From Ears to Eyes: The Impact of Augmented Text Integration on Audio Guidance

Oliver Hein[1], Sude Gökçel[2], and Florian Alt[2]

[1] University of the Bundeswehr Munich, Germany,
oliver.hein@unibw.de,
[2] LMU Munich, Germany,
sude.goekcel@campus.lmu.de, florian.alt@ifi.lmu.de

Abstract. As AR guidance technologies advance, delivering clear and effective instructions remains a key challenge. Audio-only guidance can lead to misunderstandings and increase cognitive load. This study investigates whether adding visual elements (text or symbols) to audio instructions on AR glasses improves user performance and experience. In a user study ($n = 9$), participants completed drawing tasks under three conditions: `Audio`, `Audio+Text`, and `Audio+Symbols`. We measured task time, accuracy, replay count, and user satisfaction via the UEQ-S. Both augmented methods significantly reduced replays and improved satisfaction compared to audio-only, while task time and accuracy remained similar. Text and symbols proved equally effective, offering flexible options for AR instruction design.

Keywords: Guidance · Augmented Reality · Mixed Reality.

1 Introduction

As augmented reality (AR) becomes more widespread, presenting information clearly without overwhelming users is essential. AR headsets require users to split attention between real and virtual environments, increasing cognitive load and impacting performance [7]. To address this, researchers have integrated speech recognition (SR) with AR to convert spoken instructions into visual content. However, few studies have systematically compared how different visualizations affect performance and user experience when paired with audio. This study examines whether augmenting audio instructions with text or symbols on AR glasses improves task performance and user satisfaction over audio-only guidance. We investigate (1) the overall impact of visual augmentation and (2) whether text or symbols are more effective. Participants completed drawing tasks under three conditions: `Audio`, `Audio+Text`, and `Audio+Symbols`. We measured task time, replay count, accuracy, and user satisfaction (UEQ-S). Both visual methods significantly reduced replay rates and were rated more favorably than audio-only instructions, though task time, accuracy, and differences between the two visual methods were not statistically significant. These findings suggest visual augmentation enhances clarity and satisfaction in AR without affecting efficiency, offering flexible design options for AR instruction systems.

2 Related Work

2.1 Integration and Application Scenarios of AR and SR

In education, several projects have demonstrated the benefits of combining AR and SR for language learning. For example, Wibowo et al. [17] developed an AR-based English learning tool for children aged 6–10, featuring 3D virtual objects and real-time pronunciation feedback.

Similarly, Tsai [14] showed that mobile AR paired with SR significantly improved oral English skills and learning perceptions in a study with 90 university students. Che Dalim et al. [2] also found that AR systems with speech input helped 120 young learners improve vocabulary acquisition, task completion, and enjoyment compared to traditional methods.

Beyond education, AR-SR systems have also been developed for communication support. Mirzaei et al. [8] created an AR tool for the deaf and hard of hearing, using SR to transcribe speech into real-time text near the speaker's face. This was well-received as an alternative to sign language.

Likewise, Watanabe et al. [15] developed smart glasses with multiple microphones to enhance voice direction detection and display transcribed speech with over 90% accuracy in controlled settings.

The integration of AR and SR has proven beneficial in education and accessibility, however, research focused on task performance or related to radio communication remains limited.

2.2 Enhancing Text Visibility in AR Interfaces

Existing studies have investigated the impact of font types, color contrast, text placement, and layout adaptability in AR environments. Gabbard et al. [5] noted the lack of standardized design guidelines for AR user interfaces. However, progress has been made through individual studies.

For example, Agić et al. [1] and Gattullo et al. [6] found that high-contrast combinations, such as white text on black or blue backgrounds, significantly improve legibility. They also recommended adding billboard backgrounds to improve readability against complex scenes.

Further research by Erickson et al. [3] showed that positive contrast (light text on dark backgrounds) is generally preferred, as it supports better physiological comfort on optical see-through displays.

Text placement has also been a focus. Rzayev et al. [10] compared three positions (top right, center, bottom) and two presentation styles for see-through smart glasses. Their study revealed that top-right placement resulted in lower comprehension and higher workload, while center placement demanded full attention, making it unsuitable for multitasking.

