Modeling the Impact of Head-Body Rotations on Audio-Visual Spatial Perception for Virtual Reality Applications

Edurne Bernal-Berdun, Mateo Vallejo, Qi Sun Ana Serrano and Diego Gutierrez

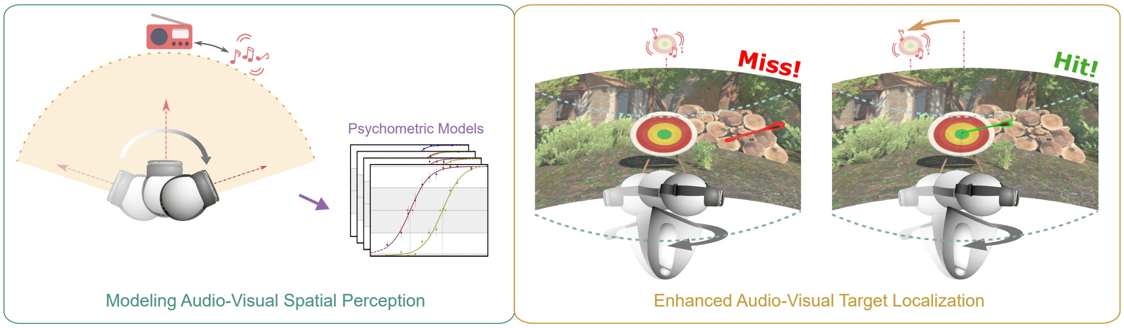
Psychometric Models
Miss!
Hit!
Modeling Audio-Visual Spatial Perception
Enhanced Audio-Visual Target Localization

Fig. 1: Fast and accurate audio-visual target localization is essential in VR applications. However, head-body rotations induce a mismatch between our auditory and visual perceptual spaces. This leads to an overestimation of the sound location with respect to the visual in the rotation direction of the head, hindering precision and accuracy. In our work, we model the spatial localization acuity when integrating audio-visual, vestibular, and proprioceptive cues during natural head-body rotations typical of VR applications. We first measure the spatial detection thresholds of perceived audio-visual mismatches under different head-body rotations, proposing psychometric models to describe spatial audio-visual perception (left). We then show how our model can be leveraged to improve target localization performance by applying an offset to the sound that compensates for the measured overestimation of the target location (right).

Abstract—Humans perceive the world by integrating multimodal sensory feedback, including visual and auditory stimuli, which holds true in virtual reality (VR) environments. Proper synchronization of these stimuli is crucial for perceiving a coherent and immersive VR experience. In this work, we focus on the interplay between audio and vision during localization tasks involving natural head-body rotations. We explore the impact of audio-visual offsets and rotation velocities on users' directional localization acuity for various viewing modes. Using psychometric functions, we model perceptual disparities between visual and auditory cues and determine offset detection thresholds. Our findings reveal that target localization accuracy is affected by perceptual audio-visual disparities during head-body rotations, but remains consistent in the absence of stimuli-head relative motion. We then showcase the effectiveness of our approach in predicting and enhancing users' localization accuracy within realistic VR gaming applications. To provide additional support for our findings, we implement a natural VR game wherein we apply a compensatory audio-visual offset derived from our measured psychometric functions. As a result, we demonstrate a substantial improvement of up to 40% in participants' target localization accuracy. We additionally provide guidelines for content creation to ensure coherent and seamless VR experiences.

Index Terms—Virtual Reality, Audio-Visual Spatial Perception

1 Introduction

Humans perceive the world by multimodal sensory feedback, both external (mainly visual and auditory) and internal (balance and body position awareness) [35,40]. This principle is also present in virtual reality (VR) environments, where multimodality has been found even to enhance the viewing experience [16,27]. However, to create a cohesive perception of these stimuli, proper synchronization between different sensory sources is essential [29,34]. While visual information is our primary cue to spatial perception, our acuity for visual target detection is remarkably influenced by sound. For instance, adding audio to visual stimuli can improve our ability to discriminate their motion [38]. Moreover, audio contributes to a better understanding of the environment, leading to a more comfortable and immersive experience [17,41].

In this paper, we focus on the interplay of audio and vision during the process of searching for and locating objects while rotating our heads. Such rotation triggers a complex three-fold mechanism across visual, audio, and proprioception sensations (i.e., body position awareness). For instance, rapid head movements introduce ballistic changes on the retina, which in turn leads to a degradation of the perceived auditory space [25]. While this effect is specific to rapid head movements, it evidences the intertwined nature of head movements and sound localization perception. Locating objects by rotating our heads is a common task in both VR and the physical world. Further, it has been found that users in VR tend to exhibit more frequent head movements to fetch targets than in the physical world [18]. However, our understanding of target localization performance in natural, multimodal (audio-visual) settings is still limited, particularly when considering various viewing modes such as stationary viewing (e.g., film-watching), target pursuit (e.g., gaming), or re-orientation (e.g., teleconferencing), all common motions when consuming VR content. We aim to study the interaction of visual, audio, and proprioceptive sensory modalities during target localization under these different viewing modes. By developing a quantitative model, we seek to gain deeper insights into these scenarios. This study is an important step towards optimizing virtual content and enhancing content creation, especially in cases requiring accurate spatial localization.

- Edurne Bernal-Berdun, Mateo Vallejo, Ana Serrano and Diego Gutierrez are with Universidad de Zaragoza - I3A. E-mail: edurnebernal | mvallejo | anase | diegog@unizar.es.
- Qi Sun is with the New York University. E-mail: qisun@nyu.edu.

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

To achieve this goal, we measure and model the directional localization acuity of VR users through psychometric functions when integrating visual, auditory, and proprioceptive cues. We explore four head-body movements (*Stationary*, *Pursuit* and *Slow/Fast re-orientation*), and examine different audio-visual offsets. Our analysis reveals differences in the point of subjective equality (when the visual and auditory stimuli are perceived as aligned) and offset detection thresholds for the different viewing modes. In particular, our results show that when the head is aligned *relative to the visual stimulus*, regardless of whether it remains static (Stationary mode) or is pursuing a target (Pursuit mode), users exhibit a high degree of sensitivity in discerning misalignments between auditory and visual cues. However, when there is relative motion between the head and the visual stimulus (Re-orientation mode), our ability to detect audio-visual offsets is notably compromised, directly impacting our audio-visual localization performance. Understanding and quantifying the mismatch in audiovisual target localization therefore carries significant implications for practical VR usage. For example, in competitive VR gaming scenarios where quick and accurate audiovisual target localization is crucial, compensating for these perceived disparities can lead to competitive advantages, as exemplified in Figure 1 and demonstrated by our application in Sec. 4.1. Our proof-of-concept experiment, implemented as an interactive VR game, shows a significant increase in localization accuracy of up to 40% by applying a compensatory audio-visual offset based on our measured psychometric functions. Our findings also have implications in VR storytelling, where spatial audiovisual synchronization is critical for creating seamless and immersive user experiences during head rotations [28, 33]. In Sec. 4.2, we derive practical guidelines to assist in achieving this goal. Additionally, applications such as training and simulation, like military or emergency response training, may greatly benefit from accurate audiovisual target localization for effective decision-making.

We believe our research opens up new avenues for exploring the boundaries of perceptual capabilities and performance in the development of virtual and augmented reality (AR) interfaces and content. Our source code and measured de-identified data can be found at https://graphics.unizar.es/projects/AV_spatial_perception/.

remo

2 Related Work

Previous studies on sound localization acuity demonstrate outstanding precision in discerning sounds with horizontal spatial separations of over 3° [14, 31]. Researchers have also explored the concept of auditory motion parallax, showing that head movements contribute to improved spatial localization [13]. Nevertheless, when it comes to judging the localization of moving audio during head rotation, a perceptual error arises, since we can only compensate for 85% of the mismatch introduced by the speed difference between head- and world- coordinates [10]. Moreover, rapid head movements can perceptually compress the auditory space and interfere with an accurate sound localization [25]. Some studies have attributed the distortion of the perceptual auditory space to a misperceived head-on-trunk orientation [24] or vestibular stimulation [45]. A similar phenomenon takes place within the visual system, where saccades, which are swift and ballistic eye movements, contribute to enhancing spatial localization. By shifting the point of focus to different stimuli within the field of view, saccades enable the formation of high-resolution images of the environment [4]. However, these rapid and ballistic eye movements can also introduce significant distortion in visual-spatial perception, resulting in a perceptual compression of the visual space [36]. Hence, it has been established that perceptual disparities are likely to emerge under diverse head movement conditions, affecting both visual and audio spatial localization separately. As such, disparities are anticipated to manifest as well in audio-visual localization.

