

Effects of textual explanations and spatially augmented visual cues on spatial visualization and mental rotation performance in desktop VR

Qian Chen^{a,*,} Yun Zhou^{a,*}, HanBing Kou^a, Hui Zhang^a, Tao Xu^{b,*}

^a Faculty of Education, Shaanxi Normal University, South Chang'an Road 199, Yanta District, Xi'an, Shaanxi, 710062, China

^b School of Software, Northwestern Polytechnical University, 127 West Youyi Road, Beilin District, Xi'an, Shaanxi, 710072, China

ARTICLE INFO

Keywords:

Desktop virtual reality (desktop VR)
Spatial ability training
Textual explanations
Visual cues
Eye-tracking

ABSTRACT

Spatial ability is fundamental to STEM learning, as many STEM tasks require the mental representation and transformation of 2D and 3D information. Desktop virtual reality (desktop VR) has therefore been increasingly adopted for spatial training due to its interactivity and 3D representations. Textual and visual cues are widely used as instructional supports in multimedia learning, yet their distinct and combined effects on spatial training in desktop VR remain insufficiently understood. This study conducted a 2×2 between-subjects experiment manipulating the presence of textual explanations and spatially augmented visual cues in a desktop VR environment to examine spatial visualization and mental rotation performance. Spatial performance, learning experience, cognitive load, and visual attention allocation were measured using tests, subjective scales, and eye-tracking data. The results showed no significant main effect of textual explanations or spatially augmented visual cues on posttest performance after controlling for pretest scores, but a significant interaction between the two. The textual-explanation-only condition outperformed the combined textual-explanation-and-visual-cue condition, suggesting that the effect of textual explanations depended on whether visual cues were also present. Conditions including spatially augmented visual cues showed comparable training-phase performance to the other conditions; however, this was not accompanied by better immediate posttest performance when the cues were unavailable. Eye-tracking results further revealed that the combined cue/interactive areas received more fixations than the text areas, but this attention pattern did not translate into better posttest performance. These findings inform the application of the signaling principle within desktop VR-supported spatial training and suggest that instructional supports need to be carefully coordinated rather than simply combined.

1. Introduction

Spatial ability is a crucial component of human cognitive capabilities (Uttal & Cohen, 2012), involving multiple cognitive processes that support learners' reasoning and performance across science, technology, engineering, and mathematics (STEM) disciplines (Stieff & Uttal, 2015; Wai et al., 2009). Solving problems in these disciplines often requires spatial reasoning, such as mentally manipulating

* Corresponding authors.

E-mail addresses: qian.chen@snnu.edu.cn (Q. Chen), zhouyun@snnu.edu.cn (Y. Zhou), kouhanbing@snnu.edu.cn (H. Kou), z0915h@snnu.edu.cn (H. Zhang), xutao@nwpu.edu.cn (T. Xu).

<https://doi.org/10.1016/j.tsc.2026.102402>

Received 23 December 2025; Received in revised form 13 July 2026; Accepted 18 October 2026

Available online 19 October 2026

1871-1871/© 2026 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

chemical structures (Rahmawati et al., 2021), interpreting biological diagrams (Castro-Alonso & Uttal, 2019), or understanding anatomical systems (Langlois et al., 2020). Spatial ability is not a unitary construct; instead, it comprises several distinct sub-components. Among these, spatial visualization and mental rotation are two particularly important sub-components that contribute to solving spatial problems (Purwanto et al., 2024). Spatial visualization supports the multi-step transformation of complex three-dimensional (3D) forms, and mental rotation enables the manipulation of object orientations in one's mind (Zhou et al., 2022).

Recent studies have increasingly demonstrated the potential of virtual technologies for training spatial ability (Di & Zheng, 2022). Desktop VR, as a screen-based 3D interactive environment, offers interactive spatial representations that enable learners to manipulate and inspect spatial objects in ways that are difficult to achieve with traditional two-dimensional (2D) materials, thereby offering a useful context for spatial ability training (Chang et al., 2018). However, from a cognitive load perspective, the 3D visual presentation and high level of interactivity in virtual environments may also impose additional information-processing demands on learners, potentially distracting attention (Mayer & Moreno, 1998; Mayer, 2024) and increasing cognitive load (Parong & Mayer, 2021). These challenges are particularly pronounced for individuals with lower spatial ability, who may struggle to identify relevant features and manage the complexity of 3D spatial tasks (Guo et al., 2025; Sun et al., 2020). Such difficulty may also elicit spatial anxiety, which could further constrain learners' available working-memory resources and processing capacity, thereby negatively affecting spatial task performance (Nori et al., 2023; Ramirez et al., 2012).

Prior work in multimedia learning demonstrates that textual explanations and visual cues can direct attention and improve learning outcomes (Imhof et al., 2013; Marwan et al., 2019). According to the signaling principle (Schneider et al., 2018; Van Gog, 2014), these instructional cues may support learning by directing learners' attention to task-relevant information and reducing unnecessary search during the processing of spatial relations. In spatial visualization and mental rotation tasks, such cueing may be particularly relevant because learners must identify task-relevant spatial relations, coordinate 2D and 3D representations, and mentally transform object orientations. Relevant research suggests that cueing can highlight essential visuospatial elements and relationships to reduce unnecessary processing (Castro-Alonso et al., 2019), and that guiding learners where to look can improve spatial problem solving among individuals with low mental rotation ability (Roach et al., 2019). Yet, most existing cueing research has examined 2D multimedia materials, such as text-picture materials (Ozcelik et al., 2009; Scheiter & Eitel, 2015), static diagrams (Jamet, 2014), animations (Boucheix et al., 2013; De Koning et al., 2010), and instructional videos (Li et al., 2023b), with far fewer studies exploring spatially augmented visual cues embedded directly within desktop VR or other 3D interactive environments (Chen et al., 2022; Stefanucci et al., 2022; Wen et al., 2024). These cues differ from traditional 2D cues in that they are volumetric, spatially anchored, or dynamically integrated into the learner's perspective. As a result, they may offer affordances for improving spatial abilities. Despite this potential, little is known about how such cues function in desktop VR, especially when combined with textual explanations. In the present study, we use the term spatially augmented visual cues to refer to visual cues embedded in the desktop VR environment and spatially aligned with task-relevant 3D information.

To address the aforementioned gap, this study experimentally investigates how textual explanations and spatially augmented visual cues influence spatial visualization and mental rotation performance within desktop VR, and further examines the interaction between these two forms of instructional support. Theoretically, this study aims to identify potential boundary conditions of signaling effects in desktop VR-supported spatial training by examining whether the effects of textual explanations and spatially augmented visual cues are additive or depend on their coordination within an interactive 3D environment.

2. Literature review and research questions

2.1. VR-supported training of spatial abilities and spatial task performance

Spatial ability refers to the capacity to mentally represent and transform 2D or 3D information (Lu & Carr, 2014). Classic taxonomies categorize spatial ability into spatial perception, mental rotation, and spatial visualization (Linn & Petersen, 1985). Spatial perception was defined as the ability to determine spatial relationships with respect to the orientation of one's own body, despite distracting information (Linn & Petersen, 1985), and is broadly understood as involving spatial judgment (Xiao & Zhang, 2021). Mental rotation refers to mentally rotating objects and identifying their 3D features (Shepard & Metzler, 1971). Spatial visualization is the skill to perceive and interpret complex spatial patterns, as well as comprehend imagined spatial movements (Mix & Cheng, 2012). These sub-components are commonly assessed through different task formats. Spatial visualization is often measured by paper-folding, cube-comparison, and 2D-3D transformation tasks, which require learners to mentally restructure complex configurations or infer how flat patterns form three-dimensional objects (Ekstrom & Harman, 1976; Hegarty, 2010; Martín-Gutiérrez et al., 2010). Mental rotation is typically measured by object- or card-rotation tasks, the Vandenberg Mental Rotations Test, and the Purdue Visualization of Rotations Test, in which learners judge whether objects or configurations match after rotation (Bodner & Guay, 1997; Shepard & Metzler, 1971; Vandenberg & Kuse, 1978). However, complex spatial tasks often involve the coordinated use of spatial visualization and mental rotation, especially when learners need to mentally transform configurations, construct or inspect 3D representations, and judge object orientations or spatial correspondences (Lin & Chen, 2016; Supli & Yan, 2024; Zhou et al., 2022).

Virtual technologies have emerged as effective media for enhancing spatial ability since they can provide interactive spatial representations and simulated 3D environments (Chang et al., 2018). VR systems can be classified into immersive VR (IVR) and non-immersive VR. IVR provides a high-fidelity simulated environment through head-mounted displays and motion-tracking technology (Mikropoulos & Natsis, 2011). Another form of VR, often referred to as desktop VR, uses screens, keyboards, mice, joysticks, and touchscreens to provide rich interaction and 3D experiences (Dodd & Antonenko, 2012; Makransky & Petersen, 2019). Prior work has described 3D virtual worlds as desktop interactive VR environments that provide an illusion of 3D space and support interaction

within the environment (Dickey, 2005). In this study, we use the term desktop VR in this narrower sense, referring to a non-immersive, screen-based 3D interactive environment rather than an HMD-based IVR system. Although less immersive, desktop VR has been shown to be highly effective for learning and, in some cases, can even outperform IVR in promoting learning (Lee & Wong, 2014; Makransky et al., 2019).

A growing body of research has confirmed the value of desktop VR for training spatial abilities and improving spatial task performance. For example, emerging research in geography education demonstrates that VR-based metaverse environments can significantly improve learners' spatial thinking skills, showing medium-sized training effects (Purwanto et al., 2024). Zhou et al. (2022) proposed a desktop VR spatial training system and found that learners who received the training showed significant improvements in spatial task performance compared to those who did not receive the training. Hauptman (2010) developed a desktop 3D virtual environment that trains users to construct and manipulate spatial images, and reported that this training effectively enhanced students' spatial thinking. Taken together, these findings indicate that desktop VR is a promising tool for developing multiple sub-components of spatial ability. Building on this foundation, the present study adopts desktop VR as the platform for delivering spatial training tasks. In this environment, learners manipulate 3D objects, coordinate 2D planar figures with 3D models, and may use textual explanations and spatially augmented visual cues during spatial training.

2.2. Effects of on-screen text in multimedia learning

In multimedia learning environments, on-screen text refers to text presented on the screen such as subtitles, annotations, labels, narration, and explanatory statements. They serve functions including complementing auditory information (Hyunjeong & Mayer, 2018), explaining key points (Stark et al., 2018), and providing prompts (Albus et al., 2021).

Although the redundancy principle suggests that presenting on-screen text may increase learners' cognitive load, in some cases, it may facilitate learning (Mayer & Fiorella, 2021). For example, Li et al. (2023) found that the textual cues and summary scaffolds in IVR improved both learning performance and mental model development. Ardaç and Unal (2008) examined how the quantity of on-screen text would affect students' learning of basic concepts in coordinate geometry within multimedia instructional units, and the results indicated that presenting the entire text was more effective for students with strong memory abilities. Recent research further shows that structured textual cues in educational animations can reduce cognitive load and enhance both learning outcomes and self-efficacy among elementary students (Du et al., 2025). However, not all text-based augmentations are effective. Turčani et al. (2024) found that keyword highlighting has limited learning benefits.

