Overdriving Visual Depth Perception via Sound Modulation in VR

Daniel Jiménez-Navarro, Colin Groth, Xi Peng, Jorge Pina, Qi Sun, Praneeth Chakravarthula, Karol Myszkowski, Hans-Peter Seidel, and Ana Serrano

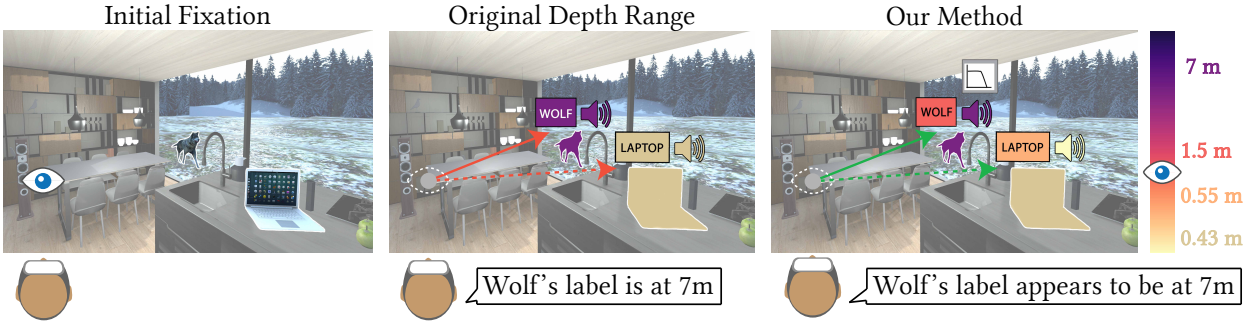

Fig. 1: This paper investigates the effect of spatially decoupling auditory and visual cues on depth perception in virtual reality (VR). *Left*: The user begins by fixating on some random object within a VR scene. *Center*: A new annotation (e.g., related to the wolf) appears in the distance, accompanied by colocated audio (indicated by the same purple color). Visual response and fixation on the wolf annotation requires a significant eye vergence angle change ($\Delta\theta$). *Right*: Using our method, the disparity variation of the visual annotation relative to the initial fixation point is reduced, thereby decreasing $\Delta\theta$. To compensate for this visual depth reduction, we manipulate the sound's depth cues to increase perceived distance. Due to the fusion of visual and auditory cues, the user perceives the audiovisual annotation shifted toward the wolf's depth. In contrast, for the laptop case, the visual reduction ($\Delta\theta$) is now compensated by adjusting the sound's depth cues to reduce perceived distance. Note how different strategies are applied depending on the visual motion performed: sound source spatial location is altered for the laptop (convergence eye movement), while the sound frequency spectrum is modified for the wolf (divergence eye movement). In both scenarios, our method preserves accurate depth perception while reducing $\Delta\theta$, which accelerates gaze retargeting (illustrated with green arrows). The right bar visualizes physical depth.

Abstract—Our ability to perceive and navigate the spatial world is a cornerstone of human experience, relying on the integration of visual and auditory cues to form a coherent sense of depth and distance. In stereoscopic 3D vision, depth perception requires fixation of both eyes on a target object, which is achieved through vergence movements, with convergence for near objects and divergence for distant ones. In contrast, auditory cues provide complementary depth information through variations in loudness, interaural differences (IAD), and the frequency spectrum. We investigate the interaction between visual and auditory cues and examine how contradictory auditory information can overdrive visual depth perception in virtual reality (VR). When a new visual target appears, we introduce a spatial discrepancy between the visual and auditory cues: the visual target is shifted closer to the previously fixated object, while the corresponding sound localization is displaced in the opposite direction. By integrating these conflicting cues through multimodal processing, the resulting percept is biased toward the intended depth location. This audiovisual fusion counteracts depth compression, thus reducing the required vergence magnitude and enabling faster gaze retargeting. Such audio-driven depth enhancement may further help mitigate the vergence–accommodation conflict (VAC) in scenarios where physical depth must be compressed. In a series of psychophysical studies, we first assess the efficiency of depth overdriving for various VR-relevant combinations of initial fixations and shifted target locations, considering different scenarios of audio displacements and their loudness and frequency parameters. Next, we quantify the resulting speedup in gaze retargeting for target shifts that can be successfully overdriven by sound manipulations. Finally, we apply our method in a naturalistic VR scenario where user interface interactions with the scene show an extended perceptual depth.

Index Terms—Multimodal perception, audiovisual integration, depth estimation, graphics application

- *Daniel Jiménez-Navarro is with Max Planck Institute for Informatics. E-mail: djimenez@mpi-inf.mpg.de*
- *Colin Groth is with New York University. E-mail: c.groth@nyu.edu*
- *Xi Peng is with University of North Carolina. E-mail: xipeng@cs.unc.edu*
- *Jorge Pina is with University of Zaragoza. E-mail: jpina@unizar.es*
- *Qi Sun is with New York University. E-mail: qisun@nyu.edu*
- *Praneeth Chakravarthula is with University of North Carolina. E-mail: cpk@cs.unc.edu*
- *Karol Myszkowski is with Max Planck Institute for Informatics E-mail: karol@mpi-inf.mpg.de*
- *Hans-Peter Seidel is with Max Planck Institute for Informatics E-mail: hpseidel@mpi-inf.mpg.de*
- *Ana Serrano is with University of Zaragoza E-mail: anase@unizar.es*

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

1 Introduction

Virtual reality (VR) devices support for multimodal content shows how the combination of visual, auditory, and sometimes haptic cues, also referred to as multisensory integration, enhances realism and improves the perception of our surroundings [23, 46]. Focusing on audio and visual modalities, a higher sense of immersion and presence can be achieved when sound is combined with visual content [42, 49, 64]. Benefits of audiovisual content are also reported in guiding navigation by drawing attention to off-screen events [38, 64], supporting interactivity by reinforcing feedback from user actions [42], and promoting stress relief in immersive interventions [49]. In such cases, audio does not merely complement visuals but actively shapes perception and behavior.

Multimodal sensory data, both visual and auditory, is also unconsciously integrated to perceive depth in the physical world. Audiovisual depth cues improve our ability to estimate distances and accurately interpret spatial layouts [4, 11, 16]. In virtual content, visual-based depth perception in stereoscopic displays has been shown to be influenced by real-world auditory sources such as loudspeakers [61]. For fully

immersive setups (e.g., AR), complementing visual depth cues with depth-scaled sound spatialization has been demonstrated to enhance relative depth perception and reduce task completion time [71].

In this work, we investigate *misplaced* (non-colocated) visual and auditory cues in VR and show that, when such conflicting cues are fused, the perceived depth can be overdriven from its visually defined position toward the sound's spatial location. We focus on newly appearing targets, which typically trigger a gaze shift from the current fixation, involving an ocular movement sequence that combines a saccade to the new target's screen position and vergence to its stereoscopic depth (Fig. 1). We experiment with reducing the depth change between the current fixation and the new target by shifting the target along its line of sight without changing its screen position (e.g., the annotations in Fig. 1). The resulting depth compression may be desirable, as it helps control potentially uncomfortable depth changes and enables faster reaction times by reducing the magnitude of vergence required.

To compensate for the compressed depth, we displace the corresponding sound location in the opposite direction with respect to the depth manipulation. We manipulate the sound sources depth by adapting their loudness and spectral frequency content accordingly, but we also displace sound laterally to act on interaural differences (IAD). Across all investigated conditions, we ensure that visual and auditory cues remain integrable, allowing them to fuse into a consistent percept that is biased toward the intended target depth location. In this context, the audiovisual fusion mechanism offers an alternative means of supporting perceived depth in scenarios where depth accuracy is reduced as a consequence of the measures required to alleviate the vergence–accommodation conflict (VAC) [25].

We design a perceptual experiment to systematically assess the efficiency of such perceived depth overdriving for various VR-relevant combinations of initial fixations and shifted target object locations under the described audio condition variations. Our results show that, if optimally designed, success probability in judging depth comparisons can decrease from about 80% with visual-only cues towards chance level (50%) with conflicting overdriving audio. Based on the outcome of the first experiment, we select the most successful scenarios of depth overdriving and demonstrate up to 50 ms speedup in gaze retargeting. Finally, we demonstrate the benefits of sound-based depth perception enhancement for dynamic interfaces in natural VR applications.

Our contributions throughout this work include:

- We show that non-colocated auditory and visual cues can be used to manipulate perceived depth, and we provide guidelines for controlling this effect in VR.
- For response-time–critical scenarios, our approach enables faster visual reactions while the underlying vergence reduction manipulation is masked by overdriving sound.
- We demonstrate practical applications of the sound overdriving method in a naturalistic VR application, with consistent results further shown in a proof-of-concept mixed reality (MR) version.

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

2 Related Work

In this section, we discuss visual depth cues, particularly for stereovision (Sec. 2.1), their limitations in VR (Sec. 2.2), the role of auditory cues in depth perception (Sec. 2.3), and finally the integration of both audio and stereovision (Sec. 2.4).

2.1 Depth Perception and Eye Movements

Human depth perception relies on both monocular and binocular cues [13]. Monocular cues such as perspective, occlusion, shading, and motion parallax convey depth through experience and context [51], while binocular stereopsis provides a direct sense of depth from retinal disparities [26]. Whereas monocular cues benefit from familiarity, stereopsis functions as a low-level, pre-attentive mechanism independent of prior knowledge [13, 26]. Depth cues may conflict, but the brain resolves inconsistencies through cue fusion, modeled as strong fusion (winner-takes-all), weak fusion (Maximum Likelihood Estimation weighting by cue reliability), or modified weak fusion (Bayesian inference with priors) [34, 40].

We focus on stereovision, where saccades and vergence are the primary eye movement mechanisms during free scene exploration. Saccades are rapid, ballistic movements that redirect gaze, preceded by a 200 ms latency and lasting 20–100 ms depending on amplitude [1, 3, 62]. Vergence involves opposite-direction eye rotations driven by depth changes between fixations, with latencies of 180–250 ms and rotations lasting 200–1000 ms depending on depth magnitude [55, 60]. In scenarios with fixation shifts at different depths, combined saccade–vergence movements can occur, producing up to 20% vergence speedup at the cost of saccades slowing by up to 50% [12, 15, 65, 66, 70].

