about to be collided) and (b) a control object present that was of similar size, shape, color, and location as the target object but not involved in the physical event. Most videos had either one or two possible event boundaries that could be probed, with an average of 1.8 events. Then, for these events, we created two videos: one in which a gray letter flashed on the target object and one in which a gray letter flashed on the control object (Figure 2C). The letters were always placed in similar locations on the control and target objects, and lasted for 66 ms (two frames in a 30 fps video) before disappearing. Videos were randomly assigned to either having an "E" or "F" flash on the screen, but were made consistent across participants. On average, the offset of the letter flash occurred approximately 35.60 ms before event boundaries, although there was variability in when these probed occurred (with some shown at the boundary [as soon as 4 ms before the event boundary] and some shown in the middle of the event [as late as 2,000 ms before the event boundary]), which we later explore.

Apparatus

This experiment was run online and therefore we were unable to control the monitor setup (e.g., refresh rate, distance to the display). As in Experiment 1, we required all participants to participate on a desktop or laptop computer, adjust the size of an image on the screen to a credit card size to calibrate stimulus presentation size, and maintain an arms-length distance throughout the experiment. In the debrief survey, all participants reported that they did not experience issues in video quality (i.e., the videos ran smoothly and clearly). Although some participants commented on the difficulty of telling apart which letter flashed, none expressed difficulty in seeing that a letter was presented (e.g., "…I also found it very hard to see the letters, but could recognize flashes.").

Simulation Probe Task

For the simulation probe task, participants were asked to quickly and accurately respond to the 66-ms letter flash by pressing the key that corresponded to the flashing letter (either an "E" or an "F"). Right before each trial, participants saw a reminder at the top of the page: "Pay attention for an 'E' or 'F' flashing on the screen." In pilot testing, we determined that the size and duration of the simulation probe made it very difficult to detect. Thus, each video was shown in full-screen mode during the actual task. The border around the video flashed red when it registered a response; otherwise, the border was black. After each video

finished, participants rated their confidence on a sliding scale from *not very confident* to *very confident*. Participants could then press the next button to start the next trial. Progress through the experiment was shown with a trial counter on the bottom of the screen.

As mentioned above, simulation probes could occur either on target or control objects shortly before the boundary between physical events. We created two different sets of videos, where the same event boundary was only ever tested in either the target or control condition. In one set, 32 targets were probed while 26 control objects were probed; in the other set, the opposite was true. In half of these trials, the letter "E" appeared and in the other half, the letter "F" appeared. All participants also saw six videos in which no probe occurred, which were randomly chosen for each participant; however, we only considered trials in which a probe was presented for our analyses. This meant in total, participants completed 64 trials, which were presented in a randomized order. Similar to Experiment 1, individual trials were excluded if participants pressed either the "E" or "F" key more than once during a video, if they responded with the wrong key press, or if they responded outside of the acceptable response time window (within 300 and 2,500 ms of the probe).

At the start of the experiment, participants were shown the experiment instructions and completed two practice trials. On the first practice trial, participants were shown a video of a ball rolling out of a tube then hitting a block that topples and collides with another ball. They were told there would not be a simulation probe in this video. On the second practice trial, the same video was shown, but now the letter "E" flashed on the block that would topple. (Participants were shown a screenshot of the letter on the block prior to the practice trial to give them an idea of what to look for; this was never shown for the actual test trials.) After the practice, participants completed a two-question quiz on the instructions to test for comprehension. If either question was answered incorrectly, the instructions and practice trials were repeated, and participants only proceeded to the main experimental trials once they obtained 100% on the comprehension test. After completing the experimental trials, participants were asked a series of debriefing questions about task difficulty, engagement, and any feedback they had about the task. The code used to run this experiment was built using psiTurk (Eargle et al., 2020; Gureckis et al., 2016).

Transparency and Openness

Deidentified data for Experiment 2, as well as the videos that were shown, are publicly accessible at https://osf.io/ma32p/. The code used to run the tasks is publicly accessible on Github (Experiment 2 spatial probe task: https://github.com/CNCLgithub/phys-env-psiturk/tree/spatial_probe). The code used to create the videos for both experiments is also publicly accessible on Github (https://github.com/CNCLgithub/physical_event_primitives) as is the code for the analyses (https://github.com/CNCLgithub/physical_events_analysis). Experiment 2 was not preregistered.