Other studies have explored the form and modality of text presentation. Yue et al. (2013) investigated the differential effects of on-screen text and narration on learning and reported that subtle differences between on-screen text and narration were advantageous for learning. Hsieh (2020) found that subtitles presented in instructional videos could enhance learners' vocabulary learning and listening comprehension abilities. Building on this line of inquiry, a growing body of research has shifted toward examining the interaction between on-screen text and other instructional factors. Zhou et al. (2024) found that when learning tasks rely on mental imagery, high visual detail combined with low text load optimizes learning, whereas high text load and low visual detail independently increase cognitive load, indicating an interaction between visual richness and textual load in 3D animation-based learning.

The effects of on-screen textual explanations appear to depend on their function in the learning task. They may facilitate learning when they clarify key concepts, summarize essential information, or connect verbal explanations with visual representations, but their benefits may be limited when they are redundant, cognitively demanding, or poorly aligned with task demands. These boundary conditions may help explain the mixed findings on-screen text. However, most prior research has focused on conceptual understanding, vocabulary acquisition, and cognitive processing in traditional multimedia learning contexts, leaving its role in spatial skill training underexplored. Given the heightened visual complexity, spatial representations, and interaction demands of desktop VR, it remains unclear whether conclusions from conventional multimedia learning can be generalized to desktop VR-supported spatial ability training. This gap highlights the need to examine whether and under what conditions textual explanations may function as scaffolding for spatial skill training in desktop VR.

2.3. The impact of visual cues on learning

In multimedia learning, cues can be broadly categorized into textual cues and visual cues. Visual cues refer to graphical or picture-based signals and may take forms such as arrows, color changes, flashing, shading, zooming, and color coding. According to the signaling principle in multimedia learning (Van Gog, 2014), adding cues into multimedia learning materials can help guide learners' attention to relevant elements of the material and support the initial selection of information for further processing. Empirical studies have shown that visual cues in multimedia learning can support learning by guiding attention to relevant information, although their effects vary across studies and outcome measures. For instance, Lin and Atkinson (2011) found that adding arrow cues to animations improved learning. Boucheix and Lowe (2010) used eye-tracking and revealed that visual cues helped focus attention on relevant areas, leading to higher test performance. However, cueing effects are not always reflected in all learning outcomes. Although Ragazou and Karasavvidis (2023) reported that visual cues in instructional videos did not significantly improve task performance, they did enhance learners' motivation. Additionally, research has identified an expertise reversal effect in multimedia learning, where visual cues tend to benefit learners with low prior knowledge (Arslan-Ari, 2018; Richter & Scheiter, 2019), but may increase cognitive load for more advanced learners.

Although early research on visual cues has primarily focused on 2D video- or image-based learning, there is growing interest in their application within VR, AR, XR, and other spatially rich learning environments. For example, Nelson et al. (2014) found that visual

cues in a virtual science inquiry environment reduced students' reported cognitive load. Han et al. (2023) demonstrated that cues in IVR effectively directed the attention of low-prior-knowledge learners and supported their selection of relevant information. In XR collaborative systems, cues enhance workspace awareness but face challenges related to cognitive load, personalization, and device compatibility (Assaf et al., 2024). Comparative studies also suggest that cue effectiveness hinges on spatial clarity, task alignment, and interpretability. For instance, Gaze Lines outperform other cue types in AR visual search tasks (Kelley et al., 2024), and in situated analytics, color cues remain effective, while combining multiple cue types may offer complementary benefits under certain conditions (Doerr et al., 2024). Research on spatial decision-making reveals that visual signals can bias spatial choices toward cued locations, indicating that visual cues may influence how attention is allocated during spatial decisions (Danziger & Rafal, 2009). Antrilli and Wang (2016) found that richer visual cues generated during object rotation facilitated mental rotation, suggesting that visual cues can support the integration of visual and motor information during spatial transformation.

Drawing on both the signaling principle (Van Gog, 2014) and empirical findings, visual cues should not be understood as automatically beneficial. Although cues may support learning by helping learners attend to relevant information and select, organize, or integrate essential material, empirical studies suggest that these benefits may be limited when cues are redundant, difficult to interpret, or unnecessary for learners who can already locate relevant information. Thus, attention guidance alone may not guarantee improved learning, which may partly explain the mixed findings on visual cues. Compared to video- or image-based media, desktop VR and other 3D interactive environments may introduce additional considerations for the use of visual cues. As visual information shifts from fixed 2D frames to interactive 3D objects and spatial layouts, learners may need to process more complex spatial relationships and coordinate multiple sources of information. However, it remains unclear how spatially augmented cues function in desktop VR and how they interact with textual scaffolds, highlighting the need for further empirical investigation.

2.4. Eye-tracking measures in spatial cognition research

Eye-tracking provides a direct and objective method for examining learners' visual-cognitive processes (Zhao et al., 2018). Based on the eye-mind assumption, gaze behavior can be interpreted as indirect evidence of learners' attention allocation and information processing during the learning task (Poole & Ball, 2006). This is also consistent with the Cognitive Theory of Multimedia Learning (CTML), which emphasizes the selection, organization, and integration of information and suggests that instructional cues may facilitate these processes by guiding attention and highlighting relations among elements (Wang et al., 2020). Research comparing learners with different spatial abilities shows that patterns can be detected through their eye-movement trajectories (Sun et al., 2020). Imagery research demonstrates that eye movements reflect spatial encoding processes (Mast & Kosslyn, 2002). Recent eye-tracking studies on spatial tasks suggest that fixation- and AOI-based measures can reveal learners' visual information processing in spatial cognition (Keskin et al., 2018), and fixation, saccade, and switching measures can be used to examine information processing and search efficiency across spatial tasks such as mental rotation and spatial visualization (Tian et al., 2022). Hsieh (2026) showed that eye-tracking indicators can reveal strategy-related differences in relational spatial reasoning, supporting the use of eye-tracking to examine visual search and attentional control in spatial tasks. Thus, eye-tracking is well suited for evaluating how textual explanations and spatially augmented visual cues influence the training of spatial visualization and mental rotation.

2.5. Conceptual framework, research questions and hypotheses

To address the aforementioned gap, the present study investigates the effects of textual explanations and spatially augmented visual cues on spatial visualization and mental rotation performance in desktop VR environments. This section serves as the conceptual framework linking desktop VR as the spatially interactive learning environment, the signaling principle and cognitive load theory as the theoretical bases, textual explanations and spatially augmented visual cues as instructional supports for spatial cognition, and spatial performance, cognitive load, and attention allocation as the educational outcomes of interest. In this study, desktop VR is the spatially interactive learning environment rather than an independent variable. It provides the context in which learners manipulate 3D objects and coordinate 2D planar figures with 3D models. Guided primarily by the signaling principle (Schneider et al., 2018; Van Gog, 2014) and cognitive load theory (Sweller et al., 1998), and informed by prior empirical findings, we propose several hypotheses while retaining exploratory research questions for effects that remain underexplored in desktop VR-supported spatial training. The signaling principle suggests that instructional cues guide learners' attention to essential information (Van Gog, 2014), and such signals may reduce unnecessary cognitive effort and thereby lower extraneous load (Van Gog, 2014). Accordingly, textual explanations are expected to make task-relevant spatial relations more explicit and support coherent spatial representations. Prior studies have shown that on-screen text can complement auditory information, explain key points, provide prompts, and improve learning performance or mental model development (Albus et al., 2021; Du et al., 2025; Hyunjeong & Mayer, 2018; Li et al., 2023a; Stark et al., 2018). Thus, we hypothesize that learners receiving textual explanations will show better spatial performance than those not receiving textual explanations (H1).

Prior studies also suggest that instructional supports may improve learners' motivational or affective experience by making learning materials more understandable or by increasing engagement (Du et al., 2025; Ragazou & Karasavvidis, 2023). Thus, we hypothesize that learners receiving textual explanations or spatially augmented visual cues will report more positive learning experiences than those receiving no instructional support (H2).

According to cognitive load theory, cognitive load includes intrinsic, extraneous, and germane load (Sweller et al., 1998). In this study, coordinating 2D planar figures with 3D models and mentally transforming spatial relations constituted a major source of intrinsic load. Therefore, textual explanations and spatially augmented visual cues were expected to influence cognitive load mainly by

affecting extraneous processing rather than by changing the intrinsic complexity of the tasks. Since textual and visual cues may reduce unnecessary search while also adding information to be processed, we cautiously hypothesized that these supports would not increase learners' overall cognitive load during desktop VR-supported spatial training (H3).

Visual cues and interactive manipulation are expected to guide attention to task-relevant visual information, as supported by the signaling principle and eye-tracking studies showing that cues help learners focus on relevant areas and select task-relevant information (Boucheix & Lowe, 2010; Han et al., 2023). Thus, we predict that learners will allocate more visual attention to the combined cue and interactive areas than to the text areas during desktop VR-supported spatial training (H4).

Finally, the split-attention principle (Ayres & Sweller, 2005) suggests that when learners need to coordinate multiple sources of information that are physically or functionally separated, additional search and integration demands may increase extraneous processing. In desktop VR-supported spatial training, learners may need to coordinate textual explanations, spatially augmented visual cues, 3D models, and interactive manipulation simultaneously. Therefore, when textual explanations and spatially augmented visual cues are presented together, they may not necessarily produce additive benefits; instead, their effectiveness may depend on how well these sources of information are coordinated. Given the limited and mixed evidence on how textual explanations and spatially augmented visual cues jointly function in desktop VR, their interaction effect was examined exploratorily.

Based on these considerations, the study focused on the following research questions and hypotheses.

RQ1. How do textual explanations (absent vs. present) and spatially augmented visual cues (absent vs. present) affect spatial performance, including spatial visualization and mental rotation, in a desktop VR-supported 3D interactive learning environment?

H1. Learners receiving textual explanations will show better spatial performance than those not receiving textual explanations.

Since evidence remains mixed, the effect of spatially augmented visual cues on spatial performance was examined exploratorily.

RQ2. How do textual explanations and spatially augmented visual cues affect learning experience, including learning motivation, satisfaction, perceived usefulness, and perceived ease of use, in a desktop VR-supported 3D interactive learning environment?

H2. Learners receiving textual explanations (H2a) or spatially augmented visual cues (H2b) will report more positive learning experiences than those receiving no instructional support.

RQ3. How do textual explanations and spatially augmented visual cues affect overall cognitive load?

H3. Textual explanations (H3a) and spatially augmented visual cues (H3b) will not increase learners' overall cognitive load during desktop VR-supported spatial training.

RQ4. How do learners allocate visual attention between text areas and combined cue/interactive areas during desktop VR-supported spatial training?

