applied sciences

MDPI

Article

Impact of Background Type and Text Orientation Angle on Glance-Based 3D Text Legibility for Mixed-Reality Applications

Ana Jović [1,*], Lea Skorin-Kapov [2] and Lidija Mandić [1]

[1] University of Zagreb, Faculty of Graphic Arts, 10000 Zagreb, Croatia; lidija.mandic@grf.unizg.hr
[2] University of Zagreb, Faculty of Electrical Engineering and Computing, 10000 Zagreb, Croatia; lea.skorin-kapov@fer.unizg.hr
* Correspondence: ana.jovic@grf.unizg.hr

Abstract

Textual information in augmented and mixed-reality (AR/MR) is often presented as two-dimensional, front-facing content. Spatially oriented 3D text provides an alternative that may allow textual information to be integrated more naturally into the surrounding environment and potentially reduce scene occlusion. However, introducing spatial orientation also raises questions about how text orientation angle and complex visual backgrounds affect text legibility. This study examined the effects of background type and text orientation angle using a glance-based methodology for assessing 3D text legibility in a controlled screen-based simulation relevant to handheld MR. Eighteen participants completed all experimental conditions (270 trials per participant) in which white 3D text was displayed for 659 ms at 15°, 30°, and 45° over three types of complex backgrounds, each represented by three images. Objective performance was assessed using accuracy and total error and complemented by Raw NASA-TLX and subjective ratings. Text orientation angle produced the strongest and most consistent effect: performance was the lowest at 15° and substantially improved at 30° and 45°, while subjective workload was the highest at 15°. Background-type effects were smaller and varied across angles and outcome measures, with no single category consistently producing the highest or lowest legibility. Considerable variability was also observed among individual background images and specific text stimuli. The findings indicate that spatial orientation is a major determinant of glance-based 3D text legibility, while the impact of background type was less consistent.

Keywords: 3D text legibility; mixed reality (MR); augmented reality (AR); glance-based legibility; complex backgrounds; text orientation; NASA-TLX; handheld AR

Check for updates

Academic Editor: Dong Hyun Jeong

Received: 28 September 2026
Revised: 24 October 2026
Accepted: 30 October 2026
Published: 3 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Text represents one of the fundamental ways of transmitting information in augmented- and mixed-reality (AR/MR) systems, where virtual information is presented in a spatial relationship with a real or simulated environment. The terminology associated with these technologies is not completely uniform. According to the reality–virtuality continuum originally defined by Milgram and Kishino (1994) [1], MR encompasses the area between a fully real and fully virtual environment, with AR representing the part of the continuum closer to the real environment. Extended reality (XR) is currently used as an umbrella term encompassing AR, MR and virtual reality (VR) [2]. At the same time, there is no completely uniform use of the term MR in the literature, and the meanings and use of the terms AR and MR partially overlap [3]. For this reason, this paper uses the designation AR/MR

when talking about systems that combine real and virtual content, while XR is used as a broader umbrella term.

In the context of XR systems, the legibility of text depends both on typographic properties such as the shape of the font, the size of the characters or their spacing, and the way the text is visually integrated with the environment. It is necessary to distinguish between the terms *readability* and *legibility*, where readability includes the ease of understanding the text while the term legibility refers to the ease of distinguishing one shape of letter from another [4,5]. In this study, we explore *legibility*, i.e., the possibility of fast and accurate recognition of individual characters, given that virtual text can be displayed over visually heterogeneous surfaces, different colors, textures and levels of illumination. In XR environments, additional difficulties arise from the characteristics of the end-user device (e.g., use of AR/MR glasses vs. a handheld device such as a smartphone), the spatial positioning of the content, and the fact that text is often not displayed frontally in relation to the observer. Kilpeläinen and Häkkinen (2023) [6] point out that text in XR rarely remains frontoparallel (i.e., front-facing) to the user, making character identification even more difficult.

The influence of the background is one of the well-recognized challenges of text legibility in MR. Unlike traditional graphical interfaces, where the designer largely control the relationship between the text and the surface on which it is displayed, AR/MR environments present backgrounds that can be unpredictable and spatially heterogeneous. This variability arises because virtual text is overlaid onto, or spatially positioned within, the real-world environment through which the user moves. Consequently, the visual properties of the background behind the text can change continuously, making it more difficult to maintain consistent text legibility. Changes in brightness, color, texture and local contrast can thus affect the visibility of different parts of the characters. Gabbard et al. (2006) [7] were among the first to systematically examine text representations over different background textures in outdoor AR and showed significant effects of background, natural lighting and the way the text is rendered on the success of the character identification task. Later work further confirmed that the relationship between text and background cannot be reduced to the nominal color of the text alone. Manghisi et al. (2017) [8], for example, showed that textured backgrounds pose a specific problem for predicting text legibility on optical see-through displays, precisely because of the interaction of virtual content with the background structure.

A similar problem occurs on conventional screens when text is displayed over complex image content. Sawyer et al. (2020) [9] examined the *glance legibility* of text over complex backgrounds, referring to the ability of users to accurately perceive and understand textual information during brief viewing periods. This concept is particularly relevant to AR/MR environments, where users may need to extract information from virtual text while simultaneously attending to the surrounding physical environment. Their study showed that various procedures for visually separating text from the background, including Gaussian blur, semi-transparent *scrim* and localized *drop shadow* procedures, can significantly improve legibility. In the AR context, Falk et al. (2021) [10] also found that legibility and readability are influenced by background properties, along with the illumination, polarity, opacity, size, and organization of the text.

These findings have led to numerous recommendations for reducing interference between text and the environment, for example, by using *billboards*, borders, shadows, or other graphic layers that increase or stabilize local contrast [11,12]. However, such an approach primarily answers the question of how to protect text from the effects of complex backgrounds, rather than addressing the extent to which different types of complex backgrounds differ in their impact on legibility. This difference is important in experimental research. If a certain type of background systematically increases or decreases the success

of text recognition, then the background represents a factor that should be controlled or explicitly included in the experimental design. On the other hand, if different types of complex scenes give comparable results under controlled conditions, it is possible to use them more flexibly without introducing significant additional variability.

The problem becomes even more complex when text is not portrayed as a two-dimensional element parallel to the plane of the screen, but is shaped as *a 3D object positioned in space*. In 3D text, perceptual conditions also depend on the spatial orientation of the characters, the visibility of their front and side surfaces, depth or extrusion, illumination, and the change in the projected geometry of the characters with the text orientation angle. Büttner et al. (2020) [13] showed that the spatial orientation of text can have a measurable effect on its legibility. In a VR environment, the authors systematically changed the rotation of 2D text from 0° to 80° in a virtual environment and determined a significant effect of the angle of rotation on the minimum size required for a legible display. The greatest deterioration was pronounced at high rotations, above 60°. More recent work further highlights the need for direct perceptual legibility measurements, as the technical specifications of XR devices alone are not sufficient to predict the actual performance of the user [6].

From a broader human-centered perspective, the usability of spatially integrated content in AR/MR not only depends on its visual legibility, but also on how virtual information is positioned, perceived, and experienced within the surrounding environment. Recent work on mobile and extended-reality applications has emphasized that user experience is multidimensional and may be influenced by perceptual and cognitive factors, spatial registration, environmental conditions, system latency, and the integration of virtual content with the physical scene [14–16]. Accordingly, evaluating spatially presented textual information benefits from combining controlled perceptual measures with subjective assessments of workload and task experience.

However, research on spatially oriented 3D text on complex backgrounds remains limited. Existing works investigating the background are mostly related to 2D text or graphic procedures for separating it from the background, while works investigating the spatial orientation of text are mainly carried out in VR or head-mounted display (HMD) systems and often use relatively simple or controlled backgrounds [11]. The research gap is particularly relevant for screen-based and handheld AR/MR. A systematic review of the literature showed a marked dominance of research conducted on HMD devices, while a significantly smaller number of studies focus on screen-based displays, despite the practical importance of smartphones and tablets. The same review identified the lack of systematic research on the legibility of 3D text and the need for research in more complex and realistic visual scenarios [12].

For the study of legibility in such conditions, the glance-based approach, which assesses the ability to recognize textual information over a short exposure time, is particularly interesting. Unlike classical measures of long-term reading, this approach allows the detection of relatively small differences in the perceptual availability of characters. Dobres et al. (2016) [17] showed that psychophysical methods with an adaptive staircase procedure can reliably detect differences in glance legibility related to the shape of the letter, size and contrasting polarity. Sawyer et al. (2017) [18] then applied the same general approach to typographic manipulations of the size, width, and shape of characters, specifying the time threshold required for the successful identification of the briefly presented text.

The glance-based approach was chosen for the analyses conducted in this study because it focuses on assessing the ability to recognize individual characters quickly and accurately, making it suitable for examining subtle changes in legibility caused by the text orientation angle and other parameters of 3D text. In order to complement the objective recognition performance by assessing the subjective burden of participants, the NASA Task

Load Index (NASA-TLX) [19] was also included in the study, along with targeted questions about subjective legibility.

The motivation is based on the fact that previous research has shown that the legibility of 2D text on complex backgrounds can be significantly improved by applying styles such as billboards, which visually separate the text from the background, but at the same time cover part of the surrounding content [11,12]. Spatially oriented 3D text could be an alternative because it can be displayed at an angle to the user, potentially occupying a smaller projected area and causing less occlusion of the environment. However, changing the angle can simultaneously reduce legibility due to the perspective shortening and change in the visibility of the front and side surfaces of the characters. Therefore, it is necessary to determine to what extent different complex backgrounds and text orientation angles together affect the legibility of 3D text.

Starting from these research gaps, the aim of this paper is to examine the influence of different types of complex backgrounds on the legibility of spatially oriented white 3D text in a controlled simulation of visual conditions relevant to handheld MR. Three types of complex backgrounds are examined—urban, natural and abstract—where each type is represented by three different image scenes to reduce the dependence of the results on one individual image. At the same time, the text displayed at three spatial angles is examined, which enables the analysis of both the main effect of the type of background and the possible relationship between the complexity of the scene and the spatial orientation of the text. Determining whether the type of complex background represents a significant and systematic source of variability is necessary to be able to distinguish the impact of the 3D text parameters from the impact of the visual environment in future studies. In this way, this research simultaneously contributes to the understanding of the relationship between 3D text and complex scenes and to the formation of a more reliable experimental framework for further research on legibility in handheld MR.

In particular, the following research questions (RQs) and hypotheses are tested when the text is viewed on a screen-based device:

RQ1: Are there any differences in the objective legibility of 3D text when positioned over urban, natural, and abstract complex backgrounds?

RQ2: Is there an interaction between the type of complex background and the text orientation angle of 3D text in their effect on objective legibility?

RQ3: Do subjective workload, estimated legibility and response confidence change depending on the text orientation angle, and do subjective assessments of reading difficulty differ between urban, natural and abstract backgrounds?

RQ4: To what extent is objective legibility related to subjective workload and other subjective assessments?

RQ5: What is the variability in legibility between individual background images within the same type of background?

H1. *The legibility of 3D text will increase with an increase in the text orientation angle from 15° to 45°, with the lowest success rate at 15° and the highest at 45°.*

H2. *Lower objectively measured legibility will be associated with a higher subjectively measured workload and less favorable subjective assessments of the task.*

The novel contributions of this work are as follows:

1. Analysis of the influence of the type of complex background on glance-based legibility of 3D text, where urban, natural and abstract backgrounds are represented by three different scenes within each type.

2. Analysis of the relationship between the type of background and the spatial orientation of 3D text, i.e., determining whether the effect of a complex background on 3D text legibility changes at different text orientation angles.
3. Comparison of objective and subjective measures by linking the performance of character recognition to the NASA-TLX workload and targeted subjective assessments.

2. Materials and Methods

This research is a continuation of our previous controlled study in which a glance-based methodology was developed to assess the legibility of spatially oriented 3D text and determine the appropriate time of its short-term display for future mixed-reality research [20]. In the previous study, an adaptive three-down, one-up staircase procedure was used, whereas in this study, the display time was pre-fixed for all participants and all experimental conditions based on the results reported in [20]. Thus, the experimental focus was shifted from determining the time threshold to directly measuring the success of recognizing 3D text displayed over different complex backgrounds.

2.1. Study Design

The research was conducted as a controlled screen-based simulation of conditions relevant to future handheld MR displays. Instead of a real-time MR application, participants were shown previously rendered visual stimuli on the screen. Such an approach enabled all participants to see identical representations and to control the background properties and spatial orientation of the 3D text independently of changes in the real environment.

The main experimental factor was the category of complex background, with three levels: *abstract* (ca), *natural* (cn) and *urban* (cu). Each category was represented by three different background images, i.e., a total of nine background images were used. The introduction of three variants within each category aimed to reduce the dependence of the results on the specific characteristics of one individual image and to enable the assessment of the consistency of the results within the wider background category.

Another experimental factor was the text orientation angle of the 3D text, with nominal values of 15°, 30°, and 45°. Each participant was exposed to all three angles and all three background categories. The experiment was organized in three blocks according to the text orientation angle. The order of the blocks was randomized for each participant to minimize the systematic influence of ordering, learning, or fatigue.

For each angle, a total of 90 experimental stimuli were presented: three background categories × three concrete images within the category × ten textual stimuli per specific image. Each participant therefore viewed 270 experimental stimuli throughout the experiment. Since 18 participants participated in the study, the complete experiment produced a total of 4860 trial-level observations.

After completing each block of 90 trials, the participant completed the Raw NASA-TLX and an additional subjective questionnaire, and then proceeded to the next randomized specific angle.

Age was included as an additional between-participant characteristic through three predefined age groups: 18–35, 36–50 and 51–70 years. It was included as an exploratory participant characteristic because visual functions relevant to rapid character recognition, particularly visual acuity and contrast sensitivity, are known to decline with age and may therefore influence performance under brief and visually complex viewing conditions [21].

2.2. Participants

A total of 18 participants were included in the study, evenly distributed in three age groups. The group of 18 to 35 years consisted of six participants (M = 26.17, SD = 5.49,

range = 21–34 years), the group of 36 to 50 years of six participants (M = 43.67, SD = 3.56, range = 37–47 years), and the group of 51 to 70 years of six participants (M = 58.00, SD = 5.51, range = 51–66 years). There were four women and two men in each age group. The total sample thus included 12 women and 6 men.

Participants were recruited using convenience sampling from individuals known to the researchers and available to participate in the study. Recruitment was continued until six participants had been included in each of the three predefined age groups, resulting in a balanced age-group distribution. Eligibility required participants to be between 18 and 70 years of age and to report normal or corrected-to-normal vision. Vision status was based on participant self-report; no formal visual-acuity, colour-vision, or other ophthalmological screening was performed. Participants were asked whether they had any uncorrected visual problems that could interfere with character recognition. No additional exclusion criteria were applied.

No a priori power analysis was conducted to determine the sample size. The sample size was determined primarily by the feasibility of the intensive within-participant experimental protocol, in which each participant completed 270 experimental trials during a session lasting approximately 90 min. Participants were distributed equally across the three predefined age groups to provide balanced age representation. The age-group analyses were not a primary basis for sample-size determination and were therefore treated as secondary exploratory analyses.

The study was conducted in accordance with the principles of the Declaration of Helsinki, and all participants provided informed consent prior to participation.

2.3. Experimental Stimuli

2.3.1. Preparation of Complex Backgrounds

Nine complex backgrounds were prepared for the research, divided into the three previously mentioned categories: abstract environment (ca), natural environment (cn) and urban environment (cu). Each category was represented by three different images, labeled ca1–ca3, cn1–cn3 and cu1–cu3.

The original backgrounds were generated using the OpenAI/ChatGPT 5.6 Sol image generation tool according to predefined visual criteria. During the generation and selection, efforts were made to achieve high consistency of images within each category, while maintaining the natural variability between the three specific examples. Special attention was paid to perspective and framing, general visual impression, distribution of details, color palette and the absence of semantically highly prominent elements, such as people, vehicles, traffic signs, logos, or other text. The generated images were also visually inspected before inclusion to exclude obvious generation artifacts.

Urban backgrounds depicted dense city facades and architectural elements in neutral daylight, with beige, brown, gray, and darker tones predominating. Natural backgrounds depicted a forest environment with vegetation, trees, stones, and moss, and a dominant green-brown palette. Abstract backgrounds were designed as unrealistic multi-colored compositions with a large number of overlapping geometric and organic shapes, textures, and local color changes. The generated images were then imported into Blender v5.1 and used as backgrounds for 3D text stimuli. The final stimuli were rendered at a resolution of 1536 × 864 px. The background images used are shown in Figure 1.

