Overdriving Visual Depth Perception via Sound Modulation in VR - Supplementary Material

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

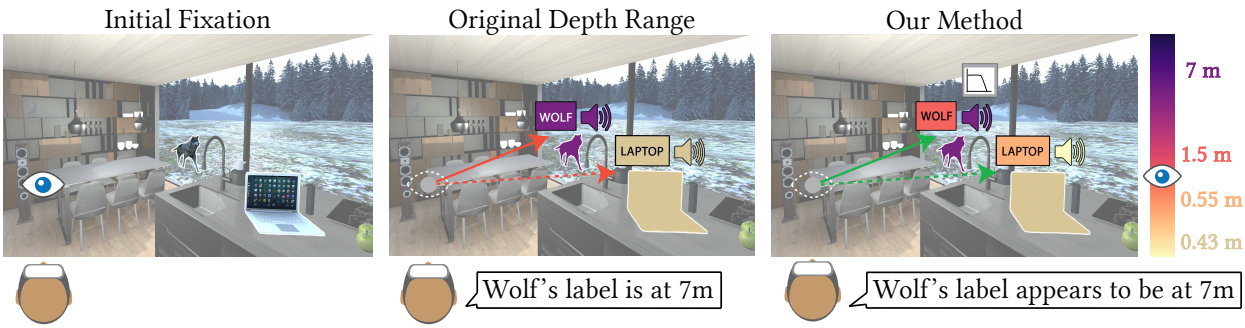

Fig. 1: This paper investigates the effect of spatially decoupling auditory and visual cues on depth perception in virtual reality (VR). *Left*: The user begins by fixating on some random object within a VR scene. *Center*: A new annotation (e.g., related to the wolf) appears in the distance, accompanied 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

This supplementary document contains the following sections:

- S1. Hardware details
- S2. Vergence range
- S3. Additional details for the visual response experiment
- S4. Additional details for the analyses performed in the main experiment, visual response experiment, and application case
- S5. Application case - Mixed Reality version
- S6. Experiment survey

S1. Hardware Details

Tab. 1 shows details of the HMD used in the experiments while Tab. 2 contains the details of the headphones.

Table 1: Specifications of the Varjo Aero HMD.

Display Resolution	2880 × 2720 pixels per eye
Display Refresh Rate	90 Hz
Eye Tracking Frequency	200 Hz
Eye Tracking Accuracy	Sub-Degree

Table 2: Specifications of the Beyerdynamic DT 770M headphones.

Frequency Response	5 Hz - 30 kHz
Noise Cancellation Technology	Passive
Nominal Impedance	80 Ohm
Sensitivity	105 dB/mW

S2. Vergence Range

The whole vergence range considered in this work is displayed in Fig. 2 for a direct visualization.

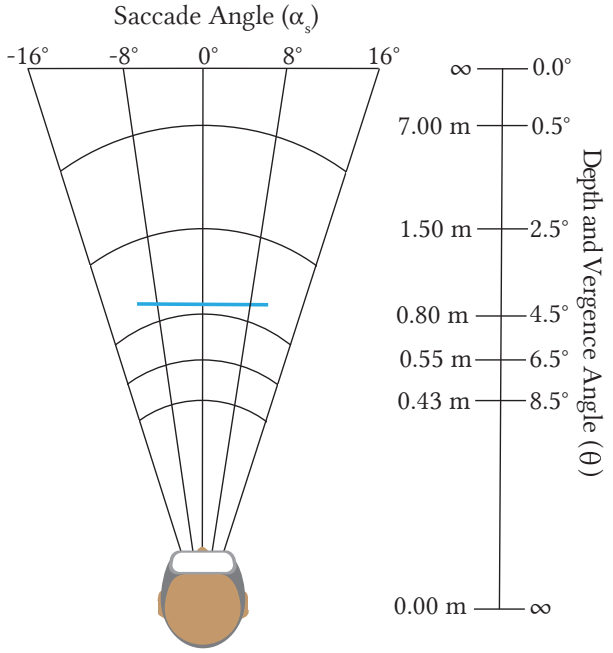

Fig. 2: Vergence range considered in this work. The vergence angle formed between both gaze rays depends on the depth of the visual target that is being fixated on. The closest distance is set to 0.43 m ($\theta = 8.5°$) and the farthest limit is set to 7 m ($\theta = 0.5°$), creating a vergence range of 8°. The focal distance of the Varjo Aero HMD at 0.85 m is highlighted (light blue). Scheme adapted from [1].