A dynamic approach by Orlosky et al. [9] adapted text placement based on lighting, aiming to mimic natural user behavior by placing text in darker visual areas. However, Klose et al. [7] critiqued this method for violating consistency in information presentation. They suggested placing text at the top in cluttered environments and at the bottom when more sustained attention is required.

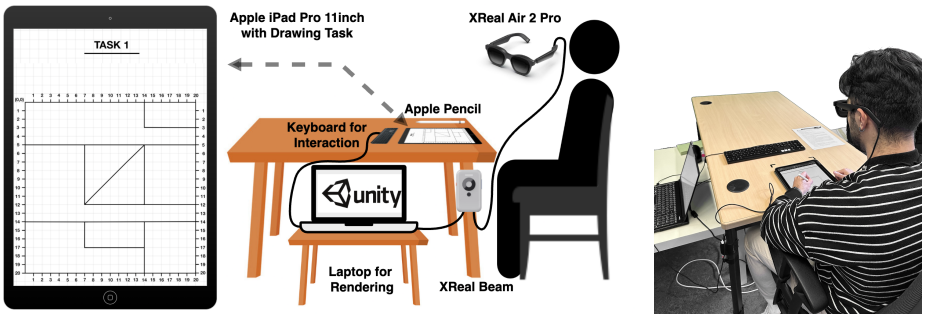
3 Methodology

This study investigates how different methods of delivering instructions affect task performance and user satisfaction when using AR glasses. To isolate the effects of instruction delivery without the limitations of current speech recognition (SR) systems, a 'Wizard of Oz' approach was used [11]. Participants received pre-recorded audio instructions along with pre-written text, simulating the output of a real-time SR system. The study aimed to answer two research questions:

RQ1 How does the addition of text-based visual instructions to audio guidance on AR glasses impact user performance and satisfaction?

RQ2 In the context of AR visually assisted tasks, how do text-based instructions compare to symbolic representations in terms of performance and preference?

We used a within-subject design, with each participant completing tasks under three conditions: `Audio`: Only audio instructions through AR glasses speakers, `Audio+Text`: Audio instructions plus corresponding written text, and `Audio+Symbols`: Audio instructions plus symbolic representations of each step instead of full sentences. These visuals are grounded in prior research suggesting that symbols can enhance comprehension, particularly in constrained or cross-cultural settings [12, 13, 16]. Participants performed drawing tasks on an iPad using an Apple Pencil. Each task involved drawing and erasing horizontal, vertical, and diagonal lines on a 20×20 grid, resulting in an abstract geometric shape (see Figure 1a). To control for learning effects, three distinct drawing patterns were created, each requiring twelve steps with equal complexity but differing in shape and sequence. In all conditions, participants could navigate between or replay instructions using arrow keys on a keyboard. The order of conditions was counterbalanced using a Latin square. Performance metrics included task completion time, replay count, and drawing accuracy, measured via image comparison. To assess subjective feedback, participants completed the short User Experience Questionnaire (UEQ-S) after each task.



(a) Sketch of the setup (b) The real setup

Fig. 1: Study setup

(a) Audio+Text (b) Audio+Symbols

Fig. 2: AR display content (black appears transparent within the AR glasses)

3.1 Apparatus

The study setup, shown in Figure 1, included an iPad Pro (11 inch) and an Apple Pencil for the drawing task. Audio instructions and visual content were delivered through XReal Air 2 Pro AR glasses with built-in speakers. These glasses were connected to the XReal Beam, which linked to a computer to project screen content spatially. A keyboard allowed participants to navigate between instructions using arrow keys. We developed a custom Unity application to present instructions in different formats on the AR glasses (see Figure 2). White text is placed head-locked in the upper middle section of the visual field to remain clearly visible without obstructing the drawing task [10, 7, 3]. In the `Audio+Symbols` condition, we used a minimal set of task-specific Unicode icons designed for clarity and spatial guidance ('←↑→↓↙↖↗↘','×' for erase). Symbols were based on common interface and signage conventions, but no formal validation was conducted. We generated audio files for spoken instructions using the tool 'Voicebooking'[3].