Results

Detection Accuracy: People Are More Accurate at Detecting Probes Placed on Objects That Lie Along the Simulated Trajectory of the Ongoing Physical Event

For our main analysis, we asked whether detection accuracy for simulation probes placed on target objects (i.e., objects that are

stationary at the moment when the simulation probe flashes but otherwise lie on the simulated trajectory of the current event) would be higher than detection accuracy for simulation probes placed on control objects (i.e., objects that are also stationary at the moment when the simulation probe flashes but lie outside of the simulated trajectory of the current event). The rationale for this hypothesis was that attention would be allocated via a form of "look ahead" simulations toward those objects that lie within the unfolding local causal context of the current physical event (though again, critically, are not actively in motion). To test this, we averaged probe detection accuracy across target and control probe trials within each participant, and used a two-sided paired t test to determine significance. We again focus on detection accuracy, rather than accuracy adjusted for false alarm rates, given the low rate of false alarms in this task (average false alarm rate of 0.5%, with only one participant false alarming to more than one nonprobe trial). In this and all following analyses, we only consider the trials in which participants responded once within the detection window (average 1.00% of trials per participant dropped for multiple key presses; average 4.41% of trials per participant dropped for out-of-range response times). Importantly, we also confirmed that when participants responded on a given trial, they were significantly above chance in correctly determining which letter appeared—$M = 77.84\%$, $t(26) = 10.70$, $p < .001$, Cohen's $d = 2.10$; one participant excluded for not responding on any trial. This was true both for simulation probes that occurred on target objects—$M = 81.46\%$, $t(26) = 13.75$, $p < .001$, Cohen's $d = 2.70$—and for simulation probes that occurred on control objects—$M = 73.69\%$, $t(26) = 5.94$, $p < .001$, Cohen's $d = 1.16$—with only a marginal advantage for determining which letter appeared on target objects—$t(26) = 2.00$, $p = .055$, Cohen's $d = 0.39$. For all subsequent analyses, we only considered trials in which participants responded with the correct letter probe (average 11.38% of trials per participant dropped for incorrect key press [e.g., "F" when the probe was "E"]).

As depicted in Figure 5A, participants' detection accuracy for simulation probes that occurred on target objects ($M = 49.31\%$) was significantly higher than that of control objects ($M = 41.71\%$), $t(27) = 3.54$, $p = .001$, Cohen's $d = 0.68$. This effect was not driven by just a few subjects; instead it was observed in 23 out of 28 participants (Figure 5B). These results suggest that participants are indeed selectively processing objects that are along the simulated trajectory of the current physical event.

Could these results simply reflect the spatial spread of attention? Indeed, when we looked at our stimuli, we found that control objects tended to be further away from the main agent object at the moment when the simulation probe occurred (Figure S5A in the online supplemental materials). Additionally, in the debrief survey, several participants mentioned focusing on the moving object as a strategy. To address this potential confound, we ran an analysis on a subset of trials where we matched the pixel-based Euclidean distance between the target object and the agent object (e.g., the red ball that will soon collide with the target object, but not the control object) and the control object and the agent object. Specifically, we only looked at trials where the distance between the center of the agent and the center of the target/control object was between 225 and 275 pixels, reducing the number of target and control object trials per participant from approximately 50 to between 10 and 20 trials. When matching for spatial distance in this way, we found a similar (in fact, numerically larger) effect as in our main result: participants' detection accuracy for simulation probes that occurred on target objects ($M =$

Figure 5

Simulation Probe Detection Accuracy in Experiment 2, Averaging the Target and Control Object Trials Within Participants
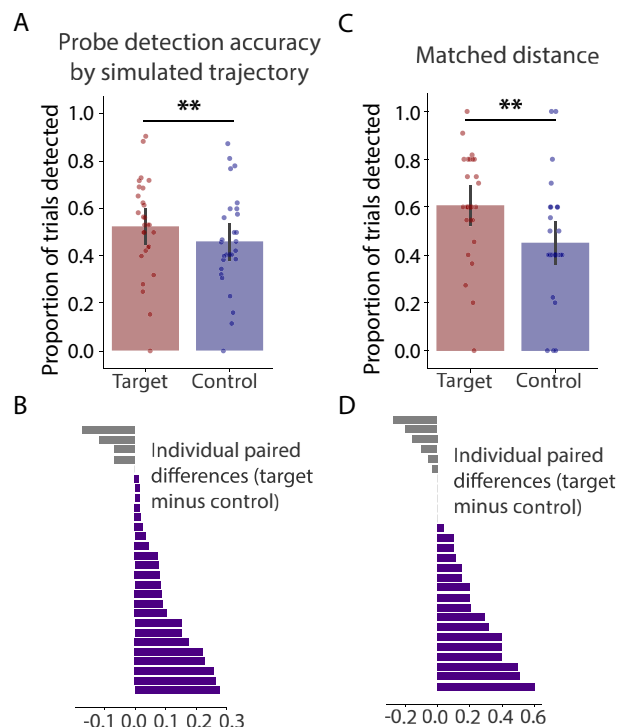
Note. (A) Simulation probes were significantly more likely to be detected when they appeared on target objects (i.e., objects within the current physical event) compared to control objects. Dots represent individual participant averages, and error bars reflect the 95% confidence interval derived from bootstrap resampling of the mean. (B) Differences at the individual participant level, sorted by the size of the effect. The expected difference (target accuracy greater than control accuracy) was found in 23 out of the 28 participants tested. (C) When matching for the Euclidean distance between the probed objects and the agent object, simulation probes were still significantly more likely to be detected if they appeared on target compared to control objects. (D) Differences at the individual participant level after matching for difference, with the expected difference found in 17 out of the 28 participants tested. See the online article for the color version of this figure.
** $p < .01$.