Audio-visual stimuli spatial localization is a multimodal process, in which both sensory signals are integrated to estimate the spatial location. This integration process is a complex paradigm in which multiple factors can affect the result. Vision usually predominates multimodal integration [32]. However, under different conditions, this predominance can change. For instance, audio dominates audio-visual integration mechanisms when the reliability of visual information is reduced [12]. As in purely auditory or visual perception, head and eye movements can also influence multimodal integration. For example, Lewald and Karnath [26] showed that if a visual stimulus is introduced during head rotation, the distortion of auditory space is mitigated. One possible explanation is that the visual stimulus suppresses the vestibular nystagmus effect, which causes a shift in the mean eye position relative to the head in the direction of rotation [1]. This hypothesis was further analyzed by Van Barneveld and John Van Opstal [46], whose results show that the perceived spatial location of auditory sources is shifted in the direction of the eyes. This suggests that the presence of a visual stimulus in audio-visual stimuli may induce an eye eccentricity that causes a perceptual audio-visual spatial disparity when integrating both stimuli.

Previous works have also studied the influence of *binding* effects between auditory and visual stimulus on multimodal integration under spatial incongruities [5]. The existence of cognitive links between auditory and visual stimuli such as time synchrony or semantic meaning can infer a binding wherein visual and audio cues are integrated as a unified audio-visual stimulus, even if they are spatially misaligned. Several studies have explored this phenomenon, attempting to identify the spatial offset threshold beyond which the unified integration is broken. However, due to the complexity of this phenomenon, factors such as the participants' training, type of sound [44], device [22], or position in the field of view [3] influence the reported thresholds, making it challenging to establish a definitive threshold. This binding effect has also been analyzed in augmented and virtual reality environments. Kytö et al. [22] reported greater spatial limits for AR than in real environments, while the study performed by Kim and Lee [19] in VR showed more conservative limits.

In this paper, we focus on measuring the perceptual localization of audio-visual stimuli featuring a spatial offset. Specifically, to the best of our knowledge, we are the first to explore the impact of users' natural self-head and -body rotation (i.e., active rotation) on this task. Our study does not specifically target the cognitive binding effect between auditory and visual stimuli but instead aims to characterize the accuracy in perceiving the relative spatial position between them while modeling the perceptual disparities introduced by head-body rotations. As our experimental setup replicates natural movements performed in VR, we show how the insight found in our study can directly be applied to enhance target localization in real applications.

3 Measuring Audio-visual Spatial Disparity

A *perceptual disparity* exists when a pair of spatially misaligned auditory and visual stimuli are perceived as being aligned or vice versa. To measure these potential disparities, we designed and performed a psychophysical experiment, under various self-motion conditions to simulate natural movements in VR. From the participants' responses, we fitted a psychometric curve that models perceptual audio-visual disparity for the different viewing modes.

Participants.

The study was conducted with 22 participants (ages 22 - 43) including seven females, with no participants identifying themselves as *non-binary*, *not listed*, or *prefer not to disclose*. All participants provided written consent and reported normal or corrected-to-normal vision and audition. The research protocol was approved by the *Comité de Ética de la Investigación de la Comunidad Autónoma de Aragón (CEICA)*.

Hardware.

We experimented with an HTC Vive Pro Eye head-mounted display (HMD) and its integrated Tobii eye-tracker. This HMD has a display resolution of 1440 × 1600 per eye, a field of view of 110 degrees, and a frame rate of 90 fps. Our experimental setting features controlled audio-visual stimuli, which minimize the potential influence of optical distortions. Furthermore, the HTC Vive Pro, which has been extensively employed in perceptual studies, has an end-to-end latency of under 80 milliseconds [6]. This latency level falls within the standard range for VR systems and consistently remains below the Just Noticeable Difference (JND) threshold for perceived audio-visual

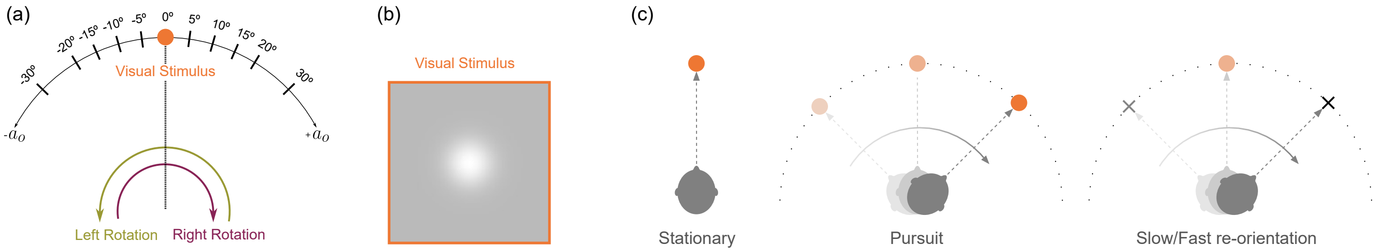

Fig. 2: **(a)** Diagram representing the two rotation directions and eleven sampled offsets applied to the auditory stimulus with respect to the visual stimulus, ranging from a_o=$-30°$ to a_o=$+30°$. **(b)** Visual stimulus shown to the participants. **(c)** Representation of the four viewing modes evaluated. *Left*: Stationary mode in which the audio-visual stimulus is presented statically right in front of the participant. *Center:* Pursuit mode in which a dynamic audio-visual stimulus, performing a semi-circular trajectory with a velocity of 50°/s, is observed by the participant instructed to follow it. *Right:* Slow and Fast re-orientation modes in which participants perform a head-body rotation at 50°/s and 25°/s while passing over an audiovisual stimulus that remains static in the middle of the semi-circular rotation trajectory. Participants' rotation velocity is guided by a fixation cross. Refer to Section 3 for further details on the experimental conditions and to Section 3.1 for an in-depth analysis of the study's results.

synchrony [47]. We performed a five-point eye-tracker calibration before each experiment phase and measured head and eye parameters at a frequency of 120 Hz. Participants remained in a standing position during the whole experiment. A retractable cord was attached to the ceiling to manage the HMD cable and avoid interference with participants' rotations. The directional audio was reproduced by the incorporated HTC Vive Pro headphones using the Steam Audio SDK, which implements a generic Head Related Transfer Function (HRTF) with 5° of resolution in the horizontal plane.

Stimulus. The visual stimulus consisted of a Gaussian patch of 5° of visual angle with a maximum contrast level of 1. It was rendered at a distance of 5 meters with respect to the participant. The stimulus was displayed against a grey background environment with a contrast level of 0.73, which is perceived as an intermediate contrast level (Figure 2.b). The auditory stimulus consisted of white noise in the whole audible band, from 20 to 20,000 Hz. We established eleven possible longitudinal offsets a_o in the location of the auditory stimulus with respect to the visual stimulus (Figure 2.a). Both visual and auditory stimuli were presented simultaneously for 2 seconds, appearing concurrently to support the perception of a cohesive audio-visual stimulus.