2.2 Limitations of Stereovision in VR

Standard VR displays lack the eye accommodation cue, leading to the vergence–accommodation conflict (VAC) [25, 56]. This mismatch between accommodation, set by the display distance, and vergence, aligned with the fused object, causes discomfort [21, 25, 56] and prolongs vergence duration [58]. A common mitigation strategy compresses scene depth by manipulating disparity, i.e., the horizontal pixel shift between left- and right-eye images [14, 22, 41]. However, disparity manipulation reduces perceived depth and can conflict with other monocular depth cues. To address excessive disparity compression and VAC, fixated regions can be shifted closer to the display while expanding the disparity range around the display (zero disparity) [5, 30, 31, 50]. For controllable elements such as annotations, disparity changes relative to fixation should be minimized, and when re-vergence is unavoidable, inducing saccades can reduce vergence duration [15]. A similar issue arises in stereoscopic film cuts, where depth differences at likely fixation points are reduced in post-production [60].

Since visual adjustments alone may not resolve excessive depth compression, we explore auditory depth cues. By shifting a new visual target closer to fixation while displacing the associated audio source in the opposite direction, we overdrive depth perception and reduce the vergence demand.

2.3 Depth Perception in Human Auditory System

Audio is an important cue for localizing the distance of audiovisual targets, providing semantic information and spatial cues such as loudness, interaural differences (IAD), frequency spectrum, and reverberation [68]. Loudness is usually the primary cue, being interpreted based on contextual comparisons rather than an inherent scale of sound intensity [59]. In addition, interaural time differences (ITD) and interaural level differences (ILD), the components of IAD, jointly contribute to spatial localization but are less effective for absolute depth estimation [9]. Moreover, the frequency spectrum changes with distance, as high frequencies attenuate more than low frequencies due to air absorption [44]. Reverberation also serves as an additional cue, where closer sounds have a higher direct-to-reverberant ratio and distant sounds contain more reflected components [8]. Ultimately, these cues are integrated in the brain to support more precise localization of sound sources in 3D space.

2.4 Depth Perception with Multimodal Cues

Colocated audio and visual cues significantly improve location judgment and reaction time compared to visual-only conditions [7, 20, 33], an effect also confirmed for depth-scaled 3D sound in AR environments [71]. The focus of this work, however, is on the integration of *non-colocated* audio and visual cues. A prominent example is the ventriloquism effect, or "visual capture," in which spatially misaligned auditory cues are perceptually pulled toward the visual signal in a winner-takes-all manner. Importantly, such audio–visual integration is constrained by the limits of tolerable misalignment, both in the angular domain, as directional separation between cues [18, 27], and in the spatial domain, as distance misalignment [19, 67]. More generally, misaligned cues interact to produce a perceived cross-modal location, analogous to visual depth cue fusion [6, 34, 40, 61] (Sec. 2.1), with visual capture [27], weak fusion [11], and Bayesian models accounting

for visual bias [4, 16] describing aspects of audiovisual depth integration. While feedback-driven recalibration can adjust perceived depth via cue reweighting [37, 39, 43], this mechanism is outside the scope of the present work, which focuses on fast-performance applications without feedback.

The impact of temporal and spatial misalignment between auditory and visual cues has been studied in monoscopic and stereoscopic displays [29, 54, 63]. Audio preceding visual onset can reduce saccade latency and enhance performance in time-critical tasks [28]. Closest to our goals is the work of Turner et al. [61], who demonstrated that spatially misaligned audio can bias perceived depth. Using a stereo 3D desktop display with a pair of loudspeakers positioned at different distances to produce a stereo effect, they found that sounds from the closer pair caused the stereoscopic display's apparent depth to be perceived as shifted toward that speaker location.

In our work, we employ a VR setup with headphone-based sound spatialization to systematically investigate how loudness, IAD, and spectral content can overdrive visual depth perception across different vergence angles and eccentricities during free scene exploration while combining saccade and vergence movements.

3 Method

Our goal is to study how sound cues dislocated from their visual counterparts can overdrive stereoscopic depth perception. To this end, our methodology exploits manipulations of sound loudness based on spatial shifts, IAD, and frequency spectrum, with different procedures for convergence and divergence eye movements.

Theoretical Background Key components in visual depth perception are stereoscopic disparity and eye vergence [48]. The *vergence angle* θ, is defined by the convergence of both eyes' gaze rays on a fixation point. For an object at distance d, it is given by $\theta = 2 \cdot \arctan(\mathrm{IPD}/2d)$, where IPD is the interpupillary distance. At close range, even small changes in distance result in significant vergence shifts. Objects positioned too close to the observer cannot be fused, leading to diplopia, while objects far away make gaze rays parallel, limiting the vergence range. On the other hand, audio-driven distance perception is primarily influenced by sound intensity that reaches the listener I, which follows the inverse square law $I = I_0/4\pi d^2$ where I_0 is the sound source intensity and d the distance between both. However, humans perceive intensity differences as relative changes [59], denoted as I_r in decibels (dB):

$$I_\mathrm{r} = 20 \cdot \log_{10}\left(\frac{R}{R_\mathrm{ref}}\right) \tag{1}$$

where R is the distance of the sound location and R_ref the reference distance.

Hypothesis We hypothesize that sound-driven depth cues can become strong enough to overdrive contradictory visual depth perception arising from disparity and vergence. Under this assumption, we deliberately move the newly appearing visual target closer to a previous fixation point to reduce the required vergence adjustment. We refer to this newly appearing target as $\mathbf{T_{test}}$ (Fig. 2). Our technique then aims to reproduce the depth percept associated with the full vergence change for the originally intended target ($\mathbf{T_{ref}}$) by manipulating the auditory cues of $\mathbf{T_{test}}$. To achieve the intended stereoscopic depth overdriving, we build on prior work [59, 68] that extensively analyzes the effectiveness of relative differences in auditory depth cues. For sound sources located close to the listener, distance-based loudness (Eq. 1) and interaural differences (IAD) are the most effective depth cues [35]. This situation corresponds to the convergence eye movements, in which the visual target approaches the observer, and our overdriving strategy places the sound source even closer. In Sec. 3.1, we discuss the convergence motion condition, where, along with distance-based loudness modulation (Eq. 1), we appropriately manipulate the spatial location of the audio source to enhance IAD (Fig. 2, left). At larger sound source distances, however, distance-based loudness (Eq. 1) and IAD cues become progressively less effective, making them unsuitable for divergence eye movements, where the distance to the visual target increases. In this case, our method aims for the fused audiovisual cue to be perceived as even farther away, reflecting a behavior distinct from the convergence condition. Therefore, we replace these cues with spectral cues, which are known to remain informative at greater distances [35]. In Sec. 3.2, we detail how low-pass filtering of the sound frequency spectrum (Fig. 2, right) produces frequency-based loudness variations that are employed in the divergence motion condition.

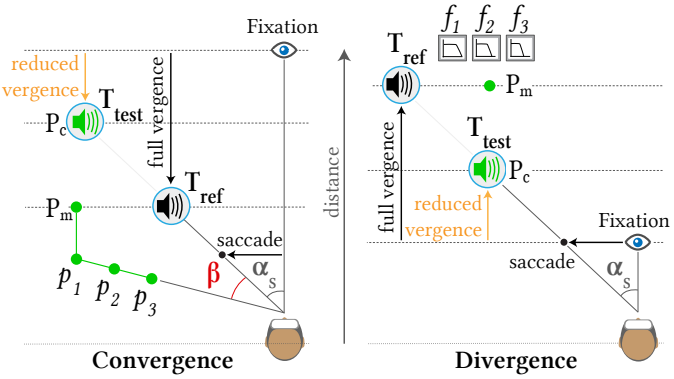

Fig. 2: Illustration of our method. Given the current fixation, the originally intended location ($\mathbf{T_{ref}}$) requires a large vergence change while a reduced change is required for our proposed visual location ($\mathbf{T_{test}}$). For convergence (left), we strengthen spatial-based loudness variations with IAD by detaching the audio signal from the visual target. For divergence (right), we consider a static spatial offset between sound and visual cues and purely alter the auditory frequency spectrum. In both cases, the magnitude of the required vergence movement can be reduced due to the closer visual target (orange arrow) from fixation. Labels $\mathrm{P_c}$, $\mathrm{P_m}$, p_{1-3} and f_{1-3} indicate the different loudness variations for $\mathbf{T_{test}}$ tested in the main experiment in Sec. 4.

3.1 Sound Manipulation for Eye Convergence

In the convergence motion condition, we aim to perceive $\mathbf{T_{test}}$ closer towards $\mathbf{T_{ref}}$'s depth. The simplest scenario to be tried is to shift $\mathbf{T_{test}}$'s sound source closer to the observer along the line of sight. In that case, the sound intensity naturally increases according to Eq. 1. However, to further enhance the overdriving effect, we also introduce lateral shifts, separating the audio source from the line of sight towards the periphery, which induce strong IAD cues. Acting on IAD alters the perceived distance of lateral audiovisual stimuli, replicating how sound behaves when objects are closer [24]. Our approach is summarized in Fig. 2, left. However, to maintain audiovisual coherence, the angular (lateral) offset between visual and auditory cues (β) must be restricted to the fusion limit $\beta_\mathrm{max} = 16°$ to prevent multimodal integration from breaking down into distinct auditory and visual stimuli [18]. For example, in Fig. 2, left, the positions p_{1-3} are constrained to the maximum lateral shift $\beta = \beta_\mathrm{max}$, thus the audio source can only move toward the observer.

