Accelerating Saccadic Response through Spatial and Temporal Cross-Modal Misalignments

Daniel Jiménez-Navarro
djimenez@mpi-inf.mpg.de
Max Planck Institute for Informatics
Saarbrücken, Germany

Xi Peng
xp2011@nyu.edu
New York University
Brooklyn, USA

Yunxiang Zhang
yunxiang.zhang@nyu.edu
New York University
Brooklyn, USA

Karol Myszkowski
karol@mpi-inf.mpg.de
Max Planck Institute for Informatics
Saarbrücken, Germany

Hans-Peter Seidel
hpseidel@mpi-inf.mpg.de
Max Planck Institute for Informatics
Saarbrücken, Germany

Qi Sun
qisun@nyu.edu
New York University
Brooklyn, USA

Ana Serrano
anase@unizar.es
Universidad de Zaragoza
Zaragoza, Spain

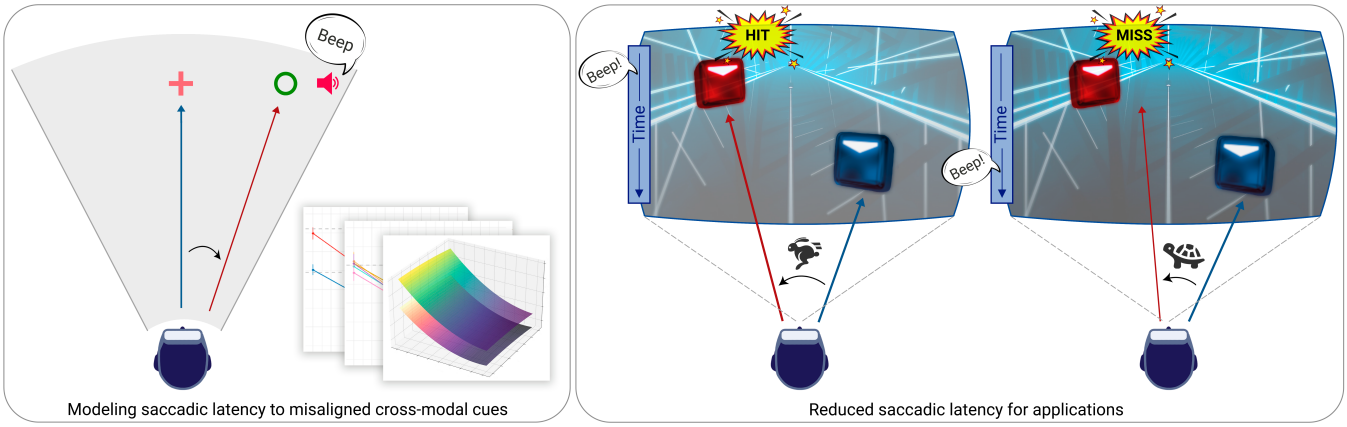

Figure 1: *Left*: Simplified diagram of our experiment exploring how the *spatial* and *temporal* relationships between visual and auditory stimuli affect saccadic latency. The results derived from this experiment provide insights into how these factors modulate saccadic latency. *Right*: Illustration of a practical application of our findings in a game similar to *Beat Saber*. In the depicted scenario, user's ability to react to a target is influenced by the timing of auditory cues. When the sound precedes the visual stimulus, following our insights, the user experiences reduced saccadic latency, enabling them to hit the target on time. Conversely, if the sound and visual stimulus are synchronized, the user's reaction time is slower, resulting in missing the target.

ABSTRACT

Human senses and perception are our mechanisms to explore the external world. In this context, visual saccades –rapid and coordinated eye movements– serve as a primary tool for awareness of our surroundings. Typically, our perception is not limited to visual stimuli alone but is enriched by cross-modal interactions, such as the combination of sight and hearing. In this work, we investigate the *temporal* and *spatial* relationship of these interactions, focusing on how auditory cues that precede visual stimuli influence saccadic latency –the time that it takes for the eyes to react and start moving towards a visual target. Our research, conducted within a virtual reality environment, reveals that auditory cues preceding visual information can significantly accelerate saccadic responses, but this effect plateaus beyond certain temporal thresholds. Additionally, while the spatial positioning of visual stimuli influences the speed of these eye movements, as reported in previous research, we find that the location of auditory cues with respect to their corresponding visual stimulus does not have a comparable effect. To validate our findings, we implement two practical applications: first, a basketball training task set in a more realistic environment

SIGGRAPH Conference Papers '24, July 27–August 01, 2024, Denver, CO, USA

ACM ISBN 979-8-4007-0525-0/24/07
https://doi.org/10.1145/3641519.3657432

with complex audiovisual signals, and second, an interactive farm game that explores previously untested values of our key factors. Lastly, we discuss various potential applications where our model could be beneficial.

CCS CONCEPTS

• **Computing methodologies** → **Perception**; **Virtual reality**; **Mixed / augmented reality**.

KEYWORDS

Audiovisual integration, cross-modal interactions, multisensory perception, saccadic latency, virtual reality

ACM Reference Format:
Daniel Jiménez-Navarro, Xi Peng, Yunxiang Zhang, Karol Myszkowski, Hans-Peter Seidel, Qi Sun, and Ana Serrano. 2024. Accelerating Saccadic Response through Spatial and Temporal Cross-Modal Misalignments. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers '24 (SIGGRAPH Conference Papers '24), July 27–August 01, 2024, Denver, CO, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3641519.3657432

1 INTRODUCTION

Human beings continuously interact with the external world via their senses, gathering and integrating information to construct their perception. This cross-modal integration, particularly between sight and hearing, profoundly influences our perception of the world. However, the relationship between these modalities is not always straightforward. For instance, some studies [Hidaka and Ide 2015; Malpica et al. 2022] have shown that in certain scenarios, auditory stimuli can unexpectedly suppress visual perception, demonstrating the complexity and sometimes counterintuitive nature of these interactions. In this work, we focus on the *spatial* and *temporal* aspects of the relationship between visual and auditory stimuli. In typical real-world situations, visual and auditory cues occur almost simultaneously, or sound is slightly delayed due to differences in the speed of light and sound travel. Existing literature often reflects these scenarios accordingly [Arnold et al. 2005; Vroomen and Keetels 2010]. In contrast, we benefit from the flexibility of immersive digital environments and virtual reality (VR) setups to investigate a variety of temporal and spatial shifts between visual and auditory stimuli, with an emphasis on sound preceding visual information. This focus is motivated by evidence showing that auditory cues preceding visual stimuli can significantly reduce the latency of saccadic eye movements [Ross and Ross 1981; Vidal and Vitu 2022]. Saccades are rapid, simultaneous movements of both eyes in the same direction, which are crucial for directing gaze and attention, impacting reaction times in diverse situations, including gaming and e-sports [Koposov et al. 2020]. A saccade, once triggered, follows a pre-programmed, close-to-ballistic motion whose velocity and duration cannot be easily changed [Arabadzhiyska et al. 2017; Bahill 1975]. Therefore, we focus on the latency of the preparatory stage when the oculomotor mechanisms are preparing for the saccade launch, which is guided both by visual and auditory cues [Kowler 2011]. In particular, we investigate for the first time how retinal eccentricities and spatio-temporal shifts between visual and auditory stimuli can influence saccadic latency in a virtual environment (Sec. 3). We conduct our experiments systematically, considering those dimensions in the ranges relevant to VR applications. Our results show that temporal shifts in auditory cues can significantly accelerate saccadic responses, but this acceleration reaches a plateau beyond certain temporal intervals. Moreover, while eccentricity consistently influences saccadic latency, as observed in previous work [Duinkharjav et al. 2022], spatial shifts in the auditory cues do not seem to significantly impact this latency.

These findings not only advance our understanding of cross-modal interactions but also have practical implications across various domains (Fig. 1). Reducing saccadic latency through finely tuned audiovisual timings is useful in a variety of applications. For user interfaces, reduced saccadic latency can lead to faster navigation and data retrieval, streamlining workflows. In augmented reality (AR) environments, faster saccadic responses can enhance real-virtual interactions, making information assimilation more intuitive. Within videogames, especially those relying on quick reflexes, our insights can inform the design of cues (visual-auditory timings) that improve players' response to in-game events, leading to a more engaging and competitive experience [Kim et al. 2019; Koposov et al. 2020]. Digital multimedia, which often relies on carefully sequenced audiovisual information to achieve specific effects, such as the perception of simultaneity or attention-redirecting effects [Ogawa et al. 2023; Serafin et al. 2018], also stands to benefit from these insights. In online multimedia, latencies and delays are common issues that can deteriorate the user experience, sometimes to the point of discomfort [Hopkins et al. 2022]. By applying our understanding of saccadic latency, these challenges can be mitigated, enhancing the overall experience for users.