S3. Visual Response Acceleration Experiment

Throughout the visual response acceleration experiment, a step scenario is always employed between the offset of the previous visual anchor and the onset of the new audiovisual cue. This setup avoids variations in visual response latency typical from other conditions such as gap or overlap modes between visual stimuli [3,4]. Therefore, saccade latency remains unaffected in our conditions as well as saccade response time since are associated to visual eccentricities [2].

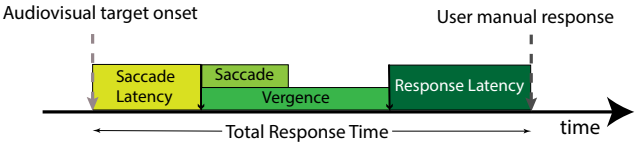

Fig. 3: Temporal representation of response time measured. Target onset indicates the start of the visual response, covering saccade latency and saccade execution time, as well as vergence execution time [1]. Additionally, button press latency is also considered in the total response time.

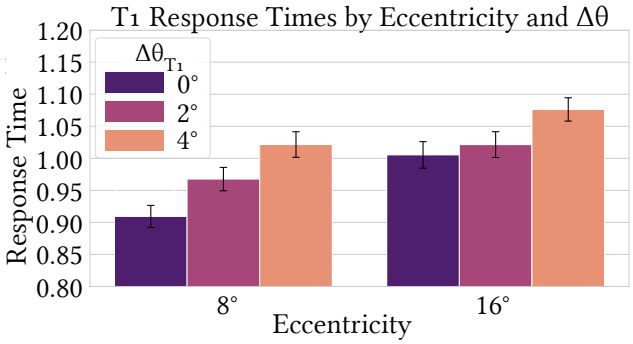

Fig. 4: Manual response times for T_1. Response times increase with larger variations of vergence angle $\Delta\theta_{T_1}$ and larger eccentricities, while error bars show standard error of the mean (SEM).

Manual response delay time acts as an offset for all measurements, so the response time variations measured are associated to temporal differences in vergence motion. We do observe in Fig. 3 the temporal scheme of the total response time measured, considering visual saccade, vergence and manual response.

In Fig. 4 and 5, we can also find timing differences for T_1 and T_2 targets employed in the experiment. Same observations can be found for these cases than for the reported $T_1 + T_2$ combination in the main manuscript. Larger variations in vergence angle cause slower response times, while higher eccentricities also slow down the visual response due to longer times in saccade latency [2].

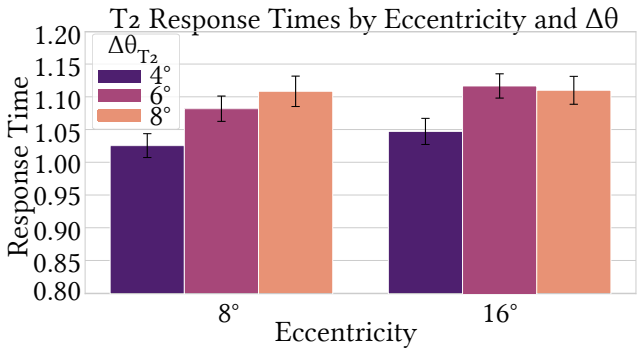

Fig. 5: Manual response times for T_2. Response times increase with larger variations of vergence angle $\Delta\theta_{T_2}$ and larger eccentricities, while error bars show standard error of the mean (SEM).