3.2 Sound Manipulation for Eye Divergence

On the other hand, in the divergence motion condition, we aim to perceive $\mathbf{T_{test}}$ farther away towards $\mathbf{T_{ref}}$'s depth. Since distant sounds are characterized by a higher proportion of low-frequency content due to atmospheric absorption [44], this effect is replicated by applying a low-pass filter to the sound signal, which naturally reduces audio loudness. In this case, loudness variations based only on spatial shifts would end up breaking the spatial limits of audiovisual fusion due to the logarithmic nature of sound loudness variations (Eq. 1) [19]. Consequently, the filtered sound is always presented matching $\mathbf{T_{ref}}$'s depth ($\mathrm{P_m}$) (see Fig. 2, right). For even a further intended perceived depth, we intensify the filtering effect by modifying the cut-off frequency (f_1, f_2, f_3). For example, for the beep-like audio signal centered at 880 Hz used in Sec. 4, the cut-off frequencies are selected within the fifth octave (740 or 660 Hz) to obtain the desired loudness variations. On the lower end, cut-off frequencies with excessive signal degradation, which causes the sound to be barely audible, are prevented. The highest

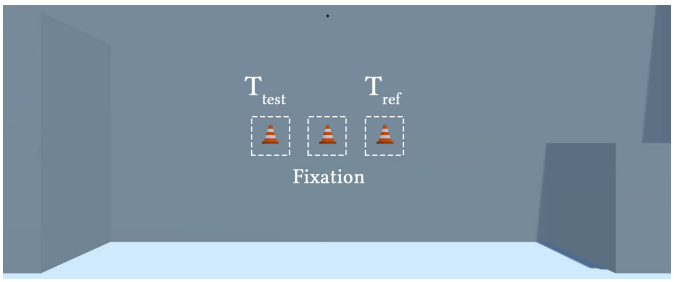

Fig. 3: User point of view during the main user study. Visual stimuli (traffic cones) appeared sequentially and only one was visible at the same time (dashed lines). All of them (Fixation, $\mathbf{T_{ref}}$ and $\mathbf{T_{test}}$) have their size increased by a factor of 4 for visualization purposes. In this case, Fixation is placed at 0.8 m, $\mathbf{T_{ref}}$ is at 1.5 m and $\mathbf{T_{test}}$ is at 7 m, corresponding to Divergence-Far condition (Tab. 1). Note that the stimulus size is fixed irrespective of distance to eliminate undesired monocular depth cues.

cut-off frequency is set to 880 Hz, yielding a loudness reduction of 3 dB (f_1), which is comparable to the p_1 sound overdriving level (Fig. 2; left) in the convergence case (Sec. 3.1). A more detailed discussion of the specific audio settings and conditions used in evaluating depth perception overdriving is provided in Sec. 4.

4 Sound-Based Depth Perception Modulation

In this section, we present our main psychophysical experiment, which examines how modifications in auditory cues influence visual depth perception. Applying the approach outlined in Sec. 3, we determine the levels of audio signal modulation that conceal visual depth shifts while maintaining audiovisual fusion.

4.1 Experimental Setup

4.1.1 Hardware

Experiments were conducted in VR using a Varjo Aero HMD (200 Hz eye-tracking, accuracy of 1° of visual angle). The sound was played by Beyerdynamic DT 770M headphones with passive noise cancellation. The experiment was designed using the Unity game engine, with the Meta XR Audio Spatializer (version 68.0.0) handling audio spatial propagation. Directional audio was rendered using a generic head-related transfer function (HRTF), which offers greater generalizability and has been reported to provide accurate sound localization in similar studies [28]. Regarding the spatializer parameters, scene reverberation was removed by applying a -60 dB intensity reduction, the pitch of audio sources remained unmodified, and the overall scene volume was reduced by -5 dB to maintain comfortable listening levels. Additional hardware details are included in **Sec. S1** of the supplementary material.

4.1.2 Audiovisual Stimuli and Depth Cues Control

As visual cues, we employed 3D traffic cones with a visual angle size of 0.7° at eye level, featuring high-contrast orange and white stripes. Conversely, the background 3D scene was gray and mostly empty. With a lifetime of 2 s, targets were shown well above the typical visual response times [15]. Simultaneous with the onset of the visual cue (traffic cone), a short beep-like sound centered at 880 Hz was played for 0.5 s, following the sound design of previous works [28, 69]. Such audiovisual stimuli were employed as both fixation and targets. Refer to Fig. 3 for more details on the sequential presentation of the stimulus.

In order to ensure experimental control over the existing visual cues, audiovisual stimuli are attached to the user's head [52] while the background scene is world-referenced. Since target locations are fixed in user field-of-view (FOV), motion parallax from depth-related motion drifts on the retina is cancelled [32, 53]. This ensures controlled saccades, as head motion does not alter target presentation. Simultaneously, head jitter and uniform motion-drift between all stimuli and the background are enabled by 6 DoF head-tracking. This design preserves vestibular feedback cues and enhances VR immersion, compared to other head-fixed setups such as chin-rest [47] or head tracking cancellation [36]. Additionally, the target scale is made depth-independent, always displaying it with 0.7° size horizontally. In our case, due to the small size of the visual cue, undesired depth indications based on texture gradients are minimal and not expected to affect our measurements. Targets' shadow-casting is disabled, and directional light ensures uniform shading of the targets in the entire depth range, leaving disparity and vergence as the only visual depth cues. Lastly, regarding audio depth cues, spatializer parameters allow us to disable audio reverberation to remove their influence on the measured overdriving effect. Since inner scene geometry changes are needed to study reverberation differences, our study is limited to audio source depth cues for a broader generalization, such as spatial location and frequency content.

Condition (Direction - Range)	Fixation	$\mathbf{T_{ref}}$	$\mathbf{T_{test}}$
Convergence - Near	0.8 m−4.5°	0.43 m−8.5°	0.55 m−6.5°
Convergence - Far	7 m−0.5°	0.8 m−4.5°	1.5 m−2.5°
Divergence - Near	0.43 m−8.5°	0.8 m−4.5°	0.55 m−6.5°
Divergence - Far	0.8 m−4.5°	7 m−0.5°	1.5 m−2.5°

Table 1: Physical depth and corresponding vergence angle (θ) related to fixation, reference target ($\mathbf{T_{ref}}$), and test target ($\mathbf{T_{test}}$) visual cues for the vergence Direction-Range combinations considered. Importantly, the placement of visual cues is adjusted according to the corresponding eccentricity for a correct vergence angle θ.

4.1.3 Procedure

All experiment sessions started with a five-dot eye-tracking calibration. During the experiment, both $\mathbf{T_{ref}}$ and $\mathbf{T_{test}}$ audiovisual stimuli were sequentially presented to the user who selected **which one was closer in depth**. Each trial began with the fixation stimulus at the center of the screen. When the fixation stimulus disappeared, the first audiovisual target was presented, eliciting the participant's visual response. The fixation stimulus then reappeared, followed by the second audiovisual target, and the trial concluded with the fixation stimulus shown once more. After all stimuli appeared, participants were instructed to indicate the target that was closer to them by button pressing, relying on both sound and visual cues. Then, the next trial starts. The order (first or second) and side (left or right) of the T_{ref} and T_{test} cases was randomized, while opposite sides for each case were always ensured. The experiment was conducted in three sessions of 20 minutes to prevent fatigue, while the order of the trials was also randomized across sessions. Over the three sessions, each participant completed a total of 240 trials with five minutes rest between sessions. The entire experiment lasted 1.5 hours.

4.1.4 Conditions and Sound Modulation

The variables we consider are *eccentricity* (E), *loudness shift* (ΔL), *vergence range* (R), and *vergence direction* (V). The *eccentricity* (E) describes the horizontal visual angle between the fixation and the visual targets. We choose $E = \{8°, 16°\}$ to align with common saccade displacements [2]. When visually shifting the focus point at different depths, the vergence movement can either diverge from the observer or converge to a closer target, setting the values for *vergence direction* (V) as $V = \{C, D\}$. Near and far distance ranges are investigated separately, setting $R = \{N, F\}$ for *vergence range* (R). All Direction-Range combinations can be seen in Tab. 1, where distances and vergence angles of visual cues are shown. Note how vergence angle variations ($\Delta\theta$) for reference ($\mathbf{T_{ref}}$) and test ($\mathbf{T_{test}}$) cases from fixation are respectively $\Delta\theta = 4°$ and $\Delta\theta = 2°$. The total vergence range considered in this work can be found in the supplementary, **Sec. S2**.

Regarding audio modality, different sound modulation levels for the test case ($\mathbf{T_{test}}$) relative to the reference case ($\mathbf{T_{ref}}$) are considered within *loudness shift* (ΔL). For convergence (Fig. 2, left), we consider ΔL = {Λ dB ($\mathbf{P_c}$), 0 dB ($\mathbf{P_m}$), +3 dB (p_1), +6 dB (p_2), +9 dB (p_3)}. For divergence (Fig. 2, right), we consider ΔL = {Λ dB ($\mathbf{P_c}$), 0 dB ($\mathbf{P_m}$), -3 dB (f_1), -8 dB (f_2), -13 dB (f_3)}. Note that Λ dB denotes case-dependent loudness variations (Tab. 1). Next, we explain each of these conditions:

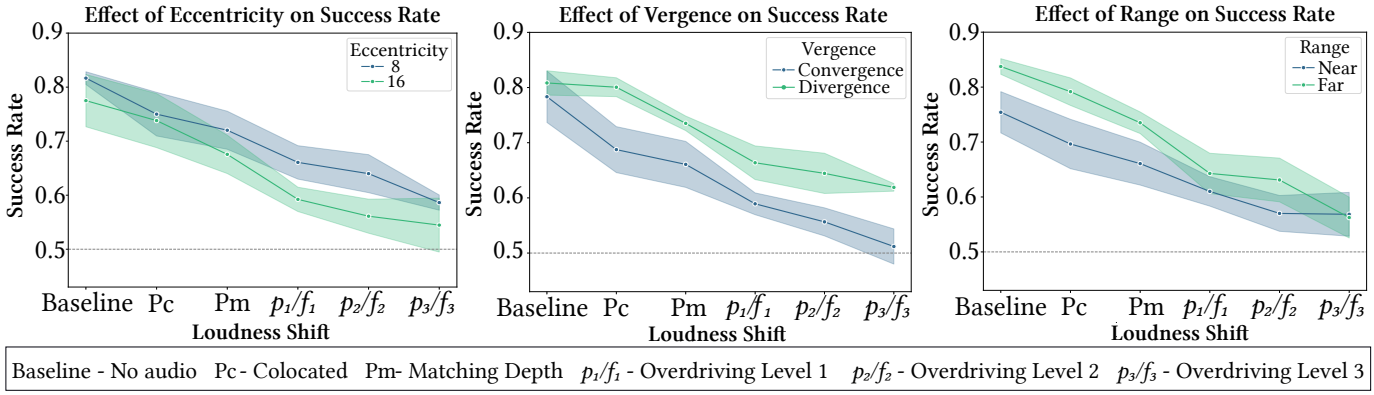

Fig. 4: Results of the main experiment on overdriving depth perception. The success rate indicates the probability of correctly selecting the stimulus visually closer to the participant: $\mathbf{T_{ref}}$ in Convergence and $\mathbf{T_{test}}$ in Divergence. We separately visualize the effect of eccentricity (left), vergence direction (center) and vergence range (right) on success rate. Data points indicate mean success rates across participants for each sound condition (linear trends for visualization only), and shaded regions indicate SEM error. The width of the shaded areas reflects variability across participants. The Baseline condition shows correct stereo acuity without any sound influence, with performance around 80%. By increasing sound modulation, success rates decrease toward chance level (dashed line), suggesting an intensified perceptual bias.