In our work, we validate these concepts through two proof-of-concept applications. The first, a basketball training task set in a more realistic environment, demonstrates how saccades can be effectively accelerated, extending our findings beyond simple experimental setups (Sec. 4.1). The second, an interactive farm game, explores the effects of previously untested values of our main factors, further illustrating the versatility of our model (Sec. 4.2).

Our collected anonymized data, model, and code are publicly available at https://avsaccadeaccel.mpi-inf.mpg.de/

2 RELATED WORK

In this section we discuss human visual saccades and the factors that affect them, focusing on their importance in graphics applications. Then, we introduce audiovisual spatio-temporal interactions and discuss their connection with the behavior of visual saccades.

2.1 Visual Saccades

Saccades, the fastest eye movements, align the fovea with the region of interest for optimal visual acuity. Understanding the dynamics of visual saccades has been beneficial for many graphics applications, informing key areas such as foveated rendering [Albert et al. 2017; Franke et al. 2021], optimization of 3D VR streamings [Chen et al. 2022], or the development of strategies to stimulate the human visual system beyond real-world capabilities [Dunn et al. 2020].

The relationship between saccade amplitude, peak velocity, and duration, typically ranging between 20-80 ms, is well-understood [Anliker 1976; Arabadzhiyska et al. 2017; Bahill 1975]. However, saccade

latency, the time from the appearance of a target stimulus to the execution of the saccade, varies more significantly and is longer, between 200-400 ms [Duinkharjav et al. 2022]. During this time, a number of tasks must be performed, such as releasing attention from the previous fixation point, sensory registration of a new target, decision-making on active attending, and programming of oculomotor mechanisms for saccade launching [Zambarbieri et al. 1995].

Saccade latency can be influenced by various visual features of the target, such as saliency [Yamagishi and Furukawa 2020], contrast [Carpenter 2004], intensity [Bell et al. 2006], and spatial frequency [Duinkharjav et al. 2022]. This means that saccade latency can be effectively reduced by enhancing the contrast of potential targets (this effect saturates for strongly suprathreshold contrasts), increasing their intensity, or assuring that spatial patterns include frequencies over 4 cycles-per-degree (while higher frequencies have not been measured following the contrast sensitivity function [Barten 1999], one can assume a latency drop over 10–20 cpd [Duinkharjav et al. 2022]). Additionally, saccade target direction and eccentricity play roles in latency variation, with horizontal saccades typically being faster than vertical ones [Dafoe et al. 2007], and different effects observed based on the target positioning in the visual field. Lower saccade latency can be expected for the target in the upper hemifield compared to the lower hemifield due to more visual processing in the latter case, as required for navigation in the environment [Dafoe et al. 2007]. This pattern is complemented by findings that accuracy improvements in saccades are more pronounced in elevation than azimuth, especially in audiovisual contexts [Corneil et al. 2002]. Such improvements are greater when the stimuli include auditory information, highlighting the significant role of sound in enhancing saccadic accuracy. The impact of target eccentricity on saccade latency remains ambiguous. Dafoe et al. [2007] observed no significant changes in latency related to eccentricity, while Zambarbieri et al. [1995] reported an increase in latency as eccentricity rose. Saccades, traditionally studied from a purely visual perspective, have been shown to suppress visual sensitivity [Ross et al. 2001], with visual space compression occurring just before the onset [Ross et al. 1997]. These effects relate to neurophysiological anticipatory procedures and decision processes in saccade generation.

The gap-overlap paradigm, illustrated in Fig. 2, is a key concept in understanding visual engagement and attention [Fischer et al. 1997; Pierrot-Deseilligny et al. 1995]. It has been widely used to measure changes in saccade latency, relating these to target luminance and anti-saccadic tasks (looking away from the target) [Kingstone and Klein 1993; Reuter-Lorenz et al. 1991]. While this paradigm has been extensively used in visual studies, it has also been applied in purely auditory studies to detect disorders such as tinnitus [Boyen et al. 2015; Fournier and Hebert 2012]. However, its potential relationship with auditory cues in cross-modal conditions has been less explored.

2.2 Audiovisual Interactions

The perceptual interaction between visual and auditory signals is a topic of extensive research, with findings indicating both facilitatory and inhibitory effects. The exact nature of this audiovisual interaction is complex, with some studies showing that auditory signals can even cause visual suppression under certain conditions [Hidaka and Ide 2015; Malpica et al. 2022]. Conversely, facilitatory effects have been also found for this audiovisual integration. For instance, reaction times to bimodal (audiovisual) stimuli are often faster than to unimodal stimuli [Teder-Sälejärvi et al. 2005], and phenomena such as the McGurk effect (blending between non-cosistent audio and visual cues in speech) demonstrate the powerful illusionary potential of audiovisual integration. Multisensory interaction is the physiological mechanism through which two different sensory stimuli can affect each other and overall perception. Focusing on the audiovisual modality, different factors have been found to affect this multisensory binding [Spence and Driver 2004], including the temporal window [Chen and Spence 2017; Vatakis and Spence 2007], and the meaning and content of stimuli for semantic congruency perception [Doehrmann and Naumer 2008].

Regarding visual saccades, the limits on the advantages of including the auditory modality remain unclear. Audiovisual integration has been observed to be beneficial in reducing visual saccade latencies [Diederich et al. 2012; Vidal et al. 2020; Wang et al. 2017] and improving overall performance. Thus, auditory cues have been shown to alter how saccades are performed [Zou et al. 2012]. Research has shown that auditory cues aligned with visual signals in time and space can reduce saccadic latency. Frens et al. [1995] observed that, while this latency reduction diminishes with spatial misalignments, the increase in latency is relatively minor, maintaining stability even with misalignments of approximately 20°. They also observed that when auditory cues lag behind visual signals, there is an increase in saccadic latency. Nevertheless, it has been shown that behavioral responses to bimodal stimuli generally outperform those to unimodal stimuli, both in speed and accuracy, regardless of the audiovisual pairing locations [Corneil et al. 2002; Teder-Sälejärvi et al. 2005; van Wanrooij et al. 2009].

Closest to our work is the work of Vidal and Vitu [2022] that modulates visual latencies by manipulating the audiovisual delay (temporal shift) and the temporal interval between fixation offset and target onset (gap-overlap paradigm). As highlighted in their findings, combining these methods is crucial for achieving maximum modulation of saccades. Our research takes this a step further by providing a comprehensive analysis of visual saccade latency, including both spatial and temporal shifts between audiovisual cues. Our contributions extend beyond this analysis, as we also validate our findings in more complex scenarios, thus bringing our insights closer to practical applications. Additionally, we conduct our tests in VR environments, incorporating stereo and spatial sound and offering a more immersive and realistic context for understanding the facilitative role of auditory information in saccadic movements.

3 MEASURING SACCADIC LATENCY UNDER CROSS-MODAL MISALIGNMENTS

In this section, we describe our main experiment, which focuses on measuring saccadic latency under various spatio-temporal misalignments of audiovisual cues. We describe our experimental setup and methodology in detail (summarized in Fig. 3), followed by an in-depth analysis of the results.