Conditions The audio stimulus location a_o was presented at various relative horizontal offsets to the visual stimulus: $0°, \pm5°, \pm10°, \pm15°, \pm20°, \pm30°$; where 0° indicates that the audio-visual stimulus was perfectly aligned, while $-, +$ indicate that the auditory stimulus was placed to the left or right, respectively, of the visual stimulus. In addition to sampling these offsets, we also designed four different viewing modes resembling common motions in VR experiences, illustrated in Figure 2.c:

1. **Stationary:** The audio-visual stimulus was presented statically in front of the participants while their head remained stationary.
2. **Pursuit:** The audio-visual stimulus moved with constant velocity following a semi-circular trajectory (from 0° to ±180°) with a 5-meter radius. The velocity was set to 50°/s, which is the mean head velocity found in VR when not fixating [42]. Both left and right rotation directions were included. Participants were instructed to follow the stimulus with their gaze, allowing for natural head and body rotation.
3. **Fast re-orientation:** The audio-visual stimulus remained static at 90° with respect to the participants' starting head orientation. An auxiliary dynamic fixation cross was added to guide rotation speed. The fixation cross followed the same trajectory as described in the Pursuit mode for the audio-visual stimulus, moving with a velocity of 50°/s. Participants were instructed to follow the cross with their gaze, while allowing for natural head and body rotation. Both left and right rotation directions were also included.
4. **Slow re-orientation:** Similar to the previous mode, the auxiliary cross moved on a semi-circular trajectory but at a slower speed of 25°/s. We created these two re-orientation conditions since our pilot studies revealed a potential influence of this velocity, which we also aim to explore (see Section 3.2 for further details).

We opt for horizontal offsets primarily because audio exhibits better spatial resolution along this axis [14]. Additionally, in both VR and the real world, most relevant content is predominantly concentrated around the equator, resulting in more frequent and rapid horizontal rotations [15, 42].

Following recent work on motion perception in VR [39, 50], our experiment allowed for natural head-body rotations instead of restricting participants' movements and artificially simulating motion. According to prior art, passive rotation such as the one induced by wheelchairs or rotatory platforms influences the spatial perception of auditory [13] and visual stimuli [48, 49]. Hence, it is crucial to investigate the effects of active self-rotation, which more closely resembles real-world scenarios where users freely navigate the VR environment. However, allowing for natural motion introduces some variability as accidental rotations may occur. To address this, we analyzed head motion and eye tracking data to ensure that participants behaved as expected and met the experiment's requirements (Section 3.2).

Procedure We conducted a two-alternative forced-choice (2AFC) experiment with a method of constant stimuli, following previous literature for detection thresholds [8, 21, 39]. Participants' task was to determine whether the auditory stimulus was to the left or to the right of the visual stimulus after each trial. We followed a full-factorial design in which all participants experienced the four different viewing modes with the eleven possible offsets a_0. Additionally, each offset a_o was sampled three times for each condition and we also explored both left and right rotation directions when applicable. This resulted in 33 trials for the Stationary viewing mode, and 66 trials for the Pursuit, Fast re-orientation, and Slow re-orientation modes. To prevent sickness caused by frequent changes in speed, participants completed all the trials for each viewing mode before moving on to the next. The order of the viewing modes and offsets within each viewing mode was randomized. Additionally, four test trials were introduced at the beginning of each experimental condition to familiarize the participants with the task and procedure. To prevent participants' fatigue, we established a break after each 33 trials (approximately four minutes). The entire session had an average duration of 45 minutes. We acquired 5,082 trials in total.

3.1 Psychometric Fitting

With the collected binary responses from the 2AFC task, we apply psychometric modeling and perform statistical analyses. We first compute the proportion of *rightward* responses (the audio being perceived as coming from the right of the visual stimulus) for each of the varying offsets of the auditory stimulus with respect to the visual stimulus. Then, we fit a psychometric function ψ to estimate each participant's thresholds for detecting audio-visual misalignments. This also allows us to determine the longitudinal location at which the auditory stimulus is perceived as fully aligned with the visual stimulus, known as the *point of subjective equality* (PSE). This corresponds to $\psi = 0.5$, which is equivalent to chance-level performance. Following previous work [39], we select a logistic function S as a core function for ψ, which is denoted as:

$$S(a_o;m,w) = \frac{1}{1+e^{-2\log\left(\frac{1}{\alpha}-1\right)\frac{a_o-m}{w}}}, \tag{1}$$

where m is the PSE at which the visual and auditory stimuli are perceived as aligned such that $S(m) = 0.5$. We set α as 0.05, thus the width w of the logistic function is defined as $S^{-1}(1-\alpha) - S^{-1}(\alpha)$. Then, our psychometric function ψ is expressed by the logistic function S and the upper (γ) and lower (λ) asymptotes as:

$$\psi(a_o;m,w,\lambda,\gamma) = \gamma + (1-\lambda-\gamma)\,S(a_o;m,w). \tag{2}$$

Asymptotes γ and λ, commonly referred to as stimulus-independent lapses, shift based on erroneous answers unrelated to a_o (e.g., loss of focus). Following Schütt et al. [37], we use Bayesian inference to fit the psychometric function ψ, assuming an equal asymptote condition where stimulus-independent errors are equally probable for both "rightward" and "leftward" answers. In accordance with standard procedure in psychometric experiments [22, 39], we express as $\text{DT}_{25\%}$ the detection threshold for audio offsets to the right of the visual stimulus ($\psi = 0.25$), and as $\text{DT}_{75\%}$ the detection threshold for audio offsets to the left of the visual stimulus ($\psi = 0.75$). Hence, the width between $\text{DT}_{25\%}$ and $\text{DT}_{75\%}$ defines the region of low sensitivity to misalignments between audio and visual stimulus known as *the confusion region*. To ensure reliable model generalization, we carefully filter out unreliable participants with generally poor task performance. Specifically, participants must achieve at least 75% answer accuracy corresponding to the extreme offset values (i.e., $a_o = -30°, 30°$). From the original pool, two participants who identified as female did not satisfy this restriction. Figure 3 and Table 1 display the fitted psychometric functions, as well as the obtained PSEs and DTs for the different viewing modes and left and right rotations. Please note that while the figures show pooled data across participants for simplicity, our analysis and insights are derived from the distribution of individual participants' psychometric functions.

3.2 Analysis and Discussion

In this section, we report and discuss the main insights of our analysis. Please refer to the supplementary material for the complete details of the statistical analysis.

Viewing mode. We first investigate the audio-visual spatial perception for the four viewing modes: Stationary, Pursuit, Fast re-orientation, and Slow re-orientation. We found that participants exhibit high detection sensitivity for audio-visual offsets in Stationary and Pursuit modes. However, in Slow and Fast re-orientation modes, they tend to overestimate the location of the auditory stimulus toward the self-rotation direction. The PSEs and DT for each condition are visualized in Table 1.

