Spatial Visual Guides: Supporting Spatial Thinking Through Multimodal Hand Interaction in VR

Prashant Rawat
Faculty of Computer Science
Dalhousie University
Halifax, Nova Scotia, Canada
prashant.rawat@dal.ca

Ga Wu
Faculty of Computer Science
Dalhousie University
Halifax, Nova Scotia, Canada
ga.wu@dal.ca

Francisco Raul Ortega
Computer Science
Colorado State University
Ft. Collins, Colorado, USA
fortega@colostate.edu

Lizbeth Escobedo
Faculty of Computer Science
Dalhousie University
Halifax, Nova Scotia, Canada
lizbeth.escobedo@dal.ca

Mayra Donaji Barrera Machuca
Department of Computer Science
University of Calgary
Calgary, Alberta, Canada
mbarrera@ucalgary.ca

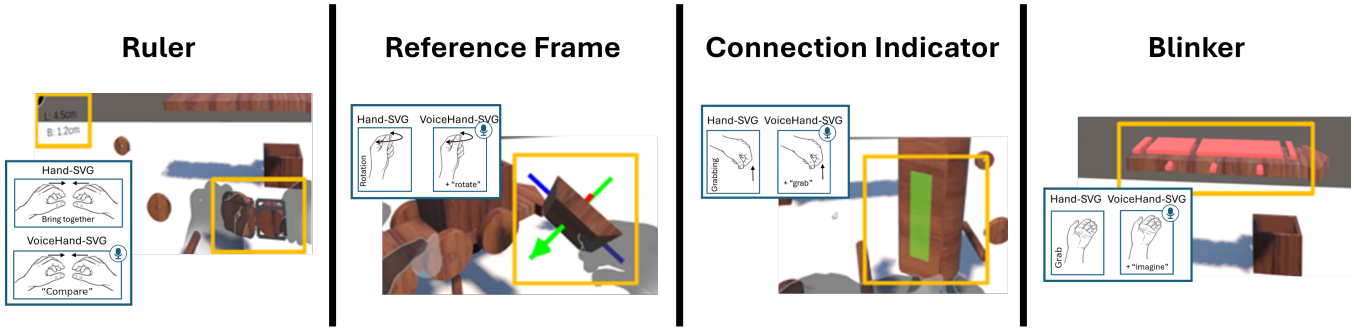


Figure 1: Spatial Visual Guides (SVG) support spatial thinking during 3D assembly by mapping Spatial Hand Actions [52] to visual guidance aligned with the spatial reasoning processes those actions externalize: dimensional comparison (*Ruler*), orientation alignment (*ReferenceFrame*), attachment compatibility (*ConnectionIndicator*), and placement preview (*Blinker*).

Abstract

Assembling objects in Virtual Reality using hand actions is cognitively demanding because users must identify relationships among components. We present **Spatial Visual Guides** (SVG), a novel user interface that provides context-sensitive visual guides designed to scaffold users' spatial reasoning during 3D assembly tasks. Users invoke **SVG** automatically through naturalistic hand actions (*Hand-SVG*) or deliberately through a multimodal combination of voice and hand input (*VoiceHand-SVG*). In a user study, 24 participants performed a 3D assembly task at two levels of complexity (easy and hard) using SVG. We found a difference in usage between the multimodal *VoiceHand-SVG* interface and *Hand-SVG*: *VoiceHand-SVG* produced fewer guide activations while remaining visible for similar durations. Moreover, participants' articulated strategies and behavioural logs helped us understand how participants used SVGs during assembly, and showed that participants used SVGs for orientation anchoring, relational comparison, and placement planning. SVGs also improved task performance compared to no visual feedback. Our results demonstrate the advantages of designing visual guides around specific spatial reasoning behaviours, and reveal trade-offs between unimodal and multimodal interaction strategies.

CCS Concepts

• **Human-centered computing** → **Gestural input**; *Virtual reality*; User studies.