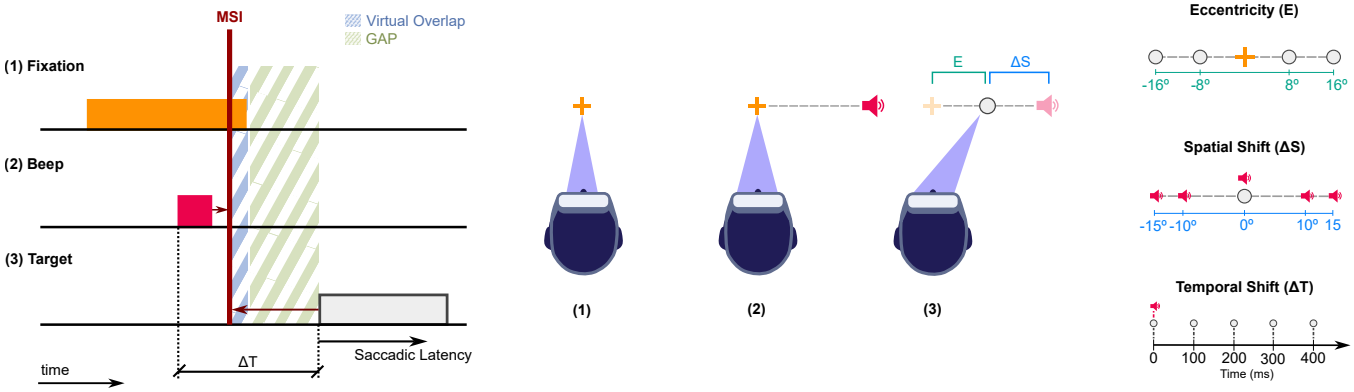

Figure 3: *Left*: Temporal Dynamics in the Experiment: The Multisensory Integration (MSI) process, influenced by the temporal shift ΔT, alters the perception of the target onset to be closer to the beep onset, thereby effectively integrating these two elements into a single event. This creates a perceived virtual overlap of the fixation and the visual target, which is effectively compensated with a real gap as inspired by by Vidal et al. [2022]. *Center*: User Study. Participants initially fixate on the central cross (1). When they confirm readiness, the experiment starts. After a random time to avoid learning effects, the beep sounds (2), followed by the visual target apparition (3) after the ΔT interval. *Right*: Graphical representation of the stimuli cases with different levels of eccentricity (E), spatial shift (ΔS), and temporal shift (ΔT).

3.1 Experimental Setup

Hardware. We used a Varjo Aero HMD and Beyerdynamic DT 770M headphones with passive noise cancelation. This HMD is equipped with an eye-tracker running at 200 Hz and with an accuracy of 1° of visual angle. Additional hardware details, including a discussion about the main sources of measurement error, can be found in Sec. S3 of the supplementary. The experiment was developed and conducted using the Unity game engine, and the Oculus Native Spatializer was used for managing audio spatial propagation. Directional audio was reproduced using a generic head related transfer function (HRTF). Existing research supports the notion that generic HRTFs provide satisfactory azimuth localization accuracy within our spatial shift range, compared to individualized HRTFs [Berger et al. 2018; Rummukainen et al. 2019; Wenzel et al. 1993]. Therefore, we believe these minor variations in spatial perception are unlikely to impact our results, while also simplifying the experimental procedure. Participants remained seated during the experiment while being able to move the head freely.

Stimuli. The target was a white circle of 0.36° placed at 6 meters from the participant, with a lifetime of 2 s [Vidal and Vitu 2022]. The Weber contrast was set to 0.3, above detection thresholds, such that it would not affect saccade latency [Duinkharjav et al. 2022]. The fixation cross was of size 0.12°, located in the middle of the screen [Vidal and Vitu 2022]. These are illustrated in Fig. 4. The auditory cues were beep-like sounds, centered at 880 Hz and 60 dB [Zambarbieri 2002; Zambarbieri et al. 1995] that stop after 150 ms. We employ the *beep and flash* scenario, a simple experimental setup involving auditory beeps and visual flashes, for its ease of control and limited cognitive interference. This scenario has been shown to tap into fundamental neurophysiological processes related to temporal resolution and integration in the brain, which are relevant across various types of more complex tasks [Shams et al. 2000]. Additionally, this scenario is particularly relevant in practical applications such as gaming environments where quick perception of simple, synchronized audiovisual cues is crucial. A more detailed discussion on this topic can be found in Sec. S1 in the supplementary.

Conditions. The variables considered in this study were visual *eccentricity* (E), *temporal shift* (ΔT), and *spatial shift* (ΔS) between the visual target and its corresponding auditory cue. For eccentricity, we tested 8 and 16 visual degrees in both left (-) and right (+) peripheries (E=[-16°, -8°, 8°, 16°]), aligning with common saccade displacements [Bahill 1975]. The spatial shifts were ΔS=[-15°, -10°, 0° (co-located), 10°, 15°] visual degrees, considering the simultaneity fusion window limits based on eccentricity levels [Godfroy et al. 2003]. For temporal shifts, we considered ΔT=[0 (no delay), 100, 200, 300, 400] ms, accounting for the temporal binding window limits in literature [Spence 2011], with audio cues always generated before the visual target onset. We selected these values to achieve effective audiovisual integration, taking relevant literature into account; a more detailed discussion is available in Sec. S1 in the supplementary. These factors are depicted in Fig. 3 (right). We considered a fully factorial design resulting in 4 (E) × 5 (ΔS) × 5 (ΔT) = 100 conditions. We additionally included four baseline cases without auditory cues (one for each eccentricity). Each participant was presented with each condition four times, totaling 416 trials.

Participants. The study was conducted with 12 participants (6 male, 6 female, none non-binary; ages 22-26). All reported normal or corrected-to-normal vision and hearing. All provided written consent, and the study protocol was approved by the Ethical Review Board of the Department of Computer Science at Saarland University. From the participant pool, 58.3% reported having no experience with VR before, while 41.7% reported having used VR equipment 5 times or less. All survey questions asked to the participants can be found in Sec. S5 in the supplementary.

Procedure. The experiment was conducted in four sessions of ten minutes each, to prevent fatigue. The order of trials was randomized across sessions. For each session, the eye-tracker was calibrated following a five-dot calibration. In each trial, participants started by looking at the fixation cross located in the middle of the screen (starting point). Afterward, participants pressed the *space* on the keyboard to indicate that they were ready. After some random time (between 750 and 1750 ms), used to avoid learning and prediction effects, the fixation cross disappeared and audiovisual cues were generated according to the corresponding E, ΔS and ΔT in the trial. Participants were asked to look at the visual target (make a saccade towards it) as soon as they perceived it. After some time (2 s), the visual target disappeared and the fixation cross spawned again, starting a new trial. This procedure is represented in Fig. 3.

3.2 Results and Discussion

Data processing and analysis. We used a two-step velocity threshold method to accurately identify the start (onset) and end (offset) of the saccades, accounting for eye-tracking noise [Arabadzhiyska et al. 2017]. Conservative thresholds of $120°/s$ and $60°/s$ were set to detect the saccade and its onset. We compute the *saccadic latency* as the difference between the target stimulus onset and the saccade onset. Following previous literature [Frens et al. 1995; Ross and Ross 1981; Vidal and Vitu 2022], we filter out saccade latencies lower than 100 ms (prediction cases), and higher than 500 ms (distraction cases). For computing the *landing error* of the saccade, we compute the distance in degrees of visual angle between the landing point (related to the saccade offset) and the location of the target [Arabadzhiyska et al. 2017]. Given the eye-tracker's accuracy of 1°, the target size of 0.36°, and typical saccadic undershoots ranging from 10% to 20% [Kapoula and Robinson 1986; Lisi et al. 2019], deviations up to 4.2° for the largest saccades are expected. Consequently, we discard saccades with a landing error exceeding 5° to filter out inaccuracies, such as those from distracted participants. For further details please see Sec. S2 in the supplementary.

We combine positive and negative eccentricities (merging -8 with 8 and -16 with 16) after verifying their similarity and in line with previous studies that consider symmetry [Charbonneau et al. 2013; Duinkharjav et al. 2022], resulting in two eccentricity levels for the analysis: 8 and 16. We employ a generalized linear mixed model (GLMM), since it provides a robust and flexible approach for analyzing data when random effects are present [Bolker et al. 2009]. Our model includes three fixed effects: *temporal shift* (ΔT), *eccentricity* (E), and *spatial shift* (ΔS), as well as their interactions. We consider the participant as a random effect. The dependent variables are the *saccadic latency* and the *landing error*. We establish the significance level at $\alpha = 0.05$. After applying the exclusion criteria outlined above, each unique combination of main factors appears 32 to 48 times (4 trials x 12 users) in our analysis, with trial count variations effectively managed by the GLMM. The main results of the analysis are discussed in the following and depicted in Fig. 5. All the statistical tests and post-hocs can be consulted in Sec. S4.2 in the supplementary.