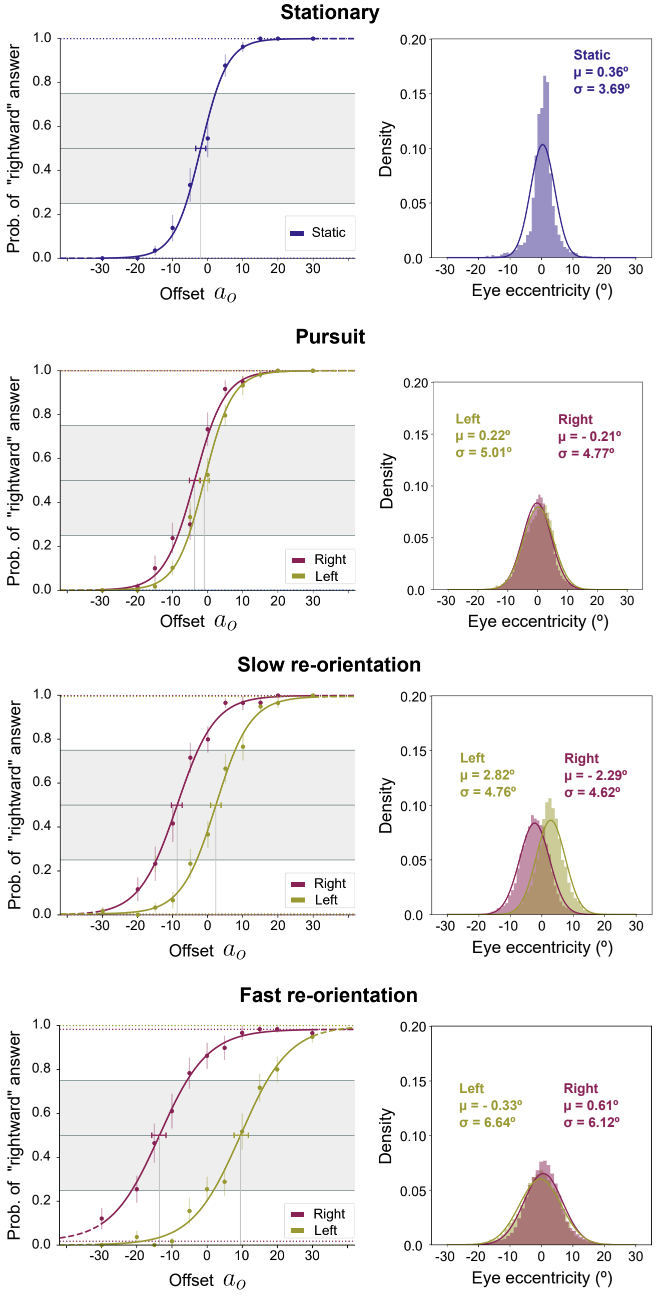

Fig. 3: *Psychometric analysis of viewing modes.* **Left**: The x-axis shows the offset a_o applied to the auditory stimulus with respect to the visual stimulus. The y-axis shows the proportion of trials in which the sound was perceived as being to the right of the visual stimulus. Each point represents the participants' pooled answers while error bars indicate the standard error. The horizontal gray line represents the point of subjective equality (PSE), the azimuthal location at which the sound is perceived as fully aligned with the visual stimulus. The horizontal error bar indicates the confidence interval for the PSE. The dotted lines show the asymptotes γ and λ. Rotation direction is color-coded in purple (right rotation) and gold (left rotation). **Right**: Gaussian distribution of eye eccentricities for each condition, with its mean μ and standard deviation σ.

In Stationary and Pursuit modes, the high sensitivity to slight misalignments is evidenced by the narrow width ($<10°$) between $\text{DT}_{25\%}$ and $\text{DT}_{75\%}$ (see first three rows of Table 1). Moreover, the PSEs are between -4° and -1°, close to the area occupied by the visual stimulus ($-2.5°$ to $2.5°$). This suggests that participants accurately localize the audio co-located with the visual. That is, we observe no spatial disparities in the perception of audio-visual stimuli when the head is relatively stationary to the visual stimulus. This observation holds both when the head and stimulus are static (Stationary), or when the head is rotating in solidarity with the stimulus (Pursuit). This also suggests that using a generic HRTF to reproduce the sound may be sufficiently precise for effectively discerning the location of the sound source, potentially reducing the need for a personalized HRTF. Having verified that our data adheres to a normal distribution (Shapiro-Wilk test, $p > 0.05$ for all conditions), we conduct a repeated measures ANOVA. We employ Dunn-Bonferroni post-hoc tests with Bonferroni correction to identify any significant variations among the PSE values across the four viewing modes. The statistical analysis, performed on per-participant PSEs, revealed no significant differences between the Stationary and Pursuit (both left and right rotations) modes ($p = 1.0$), suggesting similar behavioral patterns between them. Both conditions favor the alignment of the head with the audio-visual stimulus, which

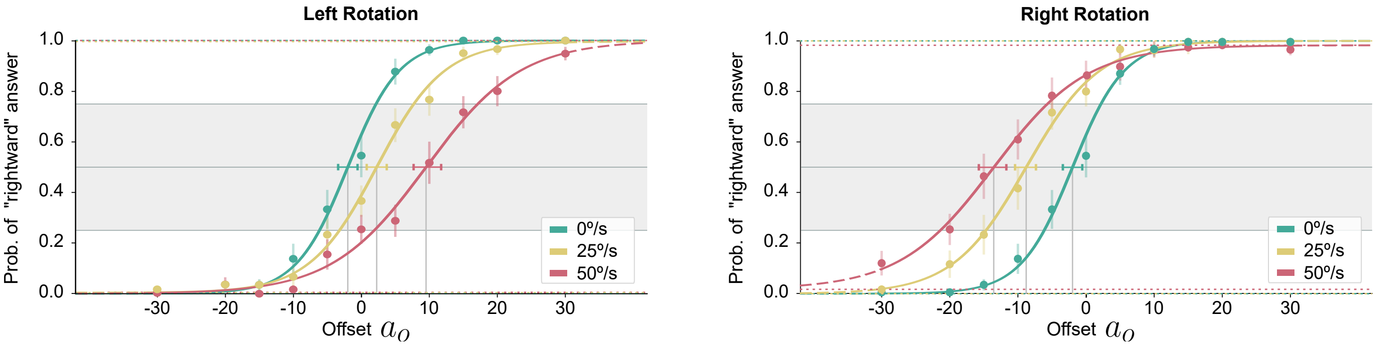

Fig. 4: *Psychometric analysis of head rotation velocities.* The x-axis shows the offset a_o applied to the auditory stimulus with respect to the visual stimulus, and the y-axis shows the probability of perceiving the auditory stimulus to the right of the visual stimulus (i.e., "rightward" answer). The standard error of the mean across participants and trials is represented as vertical error bars and a horizontal gray line indicates the PSE for each function, where the confidence interval for the PSE is represented with a horizontal error bar. Asymptotes γ and λ are denoted as horizontal dot lines. The PSE shifts toward the direction of head rotation when velocity increases.

has been reported by previous studies to improve the detection of spatial misalignments [3, 22].

On the other hand, in the Slow and Fast re-orientation modes, participants displayed a significant PSE shift towards the opposite direction of head rotation, significantly different with respect to the Stationary mode ($p < 0.05$). Specifically, participants tend to perceive the location of the auditory stimulus as shifted toward their head rotation direction. This is evidenced by the leftward/rightward shifts in the psychometric curves during right/left rotations. The shift in the curves can also be observed in the PSE and DT values shown in the last four rows of Table 1. This indicates a perceptual discrepancy between the auditory and visual stimuli in these modes, where participants perceive the stimuli as aligned when they are actually not, and vice versa. Regardless of the presence or absence of perceptual disparity, our experimental conditions consistently show a slight PSE shift towards the right, aligning with previous findings that indicate a rightward bias in human auditory spatial perception [7, 43].

The detection thresholds found for audio-visual offsets in the Stationary and Pursuit modes align with the minimum threshold angle required to discriminate the location of two sounds in non-VR/AR environments [14, 31]. These thresholds are more conservative than those observed in VR/AR [19, 22]. This indicates that participants are more acute in detecting misalignments when they receive specific instructions to identify the location of the auditory stimulus compared to when their task is limited to assessing alignment between the stimuli.

Rotation velocity. Given the perceptual disparity between the auditory and visual stimuli towards the direction of rotation in the Slow and Fast re-orientation modes, we investigate how the relative velocity of the head rotation influences this disparity. This factor represents the main distinction between these modes and the Stationary and Pursuit modes. We found that audio-visual disparity in re-orientation modes expands with increased head velocity relative to the stimulus. We analyze three velocities for both left and right rotation directions: 0°/s (Stationary), 25°/s (Slow re-orientation), and 50°/s (Fast re-orientation). We designed these velocity values to strike a balance between experiment tractability and sensitivity in detecting meaningful effects, as it was not feasible to sample higher velocities in order to maintain a fixed stimulus exposure time of two seconds within a field of view of 110°.

Figure 4 displays the psychometric functions fitted for the three different velocity conditions for both rotation directions. As shown in Table 1, the absolute offset a_o at the PSE increases with the magnitude of the rotational velocity. This indicates that higher velocities result in a greater perceived misalignment toward the direction of rotation. To assess whether the a_o at the PSE is significantly different among the three velocities, we conducted a statistical analysis using repeated measures ANOVA with Dunn-Bonferroni post-hoc tests and Bonferroni correction for multiple comparisons due to the normal distribution of our data ($p > 0.05$ for all conditions, Shappiro-Wilk test). Our analysis of the distribution of per-participant psychometric fits revealed significant differences ($p < 0.05$) between the PSE values for left and right rotation in the Slow and Fast re-orientation modes. Significant differences are also observed between 0°/s, 25°/s, and 50°/s velocities while left rotation ($p < 0.05$), although no significant difference was found between 25°/s and 50°/s velocities while rotating right ($p = 0.14$), which could be due to the rightward bias in human's auditory perception [7, 43].

Table 1: Values for the PSE, left detection threshold $\text{DT}_{25\%}$, and right detection threshold $\text{DT}_{75\%}$ for all the experimental conditions. The difference between the detection thresholds is denoted as $\text{width}_{25\%-75\%}$, which indicates the confusion area when participants cannot reliably identify the misalignment of the offset.