S4. Statistical Analyses Details

In this section, we include the tables corresponding to all pairwise comparisons in the statistical analysis for the main experiment, the

visual response experiment and the final exploring application. In Tab. 3 we show the effect of each factor on the success ratio in the main analysis. Following the results of that table, we run post-hocs analysis for statistically significant differences between *loudness shift* (ΔL) levels. Pairwise comparisons for each sound *loudness shift* considered are shown in Tab. 4. We do not run post-hocs for the *vergence direction* (V), *vergence range* (R) and *eccentricity* (E) since they have only two levels.

Table 3: Statistical analysis of the results of the main experiment. Effects of the analyzed factors on success ratio.

Factor	z-ratio	p-value
Matching Depth	-2.047	0.0406
Overdriving level 1	-5.046	$<.0001$
Overdriving level 2	-6.080	$<.0001$
Overdriving level 3	-7.473	$<.0001$
Vergence Direction	6.102	$<.0001$
Eccentricity	-3.288	0.001
Vergence Range	3.447	$<.0001$
Side	-0.390	0.6962

Table 4: Statistical analysis of the results of the main experiment. Post-hocs for the *loudness shift* factor (ΔL) on success ratios.

Comparison	z-ratio	p-value
Colocated - Matching Depth	2.047	0.243
Colocated - Overdriving Level 1	5.046	$<.0001$
Colocated - Overdriving Level 2	6.080	$<.0001$
Colocated - Overdriving Level 3	7.473	$<.0001$
Matching Depth - Overdriving Level 1	3.034	0.020
Matching Depth - Overdriving Level 2	4.089	$<.001$
Matching Depth - Overdriving Level 3	5.516	$<.0001$
Overdriving Level 1 - Overdriving Level 2	1.069	0.822
Overdriving Level 1 - Overdriving Level 3	2.523	0.085
Overdriving Level 2 - Overdriving Level 3	1.457	0.590

Binomial test comparisons are performed between all sound modulation levels and chance level (0.5) for each variable considered in the experiment, in order to find out if the proportion of success ratios for each condition is significantly different from chance level. This is shown in Tab. 5, 6, 7, 10, 9, 10.

Table 5: Statistical analysis of the results of the main experiment. Binomial test between *loudness shift* (ΔL) cases and Chance Rate for $E = 8$ condition.

Sound Level	p-value
Colocated	$<.0001$
Matching Depth	$<.0001$
Overdriving Level 1	$<.0001$
Overdriving Level 2	$<.0001$
Overdriving Level 3	0.0018

Regarding the visual response experiment, the results are shown in Tab. 11, 12, 13 showing the effect of each variable in the response times recorded for the first target (T_1), second target (T_2) and the combination of both ($T_1 + T_1$) respectively.

Table 6: Statistical analysis of the results of the main experiment. Binomial test between *loudness shift* (ΔL) cases and Chance Rate for E=16 condition.

Sound Level	p-value
Colocated	$<.0001$
Matching Depth	$<.0001$
Overdriving Level 1	$<.001$
Overdriving Level 2	0.0332
Overdriving Level 3	0.113

Table 7: Statistical analysis of the results of the main experiment. Binomial test between *loudness shift* (ΔL) cases and Chance Rate for Convergence condition.

Sound Level	p-value
Colocated	$<.0001$
Matching Depth	$<.0001$
Overdriving Level 1	0.00125
Overdriving Level 2	0.0433
Overdriving Level 3	0.702

Table 8: Statistical analysis of the results of the main experiment. Binomial test between *loudness shift* (ΔL) cases and Chance Rate for Divergence condition.

Sound Level	p-value
Colocated	$<.0001$
Matching Depth	$<.0001$
Overdriving Level 1	$<.0001$
Overdriving Level 2	$<.0001$
Overdriving Level 3	$<.0001$

Table 9: Statistical analysis of the results of the main experiment. Binomial test between *loudness shift* (ΔL) cases and Chance Rate for Near range condition.