H4. Learners will allocate more visual attention to the combined cue and interactive areas than to the text areas during desktop VR-

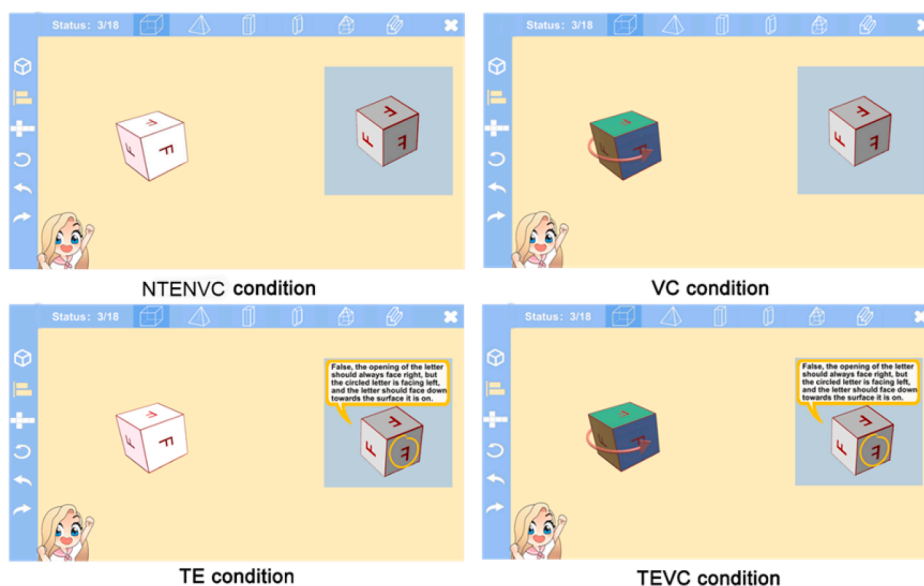


Fig. 1. The interface of four conditions (translated).

supported spatial training.

Given limited evidence, the effect of textual explanations on attention allocation was examined exploratorily.

RQ5. Do textual explanations and spatially augmented visual cues interact in their effects on spatial performance, learning experience, and overall cognitive load?

Given the limited and mixed evidence on how textual explanations and spatially augmented visual cues jointly function in desktop VR, their interaction effect was examined as an exploratory research question.

3. Methods

3.1. Participants and design

Participants were recruited through self-designed printed posters displayed on campus and recruitment announcements posted in university online social groups. A total of 100 college students (40 males and 60 females) from various majors at universities in northwestern China participated in the study. Their ages ranged from 16 to 26 years ($M = 19.62$, $SD = 2.27$).

This study employed a 2 (textual explanations: absent vs. present) $\times$ 2 (spatially augmented visual cues: absent vs. present) between-subjects design, including four conditions: no textual explanations and spatially augmented visual cues (NTENVC), spatially augmented visual cues (VC), textual explanations (TE), and textual explanations and spatially augmented visual cues (TEVC) (see Fig. 1). Textual explanations provided brief problem-solving strategies, whereas spatially augmented visual cues consisted of directional arrows and color highlights. The arrows guided rotation toward the reference figure, while color highlights helped learners identify the lettered target faces (see Fig. 1). In addition, all other instructional elements were kept consistent across groups, including the audio narration, training procedure, task difficulty, and interactive functions.

Based on Kang (2021) and an assumed medium-to-large effect size ($f = 0.40$), a power analysis using G*Power 3.1 (Faul et al., 2007) indicated a required sample size of 80. The final number of participants met this requirement. Participants were randomly assigned to one of four experimental groups, with no significant differences in age or gender distribution among the groups (see Table 1). Prior to the formal experiment, all participants provided informed consent and received compensation upon completion.

3.2. Apparatus and materials

3.2.1. Desktop VR and eye-tracking apparatus

The learning environment was implemented as a desktop VR environment rather than an immersive head-mounted display (HMD)-based VR system. Participants viewed and manipulated 3D objects on a laptop screen using a standard desktop interface. A laptop computer was used, with a screen size of 30.99 cm by 17.43 cm and a resolution of 1920×1080 . Eye movements were recorded using a Tobii Pro Nano eye tracker (sampling rate: 60 Hz), which was mounted at the bottom center of the screen using adhesive strips and connected to the laptop via USB.

3.2.2. Training materials

The learning materials and system were developed using Unity 3D and comprised a total of 18 training tasks. In this study, “training” refers to a short-term, task-based VR practice session rather than a long-term instructional curriculum. The task design was informed by an established spatial task framework targeting integrated spatial visualization and mental rotation processes (Zhou et al., 2022). Learners were required to determine whether a planar unfolded figure could be folded and rotated to match a corresponding 3D model. Thus, the judgment-based folding tasks were used to operationalize integrated spatial task performance involving spatial visualization and mental rotation, rather than to produce separate scores for the two constructs. Learners needed to mentally construct a 3D representation from a 2D unfolded figure and inspect the orientation and spatial correspondence of surfaces after folding. Each task involved 3D models containing one, two, or three instances of the letter F. Six distinct types of incorrect folding options, in which the planar figures could not be correctly folded into the target models, were intentionally designed as distractors for the judgment task and are summarized in Table 2. The training tasks, sequence, and difficulty were kept consistent across all groups. Each task comprised two interfaces: a test interface and an interactive interface (see Fig. 2).

Table 1
Demographic information.

	NTENVC (N = 25) Mean (SD)	VC (N = 25) Mean (SD)	TE (N = 25) Mean (SD)	TEVC (N = 25) Mean (SD)	statistics
Gender	0.40 (0.50)	0.40 (0.50)	0.40 (0.50)	0.40 (0.50)	$\chi^2 = 0.00$, $df = 3$, $p = 1.000$
Age	19.40 (1.98)	19.60 (2.66)	19.48 (2.18)	20.00 (2.31)	$H = 2.65$, $df = 3$, $p = 0.450$

Note: Gender: 1, Male; 0, Female.

Table 2
Six types of incorrect options used as distractors.

Type	Illustration
Incorrect orientation of the letter opening	
Incorrect facing direction of the letter	
Incorrect placement of the letter	
Incorrect opening direction of the letter relative to the target edge	
Incorrect opening direction and placement of the letter	
Incorrect facing direction and placement of the letters	

In the test interface, learners were required to complete a task to decide whether the planar unfolded figure could be folded into the corresponding 3D shape. After submitting their response, learners proceeded to the interactive interface (see Fig. 2a), where they first listened to an approximately 10-second explanatory audio. Upon completion of the audio, learners engaged in interactive practice. The explanations included the correct answer, the rationale for the judgment, and problem-solving guidance, such as how to examine adjacent relations, opposite relations, and the spatial correspondences formed after folding.

The interactive interface was divided into three areas: a progress display area, an interactive manipulation area, and a reference area (see Fig. 2b). Learners’ task progress was continuously displayed at the top. The interactive area provided six operation buttons that allowed learners to fold, unfold, and rotate the model. In addition to using these buttons, learners could also rotate the model by right-clicking and dragging it. Through this interactive practice, learners visually explored the reasons why the planar figures could or could not be folded into the corresponding 3D models, which were presented in the reference area.

3.3. Measures

3.3.1. Spatial performance test

Spatial performance was assessed using a judgment-based test consisting of 36 items, with each item scored as one point. Each item presented a 2D planar figure together with a corresponding 3D model, and participants were required to judge whether the planar figure could be folded and rotated to form the target 3D structure. The test was adapted from the FORSpatial task framework, which assesses spatial visualization and mental rotation using shape-letter combinations and has shown acceptable validity and reliability ($r = 0.66$, $p < 0.05$; Cronbach’s $\alpha = 0.841$; Zhou et al., 2022).

The spatial performance test was administered in two phases. First, participants completed the test one day prior to the formal experiment as a pretest to assess their initial spatial performance. Second, after completing the desktop VR training tasks, participants completed the same test again as a posttest to evaluate changes in spatial performance following the intervention. The same set of items was used in both phases, with the order of the questions randomized to minimize potential order effects.

The pretest scores were analyzed using a one-way analysis of variance (ANOVA), which revealed no significant differences among groups in initial spatial performance, $F(3, 96) = 0.376$, $p = 0.771$, $\eta^2 = 0.012$, indicating that participants’ spatial performance was comparable across experimental conditions before the intervention. The test showed acceptable internal consistency (Cronbach’s $\alpha = 0.774$) and preliminary construct-related validity evidence ($KMO = 0.708$; Bartlett’s test, $p < 0.001$). Test items and their difficulty and discrimination indices are reported in Appendix A.

3.3.2. Learning experience

Learning experience was measured through four dimensions: satisfaction, motivation, perceived ease of use, and perceived usefulness. The questionnaire was adapted from established scales developed by Chou and Liu (2005), Lee et al. (2010) and Hwang et al. (2013) (see Table 3). All items were rated on a 7-point Likert scale, ranging from 1 (“Completely disagree”) to 7 (“Completely agree”). The overall internal consistency of the questionnaire was high, with a Cronbach’s α of 0.887.

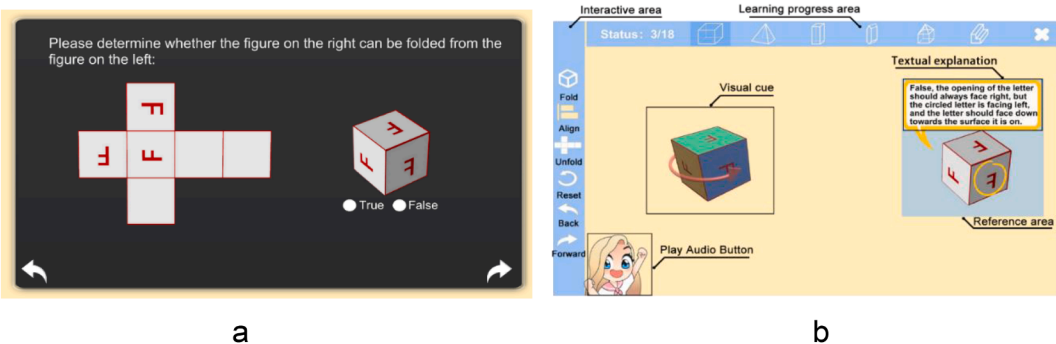


Fig. 2. Screenshots of the desktop VR system interface. Left: test interface (a); right: interactive interface (b).

Table 3

General information on the questionnaires.

Category	Dimension	Number of items	Score	Source
Basic Information		7		Self-developed scale
Spatial performance		36	36	Self-developed scale
Interactive test		18	18	Self-developed scale
Learning experience	Learning satisfaction	3	1–7	Chou and Liu (2005)
	Learning motivation	4	1–7	Lee et al. (2010)
	Perceived usefulness	3	1–7	Hwang et al. (2013)
	Perceived ease of use	3	1–7	Hwang et al. (2013)

3.3.3. Interactive test

The interactive test was process-oriented, as it captured learners' in-system practice during VR training rather than final learning outcomes. It consisted of 18 items covering six shapes at three difficulty levels, with each item worth one point for a total score of 18. Participants' responses and scores were automatically recorded by the system and constituted the interactive performance data used for subsequent analyses. The test was administered through the test interface embedded within the training system, where learners completed multiple-choice questions requiring them to judge whether an unfolded planar figure could be folded into the corresponding 3D model. The interactive test data were used to reflect learners' performance during the learning process.