Temporal shift (ΔT). The effect of ΔT is statistically significant ($p < 0.001$), indicating that varying ΔT levels effectively modulate saccade latency. When ΔT increases, saccade latency decreases. Post-hoc analysis shows no significant difference between ΔT intervals 200-300 ms ($p = 0.09$) and 300-400 ms ($p = 0.44$). This suggests a trend towards stabilization and a potential upper limit in saccade acceleration at higher ΔT levels. This finding overturns the assumption that earlier auditory cues always lead to proportionally faster saccades, suggesting a more complex relationship where the efficiency of auditory cues plateaus beyond a certain point. These results are in line with conclusions reported in the work of Vidal and Vitu [2022]. They reported a decrease in latency for ΔT = 0 compared to the no-sound case, and latencies were found to be shorter for beeps preceding the target onset ΔT = 60, 120 ms with gap compensation, showing how the modulation effect was larger with gap compensation compared to not compensated cases, reducing latency even more. Frens et al. [1995] (experiment 3) also reported how ΔT can affect the saccadic latency distribution. Using different ΔT cases (-100, -50, 0, 50), the lowest latency was found when the auditory preceded the visual ΔT=50 instead of when they were synchronous, justifying this on different speeds generating different arrival moments in receptor organs. Different onset and offset conditions for visual or auditory warning signals were tested by Ross et al. [1981] (experiment 1), changing also ΔT to trigger saccades. For the auditory warning, they found how both onset and offset warnings could facilitate saccade latency at ΔT = 100 and 300 ms but not when the audio onset was after the target onset.

Eccentricity (E). This factor significantly influences latency ($p < 0.001$), with larger eccentricities showing increased latency. Interestingly, our results reveal that this effect appears to be uniform across all ΔT levels, reinforcing and complementing previous studies. This suggests that the spatial positioning of visual stimuli plays a crucial role in saccadic responses, regardless of auditory cue timing, emphasizing the importance of visual cue placement in multisensory integration. Our results align with prior research. Duinkharjav et al. [2022] report a similar trend in their preliminary study, where they varied eccentricity (with frequency and contrast) to compare saccade latency durations towards a gabor patch stimuli. They observed that the latency-eccentricity relationship exhibited a U-shaped curve, with latency being higher at both the foveal (0°) and peripheral (20°) regions, and lower at a mid-peripheral point (10°). The same bowl shape was found for the latency-eccentricity function by Kalesnykas et al. [1994]. They measured 38 different retinal eccentricities and observed how color and intensity have small contributions to the central peak (5-15 ms). Their central robust minimum latency peak ranged from 0.75° to 12° (our 8°) and reported a height of 35-75 ms. After that, latency gradually increases towards the periphery (our 16°), becoming larger and more erratic for the largest periphery values (around 50°-60°).

Spatial shift (ΔS). Our data indicates no significant effect of this factor on saccadic latency ($p = 0.18$). However, it is important to interpret this with caution, as the absence of a statistically significant impact does not definitively rule out an effect. This finding suggests that if there is an influence of precise sound localization (co-location) on saccadic latency, it may not be consistent or robust across different scenarios. This could have significant implications for how auditory cues are used in various applications, allowing for more flexibility in sound design.

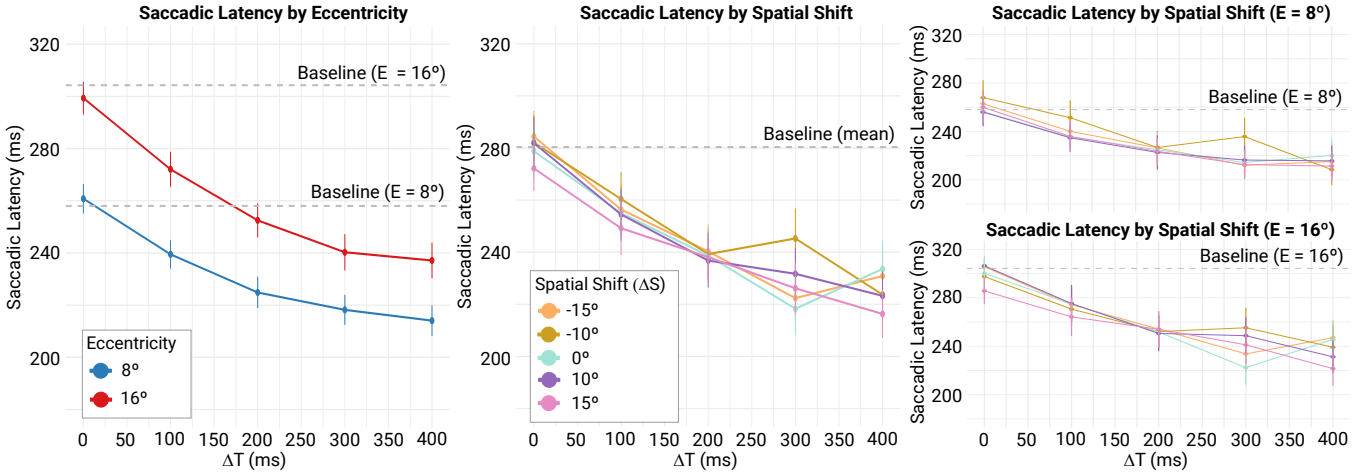

Figure 5: Saccadic latency in the main experiment for different temporal shifts depending on eccentricity (left), spatial shifts (center), and their interaction (right). Baseline cases (no sound) are represented with dashed lines. Error bars represent 95% CI.

Our findings align with those of Frens et al. [1995], who studied the effects of vertical, horizontal, and diagonal misalignments from the central fixation point. They found that spatially coincident visual and auditory stimuli reduced visual latency by approximately 50 ms compared to visual-only scenarios. However, this latency reduction diminished with increasing spatial shifts (completely vanishing at large misalignments of about 54°), and was very subtle for small shifts like those we considered. Similarly, Van Wanrooij et al.[2009] reported that spatially aligned auditory events produced shorter reaction times, but these effects lessened with greater spatial shifts, leading to a breakdown in multisensory integration and resulting in bistable saccades. Lastly, Colonius et al. [2001] reported how audio cues indeed help to reduce saccadic latencies towards visual targets up to large spatial shifts. Without temporal shift ($\Delta T = 0$), there were almost no latency differences for $\Delta S = +/-15$, while very small differences were reported again with $\Delta T = 30$. Our results, showing no significant impact of spatial shifts within our explored range on saccadic latency, are consistent with the subtle differences observed in other studies for comparable spatial misalignments. This complements our approach of restricting spatial shift values to ensure effective multisensory integration and maintain stimuli within the field of view, essential for audiovisual integration applications.

Interaction of ΔT and E. This is the only significant interaction ($p < 0.001$). Our results suggest that smaller eccentricities (E = 8°) reach a point of stabilization and maximum acceleration at an earlier ΔT compared to larger eccentricities (E = 16°). Specifically, for E = 8°, the differences in latency become smaller and statistically non-significant starting from an ΔT of 200 ms. For E = 16°, a similar pattern emerges only from a ΔT of 300 ms, as the interval between ΔT 200-300 ms still shows significant differences. This finding, particularly the stabilization points for different eccentricities, provides valuable insights for optimizing multisensory cues in environments where both spatial and temporal factors are critical.

Saccades accuracy. Landing error is shown in Fig. 6. The only significant factor is eccentricity ($p < 0.001$). However, this difference might be partly attributed to the limitations of the eye-tracker, which is less accurate at larger eccentricities [Stahl 2001]. Despite this, all the observed error consistently remained under 1°, which is roughly the accuracy of the eye-tracker. In the context of audiovisual saccades, previous research by Corneil et al. [2002] suggests that these are generally faster and more precise than saccades driven solely by visual or auditory stimuli. Our study did not record bistable unimodal responses, which are typically absent with small perceived ΔS where multisensory integration (MSI) occurs. Consequently, given the small ΔS levels employed in our experiment, we did not anticipate observing notable effects on saccade accuracy.

4 APPLICATIONS

In this section, we present two main applications of our study. First, we validate our findings through a realistic use case, linking experimental results to a practical scenario. Second, we introduce a more interactive validation using a game-like experience that tests previously unmeasured values of our main factors, demonstrating the model's broader applicability. Finally, we discuss other potential applications of our insights and model.

