Spatially Perturbed Collision Sounds Attenuate Perceived Causality in 3D Launching Events

Duotun Wang*[1], James Kubricht*[2], Yixin Zhu*[3], Wei Liang†[1], Song-Chun Zhu[3], Chenfanfu Jiang[4], and Hongjing Lu[2]

[1] Laboratory of Intelligent Information Technology, Beijing Institute of Technology
[2] Computational Vision and Learning Laboratory, UCLA
[3] Center for Vision, Cognition, Learning and Autonomy, UCLA
[4] Computer Graphics Group, UPenn

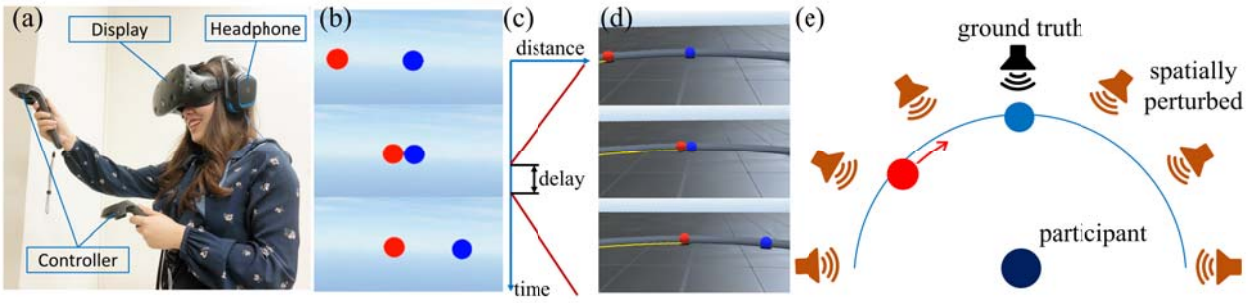

Figure 1: Illustration of (a) VR materials utilized in the present study as well as the virtual environment for (b) 2D and (d) 3D collision events. (c) In each environment, the red ball moves to the impact location, stops, and after a delay the blue ball moves forwards. (e) The collision was accompanied by an auditory collision indicator which was spatially perturbed across trials in the 3D environment.

Abstract

When a moving object collides with an object at rest, people immediately perceive a causal event: *i.e.*, the first object has *launched* the second object forwards. However, when the second object's motion is delayed, or is accompanied by a collision sound, causal impressions attenuate and strengthen. Despite a rich literature on causal perception, researchers have exclusively utilized 2D visual displays to examine the launching effect. It remains unclear whether people are equally sensitive to the spatiotemporal properties of observed collisions in the real world. The present study first examined whether previous findings in causal perception with audiovisual inputs can be extended to immersive 3D virtual environments. We then investigated whether perceived causality is influenced by variations in the spatial position of an auditory collision indicator. We found that people are able to localize sound positions based on auditory inputs in VR environments, and spatial discrepancy between the estimated position of the collision sound and the visually observed impact location attenuates perceived causality.

Index Terms: Causal perception; virtual reality; intuitive physics; visual capture; launching

1 Introduction

Consider the following visual display: a red circle moves in a straight line towards a blue circle until the edges of the two circles touch. After the two circles make contact, the blue circle moves away from the red circle along the same straight line. Although there is no directly observable information in the display signaling a causal connection between the motions of the two circles, you will most likely perceive the red circle as having *launched* the blue circle forward [20]. This is an example of causal perception: *i.e.*, the immediate, automatic, and irresistable impression of causality and animacy from low-level perceptual inputs [45]. Such impressions lie in contrast with high-level causal inference, which describes how real-world interpretations are made using logical rules and conceptual knowledge [39].

A key characteristic of causal perception is that impressions are constructed at the perceptual level and do not rely on explicit background knowledge or experience [32, 39, 43, 45]. However, causal impressions are incredibly sensitive to the spatiotemporal properties of dynamic events [8, 32, 34]. For example, in the aforementioned visual display, (1) if there is a temporal delay between when the red circle stops and the blue circle begins moving, (2) if the edges of the circles are separated or overlapped at the time of impact, or (3) if the blue ball moves perpendicular to the red ball's motion prior to impact, the impression that the red ball launched the blue ball will diminish [30, 42, 44]. Impressions of launching are also influenced by briefly observed motions of nearby shapes [44], indicating that the human perceptual system rapidly processes spatiotemporal information and visual context information to form immediate causal impressions from perceptual inputs [15, 39].

To date, researchers have exclusively utilized 2D visual displays to examine causal perception [30, 42, 44, 45]. This is largely due to the difficulty in varying the spatiotemporal characteristics of moving objects in the real world. Specifically, it is prohibitively difficult to "pause" a real-world collision at the moment of impact without some costly external apparatus: *e.g.*, using magnets and a digital controller to move metallic objects at a given speed across an opaque track. However, virtual reality (VR) provides the means to manipulate such characteristics in immersive 3D environments. A secondary manipulation that VR technology affords is the perturbation of a sound's location in 3D space. Previous work has shown that when a collision event is accompanied by an auditory cue indicating contact between two objects (*e.g.*, a *clack* sound), observers report a greater causal impression than when the sound is absent [18]. It remains unclear, however, whether the human perceptual system encodes the location of the sound when forming such impressions, as it does when it infers that a ventriloquist's voice emanates from a nearby dummy [1]. Thus, the present study sought to answer the following questions: (1) do classical findings in causal perception extend to 3D virtual environments, and (2) does the spatial position of an auditory collision indicator influence perceived causality?

Three experiments were conducted to address these questions. In each experiment, an initially moving red object collides with an initially stationary blue object, and after a 0 to 400 msec delay, the blue object begins moving forwards (see Fig. 1 (b)(d)). In Experiment 1, participants reported causal impressions in 2D launching events in the presence and absence of an auditory collision indicator. The first experiment was a direct replication of Guski and Troje's [18] previous study and was designed to determine whether their findings extend to tasks presented via a VR apparatus. In Experiment 2, participants completed an identical task but in a 3D virtual environment. The collision sound in the second experiment was always located at the ground-truth position: *i.e.*, at the location of impact. The purpose of Experiment 2 was to ensure that previous findings in audiovisual

* D. Wang, J. Kubricht and Y. Zhu contributed equally to this work.
† Corresponding author. e-mail: liangwei@bit.edu.cn

2018 IEEE Conference on Virtual Reality and 3D User Interfaces
18-22 March, Reutlingen, Germany
978-1-5386-3365-6/18/$31.00 ©2018 IEEE

causal perception extend to 3D collision situations. Experiment 3 was identical to Experiment 2, except that the spatial position of the auditory collision indicator was perturbed $\pm 90°$ around the observer in increments of 30° (see Fig. 1 (e)).

In summary, the present study made the following contributions: (1) replicated previous work of causal perception in a virtual environment to demonstrate the viability of VR technology in examinations of human perception and cognition, (2) examined the effect of spatially perturbed auditory collision indicators on impressions of causality in delayed launching events, and (3) measured how well humans can estimate sound location in a VR setup. The remainder of the paper is structured as follows: Sect. 2 discusses related work in causal perception and virtual reality, Sect. 3 describes the method and results for the aforementioned experiments, and Sect. 4 discusses our findings and proposes future directions for further work.