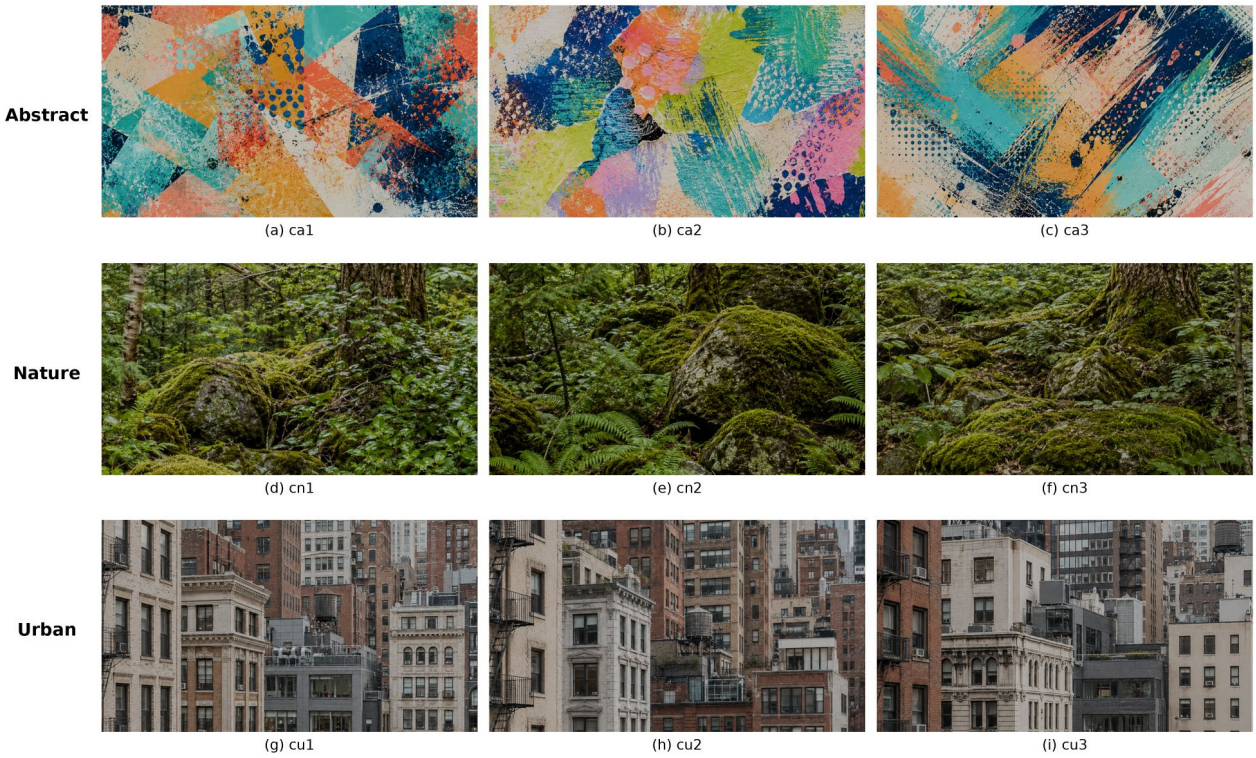


Figure 1. Complex background images used in the experiment: (**a**–**c**) abstract backgrounds (ca1–ca3), (**d**–**f**) natural backgrounds (cn1–cn3), and (**g**–**i**) urban backgrounds (cu1–cu3). The images shown are background-only Blender v5.1 renders corresponding to the backgrounds used in the experimental stimuli.

2.3.2. Characterisation of the Brightness of Background Images

Since bright 3D text was used in all experimental conditions, the tonal properties of the background images were further analyzed to more accurately describe the conditions in which the text was displayed. The analysis was performed on nine rendered backgrounds without text, created in Blender v5.1 at the same scene, camera, and rendering settings as the final experimental stimuli. Thus, the analysis was performed on the versions of the backgrounds that corresponded to the backgrounds shown to the participants during the experiment, and not on the original images before importing them into Blender v5.1. Each rendered background had a resolution of 1536 × 864 pixels. For analysis purposes, the images were converted to grayscale, after which a pixel intensity histogram was created for each image (Figure 2).

The histogram shows the distribution of dark, mid, and light tones within the image. The pixel intensity was normalized to a range of 0 to 1, with a value of 0 corresponding to black and a value of 1 to white. For each histogram, the X-axis shows the normalized grayscale intensity, while the Y-axis shows the percentage of the total number of pixels of the image belonging to each intensity interval. The histograms were calculated using 100 equal intervals, which corresponds to a bin width of 0.01.

As a simple numerical measure of the total brightness of an image, the mean grayscale intensity was additionally calculated. The obtained values showed a clear pattern between the three categories of backgrounds. Abstract backgrounds were on average the brightest (ca1 = 0.483; ca2 = 0.520; ca3 = 0.495), natural backgrounds were the darkest (cn1 = 0.292; cn2 = 0.229; cn3 = 0.254), while urban backgrounds occupied the middle range (cu1 = 0.387; cu2 = 0.360; cu3 = 0.370). The average value for the category was approximately 0.499 for abstract, 0.258 for natural and 0.372 for urban backgrounds. For abstract backgrounds, the distribution is shifted towards higher intensity values, which indicates a higher proportion of medium and light tones. Natural backgrounds show a distribution shifted towards

lower values, i.e., a higher proportion of darker tones, while urban backgrounds occupy an intermediate position between these two groups. At the same time, the histograms show that the three images within the same category are relatively similar, but not identical, which confirms that they were used as controlled variants of the same general type of background environment.

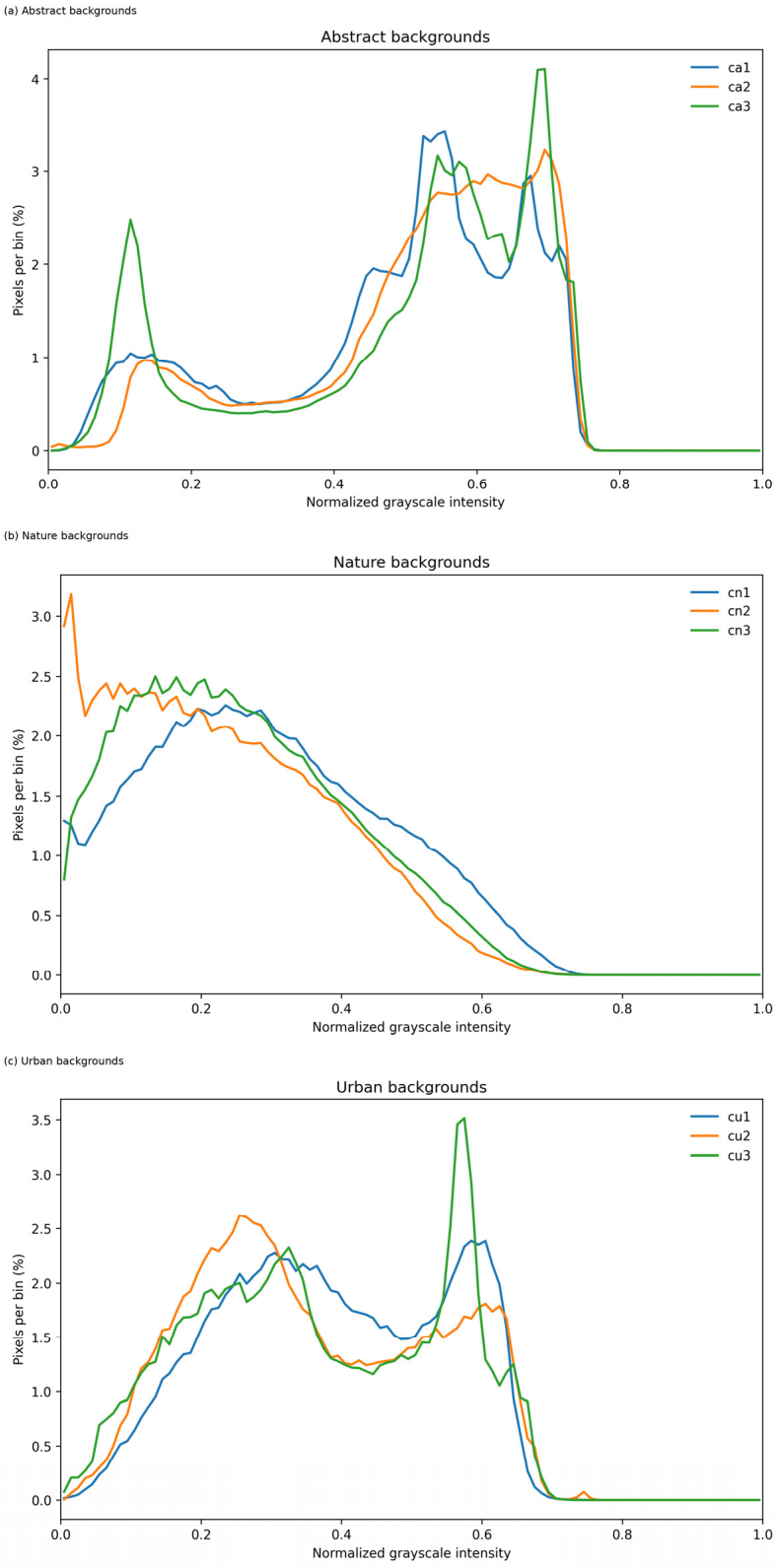


Figure 2. Histograms of normalized grayscale intensity of rendered background images: (**a**) abstract backgrounds ca1–ca3, (**b**) natural backgrounds cn1–cn3, and (**c**) urban backgrounds cu1–cu3.

These differences are important for describing the experimental conditions because the same light 3D typography was used on all backgrounds. Lighter backgrounds generally produce less contrast with light text, while darker backgrounds allow for higher contrast. However, because the backgrounds used are spatially heterogeneous and contain both dark and light areas at the same time, the average brightness and histogram of the entire image do not represent a direct measure of the local contrast between each individual character and the area of the background immediately behind it.

As all experiments were conducted on the same laptop and the same screen with the same display settings, the relative differences in the digital values of the background images were the same for all participants. The screen settings and characteristics are described separately in Section 2.6.

2.3.3. 3D Text Parameters and Rendering

For all stimuli, the same combination of typographic and spatial parameters was used to isolate the background effect. The combination of Roboto Regular + Wide tracking + Shallow depth (r_w_s) was chosen, which in a previous study [20] showed the lowest overall time threshold among the eight combinations examined.

The optical x-height of the Roboto characters was 0.39789 m in the Blender scene. For Wide tracking, the Character Spacing parameter value of 1.050 was used. The Shallow depth variant was defined as the total depth corresponding to 15% of the x-height. Since the extrusion in Blender was carried out symmetrically to the front and back of the text, the Blender Extrude parameter value for Roboto was 0.0298 m. Examples of text under different angles for each type of background are shown in Figure 3.

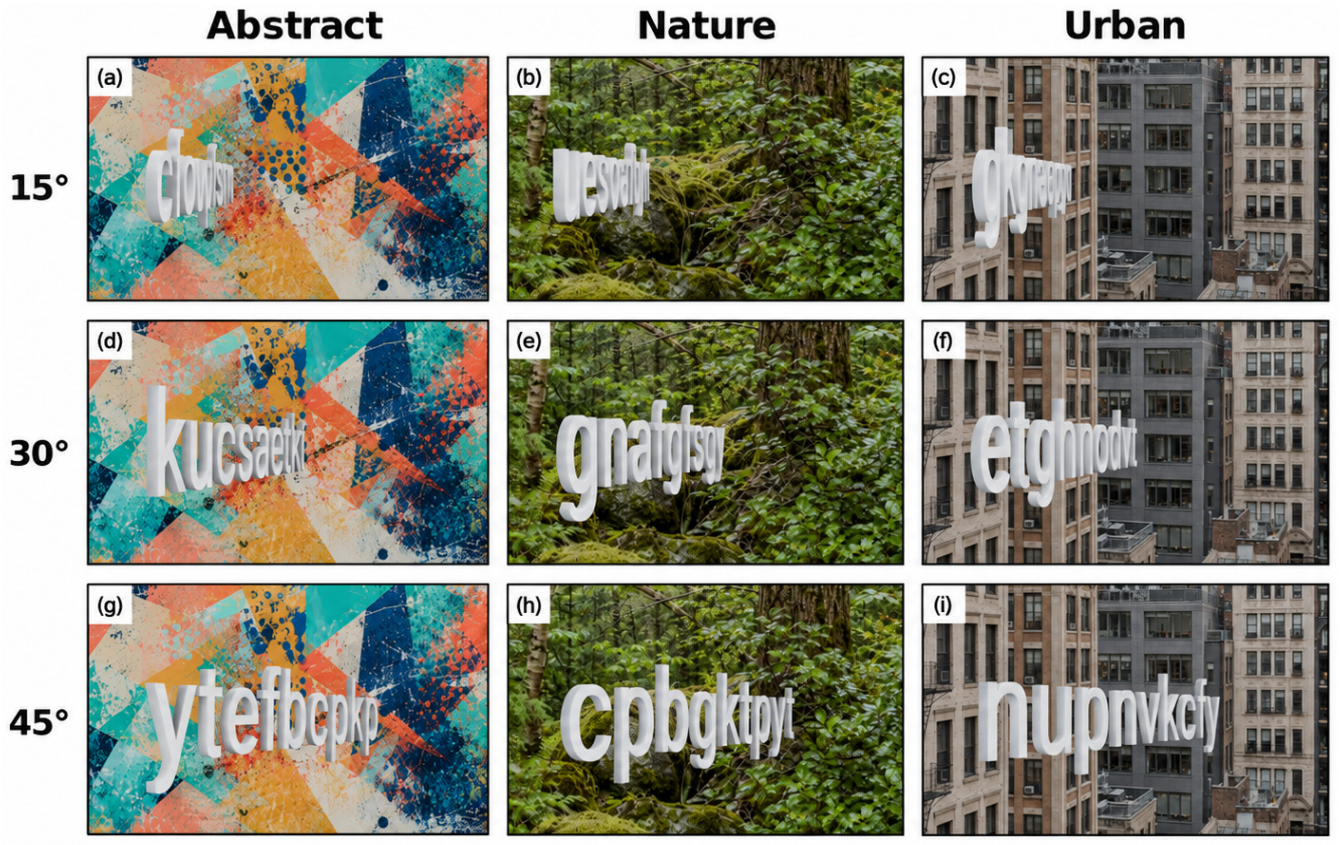


Figure 3. Examples of final experimental stimuli for the combination of Roboto Regular, Wide tracking and Shallow depth at angles of 15° (**a**–**c**), 30° (**d**–**f**) and 45° (**g**–**i**) and on three categories of complex background: abstract, natural and urban. Within each category, the same background image is shown at all three angles for easier visual comparison of the effect of spatial orientation of the text.

Three nominal text orientation angles were examined: 15°, 30° and 45°. The nominal text orientation angle (α) was defined in plan view as the acute angle between the camera viewing vector, projected onto the horizontal XY plane, and the front plane of the 3D text. Accordingly, $\alpha = 0°$ represents an exact lateral (edge-on) orientation, in which the viewing vector is parallel to the text plane, whereas $\alpha = 90°$ would represent a fully front-facing orientation. Because the camera was displaced relative to the reference axis of the Blender scene, the exact lateral orientation corresponded to a text-object rotation of 105° around the Z-axis. The tested nominal angles were therefore obtained as $\alpha = 105° - \theta Z$: Blender rotations of 90°, 75°, and 60° corresponded to nominal text orientation angles of 15°, 30°, and 45°, respectively. Thus, 15° represented the most oblique condition and 45° the condition closest to a front-facing presentation. The scheme is shown in Figure 4. Changing the angle simultaneously changes the visible area of the frontal and side surfaces of the characters, the perspective foreshortening of their width and the apparent spacing between the characters. The nominal color of the 3D text object was set to #FFFFFF. All lighting, camera and rendering settings were kept constant through experimental conditions.

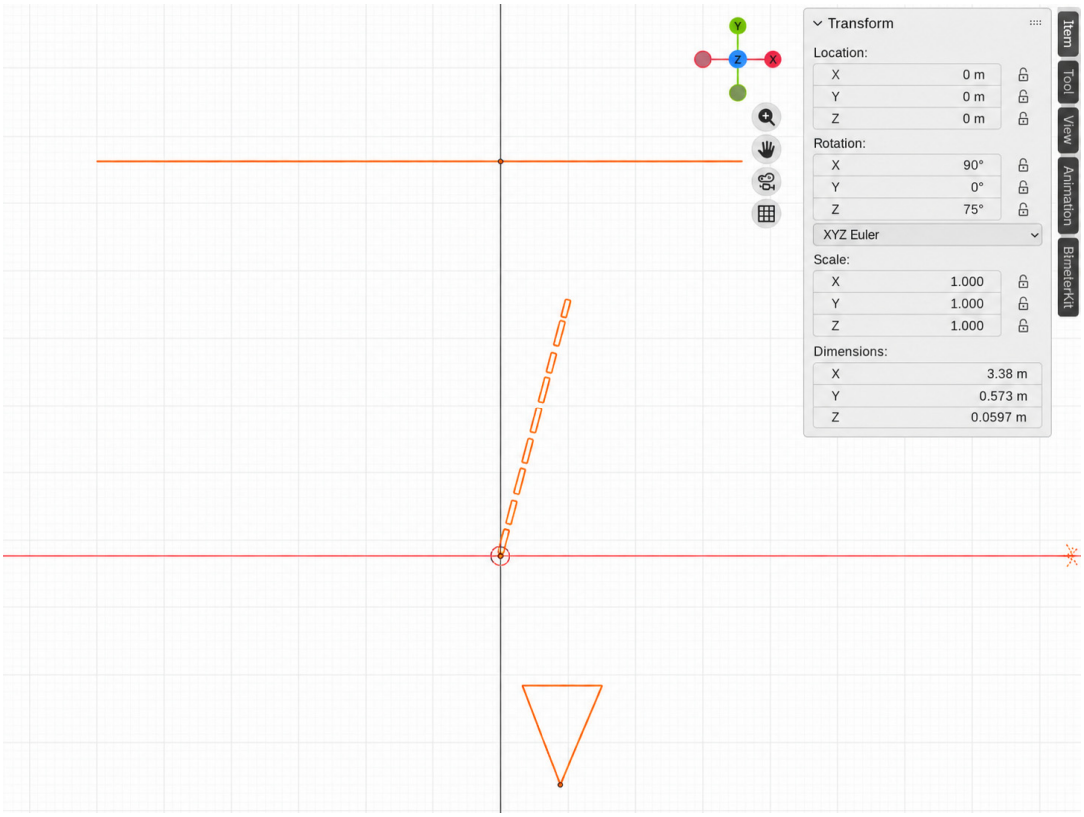


Figure 4. Top-view representation of the Blender v5.1 scene for the nominal 30° viewing-angle condition. The 3D text object is located at the scene origin and rotated around the Z-axis by 75°, while the camera is positioned below the text object and directed toward it. The background plane is visible at the top of the scene, and the light source is positioned to the right. The Transform panel shows the rotation values of the selected text object used to obtain the 30° nominal text orientation angle.