3.3.4. Eye-tracking

Fig. 3 shows the laptop and eye-tracking device used during the experiment. Participants viewed the desktop VR learning environment on a laptop screen and manipulated the 3D objects using a mouse, while their eye movements were recorded using a Tobii Pro Nano eye tracker mounted at the bottom center of the screen.

Eye-tracking measures were employed to examine cognitive load and attention allocation across the four experimental groups. The average fixation duration was used as an indirect eye-tracking proxy for overall cognitive load, as fixation-based measures have been used to infer processing demands and cognitive activity during learning (Clinton et al., 2017; Korbach et al., 2018; Van Gog & Scheiter, 2010). However, this measure was not treated as a comprehensive assessment of cognitive load. The number of fixations was used to analyze patterns of attention allocation within predefined areas of interest (AOIs), reflecting learners' selection and processing of instructional information (Poole & Ball, 2006; Ponce & Mayer, 2014). Two AOIs were defined within the interactive interface: a text area and a combined cue/interactive area (see Fig. 4). The combined cue/interactive area included the 3D manipulation space and, when present, the spatially augmented visual cues embedded in that space. Thus, fixations in this AOI were interpreted as attention to task-relevant interactive visual information rather than attention to visual cues alone.

3.4. Procedure

The experimental procedure consisted of four stages and lasted approximately 60 min in total. The first stage was completed one day prior to the formal experiment, whereas the remaining three stages were conducted consecutively in a quiet laboratory. First, participants completed a pretest questionnaire, which took approximately 15 min. Second, at the beginning of the formal experiment, participants read and signed an informed consent form. During this stage, the experimenter explained the experimental procedure and

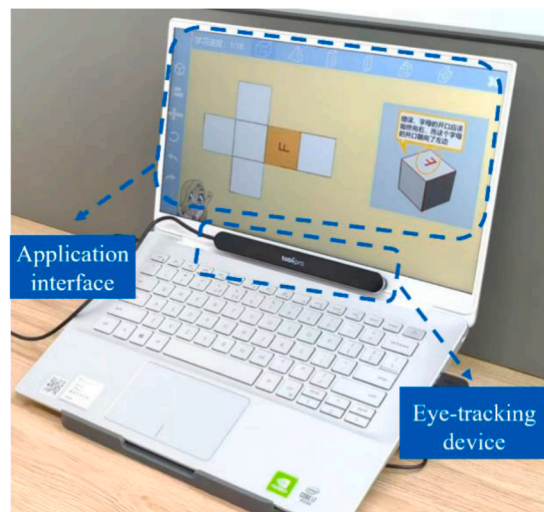


Fig. 3. Experimental setup for the desktop VR spatial training task and eye-tracking data collection.

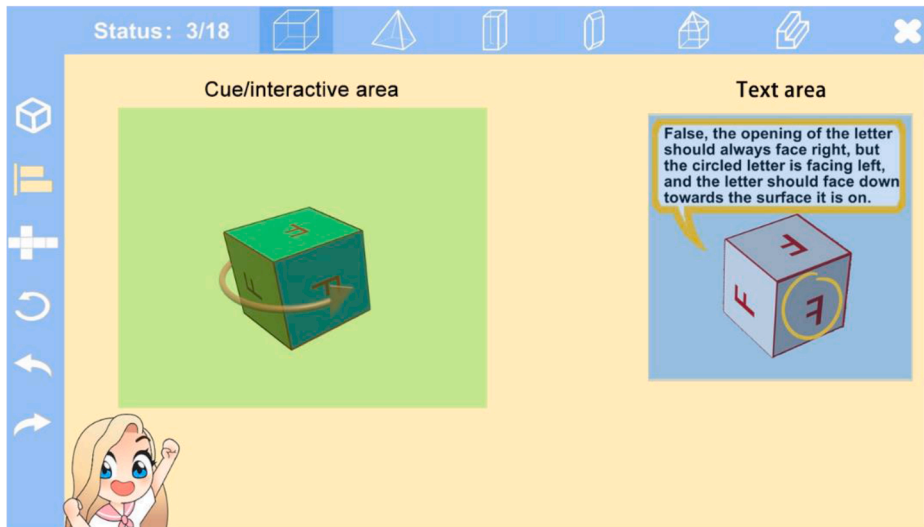


Fig. 4. The screenshot of Areas of Interest (AOIs).

requirements and addressed any questions raised by the participants. To ensure that participants became familiar with the training system and to avoid potential bias caused by operational unfamiliarity during the formal experiment, participants then watched a 2-minute and 23-second instructional video demonstrating system operation. Under the guidance of the experimenter, participants subsequently adjusted their posture and completed calibration of the eye-tracking equipment. This preparation stage took approximately 5 min. Next, after successful calibration, the formal training session began and lasted approximately 25 min. For each item, participants first completed the corresponding spatial task in the test interface, and then used the interactive interface for learning and practice. This procedure was repeated until all training tasks were completed. Finally, after completing the training tasks, participants filled out the posttest and subjective experience questionnaires. An overview of the experimental procedure is presented in Fig. 5.

3.5. Data analysis

Data were analyzed using SPSS. Normality was examined using the Shapiro-Wilk test, together with histograms, Q-Q plots, and box plots. Homogeneity of variance was assessed using Levene's test. Parametric tests were used when the relevant assumptions were met; otherwise, non-parametric tests were applied. For spatial performance, the assumptions of normality, homogeneity of variance, linearity between the covariate and the dependent variable, and homogeneity of regression slopes were met, so parametric tests were used. A paired-samples *t*-test was conducted to examine pretest-posttest improvement, and ANCOVA compared adjusted posttest scores across the four conditions with pretest scores as a covariate. The significant ANCOVA effect was followed by Bonferroni-adjusted EMM comparisons (Dunn, 1961). Two-way ANCOVA was then used to test the main and interaction effects, with significant interactions followed by simple main effects analyses (Field, 2012). Effect sizes were reported as partial eta squared (η^2) and interpreted according to Cohen's (Cohen et al., 1988, pp. 280–288) guidelines, with small, medium, and large thresholds set at 0.0099, 0.0588, and 0.1379, respectively. For the interactive test, normality and homogeneity of variance assumptions were also met; therefore, both one-way and two-way ANOVAs were used. In addition, non-parametric tests were used for learning experience and eye-tracking data as normality assumptions were not met. Kruskal-Wallis tests examined between-group differences, while Wilcoxon signed-rank tests compared the number of fixations in AOIs within conditions. These analyses should be interpreted as exploratory group-level

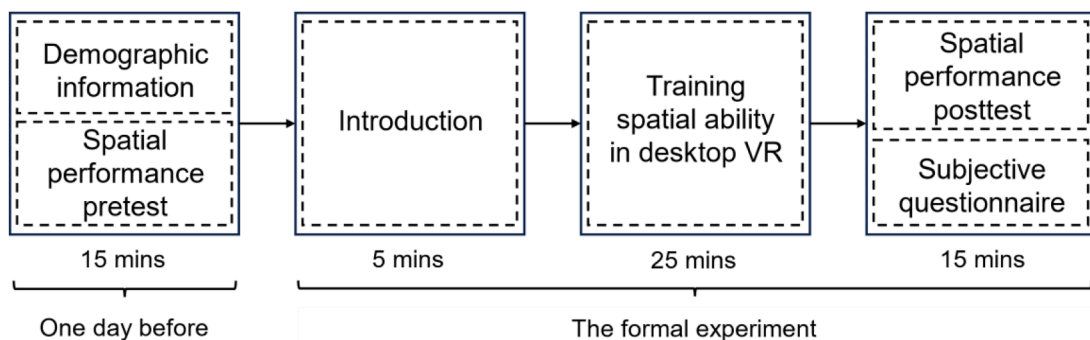


Fig. 5. Experimental procedure.

comparisons rather than formal tests of the 2×2 factorial design.

4. Results

Descriptive statistics across the four experimental conditions are presented in Table 4. The results of hypothesis testing are summarized in Appendix B.

4.1. Spatial performance

Table 4 shows the pretest and posttest descriptive statistics for spatial performance across the four groups. A paired-samples *t*-test conducted on the overall sample showed a significant improvement from pretest to posttest, $t(99) = -11.31$, $p < 0.001$, Cohen's $d = 1.13$.

After controlling for pretest scores, the group effect on posttest spatial performance was significant, $F(3, 95) = 2.999$, $p = 0.034$, $\eta^2 = 0.087$. Bonferroni-adjusted comparisons indicated that the TE group outperformed the TEVC group, $MD = 1.948$, $p = 0.032$ (see Fig. 6).

A two-way ANCOVA revealed no significant main effects of textual explanations ($F(1, 95) = 0.861$, $p = 0.356$, $\eta^2 = 0.009$) or spatially augmented visual cues ($F(1, 95) = 3.527$, $p = 0.063$, $\eta^2 = 0.036$). However, a significant interaction effect between textual explanations and spatially augmented visual cues was observed, $F(1, 95) = 4.666$, $p = 0.033$, $\eta^2 = 0.047$ (see Fig. 7). Simple main effects analyses further showed that, without spatially augmented visual cues, textual explanation significantly improved adjusted posttest scores, $F(1, 95) = 4.752$, $p = 0.032$, $\eta^2 = 0.048$. In addition, within the textual explanation condition, the simple effect of spatially augmented visual cues was significant, $F(1, 95) = 8.124$, $p = 0.005$, $\eta^2 = 0.079$.

4.2. Interactive test performance

The one-way ANOVA showed no significant differences in interactive test performance among the experimental groups, ($F(3, 96) = 0.795$, $p = 0.500$, $\eta^2 = 0.024$). The two-way ANOVA also showed no significant main effect of textual explanations ($F(1, 96) = 0.000$, $p = 1.000$, $\eta^2 = 0.000$), no significant main effect of spatially augmented visual cues ($F(1, 96) = 1.067$, $p = 0.304$, $\eta^2 = 0.011$), and no significant interaction effect ($F(1, 96) = 1.317$, $p = 0.254$, $\eta^2 = 0.014$).

4.3. Learning experience

Results of the Kruskal-Wallis test indicated no significant differences in satisfaction across the training conditions, $H = 6.081$, $p = 0.108$. However, a significant difference was observed for motivation, $H = 9.972$, $p = 0.019$. Adjusted pairwise comparisons (see Fig. 8) revealed that motivation in the textual explanations group was significantly higher than that in the textual explanations and spatially augmented visual cues group, $H = 2.817$, $p = 0.029$, with no significant differences between the other groups. Results of the Kruskal-Wallis test indicated no significant differences in perceived usefulness across the experimental conditions, $H = 5.343$, $p = 0.148$, nor in perceived ease of use, $H = 2.432$, $p = 0.488$.

4.4. Eye-tracking outcomes

4.4.1. Average fixation duration

Descriptive statistics for average fixation duration are presented in Table 5. Results of the Kruskal-Wallis test indicated no significant differences in average fixation duration across conditions, $H = 2.345$, $p = 0.504$. This result suggests that this eye-tracking proxy did not reveal significant differences in overall cognitive processing demands among the conditions.