Sound Level	p-value
Colocated	$<.0001$
Matching Depth	$<.0001$
Overdriving Level 1	$<.0001$
Overdriving Level 2	0.0118
Overdriving Level 3	0.0139

Table 10: Statistical analysis of the results of the main experiment. Binomial test between *loudness shift* (ΔL) cases and Chance Rate for Far range condition.

Sound Level	p-value
Colocated	$<.0001$
Matching Depth	$<.0001$
Overdriving Level 1	$<.0001$
Overdriving Level 2	$<.0001$
Overdriving Level 3	0.0251

Finally, in Tab. 14 we show the results of the application case in evaluating factor levels significance on the success ratio of selecting the scene item colocated with the annotation. Binomial test comparisons between each condition's success ratio and chance level (0.5) are also shown in Tab. 15.

Table 11: Statistical analysis of the results of the validation test. Effects of the analysed factors on response times for T_1, first target appearing.

Factor	z-ratio	p-value
Δ2°	2.732	0.006
Δ4°	6.196	< .0001
Eccentricity	5.895	< .0001
Side	1.504	0.133
First Vergence Direction	0.644	0.519

Table 12: Statistical analysis of the results of the validation test. Effects of the analysed factors on response times for T_2, second target appearing.

Factor	z-ratio	p-value
Δ6°	4.675	< .0001
Δ8°	4.979	< .0001
Eccentricity	1.796	0.072
Side	-0.660	0.509
First Vergence Direction	-0.766	0.444

Table 13: Statistical analysis of the results of the validation test. Effects of the analysed factors on response times for $T_1 + T_2$, considering total trial response time.

Factor	z-ratio	p-value
Δ8°	4.649	< .0001
Δ12°	7.004	< .0001
Eccentricity	4.871	< .0001
Side	0.547	0.0.585
First Vergence Direction	-0.066	0.948

Table 14: Statistical analysis of the results of the application case. Effects of the analysed factors on the success ratio of selecting the scene item at the same visual depth than the annotation.

Factor	z-ratio	p-value
Sound Level	-10.375	< .0001
Vergence Direction (D)	0.945	0.3449
Eccentricity (16)	0.601	0.5475

Table 15: Statistical analysis of the results of the application case. Binomial test between association ratios for each sound case and Chance Rate (0.5) for each scene item condition.

Annotation	Sound Level	t_stat	p-value
Bike	Colocated	0.2223	0.8247
Bike	Compensated	-18.3518	< .0001
Cat	Colocated	7.6024	< .0001
Cat	Compensated	0.6685	0.5058
Kid	Colocated	10.9000	< .0001
Kid	Compensated	-0.8933	0.3744
Toy	Colocated	-2.0525	0.0434
Toy	Compensated	-16.0645	< .0001

S5. Mixed Reality Version of Application

To further highlight overdriving effect capabilities on extending to MR content, we do replicate our application case in a MR scenario. Since the Varjo Aero headset lacks passthrough cameras for augmented reality, for this additional experiment we used the Varjo XR-4 headset, whose specifications are shown in Table 16. On the other hand, the scene setup is represented in Fig. 6.

Table 16: Specifications of the Varjo XR-4 HMD.

Display Resolution	3840 × 3744 pixels per eye
Display Refresh Rate	90 Hz
Passthrough Cameras	2 x 20 Mpx
Eye Tracking Frequency	200 Hz
Eye Tracking Accuracy	Sub-Degree