A total of 270 pre-rendered experimental stimuli were prepared. During stimulus generation, the predefined viewing-angle × background-image combinations were processed in a fixed sequence, with ten independently generated random text strings assigned to each combination. The text strings and target letters were generated randomly and were not explicitly counterbalanced across experimental conditions. The resulting allocation was fixed, and the same set of pre-rendered stimuli was used for all participants. A subsequent

balance analysis showed no statistically detectable association of target-letter vertical-zone group or target-letter occurrence count with text orientation angle or background type (Appendix E).

2.4. Glance-Based Legibility Method and Experimental Task

The experimental task was based on the glance-based legibility approach, i.e., the assessment of the ability to quickly and accurately visually recognize characters during a short-term display. The methodological approach developed in the previous study links the task of Gabbard et al. (2007) [22], in which the participant finds and counts the occurrences of the target character among random characters, with a time-limited representation, as used in the glance legibility study by Sawyer et al. (2017) [18]. The use of random strings instead of words reduces the contribution of lexical knowledge and familiar word patterns and focuses the task on visually distinguishing individual characters.

In this study, each textual stimulus consisted of a single line with nine randomly generated lowercase letters. As in the previous study [20], the target letter could appear in a sequence from zero to three times and could not appear consecutively. The design of the nine-character string was chosen as a compromise between sufficient visual search weight and the need for the stimulus to remain short enough for a glance-based display.

The course of each individual trial was the same as that used in the previous study [20], with the difference that the duration of the 3D stimulus was no longer adaptive. First, the target letter that the participant needed to find was shown at 1500 ms, followed by a 200 ms visual mask. A 3D text stimulus was then shown with a duration of exactly 659 ms, followed again by a visual mask of 200 ms. The participant then chose how many times the target letter appeared in the previously displayed sequence: 0, 1, 2 or 3 times (Figure 5). Up to 5 s was scheduled for response; the absence of response within this time was treated as an incorrect response, as in the previous study. A previous version of the experimental sequence is described in detail in [20].

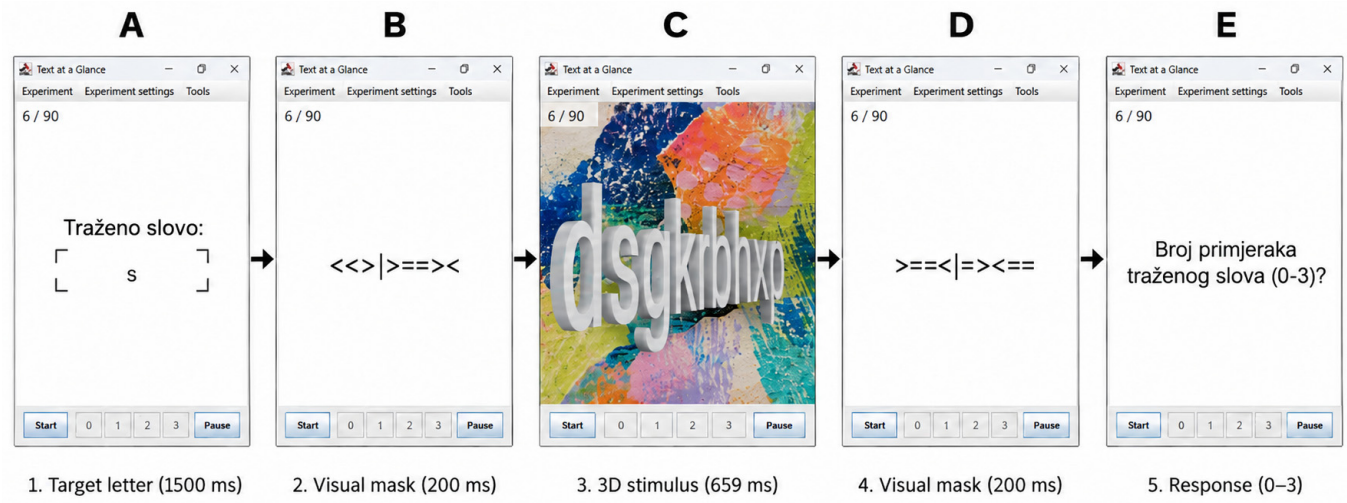


Figure 5. Sequence of events in a single glance-based 3D text legibility trial. (**A**) The target letter is shown within square brackets, indicating the character that the participant should identify and count in the subsequently presented text stimulus. "Traženo slovo" means "Target letter". (**B**) A visual mask is displayed before the stimulus to reduce visual persistence from the preceding screen. (**C**) The 3D text stimulus is briefly presented over a complex background; the example shown represents the 30° condition on an abstract background. (**D**) A second visual mask is displayed immediately after the stimulus to limit continued visual processing of the text after its disappearance. (**E**) The response screen allows the participant to report how many times the target letter occurred in the displayed stimulus, using response options from 0 to 3. "Broj primjeraka traženog slova" means "The number of examples of the target letter".

In this study, the display time of 659 ms was the same for all participants, backgrounds and angles. The value was derived from the results of our previous study reported in [20], in which the average of the last up to five reversals was 575.0 ms, and the average of the last six image displays was 743.5 ms. The arithmetic mean of 659.3 ms was proposed as a practical time for subsequent studies, with an explicit note that it should be validated by directly measuring the accuracy achieved. For implementation in the current experiment, an integer value of 659 ms was used.

2.5. Experimental Procedure

Before the start of the experiment, the task of recognizing and counting the target letter was explained to the participants. The participants were shown an image with all the font letters participating in the study and were instructed to pay attention to the construction of the letters (Figure 6). For example, the letter "f" in this type of script does not have a bottom zone. They were also shown pictures with an example of the letters under each of the angles to be measured. The letters in all the pictures were shown on a gray background.



Figure 6. Training session: participants were shown displays of the letters of the selected type of Roboto script for the purpose of becoming acquainted with the appearance of the construction of individual letter characters.

Prior to running the actual experiment, each participant was instructed to take part in a training session. The participants trained on a 15° angle display. For each complex background type, 7 images were shown, totalling 21 images for the training. The background images also changed within the background type. The training session was intended as an informal familiarization with the task and stimulus presentation procedure. Performance during the training trials was not formally evaluated, and no minimum accuracy criterion was required. After completing the training session, each participant was asked whether the task and the expected response procedure were clear, and the main experiment began only after the participant confirmed understanding.

The main experiment consisted of three experimental blocks, one for each angle of the text display. The order of the angles was randomized separately for each participant. Once the first angle was determined, the participant was shown all 90 stimuli belonging to that angle, including all three background categories and all three concrete examples within each category. This randomization affected only the presentation order of the pre-prepared stimuli and did not alter the fixed assignment of text strings to background-image and viewing-angle combinations.

At the end of the 90 views, the participant filled out the Raw NASA-TLX questionnaire and an additional subjective questionnaire that related exclusively to the just completed

block and the corresponding angle. This was followed by the next randomized angle, and the same procedure was repeated until the participant completed all three blocks. Such a design allowed NASA-TLX to be interpreted at the level of the angle or the entire block, while additional subjective questions allowed for a separate assessment of the impression of each of the three categories of background within each angle.

The order of individual stimuli within each experimental block was randomized for each subject. The participants were assigned a seed that allowed the repetition of a particular sequence of image displays.

After each block and after completing the questionnaire, the participant was allowed a short break of up to 10 min. Furthermore, the participant was able to pause the study during each of the experimental blocks. It was possible to take up to 3 breaks for each block.

The complete experimental session lasted approximately 90 min with minimal pausing periods. Each of the three experimental blocks, including the subsequent completion of the Raw NASA-TLX and the additional subjective questionnaire, required approximately 30 min.

2.6. Experimental Equipment

The experiment was conducted on the same computer and using the same research application as in the previous study [20]. A Lenovo Legion Y520 laptop with an Intel Core i5-7300HQ @ 2.50 GHz processor, 16 GB of RAM and an NVIDIA GeForce GTX 1050 Ti (4 GB) graphics card was used. The built-in 15.6-inch LED display operated at a resolution of 1920 × 1080 px and a refresh rate of 60 Hz, with the Windows 10 Home operating system.

To conduct the experiment, we developed a custom application called "Text at a Glance", developed in the Java programming language and compatible with the Java SE 8 platform. The application uses Swing for graphical user interface development, loads pre-prepared visual stimuli, manages their time-defined display and stores participants' responses. For the purposes of this study, the application was configured to display each 3D stimuli for a fixed duration of 659 ms, so the adaptive-staircase function used in our previous study [20] was not active.

The display of images in the application was implemented via the Swing component JPanel, and the display changes were implemented via the paintComponent() and repaint() methods. Time control was achieved using the Swing Timer mechanism. Since the monitor operates at 60 Hz, one refresh cycle takes approximately 16.7 ms. Therefore, the nominal 659 ms represents the software-defined presentation duration, while the actual visible duration of an individual stimulus may deviate by approximately one refresh cycle due to display refresh synchronization and operating-system event scheduling. The effective stimulus duration was not independently verified using external timing equipment such as a photodiode or high-speed camera. However, the same presentation procedure and nominal duration were used for all participants and experimental conditions.

Each experimental stimulus image had a resolution of 1536 × 864 px and was displayed without scaling, using a 1:1 pixel mapping within the application. The visible application area measured 1536 × 803 px; consequently, the lower 61 px of each stimulus image fell outside the visible area. This cropped portion did not contain the text stimulus and was not relevant to the experimental task. The displayed dimensions and cropping were identical for all participants and conditions.

Participants were seated at an approximately 50–70 cm viewing distance from the laptop display. Viewing position was not mechanically fixed using a chin rest, forehead rest, or other head-stabilization device, and no fixation marker was used. Viewing distance was not precisely measured, but the experimenter monitored participants throughout the

session to ensure that they maintained a conventional and comfortable reading distance, while allowing minor natural changes in head position.

Ambient illumination was not measured photometrically. However, all experimental sessions were conducted under the same indoor lighting conditions, using the same laptop, display settings, and fixed screen-brightness setting throughout the study. No automatic brightness adjustment was used.

2.7. Legibility Measures

2.7.1. Objective Legibility Measures

Two objective measures of legibility of 3D text were used, namely accuracy and total error, which is similar to the methodology of Gabbard et al. (2007) [22] for 2D text. Accuracy was assessed based on the correct recognition of the number of occurrences of the target letter in the displayed sequence. For each individual trial, the answer was classified as correct (*correct* = 1) if the participant's response fully corresponded to the actual number of occurrences of the target letter, and as incorrect (*correct* = 0) if there was any discrepancy. Based on this, the measure of *accuracy* was calculated as the proportion of correct answers in the observed set of trials, expressed as a percentage.

The second objective measure was the total error (*total_error*), which, unlike the binary measure of *accuracy*, also took into account the magnitude of the deviation of the response from the correct value. The error of each trial corresponded to the absolute deviation of the participants' responses from the actual number of target letters, and the values went from 0 for the correct answer to 3 for the maximum error.

Nine trials (0.19% of all 4860 trials) were unanswered within the allowed response time. According to the predefined procedure, these trials were retained in the analysis, classified as incorrect, and assigned the maximum error value of 3. There were no missing trials, and no trials were excluded from the analysis; no post hoc trial-level exclusion criteria were applied.

The *total_error* measure was calculated for each participant as the sum of the absolute deviations across the 30 trials within each combination of text orientation angle and background type. For example, if the correct number of target letters was 3, a response of 2 contributed an error of 1, whereas a response of 1 contributed an error of 2.

2.7.2. Subjective Legibility Measures

The subjective workload of the 3D text legibility task was estimated using the Raw NASA Task Load Index (Raw NASA-TLX) questionnaire [19]. Raw NASA-TLX encompasses six dimensions (items) of subjective assessment: *Mental Demand*, *Physical Demand*, *Temporal Demand*, *Performance*, *Effort*, and *Frustration*. Mental demand was related to concentration, letter recognition, and decision-making; physical demand included eye strain, body position, and hand movements; temporal demand was specifically related to the difficulty of the task due to the short duration of the display, while performance was operationalized as satisfaction with one's own performance. Effort and frustration were assessed as separate items on the same scale. Each dimension was assessed on a scale from 0 to 100 in 5-point increments.

The questionnaire was administered in Croatian. No previously validated Croatian adaptation was used. The wording was translated by the authors from the English version for the purposes of this study. The translation was reviewed for clarity and consistency with the original NASA-TLX dimensions before administration. The English wording presented in Appendix C.1 is an English-language rendering of the Croatian items administered to participants.

The questionnaire was completed separately for each of the three text orientation angles (15°, 30° and 45°) after conducting an experiment for the respective angle (i.e., after viewing the corresponding 90 images). The instruction explicitly asked for an assessment of the experience during the just-completed angle at which the text was displayed over different complex backgrounds.

The total NASA-TLX Raw score for each angle was calculated as an unweighted average of the six dimensions. Since in the applied version of the questionnaire, a higher value of the performance dimension indicated a better subjectively assessed performance, before calculating the total load, this dimension was reversely coded (100—performance) so that for all six components the higher value corresponds to higher load.

In addition to NASA-TLX, an additional subjective user experience questionnaire was used with six questions. The first five questions were assessed on a five-point numerical rating scale from 1 (strongly disagree) to 5 (strongly agree). The questions were related to: (1) the general ease of reading of the displayed text, (2) the influence of the urban background on reading difficulty, (3) the influence of natural background on reading difficulty, (4) the influence of abstract background on reading difficulty, (5) the participants' confidence in the accuracy of their own answers, and (6) the confidence that different background images were displayed within each type of background during the experiment (see Appendix C.2 for details). The last question was used as an indicator of conscious noticing of changes in individual background images within the same type. Individual questions of the additional questionnaire were analyzed separately, since the measures are conceptually different aspects of the user experience.

2.8. Statistical Methods

The statistical analysis was carried out at two complementary levels of results: at the level of individual trials and at the level of results aggregated per participant. The reason for this approach stems from the structure of the experiment itself. Namely, each participant solved a large number of individual trials, and the success in each trial depended not only on the experimental factors that were the subject of the research, such as the text orientation angle and the type of background, but were also influenced by the number of occurrences of the target letter, the identity of that letter, the specific sequence of characters (stimulus) and the individual background image. In addition, the responses of the same participant are more interrelated than the responses of different participants.

A simple comparison of average *accuracy* percentages or *total_error* can clearly show what was observed, but it cannot simultaneously separate all of the listed sources of variability that influenced the result. For example, a lower average *accuracy* at 15° may be the actual effect of the angle, but a given set of trials may contain more difficult target letters or more challenging stimuli at the same time. A statistical model may allow all individual trials to be analyzed together and estimate the effect of the factors of interest while taking into account other known sources of variability that affect the result. For this reason, a *generalised linear mixed model* (GLMM) at the level of individual trials was used for a suitable primary analysis of objective legibility.

In addition to the GLMM analysis, results were also analysed at the level of scores aggregated per participant. In them, a larger number of individual trials are summarised into a single value for an individual participant and an experimental condition, for example, in overall *accuracy* at an angle of 15° against an urban background. This analysis does not separate the sources of variability of individual trials, but it provides a direct representation of the actual observed average performance of participants and makes it possible to verify whether the same basic patterns occur when each participant is considered as a unit of

analysis. The analyses at the level of individual trials and the aggregated analyses therefore provide different but complementary views of the same data.

To assess the robustness of the background-related findings to the estimation approach, an additional frequentist binomial generalized estimating equation (GEE) analysis was performed. Its specification and results are provided in Appendix D.

2.8.1. Binomial Generalised Linear Mixed Model

The primary outcome of each individual trial is binary: the participants' response was either completely correct (*correct* = 1) or incorrect (*correct* = 0). Because each participant participated in a large number of trials, and the same participants, background images, target letters, and other stimulus components appear in a larger number of trials, the individual responses are not completely independent of each other. For such a data structure, the binomial GLMM with the logit link function is suitable [23].

GLMM allows for both *fixed* and *random* effects to be included simultaneously in the same model, which is why it is called a mixed model. Fixed effects refer to experimental factors whose specific levels we want to compare directly or whose influence we want to take into account when evaluating other effects in the model. For example, the aim of the study was to compare angles of 15°, 30° and 45° and abstract, natural and urban backgrounds, and therefore both the angle and the type of background are fixed effects. In contrast, random effects represent those sources of variability that can affect the result, but their individual levels are not the main subject of research. For example, the aim is not to estimate the separate effect of each of the 18 participants or each of the 21 target letters, but to take into account the fact that the participants differ from each other and that some letters may be easier and some more difficult to identify. It should be noted that the term *fixed effect* does not mean that the value of the factors does not change during the experiment. It means that the effects of its specifically defined levels are directly estimated in the model. For random effects, the model estimates the standard deviation of their variability. A larger estimated standard deviation indicates greater heterogeneity associated with this source, for example greater differences between stimuli or between participants. These estimates are shown on a logit scale. Table 1 lists and explains fixed and random effects as components of the GLMM.