Viewing mode	Rot. Direction	$\text{DT}_{25\%}$	PSE	$\text{DT}_{75\%}$	$\text{Width}_{25\%-75\%}$
Stationary	-	-6.14°	-1.98°	2.17°	8.32°
Pursuit	Right	-8.48°	-3.72°	1.04°	9.52°
	Left	-5.38°	-0.97°	3.44°	8.82°
Slow re-orientation	Right	-14.68°	-8.78°	-2.88°	11.80°
	Left	-3.15°	2.26°	7.68°	10.82°
Fast re-orientation	Right	-21.06°	-13.54°	-6.03°	15.04°
	Left	1.90°	9.52°	17.14°	15.23°

Head velocity control. The analysis of participants' self-motion behavior is a critical aspect of our experiment. Since we do not explicitly control nor restrain their movement, we examine whether participants' behavior aligns with the expected responses for each condition (Section 3). Specifically, we quantitatively assess the *mean head rotation velocity* for each trial using the recorded HMD data. The distribution of head velocities for each mode was analyzed to confirm experimental requisites were met (Table 2), and 3% of the trials with outlier velocities (1.5×IQR method) were removed from the data pool.

Table 2: Mean, first quantile (Q1), and third quantile (Q3) head velocities for the four experimental conditions. The head velocities met the experimental conditions of 0°/s, 50°/s, 25°/s, and 50°/s for the Stationary, Pursuit, Slow re-orientation, and Fast re-orientation respectively.

Velocity (°/s)	Q1	Mean	Q3
Stationary	0.07	0.32	0.43
Pursuit	50.06	52.7	54.96
Slow re-orientation	25.98	27.5	28.65
Fast re-orientation	48.32	50.95	53.47

Eye eccentricity Previous research has revealed the influence of eye eccentricity on audio-visual perception [46]. Therefore, we assess the eye eccentricity's impact on participants' spatial misalignment perception that is exhibited in the re-orientation modes. Our results show

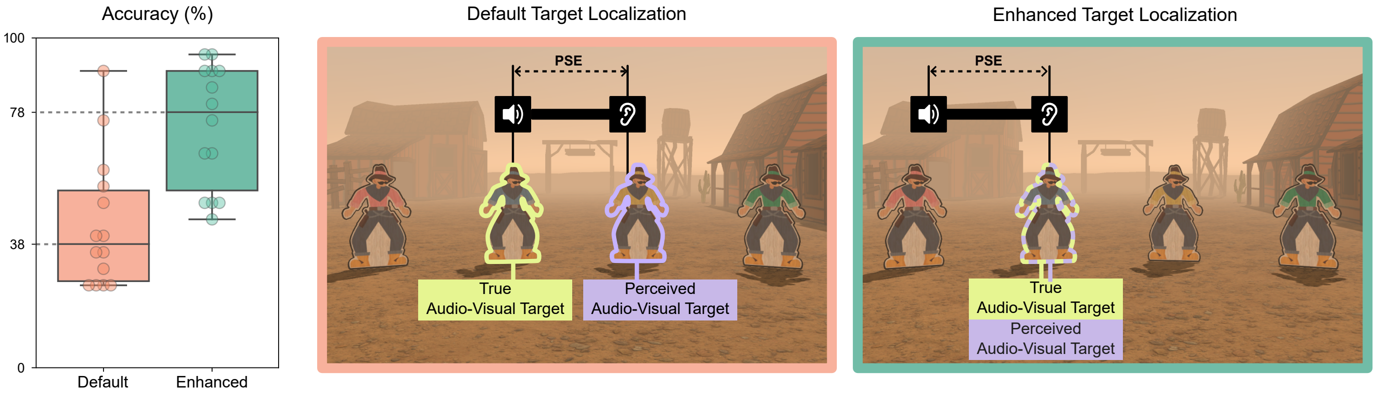

Fig. 5: *Application results.* **Left**: Participants' accuracy at identifying the target audio-visual stimulus (i.e., villain) among the non-target visual-only stimuli. It significantly increases with the compensation offset. **Middle**: Representation of the default setup where the audio is completely aligned with its corresponding visual target (yellow). However, the head-body rotation induces a shift in the PSE, causing a perceptual bias to the right (purple). **Right**: Instead, we apply a compensatory leftward shift to the auditory stimulus. The shift value is suggested by our measured PSE. As a result, the audio was accurately perceived as emitted from its corresponding visual stimulus.

that eye eccentricity is not a major cause of the audio-visual spatial disparity. We measure the longitudinal eye eccentricity, representing the difference between eye and head gaze vectors, as the large circle distance between the longitudinal eye and head coordinates for each trial timestamp. We show in Figure 3.b the distribution of eye eccentricities, which exhibit a Gaussian distribution whose sigma slightly increases when head rotation is present. In the Fast Re-orientation mode, eye eccentricities are centered around 0° for both rotation directions, resembling the distribution found in the Stationary mode. Whereas Slow re-orientation exhibits a slight shift of the distribution of eye eccentricity toward the opposite direction of rotation. Therefore, although eye eccentricity can potentially induce a slight PSE shift, particularly in the Slow re-orientation mode, our findings for the Fast re-orientation mode do not support eye eccentricity as the primary cause of the perceptual disparity with our experimental conditions. Instead, as discussed above, we attribute this effect to the *head velocity relative to the stimulus*.

4 Applications

We present two main applications of our psychometric study. Firstly, we leverage the measured points of subjective equality (PSEs) to enhance target localization. We compensate for the perceptual disparity while rotating, enhancing users' target localization. Our approach has the potential to improve task performance, reduce response times, and enhance overall user engagement. Secondly, we propose guidelines for content creators to reduce or eliminate user confusion, since inconsistencies and misalignments between auditory and visual stimuli can lead to cognitive dissonance and negatively affect user experience [20, 23].

4.1 Enhancing VR Audio-Visual Target Localization

Target acquisition and localization are essential tasks in VR applications such as training, video watching, or gaming. In those scenarios, visual targets commonly appear with audio, which may be the primary distinctive cue (e.g., diegetic sound). However, our experimental results in Section 3 suggest that users' performance in localization may be suboptimal during re-orientation with rotation velocities relative to the audio-visual target (see Section 3.2). In such cases, sound localization is biased toward the direction of head rotation and it may produce spatial confusion, particularly in scenes with multiple targets. We present a proof-of-concept experiment demonstrating how our detection thresholds can improve VR users' target localization accuracy via guided spatial sound placement. This experiment also allows us to assess our measured PSEs under more complex audio-visual scenarios.

Participants Fourteen participants were recruited. Six identified themselves as male and seven as female, and one preferred not to disclose their gender (ages 20 - 57). All participants reported normal or corrected-to-normal vision and hearing, and filled an informed consent. Participants were naive to the aim of the experiment. The research protocol was approved by the *Comité de Ética de la Investigación de la Comunidad Autónoma de Aragón (CEICA)*.

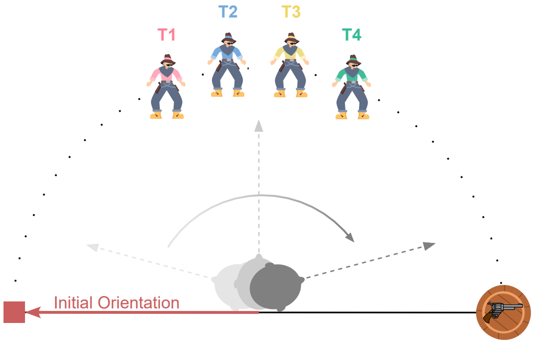

Fig. 6: *Application setup.* Participants rotate from the initial orientation 180° to take the pistol. While rotating, one of the four targets emits a sound. Once participants take the pistol, they are asked to select the visual target producing the sound (i.e., villain). The auditory stimulus activates when the visual targets enter the FoV during rotation and stop when they move out of the FoV, which happens when participants face the pistol.