Keywords

Hand Actions, 3D User Interface, Multimodal Interactions, Spatial Thinking

ACM Reference Format:
Prashant Rawat, Ga Wu, Francisco Raul Ortega, Lizbeth Escobedo, and Mayra Donaji Barrera Machuca. 2026. Spatial Visual Guides: Supporting Spatial Thinking Through Multimodal Hand Interaction in VR. In *32nd ACM Symposium on Virtual Reality Software and Technology (VRST '26), November 16–18, 2026, Sendai, Japan.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3822517.3848697

This work is licensed under a Creative Commons Attribution 4.0 International License.
VRST '26, Sendai, Japan
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2811-2/2026/11
https://doi.org/10.1145/3822517.3848697

1 Introduction

3D assembly tasks are common across multiple domains, e.g., manufacturing, construction, and digital fabrication [1], as they involve

making one object from its individual components. Modern Virtual Reality (VR) systems have emerged as a powerful medium for 3D assembly tasks, as users can create and manipulate 3D structures in a body-centric space [36, 54]. VR systems also utilize gestural interactions, which allow users to position, rotate, and connect components in a natural way [36, 43], resembling the physical manipulation people rely on when reasoning about spatial configurations [50]. Despite the affordances of VR systems, assembling objects in 3D remains cognitively demanding, as users must maintain object orientations, visualize unseen configurations, judge spatial relationships, and verify valid connections between parts [6, 8, 11, 52]. These operations place high demands on the users' spatial reasoning [46].

Past research in spatial cognition suggests that reasoning during spatial tasks often becomes visible through physical interaction [3, 52]. For example, when people solve spatial problems, they frequently perform subtle hand movements to help externalize intermediate reasoning steps, such as rotating a piece in mid-air to inspect it from different viewpoints or bringing two pieces together to compare them [61, 62]. Rawat and Barrera [52] describe these behaviours as *Spatial Hand Actions* (SHA), spontaneous manipulation movements that implicitly express ongoing spatial reasoning processes during design and assembly tasks. In other words, VR systems supporting 3D assembly tasks should allow users to coordinate perception, action, and imagined transformation through hand actions. Yet prior work incorporated visual guidance techniques such as ghost previews, alignment widgets, grids, and snap points, invoked via menus, controllers, or gaze, to expose geometric relationships during manipulation [7, 59, 69]. Other domains, such as navigation, search, and manual tasks, have explored adaptive guides that use gaze, proximity, or inferred context to aid spatial thinking [38, 40, 57, 71]. However, these visual aids rarely aligned with the embodied hand actions through which users externalize spatial reasoning while solving 3D assembling tasks. Existing VR systems rarely treat hand actions as signals of ongoing spatial reasoning and they interpret hand movements only as manipulation commands rather than as observable correlates of cognitive activity [10, 36, 43].

This paper presents **Spatial Visual Guides** (SVG), a behaviour-timed user interface for 3D assembly in VR triggered by SHA [52]. SVG provide minimal, context-sensitive cues that externalize reasoning-relevant information for three core spatial processes used during assembly: orientation, visualization, and relations (see Figure 1). By grounding activation in naturalistic hand actions, SVG avoid requiring users to learn symbolic gesture vocabularies or navigate interface menus [66], while reducing the interaction overhead associated with command-based guidance systems [30].

We implemented the **Spatial Visual Guides** (SVG) in two user interfaces, **Hand-SVG** and **VoiceHand-SVG**, with the same guides' geometry, timing parameters, and spatial computations, but with different invocation methods, reflecting two common multimodal design trade-offs: continuous, implicit assistance versus deliberate, on-demand control. **Hand-SVG** is an unimodal interface that triggers SVGs directly via SHA, allowing guidance to emerge from users' ongoing embodied spatial reasoning. **VoiceHand-SVG** is a multimodal interface in which users activate SVGs by combining a corresponding SHA with a single-word voice command. Voice commands provide an explicit channel for declaring intent while keeping the hands engaged in object manipulation, aiming to support deliberate control and reduce unintended activations through disambiguation of intent [47, 65].

We evaluated SVG in a repeated-measures user study (N=24) in which participants completed a 3D assembly task using three interfaces (**No Visual Feedback**, **Hand-SVG**, **VoiceHand-SVG**) at two difficulty levels (*hard* vs. *easy*). We measured participants' articulated spatial reasoning strategies, guide usage behaviour, and user performance and experience. Our findings revealed that **VoiceHand-SVG** triggered fewer but more sustained guide activations, keeping guides visible for a comparable period of the trial despite a far lower activation count, suggesting more selective use of assistance during moments of uncertainty. Strategy articulation further indicated that SVG supported participants' reasoning about spatial relationships by making orientation references, attachment compatibility, and placement alternatives visible during manipulation. Finally, SVG improved assembly performance over **No Visual Feedback**, as **Hand-SVG** achieved higher completion, while **VoiceHand-SVG** reduced errors.