Colocated, P_c: The audio source is located at the same spatial position as the visual target $\mathbf{T_{test}}$, so the loudness variation with respect to $\mathbf{T_{ref}}$ (Λ dB) depends on each case (Tab. 1).

Matching depth, P_m: $\mathbf{T_{test}}$'s audio source is located at the same depth as $\mathbf{T_{ref}}$. Therefore, sound cues associated to the $\mathbf{T_{test}}$ and $\mathbf{T_{ref}}$ targets have minimal differences in loudness, resulting in $\Delta L = 0$ dB. In this first overdriving attempt, the auditory depth shift tries to compensate the visual depth change, testing if users mainly rely on the auditory cue for deciding on the depth of an object.

Overdriving Level 1, 2 and 3: In convergence, it covers the increased loudness shifts of +3 dB (p_1), +6 dB (p_2) and +9 dB (p_3), achieved by spatially moving $\mathbf{T_{test}}$'s audio source closer to the user according to Eq. 1. We chose steps of +3 dB as they represent doubling the power of the sound [45], resulting in a small but noticeable change in loudness by human observers. To enhance the effect, we also maximize IAD while ensuring audiovisual fusion (Fig. 2, left). In the case of divergence, the method is boosted by modifications of the frequency spectrum and consequent loudness shifts of -3 dB (f_1), -8 dB (f_2) and -13 dB (f_3) (Fig. 2, right).

We employed a fully factorial design with $2(E) \times 2(R) \times 2(V) \times 5(\Delta L) \times 2(left/right) = 80$ conditions. Each condition was repeated three times, yielding 240 trials per participant.

Baseline Study Before running the main experiment, a pilot study was run with 10 participants (4 female, 6 male, avg. age = 24.3, SD = 3.2) to test depth perception performance for every condition considered without any sound cue (no audio, only visual stimuli). The pilot was designed and performed to validate users' ability to properly distinguish visual depth changes above the stereoacuity threshold used in our experimental setup [14]. The procedure explained in the previous section also applies in this study, and the results are shown together with the main study for better comparison (Fig. 4).

4.2 Results

A total of 14 participants took part in the main experiment (6 females, 8 males, avg. age = 23.2, SD = 1.2). All reported normal or corrected-to-normal vision and hearing, and written consent was provided. The study protocol for all experiments reported in this document was approved by the Ethical Review Board of the Department of Computer Science at Saarland University. See **Sec. S6** in the supplementary for more details on the survey. Experiment results are shown in Fig. 4. A Generalized Linear Mixed Model (GLMM) with a logit link function was applied to analyze the binary dependent variable (success at selecting the closest target), including fixed effects (*eccentricity*, *loudness shift*, *vergence range*, *vergence direction*, and *side* of the test target) and a random intercept to account for variability across participants. Post-hoc pairwise comparisons using estimated marginal means were conducted to examine differences between levels of the sound loudness shift condition. Additionally, binomial tests compared participants' performance against chance level (50%) across all sound conditions and experimental subgroups ($E = 8°$, $E = 16°$, convergence, divergence, near range and far range). Refer to **Sec. S4** in the supplementary for detailed numerical results of the analyses. The analysis revealed significant effects for *loudness shift*, *eccentricity*, *vergence direction*, and *vergence range*, while the *side* did not significantly influence success rates ($p = 0.69$), suggesting that eye dominance does not influence depth discrimination. Participants showed a high success ratio of approximately 80% in the Baseline condition, consistent with expectations that vergence differences of 2° between $\mathbf{T_{ref}}$ and $\mathbf{T_{test}}$ are well above the stereoacuity threshold [14].

Higher *loudness shift* (ΔL) led to lower success rates ($p < 0.001$). Post-hoc pairwise comparisons confirmed significant differences among several sound conditions, particularly between small (P_c and P_m) and large (p_2/f_2 and p_3/f_3) loudness shifts, supporting the notion of auditory-induced overdriving of visual depth perception. Importantly, individual differences, while moderate (group-level variance of 0.9369), were not the primary source of the effect, emphasizing the findings' generalizability across participants.

A significant effect of *eccentricity* was observed ($p < 0.001$), with lower accuracy for higher eccentricities, reflecting a reduction in stereoacuity as stimuli moved further into the periphery. This aligns with the well-documented decrease in depth perception accuracy at larger eccentricities compared to near-foveal regions [57].

The effect of *vergence direction* was also significant ($p < 0.001$), showing better performance for convergence than divergence. Such difference shows how effective interaural differences (IAD) are processed as depth cue at near distances, being more reliable than frequency-based cues for far distances even with increased loudness variations (e.g., 8–12 dB with f_2 and f_3).

For *vergence range* ($p < 0.001$), there is a stronger decrease in accuracy in near-range conditions, where distance ranges from 0.43 m to 0.8 m (corresponding to θ ranging from 8.5° to 4.5°), compared to far-range conditions, where distance ranges from 0.8 m to 7 m (corresponding to θ ranging from 4.5° to 0.5°). This may be associated with the VAC, where closer stimuli align better with natural viewing conditions than farther stimuli, although further investigation is needed to disentangle the contributions in convergence or divergence motion. Nevertheless, no difference is visible for the largest loudness shift (Overdriving Level 3) used in the following sections.

Binomial tests also confirm what is visually apparent in Fig. 4. In most conditions, performance was statistically similar to chance when

considering the strongest manipulation (Overdriving Level 3), indicating that overdriving the visual depth perception through sound made the target's depth indistinguishable from the reference. Only in conditions with the lowest eccentricity ($E = 8, p = 0.0018$) and under divergence ($p < .0001$), performance was statistically distinguishable from chance level. This suggests a reduced, but still measurable, influence of the audiovisual manipulation.

4.3 Discussion

The results of the main experiment confirm our hypothesis that auditory cues can effectively modulate perceived depth and fully overdrive stereoscopic visual depth perception. As shown in Fig. 4, increasing the magnitude of the sound manipulation progressively reduces participants' ability to distinguish between $\mathbf{T_{test}}$ and $\mathbf{T_{ref}}$ targets, with performance approaching chance level. This indicates that the audiovisual fusion process successfully masked the underlying visual disparity manipulation. Thereby, the effect was observed consistently across both near and far vergence ranges, covering depths in the full operational depth range of standard VR systems from 0.43 m to 7 m. The effectiveness of the overdriving varies depending on eccentricity and vergence direction. For lower eccentricity ($E = 8°$), the overdriving effect was weaker compared to the $E = 16°$ condition, which may be attributed to the reduced strength of IAD at smaller lateral displacements. At higher eccentricity, IAD becomes a stronger cue due to increased lateralization of the sound source, especially in near-range convergence conditions [35]. Similarly, in divergence motion, the method is less effective than in convergence. Again, IAD impact is reduced given the limited horizontal angle displacement for larger distances. Overall, the results support our hypothesis that spatially misaligned sound cues can override visual depth judgments in stereoscopic displays, provided the manipulation is strong enough and well-tuned to the perceptual sensitivity of the respective cue type (IAD vs. spectrum).

5 Visual Response Acceleration

The results of the main experiment (Sec. 4) show that moving a visual target closer to a former fixation can remain indistinguishable if the shift is properly compensated by sound modulation. In this section, we evaluate the effect of this depth reduction on the vergence response time. To do so, we apply Overdriving Level 3 sound modulation from Sec. 4 to test how the time needed to fixate on a new target significantly decreases for smaller vergence variations. Consequently, depth-overdriven targets would significantly increase user performance, as visual information is recovered faster.

5.1 Experimental Setup

To optimally measure the response behavior, we specifically place the targets' visual cues at the limits of the vergence range. This means, we focus on Convergence-Near and Divergence-Far conditions (cf. Tab. 1), which allow us to shorten the vergence range from 0.43 m to 7 m in distance (with θ ranging from 8.5° to 0.5°, $\Delta\theta_{max} = 8°$) to a new compressed range from 0.5 m to 1.5 m in distance (with θ ranging from 6.5° to 2.5°, $\Delta\theta_{max} = 4°$). Meanwhile, sound manipulation allows for overdriving perceived depth as shown in the previous section.

5.1.1 Stimuli and Procedure

Compared to the main experiment, the same scene (Fig. 3, background) and Overdriving Level 3 loudness shift (Sec. 4) were used. Regarding visual cues, here the initial fixation displayed a centered red cross, while visual targets displayed a clustered Landolt C shapes of 0.7°, as shown in Fig. 5. Such a visual target ensures peripheral masking, thus enforcing proper fixation [58].