2 Background and Related Work

The ability to perceive causality is fundamental for making sense of the dynamic world. It emerges early in human life, as 6-10 month-old infants are sensitive to cause-effect relationships in visual scenes [12, 24]. Following the classical work of Michotte [32], object collisions–and more specifically, launching events–have been demonstrated as invaluable physical systems for examining causal perception (*e.g.*, [18,30,42,44,45,52]). Their utility follows from the strong dependency of causal impressions on the spatiotemporal characteristics of perceived events. Importantly, displays are generally devoid of any high-level indicators of causal agency (but see [30]). Thus, the problem of causal inference can be modeled at the perceptual level, which is described in Sect. 2.1. Sect. 2.2 outlines recent applications of VR technology in academic studies.

2.1 Causal Perception

Recently, researchers have proposed that causality is perceived by inferring the marginal probability of a causal relationship given observational evidence in visual scenes, which has been evaluated under the *noisy Newton* framework [17,42]. Since people's estimates of relatively simple perceptual variables (*e.g.*, distance, velocity, and time) are inherently uncertain, their values must be inferred based on prior expectations. For example, people expect that objects are more likely to move slowly than quickly, and in causal collision events, the initially stationary object should move immediately after the initially moving object makes contact (*i.e.*, there is no spatial separation between objects at the moment of impact) [42]. When observing collision events, perceptual estimates of velocity, spatial separation, and temporal delay are consistently biased towards these prior expectations.

The noisy Newton framework further assumes that humans have an internal physical model encoded in neural circuitry [14] which approximates ground-truth physical principles to propagate noisy perceptual inputs forwards in time. Given a causal physical model (or schema [38]) for collision events, objects should move in accordance with the principle of conservation of momentum, and given a noncausal model, objects should move randomly. The observed speeds of the two objects (pre- and post-collision) as well as the observed temporal delay and spatial separation in a perceived collision are compared to each model's predictions to assess the relative likelihood of a causal interaction. This provides a quantitative estimate of human causal impressions across launching events that vary in their spatiotemporal characteristics. The resulting predictions align well with people's causal ratings: *i.e.*, as the the temporal delay and spatial separation in a launching event increase, causal ratings decrease [42].

Although the noisy Newton framework has demonstrated success explaining a breadth of intuitions that people have about the physical world (see [23] for a review), situations involving both visual and auditory information have yet to be modeled. It remains unclear whether certain characteristics of auditory information (*e.g.*, the spatial position of sound) are encoded by the perceptual system when forming launching impressions. The present work makes an initial stride towards determining the depth and complexity of perceptual information utilized when inferring the causal structure of the physical world.

2.2 Virtual Reality (VR)

Virtual reality technology has demonstrated itself as a low-cost and effective means to test and train humans in rare and extreme environments. For example, VR systems–and gaming systems, broadly–have been utilized to train users in disaster prevention exercises [33,37,49,53], medical emergency scenarios [2,48], firefighting simulations [5, 10, 50], aviation safety [11], and general traffic and fire safety [6,31,36]. These systems are incredibly useful for training purposes, as they allow for the simulation of dangerous situations in a safe and controlled environment.

However, previous generations of VR devices have been limited by their computational capabilities, as well as their relatively poor sound and video quality. Therefore, the aforementioned studies have primarily focused on procedural training in critical situations. Traditionally, VR systems have also suffered from their limited accessibility, as they often required large-scale testing environments and prohibitively bulky motion tracking devices operated by experienced personnel. Such drawbacks have hindered the use of VR technology in cognitive studies, where participants (and experimenters) generally lack the experience needed to work with complex VR apparatuses. Thus, previous behavioral studies in cognitive science which utilize video game and virtual reality technology have been largely restricted to problems represented at the symbolic level [22, 35, 40].

Recently, however, the virtual reality industry has addressed many of these shortcomings. With the increasing popularity of consumer-level VR devices (*e.g.*, Oculus Rift/Touch, HTC Vive, Google Daydream, *etc.*) as well as recent advancements made in general-purpose GPU implementations, the auditory and visual quality of modern VR systems have dramatically improved. Over the past two years, VR has become increasingly popular in academic research and has demonstrated itself as an established, albeit relatively new, method for administering sophisticated and detailed tasks. Examples of such applications include fine-grained earthquake simulation and disaster prevention [27], visual navigation [57], semantic planning [56], and robot grasping [19, 54]. General purpose VR platforms have also been utilized to examine human-scene interactions [28] and optimize autonomous vehicle policies [47]. In cognitive studies, researchers have demonstrated the utility of VR devices in examinations of human deceptive behavior [3], physical intuitions in novel gravity fields [55], visuomotor adaptation [29], haptic retargeting [4], and locomotion and motion perception [9]. The present study aims to further establish the viability of VR experiments in cognitive science by examining causal perception in immersive, 3D virtual environments.

3 Experiments

Participants A total of 36 participants (17 female; 19 male) were assigned to three separate experiments in the present study. Of the 36 participants, 10 (3 female; 7 male) were assigned to Experiment 1, 10 (5 female; 5 male) were assigned to Experiment 2, and 16 (9 female; 7 male) were assigned to Experiment 3. Participants were either undergraduate or graduate students. All participants had either normal or corrected-to-normal vision and reported normal hearing ability. The average age of participants was 22.5 ($SD = 2.3$), and each participant was randomly assigned to one of the three experiments.

Most participants had little or no experience interacting with VR systems. Of the 36 participants who completed the study, only 2 reported having experience using VR devices for more than 10 hours over the past year; all other participants reported having 0-5 hours of experience over the past year. Participants were also asked to report their experience playing video games. Only 2 participants indicated having experience with XBox or Playstation gaming systems. One third of the participants reported that they played PC games every week, and 10 participants reported that they played cell phone games regularly.

Ethics Statement The current experiment received approval from the Institutional Review Board and was confirmed as having no conflicts of interest. The study had minimal risk, and participation was voluntary: *i.e.*, participants could choose to halt the study at any time. Oral consent was obtained prior to each experiment session by the experimenter, and no identifying information was attached to the collected data.

Apparatus and Procedure The VR environments in the present study were designed in the Unity3D 2017.1 engine and were administered via an HTC Vive system (see Fig. 1 (a)). In each experiment, participants wore a Vive head-mounted display (HMD) which provided visual input to the user. The HMD consists of two screens (one for each eye), each providing a 1080×1200 resolution image to the user at a refresh rate of 90 Hz. Participants also held a pair of HTC Vive controllers which were tracked by two Vive base stations mounted on the walls of the experiment room. The trackpad on each controller was programmed to work as a laser pointer which was used to provide user input to the VR system. Participants traversed instruction windows and reported causal ratings by pointing to their choices and pressing the controller's trigger to indicate their selection. Auditory input was provided by a pair of Logitech G430 gaming headphones with 7.1 channel surround sound output. The headphones provided stable and accurate sound localization, which was imperative for Experiment 3.

To inhibit interference due to external stimuli, each experiment was conducted in a quiet and spacious testing room. During the experiment, participants sat on a swivel chair that was free to rotate about a 360° angle. Although they were unable to traverse the 3D environment, participants were free to change their viewing angle during each trial. Prior to each experiment, participants were provided with instructions showing them how to wear the HMD and headphones, as well as how to interact with the system using the controller. Participants were told to report any sickness or discomfort with the apparatus at any point in the experiment and that they could terminate their session at any time. Instructions for each experiment were provided via a window in the VR environment; participants read the instructions and indicated, with their controller, when they were ready to proceed to the next window. This was done to prevent any bias resulting from verbal instructions from the experimenter. Participants were, however, told to notify the experimenter if they had any questions during the instruction period.