4.1 Practical Application: Basketball Game

Our findings regarding latency acceleration are validated in a scenario that features audiovisual stimuli with enhanced semantic content, closer to real applications than the *beep and flash* paradigm used in our main experiment.

Scene and stimuli conditions. The designed task simulates a basketball training application (Fig. 7). Participants were located in the middle of a basketball court, looking at one basket. At each side of the basket, two teammates were looking towards the participant. Behind them, four referees were placed in four different locations. As visual stimuli, basketballs of a size of 3° appeared in front of the teammates as they were holding them. For simplicity and to

avoid fatigue, a single eccentricity level was considered in both directions E=[-8°, 8°]. Auditory cues were a whistling noise of 60 dB, generated by the referees with two possible spatial shifts ΔS =[0° (co-located), 15°]. A red cross in the middle of the whiteboard in the basket pole was used as a starting fixation point. Audiovisual temporal shifts were again ΔT=[0 (no delay), 100, 200, 300, 400] ms (audio-leading). This resulted in 2 (E) × 2 (ΔS) × 5 (ΔT) = 20 conditions plus 2 baseline conditions with no sound. Each condition was experienced 10 times, resulting in a total number of 220 trials per user.

Participants and procedure. A total of 7 participants who did not participate in the main experiment carried out this study (6 male, 1 female, non non-binary; ages 24-29). The procedure was the same as the main experiment, with the main differences being the scene as well as the audiovisual cues. The task of the participants was to react as quickly as possible to the ball appearing and gaze towards it (perform a saccade).

Results. After merging eccentricities according to horizontal symmetry [Charbonneau et al. 2013], we obtain a total of 140 samples per condition. Following the same data processing and analysis as in the main experiment (details in Sec. S4.2 in the supplementary), results are shown in Fig. 8. In summary, results in this more realistic scenario confirm our key findings from the main experiment: the persistence of the effect of ΔT on latency acceleration with consistent modulation of latency with temporal shifts ($p < 0.001$), and the potential upper limit in acceleration efficiency at higher temporal shifts. Furthermore, results for the tested eccentricity case (E = 8°) align with the main study, demonstrating substantial latency acceleration with temporal shifts whereas spatial shift (ΔS) remains non-significant ($p = 0.08$). Saccade accuracy, quantified through landing error, exhibits no significant differences ($p = 0.91$ for ΔT and $p = 0.28$ for ΔS), maintaining errors around 1°, consistent with eye-tracker accuracy. These results confirm the robustness and generalizability of our findings across diverse scenarios.

4.2 Interactive Application: Farm Game

Our study progresses to a more interactive validation with a virtual farm game, testing our model in new conditions. For this purpose, we first fit a model using eccentricity, spatial shift, and temporal shift as inputs and saccadic latency as the output. Choosing a polynomial based on main experiment results, we perform a 10-fold cross-validation and select a 2-degree polynomial with an R-squared value of 0.95. The model is illustrated in Fig. 9 (left) and detailed in Sec. S4.1 in the supplementary.

Scene and stimuli conditions. Participants were tasked with protecting a rooster (fixation point) from approaching cats (visual targets) at new eccentricities (E = 10°, 12°, 14°), with co-located auditory cues (meows, ΔS = 0), as depicted in Fig. 9 (right). Different temporal shifts (ΔT) were computed using our model with the goal of accelerating saccades by 20, 40, and 60 ms. The game featured a timing bar that filled if participants failed to maintain gaze on the rooster, with the rooster escaping if the bar filled completely. Players used a controller to interact with and deter the cats, creating a dynamic gaming environment typical of action sequences.

Participants and procedure. Seven new participants (5 male, 2 female, ages 22-34) were involved in the study, following the same procedures for eye-tracking calibration and trials as in the main experiment and the basketball application.

Results. The average error across conditions between the measured latencies and those predicted by our model is 8 ms, confirming the robustness of our model (please refer to Sec. S4.3 in the supplementary for detailed results). The more interactive dimension of this application corroborates the model's consistency and practical relevance, as similarly demonstrated in the basketball scenario.

4.3 Other Applications

In this section, we discuss potential scenarios in which our insights could offer valuable guidelines. Our two applications demonstrate the direct feasibility of our approach as content design guidelines for triggering faster saccadic responses and overall reaction times, suggesting adaptability across various scenarios. For instance, in AR-based navigation, auditory cues can precede visual markers, reducing cognitive load and enhancing quick visual localization. In high-stakes user interface design, such as air traffic control or complex data visualization, multisensory integration principles can optimize alert timing, aiding rapid focus shifts to critical information without overwhelming users. The feasibility of audio-visual decoupling, demonstrated in platforms like Teams and Zoom, suggests potential adjustments for varying network latencies in live streaming and esports, facilitating synchronous experiences in competitive gameplay. While our model may not be immediately deployable in these scenarios, we hope our work will inspire further research to refine and adapt these concepts for practical applications.

5 LIMITATIONS AND FUTURE WORK

While we selected the ranges of our factors based on previous literature and with the goal of maintaining multisensory integration [Godfroy et al. 2003; Hopkins et al. 2022; Vidal and Vitu 2022], larger ranges could be measured. Nevertheless, we have already observed how acceleration stabilizes within the 300-400 ms delay range. Further, the ranges of temporal shifts we measure already surpass the simultaneity perception threshold [Hopkins et al. 2022; Kim and Lee 2022]. Expanding these delays further could disrupt multisensory integration, causing the visual and auditory cues to be perceived as separate, isolated events, with the sound acting as a distractor [Berti and Schröger 2001; Tellinghuisen and Nowak 2003]. Similarly, larger spatial shifts could lead to the same issue, resulting in even visual suppression effects [Hidaka and Ide 2015; Malpica et al. 2022].

Depth perception in virtual scenarios, particularly in the context of the vergence-accommodation conflict, remains an important area of research. We employed a single depth value for all stimuli, but different distances may influence sound perception and subsequent saccade latency. Additionally, variations in depth regarding audiovisual mismatch, which we only considered horizontally, could yield different results. The influence of saccades on eye accommodation and binocular vision also remains to be further explored. Previous research suggests that saccades can accelerate eye focusing [Schor et al. 1999] and disparity vergence through neural mechanisms [Duinkharjav et al. 2023; Enright 1986, 1984].

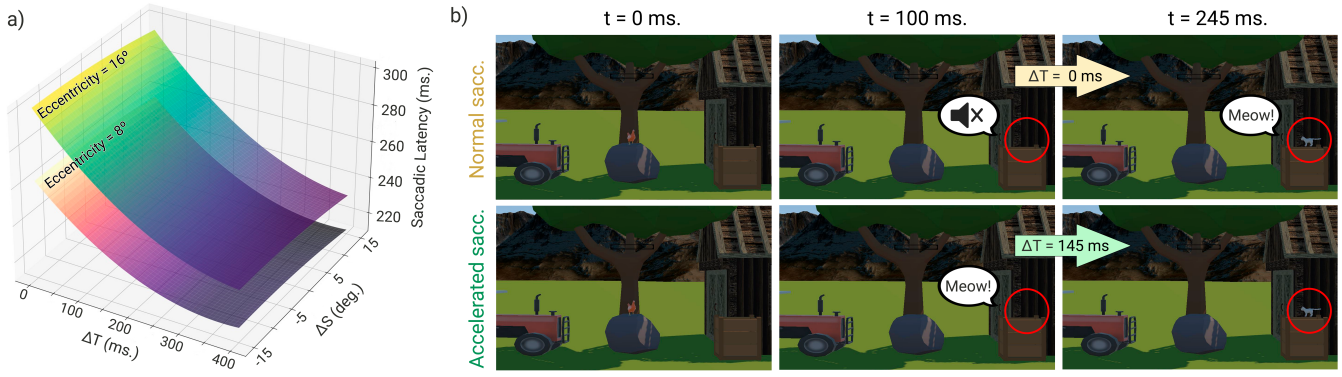

Figure 9: Latency acceleration in the interactive farm game. *Left*: Model fitted to our data as described in the text. Two representative eccentricities are depicted. *Right*: Model applied to our interactive farm game to influence saccade latency. In normal conditions (top), the sound and the visual event are presented simultaneously. In the accelerated response condition using our model (bottom), a temporal shift is employed to play the sound associated with a key game event before the visual presentation. As a result, differences in saccade latency are observed, achieving the desired acceleration as predicted by our model and prompting a faster saccadic response.