4.4.2. The number of fixations

The number of fixations was used to examine attention allocation across different AOIs within the same condition. The values in Table 6 refer to the total number of fixations within each AOI during the task, rather than values normalized by time. Results of the

Table 4
Descriptive statistics across four conditions.

Measure		NTENVC (N = 25) M (SD)	TE (N = 25) M (SD)	VC (N = 25) M (SD)	TEVC (N = 25) M (SD)
Pretest		26.48 (4.54)	27.72 (4.90)	26.80 (3.80)	26.52 (5.46)
Spatial performance (Posttest)		30.40 (3.49)	32.44 (2.43)	30.68 (3.25)	29.96 (3.42)
Interactive test		14.16 (1.25)	14.20 (1.76)	14.56 (2.14)	13.80 (1.71)
Learning experience	Satisfaction	5.63 (0.74)	6.04 (0.72)	6.08 (0.58)	5.68 (0.87)
	Motivation	5.35 (0.88)	5.83 (0.90)	5.41 (0.75)	5.24 (0.86)
	Perceived usefulness	5.71 (0.94)	6.07 (0.89)	5.88 (0.99)	5.59 (0.93)
	Perceived ease of use	6.12 (0.74)	6.36 (0.75)	6.37 (0.63)	6.20 (0.79)

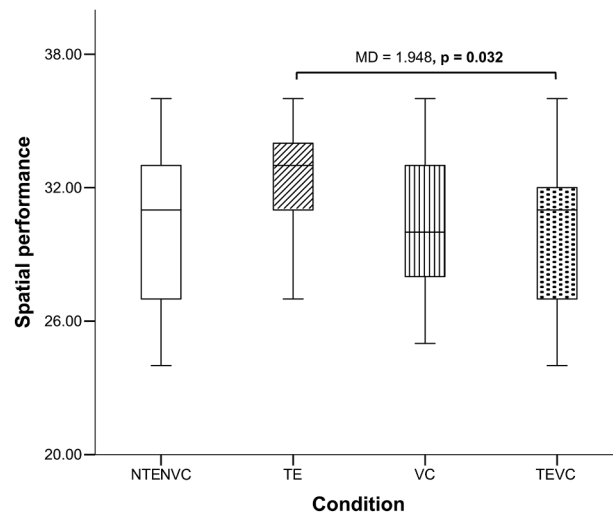


Fig. 6. Differences in the spatial performance across the four conditions.

Note. The bracket indicates a significant Bonferroni-adjusted comparison between the TE and TEVC groups.

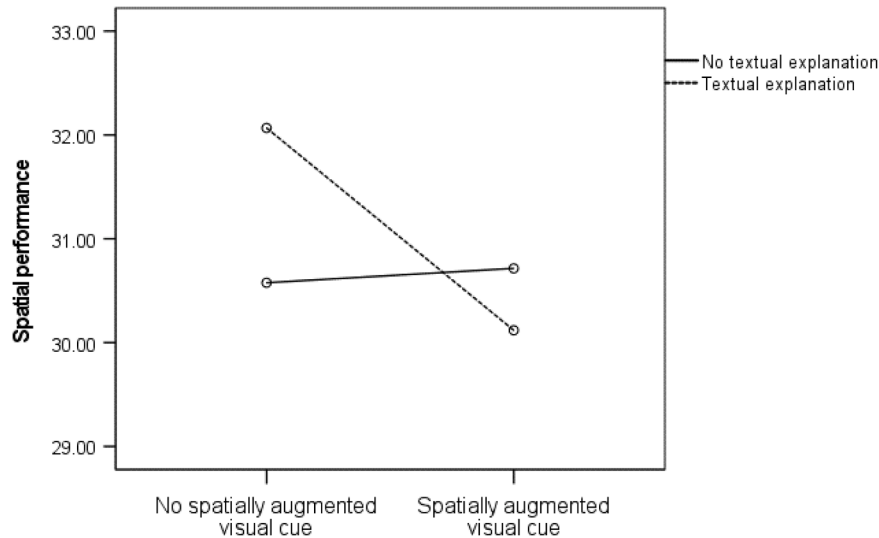


Fig. 7. Interaction effect of textual explanations and spatially augmented visual cues on spatial performance.

Wilcoxon signed-rank tests indicated that fixation counts in the combined cue/interactive areas were significantly higher than those in the text areas across the TE, VC, and TEVC conditions. Because this AOI combined the manipulation space and embedded visual cues, the result indicates attention to task-relevant interactive visual information rather than to visual cues alone.

5. Discussion

5.1. Empirical contributions

5.1.1. The effect of textual explanations depends on whether spatially augmented visual cues are present in desktop VR

The first major finding is that the effect of textual explanations on posttest spatial performance was conditional rather than independent. After controlling for pretest spatial performance, neither textual explanations nor spatially augmented visual cues showed a significant main effect on posttest performance, whereas their interaction was significant. Specifically, the textual-explanation-only condition outperformed the combined textual-explanation-and-visual-cue condition. This suggests that textual explanations may support spatial performance when they are not presented simultaneously with competing spatially augmented visual cues. Prior research showed that textual cueing can facilitate cognitive processing in desktop VR environments (Albus et al., 2021). However, the present findings do not support a general conclusion that textual explanations independently improve spatial performance across

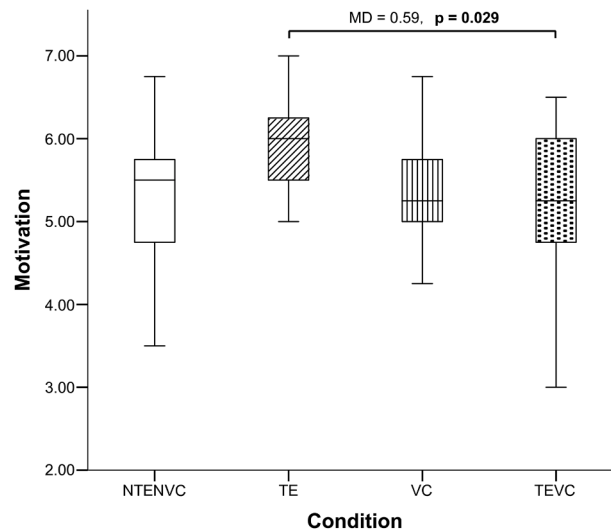


Fig. 8. Differences in the motivation across the four conditions.

Table 5

Descriptive statistics of average duration of fixation across the four conditions.

Measure	NTENVC (N = 25) M (SD)	TE (N = 23) M (SD)	VC (N = 23) M (SD)	TEVC (N = 24) M (SD)
Average duration of fixation (s)	0.31 (0.06)	0.29 (0.07)	0.30 (0.07)	0.30 (0.06)

Note: Eye-tracking recordings from two participants in the TE group, two in the VC group, and one in the TEVC group were excluded from the eye-tracking analyses due to technical issues, such as calibration failure, unstable tracking, or excessive data loss. These exclusions occurred across conditions and were not concentrated in any single group.

Table 6

Descriptive statistics and Wilcoxon signed-rank test results for AOIs across the conditions.

Condition	AOI	M (SD)	Q2 (Q1, Q4)	Z	p	r
TE (N = 23)	Text area	470.26 (250.75)	406.00 (307.00, 579.00)	-3.984 ^b	0.000	0.831
	Cue/interactive area	789.83 (440.76)	669.00 (476.00, 1111.00)			
VC (N = 23)	Text area	262.13 (143.69)	239.00 (158.00, 239.00)	-4.197 ^b	0.000	0.875
	Cue/interactive area	883.83 (376.19)	779.00 (612.00, 1173.00)			
TEVC (N = 24)	Text area	472.54 (162.31)	483.00 (312.50, 603.50)	-4.200 ^b	0.000	0.857
	Cue/interactive area	882.63 (460.34)	770.50 (565.00, 1100.75)			

Note: Eye-tracking recordings from two participants in the TE group, two in the VC group, and one in the TEVC group were excluded from the eye-tracking analyses due to technical issues, such as calibration failure, unstable tracking, or excessive data loss. These exclusions occurred across conditions and were not concentrated in any single group.

conditions. Rather, they suggest a conditional advantage of task-specific textual explanations in desktop VR-supported spatial training. In this study, textual explanations clarified correct answers, judgment rationales, and key spatial relations, including adjacent relations, opposite relations, and folded spatial correspondences. Thus, the findings should be interpreted as evidence about spatial task performance in desktop VR-supported spatial training, rather than as evidence that textual explanations improve problem solving in general.

The interactive test and posttest served different measurement purposes. The former reflected training-phase performance, whereas the latter assessed immediate spatial performance after training. Because interactive test performance did not differ significantly across conditions, the results do not indicate a clear training-phase advantage of textual explanations. Instead, the contrast between comparable training-phase performance and different posttest outcomes suggests that the arrangement of instructional supports may affect whether learners can perform effectively after the supports are removed.

Eye-tracking results showed that learners allocated significantly more visual attention to the combined cue and interactive area than to the text area across the TE, VC, and TEVC conditions. Because this AOI included both the interactive manipulation space and the embedded spatially augmented visual cues, this finding should not be interpreted as attentional capture by visual cues alone. Instead, it indicates that task-relevant interactive visual information attracted more attention than textual explanations. However,

textual explanations may still have supported learners' interpretation of spatial relations without dominating visual attention allocation. Additionally, no significant differences were found in average fixation duration across conditions, suggesting that textual explanations or spatially augmented visual cues did not increase overall cognitive load based on this eye-tracking proxy. This finding is consistent with prior evidence indicating that textual information does not necessarily increase cognitive load (Li et al., 2023a).

Although satisfaction, perceived usefulness, and perceived ease of use did not differ significantly across conditions, these findings may reflect the fact that all participants used the same desktop VR environment and completed the same spatial training tasks. Shared system-level and interaction-level features may have reduced between-group differences in learning experience (Lee et al., 2010; Makransky & Petersen, 2019). By contrast, the higher motivation reported in the textual-explanation-only condition relative to the combined condition suggests that motivation may be more sensitive to how textual explanations and spatially augmented visual cues are coordinated.

Taken together, these findings qualify prior multimedia and desktop VR research suggesting that textual or signaling cues can support learning (e.g., Albus & Seufert, 2022; King et al., 2023). Rather than showing a robust independent benefit of textual explanations, the present study found that their effectiveness in desktop VR-supported spatial training depends on how they are coordinated with spatially augmented visual cues.

5.1.2. *Spatially augmented visual cues do not produce a training-phase advantage or better posttest performance*

The second key finding is that spatially augmented visual cues did not produce a significant training-phase performance advantage or better posttest performance after controlling for pretest scores. Learners in the visual-cue conditions performed comparably to other groups during interactive training, indicating that the cues did not impair training-phase performance. However, this comparable performance was not accompanied by higher immediate posttest performance. This pattern may suggest that spatially augmented visual cues provided task-level guidance during training, but did not produce a measurable advantage beyond the supported practice phase.