Experiment Overview Experiments 1 and 2 were designed to replicate Guski and Troje's [18] primary finding (*i.e.*, provision of an auditory collision indicator in launching events increases perceived causality) in 2D and 3D VR setups, respectively. The purpose of the replication was (1) to ensure that measured performance in the response data did not arise due to the use of the VR interface and (2) to determine whether previous findings in audiovisual causal perception extend to realistic, 3D environments. The procedure for each experiment was similar to that of Guski and Troje: Participants viewed a red object (a 2D circle in Experiment 1; a 3D ball in Experiment 2) move towards a stationary blue object until the edges of both objects coincided. Once contact was made, the red object stopped, and after a delay the blue object began moving forwards. The movement of the blue object following impact was delayed with duration in the range of 0 to 400 msec in each experiment. A schematic drawing illustrating the temporal delay is shown in Fig. 1 (c). Following observation of each collision, participants reported the degree to which they perceived the red object as having *launched* the blue object forwards. The objects moved along a straight trajectory in Experiment 1 (*i.e.*, the 2D environment) but moved along a circular trajectory in Experiment 2 (*i.e.*, the 3D environment; see Fig. 1 (d)). In Experiment 2, the collision sound was always located at the ground-truth position: *i.e.*, the point of impact.

Experiment 3 was identical to Experiment 2, except that the spatial position of the collision sound was varied across the circular movement path. A circular path was utilized in Experiments 2 and 3 so that the volume of the collision sound did not vary with the magnitude of spatial perturbation: *i.e.*, if a straight trajectory was used, the volume would decrease as the magnitude of perturbation–and thus the spatial distance of the sound–increased (see Fig. 2). Given a straight trajectory in 3D space, it would be unclear whether differences in causal ratings were due to variations in the spatial position of the sound or the volume. However, since all points on a circular path are equidistant to the observer, the effect of volume on causal ratings was effectively removed. Experiment 2 also served as a control experiment for Experiment 3. Without the second experiment, it would be unclear whether causal ratings in Experiment 3 were influenced by the spatial positions of the collision sound or the extension of motion from a straight 2D trajectory to a circular 3D trajectory.

Experiment Setting To prevent any carryover effects between experiments, we employed a between-subjects design. In both Experiments 1 and 2, participants completed 36 trials in a randomized order without a break. The temporal delay was varied between 0 and 400 msec in increments of 50 msec, resulting in 9 temporal delays total. The same trials were presented to each participant twice. Eighteen trials were presented with a collision sound, and the remaining 18 were presented with no sound. Both Experiments 1 and 2 took approximately 30 minutes to complete. After viewing the collision in each trial, participants were asked "Did the red object launch the blue object?" and gave their rating on a virtual slider ranging from "Definitely No" to "Definitely Yes". Participants were told that a rating of "Definitely Yes" should correspond to "a strong impression that the red ball set the blue ball into motion by pushing it forward" and a rating of "Definitely No" should correspond to "a strong impression that the red ball did not influence the motion of the blue ball". The reading of the slider was later converted into the rating scale used in previous work [18] via the following expression:

$$\frac{s-1}{9} \leq \frac{p-p_{\min}}{p_{\max}-p_{\min}} \leq \frac{s}{9}, \tag{1}$$

where p is the reading of the slider, $p_{\min}$ and $p_{\max}$ are the minimum and maximum values of the slider, and s is the converted rating scaled from 1 to 9 with a step size of 1. Here, a rating of 1 corresponds to the lowest possible rating, and a rating of 9 corresponds to the highest.

In Experiment 3, the same 9 temporal delays were included, but the spatial position of the collision sound was also varied between -90° and 90° in increments of 30° around the observer (8 collision sound conditions total, including the without-sound condition). Each Trial was repeated twice, yielding a total of 144 trials which were separated into two blocks and presented in a randomized order. Following observation of the collision in each trial, participants were asked the same question as in Experiments 1 and 2. After completing the first two blocks, participants completed a third block in which they were asked to indicate on a slider where they estimated the collision sound came from. Nine trials without a collision sound were present in each of the first two blocks, but removed from the third block, yielding 63 trials in the third block presented in a randomized order. Participants were given the opportunity to take a break between blocks in Experiment 3. Trials in the third block of Experiment 3 differ from the trials in the preceding two blocks–as well as the trials in Experiments 1 and 2–in that participants were explicitly prompted to consider the auditory collision indicator *before* giving their causal rating. This allowed us to examine whether (1) people's position estimates of collision sounds are accurate or biased; and (2) whether increased attention to an auditory cue impacts perceived causal impressions. It took participants approximately two hours to complete the three blocks in Experiment 3.

The size and speed of the objects in the 2D and 3D environments were also matched to previous work in causal perception. In the 2D and 3D environments, each object subtended 10.1° and 10.7° of visual angle, respectively. In both environments, the objects moved at an angular speed of 13.2°/sec. The post-collision speed of the blue object was matched to the pre-collision speed of the red object, indicating a perfectly elastic collision (*i.e.*, no energy was lost to heat/friction or deformation in the collision) where each object was equally heavy. In each experiment, the size and surface material of each object were also identical (only the colors of the objects were different). In Experiment 1, each launching event lasted approximately 10 seconds. In Experiments 2 and 3, each event lasted approximately 15 seconds.

3.1 Experiment 1: Causal Perception in 2D motion

In the first experiment, participants were asked to provide causal ratings after viewing launching events in a 2D virtual environment in the presence and absence of an auditory collision indicator. The

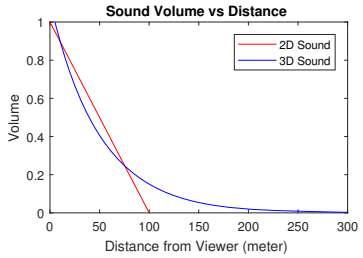

Figure 2: Synthesized sound volume plotted as a function of distance between the sound location and the observer. In Experiment 1, a linear model was used to mimic the *clack* sound described in the original study [18]. In Experiments 2 and 3, a logarithmic model was used to emulate the position and motion of the sound in the 3D environment.

aim of Experiment 1 was to compare causal impressions in a virtual environment with previous work which presented collision events using an LCD projector [18]. The movement of each circle was confined to a straight line in 2D space, and the shading was removed in the visual rendering (see Fig. 1 (b)). The design and method of Experiment 1 was identical to Guski and Troje's previous study; the only difference was the hardware used to present collision events and the question for measuring participants' causal impressions. Our aim was to provide a causal question to participants that was consistent with previous studies in the causal perception literature [45]. Thus, following observation of each collision, participants were simply asked "Did Object A launch Object B?".

2D Sound Synthesis In order to replicate the *clack* sound presented to participants in the original study [18], auditory spreading effects in the virtual environment were removed. The spatial blending setting was turned to 2D mode, and reverberation effects were turned off. A linear model was used to determine the volume of the sound as a function of distance from the observer (see Fig. 2). Taken together, the sound settings constrained both auditory channels to have equal volume, which effectively emulated the previously employed collision sound.

A bowling ball collision sound (imported to the Unity engine) was utilized in the present study. The sound lasted for approximately 10 msec in the 2D environment, and its pitch P (or frequency) was modified based on the speed of the object via the following expression:

$$P = \frac{s_t}{s_0} + b, \tag{2}$$

where s_t is the speed of the object, s_0 is the reference speed of the object (*i.e.*, the maximum linear speed of the ball), and b is the pitch offset. The pitch of the collision sound was manipulating using Eq. 2 so that impacts resulting from slower objects corresponded to lower-frequency collision sounds. Since the speed of the ball s_t in the 2D environment was constant across trials, the pitch also remained constant.