Interactions of age with angle and type of background are not included in the GLMM due to the relatively small number of participants within each age group. A possible interaction between the angle and type of background with age was modeled separately on aggregated participant scores.

The GLMM can be represented by the expression

$$\begin{aligned} log\left(\frac{P(correct=1)}{1-P(correct=1)}\right) = \beta_0 + \beta_{angle} + \beta_{background} + \beta_{angle\times background} + \beta_{age_group} + \beta_{correct_answer} + \\ u_{participant} + u_{background_image} + u_{stimulus} + u_{target_letter} + u_{participant\times angle} + u_{participant\times background} \end{aligned} \quad (1)$$

Here, β are fixed effects and u are random effects. A Bayesian approach was used to estimate the GLMM. Instead of giving only one estimated value for each effect, the Bayesian procedure provides a range of possible values along with information about which of them, according to the available data and the initial assumptions of the model (prior), are more plausible. This is useful for models with a larger number of interconnected sources of variability, as it allows for more stable estimates of effects and at the same time shows how much uncertainty there is around each estimate. For example, the model not only provides an estimate of how different the probability of an accurate answer is between angles of 15° and 30°, but also the range of values that the model considers plausible for this difference.

GLMM results are primarily presented using model-estimated marginal probabilities of correct answer. They represent the estimated probability of a correct answer for a given

level of factors after the model has taken into account the other factors involved at the same time. For example, the marginal probability for an angle of 15° is a model estimate in which the differences in background type, age group and number of target letters are balanced, making it possible to compare angles under similar other conditions.

Table 1. GLMM components and their role in the model.

Component	Type of Effect	Role in the Model
Text orientation angle	Fixed	One of the main experimental factors; three defined angles are directly compared.
Background type	Fixed	One of the main experimental factors; the differences between the three types of complex background are directly examined.
Angle × background type	Fixed	Examines whether the effect of the background type changes depending on the text orientation angle.
Age group	Fixed	Allows systemic differences between predefined age groups to be taken into account when assessing the main experimental effects.
Number of target letters (correct_answer)	Fixed	Allows to include the known difference in the difficulty of the trial related to the number of times the target letter appears.
Participant	Random	Considers individual differences in overall performance and the fact that the same participant gives multiple answers.
Single background image	Random	Considers the variability between specific images within the same type of background, without evaluating each image as a separate effect.
Identity of the target letter	Random	Considers the possibility that the 21 target letters used differ from each other according to the difficulty of recognition, without comparing each letter separately.
Specific stimulus	Random	Considers the additional variability associated with a specific set of characters (stimulus) displayed in a particular trial.
Participant × angle	Random	Allows participants to differ from each other according to their sensitivity to a change in angle.
Participant × background type	Random	Allows participants to differ from each other according to their sensitivity to different types of backgrounds.

Each major model estimate is accompanied by a 95% Bayesian credible interval (95% CrI). This interval shows a range of values that, according to the Bayesian model, covers 95% of the plausible possible values of the observed effect or difference. When comparing the two conditions, it is particularly important whether the CrI includes a value of zero. If the 95% CrI difference does not include zero, the estimated values (posterior estimates) provide clear evidence of the direction of the difference between these conditions; if the interval includes zero, the data and the model do not allow an equally clear conclusion about its direction.

Details of the computational implementation of the Bayesian GLMM, including the estimation algorithm used, the priors, the calculation of marginal probabilities, posterior differences, and the Monte Carlo procedure, are given in Appendix A.

2.8.2. Statistical Analyses of Aggregated Participant Results

For objective measures of *accuracy* and *total_error*, the results were aggregated for all nine combinations of three angles and three types of background per participant. The main effects of angle and background type and their interaction were examined. When the overall analyses showed differences between the levels of each factor, appropriate pairwise comparisons were performed. In addition to parametric analyses, non-parametric checks were also used to assess the robustness of the findings. Details of the procedures used are given in Appendix B.1.

To further examine the variability of legibility between individual images within the same type of background, three abstract (ca1–ca3), three natural (cn1–cn3) and three urban (cu1–cu3) backgrounds were separately compared at each of the three angles. The performed analysis aimed to determine whether the objective performance differed between individual images of the same type and, when differences were identified, between which images they appeared. Since each participant was exposed to all three images within each type of background, the analyses were carried out as comparisons of related measurements, with correction due to multiple comparisons. Details are given in Appendix B.2.

The differences between the three age groups were analyzed based on aggregated *accuracy* and *total_error* values. The overall differences between the age groups were examined, followed by possible differences in the effect of angle and background type depending on age. The age-related analyses were treated as secondary and exploratory because each age group included only six participants. They included overall comparisons between age groups and exploratory tests of age × angle and age × background-type interactions. Details of the analyses are provided in Appendix B.3. To further characterize the sensitivity of the study to age-related interaction effects, an additional simulation-based sensitivity analysis was performed for the age × angle and age × background-type interactions (Appendix F).

Subjective measures were analyzed to determine whether the subjective load changes depending on the text orientation angle. The total Raw NASA-TLX score and individual NASA-TLX dimensions were analyzed, as well as additional subjective assessments of ease of reading, confidence in response, and difficulty of individual types of backgrounds. For the questions from the additional questionnaire collected on the five-point numerical rating scale, the differences between angles were analyzed, and the question about the equality of individual images within the same type of background was analyzed separately as an indicator of conscious observation of background change. Details of statistical procedures are given in Appendix B.4.

The correlation between subjective and objective results was analyzed in such a way that subjective estimates for each angle were related to the values of *accuracy* and *total_error* of the same participant at the same angle. The correlations of the Raw NASA-

TLX total score, subjective estimates of ease of reading and confidence in response with both objective measures were examined, and individual NASA-TLX dimensions were additionally analyzed. Since each participant gave multiple related observations through different angles, a procedure was used that estimates the common within-participant association between the two variables and does not treat all observations as mutually independent. Details of the procedure are given in Appendix B.5.

Because the exact textual sequences were associated with particular experimental conditions, an additional stimulus-balance analysis was conducted on the 270 unique stimulus configurations. Target letters were grouped into lower-, mid-, and upper-zone letters according to their vertical typographic extent, and the distributions of these groups and of target-letter occurrence counts (0–3) were examined across text orientation angles and background types. Pearson chi-square tests were used for the main condition-level comparisons. At the finer level of individual background images, where each angle × image combination contained only ten textual stimuli, distributions were examined descriptively because of the small cell counts (see Appendix E).

3. Results

3.1. Objective Performance Overview

At the level of the participants' aggregated scores, the average *accuracy* was 75.1%, while the average *total_error* was 8.31 wrong units per 30 trials (Table 2). The lowest *accuracy* aggregated by background type was recorded at an angle of 15° (66.9%), followed by 30° (77.3%), and the highest at 45° (81.1%). The lowest *accuracy* aggregated by angle was 74.1% for abstract backgrounds, followed by 74.5% for urban backgrounds, and the highest was 76.6% for natural backgrounds. From reviewing the overall results, it is evident that the effect of angle is significantly more pronounced than the differences between background types.

Table 2. Overall objective performance by angle and background type based on aggregate scores.

Angle	Abstract, Accuracy	Natural, Accuracy	Urban, Accuracy	Average Accuracy	Average Total_Error
15°	65.0%	68.0%	67.6%	66.9%	11.17
30°	77.2%	80.7%	73.9%	77.3%	7.61
45°	80.2%	81.1%	82.0%	81.1%	6.15
Average accuracy	74.1%	76.6%	74.5%	75.1%	-
Average total_error	8.74	7.54	8.65	-	8.31

3.2. Primary Analysis of Objective Legibility at the Level of Individual Trials

3.2.1. Effect of the Text Orientation Angle

GLMM showed a pronounced effect of the text orientation angle on the probability of giving a correct answer. Marginal probabilities for individual angles were calculated by equal weighting of three types of background, three age groups and four levels of correct_answer and are shown in Table 3. The estimated probability of giving a correct answer increased from the angle of 15° to the angle of 45°. Estimated values of differences between angles and types of backgrounds (posterior contrasts) are shown in Table 4.

Table 3. GLMM estimates of the marginal probability of a correct answer by text orientation angle.

Angle	Marginal Probability	95% CrI
15°	65.8%	63.4–68.1%
30°	80.1%	77.1–82.9%
45°	81.9%	78.8–84.7%

Table 4. Posterior contrasts of the GLMM-estimated marginal probabilities of a correct answer.

Contrast	Percentage Point Difference	95% CrI
30° − 15°	+14.3 p.p.	12.0 to 16.7
45° − 15°	+16.1 p.p.	13.6 to 18.5
45° − 30°	+1.7 p.p.	−1.5 to 5.0
Urban–Abstract	+2.0 p.p.	−0.4 to 4.3
Natural–Abstract	−1.5 p.p.	−4.0 to 0.9
Natural–Urban	−3.5 p.p.	−7.0 to −0.1

Compared to the 15° angle, the estimated *accuracy* was 14.3 percentage points higher at the 30° angle (95% CrI: 12.0–16.7) and 16.1 percentage points higher at the 45° angle (95% CrI: 13.6–18.5). The difference between 45° and 30° was 1.7 percentage points, but its 95% CrI included zero (−1.5 to 5.0). The results therefore clearly confirm the lowest legibility at 15° and the highest legibility at 45°, but the GLMM shows no significant difference between the angles of 30° and 45°.

3.2.2. Effect of the Complex Background Type

GLMM's estimated marginal *accuracy* probabilities by background type, obtained by equal weighting of three angles, three age groups, and four correct_answer levels, are shown in Table 5.

Table 5. GLMM estimate of the marginal probability of correct answer by background type.

Background Type	Marginal Probability	95% CrI
Abstract	75.7%	73.8–77.6%
Natural	74.2%	71.0–77.3%
Urban	77.8%	74.6–80.7%

The clearest difference was found between natural and urban backgrounds: the estimated probability of a correct answer was 3.5 percentage points higher on urban backgrounds (95% CrI: −7.0–−0.1). The intervals of difference between urban and abstract backgrounds (+2.0 percentage points; 95% CrI: −0.4–4.3) and natural and abstract backgrounds (−1.5 percentage points; 95% CrI: −4.0–0.9) included zero, and therefore the differences between these types of backgrounds are less plausible.

The additional GEE robustness analysis (see Appendix D) yielded the same overall ranking of background types as the GLMM and confirmed the urban–natural contrast (3.3 percentage points in favor of urban backgrounds; 95% CI: 0.2–6.5).

3.2.3. Interaction of Text Orientation Angle and Background Type

Since a possible interaction between angle and background type has been established at the level of overall results, it is necessary to consider the marginal effect of the background type together with the results for individual angles. The marginal probabilities estimated by GLMM show that the relationship between the type of background and the probability of correct answer is not the same at all angles of text display (Table 6).

Table 6. GLMM estimate of marginal probabilities of correct answer by angle and background type.

Angle	Abstract (95% CrI)	Natural (95% CrI)	Urban (95% CrI)
15°	63.3% (61.2–65.4)	63.2% (60.0–66.3)	*70.8% (67.9–73.7)*
30°	77.6% (74.9–80.1)	*82.0% (77.7–85.9)*	80.8% (76.3–84.8)
45°	*86.4% (84.3–88.3)*	77.5% (72.6–82.0)	81.7% (77.3–85.8)

At the angle of 15°, the highest adjusted probability of correct answer was estimated for urban backgrounds, while at the angle of 30° the highest value was recorded for natural backgrounds. At the angle of 45°, abstract backgrounds are preferred. These results suggest that the relative effect of the background type depends on the text orientation angle of the 3D text, and that there is no universally easiest or most difficult background type regardless of the angle.

The additional GEE robustness analysis also supported an overall angle × background-type interaction (Wald $\chi^2(4) = 25.98$, $p < 0.001$), although some individual background contrasts, particularly at 30°, were sensitive to the analytical approach (Appendix D).

3.3. Analysis of Text Orientation Angle and Background Type on Aggregated Results at the Participant Level

To provide a complementary participant-level assessment of the primary trial-level analysis, the average values of *accuracy* and *total_error* measures were analyzed for nine combinations of three angles and three types of complex background.

3.3.1. *Accuracy* Measure

The ANOVA of repeated measurements showed a strong main effect of angle on *accuracy*: $F(2, 34) = 48.47$, $p < 0.001$, $\eta p^2 = 0.740$. The main effect of the background type was not statistically significant: $F(2, 34) = 1.80$, $p = 0.180$, $\eta p^2 = 0.096$. The two-way ANOVA of repeated measurements did not find a statistically significant interaction between angle and background type: $F(4, 68) = 1.98$, $p = 0.108$, $\eta p^2 = 0.104$. Although Table 2 shows that there is some interaction between angle and background type at the level of aggregated scores of respondents, there is insufficient statistical evidence that this effect is significant.

Friedman's test confirmed significant differences in *accuracy* between the angles: $\chi^2(2) = 29.60$, $p < 0.001$, $W = 0.822$. Pairwise Wilcoxon comparisons with Holm's correction showed significant differences between all three angles: 15° vs. 30°: p(Holm) < 0.001, 15° vs. 45°: p(Holm) < 0.001 and 30° vs. 45°: p(Holm) = 0.010. This confirmed a monotonous increase from 15° to 45° in the data on *accuracy* at the level of the subjects.

For the type of background, Friedman's test did not show a significant overall difference: $\chi^2(2) = 1.68$, $p = 0.432$. The average *accuracy*, according to Table 2, was 74.1% for abstract, 76.6% for natural and 74.5% for urban backgrounds. When the types of backgrounds were analyzed separately within each angle, a significant difference was found only at an angle of 30°: $\chi^2(2) = 9.12$, $p = 0.010$. At this angle, the highest *accuracy* was recorded for natural backgrounds (80.7%), followed by abstract backgrounds (77.2%), and the lowest for urban backgrounds (73.9%). Post hoc analysis showed a significant difference between natural and urban backgrounds (p(Holm) = 0.028), while other pairwise comparisons were not statistically significant. At the angles of 15° and 45°, no significant differences were found between background types.

3.3.2. *Total_Error* Measure

For the *total_error* measure, the ANOVA of repeated measurements also showed a strong main angle effect: $F(2, 34) = 46.15$, $p < 0.001$, $\eta p^2 = 0.731$. In contrast to *accuracy*, this measure also found a statistically significant main effect of the type of background: $F(2, 34) = 3.68$, $p = 0.036$, $\eta p^2 = 0.178$, where the average *total_error* was 8.74 for abstract, 7.54 for natural and 8.65 for urban backgrounds (Table 2). A significant interaction between angle and type of background was also found: $F(4, 68) = 2.85$, $p = 0.030$, $\eta p^2 = 0.143$.

Friedman's test also confirmed significant differences in *total_error* between angles: $\chi^2(2) = 30.52$, $p < 0.001$, $W = 0.848$. Pairwise Wilcoxon comparisons with Holm's correction showed significant differences between all three angles: 15° vs. 30°: p(Holm) < 0.001, 15° vs. 45°: p(Holm) < 0.001 and 30° vs. 45°: p(Holm) = 0.010. *Total_error* monotonously

decreased from the angle of 15° to the angle of 45°, which also confirmed the improvement of performance with increasing the text orientation angle.

Friedman's test confirmed a significant overall difference between the types of background: $\chi^2(2) = 6.54$, $p = 0.038$, W = 0.182. Post hoc analysis showed lower *total_error* on natural than on abstract backgrounds (p(Holm) = 0.041), while the difference between natural and urban backgrounds did not remain statistically significant after correction. As with *accuracy*, the differences between background types were most pronounced at an angle of 30°: $\chi^2(2) = 10.40$, $p = 0.006$. At this angle, the lowest average *total_error* was recorded for natural backgrounds (6.17), followed by abstract backgrounds (7.56), and the highest for urban backgrounds (9.11). Post hoc Wilcoxon's comparisons with Holm's correction showed statistically significantly lower *total_error* for natural than for urban backgrounds (p(Holm) = 0.014), while the differences between abstract and natural and abstract and urban backgrounds were not statistically significant. As for *accuracy*, no statistically significant differences between the types of background were found at angles of 15° and 45°.

3.4. Variability Between Individual Background Images

To examine whether the legibility of 3D text differs between individual images within the same type of background, three abstract, three natural and three urban images were compared separately at each text orientation angle.

3.4.1. Differences Between Individual Background Images at the Level of Participants' Aggregated Scores

The results, aggregated at the level of the participants, show a marked variability in *accuracy* between individual images of the same type of background. The highest range was recorded among urban backgrounds at the angle of 30°, where the *accuracy* ranged from 62.2% to 85.0%, i.e., it differed by 22.8 percentage points (Table 7). A similar range was recorded among natural backgrounds at the angle of 15° (58.9–80.6%; range of 21.7 percentage points). After Holm's correction of the Friedman test for nine overall comparisons, statistically significant differences were found between abstract images at the angle of 15°, natural images at the angles of 15° and 45°, and urban images at the angle of 30°.

Table 7. Recognition *accuracy* for individual background images by background type and angle.

Angle	Background Type	Figure 1	Figure 2	Figure 3	Friedman p	Holm-Corrected p
15°	Abstract	66.7%	56.1%	72.2%	*0.004*	*0.022*
15°	Natural	64.4%	80.6%	58.9%	*<0.001*	*<0.001*
15°	Urban	63.9%	75.0%	63.9%	*0.041*	0.123
30°	Abstract	75.0%	81.1%	75.6%	0.483	—
30°	Natural	75.0%	78.9%	88.3%	*0.011*	0.053
30°	Urban	74.4%	62.2%	85.0%	*<0.001*	*0.001*
45°	Abstract	85.0%	77.2%	78.3%	0.128	—
45°	Natural	77.8%	76.1%	89.4%	*<0.001*	*0.006*
45°	Urban	83.9%	86.1%	76.1%	*0.011*	0.053

Post hoc comparisons showed significant differences between the first and second and the second and third abstract paintings at the angle of 15°, between all pairs of natural images at the angle of 15°, between the second and third urban images at the angle of 30°, and between the first and third and the second and third natural images at the angle of 45°.