This finding is consistent with prior research suggesting that realistic details may act as retrieval cues, but that such cue-related benefits may not persist when the cues are no longer available (Skulmowski & Rey, 2021). In the present study, spatially augmented visual cues may have guided learners' attention to task-relevant visual information during training. However, because the cue-related AOI also included the interactive manipulation space, the eye-tracking results should be interpreted as attention to combined cue and interactive information rather than to visual cues alone. Moreover, greater attention to the combined cue/interactive area was not accompanied by higher training-phase or immediate posttest performance.

Therefore, attention to visual cue/interactive areas should not be assumed to indicate better learning outcomes. Although visual cues can help focus attention on task-relevant elements (Boucheix & Lowe, 2010), the present findings show that such attention did not translate into better independent posttest performance. Spatially augmented visual cues should thus be interpreted cautiously as potential attentional guides rather than as demonstrated supports for improved learning performance.

5.1.3. *Spatially augmented visual cues weaken the advantage observed in the textual-explanation-only condition*

The third major finding concerns the significant interaction between textual explanations and spatially augmented visual cues. Follow-up simple main effects analyses showed that textual explanations significantly improved adjusted posttest scores only when spatially augmented visual cues were absent. Within the textual explanation condition, learners achieved higher adjusted posttest scores when visual cues were absent than when they were present. These results suggest that the benefit of textual explanations was conditional, and that adding spatially augmented visual cues weakened the performance advantage observed in the textual-explanation-only condition.

One possible explanation is that, in the combined condition, learners had to coordinate textual explanations, the 3D model, spatially augmented visual cues, and interactive manipulation simultaneously. This may have increased the need to coordinate multiple information sources and reduced the advantage associated with textual guidance. From the perspective of the split-attention effect (Mayer & Moreno, 1998; Mayer, 2024), simultaneous presentation of textual and visual supports may have divided learners' attention rather than helping them integrate spatial relations more effectively. However, this interpretation should remain tentative because the present eye-tracking analysis did not separately isolate attention to visual cues from attention to the interactive manipulation space.

Prior multimedia research has shown that combining text and visual cues can support learning under some conditions (e.g., Wang et al., 2020). However, these studies differed from the present study in learning content, task type, media format, and research design. The present results should therefore be interpreted cautiously: in this desktop VR-supported spatial training task, combining textual explanations with spatially augmented visual cues did not produce additive benefits. This interaction emphasizes the need for careful instructional design in desktop VR learning environments. Rather than assuming that more instructional support is always beneficial, designers should consider that different forms of support may not produce additive benefits when presented simultaneously and may require careful coordination.

5.2. *Theoretical implications*

Rather than merely extending the signaling principle from conventional multimedia learning to desktop VR, this study refines the principle by identifying boundary conditions under which different forms of signals support or fail to support spatial learning in interactive 3D environments. While the signaling principle suggests that cues can guide attention and facilitate deeper learning (Schneider et al., 2018), the present findings indicate that attention guidance alone may not be sufficient for improving spatial learning

outcomes in desktop VR-supported spatial training. After controlling for pretest spatial performance, neither textual explanations nor spatially augmented visual cues showed a significant main effect on posttest performance; instead, their interaction was significant. Moreover, visual attention to cue/interactive areas did not translate into better training-phase or posttest performance. This distinction between attention guidance, training-phase performance, and post-training performance qualifies a key assumption of the signaling principle (Van Gog, 2014): cues may attract attention, but they do not necessarily improve posttest performance. Extending Li et al.'s (2023) observation that cue effects may vary across media, learning environments, and instructional combinations, the present findings further suggest that signaling effectiveness in desktop VR depends on how instructional cues are coordinated with textual explanations and learners' attentional processing.

5.3. Practical implications

The findings suggest that instructional supports in desktop VR-supported spatial training should be carefully coordinated rather than simply combined. Designers should consider aligning textual explanations with learners' interactive exploration so that key spatial relations are made explicit, while being cautious about presenting spatially augmented visual cues simultaneously with textual explanations. Because the textual-explanation-only condition outperformed the combined condition, textual guidance may be less useful when it is presented simultaneously with spatially augmented visual cues. Based on the average fixation duration measure, there was no eye-tracking evidence that textual guidance increased overall cognitive load. However, this finding should be interpreted cautiously because cognitive load was assessed using a single indirect eye-tracking proxy.

Spatially augmented visual cues should also be used selectively. Although learners attended more to the combined cue/interactive areas, this attention did not produce a training-phase advantage or better posttest performance. Therefore, spatially augmented visual cues should not be treated as beneficial simply because they attract attention. Designers may present cues sequentially with textual explanations, or gradually fade cues to encourage independent spatial reasoning. These implications are best interpreted within the context of the present desktop VR-supported spatial training task and do not indicate that the same design recommendations are specific to VR or generalizable to other learning environments.

6. Limitations and future directions

This study has several limitations that suggest directions for future research. First, it examined signaling strategies in short-term spatial performance training, whereas spatial ability development typically requires long-term practice. Potential practice effects cannot be fully excluded because the same items were used in the pretest and posttest, although all groups followed the same procedure and the analyses focused on between-group comparisons. In addition, the interactive test and posttest measured different aspects of performance and should not be treated as directly comparable indicators of internalization or transfer. As the learning experience and eye-tracking data did not meet parametric assumptions, these non-parametric results should be interpreted as exploratory group-level comparisons. Future studies should employ longitudinal designs, delayed posttests, and parallel test forms to better distinguish training effects from possible practice effects and examine whether the effects of textual explanations and spatially augmented visual cues are sustained over time. Although the spatial performance test showed acceptable internal consistency and preliminary construct-related validity evidence, items with relatively low discrimination should be further refined. Second, the study was conducted in a desktop VR environment, and it remains unclear whether the effects generalize to IVR settings, where higher immersion and interaction may alter the role of instructional supports. The findings should be interpreted in relation to desktop VR-supported 3D interactive learning environments rather than fully immersive HMD-based VR. Future research should investigate whether these findings hold in IVR environments. Third, although cognitive load theory was used to interpret learners' processing demands, cognitive load was measured only through average fixation duration, an indirect eye-tracking proxy. This measure did not allow us to comprehensively assess cognitive load or distinguish intrinsic, extraneous, and germane load. Moreover, eye-tracking findings should also be interpreted cautiously. The 60 Hz sampling rate limited the capture of fine-grained attention shifts. Some recordings were excluded due to quality issues, and although these exclusions showed no apparent concentration across groups, the potential influence of the missing data cannot be entirely ruled out. The randomness of the missing eye-tracking data was not formally tested. Future studies should combine validated scales with eye-tracking to differentiate cognitive load types, utilize higher-resolution devices with stricter calibration to reduce data loss, and incorporate complementary metrics (e.g., pupillometry) or multi-modal data. Fourth, the combined cue/interactive area included both spatially augmented visual cues and the interactive manipulation space. Thus, the current eye-tracking analysis could not fully separate attention to visual cues from attention driven by interaction itself. Future studies could define more fine-grained AOIs to isolate cue-specific attentional effects. In addition, individual differences in spatial ability may moderate the effectiveness of instructional supports and should be examined in future adaptive signaling designs. Finally, this study manipulated the presence of task-specific textual explanations but did not independently vary the type of the textual content. The findings cannot fully disentangle the effect of text presence from the effect of explanatory content. Future studies should compare different forms of textual support to determine which textual features most effectively support spatial training in desktop VR. Moreover, because this study focused on spatial visualization and mental rotation tasks, future research should examine whether similar effects occur in non-spatial problem-solving domains.

Funding statement

This work was supported by the Research Projects of the National Natural Science Foundation of China under Grant [62377039];

2025 Shaanxi Normal University Graduate Education Comprehensive Reform Research and Practice Project under Grant [GERP-25-01]; and Higher Education Research Fund of Northwestern Polytechnical University in 2025, No. GJJJM202504.

Consent to participate

All participants signed informed consent agreement.

Artificial intelligence generated content

None.

Ethics approval statement

This research complies with the WMA Declaration of Helsinki regarding participant recruitment, data collection, data use, data confidentiality, and protection of participants' rights. All participants were informed of the voluntary nature of their participation and provided written informed consent. They were assured of anonymity and informed about the purpose of the research, how their data would be used, and that no risks were associated with their involvement.

CRedit authorship contribution statement

Qian Chen: Writing – original draft, Visualization, Validation, Formal analysis, Conceptualization. **Yun Zhou:** Writing – review & editing, Writing – original draft, Validation, Supervision, Funding acquisition, Formal analysis, Conceptualization. **HanBing Kou:** Formal analysis. **Hui Zhang:** Formal analysis. **Tao Xu:** Writing – review & editing, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare no conflicts of interest.

Acknowledgments

We would like to thank Ms Jialin Yu and Mr. Dong Yang for their technical support in VR room.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at [doi:10.1016/j.tsc.2026.102402](https://doi.org/10.1016/j.tsc.2026.102402).

Data availability

Data will be made available on request.

References

- Albus, P., & Seufert, T. (2022). Signaling in 360° desktop virtual reality influences learning outcome and cognitive load. *Frontiers in Education*, 7, Article 916105. <https://doi.org/10.3389/educ.2022.916105>
- Albus, P., Vogt, A., & Seufert, T. (2021). Signaling in virtual reality influences learning outcome and cognitive load. *Computers & Education*, 166, Article 104154. <https://doi.org/10.1016/j.compedu.2021.104154>
- Antrilli, N. K., & Wang, S. H. (2016). Visual cues generated during action facilitate 14-month-old infants' mental rotation. *Journal of Cognition and Development*, 17(3), 418–429. <https://doi.org/10.1080/15248372.2015.1058262>
- Ardaç, D., & Unal, S. (2008). Does the amount of on-screen text influence student learning from a multimedia-based instructional unit? *Instructional Science*, 36(1), 75–88. <https://doi.org/10.1007/s11251-007-9035-4>
- Arslan-Ari, I. (2018). Learning from instructional animations: How does prior knowledge mediate the effect of visual cues? *Journal of Computer Assisted Learning*, 34(2), 140–149. <https://doi.org/10.1111/jcal.12222>
- Assaf, R., Mendes, D., & Rodrigues, R. (2024). Cues to fast-forward collaboration: A survey of workspace awareness and visual cues in XR collaborative systems. *Computer Graphics Forum*, 43(2), Article e15066. <https://doi.org/10.1111/cgf.15066>
- Ayres, P., & Sweller, J. (2005). The split-attention principle in multimedia learning. *The Cambridge Handbook Of Multimedia Learning*, 2, 135–146. <https://doi.org/10.1017/CBO9780511816819.009>
- Bodner, G. M., & Guay, R. B. (1997). The Purdue visualization of rotations test. *The Chemical Educator*, 2(4), 1–17. <https://doi.org/10.1007/s00897970138a>
- Boucheix, J.-M., & Lowe, R. K. (2010). An eye tracking comparison of external pointing cues and internal continuous cues in learning with complex animations. *Learning and Instruction*, 20(2), 123–135. <https://doi.org/10.1016/j.learninstruc.2009.02.015>
- Boucheix, J. M., Lowe, R. K., Putri, D. K., & Groff, J. (2013). Cueing animations: Dynamic signaling aids information extraction and comprehension. *Learning and Instruction*, 25, 71–84. <https://doi.org/10.1016/j.learninstruc.2012.11.005>
- Castro-Alonso, J. C., Ayres, P., & Sweller, J. (2019). Instructional visualizations, cognitive load theory, and visuospatial processing. *Visuospatial processing for education in health and natural sciences* (pp. 111–143). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-20969-8_5