55.11%) was significantly higher than that of control objects ($M = 41.29\%$), $t(27) = 3.26$, $p = .003$, Cohen's $d = 0.63$ (Figure 5C). This effect was again evident in the majority of participants (17 out of 28; Figure 5D), and robust across the range of spatial distance values that were tested (Figure S5B in the online supplemental materials).

We also investigated whether the particular letters we used as probes mattered for our effects. To our surprise, we observed that detection accuracy as well as the magnitude of the effect were impacted by the identity of the letter shown (see Figure S6A and S6B in the online supplemental materials). Even though our main aggregate results comparing target and control objects are not impacted by this difference between the two letter probes, this result indicates that future studies should consider the detectability of

letters when designing their experiments (Mueller & Weidemann, 2012).

Confidence and Response Time Also Reflect Simulated Trajectory of Objects

We also tested whether participants' confidence and response times reflect any signature of simulated trajectories within an event. First, we investigated confidence in probe detection accuracy averaged across target and control probe trials within each participant and used a two-sided paired *t* test to determine significance. As depicted in Figure S7A in the online supplemental materials, participants' confidence in their performance at detecting simulation probes appearing on target objects (*M* = 61.86) was higher than their confidence in detecting simulation probes appearing on control objects (*M* = 57.47), *t*(27) = 3.28, *p* = .003. Second, to analyze response time data, we considered participants who responded to at least one target and control trial, resulting in only 26 participants. Trending in the direction we expect, we found participants' response times were numerically smaller on target objects (*M* = 1,603.07 ms) relative to controls (*M* = 1,635.37 ms), *t*(25) = −1.83, *p* = .078, but this relationship did not reach significance (Figure S7B in the online supplemental materials). These results are consistent with our main detection accuracy findings, suggesting that objects along the simulated trajectory of the current event are prioritized.

Connecting Experiments 1 and 2: Detection Accuracy at Event Boundaries Versus Event Middles

Thus far, the results from Experiment 2 have shown that physical events enable selective processing of objects that are on the future trajectory of the current physical event, relative to control objects. At a finer level of granularity, the ability to detect a simulation probe may depend on when during the time course of an event it occurs, much like the time course of the detectability of temporal probes in Experiment 1. Because simulation should be tied to the current physical event, we may expect the detection accuracy of simulation probes to be higher during an event (i.e., event middle) compared to at an event boundary.

In Experiment 2, simulation probes were typically placed just before physical event boundaries, but sometimes, they appeared much earlier in time—that is, during the middle of an event. This variability in the temporal placement of simulation probes gave us the opportunity to explore whether detection accuracy was modulated by temporal proximity to the event boundary. Experiment 1 demonstrated that temporal probe detection accuracy is at baseline levels around 600 ms before the event boundary, but substantially drops within 200 ms of the event boundary. Therefore, we labeled probes that occurred 500 ms or more before the event boundary "event middle simulation probes" and probes that occurred closer in time to the boundary "event boundary simulation probes." (Note that too few trials would be included if we only looked at probes occurring within 200 ms of the boundary.) We first examined trials in which the simulation probe appeared on a target object ("target trials"), averaging probe detection accuracy for event middle simulation probes versus event boundary simulation probes within each participant and using a two-sided paired *t* test to determine significance. Participants' detection accuracy for target simulation probes that occurred at event boundaries (*M* = 47.17%) was

Figure 6

Simulation Probe Detection Accuracy in Experiment 2 Based on When During the Physical Event the Probe Appeared
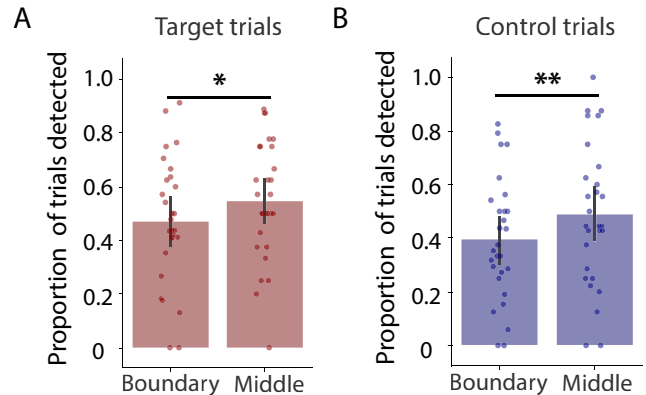
Note. (A) Target simulation probes were significantly more likely to be detected when they appeared at event middles (more than 500 ms from an event boundary) compared to near event boundaries. Dots represent individual participant averages, and error bars reflect the 95% confidence interval derived from bootstrap resampling of the mean. (B) Control probes were also significantly more likely to be detected when they appeared at event middles compared to event boundaries. See the online article for the color version of this figure.
* *p* < .05. ** *p* < .01.