Training Session Participants began the experiment by reading through a set of instructions presented on windows (or screens) in the virtual environment. Participants were provided with a static depiction of a launching event (pre- and post-collision), and were briefed about the task. Once participants finished reading through the instructions, they were told that they would begin the experiment by completing a set of practice trials. Following Guski and Troje's experimental procedure, participants were informed that the practice trials would consist of both good *and* bad examples of launching, as well as examples that are somewhere in between. This was done to establish the bounds of participants' individual rating scales: i.e., what should be perceived as launching and what should not. Information about the presence of an auditory collision indicator was not provided to the participants.

There were 6 practice trials at the beginning of the training session, depicting low (0 msec), medium (200 msec) and high (400 msec) temporal delays in the presence and absence of a collision sound. The trials were presented in a randomized order, and the

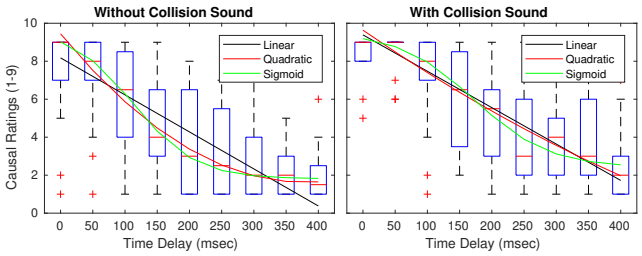

Figure 3: Box plot of causal ratings in Experiment 1 in the (left) absence and (right) presence of a collision sound. Red horizontal lines indicate median causal ratings, and the bottom and top edges of the blue boxes indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points that were not considered outliers, and red '+' symbols indicate outliers. The x-axis shows the temporal delay between the moment of impact and the start of the blue circle's motion. The black, red and green lines are linear, quadratic and logistic regression plots, respectively.

ground-truth configuration (temporal delay and presence of a collision sound) was not explicitly provided to participants. No feedback was provided following completion of each practice trial. Once participants finished the set of practice trials, they continued to the testing session.

Testing Session After completing the practice trials, participants were presented with 36 testing trials in a randomized order. The stimulus parameters (*i.e.*, the temporal delays and presence/absence of an auditory collision indicator) were identical to the parameters used in the original study by Guski and Troje [18]. Participants answered the same question as in the practice trials, and provided their causal ratings. Ratings from each trial were converted to a 1-9 scale (step size of 1) using Equation 1 and logged into the VR system. Participants did not receive feedback.

Results Consistent with previous findings, median causal ratings in Experiment 1 declined as temporal delay increased in both conditions (*i.e.*, presence or absence of a collision sound; see Fig. 3). Since participants' causal ratings in the with-sound and without-sound conditions were skewed towards moderate values, the assumption that the ratings followed a normal distribution were not satisfied. Thus, non-parametric statistical analyses were performed on causal ratings. A Friedman test was conducted on measured causal ratings, indicating a significant effect of the presence of a collision sound: $\chi^2(1) = 16.9$; $p < 0.001$. This result is in agreement with previous findings in audiovisual causal perception.

Linear, quadratic and logistic regression were performed on median causal ratings as a function of temporal delay. Note that although causal ratings were skewed, the median prediction errors satisfied the assumption of normality in the regression analyses. For each model, we calculated both the correlation and Bayesian information criterion (*BIC*; lower value indicates superior fit), which is penalized for model complexity: *i.e.*, the number of free parameters in each model. In the absence of a collision sound, causal ratings showed a strong quadratic ($r^2 = 0.98$, $BIC = -12.77$) and logistic ($r^2 = 0.995$, $BIC = -25.03$) trend, whereas the linear trend was less pronounced ($r^2 = 0.69$, $BIC = -0.04$). In contrast, ratings in the presence of sound showed a strong linear ($r^2 = 0.94$, $BIC = -6.00$), quadratic ($r^2 = 0.92$, $BIC = -4.36$) and logistic ($r^2 = 0.96$, $BIC = -3.95$) trend. Although the quadratic and logistic trends were both pronounced in the with-sound condition, the curve was nearly linear. Since the linear trend had the smallest *BIC* value, we concluded that ratings in the with-sound condition were best fit by a linear function. This is further evidenced by the smaller logistic regression slope in the with- versus without-sound condition: b = 3.01 versus 3.60, respectively. The present results agree with Guski and Troje's previous findings: *i.e.*, ratings in the presence and absence of an auditory collision indicator were best fit by a linear and quadratic function, respectively.

Although the overall trends in our results were consistent with Guski and Troje's study, there were some discrepancies. While we

found no observable difference between causal ratings in the with-sound and without-sound conditions at a temporal delay of 400 msec (see Fig. 3), Guski and Troje reported with-sound causal ratings approximately twice as large as without-sound ratings ([18]; Fig. 3, pg. 794). One potential reason is that the studies used different wording when instructing participants to provide their rating judgments. While participants in the present study were simply asked "Did the red ball launch the blue ball?", Guski and Troje's study asked "How probable is it that the movement of the blue object (disk or ball) is caused by a perceivable event immediately before?". The two questions differ in one critical aspect: *i.e.*, the type of causal event (sound or motion) was left open to participants in the Guski & Troje study, whereas the auditory cue was not alluded to with the wording used in the current experiment. Since their experimental question was designed to call attention to the auditory cue (implied by "a perceivable event immediately before"), it was likely to have biased participants to give higher ratings in the with-sound condition, regardless of temporal delay.

Taken together, results from Experiment 1 confirm that the impact of auditory collision indicators on perceived causality extends to launching events presented in a VR system. In agreement with previous findings [13,18,21,25,26,41,46,51], the present results further demonstrate that visual and auditory events appear to be linked together given that they occur within 200 msec of one another.

3.2 Experiment 2: Causal Perception in 3D motion

The second experiment was identical to Experiment 1, except that launching events were presented in a 3D virtual environment. The primary purpose of Experiment 2 was to determine whether previous findings in causal perception extend to 3D virtual environments. Instead of circles, participants viewed red and blue balls, which were rendered using 3D meshes with natural lighting and shading. A grey tile floor and a curved wall indicating the edge of the circular trajectory were placed into the environment to facilitate motion perception in 3D space. The curved path of each object was also highlighted (via a yellow line that trailed each ball) in order to facilitate perception of the circular trajectory (see Fig. 1 (d)).

3D Sound Synthesis The virtual environment employed in Experiment 2 not only affords 3D vision, but also 3D auditory perception: *i.e.*, the Unity engine provides native tools and configurations to accurately emulate a sound's position and motion in 3D space. This was achieved by setting the spatial blend option in the Unity engine to 3D mode and utilizing the standard *surrounding 7.1* audio environment offered by the *Microsoft HRTF Spatializer*. The pitch of the collision sound was calculated using Equation 2, and reverberation effects were present but set to a low value. A logarithmic roll off mode was used to determine the relationship between volume and the distance of the sound to the observer (see Fig. 2). A logarithmic model was utilized so that small differences in spatial distance (specifically, the distance between each observer's ear) led to a relatively large change in volume. This was especially important for Experiment 3, where observable differences in volume between each ear were needed to localize auditory signals in 3D space. The collision sound in the 3D environment lasted approximately 6 msec.