The marked differences between the individual images raised the question of whether they could partly result from the different severity of the randomised textual stimuli. Of

particular importance is the number of occurrences of the target letter (correct_answer), which shows a strong effect on the probability of a correct answer (see Table 8, second column). Therefore, additional analysis was carried out with the GLMM to estimate the variability between individual images after keeping this factor fixed in the model.

Table 8. *Accuracy* and success range of specific stimuli according to the number of target letters.

Number of Target Letters (Correct_Answer)	Average Accuracy	Range of Accuracy of Individual Stimuli
0	94.1%	61.1–100.0%
1	85.2%	38.9–100.0%
2	69.3%	16.7–100.0%
3	49.4%	0.0–100.0%

3.4.2. Variability Related to Background Image and Specific Stimulus

The GLMM additionally enabled the estimation of residual variability associated with the individual background image, target letter and specific stimulus after the angle, background type, their interaction, age group and number of target letter occurrences were taken into account in the fixed part of the model. In Table 9, the model-estimated standard deviations of the random effects are shown.

Table 9. Estimated standard deviation (on the logit scale) of the random effects of GLMM.

Random Effect	GLMM, Estimated SD	95% CrI
Participant	*0.457*	0.326–0.641
Single background image	*0.242*	0.133–0.441
Target letter	*0.605*	0.447–0.819
Specific stimulus	*0.740*	0.680–0.805
Participant × angle	0.229	0.188–0.278
Participant × background type	0.200	0.164–0.243

The random effect that was found to have the greatest influence on the results of the model was the specific stimulus (SD = 0.740), followed by the identity of the target letter (SD = 0.605) and then the participant (0.457). The variability associated with the individual background image was smaller, but still present (SD = 0.242).

Large differences between stimuli can also be observed descriptively when the results of individual trials are grouped according to an equal number of target letters (Table 8). Although the target letter number factor is strongly related to the difficulty of the task, the very wide success ranges of individual stimuli within the same value of correct_answer show that the target letter alone does not explain all the variability in *accuracy*.

The presented results indicate two related levels of variability. First, individual background images of the same type may be associated with significantly different legibility, and some of these differences are retained even after the number of target letters is kept fixed. Second, an even greater source of heterogeneity is the concrete textual stimulus, which indicates the importance of the characteristics of the character sequence itself and its relationship with the specific background image. Mechanisms that could explain this finding are discussed in Section 4.2.

Because stimulus-specific variability was substantial, the distribution of systematic target-letter characteristics across experimental conditions was additionally examined. Across the 270 unique stimulus configurations, neither target-letter vertical-zone group nor target-letter occurrence count showed a statistically detectable association with text orientation angle or background type (Appendix E). Thus, the main experimental conditions

were reasonably balanced with respect to these two characteristics. At the level of individual background images, however, each angle × image combination contained only ten textual stimuli, and greater local variation in target-letter composition was present.

3.5. Age Differences in Objective Performance

Age differences were analyzed as a secondary exploratory analyses of objective performance. As mentioned earlier, participants were divided into three age groups: 18–35, 36–50 and 51–70 years.

3.5.1. Results for *Accuracy* and *Total_Error*

The average *accuracy* shows a gradual decline with age, as shown in Table 10. The youngest group achieved an average *accuracy* of 81.2%, the middle group 75.5%, and the oldest 68.5%.

Table 10. Objective performance by age group.

Age Group	Accuracy, Mean (SD)	Total_Error, Mean (SD)
18–35	81.23% (3.37)	6.19 (1.52)
36–50	75.49% (8.83)	8.00 (3.08)
51–70	68.52% (6.53)	10.74 (2.24)

For *accuracy*, the one-way ANOVA showed a statistically significant effect of the age group: $F(2, 15) = 5.54$, $p = 0.016$, $\eta^2 = 0.425$, which was confirmed by Welch's ANOVA: $F(2, 8.61) = 8.65$, $p = 0.009$ and the Kruskal–Wallis test: $H(2) = 7.35$, $p = 0.025$. Post hoc analyses showed that the youngest group was significantly more accurate than the oldest. The difference between the 18–35 and 51–70 groups was approximately 12.7 percentage points and was significant according to the Tukey test ($p = 0.012$) and the Games–Howell method ($p = 0.008$), as well as for the non-parametric Mann–Whitney U-test with Holm correction ($p = 0.025$). The differences between the youngest and middle groups and between the middle and oldest groups did not reach statistical significance.

A similar pattern was found for *total_error*. The one-way ANOVA showed a significant effect of the age group: $F(2, 15) = 5.63$, $p = 0.015$, $\eta^2 = 0.429$, and Welch's ANOVA was also significant: $F(2, 9.27) = 7.95$, $p = 0.010$, as well as the Kruskal–Wallis test: $H(2) = 7.94$, $p = 0.019$. The youngest group had the lowest and the oldest had the highest *total_error*. Post hoc analysis again showed a significant difference between the 18–35 and 51–70 age groups (Tukey $p = 0.012$; Games–Howell $p = 0.007$; Mann–Whitney U-test with Holm correction $p = 0.025$), while other pairwise comparisons were not statistically significant.

3.5.2. Interaction of Age and Text Orientation Angle

Descriptively, the same age pattern appeared at all three text-orientation angles (Table 11).

Table 11. Legibility results of age group and angle interaction.

Age Group	15°, Accuracy	15°, Total Error	30°, Accuracy	30°, Total Error	45°, Accuracy	45°, Total Error
18–35	73.0%	8.89	83.3%	5.56	87.4%	4.11
36–50	68.3%	10.44	78.0%	7.28	80.2%	6.28
51–70	59.3%	14.17	70.6%	10.00	75.7%	8.06

The focus of mixed ANOVA was on checking the age × angle interaction. No statistically significant age × angle interaction was detected for *accuracy*, $F(4, 30) = 0.38$, $p = 0.820$, or *total_error*, $F(4, 30) = 0.55$, $p = 0.703$. Given the small number of participants within each age group, these results should not be interpreted as evidence that age-related differences in the effect of text orientation angle are absent.

3.5.3. Interaction of Age and Type of Background

Average legibility scores by background type were also the highest in the youngest and the lowest in the oldest group (Table 12).

Table 12. Legibility results of age group and background type interaction.

Age Group	Abstract, Accuracy	Abstract, Total_Error	Natural, Accuracy	Natural, Total_Error	Urban, Accuracy	Urban, Total_Error
18–35	80.0%	6.78	83.7%	5.11	80.0%	6.67
36–50	73.5%	8.72	76.5%	7.44	76.5%	7.83
51–70	68.9%	10.72	69.6%	10.06	67.0%	11.44

No statistically significant age × background-type interaction was detected by mixed ANOVA for *accuracy*, $F(4, 30) = 0.64$, $p = 0.636$, or *total_error*, $F(4, 30) = 0.56$, $p = 0.692$.

The simulation-based sensitivity analysis of age × angle and age × background-type interactions indicated that the observed interaction magnitudes were substantially smaller than those associated with approximately 80% detection probability in the present sample (Appendix F). The study therefore had limited sensitivity to small or moderate age-related interactions.

3.6. Subjective Workload and Additional User Assessments

3.6.1. Raw NASA-TLX According to the Text Orientation Angle

The total subjective workload measured for all participants via the Raw NASA-TLX questionnaire was highest at the angle of 15° and decreased at 30° and 45° (Table 13).

Table 13. Results of the Raw NASA-TLX Raw questionnaire by text orientation angle.

Measure	15° Mean (SD)	30° Mean (SD)	45° Mean (SD)
Raw NASA-TLX total	56.30 (16.13)	51.71 (15.60)	49.26 (15.95)
Mental demand	67.78 (28.71)	60.28 (24.88)	58.33 (24.49)
Physical demand	47.78 (34.01)	45.56 (31.48)	38.89 (29.38)
Temporal demand	76.94 (21.77)	73.89 (20.11)	71.11 (24.35)
Performance	65.56 (19.77)	70.28 (14.50)	70.28 (21.45)
Effort	73.33 (20.72)	70.28 (20.54)	70.00 (19.85)
Frustration	37.50 (26.14)	30.56 (26.23)	27.50 (25.16)

For the Raw NASA-TLX total score, a significant angle effect was determined: ANOVA $F(2, 34) = 7.37$, $p = 0.002$, $\eta p^2 = 0.302$. Friedman's test gave the same conclusion: $\chi^2(2) = 10.62$, $p = 0.005$, $W = 0.295$. The result remained statistically significant even after Holm's correction for a set of NASA-TLX analyses. Post hoc pairwise comparisons with Wilcoxon's test with Holm's correction showed a higher total workload at the angle of 15° than at the angle of 30° (p(Holm) = 0.037) and 45° (p(Holm) = 0.005), while the angles of 30° and 45° did not differ significantly (p(Holm) = 0.177). Paired t-tests after ANOVA gave the same pattern.

Of the individual NASA-TLX dimensions, the most pronounced effect was determined for the measure of *frustration*: ANOVA $F(2, 34) = 6.73$, $p = 0.003$, $\eta p^2 = 0.284$ and Friedman's test: $\chi^2(2) = 15.55$, $p < 0.001$, $W = 0.432$. Frustration was the highest at the angle of 15° and the lowest at the angle of 45°, and the difference between 15° and 45° remained significant after a post hoc comparison with the Wilcoxon test (p(Holm) = 0.004). Mental and physical demands were also descriptively the highest at 15°, but their effects did not remain statistically significant after correction for multiple testing. No significant differences between angles were found for temporal demand, performance and effort.

In addition, it should be noted that the SDs of individual NASA-TLX dimensions were quite large, especially for physical demand, mental demand and frustration. The results indicate that there is a large individual difference in the subjective experience of the task, although at the average level a clear pattern between the angles can be established.

3.6.2. Subjective Assessment of Ease of Reading and Confidence in Response

The estimate of general ease of reading clearly differed between angles. The average value increased from 2.44 at the 15° angle to 3.61 at the 30° angle and 3.83 at the 45° angle (Table 14).

Table 14. Subjective ease of reading and confidence in response (scale 1–5; see Appendix C.2).

Subjective Assessment	15° Mean (SD)	30° Mean (SD)	45° Mean (SD)
The text was generally easy to read	2.44 (0.98)	3.61 (1.09)	3.83 (0.99)
Confidence in the accuracy of the answer	2.28 (0.89)	3.11 (0.83)	3.06 (0.73)

To assess ease of reading, Friedman's test showed significant differences between the angles: $\chi^2(2) = 17.44$, $p < 0.001$, W = 0.485. Post hoc comparisons showed that the text was estimated to be significantly more difficult to read at the angle of 15° than at the angles of 30° (p(Holm) = 0.004) and 45° (p(Holm) = 0.003). No significant difference was found between the angles of 30° and 45° (p(Holm) = 0.417). A very similar pattern emerged for participants' confidence in the accuracy of their own answers: $\chi^2(2) = 18.47$, $p < 0.001$, W = 0.513. Participants were less confident in their answers at the angle of 15° than at the angle of 30° (p(Holm) = 0.004) and the angle of 45° (p(Holm) = 0.003), while estimates between the angles of 30° and 45° did not differ significantly (p(Holm) = 0.705).

3.6.3. Subjective Assessment of the Impact of the Type of Background

Participants separately assessed the extent to which urban, natural and abstract backgrounds made the text difficult to read. Average estimates were relatively low and did not show a stable pattern of differences between angles (Table 15).

Table 15. Subjective assessment of how much a background type makes it difficult to read (scale 1–5; see Appendix C.2).

Estimated Background Interference	15°	30°	45°
Urban	2.22	1.94	1.89
Natural	2.33	2.06	2.06
Abstract	2.28	2.17	2.50

For urban, natural, and abstract complex backgrounds, no significant differences were found between the angles after correction for multiple testing. When the three types of background were compared with each other within each angle, no significant differences were also found at the angles of 15° and 30°. At 45°, the total Friedman test reached nominal statistical significance (p = 0.048), but no pairwise comparison remained significant after Holm's correction. Therefore, there is no reliable evidence that participants consistently perceived one type of complex background as more challenging than the other.

3.7. The Link Between Subjective and Objective Measures

3.7.1. Main Indicators

The Raw NASA-TLX total score showed a strong correlation with objective performance. Higher subjective workload was associated with lower *accuracy*: repeated-measures

correlation coefficient (see Appendix B.5) $r_{rm} = -0.634$, $p < 0.001$ and with higher *total_error*: $r_{rm} = 0.608$, $p < 0.001$ (Table 16). Thus, when the same participant achieved lower objective performance, they subjectively declared that they perceived the condition as more demanding. An even greater correlation was found between subjective assessment of ease of reading and objective *accuracy*: $r_{rm} = 0.694$, $p < 0.001$. A higher estimate of ease of reading was also associated with lower *total_error*: $r_{rm} = -0.684$, $p < 0.001$. Participants' confidence in the accuracy of their own answers was also significantly associated with objective performance. Higher security was associated with higher *accuracy*: $r_{rm} = 0.585$, $p < 0.001$ and lower *total_error*: $r_{rm} = -0.562$, $p < 0.001$.

Table 16. Correlation between the main subjective and objective measures.

Subjective Measure	Accuracy, r_{rm}	p	Total_Error, r_{rm}	p
Raw NASA-TLX total	−0.634	<0.001	0.608	<0.001
Ease of reading	0.694	<0.001	−0.684	<0.001
Confidence in response	0.585	<0.001	−0.562	<0.001

3.7.2. Individual Dimensions of NASA-TLX

Among the individual Raw NASA-TLX dimensions' results, the strongest association was found for frustration. Greater frustration was associated with lower *accuracy* ($r_{rm} = -0.557$, $p < 0.001$) and higher *total_error* ($r_{rm} = 0.515$, $p = 0.001$). Subjectively estimated performance was also positively associated with *accuracy* ($r_{rm} = 0.423$, $p = 0.009$) and negatively with *total_error* ($r_{rm} = -0.380$, $p = 0.021$). After Holm's correction for the six dimensions of NASA-TLX, the association of frustration remained significant for both objective measures, while performance remained significantly related to *accuracy*. The associations between mental demand, physical demand, temporal demand and effort were weaker and did not show a significant association after correction for multiple testing.

3.7.3. Subjective Assessment of Individual Types of Background

When the subjective assessment of reading difficulty for a particular type of background was associated with objective performance on the corresponding type, only a weak overall association with *accuracy* was obtained: $r_{rm} = -0.200$, $p = 0.016$, and the association with *total_error* was not statistically significant: $r_{rm} = 0.133$, $p = 0.111$. Analyses of individual types of backgrounds also did not show a consistent pattern.

3.8. Awareness of the Change in Individual Background Images

An additional question was asked whether the participants were aware that within each of the three types of background, three different images alternated during the experiment. The answers were almost entirely concentrated on the lowest value of the scale (1 out of 5). The mean value was 1.11 at the angle of 15°, 1.06 at the angle of 30° and 1.06 at the angle of 45°. At the angle of 15°, the value of 1 was chosen by 16 of 18 participants, while at the angles of 30° and 45° it was chosen by 17 of 18 participants. No statistically significant difference was found between the angles: $\chi^2(2) = 2.00$, $p = 0.368$.

The participants therefore showed virtually no awareness that the individual background images within the same background type were changing. Only two participants gave the vague impression after the experiment that "something seemed different" to them, but they were also unaware of the systematic alteration of the three different images before this was explained to them at the end of the experiment. This finding is particularly interesting in relation to the results presented in Section 3.4. Although measurable differences in objective legibility were found between individual images within the same type of background, the participants did not consciously register these changes.

4. Discussion

4.1. Summary of Main Findings and Answers to Research Questions and Hypotheses

The results of the study show that the text orientation angle had the strongest and most consistent effect on the legibility of 3D text, while the influence of the type of background was weaker and depended on the text orientation angle and differed with respect to the considered measure of objective performance. At the same time, significant variability was found between individual images of the same type of background and a clear correlation between objective results and general subjective assessments of the complexity of the task.

The present study should be viewed as an extension of our previous methodological work [20]. In that study, the primary objective was to develop the glance-based procedure and estimate an appropriate short-term presentation duration for spatially oriented 3D text under controlled, high-contrast conditions. The present study uses that methodological framework and the previously derived fixed duration of 659 ms, but shifts the research focus to performance over heterogeneous complex backgrounds. Its contribution therefore lies primarily in examining whether background type and individual background exemplars introduce systematic variability, how these effects relate to text orientation, and how objective performance corresponds to subjective workload and perceived difficulty.

Below is a summary of the answers to the research questions and hypotheses supported by the results.

RQ1—Are there differences in the objective legibility of 3D text when positioned over urban, natural and abstract complex backgrounds?