significantly lower than for probes that occurred at event middles (*M* = 54.79%), *t*(27) = −2.70, *p* = .012, Cohen's *d* = −0.52 (Figure 6A). These results show that detection accuracy decreases for target simulation probes at event boundaries, relative to event middles, and are broadly consistent with the event cognition literature (Zacks et al., 2007). We emphasize that our findings are based on measuring spontaneously unfolding processes in visual perception, and underscore the importance of the temporal structure imposed by physical events on perception.

However, if the temporal dependence of the ability to detect simulation probes is exclusively due to event-driven prediction (e.g., as assumed in Eisenberg et al., 2018), we may expect to not see a temporal dependence for objects that are irrelevant to the ongoing event. To test this, we examined participants' ability to detect probes that occurred on control objects ("control trials"). We found that participants' detection accuracy for control probes that occurred at event boundaries (*M* = 39.39%) was also significantly lower than for probes that occurred at event middles (*M* = 48.63%), *t*(27) = −2.98, *p* = .006, Cohen's *d* = −0.57 (Figure 6B). Thus, probe detection accuracy is enhanced at event middles not just for those objects that are relevant to the current physical event, but also for control objects that are in the vicinity of the target object. These results suggest that the decreased ability to detect simulation probes at event boundaries may reflect a more global effect of attention differences, instead of, or in addition to, event-driven prediction.

General Discussion

In this article, we conducted two performance-based tasks to gain insight into the spontaneous formation and contents of physical event representations in visual perception. Our first experiment

suggested that physical events are spontaneously segmented in vision, and allowed us to reveal the temporal profile of how such segmentation occurs during ongoing perception, at subsecond resolution and at the level of individual event types. Our second experiment queried the contents of physical events based on the principle of prediction. We found that objects along the simulated trajectory of the current physical event have an attentional advantage over similar objects outside of such predicted trajectory. Together, these results provide an empirical basis for understanding physical events as units of visual perception during ongoing experience, starting to bridge the knowledge gap relative to our understanding of visual object representations and highlighting new avenues for exploration in intuitive physics.

We note that our study is not unique in the study of physical events in intuitive physics; several studies have used physical events, but typically in the form of presegmented singleton events (e.g., animations including a single event or interaction) which act as stimuli to study some other cognitive or perceptual phenomena (Gerstenberg et al., 2017, 2021; Schwettmann et al., 2019). As far as we can tell, physical events are not the "object" of study in this work, that is, they do not address the structure of physical events, including how they are segmented to begin with and their contents. Thus, our work makes an important advance in understanding how physical events impose temporal structure and influence the contents of intuitive physics, providing the scaffolding for these effects to be observed.

The task in Experiment 1 was inspired by work in audition and music cognition (Repp, 1992, 1998), which found that participants are less likely to detect musical phrase lengthenings when they are placed at boundaries. Transferring this method to the domain of vision allowed us to determine when event boundaries are perceived during a series of physical interactions, without informing participants to attend to physical event structure. Our results mirror that of Ji and Papafragou (2022), who found that participants are less likely and slower to detect visual disruptions at event boundaries compared to event middles (particularly for so-called "bounded" events, relative to "unbounded" events, an important distinction made in the linguistic theories of events). In the current study, we tested not just event boundaries and middles, but also other time-points in between, allowing us to quantify event segmentation in a temporally precise manner. This approach provides a data-driven "behavioral signature" of event boundary status that should be broadly useful across psychological and cognitive neuroscience studies of event perception. A key question left unanswered in the current study is why participants are worse at detecting temporal probes at physical event boundaries. One possibility is that these event boundaries simply correspond to moments of perceptual change (Hard et al., 2011; Newtson et al., 1977). However, our analysis controlling for changes in pixels indicates that a purely low-level explanation is unlikely to account for this result. Rather, the lower temporal probe detection accuracy may reflect instantiation of a new event model, which likely consumes additional cognitive resources and might therefore impact task accuracy. Indeed, this fits with our results showing that the impact of an event boundaries on probe detection accuracy is not just at the moment of the boundary, but also during the lead up and follow up of the events. A related possibility is that event transitions "feel" subjectively longer (therefore making the temporal probes not feel out of place), as suggested in time perception research: Time is judged to have lasted longer when multiple events occur (Faber & Gennari, 2015), and