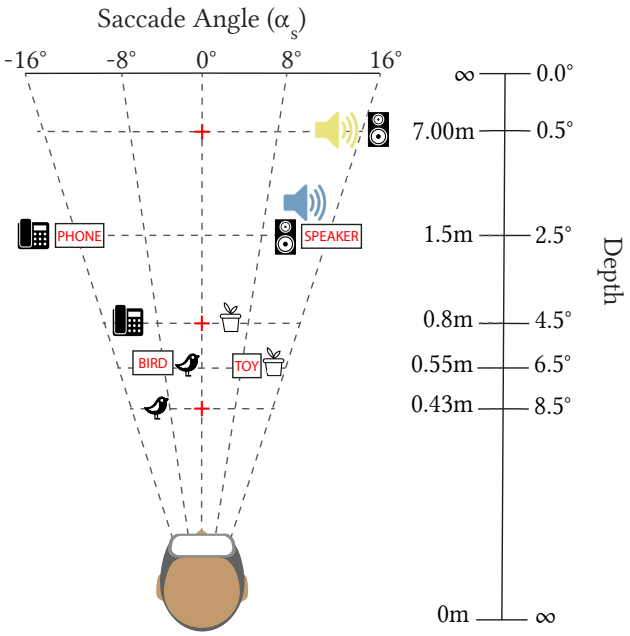

Fig. 6: Spatial placement of the scene items and their corresponding annotations in the MR experiment. As in the VR case, colocated (blue) and overdriving (yellow) sound conditions are presented while annotations are in a single location for each scene item. Initial fixations are red crosses.

New scene items are used to cover the whole vergence range such as birds, toys, phones and speakers. Regarding audio cues, they are chosen according to the scene item: a bird song for **BIRD**; a squeak for **TOY**; a phone vibration for **PHONE**; and a vinyl needle skip for **SPEAKER**. As in the VR application, annotations are presented for each scene item and participants need to choose the scene item to which the annotation refers to. Same cases are presented under the procedure explained in for the application case in the main manuscript.

A table head-chin rest was used to avoid users performing head rotations during the experiment, making sure that annotations appeared at the intended angles, aligned with the real-world objects. The chin-rest height was fixed to 35 cm from the scene items surface for all users, equalizing perspective and height of the annotations across participants.

In this case, 5 new participants (male, avg. age = 27.4, SD = 3.2) performed the experiment and the results are shown in the main manuscript. Similar trends can be seen as the ones explained in the VR application, highlighting how the overdriven effect still works in a MR scene, where only annotations are rendered content and under a less controlled scenario (luminance changes, head jitter, passthrough latency... etc.). However, it is worth pointing out some differences: while the object occupying the furthest position in the VR experiment (**BIKE**) resulted in significantly different results than the two middle ones (**KID** and **CAT**), this is no longer the case for the MR application. Similarly, the trend seen for the closest object in the MR experiment (**BIRD**) shifts towards the rest, with the accuracy rate rising for both the colocated and overdriven sound in comparison to its VR counterpart (**TOY**). This indicates more consistent behavior across all objects in the MR setup, with overdriven audio still strongly affecting perceived depth.

Statistical Analysis results of this MR application version are shown in Table 17 and 18. As can be seen, the sound level is statistically significant in associating annotations to scene items. Additionally, comparisons with the chance level show statistical significance on **TOY**, **PHONE** and **SPEAKER** colocated cases, while the overdriven counterparts manage to mask visual depth.

Table 17: Statistical analysis of the results of the application case in MR. Effects of the analysed factors on the success ratio of selecting the scene item at the same visual depth as the annotation.

Factor	z-ratio	p-value
Sound Level	-6.749	$< .0001$
Vergence Direction (D)	2.280	0.023
Eccentricity (16)	3.257	0.001

Table 18: Statistical analysis of the results of the exploring application in MR. Binomial test between association ratios for each sound case and Chance Rate (0.5) for each scene item condition.

Annotation	Sound Level	t_stat	$p-value$
Bird	Colocated	0.9475	0.3492
Bird	Compensated	-2.3333	0.0249
Phone	Colocated	10.0767	$< .0001$
Phone	Compensated	1.2748	0.2099
Speaker	Colocated	10.0767	$< .0001$
Speaker	Compensated	0.3126	0.7562
Toy	Colocated	8.3267	$< .0001$
Toy	Compensated	0.6276	0.5339

S6. Experiment survey

In this section, questions and explanations presented to participants are shown, as well as response options provided in the demographic survey completed before the experiment session.