Differences were found between types of complex backgrounds, but they were not consistent for all measures and angles of view. At the level of results aggregated by participants, the lowest average *accuracy* was recorded for abstract backgrounds (74.1%), followed by urban backgrounds (74.5%), and the highest for natural backgrounds (76.6%) (Table 2). However, the effect of background type on the *accuracy* measure was not statistically significant (ANOVA $p = 0.180$, Section 3.3.1). For *total_error*, the worst result was 8.74 for abstract backgrounds, followed by 8.65 for urban backgrounds, and the best was 7.54 for natural backgrounds (Table 2). A statistically significant effect of the background type was found (ANOVA $p = 0.036$, Section 3.3.2). The results of the GLMM, which took into account several sources of variability at the same time, showed a slightly different overall ranking: the highest marginal probability of correct answer was estimated for urban backgrounds (77.8%), followed by abstract backgrounds (75.7%), and the lowest for natural backgrounds (74.2%) (Table 5). Of all the contrasts between the types of backgrounds, only the difference between urban and natural backgrounds had a 95% CrI that did not include zero (Table 4). The difference between the aggregated and GLMM-based rankings can be explained by their distinct estimands: aggregated accuracy reflects the actual combination of trials presented in each condition, whereas GLMM marginal probabilities represent standardized predictions that balance the other fixed factors and set random effects to their population means. The alternative GEE analysis supported the GLMM-based ranking, although the differences between background types remained relatively small (Appendix D).

Overall, the results show that the type of background may have some influence on objective legibility, but they do not support the existence of a single order of their complexity that would be valid regardless of the measure and sources of variability.

RQ2—Is there an interaction between the type of complex background and the text orientation angle of 3D text in their effect on objective legibility?

The results provide partial and measure-dependent support for an interaction between angle and background type, as the evidence was not fully consistent across the different outcome measures and analytical approaches. The GLMM has shown that the relative performance of individual types of background changes depending on the text orientation

angle, with the best being the urban background for 15° (70.8%), natural background for 30° (82.0%), and abstract background for 45° (86.4%) (Table 6). Analysis at the level of aggregated results of participants showed a statistically significant angle × background type interaction for *total_error* (Section 3.3.2), but for *accuracy* the interaction was not statistically significant (Section 3.3.1). This difference is relevant because *accuracy* distinguishes only between fully correct and incorrect responses, whereas *total_error* also captures how far an incorrect response deviates from the correct value. The stronger interaction for *total_error* may therefore indicate that some angle–background combinations affect the precision of responses when errors occur, with incorrect responses remaining closer to the correct value in some conditions and showing larger deviations in others. The analysis by individual angles showed significant differences between the background types for both measurements only at the angle of 30°, where the natural background achieved significantly higher *accuracy* than the urban one, while for the other combinations of background types there were no significant differences (Sections 3.3.1 and 3.3.2). Taken together, the results suggest that the effect of background type may vary with text orientation angle, but that this interaction is not expressed equally strongly across the two objective measures. Moreover, the additional GEE analysis supported an overall interaction, particularly through changes in the relative performance of background types at 45°, while individual contrasts at 30° were less robust (Appendix D).

RQ3—Do subjective workload, estimated legibility and response confidence change depending on the text orientation angle, and do subjective assessments of reading difficulty differ between urban, natural, and abstract backgrounds?

Subjective workload, estimated legibility and response confidence clearly differ depending on the angle. The 15° angle was associated with significantly higher overall workload and frustration (Table 13) and significantly lower reading ease and lower confidence in one's own responses compared to the 30° and 45° angles (Table 14). At the 30° and 45° angles, subjective estimates were more favorable and similar to each other, with no significant differences. Subjective assessments of reading difficulties associated with urban, natural, and abstract backgrounds showed no statistically significant differences (Table 15).

RQ4—To what extent is objective legibility associated with subjective workload and other subjective assessments?

Objective and general subjective measures of legibility showed a strong correlation. Higher subjective workload was associated with lower *accuracy* and higher *total_error*, while higher legibility and greater confidence in responses were associated with higher *accuracy* and lower *total_error* (Table 16). For the results of individual subjective dimensions, the strongest correlation with objective measures was found for frustration, where greater frustration indicated significantly lower *accuracy* and higher *total_error*, and a positive correlation for the *accuracy* measure was also established for the subjective dimension of performance (Section 3.7.2). It was shown that the correlation between objective results and the subjective assessment of difficulty reading a particular type of background is weak (Section 3.7.3).

RQ5—What is the variability in legibility between individual background images within the same background type?

The variability in legibility between individual background images is considerable. In several combinations of background type and angle, individual images of the same type led to statistically significantly different objective performance, with the highest *accuracy* ranges exceeding 20 percentage points (Table 7). In doing so, the GLMM confirmed the existence of variability associated with an individual background image, but showed even greater variability related to the identity of the target letter and, most of all, to a specific text stimulus (Table 9). This suggests that the differences in legibility between individual

images result not only from the general visual characteristics of the background, but also from the relationship between the specific background image and the sequence of displayed characters. The position and shape of the target letter, the structure of the surrounding characters, and the local visual properties of the background may together increase or decrease the severity of a particular stimulus. The additional result from Section 3.8 shows that, despite the differences in *accuracy* achieved between individual images of the same type of background, participants were almost unaware that three different images were alternating.

H1. *Legibility of 3D text will increase with an increase in the text orientation angle from 15° to 45°, with the lowest success rate at 15° and the highest at 45°.*

H1 is supported. Objective performance increased from the angle of 15° to the angle of 45°, with the angle of 15° consistently representing the most challenging condition and the angle of 45° the easiest (Table 2). The performance assessment with the GLMM gave the best score at the angle of 45° and the worst at 15° (Table 3), while the difference between the angles of 30° and 45° was not significant. The analysis of the results aggregated at the level of participants showed statistically significant differences between all three angles (including a significant difference between the angles of 30° and 45°) for both objective measures (Sections 3.3.1 and 3.3.2).

H2. *Lower objectively measured legibility will be associated with a higher subjectively measured workload and less favorable subjective assessments of the task.*

H2 is supported. Significantly lower objective legibility *accuracy* was associated with higher Raw NASA-TLX total load and frustration, lower subjective legibility, and lower confidence in one's own responses. Significantly higher *total_error* was associated with higher workload and lower subjective assessments of legibility and confidence in one's own responses (Table 16).

4.2. Variability Between Individual Background Images and Textual Stimuli

One of the most important findings of this study is the significant variability of legibility between individual background images of the same type. The GLMM allowed for additional separation of several sources of this variability, as shown in Table 9. At the same time, the variability associated with a specific textual stimulus was significantly higher than the variability associated with a single background image alone. This shows that the difficulty of a single trial probably results from a combination of several factors: the number of target letters, the identity of the requested letter, the structure of the entire string of characters, and the visual properties of the background on which the characters are displayed.

The stimulus component can therefore be broken down into several sources of variability that together affect the final result. First, in complex backgrounds, *local contrast* can play an important role in legibility. The same background image may contain areas of different color, texture and visual complexity. Therefore, individual letters within the same sequence may not have the same visibility conditions. A target letter displayed over a homogeneous and contrasting area may be more easily recognizable than the same letter displayed over an area of similar color, pronounced texture or strong local edges. Secondly, the source of variability can also be represented by *the structure of the character sequence* itself, namely the position of the target letter in the sequence of characters, the characteristics of neighboring characters and their visual similarity. All these factors can affect the difficulty of a particular stimulus.

This consideration could also explain the observed large differences in results between stimuli even when the number of target letters is the same (see Table 8). The results therefore

suggest that the effect of a complex background on legibility should be viewed primarily as a local interaction of text and background, and secondarily as a global property of a single image.

It is important to emphasize that the determined stimulus variability does not call into question the robust results of the study. Namely, textual stimuli were randomized between conditions, and success estimates were based on a larger number of trials, which significantly reduces the possibility that individually light or heavy sequences systematically favor a particular experimental condition. The effect of the text orientation angle is particularly strong. In contrast, the effects of the background type show greater sensitivity to the structure of the model, which further supports their more cautious interpretation. Interestingly, objective differences between individual background images were found even though participants generally did not consciously notice that images within the same type of background were changing. This further suggests the possibility that the local visual properties of the background affect character recognition without a clear conscious assessment of the source of difficult legibility.

The existing limitation of the study is that the experiment did not allow for a complete separation of the inherent difficulty of a specific sequence of characters (stimulus) from its interaction with a specific background image because the same sequences were not systematically displayed over several different background images. The reason for this was to eliminate the possibility of remembering the sequence and thus making it easier to establish the correct answer in the participants. Therefore, the role of local contrast and sequence structure should be considered as an explanation that is consistent with the observed results, but which requires further direct experimental verification.

This finding also provides an interesting perspective on the role of contrast. Previous studies and reviews have identified text–background contrast as one of the major determinants of text legibility in AR, particularly for 2D text [7,11,12]. Gabbard et al. [7] further demonstrated that the effect of a complex background cannot be characterized adequately by its overall visual complexity alone. In their study, performance on a globally textured brick-wall background was relatively good when the text was positioned over a locally homogeneous region, suggesting that the visual properties of the background immediately surrounding the text may be more important than those of the image as a whole. This observation was subsequently emphasized by Cauz et al. [11], who related local background structure to variations in text–background contrast.

A comparable pattern can be observed in the present study. Natural backgrounds were substantially darker than urban and abstract backgrounds and therefore provided, on average, greater potential contrast with the light text, yet they did not consistently produce the highest recognition performance. Conversely, the brightest abstract backgrounds produced the highest model-estimated *accuracy* at the 45° text orientation angle. These findings should not be interpreted as evidence that contrast is unimportant for 3D text, since local text–background contrast was not directly manipulated or measured. Rather, they suggest that global background brightness and overall text–background contrast alone are insufficient to predict the legibility of spatially oriented 3D text. In this case, local contrast operates together with the spatial orientation and projected geometry of the characters, the visibility of their frontal and side surfaces, and the local visual structure of the background.

4.3. Influence of Age on Legibility

The secondary age-group analysis suggested lower objective performance with increasing age. According to Table 10, the youngest group (18–35 years) achieved the highest average *accuracy* (81.2%) and the lowest *total_error* (6.2), the middle age group (36–50 years) took an intermediate position (75.5%; *total_error* = 8.0), while the group from 51 to 70 years

achieved the lowest *accuracy* (68.5%) and the highest *total_error* (10.7). However, a statistically significant difference was found primarily between the youngest and oldest groups, while the differences between the youngest and middle and the middle and oldest groups did not reach the level of significance.

One of the possible reasons is the high variability of results within the mean age group. In the case of *accuracy*, the 36–50 years group had the highest standard deviation (SD = 8.83), significantly higher than the 18–35 years group (SD = 3.37), and a similar pattern can be seen for *total_error*. Due to the greater overlap of the results of the middle group with both neighboring age groups, these differences are more difficult to statistically clearly separate, although their mean values follow the expected age gradient.

An additional limitation is the small number of participants, with only six participants in each age group. The simulation-based sensitivity analysis confirmed that the present design had adequate sensitivity primarily to relatively large age-related interaction effects (Appendix F). The absence of statistically significant age × angle and age × background-type interactions should therefore not be interpreted as evidence that smaller or moderate interactions are absent.

4.4. The Relationship Between Subjective Assessments and Sources of Task Difficulty

Subjective estimates of the overall difficulty of the task corresponded well with objective performance. Higher workload was associated with lower *accuracy* and higher error, while greater ease of reading and confidence in response were associated with better objective performance (Tables 13–15). Participants therefore scored relatively well when the task was easier or more difficult to perform.

However, subjective assessments of the difficulty of reading by a particular type of background showed only a weak and inconsistent association with objective results. Such a finding can be related to the structure of variability determined by the GLMM (Table 9). The variability associated with a specific textual stimulus and the identity of the target letter was significantly higher than the variability associated with an individual background image. This suggests that the background was not the only, nor necessarily the dominant, source of the difficulty of a particular trial. Legibility was probably due to the joint action of the properties of the text, the target letter and the visual background.

As a result, participants can confidently feel that a particular condition was more challenging, but at the same time, it is difficult to single out the background itself as the cause of this difficulty. This result supports the combination of subjective assessments with objective measures, as the subjective experience reflects well the overall difficulty of the task, but may not reliably identify the individual visual factor contributing to it.

From a practical perspective, the weak association between subjective background-difficulty ratings and objective performance, together with the relatively small and non-uniform objective differences between background types, suggests that no broad background type can be considered consistently unsuitable for text presentation. Rather than avoiding a particular category of background, designers should pay greater attention to the specific visual characteristics of the background and to the text orientation angle under which the text is presented.

4.5. Complementarity of Accuracy and Total_Error Measures

Accuracy and *total_error* capture related but distinct aspects of performance. *Accuracy* is a strict binary measure: any deviation from the correct number of target letters is treated as an incorrect response, regardless of whether the participant missed by one or by several units. In contrast, *total_error* preserves information about the magnitude of that

deviation. Consequently, two conditions may show a similar proportion of completely correct responses while differing in how large the errors are when they occur.

This distinction is relevant to the angle × background-type interaction observed in the present study. The interaction was not statistically significant for aggregated accuracy, but it was significant for *total_error*. One possible interpretation is that background conditions do not consistently alter the probability of producing a fully correct response, but may influence the precision of responses when errors occur. In some conditions, incorrect responses may still remain close to the correct number of target letters, whereas in more difficult conditions larger deviations may occur, suggesting greater uncertainty in estimating the number of targets. In this sense, *total_error* may be more sensitive to subtle changes in task difficulty that are not captured by binary accuracy alone.

The two measures should therefore be interpreted as complementary rather than interchangeable indicators of objective legibility. The stronger background-related pattern observed for *total_error* suggests that the effects of complex backgrounds may be expressed partly through error severity and not solely through the frequency of completely incorrect responses.

4.6. Study Limitations and Future Research Challenges

The research was conducted as a controlled screen-based simulation and not in a real handheld MR environment. Such an approach allowed for precise control of the experimental conditions, but did not capture the dynamic factors of real use, such as device and user movement, continuously changing viewpoints, variable viewing distance, dynamic illumination, real-time camera capture, and compositing and latency. In real MR use, these factors may continuously change the visual relationship between the text and its surrounding environment and thereby affect its legibility. The results therefore need to be further validated under real MR conditions and should therefore not be generalised directly to all handheld or head-mounted MR systems.

Viewing distance was monitored only approximately (about 50–70 cm) and was not mechanically fixed or precisely recorded. Consequently, exact retinal character size and visual angle cannot be reported, and minor participant-to-participant and within-session variation in retinal size and viewing geometry may have occurred.

The preparation of the backgrounds sought to avoid distinct differences in general brightness and visual impression, but retained the inherent properties of urban, natural and abstract scenes. Therefore, the identified differences between background types cannot be attributed to a single isolated pictorial property, but should be interpreted as the result of a combination of several characteristics of a particular visual environment. In addition, each background type was represented by three images, which limits the extent to which the observed differences can be generalized to urban, natural, and abstract backgrounds as broad visual classes. The present results therefore do not support strong category-level conclusions about relative legibility and do not identify any of the three tested background types as consistently unsuitable for text presentation. Future studies should focus on a more detailed characterization of local background properties, particularly local text–background contrast and other visual features in the region occupied by the text, together with tighter control of stimulus generation and selection.

At the level of individual background images, each angle × image combination was represented by only ten textual stimuli. Consequently, complete balance in target-letter characteristics and intrinsic sequence difficulty could not be ensured at this level (see Appendix E), and some individual-image differences may partly reflect stimulus composition. A substantially larger number of stimuli per combination would improve such balance but would also considerably increase experimental duration and participant burden.

In addition, the present findings are specific to the tested 3D text configuration and should not be generalized directly to other typefaces, tracking settings, extrusion depths, or text colors. These parameters were kept fixed to isolate the effects of background type and text orientation and to keep the experimental duration feasible.

The nominal display time of 659 ms was realized on a 60 Hz refresh-rate display, so the actual duration of each display could vary by approximately one refresh cycle (about 16.7 ms). This timing uncertainty was not independently measured using external timing equipment. However, because the same hardware, software, and timing procedure were used for all participants and experimental conditions, such variation would be expected to introduce primarily non-systematic timing noise rather than a systematic bias favoring a particular angle or background condition. Across the large number of randomized trials, this small trial-to-trial variation is therefore unlikely to account for the observed between-condition differences, although it remains a limitation of the absolute precision of stimulus-duration control.

Uncertainty estimates from the primary Bayesian GLMM rely on a mean-field variational approximation, which may underestimate posterior uncertainty for weaker effects; for this reason, the principal background-related findings were additionally evaluated using an alternative frequentist GEE analysis (Appendix D).

From a practical perspective, the strong orientation effect observed in this controlled screen-based simulation suggests a possible direction for future AR/MR design, but should not be interpreted as a directly validated design recommendation. One such direction is a distance-dependent level-of-detail (LoD) approach [24], in which short 3D labels could be presented at greater distances in a more oblique spatial orientation and progressively become more front-facing as the user approaches the referenced object. At close viewing distances, the label could transition to a two-dimensional presentation containing more detailed information, thereby adapting both the spatial presentation and the amount of textual information to viewing distance. Because this adaptive transition was not tested in the present study, its effectiveness should be evaluated directly under more realistic and dynamic AR/MR conditions.

5. Conclusions

This study showed that the legibility of short-term displayed 3D text on complex backgrounds was most strongly influenced by the angle of display. The lowest performance was consistently found at the angle of 15°, while the results at the angles of 30° and 45° were significantly better. The effect of the background type was weaker and did not show a uniform order of difficulty independent of the angle and the measure used. At the same time, significant variability was found between individual background images. Subjective assessments of load, ease of reading, and confidence in response correlated with objective performance, while secondary exploratory analyses suggested a gradual decline in objective performance with increasing age, although the small number of participants per age group limits conclusions regarding age-related effects and interactions.