Stimuli To simulate realistic application scenarios, we designed and conducted an experiment using a cartoonish-western video game. The game consists of a target identification task, wherein participants must accurately identify the villain amongst four possible characters (see Figure 6). The villain is distinguished from the null targets as the only one generating a reloading of a pistol sound, while the null-targets are purely visual. The four targets are presented at ten meters from the participant with a separation of 15° and a size of 5° of visual angle. The sound associated with the villain can appear either aligned or shifted to the left. We leverage the thresholds found in our previous study to compensate for the shift of the PSE. By placing the sound at the PSE (i.e., to the left of the visual stimulus when the participant is rotating to the right), participants will correctly associate the sound with the visual stimulus. Following a conservative approach, we use the PSE found for the Fast re-orientation mode (-13.54°), although participants are likely to move faster than 50°/s.

Procedure We focus on the right or clockwise rotation due to the high perceptual PSE shift. From an initial orientation of 0°, participants were instructed to rotate to the right to grab a pistol at 180° (see Figure 6). While they were rotating, the villain emitted a sound. Once they acquired the pistol, they were asked to select the target they thought was the villain. The villain can be either one of the four visual stimuli and his sound can be completely aligned or present the offset to the left. Each combination of target, and offset or non-offset is sampled five times. Therefore, our experiment consists of 40 trials, in which the fourteen participants repeated each offset and non-offset condition 20 times, resulting in 280 samples per condition. We ran a power analysis [9] which indicated that our sample size (N $>$ 12) is sufficient to detect large and medium effect sizes greater than 0.25 with a significant level of $\alpha = 0.05$. The order of the trial conditions was randomly selected. We included five additional trials to familiarize participants with the task. The entire experiment lasted ten minutes on average.

Results The participants' ability to accurately identify the villain among the non-target individuals is shown in Figure 5. Accuracy, defined as the median proportion of correct identifications, was 38% for the default target localization, indicating a poor localization of the visual stimulus associated with the audio when both are spatially aligned. However, when we placed the sound at the PSE location, thus rectifying the overestimation of sound in the rotation direction, the accuracy significantly increased, reaching 78% of correct answers. We perform the non-parametric Friedman test since our data does not follow a normal distribution (Shapiro-Wilk, $p < 0.05$), revealing a significant difference ($\chi^2 = 7.14$, $p < 0.05$) between conditions' accuracy. Notably, when the audio appears completely aligned with the villain, participants consistently selected the visual stimulus positioned to the right of the villain as the source of the sound (see default target localization in Figure 5). This behavior can be observed in the discernible shift towards the upper region of the diagonal in the confusion matrix presented in Figure 7, which shows the distribution of participants' answers for each visual stimulus for the four different possible villains. This finding confirms the presence of sound overestimation in the rotation direction, as observed in our previous study detailed on Section 3.2. On the other hand, when we place the auditory stimulus at the PSE, the distribution of participants' answers concentrates in the diagonal, indicating that the perceptual disparity has been compensated. The mean rotation velocity of participants was 140.52°/s, which as expected, is notably higher than the measured in our previous study. This speed can be attributed to the design of this application, where we did not impose any restrictions on the participants' rotation speed, mimicking real-world conditions where individuals naturally choose their movement speed. This showcases how even with the rough approximation of the PSE at 50°/s, we can achieve a significant increment in target localization accuracy on real-unconstrained applications.

Our results suggest that compensating for the perceptual audio-visual disparity is possible by placing the sound at the PSE. This allows for improving the target localization ability of users in VR applications. Moreover, no participant reported noticing alterations in audio placement, thus we achieved enhancing target localization while being completely imperceptible for participants. This proof-of-concept experiment further confirms the generalizability of our findings to natural and novel conditions, showcasing their potential to optimize existing VR applications with uncontrolled user head rotations.

4.2 Guidelines for Content Generation

Our findings regarding spatial disparities in audio-visual perception can guide the design of VR content. For instance, we can predict *confusion regions* in 360° scenes, where users may have difficulty identifying the origin of a sound when rotating their heads. This may create confusion when determining which visual stimulus is responsible for the sound, especially in complex and cluttered scenes, which in turn may lead to decreased engagement.

We illustrate this in Figure 8 for a promotional video of the popular game Clash of Clans obtained from YouTube[1]. We represent the psychometric curves fitted for our *Fast re-orientation* condition as a heatmap, where warmer colors indicate a higher likelihood of wrongly identifying the corresponding region as the sound source. The confusion region for both rotation directions with a head velocity of 50°/s

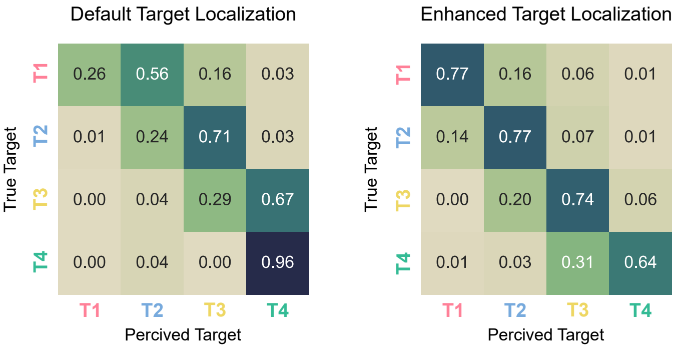

Fig. 7: Confusion matrix representing the distribution of participants' answers. **Left:** Default target localization in which the audio is completely aligned with the visual stimulus. **Right:** Enhanced target localization where audio presents our compensation offset. The x-axis represents the visual stimulus perceived by participants as the audio-visual stimulus, and the y-axis represents the actual audio-visual stimulus (i.e., villain). When no correction is applied, participants tend to select visual stimuli to the right of the actual audio-visual target. However, when we apply the offset (values at the diagonal), participants accurately identify the audio-visual target.

(mean head rotation velocity for re-orientation movements [42]) is illustrated by the pink areas overlayed on the frame, ranging from $-17.4°$ to $-1.9°$ and from $21.06°$ to $6.03°$ with respect to the sound source (see Table 1). Hence, any visual stimulus within that particular area has the potential to be linked as the source of the auditory stimulus.

In addition to the user rotating their head, other potential sources of audio-visual misalignments can hamper the VR experience, such as latency issues, transmission delays, or errors introduced during the encoding and decoding of the audio and video signals. The temporal synchronization of audio and visual signals is heavily influenced by these factors, and the presence of moving characters may result in spatial misalignments. Our experiments reveal the asymmetry of the permissible spatial errors between auditory and visual stimuli must not exceed $\approx 6°$ to the right of the visual stimulus, and $\approx 2°$ to the left. These conservative limits correspond to the detection thresholds observed in the *Stationary* mode, and thus apply even when users are static, looking straight ahead to an audio-visual stimulus.

5 Limitations and Future Work

While our research offers new insights regarding the spatial acuity of VR users during natural head-body rotations, there are still potential areas for future research and improvement. Currently, our psychometric models focus exclusively on spatial aspects and do not account for semantic factors. We briefly examined edge parameters for audio volume, bandwidth, and contrast level during pilot studies, and the effect appears consistent across various conditions. Our enhanced target localization application confirms consistent effects across various environments, using more complex stimuli than those employed during our measurements. Nevertheless, future work should explore how these different aspects may alter spatial perception. Due to the need for tractability in our experiments, we sampled a sparse set of three velocities to study the impact of rotation speed, revealing a linear increase in the PSE shift. However, it is also possible that the perceptual shift observed during rotations does not continue to increase indefinitely, but instead reaches a maximum value. Further investigations should therefore focus on refining the model to capture the potential non-linear relationship and/or identify the upper limit of the perceptual shift.