- Castro-Alonso, J. C., & Uttal, D. H. (2019). Spatial ability for university biology education. In S. Nazir, A.-M. Teperi, & A. Polak-Sopińska (Eds.), *Advances in human factors in training, education, and learning sciences: 785. Advances in human factors in training, education, and learning sciences* (pp. 283–291). Springer International Publishing. https://doi.org/10.1007/978-3-319-93882-0_28.
- Chang, C.-W., Heo, J., Yeh, S.-C., Han, H.-Y., & Li, M. (2018). The effects of immersion and interactivity on college students' acceptance of a novel VR-supported educational technology for mental rotation. *IEEE Access*, 6, 66590–66599. <https://doi.org/10.1109/ACCESS.2018.2878270>
- Chen, Q., Deng, L., Xu, T., & Zhou, Y. (2022). Visualized cues for enhancing spatial ability training in virtual reality. In 2022 *IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)* (pp. 299–300). IEEE. <https://doi.org/10.1109/VRW55335.2022.00067>.
- Chou, S., & Liu, C. (2005). Learning effectiveness in a Web-based virtual learning environment: A learner control perspective. *Journal of Computer Assisted Learning*, 21(1), 65–76. <https://doi.org/10.1111/j.1365-2729.2005.00114.x>
- Clinton, V., Cooper, J. L., Michaelis, J., Alibali, M. W., & Nathan, M. J. (2017). Revising visuals based on instructional design principles: Effects on cognitive load and learning. *Eye-Tracking Technology Applications In Educational Research*, 195–218.
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Routledge. <https://doi.org/10.4324/9780203771587>. <https://www.taylorfrancis.com/books/9781134742707>. (Accessed 13 May 1988)
- Danziger, S., & Rafal, R. (2009). The effect of visual signals on spatial decision making. *Cognition*, 110(2), 182–197. <https://doi.org/10.1016/j.cognition.2008.11.005>
- de Koning, B. B., Tabbers, H. K., Rikers, R. M., & Paas, F. (2010). Attention guidance in learning from a complex animation: Seeing is understanding? *Learning and Instruction*, 20(2), 111–122. <https://doi.org/10.1016/j.learninstruc.2009.02.010>
- Di, X., & Zheng, X. (2022). A meta-analysis of the impact of virtual technologies on students' spatial ability. *Educational Technology Research and Development*, 70(1), 73–98. <https://doi.org/10.1007/s11423-022-10082-3>
- Dickey, M. D. (2005). Brave new (interactive) worlds: A review of the design affordances and constraints of two 3D virtual worlds as interactive learning environments. *Interactive Learning Environments*, 13(1–2), 121–137. <https://doi.org/10.1080/10494820500173714>
- Dodd, B. J., & Antonenko, P. D. (2012). Use of signaling to integrate desktop virtual reality and online learning management systems. *Computers & Education*, 59(4), 1099–1108. <https://doi.org/10.1016/j.compedu.2012.05.016>
- Doerr, N., Lee, B., Baricova, K., Schmalstieg, D., & Sedlmair, M. (2024). Visual highlighting for situated brushing and linking. *Computer Graphics Forum*, 43(3), Article e15105. <https://doi.org/10.1111/cgf.15105>
- Du, L., Tang, X., & Wang, J. (2025). Different types of textual cues in educational animations: Effect on science learning outcomes, cognitive load, and self-efficacy among elementary students. *Education and Information Technologies*, 30(3), 3573–3596. <https://doi.org/10.1007/s10639-024-12929-z>
- Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293), 52–64. <https://doi.org/10.1080/01621459.1961.10482090>
- Ekstrom, R. B., & Harman, H. H. (1976). *Manual for kit of factor-referenced cognitive tests*, 1976. Educational testing service.
- Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2), 175–191. <https://doi.org/10.3758/BF03193146>
- Field, A. (2012). *Discovering statistics using SPSS: And sex and drugs and rock'n'roll* (3. ed., repr). Sage.
- Guo, S., Sun, Y., & Wang, X. (2025). Does spatial ability affect students' learning performance in virtual reality environments? A study on desktop virtual reality systems. *Thinking Skills and Creativity*, 58, Article 101923. <https://doi.org/10.1016/j.tsc.2025.101923>
- Han, J., Liu, G., & Zheng, Q. (2023). Prior knowledge as a moderator between signaling and learning performance in immersive virtual reality laboratories. *Frontiers in Psychology*, 14, Article 1118174. <https://doi.org/10.3389/fpsyg.2023.1118174>
- Hauptman, H. (2010). Enhancement of spatial thinking with Virtual Spaces 1.0. *Computers & Education*, 54(1), 123–135. <https://doi.org/10.1016/j.compedu.2009.07.013>
- Hegarty, M. (2010). Components of spatial intelligence. In *Psychology of learning and motivation*, 52 pp. 265–297. Academic Press.
- Hsieh, C. (2026). Strategic limitations in preschoolers' relational spatial reasoning: Insights from eye-tracking and instructional intervention. *British Journal of Educational Psychology*, 96(2), 780–801. <https://doi.org/10.1111/bjep.70051>
- Hsieh, Y. (2020). Effects of video captioning on EFL vocabulary learning and listening comprehension. *Computer Assisted Language Learning*, 33(5–6), 567–589. <https://doi.org/10.1080/09588221.2019.1577898>
- Hwang, G.-J., Yang, L.-H., & Wang, S.-Y. (2013). A concept map-embedded educational computer game for improving students' learning performance in natural science courses. *Computers & Education*, 69, 121–130. <https://doi.org/10.1016/j.compedu.2013.07.008>
- Hyunjeong, L., & Mayer, R. E. (2018). Fostering learning from instructional video in a second language. *Applied Cognitive Psychology*, 32(5), 648–654. <https://doi.org/10.1002/acp.3436>
- Imhof, B., Scheiter, K., Edelmann, J., & Gerjets, P. (2013). Learning about locomotion patterns: Effective use of multiple pictures and motion-indicating arrows. *Computers & Education*, 65, 45–55. <https://doi.org/10.1016/j.compedu.2013.01.017>
- Jamet, E. (2014). An eye-tracking study of cueing effects in multimedia learning. *Computers in Human Behavior*, 32, 47–53. <https://doi.org/10.1016/j.chb.2013.07.008>
- Kang, H. (2021). Sample size determination and power analysis using the G*Power software. *Journal of Educational Evaluation for Health Professions*, 18, 17. <https://doi.org/10.3352/jeehp.2021.18.17>
- Kelley, B., Wickens, C., Clegg, B., Warden, A. C., & Ortega, F. R. (2024). Guiding gaze: Comparing cues for visual search. In 2024 *IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)* (pp. 1053–1054). <https://doi.org/10.1109/VRW62533.2024.00323>
- Keskin, M., Ooms, K., Dogru, A. O., & Maeyer, P. D. (2018). Digital sketch maps and eye tracking statistics as instruments to obtain insights into spatial cognition. *Journal of Eye Movement Research*, 11(3), 23. <https://doi.org/10.16910/jemr.11.3.4>
- King, J., Marcus, T., & Markant, J. (2023). Individual differences in selective attention and engagement shape students' learning from visual cues and instructor presence during online lessons. *Scientific Reports*, 13(1), 5075. <https://doi.org/10.1038/s41598-023-32069-7>
- Korbach, A., Brünken, R., & Park, B. (2018). Differentiating different types of cognitive load: A comparison of different measures. *Educational Psychology Review*, 30(2), 503–529.
- Langlois, J., Bellemare, C., Toulouse, J., & Wells, G. A. (2020). Spatial abilities training in anatomy education: A systematic review. *Anatomical Sciences Education*, 13(1), 71–79. <https://doi.org/10.1002/ase.1873>
- Lee, E. A.-L., Wong, K. W., & Fung, C. C. (2010). How does desktop virtual reality enhance learning outcomes? A structural equation modeling approach. *Computers & Education*, 55(4), 1424–1442. <https://doi.org/10.1016/j.compedu.2010.06.006>
- Lee, E. A.-L., & Wong, K. W. (2014). Learning with desktop virtual reality: Low spatial ability learners are more positively affected. *Computers & Education*, 79, 49–58. <https://doi.org/10.1016/j.compedu.2014.07.010>
- Li, W., Feng, Q., Zhu, X., Yu, Q., & Wang, Q. (2023a). Effect of summarizing scaffolding and textual cues on learning performance, mental model, and cognitive load in a virtual reality environment: An experimental study. *Computers & Education*, 200, Article 104793. <https://doi.org/10.1016/j.compedu.2023.104793>
- Li, W., Wang, F., & Mayer, R. E. (2023b). How to guide learners' processing of multimedia lessons with pedagogical agents. *Learning and Instruction*, 84, Article 101729. <https://doi.org/10.1016/j.learninstruc.2022.101729>
- Lin, C. H., & Chen, C. M. (2016). Developing spatial visualization and mental rotation with a digital puzzle game at primary school level. *Computers in Human Behavior*, 57, 23–30. <https://doi.org/10.1016/j.chb.2015.12.026>
- Lin, L., & Atkinson, R. K. (2011). Using animations and visual cueing to support learning of scientific concepts and processes. *Computers & Education*, 56(3), 650–658. <https://doi.org/10.1016/j.compedu.2010.10.007>
- Linn, M. C., & Petersen, A. C. (1985). Emergence and characterization of sex differences in spatial ability: A meta-analysis. *Child Development*, 56(6), 1479. <https://doi.org/10.2307/1130467>
- Lu, W., & Carr, M. (2014). Working memory and strategy use contribute to gender differences in spatial ability. *Educational Psychologist*, 49(4), 261–282. <https://doi.org/10.1080/00461520.2014.960568>