3.2 Procedure

The study followed the process shown in Figure 3. Upon arrival, participants were briefed on the tasks and data collection, completed a consent form, and filled out a demographic questionnaire (age, gender, visual conditions, AR familiarity). They sat down at the desk, put on the AR glasses and received general drawing instructions. A five-minute practice session ensured they were comfortable with the tools and task format. During the experiment, participants completed drawing tasks under the three instruction conditions in counterbalanced order. Each task involved twelve instructions to draw and erase lines. After each task, participants completed the UEQ-S to evaluate the instruction method. At the end, they ranked the methods by preference, gave optional feedback, and provided personal details for compensation. Sessions lasted about 45 minutes.

[3] `https://www.voicebooking.com/`

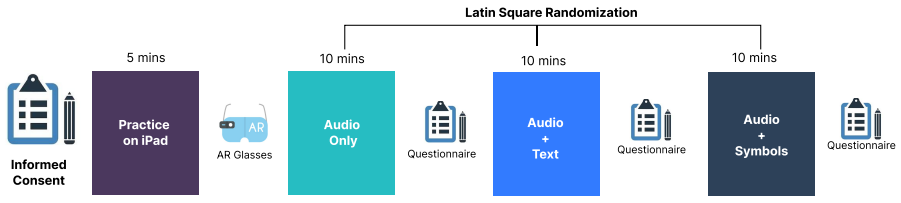


Fig. 3: Study procedure

3.3 Participants

Participants were recruited over a two-week period. Recruitment methods included emails through the central information service of our university, as well as invitations on university's Slack channel. Participants were reimbursed with a compensation of 12 Euros. Participants were required to have no major visual impairments and not to wear glasses. However, they were eligible to participate if they wear contact lenses. The study had 9 participants. 6 of them identified as female and 3 as male ($Age\ Mean = 23.11$, $SD = 0.56$). 5 participants had corrected eye sight, 2 participants had not corrected full eye sight, and 2 participants reported their eye sight as *'unknown'*. 5 out of 9 participants reported to have no experience familiarity with AR or VR technologies and the remaining 4 participants rated their familiarity as *'minimal experience'*.

3.4 Limitations

This study used a Wizard of Oz setup to simulate speech recognition (SR), enabling consistent, error-free instructions. While this isolates the effect of visual augmentation, it omits real-world SR challenges like delays and noise, limiting ecological validity.

The small, homogeneous sample (n = 9) reduces statistical power and generalizability. Some participants lacked corrected vision, which may have impacted their ability to interpret visual content, especially symbols.

Symbols were not formally validated, and some confusion was reported, suggesting a need for user-centered design in future work. The drawing task offers control but may not reflect the complexity of real-world AR scenarios such as industrial or educational applications.

Finally, the image comparison method for accuracy was overly sensitive to minor misalignments; more robust metrics should be considered in future studies.

4 Results

4.1 Objective Data Analysis

The small sample size resulted in a non-normal distribution of the data, therefore we performed a non-parametric Friedman's test to test the significance of the differences. Additionally, we applied Wilcoxon signed-rank test with Bonferroni correction for the post-hoc analysis [4].

Task Completion Time. We compared the median completion times and interquartile range (IQR). The Audio method had a median of 303.81 seconds with IQR = 126.70, indicating a relatively consistent time range among participants. The Audio+Text method showed a slightly lower median completion time of 280.84 seconds, with a wider variability (IQR = 165.57). The Audio+Symbols method had a median of 286.43 seconds and an IQR of 150.58. We conducted a Friedman test, which revealed no statistically significant difference in completion times across methods ($p = .45$).

Replay Count. During task execution, we tracked if participants would replay or revisit instructions. We analyzed the sum, median, and interquartile range (IQR) of this data. The `Audio` condition resulted in 54 total replays, with a median of 5 and an IQR of 5, indicating high replay frequency and variability. `Audio+Text` showed the lowest replay rate, with only 2 total replays, a median of 0, and no variability (IQR = 0), suggesting consistently clear instructions. `Audio+Symbols` had 11 replays in total, a median of 1, and an IQR of 2, reflecting moderate replay rates and variability.