Procedure The 6 practice trials in the training session of Experiment 2 were the same as in Experiment 1. The instructions and causal question in each trial were also identical to the previous experiment, except an additional window of instructions was provided to ensure participants' bodies were oriented towards the same position at the beginning of the training session. In the additional instructions window, participants were told to look for a fixation cross ('+' symbol) located near the left edge of the blue ball and face towards it by rotating their swivel chair clockwise. A button was placed over the fixation cross, and after it was pressed, the session began. This manipulation was made to ensure that participants observed each collision from the same viewing angle in the 3D environment.

Following the practice trials, participants completed 36 testing trials with the same temporal delay parameters as in Experiment 1. The testing trials in Experiment 2 were identical to the trials in Experiment 1, except that the experiment was paused for 2 seconds before the beginning of each trial to give participants time to prepare.

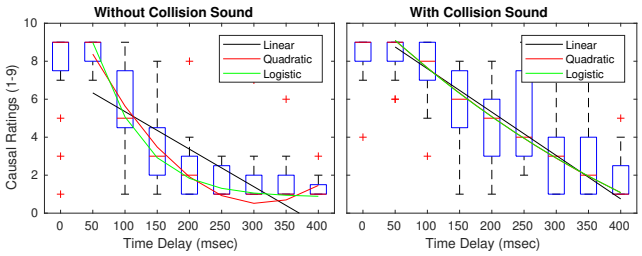

Figure 4: Box plot of causal ratings in Experiment 2 in the (left) absence and (right) presence of a collision sound. Red horizontal lines indicate median causal ratings, and the bottom and top edges of the blue boxes indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points that were not considered outliers, and red '+' symbols indicate outliers. The x-axis shows the temporal delay between the moment of impact and the start of the blue ball's motion. The black, red and green curves are linear, quadratic, and logistic regression plots, respectively. Compared to Experiment 1, all regressions begin at a temporal delay of 50 msec instead of 0 msec.

Participants also oriented their bodies towards the fixation cross prior to the testing session according to the same procedure employed in the training session.

Results Median causal ratings in Experiment 2 declined as temporal delay increased in both the with- and without-sound conditions (see Fig. 4). Results from a Friedman test indicate that causal ratings in the presence of a collision sound were significantly greater than ratings in the absence of a collision sound: $\chi^2(1) = 51.6$; $p < 0.001$. Comparing results from Experiments 1 and 2, the effect of the collision sound was greater in the 3D environment compared with the 2D environment, as evidenced by the larger Friedman test statistic. The logistic regression slope in the with-sound condition was also smaller than in the without-sound condition ($b = 0.32$ versus $b = 2.66$), and the difference between slopes was more pronounced than in Experiment 1.

It is important to note that causal ratings in the 3D environment did not decrease until after a temporal delay of 50 msec, which agrees with Michotte's classical work [32]. Thus, regression analyses were performed starting from a temporal delay of 50 msec. Regression results for the without-sound condition agree with the results in Experiment 1: *i.e.*, causal ratings in the absence of sound were best fit by a quadratic ($r^2 = 0.97$, $BIC = -8.83$) and a logistic ($r^2 = 0.997$, $BIC = -27.47$) trend rather than a linear one ($r^2 = 0.72$, $BIC = 7.56$). Similar to Experiment 1, ratings in the presence of a collision sound also showed a strong linear trend ($r^2 = 0.99$, $BIC = -18.09$), but the quadratic ($r^2 = 0.995$, $BIC = -23.19$) and logistic ($r^2 = 0.995$, $BIC = -19.22$) trend were also pronounced.

There were two notable differences between causal ratings in Experiment 2 (3D environment) and Experiment 1 (2D display): (1) Causal ratings did not decrease until a temporal delay of 50 msec, and (2) quadratic and logistic *BIC* values in the with-sound condition were lower than the linear *BIC* value. The first difference was likely due to the immersive characteristics of the 3D environment: *i.e.*, since the motions of the objects appeared more natural, causal ratings were raised to ceiling levels for small (¡50 ms) temporal delays. The second difference arises because the slope in the quadratic and logistic regression models were so small that the curves were essentially linear. Although the two models were penalized for having an extra free parameter in the BIC calculation, their predictions were slightly more accurate leading to roughly equivalent BIC values. This remains consistent with Experiment 1's results as well as Guski and Troje's: *i.e.*, without-sound causal ratings are best fit by a nonlinear (quadratic or logistic) model, and with-sound ratings are effectively fit by a linear one.

3.3 Experiment 3: Spatially Perturbed Collision Sounds

The third experiment was designed to examine whether the spatial position of an auditory collision indicator influences perceived

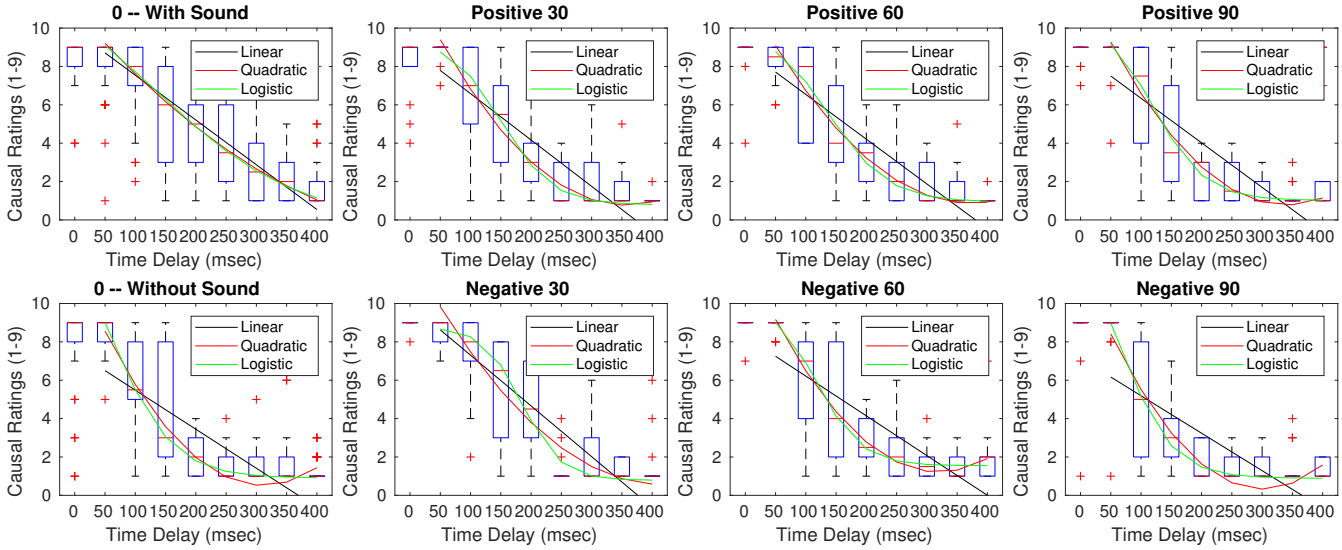

Figure 5: Box plot of causal ratings in Experiment 3. Red horizontal lines indicate median causal ratings, and the bottom and top edges of the blue boxes indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points that were not considered outliers, and red '+' symbols indicate outliers. The x-axis shows the time delay between when the red ball stopped and the blue ball began moving. The black, red and green curves are linear, quadratic and logistic regression plots, respectively.

causality. The experiment setting was identical to the setting in Experiment 2 (3D environment), except the spatial position of the collision sound was varied between $\theta = -90°$ and 90° of visual angle from the ground-truth impact location in increments of 30°. In addition, we measured how well participants could estimate the spatial locations of the collision sounds from auditory inputs. The position of the sound in 3D space for each of the angular perturbations θ was calculated using the following expressions:

$$x_s = x_0 - r\ \sin(\theta) \tag{3}$$

$$y_s = y_0 \tag{4}$$

$$z_s = r\ \cos(\theta) \tag{5}$$

where (x_s, y_s, z_s) is the spatial position of collision sound and (x_0, y_0, z_0) is the ground-truth sound position: *i.e.*, the location of impact between the two objects. The Unity engine was employed to vary the position of collision sound, and the pitch was calculated using Equation 2. Once again, the volume followed a logarithmic trend over distance (Fig. 2).