- Makransky, G., & Petersen, G. B. (2019). Investigating the process of learning with desktop virtual reality: A structural equation modeling approach. *Computers & Education*, 134, 15–30. <https://doi.org/10.1016/j.compedu.2019.02.002>
- Makransky, G., Terkildsen, T. S., & Mayer, R. E. (2019). Adding immersive virtual reality to a science lab simulation causes more presence but less learning. *Learning and Instruction*, 60, 225–236. <https://doi.org/10.1016/j.learninstruc.2017.12.007>
- Martín-Gutiérrez, J., Saorín, J. L., Contero, M., Alcañiz, M., Pérez-López, D. C., & Ortega, M. (2010). Design and validation of an augmented book for spatial abilities development in engineering students. *Computers & Graphics*, 34(1), 77–91. <https://doi.org/10.1016/j.cag.2009.11.003>
- Marwan, S., Lytle, N., Williams, J. J., & Price, T. (2019). The impact of adding textual explanations to next-step hints in a novice programming environment. In *Proceedings of the 2019 ACM Conference on Innovation and Technology in Computer Science Education* (pp. 520–526). <https://doi.org/10.1145/3304221.3319759>
- Mast, F. W., & Kosslyn, S. M. (2002). Eye movements during visual mental imagery. *Trends in Cognitive Sciences*, 6(7), 271–272. [https://doi.org/10.1016/S1364-6613\(02\)01931-9](https://doi.org/10.1016/S1364-6613(02)01931-9)
- Mayer, R. E. (2024). The past, present, and future of the cognitive theory of multimedia learning. *Educational Psychology Review*, 36(1), 8. <https://doi.org/10.1007/s10648-023-09842-1>
- Mayer, R. E., & Moreno, R. (1998). A split-attention effect in multimedia learning: Evidence for dual processing systems in working memory. *Journal of Educational Psychology*, 90(2), 312. <https://doi.org/10.1037/0022-0663.90.2.312>
- Mikropoulos, T. A., & Natsis, A. (2011). Educational virtual environments: A ten-year review of empirical research (1999–2009). *Computers & Education*, 56(3), 769–780. <https://doi.org/10.1016/j.compedu.2010.10.020>
- Mix, K. S., & Cheng, Y.-L. (2012). The relation between space and math. In *Advances in child development and behavior* (Vol. 42, pp. 197–243). Elsevier. <https://doi.org/10.1016/B978-0-12-394388-0.00006-X>
- Nelson, B. C., Kim, Y., Foshee, C., & Slack, K. (2014). Visual signaling in virtual world-based assessments: The SAVE Science project. *Information Sciences*, 264, 32–40. <https://doi.org/10.1016/j.ins.2013.09.011>
- Nori, R., Zucchelli, M. M., Palmiero, M., & Piccardi, L. (2023). Environmental cognitive load and spatial anxiety: What matters in navigation? *Journal of Environmental Psychology*, 88, Article 102032. <https://doi.org/10.1016/j.jenvp.2023.102032>
- Ozcelik, E., Karakus, T., Kursun, E., & Cagiltay, K. (2009). An eye-tracking study of how color coding affects multimedia learning. *Computers & Education*, 53(2), 445–453. <https://doi.org/10.1016/j.compedu.2009.03.002>
- Parong, J., & Mayer, R. E. (2021). Cognitive and affective processes for learning science in immersive virtual reality. *Journal of Computer Assisted Learning*, 37(1), 226–241. <https://doi.org/10.1111/jcal.12482>
- Ponce, H. R., & Mayer, R. E. (2014). An eye movement analysis of highlighting and graphic organizer study aids for learning from expository text. *Computers in Human Behavior*, 41, 21–32. <https://doi.org/10.1016/j.chb.2014.09.010>
- Poole, A., & Ball, L. (2006). Eye tracking in human-computer interaction and usability research: Current status and future prospects. In *Encyclopedia of human computer interaction*, pp. 211–219.
- Purwanto, Hamdan, A., Putra, A. K., Aripriharta, Tan, I., & Farihah, S. N. (2024). Geo-Virtual Reality (GVR): The creative materials to construct spatial thinking skills using virtual learning based metaverse technology. *Thinking Skills and Creativity*, 54, Article 101664. <https://doi.org/10.1016/j.tsc.2024.101664>
- Ragazou, V., & Karasavvidis, I. (2023). Considering visual cueing and practice type for software training's optimisation: A study of domain experts. *Interactive Technology and Smart Education*, 20(1), 36–57. <https://doi.org/10.1108/ITSE-05-2021-0091>
- Rahmawati, Y., Dianhar, H., & Arifin, F. (2021). Analysing students' spatial abilities in chemistry learning using 3d virtual representation. *Education Sciences*, 11(4), 185. <https://doi.org/10.3390/educsci11040185>
- Ramirez, G., Gunderson, E. A., Levine, S. C., & Beilock, S. L. (2012). Spatial anxiety relates to spatial abilities as a function of working memory in children. *Quarterly Journal of Experimental Psychology*, 65(3), 474–487. <https://doi.org/10.1080/17470218.2011.616214>
- Richter, J., & Scheiter, K. (2019). Studying the expertise reversal of the multimedia signaling effect at a process level: Evidence from eye tracking. *Instructional Science*, 47(6), 627–658. <https://doi.org/10.1007/s11251-019-09492-3>
- Roach, V. A., Fraser, G. M., Kryklywy, J. H., Mitchell, D. G., & Wilson, T. D. (2019). Guiding low spatial ability individuals through visual cueing: The dual importance of where and when to look. *Anatomical Sciences Education*, 12(1), 32–42. <https://doi.org/10.1002/ase.1783>
- Scheiter, K., & Eitel, A. (2015). Signals foster multimedia learning by supporting integration of highlighted text and diagram elements. *Learning and Instruction*, 36, 11–26. <https://doi.org/10.1016/j.learninstruc.2014.11.002>
- Schneider, S., Beege, M., Nebel, S., & Rey, G. D. (2018). A meta-analysis of how signaling affects learning with media. *Educational Research Review*, 23, 1–24. <https://doi.org/10.1016/j.edurev.2017.11.001>
- Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science, New Series*, 171(3972), 701–703. <https://www.science.org/doi/10.1126/science.171.3972.701>
- Skulmowski, A., & Rey, G. D. (2021). Realism as a retrieval cue: Evidence for concreteness-specific effects of realistic, schematic, and verbal components of visualizations on learning and testing. *Human Behavior and Emerging Technologies*, 3(2), 283–295. <https://doi.org/10.1002/hbe2.209>
- Stark, L., Brinken, R., & Park, B. (2018). Emotional text design in multimedia learning: A mixed-methods study using eye tracking. *Computers & Education*, 120, 185–196. <https://doi.org/10.1016/j.compedu.2018.02.003>
- Stefanucci, J. K., Brickler, D., Finney, H. C., Wilson, E., Drew, T., & Creem-Regehr, S. H. (2022). Effects of simulated augmented reality cueing in a virtual navigation task. *Frontiers in Virtual Reality*, 3, Article 971310. <https://doi.org/10.3389/fvrv.2022.971310>
- Stieff, M., & Uttal, D. (2015). How much can spatial training improve stem achievement? *Educational Psychology Review*, 27(4), 607–615. <https://doi.org/10.1007/s10648-015-9304-8>
- Sun, Y., Lu, X., & Wang, Y. (2020). Using eye tracking to explore differences between high and low map-based spatial ability. *Journal of Geography*, 119(6), 215–225.
- Supli, A. A., & Yan, X. (2024). Exploring the effectiveness of augmented reality in enhancing spatial reasoning skills: A study on mental rotation, spatial orientation, and spatial visualization in primary school students. *Education And Information Technologies*, 29(1), 351–374. <https://doi.org/10.1007/s10639-023-12255-w>
- Sweller, J., Van Merriënboer, J. J., & Paas, F. G. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, 10(3), 251–296. <https://doi.org/10.1023/A:1022193728205>
- Mayer, R. E., & Fiorella, L. (2021). *The Cambridge Handbook of Multimedia Learning*. Cambridge University Press. <https://doi.org/10.1017/9781108894333>. <https://www.cambridge.org/core/product/identifier/9781108894333/type/book>. (Accessed 19 November 2021)
- Tian, Y., Yang, T., & Dong, W. (2022). Does culture shape our spatial ability? An investigation based on eye tracking. *ISPRS International Journal of Geo-Information*, 11(11), 578. <https://doi.org/10.3390/ijgi11110578>
- Turčani, M., Balogh, Z., & Kohútek, M. (2024). Evaluating computer science students reading comprehension of educational multimedia-enhanced text using scalable eye-tracking methodology. *Smart Learning Environments*, 11(1), 29. <https://doi.org/10.1186/s40561-024-00318-5>
- Uttal, D. H., & Cohen, C. A. (2012). Spatial thinking and STEM education. In *The Psychology of Learning and Motivation: The Psychology of Learning and Motivation* (pp. 147–181). doi:10.1016/B978-0-12-394293-7.00004-2.
- Van Gog, T. (2014). 11–The signaling (or cueing) principle in multimedia learning. *The Cambridge Handbook Of Multimedia Learning*, 263.
- Van Gog, T., & Scheiter, K. (2010). Eye tracking as a tool to study and enhance multimedia learning. *Learning and Instruction*, 20(2), 95–99. <https://doi.org/10.1016/j.learninstruc.2009.02.009>
- Vandenberg, S. G., & Kuse, A. R. (1978). Mental rotations, a group test of three-dimensional spatial visualization. *Perceptual And Motor Skills*, 47(2), 599–604. <https://doi.org/10.2466/pms.1978.47.2.599>
- Wai, J., Lubinski, D., & Benbow, C. P. (2009). Spatial ability for STEM domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance. *Journal of Educational Psychology*, 101(4), 817–835. <https://doi.org/10.1037/a0016127>
- Wang, X., Lin, L., Han, M., & Spector, J. M. (2020). Impacts of cues on learning: Using eye-tracking technologies to examine the functions and designs of added cues in short instructional videos. *Computers in Human Behavior*, 107, Article 106279. <https://doi.org/10.1016/j.chb.2020.106279>

- Wen, P., Lu, F., & Mohamad Ali, A. Z. (2024). Using attentional guidance methods in virtual reality laboratories reduces students' cognitive load and improves their academic performance. *Virtual Reality*, 28(2), 110. <https://doi.org/10.1007/s10055-024-01012-0>
- Xiao, N., & Zhang, X. (2021). Interest in spatial activities predicts young children's spatial ability development: A two-year longitudinal study. *Contemporary Educational Psychology*, 64, Article 101943. <https://doi.org/10.1016/j.cedpsych.2021.101943>
- Yue, C. L., Bjork, E. L., & Bjork, R. A. (2013). Reducing verbal redundancy in multimedia learning: An undesired desirable difficulty? *Journal of Educational Psychology*, 105(2), 266–277. <https://doi.org/10.1037/a0031971>
- Zhai, X., Fang, Q., Dong, Y., Wei, Z., Yuan, J., Cacciolatti, L., & Yang, Y. (2018). The effects of biofeedback-based stimulated recall on self-regulated online learning: A gender and cognitive taxonomy perspective. *Journal of Computer Assisted Learning*, 34(6), 775–786. <https://doi.org/10.1111/jcal.12284>
- Zhou, Y., Xu, T., Yang, H., & Li, S. (2022). Improving spatial visualization and mental rotation using FORSpatial through shapes and letters in virtual environment. *IEEE Transactions on Learning Technologies*, 15(3), 326–337. <https://doi.org/10.1109/TLT.2022.3170928>
- Zhou, Y., Yi, F., Dong, B., Zhang, G., Zhang, Y., & Xu, T. (2024). Mental imagery scaffolding: The effects of detail richness and text load on geography learning. *Education and Information Technologies*, 29(13), 16929–16956. <https://doi.org/10.1007/s10639-024-12540-2>