rhythms are reproduced as longer at event boundaries (Ongchoco et al., 2023). If the perceived duration of event transitions is related to reduced performance on our temporal probe detection task, then we might expect participants to false alarm more (or "hallucinate" a temporary slowdown) at event boundaries. Alternatively, participants could have a response bias against probes near event boundaries, in which case we would expect to find more false alarms at event middles than event boundaries. In the current study, there were too few false alarms for us to consider these two hypotheses. Future work should therefore more precisely determine the cause(s) of reduced probe detection accuracy leading up to and following event boundaries.

Infant cognition research has utilized physical event types to study how physical reasoning develops early in life. One surprising and consistent finding is that infants treat containment and occlusion differently (Baillargeon, 2004; Baillargeon et al., 2010; Hespos & Baillargeon, 2001, 2006; Lin et al., 2021). Paralleling this result, in an exploratory analysis we found that detection accuracy showed temporally symmetrical decay and growth relative to containment events, but this pattern was temporally asymmetric for occlusion events. The mechanism for why this may be the case, and why falling but not collision events also show temporal symmetry, remains a topic for future consideration. One speculation is that the two events that showed rapid temporal decay and then slower growth (occlusion and collision events) are those events that can be distinguished by infants relatively early in development, between 4 and 6 months of age (Hespos & Baillargeon, 2001, 2006; Saxe & Carey, 2006). Alternatively, it may be that these results do not reflect developmental origins, but rather the dynamics that are afforded by a particular event type. Similar to recent work exploring intuitions about magic tricks (Lewry et al., 2021), future research should systematically study parallels between intuitive physics within the adult visual system and its developmental trajectory to better understand how physical event representations are represented in the mind.

Experiment 2 adapted and tested a key idea from the event cognition literature (Yates et al., 2023; Zacks et al., 2011) in visual perception of physical events: Events are defined by their ability to predict what is going to happen next. For physical events, prediction entails simulating the trajectory of a moving object and anticipating which new objects will interact with the moving object (a form of coarse simulations, or "looking ahead"). Participants were better at detecting probes that occurred on objects along the simulated trajectory of the current event. We note that because the vertical position and trajectory of objects changed on a video-by-video basis (i.e., sometimes a ball moved from left to right, sometimes from right to left, and sometimes from up to down), these results are not likely due to a simple heuristic of predicting along a single path. Physical events and simulation within them may therefore be a general principle that determines the contents of physical events, including the relevant features of objects (Strickland & Scholl, 2015), which objects themselves are relevant, and perhaps even inducing invisible ones into perception (Little & Firestone, 2021).

Although we describe this effect as the result of "look ahead" simulations, we do not have eye-tracking data to say whether or not participants made predictive fixations towards to the target object during the task. Nonetheless, because target and control objects were probed equally often in the experiment, any predictive looking due to task or goals should have been equally likely for either object (Hayhoe & Ballard, 2005). An explanation that our results are due to

distance to the moving object does not hold because our effects were still present (and numerically larger) when restricting our analyses to target/control objects that were at a similar distance to the moving object. Thus, we believe that these results reflect event-driven effects on object-based attention (Chen, 2012). To further establish this conclusion, future work should explore whether these effects continue to hold under a broader degree of variation in the observed motion, due to not just the scene dynamics but also scene-extrinsic properties such as the viewpoint (e.g., oblique angle, "birds-eye" view, toward or away from camera) and perhaps variation in lighting. We also replicated the finding that probes placed at event boundaries are more difficult to detect than probes placed at event middles (Huff et al., 2012; Pradhan & Kumar, 2022). In prior work, probes were placed either on targets that participants were explicitly told to pay attention to (Huff et al., 2012), or in the center or four corners of the stimulus display (Pradhan & Kumar, 2022). Thus, our study is unique is showing that probes are better detected when placed on target and control objects based on the physical events themselves, rather than on explicit instructions from experimenters. This is despite the fact that probes placed at event boundaries were in closer proximity to the object that was moving in the scene; if we instead found that probes placed at event boundaries were better detected, this could be explained by a simple spread of attention from the moving object to the objects nearby. Interestingly, this effect was not specific to the target objects. Thus, the decreased ability to detect probes at event boundaries appears to be a more global impact of attention, rather than something that is specific to the ongoing event. Together, these results underscore the importance of having fine-grained controls not only in terms of temporal aspects of a scene (e.g., as in Eisenberg et al., 2018; Ji & Papafragou, 2022), but also the object-level contents of what might be influenced by events.