Procedure There were 12 practice trials total comprised of three different temporal delays (low [50 msec], medium [200 msec], and high [400 msec]) and 4 different sound settings (no sound and sound at θ=-90°, 0°, and 90°). Note that the three delays used in the training trials in Experiment 3 were different from the delays used in Experiments 1 and 2. This was due to the previous finding that causal ratings did not begin to decrease until after 50 msec of temporal delay. The order of the training trials was randomized.

Following completion of the practice trials, participants proceeded to the testing trials outlined at the beginning of this section. The testing trials were divided into three blocks. The question in the first two blocks was the causal rating question as in Experiment 2. In the third block, participants were also asked to use their controller to indicate (on a slider) where in the 3D environment they believed the collision sound came from.

Results We first compared causal ratings in conditions with symmetric collision sound positions which were spatially perturbed (*i.e.*, ±30°, ±60°, and ±90°). This comparison aimed to determine whether causal impressions depended on whether the sound came from past or future object locations (relative to the impact location). Results from a set of Friedman tests indicate that participants' causal ratings were the same between ±30° ($\chi^2[1] = 0.3$, $p = 0.56$), ±60° ($\chi^2[1] = 0.4$, $p = 0.55$), and ±90° ($\chi^2[1] = 2.9$, $p = 0.09$). Thus, data in the symmetric sound positions were aggregated and then compared to ratings in the (ground truth; θ = 0°) with- and without-sound conditions. We found that causal ratings in each of the perturbed location groups were significantly different from ratings in the with-sound condition at the ground-truth location: $\chi^2(1) = 14.6$, $\chi^2(1) = 18.8$, and $\chi^2(1) = 34.22$ for the ±30°, ±60°, and ±90°conditions, respectively. Although ratings in the ±90° sound conditions were not statistically different from ratings in the without-sound condition ($\chi^2[1] = 0.4$, $p = 0.5$), ratings in the ±30°and ±60°sound conditions were significantly different: $\chi^2(1) = 9.5$, $p < 0.01$ and $\chi^2(1) = 5.7$, $p = 0.02$, respectively. These results indicate that while a collision sound located directly to the left or right of an observer had no impact on perceived causality, a sound located 30° or 60° to the left or right of an observer (measured in angular distance from the impact location) had an effect, but not as much as a collision sound located at the ground-truth position.

Linear, quadratic, and logistic regression were performed on causal ratings in each of the sound location conditions. Squared correlation and *BIC* values for each condition are depicted in Table 1. The regression results agree with those of Experiment 2 in that a quadratic and logistic trend fit best to causal ratings in the without-sound condition, whereas linear, quadratic and logistic trends were roughly equivalent when the sound came from the ground-truth position. Interestingly, the linear trend became more prominent as the position of the sound shifted from ±90° to the ground-truth position. Also, in the ±60° and ±90° sound conditions, the linear trend was

Table 1: Squared correlation coefficients and BIC for linear, quadratic and logistic regression analyses conducted on causal ratings in each of the sound position conditions. The quadratic and logistic fits appear to fit equally well in each sound position condition, whereas the linear fit appears to improve as the sound location approaches the ground-truth position.

Angular Sound	Linear		Quadratic		Logistic		
Position (θ)	r^2	*BIC*	r^2	*BIC*	r^2	*BIC*	Slope *b*
No Sound	0.73	7.48	0.98	-11.57	0.998	-29.55	-0.32
−90°	0.67	9.20	0.96	-6.47	0.992	-14.83	-4.04
90°	0.82	5.74	0.97	-7.42	0.98	-6.05	-4.04
−60°	0.77	5.16	0.99	-15.04	0.99	-12.30	-4.63
60°	0.86	3.20	0.96	-4.45	0.97	-2.58	-3.54
−30°	0.88	3.99	0.94	0.64	0.99	-6.63	-5.39
30°	0.86	4.18	0.98	-8.85	0.99	-10.46	-4.15
0°	0.98	-13.04	0.99	-20.85	0.995	-18.57	-1.46

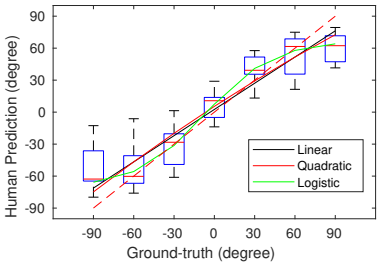

Figure 6: Participants' sound location predictions versus ground-truth positions. The black, red, and green lines indicate a linear, quadratic and logistic regression, respectively. The dashed red line depicts ideal causal predictions.

more pronounced when the sound came from a position in the right half of participants' visual field. Although causal ratings were not statistically different between sound locations in the left/right halves of the visual field, the regression results suggest that causal ratings decreased at a more linear rate (as a function of temporal delay) when the sound came from an angular position of 60 and 90° to the right of the impact location ($BIC = 3.21$ and 5.74) compared to 60 and 90° to the left ($BIC = 6.16$ and 9.20; see Table 1). This difference was not observed in the ±30°sound conditions.

We also examined how well participants were able to estimate spatial locations of collision sounds in the VR environment. Fig. 6 depicts causal predictions plotted against ground-truth sound positions. Although participants were generally accurate in predicting the direction from which the collision sounds were heard, their estimates were biased towards the visually perceived impact location when the sound was relatively far away ($|\theta| \geq 60°$). Linear, quadratic, and logistic regression were performed on participants' location predictions, which indicated a superior fit by the logistic trend (logistic: $r^2 = 0.997$; linear: $r^2 = 0.96$; quadratic: $r^2 = 0.96$). We further examined sound location predictions when the sound was generated at the ground-truth impact location. Interestingly, participants' position estimates were biased in the direction of the objects' motion: $t(143) = 2.3, p = .03$. Taken together, the results of Experiment 3 indicate that humans are capable of spatially locating sounds in 3D virtual environments, and spatial auditory information is utilized by the perceptual system when forming immediate causal impressions.

In Experiment 1, we called attention to the difference between the present causal rating question and the question asked by Guski and Troje. However, the manipulation made to the task in the third block of Experiment 3 (*i.e.*, the collision sound location component) alluded to the auditory collision sound just as Guski and Troje's question did. Similar to their findings, causal ratings in the with-sound condition were approximately twice the magnitude of ratings in the without-sound condition when participants were additionally tasked with locating the collision sound in 3D space (see Fig. 7). This finding suggests that any minor discrepancies between our results and those of Guski and Troje arose due to differences in the causal questions that were asked.