E1. How does the text work?

Each trial consists on different audiovisual stimuli being presented to you under different conditions and your task is to look at all of them. A traffic cone is used as visual target while the auditory cue is provided by a beep. The trial always starts by pressing the space bar. First of all, one traffic cone will appear at the center of the screen, we refer to it as the fixation point. After some time looking at this central traffic cone, it will disappear and another one appears at one side. As soon as you detect or perceive the visual target, you have to look at it, so perform the saccade from the fixation point to the visual target. After a while, the visual target disappears and the fixation point spawns again, so you look at the fixation point. Then, the second traffic cone appears on the opposite side, making you to saccade towards it as happened with the first one. Finally, this second traffic cone disappears and the central one appears, returning to the initial situation. After this procedure, you need to decide which traffic cone was closer to you in depth considering both visual and auditory information, the one on the right or the one on the left. To answer this, just press the left or right arrow on the keyboard accordingly. After providing this answer, press the space bar again to start the new trial. Each condition has 30 trials and there are 8 conditions, after which the experiment ends. If you have any questions, please ask the experimenter now.

☐ *Understood*

E2. Consent for participation in the study

I agree to participate in the research study. I understand the purpose and nature of this study and I am participating voluntarily. I understand that I can withdraw from the study at any time, without any penalty or consequences. I grant permission for the data generated from this questionnaire to be used in the researcher's publications on this topic. The generated data will be stored anonymously under a randomly generated unique ID. Any information that is obtained in connection with this study and that may be identified with you will remain confidential and will be disclosed only with your permission.

☐ *I agree*

Q1. Subject anonymous ID

_____

Q2. Age

_____

Q3. Gender

☐ *Male* ☐ *Female* ☐ *Rather not to say* ☐ *Other*

Q4. Home Country

_____

Q5. Do you have any visual impairments?

☐ *Yes* ☐ *No*

Q6. If you answered "Yes" to the previous question, please specify your condition (e.g. poor distance vision):

_____

Q7. If you have any visual impairments, do you have it corrected?

☐ *Yes* ☐ *No*

Q8. If you answered "Yes" to the previous question, please specify how you have it corrected (e.g. glasses):

_____

Q9. Do you have any auditory impairments?

☐ *Yes* ☐ *No*

Q10. If you answered "Yes" to the previous question, please specify your condition (e.g. age-relating hearing loss):

_____

Q11. If you have any auditory impairments, do you have it corrected?

☐ *Yes* ☐ *No*

Q12. If you answered "Yes" to the previous question, please specify how you have it corrected (e.g. ear-mounted device):

_____

Q13. Do you play video games?

☐ *No* ☐ *Yes, sporadically* ☐ *Yes, often* ☐ *Yes, everyday*

Q14. Specify your experience with Virtual Reality

☐ *None, I have never used a virtual reality device.*
☐ *Basic, I have used virtual reality devices less than 5 times.*
☐ *Experienced, I have used virtual reality devices several times.*
☐ *Professional, I use virtual reality devices on a daily basis.*

Q15. If you have already tried virtual reality, please specify those that apply

☐ *I have tried desktop-based devices like Oculus, HTC Vive, or Play Station VR.*
☐ *I have tried smartphone-based devices like Google Cardboard*
☐ *I use virtual reality devices everyday*
☐ *I suffered fatigue, dizziness or eyestrain when using virtual reality devices*

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; and by an academic gift from Meta. The authors would like to thank the participants of the experiments for their participation.

REFERENCES

[1] 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

[2] 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. 2

[3] R. Kalesnykas and P. Hallett. The differentiation of visually guided and anticipatory saccades in gap and overlap paradigms. *Experimental Brain Research*, 68:115–121, 1987. 2

[4] A. Kingstone and R. M. Klein. Visual offsets facilitate saccadic latency: does predisengagement of visuospatial attention mediate this gap effect? *Journal of Experimental Psychology: Human Perception and Performance*, 19(6):1251, 1993. 2