There are several limitations to note in this study. First, there was a difference in the number of boundary versus nonboundary trials that were tested within each participant in Experiment 1. This was part of our initial design decision, given that we wanted to sample as many different timepoints in each video as possible, but likely impacted power (e.g., the analysis in Figure 3D focusing on boundary trials). Second, despite interesting event-specific effects in Experiment 1, we did not have the power to break Experiment 2 data down into individual event types because of the smaller participant sample size and number of videos tested. Third, in Experiment 2, we only investigated attention allocation at time points near event boundaries or about 1 s away with a single control object. Further mapping out how attention is directed, and when it is heightened, during physical events (as in Experiment 1) will help uncover more of the internal structure of physical events. Moreover, there are many other aspects of the probe detection paradigm that could be further examined in future studies, such as how long probes appear on the screen and the type of the response required. For example, Pradhan and Kumar (2022) only found a difference in probe detection accuracy at event boundaries when participants were instructed to perform a more difficult go/no-go task. Thus, we might expect that decreasing the difficulty of the probe detection task, such as by increasing the time over which a probe is on-screen, would lead to a recovery in participant's probe detection accuracy at event boundaries. Finally, both of our experiments were tested online, and therefore we were unable to precisely control stimulus presentation timing, participant viewing distance, and other factors that may have influenced participants' ability to detect probes. Nonetheless, the fact that the majority of participants in both experiments showed higher detection rates than false alarm rates and consistent results across participants can be seen as evidence for the robustness of our effects.

Our work underscores the existence and structure of physical events in visual perception, which should inform computational accounts of intuitive physics (Bear et al., 2021; Piloto et al., 2022). An influential framework suggests that the mind builds runnable mental models via an "intuitive physics engine" similar to probabilistic or approximate forms of physics engines from computer graphics (Battaglia et al., 2013). Recent work observed the implausibility of simulating the trajectories of all objects in a scene all the way to the end of their motions (Bass et al., 2021; Ludwin-Peery et al., 2021), but their solutions either considered a more heuristic-driven (and less simulation-like) basis for intuitive physics or suggested ad hoc remedies. Our findings motivate an alternative for visual perception: minimally expressed, selective simulations that are sufficient to explain and predict the trajectories of objects within their local causal contexts. Future computational work should implement such "event-driven intuitive physics engines" to explore the mechanistic basis of physical events in perception.

References

Adobe Inc. (2019). *Adobe Premiere Pro (CC 2019)*. https://www.adobe.com/products/premierepro.html

Baillargeon, R. (2004). Infants' reasoning about hidden objects: Evidence for event-general and event-specific expectations. *Developmental Science*, 7(4), 391–414. https://doi.org/10.1111/j.1467-7687.2004.00357.x

Baillargeon, R., Li, J., Gertner, Y., & Wu, D. (2010). How do infants reason about physical events? In U. Goswami (Ed.), *The Wiley-Blackwell handbook of childhood cognitive development* (pp. 11–48). Blackwell Publishing.

Baker, L. J., & Levin, D. T. (2015). Event perception as a control process for visual awareness. *Visual Cognition*, 23(7), 814–816. https://doi.org/10.1080/13506285.2015.1093246

Bass, I., Smith, K. A., Bonawitz, E., & Ullman, T. D. (2021). Partial mental simulation explains fallacies in physical reasoning. *Cognitive Neuropsychology*, 38(7–8), 413–424. https://doi.org/10.1080/02643294.2022.2083950

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 110(45), 18327–18332. https://doi.org/10.1073/pnas.1306572110

Bear, D. M., Wang, E., Mrowca, D., Binder, F. J., Tung, H. Y. F., Pramod, R. T., Holdaway, C., Tao, S., Smith, K., Sun, F.-Y., Fei-Fei, L., Kanwisher, N., Tenenbaum, J. B., Yamins, D. L. K., & Fan, J. E. (2021). *Physion: Evaluating physical prediction from vision in humans and machines*. arXiv. https://arxiv.org/abs/2106.08261

Bradski, G. (2000). The OpenCV library. *Dr. Dobb's Journal of Software Tools*, 25(11), 122–125. https://elibrary.ru/item.asp?id=4934581

Chen, Z. (2012). Object-based attention: A tutorial review. *Attention, Perception, & Psychophysics*, 74(5), 784–802. https://doi.org/10.3758/s13414-012-0322-z

Community, B. O. (2018). *Blender—A 3D modelling and rendering package*. Blender Foundation. https://www.blender.org

Eargle, D., Gureckis, T., Rich, A. S., McDonnell, J., & Martin, J. B. (2020). *Psiturk: An open platform for science on Amazon Mechanical Turk*. Zenodo. https://doi.org/10.5281/zenodo.3598652