Investigating how saccade latency reduction affects these processes in the presence of temporally shifted audio could provide valuable insights.

In our experiment, we consider the *beep and flash* paradigm. Additionally, we validate our insights in a more complex scene with semantic audiovisual signals. However, future work could further investigate semantic correspondences between audio and visual cues (speech, playing an instrument, object-based videos), as well as explore how participants' emotional states, induced through various means, might influence the saccade acceleration process.

In future research, additional factors such as motion, cognitive load, contextual information, and task complexity could be explored. Regarding complex scenarios with multiple visual and auditory targets, principles like the cocktail party effect [Kaya and Elhilali 2017; Mangun 1995] show that we can prioritize primary stimuli, suggesting that our observed effect may hold in more intricate environments. We hope to establish a foundational framework and set the stage for future investigations in these directions.

6 CONCLUSION

In this work, we focus on investigating the effect of *temporal* and *spatial* alignment of visual and auditory stimuli in saccadic latency in immersive environments. Making use of VR, we study various temporal and spatial shifts between these stimuli, specifically when auditory cues precede visual information. In particular, we explore three main factors: visual eccentricity, spatial shift, and temporal shift to achieve saccade latency acceleration.

The main insights from our research reveal that temporal shifts in auditory cues can significantly accelerate saccadic responses, although this acceleration reaches a plateau beyond certain temporal intervals. Interestingly, while visual eccentricity consistently affects saccadic latency as described in previous work [Duinkharjav et al. 2022; Kalesnykas and Hallett 1994], spatial shifts in auditory cues do not seem to significantly impact this latency. Our findings align with previous research while extending our understanding of these cross-modal interactions by systematically considering these dimensions in VR environments and for different retinal eccentricities and sound source localizations. In our proof-of-concept game experiences, we demonstrate the practicality of our insights, validating them in more complex and interactive scenarios and showing that saccades can be effectively accelerated beyond experimental setups.

In conclusion, we believe that our work contributes significantly to the understanding of cross-modal interactions and their influence on saccadic latency. By investigating both spatial and temporal factors in audiovisual cues, we have provided new insights into this complex interplay. The applications of our research are broad, with potential benefits in user interfaces, AR, video games, digital multimedia, and online experiences. We hope that our work will not only serve as a useful guideline for future developments in these areas but also inspire further research in this direction.

ACKNOWLEDGMENTS

This work has been supported by grant PID2022-141539NB-I00, funded by MICIU/AEI/10.13039/501100011033 and by ERDF, EU; by the National Science Foundation grants #2225861 and #2232817; and by an academic gift from Meta. The authors would like to thank Edurne Bernal and Daniel Martin for their help with figures and the participants of the experiments for their participation.

REFERENCES

Rachel Albert, Anjul Patney, David Luebke, and Joohwan Kim. 2017. Latency Requirements for Foveated Rendering in Virtual Reality. *ACM Trans. Appl. Percept.* 14, 4, Article 25 (sep 2017), 13 pages. https://doi.org/10.1145/3127589

James Anliker. 1976. Eye Movement: On-Line Measurement, Analysis, and Control. *R. A. Monty & J. W. Senders (Eds.), Eye movements and psychological processes* (1976), 185–202.

Elena Arabadzhiyska, Okan Tarhan Tursun, Karol Myszkowski, Hans-Peter Seidel, and Piotr Didyk. 2017. Saccade landing position prediction for gaze-contingent rendering. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 1–12.

Derek H Arnold, Alan Johnston, and Shinya Nishida. 2005. Timing sight and sound. *Vision Research* 45, 10 (2005), 1275–1284.

A Terry Bahill. 1975. Most naturally occurring human saccades have magnitudes of 15 deg or less. *Invest. Ophthalmol* 14 (1975), 468–469.

Peter G.J. Barten. 1999. *Contrast sensitivity of the human eye and its effects on image quality.* SPIE – The International Society for Optical Engineering. https://doi.org/10.1117/3.353254

AH Bell, MA Meredith, AJ Van Opstal, and DougP Munoz. 2006. Stimulus intensity modifies saccadic reaction time and visual response latency in the superior colliculus. *Experimental Brain Research* 174 (2006), 53–59.

Christopher C Berger, Mar Gonzalez-Franco, Ana Tajadura-Jiménez, Dinei Florencio, and Zhengyou Zhang. 2018. Generic HRTFs may be good enough in virtual reality. Improving source localization through cross-modal plasticity. *Frontiers in neuroscience* 12 (2018), 21.

Stefan Berti and Erich Schröger. 2001. A comparison of auditory and visual distraction effects: behavioral and event-related indices. *Cognitive brain research* 10, 3 (2001), 265–273.

Benjamin M Bolker, Mollie E Brooks, Connie J Clark, Shane W Geange, John R Poulsen, M Henry H Stevens, and Jada-Simone S White. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in ecology & evolution* 24, 3 (2009), 127–135.

Kris Boyen, Deniz Baskent, and Pim van Dijk. 2015. The gap detection test: can it be used to diagnose tinnitus? *Ear and hearing* 36, 4 (2015), e138–e145.

R.H.S. Carpenter. 2004. Contrast, Probability, and Saccadic Latency: Evidence for Independence of Detection and Decision. *Current Biology* 14, 17 (2004), 1576–1580.

Geneviève Charbonneau, Marie Véronneau, Colin Boudrias-Fournier, Franco Lepore, and Olivier Collignon. 2013. The ventriloquist in periphery: impact of eccentricity-related reliability on audio-visual localization. *Journal of Vision* 13, 12 (2013), 20–20.

Shaoyu Chen, Budmonde Duinkharjav, Xin Sun, Li-Yi Wei, Stefano Petrangeli, Jose Echevarria, Claudio Silva, and Qi Sun. 2022. Instant reality: Gaze-contingent perceptual optimization for 3d virtual reality streaming. *IEEE Transactions on Visualization and Computer Graphics* 28, 5 (2022), 2157–2167.

Yi-Chuan Chen and Charles Spence. 2017. Assessing the role of the 'unity assumption'on multisensory integration: A review. *Frontiers in psychology* 8 (2017), 445.

Hans Colonius and Petra Arndt. 2001. A two-stage model for visual-auditory interaction in saccadic latencies. *Perception & psychophysics* 63, 1 (2001), 126–147.

BD Corneil, M Van Wanrooij, DP Munoz, and AJ Van Opstal. 2002. Auditory-visual interactions subserving goal-directed saccades in a complex scene. *Journal of Neurophysiology* 88, 1 (2002), 438–454.

Joan M Dafoe, Irene T Armstrong, and Doug P Munoz. 2007. The influence of stimulus direction and eccentricity on pro-and anti-saccades in humans. *Experimental Brain Research* 179 (2007), 563–570.

Adele Diederich, Annette Schomburg, and Hans Colonius. 2012. Saccadic reaction times to audiovisual stimuli show effects of oscillatory phase reset. (2012).

Oliver Doehrmann and Marcus J Naumer. 2008. Semantics and the multisensory brain: how meaning modulates processes of audio-visual integration. *Brain research* 1242 (2008), 136–150.

Budmonde Duinkharjav, Praneeth Chakravarthula, Rachel Brown, Anjul Patney, and Qi Sun. 2022. Image Features Influence Reaction Time: A Learned Probabilistic Perceptual Model for Saccade Latency. 41, 4, Article 144 (jul 2022), 15 pages.

Budmonde Duinkharjav, Benjamin Liang, Anjul Patney, Rachel Brown, and Qi Sun. 2023. The Shortest Route is Not Always the Fastest: Probability-Modeled Stereoscopic Eye Movement Completion Time in VR. *ACM Trans. Graph.* 42, 6, Article 220 (2023), 14 pages.

David Dunn, Okan Tursun, Hyeonseung Yu, Piotr Didyk, Karol Myszkowski, and Henry Fuchs. 2020. Stimulating the human visual system beyond real world performance in future augmented reality displays. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 90–100.