Although we do not explore a dynamically adaptive offset compensation, future work could investigate the dynamic adjustment of the offset based on head rotation velocity. This would require a preliminary user study to assess if the effect is persistent under this condition, and a head velocity predictor to estimate the next rotation velocity to apply the correct offset. Nevertheless, we show how even without dynamically adapting the offset, we already mitigate the perceived

[1] https://www.youtube.com/watch?v=yVLfEHXQk08

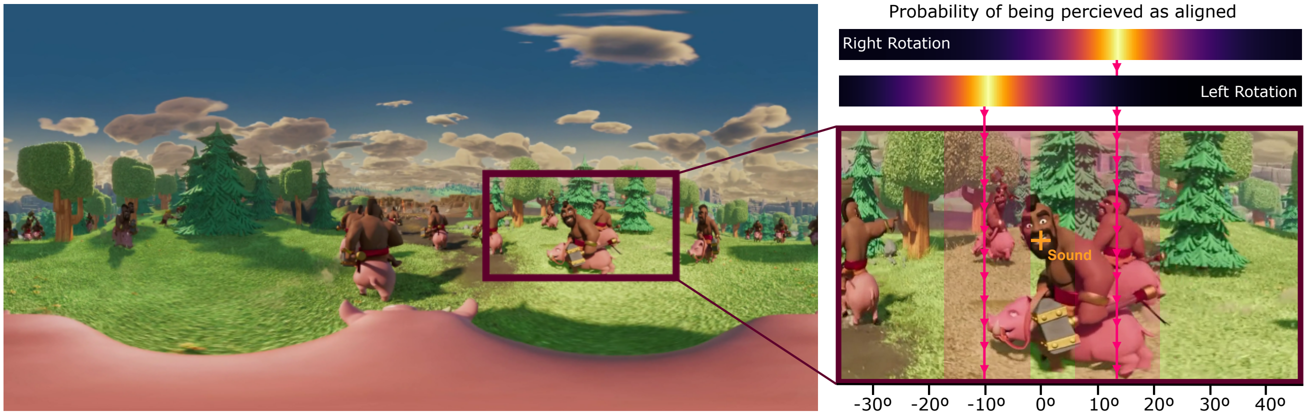

Fig. 8: *Guidelines for content generation.* **Left:** Frame (equirectangular projection) of a 360° promotional video of the popular game Clash of Clans. **Right:** Close-up view of the region marked on the frame. The orange cross indicates the source of the sound emitted by the depicted character. The probability of actually perceiving the character as the source of the sound according to our measured psyschometric functions is represented as a heatmap on top for both left and right rotation directions. Warmer colors indicate a higher probability and vertical dashed lines (dark pink) display the PSEs. All other characters near the dashed lines are likely to be wrongly perceived as the source of the sound. See Table 1 for exact PSE and DT values.

audio-visual mismatch in unconstrained scenarios. Furthermore, exploring whether training can partially mitigate this perceived offset presents an intriguing avenue for research. Prior studies indicate that the brain may exhibit a degree of plasticity, potentially enabling it to adapt to spatial misalignments between audio and visual stimuli [2, 30].

Additionally, our current model is established by assuming individual targets and isolated movements. However, real-world situations frequently involve users performing complex actions that encompass multiple targets or continuous movement sequences. Therefore, an essential future avenue is to broaden our measurement to those scenarios. One potential direction is to utilize multiple points of reference to measure and minimize the corresponding perceptual discrepancies, such as accounting for spatial triangulation [11]. Furthermore, our research opens the possibility of personalized psychometric functions that consider individual differences. By incorporating a calibration step into VR applications, it becomes possible to tailor the experience to each user's specific perceptual characteristics, leading to individualized granularity in spatial acuity.

6 Conclusion

In this work, we have focused on measuring the spatial localization acuity of audio-visual content during head-body rotations in VR environments. While some previous works have studied audio-visual mismatches in the context of binding effects [5], we attempt to develop the first model for directional acuity localization in VR users. Our study addresses an essential gap in the VR literature by investigating the relationship between audio and vision during target localization tasks involving natural head-body rotations. This has practical implications for VR applications where users encounter dynamic audio-visual environments. We study the impact of different viewing modes, closely resembling common movements in VR applications. To our knowledge, we are the first to explore this effect across these viewing modes, providing valuable insights into the integration of auditory and visual cues under natural head-body rotations. By consolidating our experiments incorporating natural head-body rotations, we gain a deeper understanding of how active movements, which stimulate the vestibular system and trigger strong proprioceptive signals, can impact the perceived spatialization of audio-visual stimulus. Through our experiments, we have found for the first time a consistent overestimation of the auditory stimulus toward the rotation direction. This perceptual shift phenomenon is significantly influenced by the relative motion between the target and users, highlighting the relevance of considering dynamic factors in spatial perception. Therefore we systematically characterize how target localization accuracy is influenced by perceptual audio-visual disparities related to stimuli-head relative motion, a novel insight in the field. Leveraging VR technology, we ensure control and reproducibility to study perception in ecologically valid environments. Further, previous works even suggest that audio-visual localization effects in real environments may also be reflected in VR, making it a powerful tool for exploring real-world perception. We aim to establish a valuable foundation by reporting and measuring this perceptual audio-visual discrepancy across various typical motions and rotation speeds in VR.

Looking ahead, we envision that our findings pave the way for future exciting research to advance personalized spatial perception in complex and multisensory virtual environments. We hope to take a step closer to the ultimate goal that users with VR/AR may surpass our performance limits even in the physical world, with the co-pilot of computational optimization of virtual content. To this end, our data, analysis, and code are available at https://graphics.unizar.es/projects/AV_spatial_perception/.

Acknowledgments

We extend our gratitude to the members of the Graphics and Imaging Lab for their support and collaboration. We would also like to thank the anonymous reviewers for their insightful comments and the participants in the experiment. We acknowledge Freepik.com for the design of the cowboy character used for the target localization application. Our work has received funding from the European Union's Horizon 2020 research and innovation program, through the CHAMELEON project (European Research Council, grant agreement No. 682080), the Spanish Agencia Estatal de Investigación (Project PID2022-141539NB-I00 funded by MCIN/AEI/10.13039/501100011033/FEDER, EU), and the National Science Foundation (NSF) grants #2225861 and #2232817. Additionally, Edurne Bernal-Berdun was supported by a Gobierno de Aragon predoctoral grant (2021–2025).

References

[1] M. D. Arnoult. Post-rotatory localization of sound. *The American Journal of Psychology*, 63(2):229–236, 1950. 2

[2] C. C. Berger, M. Gonzalez-Franco, A. Tajadura-Jiménez, D. Florencio, and Z. Zhang. Generic hrtfs may be good enough in virtual reality. improving source localization through cross-modal plasticity. *Frontiers in Neuroscience*, p. 21, 2018. 8

[3] D. Berghi, H. Stenzel, M. Volino, A. Hilton, and J. Philip Jackson. Audio-visual spatial alignment requirements of central and peripheral object events. In *IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops*, pp. 666–667, 2020. 2, 5

[4] R. H. Carpenter. *Movements of the Eyes, 2nd Rev.* 1988. 2

[5] L. Chen and J. Vroomen. Intersensory binding across space and time: a tutorial review. *Attention, Perception, & Psychophysics*, 75:790–811, 2013. 2, 8

[6] M. L. Chénéchal and J. C. Goldman. HTC Vive Pro Time Performance Benchmark for Scientific Research. In *International Conference on Artificial Reality and Telexistence and Eurographics Symposium on Virtual Environments*, 2018. 2

[7] A. Dufour, P. Touzalin, and V. Candas. Rightward shift of the auditory subjective straight ahead in right- and left-handed subjects. *Neuropsychologia*, 45(2):447–453, 2007. 5

[8] B. Duinkharjav, K. Chen, A. Tyagi, J. He, Y. Zhu, and Q. Sun. Color-perception-guided display power reduction for virtual reality. *ACM Transactions on Graphics (TOG)*, 41(6):1–16, 2022. 3

[9] E. Erdfelder, F. Faul, and A. Buchner. Gpower: A general power analysis program. *Behavior Research Methods, Instruments, & Computers*, 28:1–11, 1996. 7

[10] T. Freeman, J. Culling, M. Akeroyd, and W. Brimijoin. Auditory compensation for head rotation is incomplete. *Journal of Experimental Psychology: Human Perception and Performance*, 43, 11 2016. 2

[11] S. S. Fukusima, J. M. Loomis, and J. A. Da Silva. Visual perception of egocentric distance as assessed by triangulation. *Journal of experimental psychology: Human perception and performance*, 23(1):86, 1997. 8

[12] P. Gao, K. Matsumoto, T. Narumi, and M. Hirose. Visual-auditory redirection: Multimodal integration of incongruent visual and auditory cues for redirected walking. In *IEEE International Symposium on Mixed and Augmented Reality*, pp. 639–648, 2020. 2