The participant's task was to react to the newly appearing audiovisual targets and provide **the orientation of the central Landolt C as fast as possible** by button-pressing after appearance. In each trial, participants started by looking at the fixation cross, which eventually disappeared, showing the first audiovisual target T_1. After the first response was introduced, the second target T_2 was sequentially generated at the other extreme of the vergence range. We measure the response time between the onset of both audiovisual targets, T_1 and T_2, and the manual response provided for each of them. These times include saccade and vergence motions and cognitive processing. Notably, if the response provided for T_1 or T_2 did not match the orientation of the central Landolt C, the trial was rejected and the response time was discarded. The temporal scheme showing this user response time is shown in the supplementary **Sec. S3**.

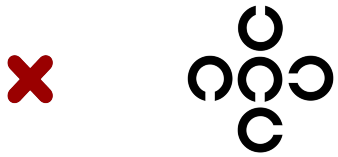

Fig. 5: Red cross (left) used as initial fixation and clustered Landolt C shapes (right) used as target to measure vergence time. Surrounding Landolt C's pattern is used to avoid orientation discrimination using peripheral vision, ensuring correct foveation [58].

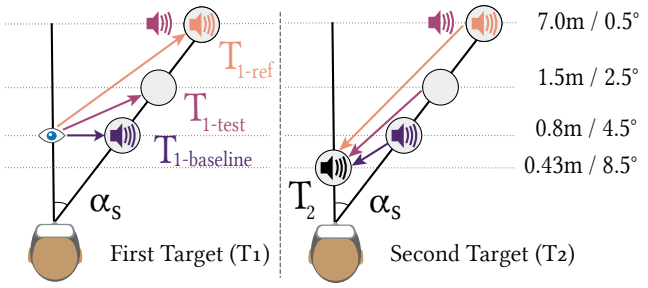

Fig. 6: Scheme showing the audiovisual cues location used in Divergence-Far condition cases. *(left)* After the initial fixation at 0.8 m, the subsequent fixation at the first target T_1 requires different divergence changes ($\Delta\theta_{T_1} = 0°$, 2°, or 4°) depending on the case. *(right)* Consequently, corresponding convergence motions to fixate at the second target T_2 at 0.43 m ($\Delta\theta_{T_2} = 4°$, 6°, or 8°, respectively) are required. Note that the $T_{1\text{-baseline}}$ and $T_{1\text{-ref}}$ cases have audio co-located with the visual target. In the $T_{1\text{-test}}$ case, overdriving sound is applied toward the original position at $T_{1\text{-ref}}$ following main experiment results (Sec. 4). $T_{1\text{-baseline}}$ is included for further comparison. A representation of the recorded gaze trajectories on angular displacements for each T_1 case is shown in Fig. 7.

5.1.2 Conditions and Participants

Aside from the two vergence Direction-Range (V-R) conditions considered (Convergence-Near and Divergence-Far), we also test three different cases (C) for the depth of the first target (T_1) from fixation: $T_{1-\text{baseline}}$ ($\Delta\theta_{T_1} = 0°$, pure saccade); $T_{1-\text{test}}$ ($\Delta\theta_{T_1} = 2°$, $\mathbf{T_{test}}$ with overdriven depth referred in previous sections); and $T_{1-\text{ref}}$ ($\Delta\theta_{T_1} = 4°$, $\mathbf{T_{ref}}$ in previous sections), as displayed in Fig. 6, left. Consequently, the second target T_2 ($E = 0°$) at the opposite extreme of the range with colocated audio requires corresponding revergence motions according to each T_1 case, being respectively $\Delta\theta_{T_2} = 4°$, $\Delta\theta_{T_2} = 6°$ and $\Delta\theta_{T_2} = 8°$ (Fig. 6, right). See Fig. 6 for an illustration of the Divergence-Far condition. Notably, while Divergence-Far condition requires a divergence-convergence motion sequence to fixate at T_1 and T_2, the opposite convergence-divergence motion sequence is required for the Convergence-Near condition. Lastly, 8° and 16° eccentricity values (E) are also considered. Each of the 24 combinations – $2(E) \times 2(V-R) \times 3(C) \times 2(left/right)$ – is presented 8 times, resulting in a total of 192 trials per participant. The 192 trials are conducted in two sessions with five minutes rest in between, and a total experiment length of one hour. A total of 5 participants took part in the experiment (2 females, 3 males, avg. age = 23.8, SD = 2.6).

5.2 Results and Discussion

Results of the accumulated response T_1+T_2 are shown in Fig. 8, while additional details and discussion about the individual results for the first (T_1) and second (T_2) targets are shown separately in the supplemen-

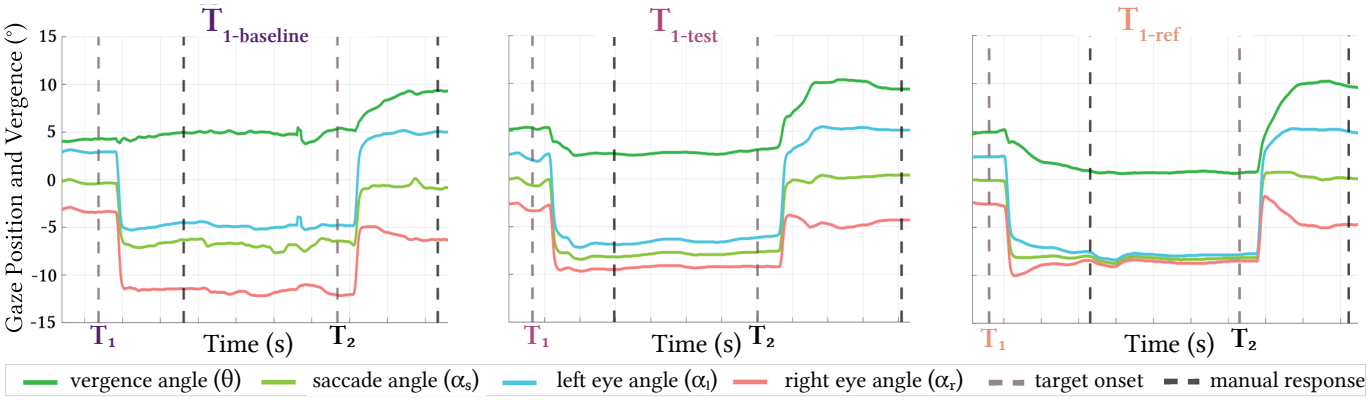

Fig. 7: Visualization of angular eye gaze and vergence trajectories over time, as recorded by eye tracking after applying a moving average filter on the raw data (kernel size of 10 data samples). Each graph shows one experiment trial, covering the sequence of T_1 and T_2 (targets' onsets trigger the visual response and press-button action) as illustrated in Fig. 6 for the Divergence-Far condition. Here, the three different cases ($T_{1-\text{baseline}}, T_{1-\text{test}}, T_{1-\text{ref}}$) of the first audiovisual target T_1 ($E = -8°$) are compared. In all cases, the initial fixation point is placed at 0.8 m ($\theta = 4.5°$). As $T_{1-\text{baseline}}$ is located at the same depth as the initial fixation, no vergence motion is required. On the other hand, $T_{1-\text{test}}$, at 1.5 m, and $T_{1-\text{ref}}$, at 7 m, require respective divergence motions of $-2°$ and $-4°$ ($\Delta\theta_{T_1}$) to focus correctly at the new depths. Afterward, all cases require the corresponding convergence to focus on T_2 ($E = 0°$) located at 0.43 m ($\theta = 8.5°$). Note the relative differences between left and right eye angular positions showing disparity changes related to the depth being fixated on each T_1 case. Each tick in the horizontal axis represents 0.5 s.

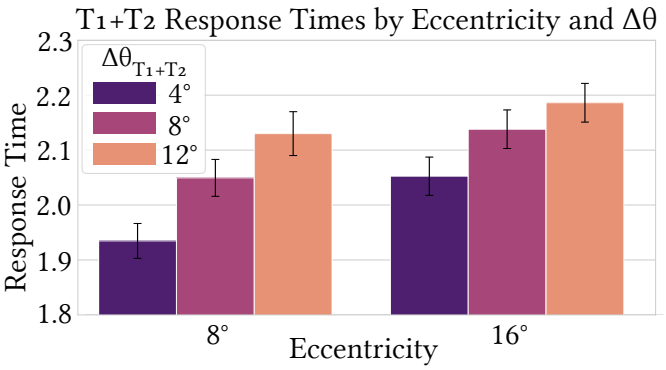

Fig. 8: Accumulated response times for the T_1 + T_2 sequence. Response times increase with larger variations of vergence angle $\Delta\theta_{T_1+T_2}$ and larger eccentricity, while error bars show standard error of the mean (SEM).

Fig. 9: Representation of the driving scene from the user perspective. Annotations used as targets correspond to different item categories present in the scene (**BIKE**, **CAT**, **TOY**, **KID**). During the experiment, annotations are shown one after the other, while participants choose the scene item the annotation is likely paired with.

tary, **Sec. S3**. A Linear Mixed Model (GLMM) was used to analyze the dependent variables (response times for T_1, T_2, and accumulated T_1+T_2), including fixed effects (E, T_1 side, $\Delta\theta$, vergence direction) and a random intercept to account for variability across participants.

A significant effect of eccentricity ($p < 0.001$) shows that larger eccentricities increase saccade latency, showing a slower reaction time towards periphery [28]. We also observe differences for $\Delta\theta$ ($p < 0.001$), where larger response times are found for vergence motions with larger $\Delta\theta$, consequence of larger disparity variations [60]. No significant differences were observed when comparing convergence vs. divergence motion ($p = 0.948$), in line with previous work [15]. The side of appearance ($left/right$) did not significantly influence response times ($p = 0.585$), suggesting that ocular dominance does not affect performance in this task either.