Eisenberg, M. L., Zacks, J. M., & Flores, S. (2018). Dynamic prediction during perception of everyday events. *Cognitive Research: Principles and Implications*, 3(1), Article 53. https://doi.org/10.1186/s41235-018-0146-z

Faber, M., & Gennari, S. P. (2015). In search of lost time: Reconstructing the unfolding of events from memory. *Cognition*, 143, 193–202. https://doi.org/10.1016/j.cognition.2015.06.014

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2021). A counterfactual simulation model of causal judgments for physical events. *Psychological Review*, *128*(5), 936–975. https://doi.org/10.1037/rev0000281

Gerstenberg, T., Peterson, M. F., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2017). Eye-tracking causality. *Psychological Science*, *28*(12), 1731–1744. https://doi.org/10.1177/0956797617713053

Gibson, J. J. (1954). The visual perception of objective motion and subjective movement. *Psychological Review*, *61*(5), 304–314. https://doi.org/10.1037/h0061885

Gibson, J. J. (1975). Events are perceivable but time is not. In J. T. Fraser & N. Lawrence (Eds.), *The study of time II* (pp. 295–301). Springer.

Guo, L., Courtney, S. M., & Fischer, J. (2020). Knowledge of objects' physical properties implicitly guides attention during visual search. *Journal of Experimental Psychology: General*, *149*(12), 2332–2343. https://doi.org/10.1037/xge0000776

Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., Halpern, D., Hamrick, J. B., & Chan, P. (2016). Psiturk: An open-source framework for conducting replicable behavioral experiments online. *Behavior Research Methods*, *48*(3), 829–842. https://doi.org/10.3758/s13428-015-0642-8

Hamrick, J. B., Battaglia, P. W., Griffiths, T. L., & Tenenbaum, J. B. (2016). Inferring mass in complex scenes by mental simulation. *Cognition*, *157*, 61–76. https://doi.org/10.1016/j.cognition.2016.08.012

Hard, B. M., Recchia, G., & Tversky, B. (2011). The shape of action. *Journal of Experimental Psychology: General*, *140*(4), 586–604. https://doi.org/10.1037/a0024310

Hayhoe, M., & Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences*, *9*(4), 188–194. https://doi.org/10.1016/j.tics.2005.02.009

Hespos, S. J., & Baillargeon, R. (2001). Infants' knowledge about occlusion and containment events: A surprising discrepancy. *Psychological Science*, *12*(2), 141–147. https://doi.org/10.1111/1467-9280.00324

Hespos, S. J., & Baillargeon, R. (2006). Décalage in infants' knowledge about occlusion and containment events: Converging evidence from action tasks. *Cognition*, *99*(2), B31–B41. https://doi.org/10.1016/j.cognition.2005.01.010

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

Ji, Y., & Papafragou, A. (2022). Boundedness in event cognition: Viewers spontaneously represent the temporal texture of events. *Journal of Memory and Language*, *127*, Article 104353. https://doi.org/10.1016/j.jml.2022.104353

Karakose-Akbiyik, S., Caramazza, A., & Wurm, M. F. (2023). A shared neural code for the physics of actions and object events. *Nature Communications*, *14*(1), Article 3316. https://doi.org/10.1038/s41467-023-39062-8

Lewry, C., Curtis, K., Vasilyeva, N., Xu, F., & Griffiths, T. L. (2021). Intuitions about magic track the development of intuitive physics. *Cognition*, *214*, Article 104762. https://doi.org/10.1016/j.cognition.2021.104762

Lin, Y., Li, J., Gertner, Y., Ng, W., Fisher, C. L., & Baillargeon, R. (2021). How do the object-file and physical-reasoning systems interact? Evidence from priming effects with object arrays or novel labels. *Cognitive Psychology*, *125*, Article 101368. https://doi.org/10.1016/j.cogpsych.2020.101368

Little, P. C., & Firestone, C. (2021). Physically implied surfaces. *Psychological Science*, *32*(5), 799–808. https://doi.org/10.1177/0956797620939942

Ludwin-Peery, E., Bramley, N. R., Davis, E., & Gureckis, T. M. (2021). Limits on simulation approaches in intuitive physics. *Cognitive Psychology*, *127*, Article 101396. https://doi.org/10.1016/j.cogpsych.2021.101396

McCloskey, M. (1983). Intuitive physics. *Scientific American*, *248*(4), 122–130. https://doi.org/10.1038/scientificamerican0483-122

Michotte, A. E. (1963). *The perception of causality* (T. R. Miles, Trans.). Methuen.