[13] D. Genzel, M. Schutte, W. O. Brimijoin, P. R. MacNeilage, and L. Wiegrebe. Psychophysical evidence for auditory motion parallax. *Proc. Nat. Acad. Sci. U. S. A.*, 115(16):4264–4269, 2018. 2, 3

[14] D. W. Grantham, B. W. Y. Hornsby, and E. A. Erpenbeck. Auditory spatial resolution in horizontal, vertical, and diagonal planes. *The Journal of the Acoustical Society of America*, 114(2):1009–1022, 07 2003. 2, 3, 5

[15] Z. Hu, A. Bulling, S. Li, and G. Wang. Ehtask: Recognizing user tasks from eye and head movements in immersive virtual reality. *IEEE Transactions on Visualization and Computer Graphics*, 29(4):1992–2004, 2023. 3

[16] H. Huang, M. Solah, D. Li, and L.-F. Yu. Audible panorama: Automatic spatial audio generation for panorama imagery. In *In CHI Conference on Human Factors in Computing Systems*, pp. 1–11, 2019. 1

[17] T. Iachini, L. Maffei, F. Ruotolo, V. P. Senese, G. Ruggiero, M. Masullo, and N. Alekseeva. Multisensory assessment of acoustic comfort aboard metros: a virtual reality study. *Applied Cognitive Psychology*, 26(5):757–767, 2012. 1

[18] E. Kim and G. Shin. User discomfort while using a virtual reality headset as a personal viewing system for text-intensive office tasks. *Ergonomics*, 64(7):891–899, 2021. 1

[19] H. Kim and I.-K. Lee. Studying the effects of congruence of auditory and visual stimuli on virtual reality experiences. *IEEE Trans. on Visualization and Computer Graphics*, 28(5):2080–2090, 2022. 2, 5

[20] H. Kim, L. Remaggi, P. J. Jackson, and A. Hilton. Immersive spatial audio reproduction for vr/ar using room acoustic modelling from 360° images. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 120–126, 2019. 6

[21] B. Krajancich, P. Kellnhofer, and G. Wetzstein. A perceptual model for eccentricity-dependent spatio-temporal flicker fusion and its applications to foveated graphics. *ACM Transactions on Graphics (TOG)*, 40(4):1–11, 2021. 3

[22] M. Kytö, K. Kusumoto, and P. Oittinen. The ventriloquist effect in augmented reality. In *IEEE International Symposium on Mixed and Augmented Reality*, pp. 49–53, 2015. 2, 4, 5

[23] P. Larsson, A. Väljamäe, D. Västfjäll, A. Tajadura-Jiménez, and M. Kleiner. Auditory-induced presence in mixed reality environments and related technology. *The engineering of mixed reality systems*, pp. 143–163, 2010. 6

[24] G. Lester and R. Morant. The role of the felt position of the head in the audiogyral illusion. *Acta Psychologica*, 31:375–384, 1969. 2

[25] J. Leung, D. Alais, and S. Carlile. Compression of auditory space during rapid head turns. *Proc. Nat. Acad. Sci. U. S. A.*, 105(17):6492–6497, 2008. 1, 2

[26] J. Lewald and H.-O. Karnath. Sound lateralization during passive whole-body rotation. *European Journal of Neuroscience*, 13(12):2268–2272, 2001. 2

[27] M. Melo, G. Gonçalves, P. Monteiro, H. Coelho, J. Vasconcelos-Raposo, and M. Bessa. Do multisensory stimuli benefit the virtual reality experience? a systematic review. *IEEE Trans. on Visualization and Computer Graphics*, 28(2):1428–1442, 2020. 1

[28] M. Narbutt, S. O'Leary, A. Allen, J. Skoglund, and A. Hines. Streaming vr for immersion: Quality aspects of compressed spatial audio. In *International Conference on Virtual System & Multimedia*, pp. 1–6, 2017. 2

[29] A. R. Nidiffer, A. Diederich, R. Ramachandran, and M. T. Wallace. Multisensory perception reflects individual differences in processing temporal correlations. *Scientific Reports*, 8(1):14483, 2018. 1

[30] B. Odegaard, D. R. Wozny, and L. Shams. A simple and efficient method to enhance audiovisual binding tendencies. *PeerJ*, 5, 2017. 8

[31] D. R. Perrott and K. Saberi. Minimum audible angle thresholds for sources varying in both elevation and azimuth. *The Journal of the Acoustical Society of America*, 87(4):1728–1731, 04 1990. 2, 5

[32] M. I. Posner, M. J. Nissen, and R. M. Klein. Visual dominance: an information-processing account of its origins and significance. *Psychological Review*, 83(2):157, 1976. 2

[33] T. Potter, Z. Cvetkovic, and E. De Sena. On the relative importance of visual and spatial audio rendering on vr immersion. *Frontiers in Signal Processing*, 2022. 2

[34] I. Poupyrev, T. Ichikawa, S. Weghorst, and M. Billinghurst. Egocentric object manipulation in virtual environments: empirical evaluation of interaction techniques. In *Computer Graphics Forum*, vol. 17, pp. 41–52, 1998. 1

[35] J. Prinz. Is the mind really modular. *Contemporary Debates in Cognitive Science*, 14:22–36, 2006. 1

[36] J. Ross, M. C. Morrone, and D. C. Burr. Compression of visual space before saccades. *Nature*, 386(6625):598–601, 1997. 2

[37] H. H. Schütt, S. Harmeling, J. H. Macke, and F. A. Wichmann. Painfree and accurate bayesian estimation of psychometric functions for (potentially) overdispersed data. *Vision Research*, 122:105–123, 2016. 4

[38] A. R. Seitz, R. Kim, and L. Shams. Sound facilitates visual learning. *Current Biology*, 16(14):1422–1427, 2006. 1

[39] A. Serrano, D. Martin, D. Gutierrez, K. Myszkowski, and B. Masia. Imperceptible manipulation of lateral camera motion for improved virtual reality applications. *ACM Trans. on Graph.*, 39(6), 2020. 3, 4

[40] L. Shams and R. Kim. Crossmodal influences on visual perception. *Physics of Life Reviews*, 7(3):269–284, 2010. 1

[41] L. Shams and A. R. Seitz. Benefits of multisensory learning. *Trends in Cognitive Sciences*, 12(11):411–417, 2008. 1

[42] V. Sitzmann, A. Serrano, A. Pavel, M. Agrawala, D. Gutierrez, B. Masia, and G. Wetzstein. Saliency in vr: How do people explore virtual environments? *IEEE Trans. on Visualization and Computer Graphics*, 24(4):1633–1642, 2018. 3, 7

[43] Y. Sosa, W. A. Teder-Sälejärvi, and M. E. McCourt. Biases of spatial attention in vision and audition. *Brain and Cognition*, 73(3):229–235, 2010. 5

[44] H. Stenzel and P. j. b. Jackson. Perceptual thresholds of audio-visual spatial coherence for a variety of audio-visual objects. *Journal of the Audio Engineering Society*, 2018. 2

[45] W. R. Thurlow and T. P. Kerr. Effect of a moving visual environment on localization of sound. *The American Journal of Psychology*, pp. 112–118, 1970. 2

[46] D. C. P. B. M. Van Barneveld and A. John Van Opstal. Eye position determines audiovestibular integration during whole-body rotation. *European Journal of Neuroscience*, 31(5):920–930, 2010. 2, 5

[47] A. Vatakis and C. Spence. Audiovisual synchrony perception for music, speech, and object actions. *Brain Research*, 1111(1):134–142, 2006. 3

[48] M. Wexler and J. J. van Boxtel. Depth perception by the active observer. *Trends in Cognitive Sciences*, 9(9):431–438, 2005. 3

[49] A. Yoonessi and C. L. Baker. Contribution of motion parallax to segmentation and depth perception. *Journal of vision*, 11 9:13, 2011. 3

[50] J. Zhang, E. Langbehn, D. Krupke, N. Katzakis, and F. Steinicke. Detection thresholds for rotation and translation gains in 360 video-based telepresence systems. *IEEE transactions on visualization and computer graphics*, 24(4):1671–1680, 2018. 3