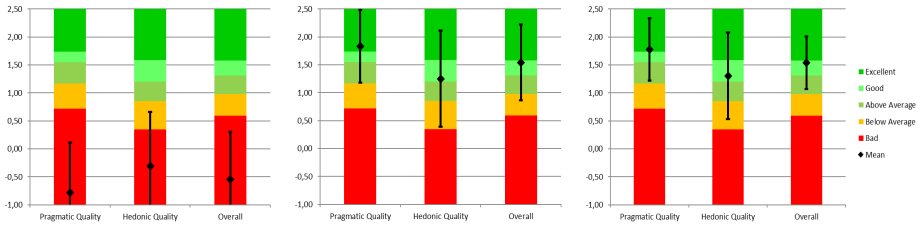
Friedman's test confirmed a significant difference between conditions ($p = .0006$). Wilcoxon signed-rank tests showed significantly fewer replays for Audio+Symbols ($p = .01$) and Audio+Text ($p = .0039$) compared to Audio-only. However, the difference between Audio+Text and Audio+Symbols was not statistically significant ($p = .07$).

Accuracy. We measured accuracy as the percentage overlap between participants' final drawings and the original, reflecting how closely they followed instructions. To remove gray grid lines, we applied a binary threshold, isolating drawn lines. An erosion filter (9×9 kernel) increased line thickness to account for minor pixel misalignments. Accuracy was then calculated as the percentage of overlapping black pixels between the reference and participant drawings.

The `Audio` condition had a median accuracy of 61.93% (IQR = 13.31), `Audio+Text` showed slightly higher accuracy at 64.7% with lower variability (IQR = 5.37), and `Audio+Symbols` had a median of 62.49% (IQR = 12). Despite minor differences, Friedman's test showed no statistically significant differences among conditions ($p = .64$).

4.2 Subjective Data Analysis

Subjective data is based on the responses from the UEQ-S, where users evaluated the different aspects of the three conditions. Additionally, we collected feedback on users' preference rankings of the three instruction methods and their reasoning behind these statements.



(a) Audio (b) Audio+Text (c) Audio+Symbols

Fig. 4: UEQ-s results with pragmatic, hedonic and overall qualities for each condition

Table 1: Post-hoc test results for each UEQ-S attribute ($p \leq .0166$ marked with $\star$)

	Audio vs Audio+Text	**Audio vs Audio+Symbols**	**Audio+Text vs Audio+Symbols**
Supportive/Obstructive	.0078 $\star$	.03	.21
Easy/Complicated	.0117 $\star$	.0117 $\star$	1.0
Efficient/Inefficient	.0078 $\star$	.0078 $\star$	.08
Clear/Confusing	.0039 $\star$	.03	.44
Exciting/Boring	.03	.09	1.0
Interesting/Not Interesting	.0164 $\star$	.02	.32
Inventive/Conventional	.21	.07	.58
Leading Edge/Usual	.07	.11	.70

UEQ-S. As shown in Figure 4, `Audio+Text` and `Audio+Symbols` received significantly higher user experience ratings than `Audio`. The `Audio` method scored poorly, with mean pragmatic and hedonic ratings of -0.77 and -0.30, respectively (overall: -0.54). In contrast, `Audio+Text` scored 1.83 for pragmatic quality and 1.25 for hedonic quality, while `Audio+Symbols` scored 1.77 and 1.30, respectively. Both augmented methods had an overall rating of 1.54, indicating *'Good'* to *'Excellent'* perceived quality.

Table 1 summarizes the results of the Wilcoxon signed-rank post-hoc tests, comparing each condition across all UEQ-S attributes. Significance is indicated by p-values below the Bonferri corrected significance level ($p = .0166$). *Supportiveness*: `Audio+Text` was rated significantly more supportive than `Audio` ($p = .0078$); no significant differences between other pairs. *Ease of Use*: Both augmented methods were rated significantly easier than `Audio` ($p = .0117$); no significant difference between them. *Efficiency*: Both augmented methods outperformed `Audio` ($p = .0078$); again, no significant difference between them. *Clarity*: `Audio+Text` was received significantly clearer than `Audio` ($p = .0039$); differences with `Audio+Symbols` were not significant. For *Excitement*, augmented methods received higher ratings, but the difference was not statistically significant ($p = .066$). Regarding *Interest*, `Audio+Text` was rated significantly more interesting than `Audio` ($p = .0164$). For both *Inventiveness* and *Leading-edge*, the augmented methods were rated higher, but differences were not statistically significant ($p = .11$ and $p = .28$, respectively).