JT Enright. 1986. Facilitation of Vergence Changes by Saccades: Influences of Misfocused Images and of Disparity Stimuli in Man. *The Journal of physiology* 371, 1 (1986), 69–87.

J. T. Enright. 1984. Changes in Vergence Mediated by Saccades. *The Journal of Physiology* 350, 1 (1984), 9–31. https://doi.org/10.1113/jphysiol.1984.sp015186

B Fischer, S Gezeck, and K Hartnegg. 1997. The analysis of saccadic eye movements from gap and overlap paradigms. *Brain Research Protocols* 2, 1 (1997), 47–52.

Philippe Fournier and Sylvie Hebert. 2012. Gap detection deficits in humans with tinnitus as assessed with the acoustic startle paradigm: Does tinnitus fill in the gap? *Hearing research* 295 (06 2012).

Linus Franke, Laura Fink, Jana Martschinke, Kai Selgrad, and Marc Stamminger. 2021. Time-Warped Foveated Rendering for Virtual Reality Headsets. In *Computer Graphics Forum*, Vol. 40. Wiley Online Library, 110–123.

Maarten A Frens, A John Van Opstal, and Robert F Van der Willigen. 1995. Spatial and temporal factors determine auditory-visual interactions in human saccadic eye movements. *Perception & psychophysics* 57 (1995), 802–816.

Martine Godfroy, Corinne Roumes, and Pierre Dauchy. 2003. Spatial variations of visual—auditory fusion areas. *Perception* 32, 10 (2003), 1233–1245.

Souta Hidaka and Masakazu Ide. 2015. Sound can suppress visual perception. *Scientific reports* 5, 1 (2015), 10483.

Torin Hopkins, Suibi Che-Chuan Weng, Rishi Vanukuru, Emma Wenzel, Amy Banic, Mark D Gross, and Ellen Yi-Luen Do. 2022. Studying the Effects of Network Latency on Audio-Visual Perception During an AR Musical Task. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 26–34.

RP Kalesnykas and PE Hallett. 1987. The differentiation of visually guided and anticipatory saccades in gap and overlap paradigms. *Experimental Brain Research* 68 (1987), 115–121.

RP Kalesnykas and PE Hallett. 1994. Retinal eccentricity and the latency of eye saccades. *Vision research* 34, 4 (1994), 517–531.

Z Kapoula and DA Robinson. 1986. Saccadic undershoot is not inevitable: Saccades can be accurate. *Vision research* 26, 5 (1986), 735–743.

Emine Merve Kaya and Mounya Elhilali. 2017. Modelling auditory attention. *Philosophical Transactions of the Royal Society B: Biological Sciences* 372, 1714 (2017), 20160101.

Hayeon Kim and In-Kwon Lee. 2022. Studying the effects of congruence of auditory and visual stimuli on virtual reality experiences. *IEEE Transactions on Visualization and Computer Graphics* 28, 5 (2022), 2080–2090.

Joohwan Kim, Josef Spjut, Morgan McGuire, Alexander Majercik, Ben Boudaoud, Rachel Albert, and David Luebke. 2019. Esports Arms Race: Latency and Refresh Rate for Competitive Gaming Tasks. *Journal of Vision* 19, 10 (2019), 2–2.

Alan Kingstone and Raymond M Klein. 1993. Visual offsets facilitate saccadic latency: does predisengagement of visuospatial attention mediate this gap effect? *Journal of Experimental Psychology: Human Perception and Performance* 19, 6 (1993), 1251.

Denis Koposov, Maria Semenova, Andrey Somov, Andrey Lange, Anton Stepanov, and Evgeny Burnaev. 2020. Analysis of the Reaction Time of eSports Players through the Gaze Tracking and Personality Trait. In *2020 IEEE 29th International Symposium on Industrial Electronics (ISIE)*. 1560–1565.

Eileen Kowler. 2011. Eye Movements: The Past 25 Years. *Vision Research* 51, 13 (2011), 1457–1483.

Matteo Lisi, Joshua A Solomon, and Michael J Morgan. 2019. Gain control of saccadic eye movements is probabilistic. *Proceedings of the National Academy of Sciences* 116, 32 (2019), 16137–16142.

Sandra Malpica, Ana Serrano, Julia Guerrero-Viu, Daniel Martin, Edurne Bernal, Diego Gutierrez, and Belen Masia. 2022. Auditory stimuli degrade visual performance in virtual reality. In *ACM SIGGRAPH 2022 Posters*. 1–2.

George R Mangun. 1995. Neural mechanisms of visual selective attention. *Psychophysiology* 32, 1 (1995), 4–18.

Kumpei Ogawa, Kazuyuki Fujita, Shuichi Sakamoto, Kazuki Takashima, and Yoshifumi Kitamura. 2023. Exploring Visual-Auditory Redirected Walking using Auditory Cues in Reality. *IEEE Transactions on Visualization and Computer Graphics* (2023).

Charles Pierrot-Deseilligny, Sophie Rivaud, Bertrand Gaymard, René Müri, and Anne-Isabelle Vermersch. 1995. Cortical control of saccades. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society* 37, 5 (1995), 557–567.

Patricia A Reuter-Lorenz, Howard C Hughes, and Robert Fendrich. 1991. The reduction of saccadic latency by prior offset of the fixation point: an analysis of the gap effect. *Perception & psychophysics* 49, 2 (1991), 167–175.

John Ross, M Concetta Morrone, and David C Burr. 1997. Compression of visual space before saccades. *Nature* 386, 6625 (1997), 598–601.

John Ross, M Concetta Morrone, Michael E Goldberg, and David C Burr. 2001. Changes in visual perception at the time of saccades. *Trends in neurosciences* 24, 2 (2001), 113–121.

Susan M Ross and Leonard E Ross. 1981. Saccade latency and warning signals: effects of auditory and visual stimulus onset and offset. *Perception & Psychophysics* 29 (1981), 429–437.

Olli Rummukainen, Thomas Robotham, Axel Plinge, Frank Wefers, Jürgen Herre, Emanuël Habets, et al. 2019. Listening tests with individual versus generic head-related transfer functions in six-degrees-of-freedom virtual reality. In *Audio for Virtual, Augmented and Mixed Realities: Proceedings of ICSA 2019; 5th International Conference on Spatial Audio; September 26th to 28th, 2019, Ilmenau, Germany*. 55–62.

Clifton M Schor, Lori A Lott, David Pope, and Andrew D Graham. 1999. Saccades Reduce Latency and Increase Velocity of Ocular Accommodation. *Vision Research* 39, 22 (Nov. 1999), 3769–3795. https://doi.org/10.1016/S0042-6989(99)00094-2

Stefania Serafin, Michele Geronazzo, Cumhur Erkut, Niels C Nilsson, and Rolf Nordahl. 2018. Sonic interactions in virtual reality: State of the art, current challenges, and future directions. *IEEE computer graphics and applications* 38, 2 (2018), 31–43.

Ladan Shams, Yukiyasu Kamitani, and Shinsuke Shimojo. 2000. What you see is what you hear. *Nature* 408, 6814 (2000), 788–788.

Charles Spence. 2011. Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics* 73 (2011), 971–995.

Charles Spence and Jon Driver. 2004. *Crossmodal space and crossmodal attention*. Oxford University Press.

John S Stahl. 2001. Eye-head coordination and the variation of eye-movement accuracy with orbital eccentricity. *Experimental brain research* 136 (2001), 200–210.

Wolfgang A Teder-Sälejärvi, Francesco Di Russo, John J McDonald, and Steven A Hillyard. 2005. Effects of spatial congruity on audio-visual multimodal integration.

Journal of cognitive neuroscience 17, 9 (2005), 1396–1409.

Donald J Tellinghuisen and Erin J Nowak. 2003. The inability to ignore auditory distractors as a function of visual task perceptual load. *Perception & psychophysics* 65 (2003), 817–828.

Marc van Wanrooij, Andrew Bell, Douglas Munoz, and John Opstal. 2009. The effect of spatial–temporal audiovisual disparities on saccades in a complex scene. *Experimental Brain Research* 198 (09 2009), 425–437.

Argiro Vatakis and Charles Spence. 2007. Crossmodal binding: Evaluating the "unity assumption" using audiovisual speech stimuli. *Perception & psychophysics* 69 (2007), 744–756.