4 Conclusions and Future Work

In the present study, we demonstrated that (1) classical results in causal perception [18, 32, 42] can be replicated in novel virtual environments, (2) humans are able to perceive the virtual location based on auditory inputs, although estimates were biased in some cases, and (3) variations in the position of an auditory collision indicator lessen the impression of causality in dynamic scenes.

The third result is best understood in the context of visual capture: i.e., the prevalence of visual information in sensory integration [16]. One common example is the ventriloquist effect, whereby a puppet is made to appear as if it is speaking by a nearby performer [1]. If the performer were standing across the room from the puppet, it would be clear that it was actually the performer speaking. This corresponds with causal ratings in the ±90°conditions, where the sound appeared to have little to no impact. Naturally, as the sound became closer to the ground-truth position, ratings correspondingly increased. It would be interesting for future work to further explore

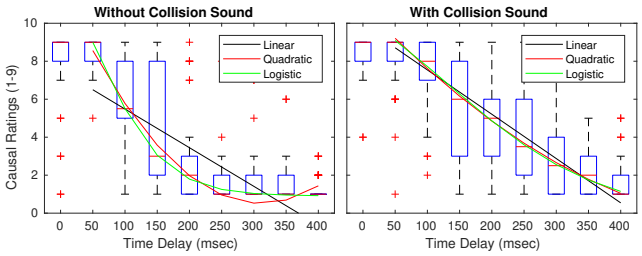

Figure 7: Box plot of causal ratings in the third block of Experiment 3 in the (left) absence and (right) presence of a collision sound. The participants were explicitly prompted to consider the auditory collisions indicator *before* giving the causal ratings. Red horizontal lines indicate median causal ratings, and the bottom and top edges of the blue boxes indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points that were not considered outliers, and red '+' symbols indicate outliers. The x-axis shows the temporal delay between the moment of impact and the start of the blue ball's motion. The black, red and green curves are linear, quadratic and logistic regression plots, respectively.

the ±30°region around the impact location, as this is where visual capture effects appear most salient.

One subtle result from Experiment 3 was that the relationship between causal ratings and temporal delay appeared to be different depending on where in the visual field the auditory collision indicator came from: *i.e.*, sounds from the right periphery of the visual field corresponded to causal decreases that were relatively linear compared with their paired sound locations in the left periphery. This could suggest that attentional resources are allocated asymmetrically across the visual field, perhaps with more attention allocated to where the objects are headed, and less attention allocated to where the objects have already been. However, previous work has shown that the ventriloquist effect does not depend on the direction of deliberate visual attention [7], suggesting that the current result might be due to the HMD orientation at the moment of impact being biased away from the impact location. Further work should examine whether the same behavioral trends occur when said orientation is controlled for.

Taken together, the present results demonstrate the viability of VR technology in studying human perception and cognition. While the historically high cost and inaccessibility of VR systems have inhibited their use in cognitive science in the past, the relatively low cost and intuitive interfaces of modern systems provide an effective means to construct unique and extraordinary testing environments for a breadth of human cognitive studies. However, it remains unclear whether certain cognitive tasks are better suited for VR environments than others: *e.g.*, the absence of haptic feedback in tasks involving interaction with virtual objects biases inferred physical attributes [55]. Thus, future work should further examine the strengths and weaknesses of VR implementations in human cognitive studies.

It is important to note that the experiments conducted herein presented dynamic events which were passively perceived by participants. One potential direction for future work is to determine whether user input to the VR environment (*e.g.*, setting the red ball into motion using a controller) influences perceived causality. In this situation, the speed of the initially moving object would be set by each participant and could be saved and presented later in passive-perception trials. Another potentially interesting manipulation would be to test whether varying the pitch of the collision sound as a function of the object's speed has an effect on perceived causality. This could provide further insight on the depth and sophistication of the auditory information utilized by the perceptual system when forming causal impressions.

Acknowledgments

The authors wish to thank Hanlin Zhu, Shu Wang, and Feng Gao for assisting the experiments at UCLA. The work reported herein was supported by DARPA XAI grant N66001-17-2-4029, ONR MURI grant N00014-16-1-2007, NSF grant BCS-1353391, and a NSF Graduate Research Fellowship.

REFERENCES

[1] D. Alais and D. Burr. The ventriloquist effect results from near-optimal bimodal integration. *Current Biology*, 14(3):257–262, 2004.

[2] P. B. Andreatta, E. Maslowski, S. Petty, W. Shim, M. Marsh, T. Hall, S. Stern, and J. Frankel. Virtual reality triage training provides a viable solution for disaster-preparedness. *Academic emergency medicine*, 17(8):870–876, 2010.

[3] C. Aravena, M. Vo, T. Gao, T. Shiratori, and L.-F. Yu. Perception meets examination: Studying deceptive behaviors in vr. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*, 2017.

[4] M. Azmandian, M. Hancock, H. Benko, E. Ofek, and A. D. Wilson. Haptic retargeting: Dynamic repurposing of passive haptics for enhanced virtual reality experiences. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1968–1979, 2016.

[5] P. Backlund, H. Engstrom, C. Hammar, M. Johannesson, and M. Lebram. Sidh-a game based firefighter training simulation. In *Information Visualization*, pp. 899–907, 2007.

[6] P. Backlund, H. Engstrom, M. Johannesson, and M. Lebram. Games and traffic safety-an experimental study in a game-based simulation environment. In *Information Visualization*, pp. 908–916, 2007.

[7] P. Bertelson, J. Vroomen, B. de Gelder, and J. Driver. The ventriloquist effect does not depend on the direction of deliberate visual attention. *Perception and Psychophysics*, 62(2):321–332, 2000.

[8] D. G. Boyle. A contribution to the study of phenomenal causation. *Quarterly Journal of Experimental Psychology*, 12(3):171–179, 1960.

[9] G. Bruder and F. Steinicke. Threefolded motion perception during immersive walkthroughs. In *Proceedings of the 20th ACM symposium on virtual reality software and technology*, pp. 177–185, 2014.

[10] M. Cha, S. Han, J. Lee, and B. Choi. A virtual reality based fire training simulator integrated with fire dynamics data. *Fire Safety Journal*, 50:12–24, 2012.

[11] L. Chittaro and F. Buttussi. Assessing knowledge retention of an immersive serious game vs. a traditional education method in aviation safety. *TVCG*, 21(4):529–538, 2015.

[12] L. B. Cohen and L. M. Oakes. How infants perceive a simple causal event. *Developmental Psychology*, 29(3):421–433, 1993.

[13] N. F. Dixon and L. Spitz. The detection of auditory visual desynchrony. *Perception*, 9(6):719–721, 1980.

[14] J. Fischer, J. G. Mikhael, J. B. Tenenbaum, and N. Kanwisher. Functional neuroanatomy of intuitive physical inference. *Proceedings of the National Academy of Sciences (PNAS)*, 113(34):E5072–E5081, 2016.

[15] J. A. Fugelsang, M. E. Roser, P. M. Corballis, M. S. Gazzaniga, and K. N. Dunbar. Brain mechanisms underlying perceptual causality. *Cognitive Brain Research*, 24(1):41–47, 2005.

[16] J. J. Gibson. *The senses considered as perceptual systems*. Houghton Mifflin, Oxford, England, 1966.

[17] T. L. Griffiths and J. B. Tenenbaum. Theory-based causal induction. *Psychological Review*, 116(4):661–716, 2009.

[18] R. Guski and N. F. Troje. Audiovisual phenomenal causality. *Attention, Perception, & Psychophysics*, 65(5):789–800, 2003.