For a better understanding of the visual response, different visual performance profiles can be found in Fig. 7, where each graph corresponds to the different cases of T_1 ($T_{1-\text{baseline}}$, $T_{1-\text{test}}$ and $T_{1-\text{ref}}$) in the Divergence-Far condition. To visualize the real vergence angle variations corresponding to different fixation depths, we can analyze the recorded eye tracking data accordingly. Since we focus on purely horizontal visual motion, both saccades and vergence, left and right gaze rays are projected and normalized to the horizontal plane. Consequently, the vergence angle (θ) can be computed as the difference between the left and right gaze horizontal angles: $\theta = \alpha_l - \alpha_r$ [15]. Angles are computed from the corresponding eye gaze direction vector: $\alpha = atan2(x,z)$; being x and z the horizontal and forward components, respectively, in a left-hand coordinate system and relative to head pose. Similarly, the saccade angle (α_s) is obtained from the combined gaze direction vector. In Fig. 7, a moving average filter (kernel size of 10 samples) is applied to smooth the signal for visualization reasons. In short, we confirmed that reduced vergence variations lead to significantly faster visual responses. Combined with our main experiment showing effective sound-based visual depth overdriving, these results suggest that our method can not only reproduce the depth percept of the full vergence change but also improve response times in VR scenarios.

6 Application Case

We evaluate our method in a naturalistic application scenario resembling a driving scene and an annotation task, where text descriptions appear to provide information about objects in the environment. Here, we can apply our methodology to shift the perception of one annotation at a fixed visual location to be associated with multiple distant objects.

6.1 Experimental Setup

6.1.1 Stimuli and Procedure

The scene displays a low-poly city scene where the user takes the perspective of a driver with a multitude of items in the street (see Fig. 9). The location of the scene items was chosen to cover the full

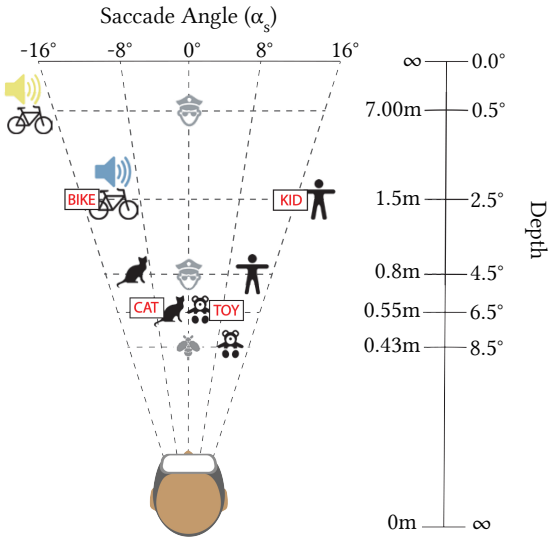

Fig. 10: Spatial placement of the scene items and their corresponding annotations. Each annotation is positioned at the same depth as the scene item that is closer to the previous fixation, while two sound conditions are tested. For example, both the colocated (blue) and overdriving (yellow) sound conditions are represented for the bike item annotation case. As observed, annotations fall within a vergence range of 4° ([2.5°, 6.5°]), while the scene items span the initial 8° range ([0.5°, 8.5°]). Initial fixation items are marked in grey.

visual vergence range of former experiments (cf. Tab. 1): bicycles at Divergence-Far ($E = -16°$), kids at Convergence-Far ($E = 16°$), cats at Divergence-Near ($E = -8°$), and toys at Convergence-Near ($E = 8°$). In this case, a policeman or a bee (only for very close fixations) acted as a fixation item. We chose 2D annotations as audiovisual targets (2° size), which describe the items they refer to. Auditory cues were chosen in context: a ring bell for **BIKE**, a "hello" call for **KID**, a meow sound for **CAT**, and a squeak for **TOY**. At the beginning of each trial, the corresponding fixation item was shown. Afterward, the visual annotation spawned with a lifetime of 2 s and the audio cue was played for 0.5 s. We asked participants to indicate **which of the same two scene items the annotation was referring to** by button pressing.

6.1.2 Conditions and Participants

The visual part of the audiovisual annotation always spawns at a fixed depth, the same depth as the scene item closer to the respective fixation. For the sound, two conditions are tested: audio colocated (reference condition) and audio modulation based on the results of Sec. 4 (overdriving condition). Scene items spatial distribution, annotations visual depth and an example of these two sound conditions are shown in Fig. 10. Selected Overdriving levels (ΔL) are +6 dB for **TOY**, +9 dB for **KID**, -8 dB for **BIKE** and -13 dB for **CAT**. Participants completed 64 trials (8 conditions × 8 repetitions) in a single 15-minute session. All conditions were randomized and 10 new participants (4 females, 6 males, avg. age = 26.4, SD = 3.9) completed the study.

6.2 Results and Discussion

Again, we use a GLMM with a logit link function to analyze the binary dependent variable (success at associating the annotation to the item at the same visual depth), including fixed effects (E, ΔL, and vergence direction) and a random intercept to account for variability across participants. As shown in Fig. 11, sound modulation significantly impacts user decisions when associating annotations with scene objects ($p < 0.001$). Binomial tests are computed comparing participants' performance against chance level (50%) for each experiment condition. Details can be found in the supplementary, **Sec. S4**. Despite observing a clear effect of sound modulation on associating annotation with the scene item, different scenarios are found for each scene item category.

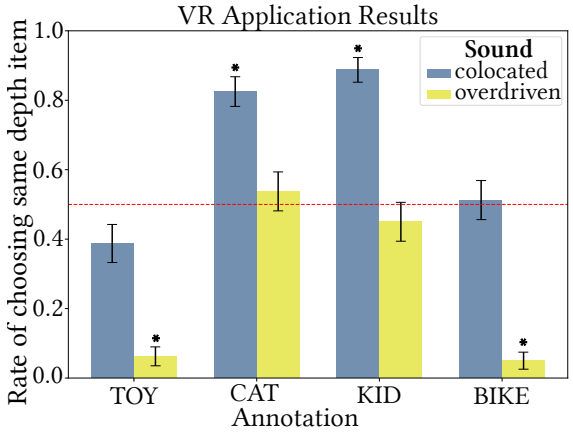

Fig. 11: Results obtained in the application experiment. The Y-axis indicates the ratio of selecting the scene item at the same visual depth as the annotation. Low ratios indicate the tendency to perceive the annotation to be related to the scene item farther away from the fixation. Error bars show SEM and stars show statistical significance against chance level, represented by the horizontal red line.

For the **CAT** and **KID** annotations, participants correctly indicated the associated objects (> 80% accuracy, being both significantly different from chance level) with colocated sound. Here, the perceived depth of the annotations could be overdriven by displaced sound, thus not being significantly different from chance level ($t\_stat = 0.6685, p = 0.5058$ for **CAT**; $t\_stat = -0.8933, p = 0.3744$ for **KID**), meaning that the annotation is equally associated with the scene item at a different visual depth and farther from fixation. Although the **CAT** condition (Divergence, $E = -8°$) is the most critical condition according to main experiment results (Fig. 4), it shows the expected effect behaviour on perceived depth by overdriving visual cues. Note that annotations' visual depth does not vary between colocated and overdriven conditions. Hence, sound compensation replicates the main experiment results for these two cases. Conversely, for the **TOY** and **BIKE** annotations, colocated audio reinforced position ambiguity since ratios are not significantly different from chance level ($t\_stat = -2.0525, p = 0.0434$ for **TOY**; $t\_stat = 0.2223, p = 0.8247$ for **BIKE**)), while sound modulation allows to perceptually recover the initial larger vergence range. In the **TOY** case, the short distance between the items seems to diminish the effect of collocated sound and obscure the clearly visible depth differences provided by stereovision (in the absence of sound), as also observed in the main experiment. In the **BIKE** case, the annotation's specific on-screen placement may have biased participants, causing them to overlook consistent visual and auditory depth cues. As conclusion, sound modulation has been reported to be effective in two different scenarios: hiding clear visual vergence thus confusing item association as in **CAT** and **KID** cases and the main experiment; but also resolving item association when colocated sound makes it confusing, as in **BIKE** and **TOY** cases. Therefore, we observe how the impact of the overdriving effect can be influenced by the particular scene configuration and depends on the relative strength of the visual and auditory cues as postulated by depth cue fusion models [4, 11, 16].

Mixed Reality Application Version Additionally, a similar application was implemented in Mixed Reality (MR) to explore further use cases. Different scene items were used as shown in Fig. 12, and results displayed in Fig. 13 show similar trends as those reported in this section. Overdriving sound is found to hide visual depth differences for **TOY**, **PHONE**, and **SPEAKER**, similarly to **CAT** and **KID** cases in the VR version. Different behavior is observed for the **BIRD** case, the closest one in depth, being consistent with the **TOY** case in VR. Statistical analyses also confirm significant differences between sound conditions, highlighting the potential of the overdriving effect. Further details of this MR version of the application are explained in **Sec. S5** in the supplementary.

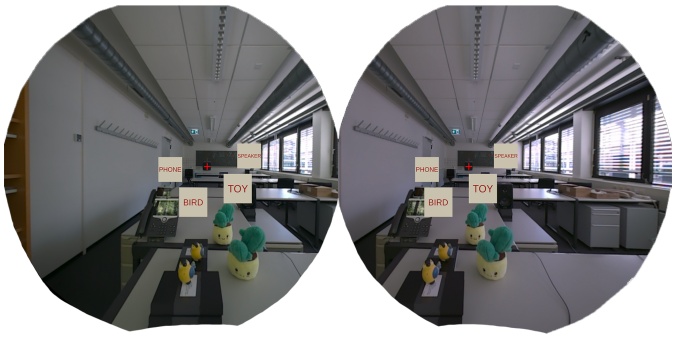

Fig. 12: MR application version from the participant's point of view, showing left and right views. Virtual annotations' visual cues, scaled by a factor of 3 for visualization purposes, are shown within the real scene for each scene item (**BIRD**, **TOY**, **PHONE**, and **SPEAKER**), as well as the central red cross used as fixation.