Manuel Vidal, Andrea Desantis, and Laurent Madelain. 2020. Irrelevant auditory and tactile signals, but not visual signals, interact with the target onset and modulate saccade latencies. *Plos one* 15, 2 (2020), e0221192.

Manuel Vidal and Françoise Vitu. 2022. Multisensory temporal binding induces an illusory gap/overlap that reduces the expected audiovisual interactions on saccades but not manual responses. *Plos one* 17, 4 (2022), e0266468.

Jean Vroomen and Mirjam Keetels. 2010. Perception of intersensory synchrony: a tutorial review. *Attention, Perception, & Psychophysics* 72, 4 (2010), 871–884.

Chin-An Wang, Gunnar Blohm, Jeff Huang, Susan E Boehnke, and Douglas P Munoz. 2017. Multisensory integration in orienting behavior: Pupil size, microsaccades, and saccades. *Biological psychology* 129 (2017), 36–44.

Elizabeth M Wenzel, Marianne Arruda, Doris J Kistler, and Frederic L Wightman. 1993. Localization using nonindividualized head-related transfer functions. *The Journal of the Acoustical Society of America* 94, 1 (1993), 111–123.

Shimpei Yamagishi and Shigeto Furukawa. 2020. Factors influencing saccadic reaction time: Effect of task modality, stimulus saliency, spatial congruency of stimuli, and pupil size. *Frontiers in Human Neuroscience* 14 (2020), 571893.

Daniela Zambarbieri. 2002. The latency of saccades toward auditory targets in humans. *Progress in Brain Research* 140 (2002), 51–59.

Daniela Zambarbieri, Giorgio Beltrami, and Maurizio Versino. 1995. Saccade latency toward auditory targets depends on the relative position of the sound source with respect to the eyes. *Vision research* 35, 23-24 (1995), 3305–3312.

Heng Zou, Hermann J Müller, and Zhuanghua Shi. 2012. Non-spatial sounds regulate eye movements and enhance visual search. *Journal of Vision* 12, 5 (2012), 2–2.

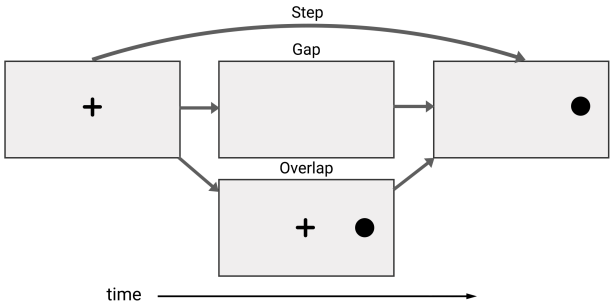

Figure 2: Visual gap-overlap Paradigm. This paradigm involves three distinct scenarios. In the *step* case, the target (represented by the filled circle) appears immediately after the offset of the fixation point (denoted by the cross). In the *gap* case, there is a delay between the disappearance of the fixation point and the appearance of the target, which typically results in faster saccade responses. Conversely, in the *overlap* case, the target appears before the fixation point disappears, leading to slower saccades. The gap scenario facilitates quicker saccades by reducing attention engagement at the fixation point [Kingstone and Klein 1993], while the overlap condition holds attention longer on the fixation point, requiring active disengagement and thus delaying the saccade initiation [Kalesnykas and Hallett 1987].

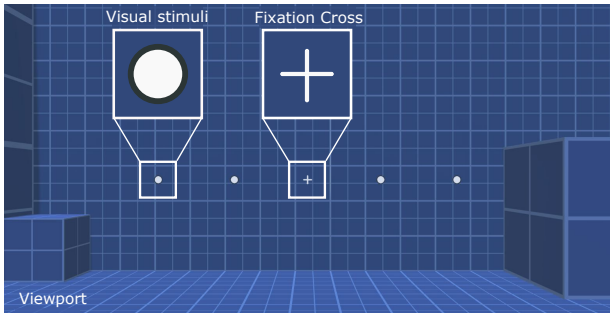

Figure 4: Illustration of the experimental view from the participant's perspective: Central to the view is a fixation cross with a 0.12° diameter, surrounded by visual targets measuring 0.36° in diameter, positioned at eccentricities of E=[-16°, -8°, 8°, 16°]. These elements are visually enlarged for visualization purposes. The purpose of the surrounding virtual environment is to provide some visual cues for depth perception and a higher feeling of immersion (adapted from Speed Tutor Assets from Unity Asset Store).

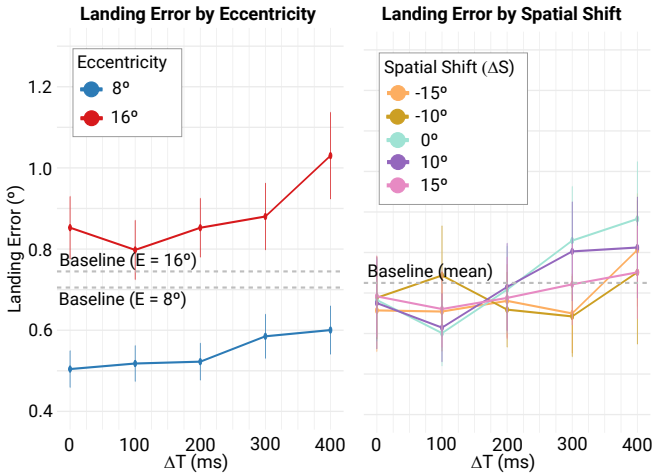

Figure 6: Saccade landing error in the main experiment for different temporal shifts depending on eccentricity (left) and spatial shifts (right). Baseline cases (no sound) are represented with dashed lines. The increased error at larger eccentricities can be partly due to reduced eye-tracker accuracy, yet all errors stay below 1°. Error bars represent 95% CI.

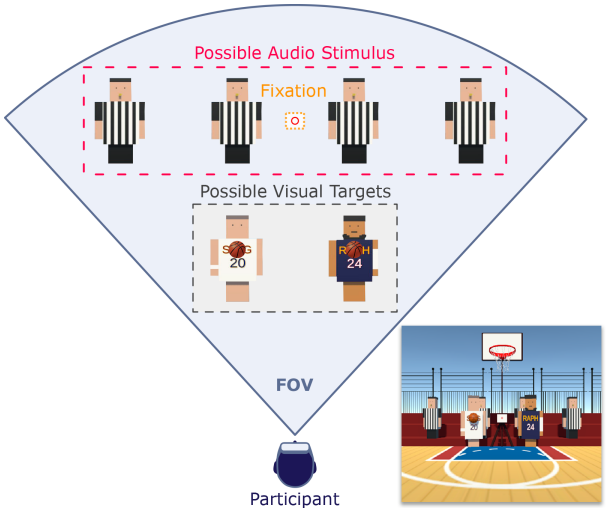

Figure 7: Illustration of our Basketball training application. The fixation point is a red cross located under the basket in the middle of the whiteboard. The visual target is a basketball being held by teammates while the auditory cue is a whistle-blowing generated by the referees on the back. The basketball court asset is obtained from Unity Asset Store.

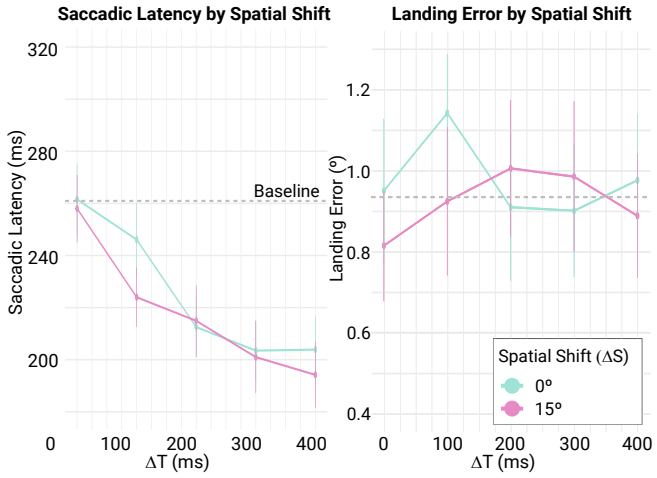

Figure 8: Saccadic latency (left) and landing error (right) in the basketball application for different temporal shifts depending on the spatial shifts. Baseline cases (no sound) are represented with dashed lines. Error bars represent 95% CI.