Future research should include a larger sample of participants in order to increase the reliability of the estimates, especially for age differences and weaker interaction effects. Future experimental design should also include a non-complex monochrome background with at least one type of complex background, to directly compare the legibility of 3D text in these two types of visual environments. Taken together with previous recommendations that generally favor front-facing 2D text, the present findings suggest that 3D text should not be treated as a single display condition. Its legibility appears to depend strongly on its spatial configuration, indicating that future comparisons between 2D and 3D text should explicitly control orientation, extrusion depth, projected character geometry, and

background conditions. The present findings therefore provide a controlled basis for future validation of spatially oriented 3D text under real AR/MR conditions.

Author Contributions: Conceptualization, A.J. and L.M.; methodology, A.J. and L.S.-K.; software, A.J.; validation, A.J., L.S.-K. and L.M.; formal analysis, A.J.; investigation, A.J.; resources, A.J. and L.M.; data curation, A.J.; writing—original draft preparation, A.J.; writing—review and editing, L.S.-K. and L.M.; visualization, A.J.; supervision, L.S.-K. and L.M.; project administration, L.S.-K.; funding acquisition, L.S.-K. All authors have read and agreed to the published version of the manuscript.

Funding: Research was funded by the European Union—Next Generation EU as part of the project "Application of Immersive Technologies, Geolocation, and Digital Twins for Improving Training and Field Support in Firefighting (IGNIS)", grant number NPOO 581-16942. The views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and was approved by the Ethics Committee of the Univeristy of Zagreb, Faculty of Graphic Arts on 21 March 2025.

Informed Consent Statement: Informed consent was obtained from all participants involved in the study.

Data Availability Statement: The data used in this study may be made available upon reasonable request. Interested researchers should contact the first and corresponding author.

Acknowledgments: During the preparation of this manuscript, the authors used ChatGPT (GPT-5.6 Sol from OpenAI) as a background image generation tool and AI-assisted writing tool. It was used to assist with the language editing and refinement of the manuscript text. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AR	Augmented Reality
CrI	Credible Interval
GLMM	Generalized Linear Mixed Model
HMD	Head-Mounted Display
MR	Mixed Reality
NASA-TLX	NASA Task Load Index
RQ	Research Question
SD	Standard Deviation
VR	Virtual Reality
XR	Extended Reality

Appendix A. Bayesian GLMM Technical Details

Appendix A.1. Model Implementation and Evaluation

The binomial GLMM was estimated using the Python programming language using the BinomialBayesMixedGLM class from the *statsmodels* library [25]. The dependent variable was the accuracy of the single trial (correct: 0 = incorrect answer, 1 = correct answer), and the model used the logit link function, according to the expression (1) in the main text. More specifically, the analyses were performed in the Python 3.13.5 programming language, with the libraries NumPy 2.3.5, pandas 2.2.3, SciPy 1.17.0 and statsmodels 0.14.6. For the fixed part of the final model according to the expression (1), an angle of 15°, an

abstract background, an age group of 18–35 years and correct_answer = 0 were used as reference levels.

The mixed model included fixed and random effects as mentioned in Table 1, and was estimated by the fit_vb method, which uses a *mean-field variational Bayes* approximation of the posterior distribution. The variational Bayesian inference replaces the hard-to-compute posterior distribution with a simpler family of approximate distributions, whose parameters are determined by optimization, in this case by minimizing the Kullback–Leibler divergence between the approximate and target posterior distributions [26]. Therefore, the posterior estimates of values and 95% CrI obtained by the GLMM should be interpreted as approximations of the full Bayesian posterior distribution. The estimation using the fit_vb method did not change the other default settings of the method.

The suitability of the binomial GLMM was based on the binary trial-level outcome and the repeated and crossed structure of the experimental data. Numerical reproducibility was ensured using a fixed random seed, while the stability of the principal background-related conclusions with respect to the estimation framework was assessed through the alternative GEE analysis reported in Appendix D.

Appendix A.2. Priori Distributions

For fixed-effect coefficients, a normal prior with a mean of 0 and a standard deviation of 5 was used: $\beta_j \sim N(0, 5^2)$, which in the implementation of BinomialBayesMixedGLM corresponds to the parameter fe_p = 5. Such a prior is wide on the logit scale and provides weak prior information, while at the same time reducing the possibility of extreme and unstable estimates of coefficients when the data themselves are not sufficiently informative.

For random effects, it is assumed that the individual levels within each component follow a normal distribution around zero: $u_{ik} \sim N(0, s_k{}^2)$, where k denotes the random effect component (e.g., participant, target letter, or specific stimulus), and the individual level within that component (e.g., a specific participant or a specific target letter), $u_{ik,}$ its random effect, and s_k the standard deviation of that component. For the logarithm of the standard deviation of each component, the normal prior was used: $\log(s_k) \sim N(0, 0.5^2)$, which in the implementation of BinomialBayesMixedGLM corresponds to the parameter vcp_p = 0.5. In this way, the possibility of extremely small or large estimates of standard deviations (SD) of random effects is limited and contributes to a more stable estimation of models with multiple sources of random variability. The same specification of the prior was used for all k random effects.

Appendix A.3. Model-Estimated Marginal Probabilities and Posterior Contrasts

For easier interpretation of the results on the probability scale, and not on the logit scale, the GLMM results were converted into a model-estimated marginal probability of a correct answer. For the factor of interest, the predictions were calculated for each of its levels, and then averaged over the levels of other fixed factors with equal weights. For example, the marginal probability for a particular angle was obtained by averaging model predictions over three types of backgrounds, three age groups and four levels of correct_answer. When calculating population marginal probabilities, random effects were set to their population mean, i.e., to zero on the random effects scale. Such a procedure results in comparisons in which the distribution of other fixed factors is the same for all compared levels. For the interaction, angle × background type, marginal probabilities were calculated separately for each angle and background type combination, averaging across age groups and correct_answer levels.

The posterior contrasts shown in Table 3 were calculated directly within each posterior realization. For example, the contrast of $30° - 15°$ was obtained by subtracting the marginal

probability estimated for an angle of 15° from the marginal probability estimated for an angle of 30° within the same realization. In this way, the posterior distribution of the difference between the conditions was obtained. The contrasts on the probability scale are shown in percentage points.

Appendix A.4. Monte Carlo Evaluation and Credibility Intervals

In order to transfer the uncertainty of the estimated fixed effects to marginal probabilities and their contrasts, 1,000,000 Monte Carlo realizations were generated from the approximate posterior distribution of fixed effects. A single Monte Carlo realization represents a randomly sampled possible set of values of fixed-effects coefficients from their estimated posterior distribution. A large number of Monte Carlo realizations were used to reduce numerical variability in the calculation of marginal probabilities, posterior contrasts and probability interval limits. For the purpose of numerical reproducibility of the procedure, a fixed random number generator with a seed value of 20,260,913 was used, applied in the initialization of both the variational Bayesian procedure and Monte Carlo sampling of the posterior distribution.

For each realization, the corresponding linear predictors were calculated, then converted to probabilities using the inverse logit function, and then averaged according to the marginalization procedure described in Appendix A.3. Posterior contrasts were calculated from the marginal probabilities obtained within each realization.

The estimated values of marginal probabilities and contrasts were obtained from the corresponding Monte Carlo distributions. The approximate 95% CrI were determined by the 2.5 and 97.5 percentiles of these distributions. For the contrast between the two conditions, the difference was considered significant in the Bayesian sense if the corresponding 95% CrI did not include zero. This means that, according to the estimated Bayesian model, 95% of the plausible values of the difference were on the same side of zero. Such an estimate is conceptually different from the statistical significance determined by the $\alpha = 0.05$ criterion for the frequentistic analyses, and no p-values were used for GLMMs.

The presented marginal probabilities and posterior contrasts were determined as the mean values of the corresponding Monte Carlo distributions. For the components of random effects, BinomialBayesMixedGLM estimates the parameters on the standard deviation logarithm scale. The estimated standard deviations shown in the results were obtained by the transformation: $\widehat{s_k} = \exp\left(\widehat{\log(s_k)}\right)$, and the limits of the approximate 95% CrI by the transformation: $\exp\left(\widehat{log(s_k)} \pm 1.96\ SD_{\log(s_k)}\right)$.

Appendix B. Details of Statistical Analyses of Aggregated Scores by Participant

Appendix B.1. Angle and Type of Wallpaper

For participant-level analyses, the measures *accuracy* and *total_error* were aggregated for all nine combinations of three angles and three types of background. For each combination, the angle × background type, the score of each participant was based on 30 trials, or three individual background images with 10 trials each.

The main effect of the angle was analyzed on the results aggregated across the three types of background, and the main effect of the type of background was analyzed on the results aggregated across the three angles. Both main effects were analysed by the ANOVA of repeated measurements. The interaction of angle × types of background was analysed by the two-way ANOVA of repeated measurements on nine combinations of these two within-participant factors.

For non-parametric verification of the main effects, Friedman's test was additionally used. When Friedman's test showed the total difference between the levels of factors, pairwise comparisons were performed using the Wilcoxon signed-rank test with Holm's correction of *p*-values within the appropriate group of paired comparisons.

Appendix B.2. Variability of Individual Background Images

The variability between the individual images was analyzed separately for each combination of background type and angle. Three abstract (ca1–ca3), three natural (cn1–cn3) and three urban (cu1–cu3) backgrounds were compared at each of the three angles, giving a total of nine groups of comparisons.

For each combination, the participant × angle × individual background image, the *accuracy* value was calculated based on 10 trials. Since each participant was exposed to all three images within each type of background and angle, the Friedman test for related measurements was used to compare the three images. In order to control the multiple tests, the *p*-values of the nine Friedman tests were corrected jointly by the Holm procedure. For the groups in which the difference between the individual images was determined after this correction, pairwise Wilcoxon comparisons were performed. The three paired comparisons within each such group formed a separate family of tests and were again applied to them by the Holm correction.

Appendix B.3. Age Groups

To analyze the total differences between age groups, the total *accuracy* score was calculated for each participant and *total_error* was aggregated across all nine combinations of angle and background type. The differences between the three independent age groups in *accuracy* and *total_error* were analyzed by the one-way ANOVA. The normality of the distributions was estimated by the Shapiro–Wilk test, and the homogeneity of the variances by the Levene test. Due to the small number of participants per age group and the possibility of unequal variances, the results were additionally verified by the Welch ANOVA and the non-parametric Kruskal–Wallis test.

For significant overall results, appropriate post hoc paired comparisons were performed: Tukey's HSD after standard ANOVA, the Games–Howell method after Welch's ANOVA, and Mann–Whitney's U-test after the Kruskal–Wallis test. For three Mann–Whitney comparisons between age groups, Holm's correction was applied.

In order to examine whether the effect of experimental factors changes depending on the age group, the mixed two-way ANOVA: age × angle and age × background type were additionally performed, separately for *accuracy* and *total_error*. Age was a between-participant factor, and angle or background type was a within-participant factor. In the age × angle analysis, the results for each angle were aggregated across three types of background, while in the age × background type analysis, the results for each type of background were aggregated across three angles.

Appendix B.4. Subjective Measures

The differences in the Raw NASA-TLX total score between the three angles were analyzed by the ANOVA of repeated measurements and the Friedman test as a non-parametric check. In the same way, the individual NASA-TLX dimensions were analyzed in an exploratory manner. The Raw NASA-TLX total score and six individual dimensions formed a family of seven analyzed outcomes, and Holm correction was applied to their total *p*-values due to multiple testing.

For outcomes in which the effect of the angle was determined, pairwise comparisons between the angles of 15°, 30° and 45° were performed. Paired t-tests were used for

parametric analyses, and Wilcoxon signed-rank tests were used for non-parametric analyses. Three pairwise comparisons between angles formed a separate family of tests, and Holm correction was applied to their p-values.

Subjective estimates of ease of reading and certainty in response, collected on a five-point numerical rating scale of 1 to 5 after each angle, were analyzed by the Friedman test. When the total difference between the angles was determined, three pairwise Wilcoxon comparisons were performed with Holm's correction.

Subjective estimates of reading difficulties associated with abstract, natural, and urban backgrounds were also analyzed as related measurements. The differences between the three types of background were examined separately within each angle by the Friedman test, and if necessary, pairwise Wilcoxon comparisons were performed with Holm's correction.

The answers to the question whether individual images within the same type of background are perceived as equal were primarily analyzed descriptively as an indicator of consciously noticing a change in a particular background image. Additionally, Friedman's test was used to check whether the answers differed between the three angles.

Appendix B.5. The Link Between Subjective and Objective Measures

The correlation between subjective and objective results was analyzed by *repeated-measures correlation*, r_{rm} [27]. Unlike the standard Pearson or Spearman correlation, this procedure takes into account that multiple observations originate from the same person and assesses the common within-participant association between the two variables.

For the analyses of the subjective assessment of the general difficulty of the task, the basic unit of observation was the combination of participant × angle. Each of the 18 participants thus contributed three related observations, for a total of 54 observations. The subjective estimate at each angle was related to the *accuracy* and *total_error* values of the same participant at the same angle, where the objective results were aggregated across three types of background. The correlations of the Raw NASA-TLX total score, subjective assessment of ease of reading and confidence in response with both objective measures were examined.

In addition, the correlations of individual NASA-TLX dimensions with objective measurements were analyzed. Due to the larger number of these comparisons, their p-values were corrected by the Holm procedure.

The correlation between the subjective assessment of the difficulty of reading with a particular type of background and objective performance was analysed at the level of participant × angle × type of background. In this analysis, the subjective assessment of a certain type of background was associated with the values of *accuracy* and *total_error* of the same participant at the same angle and for the same type of background.

Appendix B.6. Statistical Software

Statistical analyses of aggregated results were performed in Python 3.13.5 using the SciPy 1.17.0 and statsmodels 0.14.6 libraries, in addition to NumPy 2.3.5 and pandas 2.2.3 for data processing and aggregation. ANOVAs of repeated measurements were performed using the AnovaRM implementation from the statsmodels library. Friedman's, Wilcoxon's, Shapiro–Wilk's, Levene's, Kruskal–Wallis, Mann–Whitney's, and paired t-tests were performed using SciPy library functions. Welch's ANOVA was performed using anova_oneway from the statsmodels library with the assumption of unequal variances, and Holm's correction was performed using multipletests(method="holm"). Games–Howell comparisons were calculated directly in Python using Welch's standard errors and degrees of freedom, and the studentized range distribution from the SciPy library. Mixed ANOVAs and repeated measurement correlations were calculated directly in Python according to

the corresponding statistical formulations. The binomial GEE robustness analysis was performed using the GEE class from the statsmodels library, with a binomial family and logit link, an Exchangeable working correlation structure, and bias-reduced covariance estimation specified through fit(cov_type="bias_reduced").

Appendix C. Questionnaires Used in the Study (Translated from Croatian)

Appendix C.1. Raw NASA-TLX Questionnaire—Subjective Workload Assessment During the Task

Text orientation angle:
Name and surname:
Date:

Instructions:

You have just completed a part of the study in which the text was presented at one text orientation angle over different complex backgrounds. Please assess your experience during the part of the experiment you have just completed. For each item, select a value from 0 to 100 in increments of 5, where 0 indicates a very low level and 100 indicates a very high level.

No.	Dimension	Question	Response Scale
1.	Mental Demand	How mentally demanding was the task (concentration, letter recognition, and decision-making)?	0–100, in increments of 5
2.	Physical Demand	How physically demanding was the task (eye strain, body position, and hand movements)?	0–100, in increments of 5
3.	Temporal Demand	To what extent did the short duration of the word display make successful completion of the task more difficult?	0–100, in increments of 5
4.	Performance	How satisfied were you with your own performance in completing the task?	0–100, in increments of 5
5.	Effort	How much effort did you have to exert to achieve your level of performance?	0–100, in increments of 5
6.	Frustration	To what extent did you feel insecure, discouraged, irritated, stressed, or angry during the task?	0–100, in increments of 5

Scale values: 0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100.

Appendix C.2. Additional Subjective User-Experience Questionnaire

No.	Item	Response Scale
1.	The displayed text was generally easy to read.	1 = Strongly disagree; 5 = Strongly agree
2.	The urban background made the text more difficult to read.	1 = Strongly disagree; 5 = Strongly agree
3.	The natural background made the text more difficult to read.	1 = Strongly disagree; 5 = Strongly agree
4.	The abstract (colorful) background made the text more difficult to read.	1 = Strongly disagree; 5 = Strongly agree
5.	I was generally confident in the accuracy of my responses.	1 = Strongly disagree; 5 = Strongly agree
6.	How confident are you that different background images were presented within each of the three background categories (urban, natural, and abstract)?	1 = Not at all confident that different images were presented; 5 = Completely confident that different images were presented

Note: The questionnaires were administered in Croatian. The English wording presented here is a translation of the administered items. The meaning of the response scale was explained orally to participants. For Items 1–5 of the additional questionnaire, 1 indicated *strongly disagree* and 5 indicated *strongly agree*. For Item 6, participants were explicitly instructed that 1 indicated no confidence that different images had been presented within a background category, whereas 5 indicated complete confidence that different images had been presented.

Appendix D. Robustness Analysis Using Generalized Estimating Equations

To assess whether the principal background-related findings depended on the mean-field variational Bayesian approximation used in the primary GLMM, an additional frequentist binomial generalized estimating equations (GEE) analysis was conducted using the same 4860 trial-level observations [28]. The analysis used a logit link function, with participants treated as clustering units and an exchangeable working correlation structure to account for repeated responses within participants. Bias-reduced covariance estimation [29] was used to improve the estimation of uncertainty given the relatively small number of participant clusters (18 clusters).