[19] K. Hertkorn, M. A. Roa, M. Brucker, P. Kremer, and C. Borst. Virtual reality support for teleoperation using online grasp planning. In *IROS*, pp. 2074–2074, 2013.

[20] D. Hume. *A treatise of human nature*. Oxford University Press, Oxford, England, 1738/1978.

[21] A. Kohlrausch and S. van de Par. Experimente zur wahrnehmbarkeit von asynchronie in audio-visuellen stimuli. *Fortschritte der Akustik*, 26:316–317, 2000.

[22] M. D. Kozlov and M. K. Johansen. Real behavior in virtual environments: Psychology experiments in a simple virtual-reality paradigm using video games. *Cyberpsychology, behavior, and social networking*, 13(6):711–714, 2010.

[23] J. R. Kubricht, K. J. Holyoak, and H. Lu. Intuitive physics: Current research and controversies. *Trends in cognitive sciences*, 21(10):749–759, 2017.

[24] A. M. Leslie and S. Keeble. Do six-month-old infants perceive causality? *Cognition*, 25(3):265–288, 1987.

[25] J. Lewald, W. H. Ehrenstein, and R. Guski. Spatio-temporal constraints for auditory–visual integration. *Behavioural brain research*, 121(1):69–79, 2001.

[26] D. J. Lewkowicz. Perception of auditory–visual temporal synchrony in human infants. *Journal of Experimental Psychology: Human Perception and Performance*, 22(5):1094, 1996.

[27] C. Li, W. Liang, C. Quigley, Y. Zhao, and L.-F. Yu. Earthquake safety training through virtual drills. *TVCG*, 23(4):1275–1284, 2017.

[28] J. Lin, X. Guo, J. Shao, C. Jiang, Y. Zhu, and S.-C. Zhu. A virtual reality platform for dynamic human-scene interaction. In *SIGGRAPH ASIA 2016 Virtual Reality meets Physical Reality: Modelling and Simulating Virtual Humans and Environments*, p. 11, 2016.

[29] J. Lin, Y. Zhu, J. Kubricht, S. Zhu, and H. Lu. Visuomotor adaptation and sensory recalibration in reversed hand movement task. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*, pp. 2579–2584, 2017.

[30] R. Mayrhofer and M. R. Waldmann. Indicators of causal agency in physical interactions: The role of the prior context. *Cognition*, 132(3):485–490, 2014.

[31] J. McComas, M. MacKay, and J. Pivik. Effectiveness of virtual reality for teaching pedestrian safety. *CyberPsychology & Behavior*, 5(3):185–190, 2002.

[32] A. Michotte. *The perception of causality*. Basic Books, New York, NY, 1963.

[33] A. C. A. Mól, C. A. F. Jorge, and P. M. Couto. Using a game engine for vr simulations in evacuation planning. *Computer Graphics and Applications*, 28(3):6–12, 2008.

[34] T. Natsoulas. Principles of momentum and kinetic energy in the perception of causality. *The American Journal of Psychology*, 74(3):394–402, 1961.

[35] A.-H. Olivier, J. Bruneau, G. Cirio, and J. Pettré. A virtual reality platform to study crowd behaviors. *Transportation Research Procedia*, 2:114–122, 2014.

[36] L. S. Padgett, D. Strickland, and C. D. Coles. Case study: using a virtual reality computer game to teach fire safety skills to children diagnosed with fetal alcohol syndrome. *Journal of Pediatric Psychology*, 31(1):65–70, 2005.

[37] M. Reznek, P. Harter, and T. Krummel. Virtual reality and simulation: training the future emergency physician. *Academic Emergency Medicine*, 9(1):78–87, 2002.

[38] L. J. Rips. Causation from perception. *Perspectives on Psychological Science*, 6(1):77–97, 2011.

[39] M. E. Roser, J. A. Fugelsang, K. N. Dunbar, P. M. Corballis, and M. S. Gazzaniga. Dissociating processes supporting causal perception and causal inference in the brain. *Neuropsychology*, 19(5):591–602, 2005.

[40] A. Rovira, D. Swapp, B. Spanlang, and M. Slater. The use of virtual reality in the study of people's responses to violent incidents. *Frontiers in Behavioral Neuroscience*, 3, 2009.

[41] I. Rudloff. Untersuchungen zur wahrgenommenen synchronität von bild und ton bei film und fernsehen. *Master's thesis, Ruhr-Universität Bochum, Germany*, 1997.

[42] A. N. Sanborn, V. K. Mansinghka, and T. L. Griffiths. Reconciling intuitive physics and newtonian mechanics for colliding objects. *Psychological Review*, 120(2):411–437, 2013.

[43] A. Schlottmann and D. R. Shanks. Evidence for a distinction between judged and perceived causality. *The Quarterly Journal of Experimental Psychology*, 44(2):321–342, 1992.

[44] B. J. Scholl and K. Nakayama. Causal capture: Contextual effects on the perception of collision events. *Psychological Science*, 13(6):493–498, 2002.

[45] B. J. Scholl and P. D. Tremoulet. Perceptual causality and animacy. *Trends in Cognitive Sciences*, 4(8):299–309, 2000.

[46] R. Sekuler. Sound alters visual motion perception. *Nature*, 385:308–308, 1997.

[47] S. Shah, D. Dey, C. Lovett, and A. Kapoor. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and Service Robotics*, 2017.

[48] S. Stansfield, D. Shawver, A. Sobel, M. Prasad, and L. Tapia. Design and implementation of a virtual reality system and its application to training medical first responders. *Presence: Teleoperators and Virtual Environments*, 9(6):524–556, 2000.

[49] I. Tarnanas and G. C. Manos. Using virtual reality to teach special populations how to cope in crisis: the case of a virtual earthquake. *Studies in health technology and informatics*, 81:495–501, 2001.

[50] D. L. Tate, L. Sibert, and T. King. Virtual environments for shipboard firefighting training. In *Virtual Reality Annual International Symposium*, pp. 61–68, 1997.

[51] S. Van de Par and A. Kohlrausch. Sensitivity to auditory-visual asynchrony and to jitter in auditory-visual timing. In *Human vision and electronic imaging*, pp. 234–242, 2000.

[52] P. A. White. Perception of forces exerted by objects in collision events. *Psychological review*, 116(3):580–601, 2009.

[53] M. Xi and S. P. Smith. Simulating cooperative fire evacuation training in a virtual environment using gaming technology. In *Virtual Reality*, pp. 139–140, 2014.

[54] X. Yan, M. Khansari, Y. Bai, J. Hsu, A. Pathak, A. Gupta, J. Davidson, and H. Lee. Learning grasping interaction with geometry-aware 3d representations. *arXiv preprint arXiv:1708.07303*, 2017.

[55] T. Ye, S. Qi, J. Kubricht, Y. Zhu, H. Lu, and S.-C. Zhu. The martian: Examining human physical judgments across virtual gravity fields. *TVCG*, 23(4):1399–1408, 2017.

[56] Y. Zhu, D. Gordon, E. Kolve, D. Fox, L. Fei-Fei, A. Gupta, R. Mottaghi, and A. Farhadi. Visual semantic planning using deep successor representations. In *ICCV*, 2017.

[57] Y. Zhu, R. Mottaghi, E. Kolve, J. J. Lim, A. Gupta, L. Fei-Fei, and A. Farhadi. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *ICRA*, pp. 3357–3364, 2017.