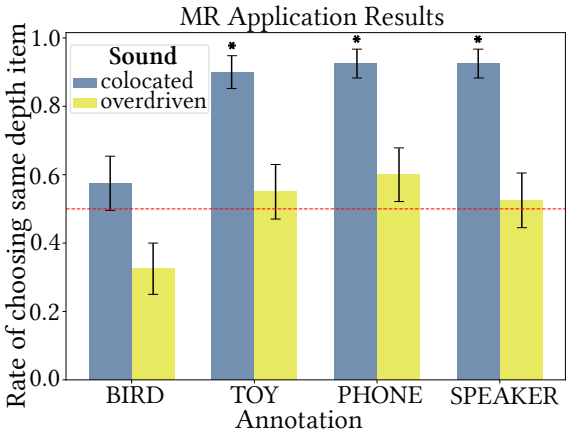

Fig. 13: Results obtained in the MR version of the application case, displayed in the same format as its VR counterpart in Fig.11. Overdriving sound shows similar trends with respect to the colocated condition as in VR application, showing effect potential also in MR.

7 Limitations and Future Work

While our approach demonstrates the effectiveness of auditory cues in modulating depth perception, limitations and avenues for future exploration remain.

Statistical power and results generalizability The response time experiment in Sec. 5 includes a limited number of participants, as its goal is not to establish a novel effect but to verify that a well-documented relationship (faster responses for smaller vergence changes [15]) is preserved when using our sound-based overdriving method. This confirmatory setup allows us to assess whether vergence range differences relevant to VR, and consequently the presence of the VAC, remain associated with faster visual responses once the overdriving effect is perceptually achieved. For the remaining experiments, larger participant pools would naturally increase statistical power; however, sample sizes comparable to those used in this work have previously been shown to be sufficient to reliably detect differences using similar experimental designs and statistical analyses [15, 28].

Seamless initial depth recovery. In our study, we considered quickly transient targets as well as short-duration sound signals. However, for a visual target shown for a prolonged time after the audio stops, imperceptible disparity manipulations might allow for the recovery of the initial depth during extended user fixation [31]. This would enable scene coherence restoration for a long-standing effect without any auditory cue. We have observed linear-constant disparity manipulation over the line of sight being particularly promising.

Feedback-driven recalibration and habituation effects. Our experiments intentionally avoid feedback or task interaction that could induce perceptual recalibration [43], allowing us to isolate the immediate effect of audiovisual decoupling. Although recalibration may arise in prolonged or interactive scenarios, prior work indicates that distance misperception in VR typically persists due to limited visual cues [17], while relative audio-loudness differences continue to modulate visual depth perception during exposure [10]. This indicates that users do not automatically adapt to spatial inconsistencies in common VR settings. Our method remains particularly effective in time-critical scenarios such as annotations, alerts, or object spawning, where stimuli appear briefly and users have limited opportunity to recalibrate. In these contexts, the proposed audio-based depth manipulation can enhance spatial interpretation and visual response without long-term perceptual side effects. Future work should investigate the longevity of the effect, including whether gradual cue adaptation or feedback mechanisms diminish the perceptual benefit over time.

Extent of vergence changes and range. We overdrive vergence angle changes of 2° between $\mathbf{T_{ref}}$ and $\mathbf{T_{test}}$, sufficient to elicit measurable visual responses (Sec. 5). Larger vergence shifts ($> 2^\circ$) may be harder to overdrive by sound. Although visual response speedup could increase further, bringing the visual target closer to the fixation depth while the intended location remains farther away makes full concealment by auditory cues unlikely. Also, while near and far vergence regions were studied to explore the effectiveness of the overdriving effect towards the vergence range limits; future work could explore how it may behave in mid-range regions.

Occlusions and scene composition. Visual occlusions as additional depth cues are a natural limitation of our approach since they indicate the relative distance positioning of scene elements. Altering visual depth with our method may disrupt these relationships, potentially leading to implausible scene interpretations.

Impact on long-term fatigue. Our technique inherently reduces vergence magnitude when fixating on newly appearing objects closer to the HMD focal distance (Sec. 6), offering clear potential to mitigate the VAC. This mitigation could be especially beneficial during prolonged VR sessions, where sustained VAC significantly contributes to discomfort. However, further systematic studies are needed to evaluate our technique's impact on long-term fatigue.

8 Conclusion

In this work, we investigate the impact of auditory cues on stereoscopic depth perception in VR. Specifically, we demonstrate that stereoscopic depth perception can be overdriven through controlled auditory manipulations based on maximizing sound-based depth cues while carefully ensuring that the integrity of audiovisual cues is preserved. As expected, stronger auditory manipulations induce more consistent depth overdriving, highlighting the potential of multimodal integration as a complementary approach to existing visual-only correction methods. The overdriving effect and its modulation levels are validated in a realistic scenario, showing how the perceived depth range can be effectively extended. Furthermore, by reducing vergence demands through auditory overdriving, faster depth adjustments and gaze realignment are enabled. These findings provide perceptual guidance for the design of dynamic audiovisual interfaces in AR and VR, where components may be distributed across varying depths. A more detailed exploration of these aspects is left for future work.

Acknowledgments

This work has been supported by grant PID2022-141539NB-I00, funded by MICIU/AEI/10.13039/501100011033 and by ERDF, EU; by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 101220555); by the National Science Foundation grants 2107454, 2225861 and 2232817; by the German Science Foundation (DFG grant 557564533); and by an academic gift from Meta. The authors would like to thank the participants of the experiments for their participation.

References

[1] E. Arabadzhiyska, O. T. Tursun, K. Myszkowski, H.-P. Seidel, and P. Didyk. Saccade landing position prediction for gaze-contingent rendering. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 36(4), 2017. 2

[2] A. T. Bahill, D. Adler, and L. Stark. Most naturally occurring human saccades have magnitudes of 15 degrees or less. *Investigative Ophthalmology & Visual Science*, 14(6):468–469, 1975. 4

[3] A. T. Bahill, M. R. Clark, and L. Stark. The main sequence, a tool for studying human eye movements. *Mathematical Biosciences*, 24(3-4):191–204, 1975. 2

[4] P. Battaglia, R. Jacobs, and R. Aslin. Bayesian integration of visual and auditory signals for spatial localization. *Journal of the Optical Society of America A*, 20:1391–1397, 2003. 1, 3, 8

[5] M. Bernhard, C. Dell'mour, M. Hecher, E. Stavrakis, and M. Wimmer. The effects of fast disparity adjustment in gaze-controlled stereoscopic applications. In *Proc. Symp. on Eye Tracking Research and Appl. (ETRA)*, pp. 111–118, 2014. 2

[6] J. S. Berry, D. A. Roberts, and N. S. Holliman. 3d sound and 3d image interactions: a review of audio-visual depth perception. *Human Vision and Electronic Imaging XIX*, 9014:33–48, 2014. 2

[7] R. Bolia, W. D'Angelo, and R. McKinley. Aurally aided visual search in three-dimensional space. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 41(4):664–669, 1999. 2

[8] A. Bronkhorst and T. Houtgast. Auditory distance perception in rooms. *Nature*, 397(6719):517—520, 1999. 2

[9] D. S. Brungart and W. M. Rabinowitz. Auditory localization of nearby sources. head-related transfer functions. *The Journal of the Acoustical Society of America*, 106(3):1465–1479, 1999. 2

[10] P. Bruns. The ventriloquist illusion as a tool to study multisensory processing: An update. *Frontiers in Integrative Neuroscience*, 13:51, 2019. 9

[11] J. J. Clark and A. L. Yuille. *Data Fusion for Sensory Information Processing Systems*. Kluver Academic, Norwell, Mass, 1990. 1, 2, 8

[12] H. Collewijn, C. J. Erkelens, and R. M. Steinman. Voluntary binocular gaze-shifts in the plane of regard: Dynamics of version and vergence. *Vision Research*, 35(23):3335–3358, 1995. 2

[13] J. Cutting and P. Vishton. Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth. In W. Epstein and S. Rogers, eds., *Perception of Space and Motion (Handbook of Perception and Cognition)*, pp. 69–117. Academic Press, 1995. 2

[14] P. Didyk, T. Ritschel, E. Eisemann, K. Myszkowski, and H.-P. Seidel. A perceptual model for disparity. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 30(4):96, 2011. 2, 5

[15] B. Duinkharjav, B. Liang, A. Patney, R. Brown, and Q. Sun. The shortest route is not always the fastest: Probability-modeled stereoscopic eye movement completion time in vr. *ACM Trans. Graph. (Proc. SIGGRAPH Asia)*, 42(6), article no. 220, 2023. 2, 4, 7, 9

[16] M. Ernst and H. Bülthoff. Merging the senses into a robust percept. *Trends in Cognitive Sciences*, 8:162–9, 2004. 1, 3, 8

[17] D. J. Finnegan. *Compensating for distance compression in virtual audio-visual environments*. PhD thesis, University of Bath, 2017. 9

[18] M. Godfroy, C. Roumes, and P. Dauchy. Spatial variations of visual—auditory fusion areas. *Perception*, 32(10):1233–1245, 2003. 2, 3

[19] M. Gorzel, D. Corrigan, G. Kearney, J. Squires, and F. Boland. Distance perception in virtual audio-visual environments. In *25th UK Conference of the Audio Engineering Society: Spatial Audio In Today's 3D World (2012)*, pp. 1–8, 2012. 2, 3

[20] M. Grohn, T. Lokki, and T. Takala. Comparison of auditory, visual and audio-visual navigation in a 3d space. *ACM Transactions on Applied Perception*, 2(4):564–570, 2005. 2

[21] C. Groth, S. Fricke, S. Castillo, and M. Magnor. Wavelet-based fast decoding of 360° videos. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, pp. 1–9, 2023. doi: 10.1109/TVCG.2023.3247080 2

[22] C. Groth, M. Magnor, S. Grogorick, M. Eisemann, and P. Didyk. Cybersickness reduction via gaze-contingent image deformation. *ACM Transactions on Graphics (TOG)*, 43(4), 2024. doi: 10.1145/3658138 2

[23] C. Groth, T. Scholz, S. Castillo, J.-P. Tauscher, and M. Magnor. Instant hand redirection in virtual reality through electrical muscle stimulation-triggered eye blinks. In *ACM Symposium on Virtual Reality Software and Technology (VRST)*, number 37, pp. 1–11, 2023. doi: 10.1145/3611659.3615717 1