Mueller, S. T., & Weidemann, C. T. (2012). Alphabetic letter identification: Effects of perceivability, similarity, and bias. *Acta Psychologica*, *139*(1), 19–37. https://doi.org/10.1016/j.actpsy.2011.09.014

Newtson, D., & Engquist, G. (1976). The perceptual organization of ongoing behavior. *Journal of Experimental Social Psychology*, *12*(5), 436–450. https://doi.org/10.1016/0022-1031(76)90076-7

Newtson, D., Engquist, G. A., & Bois, J. (1977). The objective basis of behavior units. *Journal of Personality and Social Psychology*, *35*(12), 847–862. https://doi.org/10.1037/0022-3514.35.12.847

Ongchoco, J. D. K., Yates, T. S., & Scholl, B. (2023). Event segmentation structures temporal experience: Simultaneous dilation and contraction in rhythmic reproductions. *Journal of Experimental Psychology: General*, *152*(11), 3266–3276. https://doi.org/10.1037/xge0001447

Papafragou, A., & Ji, Y. (2023). Events and objects are similar cognitive entities. *Cognitive Psychology*, *143*, Article 101573. https://doi.org/10.1016/j.cogpsych.2023.101573

Piloto, L. S., Weinstein, A., Battaglia, P., & Botvinick, M. (2022). Intuitive physics learning in a deep-learning model inspired by developmental psychology. *Nature Human Behaviour*, *6*(9), 1257–1267. https://doi.org/10.1038/s41562-022-01394-8

Pradhan, R., & Kumar, D. (2022). Event segmentation and event boundary advantage: Role of attention and postencoding processing. *Journal of Experimental Psychology: General*, *151*(7), 1542–1555. https://doi.org/10.1037/xge0001155

Pramod, R., Cohen, M. A., Tenenbaum, J. B., & Kanwisher, N. (2022). Invariant representation of physical stability in the human brain. *eLife*, *11*, Article e71736. https://doi.org/10.7554/eLife.71736

Radvansky, G. A., & Zacks, J. M. (2017). Event boundaries in memory and cognition. *Current Opinion in Behavioral Sciences*, *17*, 133–140. https://doi.org/10.1016/j.cobeha.2017.08.006

Repp, B. H. (1992). Probing the cognitive representation of musical time: Structural constraints on the perception of timing perturbations. *Cognition*, *44*(3), 241–281. https://doi.org/10.1016/0010-0277(92)90003-Z

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), 791–811. https://doi.org/10.1037/0096-1523.24.3.791

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. https://doi.org/10.1080/15326900701399913

Roussel, R., Cani, M.-P., Léon, J.-C., & Mitra, N. J. (2019). Designing chain reaction contraptions from causal graphs. *ACM Transactions on Graphics*, *38*(4), 1–14. https://doi.org/10.1145/3306346.3322977

Saxe, R., & Carey, S. (2006). The perception of causality in infancy. *Acta Psychologica*, *123*(1–2), 144–165. https://doi.org/10.1016/j.actpsy.2006.05.005

Schwettmann, S., Tenenbaum, J. B., & Kanwisher, N. (2019). Invariant representations of mass in the human brain. *eLife*, *8*, Article e46619. https://doi.org/10.7554/eLife.46619

Spelke, E. S. (1990). Principles of object perception. *Cognitive Science*, *14*(1), 29–56. https://doi.org/10.1207/s15516709cog1401_3

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental Science*, *10*(1), 89–96. https://doi.org/10.1111/j.1467-7687.2007.00569.x

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), 570–580. https://doi.org/10.1037/a0037750

Tomar, S. (2006). Converting video formats with FFmpeg. *Linux Journal*, *2006*(146), Article 10. https://dl.acm.org/doi/abs/10.5555/1134782.1134792

Van Rossum, G., & Drake, F. L. (2009). *Python 3 reference manual*. CreateSpace.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … van Mulbregt, P. (2020). Scipy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Wong, K. W., Bi, W., Soltani, A. A., Yildirim, I., & Scholl, B. J. (2023). Seeing soft materials draped over objects: A case study of intuitive physics in perception, attention, and memory. *Psychological Science*, *34*(1), 111–119. https://doi.org/10.1177/09567976221109194

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. https://doi.org/10.3758/s13423-023-02311-4

Yousif, S. R., & Scholl, B. J. (2019). The one-is-more illusion: Sets of discrete objects appear less extended than equivalent continuous entities in both space and time. *Cognition*, *185*, 121–130. https://doi.org/10.1016/j.cognition.2018.10.002

Zacks, J. M. (2020). Event perception and memory. *Annual Review of Psychology*, *71*(1), 165–191. https://doi.org/10.1146/annurev-psych-010419-051101

Zacks, J. M., Kurby, C. 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. https://doi.org/10.1162/jocn_a_00078

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), 273–293. https://doi.org/10.1037/0033-2909.133.2.273

Zheng, Y., Zacks, J. M., & Markson, L. (2020). The development of event perception and memory. *Cognitive Development*, *54*, Article 100848. https://doi.org/10.1016/j.cogdev.2020.100848