The model included text orientation angle, background type, their interaction, age group, and the number of target-letter occurrences (correct_answer) as categorical predictors. Target-letter identity was additionally included as a categorical adjustment factor, while variation among the three individual images within each background type was represented using centered within-type image contrasts. Unlike the primary Bayesian GLMM, this approach did not rely on a variational approximation to the posterior distribution. However, it also did not reproduce the full crossed random-effects structure of the GLMM and was therefore used as a complementary sensitivity analysis rather than as a replacement for the primary model.

The resulting standardized marginal probabilities are shown in Table A1. The overall estimates were 77.9% for urban backgrounds (95% CI: 73.7–82.1%), 77.4% for abstract backgrounds (73.0–81.9%), and 74.6% for natural backgrounds (72.0–77.2%). Thus, the overall ranking was consistent with that obtained from the primary GLMM.

The urban–natural contrast was 3.3 percentage points in favor of urban backgrounds (95% CI: 0.2–6.5), while the urban–abstract contrast was 0.5 percentage points (95% CI: −1.9–2.8) and the natural–abstract contrast was −2.9 percentage points (95% CI: −6.3–0.6). Accordingly, as in the primary GLMM, the urban–natural comparison was the only overall background contrast whose interval excluded zero, although the lower confidence limit was close to zero.

Table A1. GEE-estimated marginal probabilities of a correct response by text orientation angle and background type.

Angle	Abstract (95% CI)	Natural (95% CI)	Urban (95% CI)
15°	65.6% (60.9–70.3)	63.9% (60.3–67.4)	*70.2% (65.6–74.8)*
30°	78.7% (72.2–85.1)	*82.1% (78.0–86.2)*	81.6% (77.0–86.2)
45°	*88.0% (84.0–92.0)*	77.8% (73.5–82.1)	81.9% (76.6–87.2)
Overall	77.4% (73.0–81.9)	74.6% (72.0–77.2)	*77.9% (73.7–82.1)*

Note: Marginal probabilities were standardized by equal weighting of age groups and correct_answer levels. Target-letter identity was represented using sum-to-zero coding, and predictions were evaluated at the centered value of the target-letter contrasts and the within-background-type image contrasts, corresponding to zero contributions of these terms to the linear predictor. Overall background estimates additionally used equal weighting of the three text orientation angles.

The GEE analysis also indicated an overall angle × background-type interaction (Wald $\chi^2(4) = 25.98$, $p < 0.001$). Examination of the angle-specific estimates showed that not all simple contrasts were equally robust. At 30°, none of the three pairwise background contrasts had a 95% CI excluding zero. In contrast, the most pronounced differences occurred at 45°, where the estimated probability for abstract backgrounds exceeded that for urban backgrounds by 6.1 percentage points (95% CI: 3.1–9.1) and that for natural backgrounds by 10.2 percentage points (95% CI: 6.7–13.7).

Direct difference-in-differences contrasts further supported the interaction. The change in the urban–abstract contrast from 15° to 45° was −10.7 percentage points (95% CI: −16.3 to −5.1), while the corresponding change in the natural–abstract contrast was −8.4 percentage points (95% CI: −15.9 to −1.0). These results indicate that the relative effect of background type changed with text orientation angle, while also showing that individual simple contrasts, particularly those at 30°, were more sensitive to the analytical approach than the overall interaction pattern.

The alternative analysis supported the main conclusion that background-related effects were relatively small and did not establish a universally easiest or most difficult background type. It also provided supporting evidence for an angle × background-type interaction using an alternative frequentist estimation framework, while indicating that the interpretation of individual background contrasts should remain cautious.

Appendix E. Balance of Target-Letter Characteristics Across Experimental Conditions

To evaluate whether systematic differences in target-letter characteristics could have contributed to the observed condition effects, an additional balance analysis was performed on the 270 unique experimental stimulus configurations. Each stimulus configuration was counted once, rather than using all 4860 participant-level observations, because the same stimulus set was presented to all participants. Two characteristics were examined: (1) the vertical typographic zone of the target letter and (2) the number of occurrences of the target letter in the nine-character sequence (correct_answer).

Target letters were grouped according to their vertical extent in the typeface used in the experiment: lower-zone letters (*g*, *p*, *y*), upper-zone letters (*t*, *d*, *b*, *f*, *k*, *h*), and mid-zone

letters (*a*, *e*, *c*, *o*, *r*, *s*, *u*, *v*, *x*, *n*, *m*, *z*). The distribution of these groups and of target-letter occurrence counts is shown in Table A2.

Table A2. Distribution of target-letter vertical-zone groups and occurrence counts across the 270 unique experimental stimulus configurations.

Experimental Condition	*n*	Lower-Zone, *n* (%)	Mid-Zone, *n* (%)	Upper-Zone, *n* (%)	0 Occurrences, *n* (%)	1 Occurrences, *n* (%)	2 Occurrences, *n* (%)	3 Occurrences, *n* (%)
Text Orientation Angle								
15°	90	12 (13.3%)	57 (63.3%)	21 (23.3%)	27 (30.0%)	20 (22.2%)	23 (25.6%)	20 (22.2%)
30°	90	14 (15.6%)	48 (53.3%)	28 (31.1%)	21 (23.3%)	21 (23.3%)	26 (28.9%)	22 (24.4%)
45°	90	13 (14.4%)	53 (58.9%)	24 (26.7%)	23 (25.6%)	24 (26.7%)	23 (25.6%)	20 (22.2%)
Background Type								
Abstract	90	11 (12.2%)	46 (51.1%)	33 (36.7%)	23 (25.6%)	23 (25.6%)	23 (25.6%)	21 (23.3%)
Natural	90	14 (15.6%)	53 (58.9%)	23 (25.6%)	25 (27.8%)	22 (24.4%)	26 (28.9%)	17 (18.9%)
Urban	90	14 (15.6%)	59 (65.6%)	17 (18.9%)	23 (25.6%)	20 (22.2%)	23 (25.6%)	24 (26.7%)
Overall	*270*	*39 (14.4%)*	*158 (58.5%)*	*73 (27.0%)*	*71 (26.3%)*	*65 (24.1%)*	*72 (26.7%)*	*62 (23.0%)*

In Table A2, rows for text orientation angle and background type are alternative marginal summaries of the same 270 unique stimulus configurations and should therefore not be summed. Pearson chi-square tests showed no statistically detectable association between target-letter zone and text orientation angle, $\chi^2(4) = 1.94$, $p = 0.747$, or background type, $\chi^2(4) = 7.44$, $p = 0.114$. Target-letter occurrence count was likewise not associated with text orientation angle, $\chi^2(6) = 1.57$, $p = 0.955$, or background type, $\chi^2(6) = 1.77$, $p = 0.939$.

In Table A3, image-level distributions are reported descriptively because each angle × image cell contained only ten unique stimuli.

Table A3. Distribution of target-letter characteristics across the 27 viewing-angle × individual-background-image cells (10 unique stimuli per cell).

Characteristic	Minimum	Q1	Median	Q3	Maximum
Lower-zone target letters	0	1	1	2	3
Mid-zone target letters	2	5	6	7	8
Upper-zone target letters	1	1	2	4	7
Target-letter occurrence count = 0	0	2	3	4	5
Target-letter occurrence count = 1	0	2	2	3	4
Target-letter occurrence count = 2	1	2	2	4	6
Target-letter occurrence count = 3	0	1	2	4	5
Distinct target letters	6	8	8	8	10

Some local variation in target-letter composition was present across the 27 angle × individual-background-image cells, which is expected given the limited number of ten textual stimuli per cell. Nevertheless, the typical cell contained eight distinct target letters, and the interquartile ranges indicate that the more extreme minimum and maximum values were not representative of most cells.

Appendix F. Simulation-Based Sensitivity Analysis

Because the study included only six participants in each of the three age groups, an additional simulation-based sensitivity analysis was conducted to quantify the magnitude of age-related interaction effects that the study design was reasonably capable of detecting.

The analysis focused on the age × angle and age × background-type interactions for both *accuracy* and *total_error*, corresponding to the mixed-ANOVA analyses reported in Sections 3.5.2 and 3.5.3.

The sensitivity analysis reproduced the participant-level structure used in the original mixed ANOVAs. For the age × angle analysis, each participant was represented by three measurements corresponding to the 15°, 30°, and 45° conditions, aggregated across background types. For the age × background-type analysis, each participant was represented by three measurements corresponding to abstract, natural, and urban backgrounds, aggregated across text orientation angles. The design therefore consisted of 18 participants, with six participants in each age group and three repeated measurements per participant.

For each outcome and interaction, the within-participant covariance structure was estimated from the observed participant-level data after accounting for age-group means. Simulated datasets were then generated using this estimated covariance structure while retaining the original sample size and repeated-measures design. Because a 3 × 3 interaction has four degrees of freedom and can take several different forms, four orthogonal interaction patterns, spanning the four-dimensional interaction space, were examined. The magnitude of each pattern was progressively increased, and 100,000 simulated datasets were evaluated for each scenario using the same mixed-ANOVA interaction test as in the original analysis, with $\alpha = 0.05$.

Interaction magnitude was expressed as the maximum difference-in-differences: the largest difference between any two age groups in the change in the outcome between any two levels of the within-participant factor. For example, if one age group improved by 15 percentage points between two text orientation angles and another improved by 5 percentage points, the corresponding difference-in-differences would be 10 percentage points. For each interaction pattern, the sensitivity threshold was defined as the effect magnitude at which the interaction was detected in approximately 80% of the simulated datasets.

The results are presented in Table A4.

Table A4. Observed age-related interaction magnitudes and simulated effect magnitudes associated with an approximately 80% detection probability.

Interaction	Outcome	Observed Maximum Difference-in-Differences	Approximate Magnitude Required for 80% Detection
Age × angle	*accuracy*	4.6 percentage points	10.2–15.0 percentage points
Age × angle	*total_error*	1.94 units	3.69–5.20 units
Age × background type	*accuracy*	4.8 percentage points	10.1–12.7 percentage points
Age × background type	*total_error*	1.61 units	3.54–4.55 units

The observed interaction magnitudes were substantially smaller than those associated with an approximately 80% probability of detection under the present design. These results indicate that the study was primarily sensitive to relatively large age-related interaction effects. The nonsignificant age × angle and age × background-type interactions are therefore compatible with the absence of large interactions, but they do not provide strong evidence that smaller or moderate age-related interactions are absent. The sensitivity estimates should be interpreted as approximate because the covariance structure used in the simulations was itself estimated from the present sample of 18 participants.

References

1. Milgram, P.; Kishino, F. A Taxonomy of Mixed Reality Visual Displays. *IEICE Trans. Inf. Syst.* **1994**, *77*, 1321–1329.
2. Dörner, R.; Broll, W.; Grimm, P.; Jung, B. (Eds.) *Virtual and Augmented Reality (VR/AR): Foundations and Methods of Extended Realities (XR)*; Springer: Cham, Switzerland, 2022.
3. Speicher, M.; Hall, B.D.; Nebeling, M. What is Mixed Reality? In Proceedings of the ACM CHI Conference on Human Factors in Computing Systems (CHI'19), Glasgow, UK, 4–9 May 2019; pp. 1–15. [CrossRef]

4. Richardson, J.T.E. *The Legibility of Serif and Sans Serif Typefaces Reading from Paper and Reading from Screens*; Springer International Publishing: Cham, Switzerland, 2022.
5. Mills, C.B.; Weldon, L.J. Reading text from computer screens. *ACM Comput. Surv. (CSUR)* **1987**, *19*, 329–357. [CrossRef]
6. Kilpeläinen, M.; Häkkinen, J. An Effective Method for Measuring Text Legibility in XR Devices Reveals Clear Differences Between Three Devices. *Front. Virtual Real.* **2023**, *4*, 1243387. [CrossRef]
7. Gabbard, J.L.; Swan, J.E.; Hix, D. The Effects of Text Drawing Styles, Background Textures, and Natural Lighting on Text Legibility in Outdoor Augmented Reality. *Presence Teleoperators Virtual Environ.* **2006**, *15*, 16–32. [CrossRef]
8. Manghisi, V.M.; Gattullo, M.; Fiorentino, M.; Uva, A.E.; Marino, F.; Bevilacqua, V.; Monno, G. Predicting Text Legibility over Textured Digital Backgrounds for a Monocular Optical See-Through Display. *Presence Teleoperators Virtual Environ.* **2017**, *26*, 1–15. [CrossRef]
9. Sawyer, B.D.; Wolfe, B.; Dobres, J.; Chahine, N.; Mehler, B.; Reimer, B. Glanceable, legible typography over complex backgrounds. *Ergonomics* **2020**, *63*, 864–883. [CrossRef] [PubMed]
10. Falk, J.; Eksvard, S.; Schenkman, B.; Andren, B.; Brunnstrom, K. Legibility and readability in Augmented Reality. In *Proceedings of the 2021 13th International Conference on Quality of Multimedia Experience, QoMEX 2021*; IEEE: New York, NY, USA, 2021; pp. 231–236. [CrossRef]
11. Cauz, M.; Clarinval, A.; Dumas, B. Text readability in augmented reality: A multivocal literature review. *Virtual Real.* **2024**, *28*, 59. [CrossRef]
12. Jović, A.; Stanimirović, I.Ž.; Mikota, M.; Mandić, L. Design recommendations and future prospects for text legibility in augmented reality. *JGED* **2026**, *17*, 19–34. [CrossRef]
13. Büttner, A.; Grunvogel, S.M.; Fuhrmann, A. The influence of text rotation, font and distance on legibility in VR. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*; IEEE: New York, NY, USA, 2020; pp. 662–663. [CrossRef]
14. De Luca, V.; Barba, M.C.; D'Errico, G.; Nuzzo, B.L.; De Paolis, L.T. A user experience analysis for a mobile Mixed Reality application for cultural heritage. *Virtual Real.* **2023**, *27*, 2821–2837. [CrossRef]
15. Hasanović, Z.; Jovanović, M.; Đurić, I. The Influence of Spatial Localization and Latency on the User Experience in Mixed Reality Applications for Cultural Historic Tourism. *Nexus Netw. J.* **2026**, *28*, 441–463. [CrossRef]
16. Quiñones, D.; Rojas, L.F.; Olavarría, R.; Cubillos, C.; Muñoz-La Rivera, F. A Methodology for Evaluating User Experience in Human-Centered Extended Reality Applications. *Biomimetics* **2026**, *11*, 182. [CrossRef] [PubMed]
17. Dobres, J.; Chahine, N.; Reimer, B. Effects of ambient illumination, contrast polarity, and letter size on text legibility under glance-like reading. *Appl. Ergon.* **2017**, *60*, 68–73. [CrossRef] [PubMed]
18. Sawyer, B.D.; Dobres, J.; Chahine, N.; Reimer, B. The Cost of Cool: Typographic Style Legibility in Reading at a Glance. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*; SAGE Publications: Los Angeles, CA, USA, 2017; Volume 61, pp. 833–837. [CrossRef]
19. Hart, S.G.; Staveland, L.E. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Human Mental Workload*; Hancock, P.A., Meshkati, N., Eds.; North-Holland: Amsterdam, The Netherlands, 1988; pp. 139–183. [CrossRef]
20. Jović, A.; Skorin-Kapov, L.; Mandić, L. Proposal of a glance-based method for evaluating 3D text legibility and estimating its display duration for mixed reality. *Zenodo* **2026**. [CrossRef]
21. de Fez, D.; Molina-Martín, A.; Luque, M.J.; Piñero, D.P. Color and Contrast Sensitivity Threshold Norms Across Age Decades in Normal Subjects. *Technologies* **2026**, *14*, 621. [CrossRef]
22. Gabbard, J.L.; Swan, J.E.; Hix, D.; Kim, S.-J.; Fitch, G. Active Text Drawing Styles for Outdoor Augmented Reality: A User-Based Study and Design Implications. In *Proceedings of the IEEE Virtual Reality Conference (VR 2007)*; IEEE: New York, NY, USA, 2007; pp. 35–42. [CrossRef]
23. Bolker, B.M.; Brooks, M.E.; Clark, C.J.; Geange, S.W.; Poulsen, J.R.; Stevens, M.H.H.; White, J.-S.S. Generalized linear mixed models: A practical guide for ecology and evolution. *Trends Ecol. Evol.* **2009**, *24*, 127–135. [CrossRef] [PubMed]
24. Wysopal, A.; Ross, V.; Passananti, J.; Yu, K.; Huynh, B.; Höllerer, T. Level-of-Detail AR: Dynamically Adjusting Augmented Reality Level of Detail Based on Visual Angle. In *2023 IEEE Conference Virtual Reality and 3D User Interfaces*; IEEE: New York, NY, USA, 2023; pp. 63–71. [CrossRef]
25. Seabold, S.; Perktold, J. Statsmodels: Econometric and statistical modeling with Python. In Proceedings of the 9th Python in Science Conference, Austin, TX, USA, 28 June–3 July 2010; pp. 92–96.
26. Blei, D.M.; Kucukelbir, A.; McAuliffe, J.D. Variational Inference: A Review for Statisticians. *J. Am. Stat. Assoc.* **2017**, *112*, 859–877. [CrossRef]
27. Bakdash, J.Z.; Marusich, L.R. Repeated Measures Correlation. *Front. Psychol.* **2017**, *8*, 456. [CrossRef] [PubMed]

28. Liang, K.-Y.; Zeger, S.L. Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika* **1986**, *73*, 13–22. [CrossRef]
29. Mancl, L.A.; DeRouen, T.A. A Covariance Estimator for GEE with Improved Small-Sample Properties. *Biometrics* **2001**, *57*, 126–134. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.