Overall, participants found the augmented methods more supportive, easier, and clearer, with `Audio+Text` slightly favored, though not always significantly, over `Audio+Symbols`.

Ranking and Feedback. `Audio+Text` and `Audio+Symbols` were most preferred, each ranked first by 4 out of 9 participants, while `Audio` was least favored, ranked last by 5 participants. Notably, `Audio+Text` was never ranked last. Feedback showed that symbols were concise and helpful for spatial guidance but sometimes confusing due to mismatches with the audio. `Audio+Text` was praised for its clarity and alignment with audio, making it easier to follow. While some appreciated the coordinate information, others found it mentally taxing. Overall, `Audio` was seen as the least effective due to the lack of visual support.

5 Discussion

This study examined whether augmenting audio instructions in AR with text or symbols improves task performance and user satisfaction. Both visual methods significantly reduced instruction replays and improved subjective experience compared to audio-only guidance, though no significant differences were found in task time or drawing accuracy.

RQ1: Visual augmentation improves perceived clarity and satisfaction without affecting task efficiency. Participants rated both augmented conditions as more supportive, efficient, and easier to follow than audio-only guidance. Text was slightly preferred for its alignment with spoken instructions, while symbols were appreciated for compactness and spatial cues, though occasionally seen as unclear when not well matched to audio.

RQ2: No significant difference between text and symbols, though text was favored for clarity. Although performance metrics did not differ significantly, users preferred text for its semantic precision. These findings align with prior work showing that visual text improves clarity and reduces cognitive load [7, 3, 6, 1], while symbols can support spatial understanding but risk ambiguity without careful design [12, 13]. The lack of performance gains echoes studies suggesting these may only emerge in more complex or sustained tasks [17].

The Wizard of Oz setup ensured consistency but excluded real-world SR issues like latency and errors, limiting ecological validity. The small, homogenous sample (n = 9) further reduces statistical power and generalizability, possibly obscuring differences between visual methods. Uncorrected or unknown vision among participants may have also influenced perception of visual cues.

Symbol clarity was another limitation—icons were not validated, and some participants reported confusion. Future work should include user-centered symbol design and validation. The drawing task, while controlled, may not reflect real-world AR use cases; future studies should examine instruction support in applied settings like manufacturing, education, or remote collaboration. Additionally, eye-tracking could help reveal how users engage with visual cues during tasks, offering deeper insight into the mechanisms behind observed benefits.

In sum, visual augmentation improves instruction clarity and user satisfaction in AR without adding cognitive load. However, these results are preliminary. Larger, more diverse studies, with validated visuals, live SR integration, and gaze analysis, are needed to confirm and extend these findings.

6 Conclusion

This study found that augmenting audio instructions with text or symbols in AR significantly reduced replays and improved user satisfaction, with no impact on task time or accuracy. Both visual methods performed equally well, offering flexibility for design based on context and user needs. Future research should explore more complex AR tasks, diverse user groups, and real-time speech recognition to better reflect real-world conditions.

Bibliography

[1] Ana Agić, Ana Agić, Lidija Mandić, Lidija Mandić, Nikolina Stanić Loknar, and Nikolina Stanić Loknar. Legibility of typefaces and preferences of text/background color variations in virtual environment. *Proceedings*, 2022. https://doi.org/10.24867/grid-2022-p92.

[2] Che Samihah Che Dalim, Mohd Shahrizal Sunar, Arindam Dey, and Mark Billinghurst. Using augmented reality with speech input for non-native children's language learning. *International Journal of Human-Computer Studies*, 134:44–64, 2020. ISSN 1071-5819. https://doi.org/https://doi.org/10.1016/j.ijhcs.2019.10.002. URL https://www.sciencedirect.com/science/article/pii/S1071581918303161.

[3] A. Erickson, K. Kim, G. Bruder, and G. Welch. A review of visual perception research in optical see-through augmented reality. *ICAT-EGVE*, 2020. https://doi.org/10.2312/egve.20201256.

[4] Andy Field and Graham Hole. *How to Design and Report Experiments*. SAGE Publications Ltd, London, 2003.