[24] W. M. Hartmann. Localization and lateralization of sound. *Binaural Hearing: With 93 Illustrations*, pp. 9–45, 2021. 3

[25] D. M. Hoffman, A. R. Girshick, K. Akeley, and M. S. Banks. Vergence–accommodation conflicts hinder visual performance and cause visual fatigue. *Journal of vision*, 8(3):33–33, 2008. 2

[26] I. P. Howard. *Perceiving in Depth, Volume 2: Stereoscopic Vision*. Oxford University Press, 2012. 2

[27] C. E. Jack and W. R. Thurlow. Effects of degree of visual association and angle of displacement on the "ventriloquism" effect. *Perceptual and Motor Skills*, 37(3):967–979, 1973. 2

[28] D. Jiménez Navarro, X. Peng, Y. Zhang, K. Myszkowski, H.-P. Seidel, Q. Sun, and A. Serrano. Accelerating saccadic response through spatial and temporal cross-modal misalignments. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–12, 2024. 3, 4, 7, 9

[29] D. Jiménez-Navarro, A. Serrano, and S. Malpica. Minimally disruptive auditory cues: their impact on visual performance in virtual reality. *The Visual Computer*, 41(7):5059–5073, 2025. 3

[30] G. R. Jones, D. Lee, N. S. Holliman, and D. Ezra. Controlling perceived depth in stereoscopic images. In *SPIE vol. 4297*, pp. 42–53, 2001. 2

[31] P. Kellnhofer, P. Didyk, K. Myszkowski, M. M. Hefeeda, H.-P. Seidel, and W. Matusik. Gazestereo3d: seamless disparity manipulations. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 35(4), article no. 68, 2016. 2, 9

[32] P. Kellnhofer, P. Didyk, T. Ritschel, B. Masia, K. Myszkowski, and H.-P. Seidel. Motion parallax in stereo 3d: Model and applications. *ACM Transactions on Graphics (TOG)*, 35(6):1–12, 2016. 4

[33] J. Kim, J. Spjut, M. McGuire, A. Majercik, B. Boudaoud, R. Albert, and D. Luebke. Esports arms race: Latency and refresh rate for competitive gaming tasks. *Journal of Vision*, 19(10):2–2, 2019. 2

[34] D. C. Knill and W. Richards. *Perception as Bayesian Inference*. Cambridge University Press, 1996. 2

[35] A. J. Kolarik, B. C. Moore, P. Zahorik, S. Cirstea, and S. Pardhan. Auditory distance perception in humans: a review of cues, development, neuronal bases, and effects of sensory loss. *Attention, Perception, & Psychophysics*, 78:373–395, 2016. 3, 6

[36] R. Konrad, A. Angelopoulos, and G. Wetzstein. Gaze-contingent ocular parallax rendering for virtual reality. *ACM Transactions on Graphics (TOG)*, 39(2):1–12, 2020. 4

[37] A. Kramer, B. Röder, and P. Bruns. Feedback modulates audio-visual spatial recalibration. *Frontiers in integrative neuroscience*, 13:74, 2020. 3

[38] J. Kudnick, M. Weier, C. Groth, B. Fu, and R. Horst. Warpvision: Using spatial curvature to guide attention in virtual reality. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 31(11), 2025. doi: 10.1109/TVCG.2025.3616806 1

[39] D. P. Kumpik, C. Campbell, J. W. Schnupp, and A. J. King. Re-weighting of sound localization cues by audiovisual training. *Frontiers in neuroscience*, 13:1164, 2019. 3

[40] M. S. Landy, L. T. Maloney, E. B. Johnston, and M. Young. Measurement and modeling of depth cue combination: In defense of weak fusion. *Vision Research*, 35(3):389–412, 1995. 2

[41] M. Lang, A. Hornung, O. Wang, S. Poulakos, A. Smolic, and M. Gross. Nonlinear disparity mapping for stereoscopic 3D. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 29(4):75, 2010. 2

[42] S. Latif, H. Tarner, and F. Beck. Talking realities: audio guides in virtual reality visualizations. *IEEE Computer Graphics and Applications*, 42(1):73–83, 2021. 1

[43] W.-Y. Lin, R. Venkatakrishnan, R. Venkatakrishnan, S. V. Babu, C. Pagano, and W.-C. Lin. An empirical evaluation of the calibration of auditory distance perception under different levels of virtual environment visibilities. In *2024 IEEE conference virtual reality and 3D user interfaces (VR)*, pp. 690–700. IEEE, 2024. 3, 9

[44] A. D. Little, D. H. Mershon, and P. H. Cox. Spectral content as a cue to perceived auditory distance. *Perception*, 21(3):405–416, 1992. 2, 3

[45] L. E. Marks. A theory of loudness and loudness judgments. *Psychological Review*, 86(3):256, 1979. 5

[46] D. Martin, S. Malpica, D. Gutierrez, B. Masia, and A. Serrano. Multimodality in vr: A survey. *ACM Comput. Surv.*, 54(10s), article no. 216, 2022. 1

[47] Y. Mikawa and T. Fukiage. Low-latency ocular parallax rendering and investigation of its effect on depth perception in virtual reality. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 4

[48] M. Mon-Williams, J. R. Tresilian, and A. Roberts. Vergence provides

veridical depth perception from horizontal retinal image disparities. *Experimental brain research*, 133:407–413, 2000. 3

[49] A. C. Naef, M.-M. Jeitziner, S. E. Knobel, M. T. Exl, R. M. Müri, S. M. Jakob, T. Nef, and S. M. Gerber. Investigating the role of auditory and visual sensory inputs for inducing relaxation during virtual reality stimulation. *Scientific reports*, 12(1):17073, 2022. 1

[50] T. Oskam, A. Hornung, H. Bowles, K. Mitchell, and M. H. Gross. OSCAM-optimized stereoscopic camera control for interactive 3D. *ACM Trans. Graph. (Proc. SIGGRAPH Asia)*, 30(6):189, 2011. 2

[51] S. Palmer. *Vision Science: Photons to Phenomenology*. MIT Press, 1999. 2

[52] F. Prummer, M. Shereef Abdelwahab, F. Weidner, Y. Abdrabou, and H. Gellersen. It's not always the same eye that dominates: Effects of viewing angle, handedness and eye movement in 3d. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–10, 2025. 4

[53] B. Rogers and M. Graham. Motion parallax as an independent cue for depth perception. *Perception*, 8(2):125–134, 1979. 4

[54] S. M. Ross and L. E. Ross. Saccade latency and warning signals: effects of auditory and visual stimulus onset and offset. *Perception & Psychophysics*, 29:429–437, 1981. 3

[55] J. Semmlow and P. Wetzel. Dynamic contributions of the components of binocular vergence. *Journal of the Optical Society of America*, 69(5):639–645, 1979. 2

[56] T. Shibata, J. Kim, D. M. Hoffman, and M. S. Banks. The zone of comfort: Predicting visual discomfort with stereo displays. *Journal of Vision*, 11(8):11–11, 2011. 2

[57] J. Siderov and R. S. Harwerth. Stereopsis, spatial frequency and retinal eccentricity. *Vision Research*, 35(16):2329–2337, 1995. 5

[58] D. P. Spiegel and I. M. Erkelens. Vergence-accommodation conflict increases time to focus in augmented reality. *Journal of the Society for Information Display*, 32(5):194–205, 2024. 2, 6

[59] T. Z. Strybel and D. R. Perrott. Discrimination of relative distance in the auditory modality: the success and failure of the loudness discrimination hypothesis. *The Journal of the Acoustical Society of America*, 76 1:318–20, 1984. 2, 3

[60] K. Templin, P. Didyk, K. Myszkowski, M. M. Hefeeda, H.-P. Seidel, and W. Matusik. Modeling and optimizing eye vergence response to stereoscopic cuts. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 33(4), 2014. 2, 7

[61] A. Turner, J. Berry, and N. Holliman. Can the perception of depth in stereoscopic images be influenced by 3d sound? In *Stereoscopic Displays and Applications XXII*, vol. 7863, pp. 71–80. SPIE, 2011. 1, 2, 3

[62] R. J. van Beers. Saccadic eye movements minimize the consequences of motor noise. *PloS one*, 3(4):e2070, 2008. 2

[63] N. Van der Stoep, C. Spence, T. Nijboer, and S. Van der Stigchel. On the relative contributions of multisensory integration and crossmodal exogenous spatial attention to multisensory response enhancement. *Acta Psychologica*, 162:20–28, 2015. 3

[64] M. Wu, Y. F. Cheng, and D. Lindlbauer. New ears: An exploratory study of audio interaction techniques for performing search in a virtual reality environment. In *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 386–395. IEEE, 2024. 1

[65] Q. Yang, M. P. Bucci, and Z. Kapoula. The Latency of Saccades, Vergence, and Combined Eye Movements in Children and in Adults. *Investigative Ophthalmology & Visual Science*, 43(9):2939–2949, Sept. 2002. 2

[66] Q. Yang and Z. Kapoula. Saccade–vergence dynamics and interaction in children and in adults. *Experimental Brain Research*, 156(2):212–223, 2004. 2

[67] P. Zahorik. Auditory and visual distance perception: The proximity-image effect revisited. *Journal of The Acoustical Society of America - J ACOUST SOC AMER*, 113:2270–2270, 2003. 2

[68] P. Zahorik, D. Brungart, and A. Bronkhorst. Auditory distance perception in humans: A summary of past and present research. *Acta Acustica united with Acustica*, 91:409–420, 2005. 2, 3

[69] D. Zambarbieri. The latency of saccades toward auditory targets in humans. *Progress in Brain Research*, 140:51–59, 2002. 4

[70] D. S. Zee, E. J. Fitzgibbon, and L. M. Optican. Saccade-vergence interactions in humans. *Journal of Neurophysiology*, 68(5):1624–1641, 1992. 2

[71] Z. Zhou, A. D. Cheok, Y. Qiu, and X. Yang. The role of 3-d sound in human reaction and performance in augmented reality environments. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 37(2):262–272, 2007. 2