[5] Joseph L. Gabbard, Joseph L. Gabbard, Missie Smith, Missie Smith, Coleman Merenda, Coleman Merenda, Gary Burnett, Gary Burnett, David R. Large, and David R. Large. A perceptual color-matching method for examining color blending in augmented reality head-up display graphics. *IEEE Transactions on Visualization and Computer Graphics*, 2020. https://doi.org/10.1109/tvcg.2020.3044715.

[6] Michele Gattullo, Michele Gattullo, Antonio Emmanuele Uva, Antonio Emmanuele Uva, Michele Fiorentino, Michele Fiorentino, Giuseppe Monno, and Giuseppe Monno. Effect of text outline and contrast polarity on ar text readability in industrial lighting. *IEEE Transactions on Visualization and Computer Graphics*, 2015. https://doi.org/10.1109/tvcg.2014.2385056.

[7] Elisa Maria Klose, Elisa Maria Klose, Nils Adrian Mack, Nils Adrian Mack, Jens Hegenberg, Jens Hegenberg, Ludger Schmidt, and Ludger Schmidt. Text presentation for augmented reality applications in dual-task situations. *IEEE Conference on Virtual Reality and 3D User Interfaces*, 2019. https://doi.org/10.1109/vr.2019.8797992.

[8] Mohammadreza Mirzaei, Mohammad Reza Mirzaei, Seyed Ghorshi, Seyed Ghorshi, Mohammad Mortazavi, and Mohammad Mortazavi. Using augmented reality and automatic speech recognition techniques to help deaf and hard of hearing people. *Virtual Reality International Conference*, 2012. https://doi.org/10.1145/2331714.2331720.

[9] J. Orlosky, K. Kiyokawa, and H. Takemura. Managing mobile text in head mounted displays: studies on visual preference and text placement. *ACM SIGMOBILE Mob. Comput. Commun. Rev.*, 18:20–31, 2014. https://doi.org/10.1145/2636242.2636246.

[10] Rufat Rzayev, Rufat Rzayev, Paweł W. Woźniak, Paweł W. Woźniak, Tilman Dingler, Tilman Dingler, Niels Henze, and Niels Henze. Reading

on smart glasses: The effect of text position, presentation type and walking. *International Conference on Human Factors in Computing Systems*, 2018. https://doi.org/10.1145/3173574.3173619.

[11] Stephan Schlögl, Gavin Doherty, and Saturnino Luz. Wizard of oz experimentation for language technology applications: Challenges and tools. *Interacting with Computers*, 27(6):592–615, May 2014. ISSN 1873-7951. https://doi.org/10.1093/iwc/iwu016. URL http://dx.doi.org/10.1093/iwc/iwu016.

[12] Xueqing Shi. Research on graphic symbol expression in visual communication design. In *Proceedings of the 2015 4th National Conference on Electrical, Electronics and Computer Engineering*, pages 444–447. Atlantis Press, 2015/12. ISBN 978-94-6252-150-6. https://doi.org/10.2991/nceece-15.2016.87. URL https://doi.org/10.2991/nceece-15.2016.87.

[13] Yulia Pavlovna Ten. Symbol as universal non-verbal means of intercultural communication in the time of globalization. *Journal of Teaching English for Specific and Academic Purposes*, 2:33–43, 2014. URL https://api.semanticscholar.org/CorpusID:55716690.

[14] Shu-Chiao Tsai. Learning with mobile augmented reality- and automatic speech recognition-based materials for english listening and speaking skills: Effectiveness and perceptions of non-english major english as a foreign language students. *Journal of Educational Computing Research*, 2022. https://doi.org/10.1177/07356331221111203.

[15] Daiki Watanabe, Y. Takeuchi, T. Matsumoto, H. Kudo, and N. Ohnishi. Communication support system of smart glasses for the hearing impaired. pages 225–232, 2018. https://doi.org/10.1007/978-3-319-94277-3_37.

[16] Steven M. Weisberg, Steven A. Marchette, and A. Chatterjee. Behavioral and neural representations of spatial directions across words, schemas, and images. *The Journal of Neuroscience*, 38:4996 – 5007, 2017. https://doi.org/10.1523/JNEUROSCI.3250-17.2018.

[17] D. Wibowo, Ika Kusumaning Putri, and Leni Saputri. Integration of augmented reality and voice recognition in learning english for children. *Journal of Applied Intelligent System*, 2022. https://doi.org/10.33633/jais.v7i2.6119.