Our main contributions are twofold. First, we couple guide invocation for 3D assembly tasks with natural SHA [52], so that guidance appears at moments when the user's behaviour suggests an ongoing spatial reasoning process. Second, we examine whether each guide supports the specific spatial-reasoning process it targets, such as orientation, visualization, and relational reasoning, using behavioural logs and participants' articulated strategies. We claim that the SVGs scaffold specific spatial-reasoning behaviours during assembly and improve task outcomes. Our evidence for the former comes from participants' articulated strategies and behavioural logs, and for the latter from performance measures. Our findings indicate that guidance grounded in natural hand actions can scaffold spatial reasoning during 3D assembly, and that adding an explicit multimodal control interaction alters when and how users choose to invoke that guidance.

2 Related Work

2.1 Spatial Thinking and 3D Design

Spatial ability refers to a set of cognitive processes that allow individuals to reason about the structure, orientation, and relationships of objects in 3D space [17, 29, 37]. Across disciplines, many frameworks converge on three core spatial factors: spatial visualization to transform objects mentally, spatial relations to judge relative alignments, and spatial orientation to maintain consistent reference frames when interacting with 3D objects [32, 52]. Existing ways to measure them include the Mental Rotation Test (MRT) [41] and the Purdue Spatial Visualization Test: Rotations (PSVT:R) [9], but these tests quantify performance outcomes rather than capturing how spatial reasoning unfolds during a task. In HCI, past work has analyzed users' manipulation behaviour, gesture use, and verbalized strategies during spatial problem-solving tasks and design activities [2, 33, 51, 61], as spatial reasoning influences user performance in creative and design-oriented tasks [6, 15].

Other past work examines how spatial reasoning becomes visible through bodily action. Studies in embodied cognition show

that people frequently externalize spatial reasoning through spontaneous hand movements during problem solving [2, 52, 62]. These movements reflect intermediate cognitive operations such as mental rotation, comparison, and spatial structuring. This work treats hand actions as observable behavioural signals of reasoning rather than as direct read-outs of internal cognitive state, a distinction we preserve throughout. Rawat and Barrera [52] formalized this phenomenon in the context of 3D design interaction by identifying SHA, naturalistic hand movements that emerge while users reason about spatial relationships during design tasks. Yet, most VR interfaces treat hand tracking as a manipulation channel rather than as a signal of cognitive activity. Consequently, systems rarely leverage spontaneous hand behaviour as triggers for delivering context-sensitive guidance. Our work treats hand actions as indicators of when users may benefit from spatial guidance, delivering visual support at moments when spatial reasoning is occurring.

2.2 Visual Guidance for 3D Interaction

Many immersive systems incorporate visual guidance techniques to reduce spatial ambiguity during interaction. Examples include ghost models that preview target configurations [44], grid-based references that provide spatial structure in the scene [26, 59], alignment widgets that help users orient objects relative to axes or planes [22], and path-based hints that guide users toward correct matching operations [5, 7]. These approaches demonstrate that visual guides can reduce cognitive effort associated with alignment, positioning, and structural uncertainty during 3D manipulation tasks. See the supplementary material for a structured comparison among these guides.

Other work moves beyond persistent or manually invoked aids toward adaptive cueing, in which the system decides what to show and when, based on task state or user behaviour. For example, Liu et al. [38] dynamically update the set of cues shown for a bimanual, unordered placement task in VR, finding that updating cues based on hand proximity, especially on eye gaze, improves performance over a static cue set. Seeliger et al. [57] use machine learning to make navigation depend on their spatial ability. Lystbæk et al. [40] place spatial gaze markers to help users resume interrupted tasks in AR task switching. These systems demonstrate the value of guidance on behavioural or contextual signals. However, they mostly target navigation, search, ordered placement, and guidance based on gaze or proximity, rather than the embodied hand actions that externalize spatial reasoning during assembly. Our work complements these by coupling invocation to SHA and by examining whether the resulting guidance supports specific spatial-reasoning processes, not only task completion.

2.3 Multimodal Interfaces for Spatial Interaction

Multimodal interfaces combine multiple interaction modalities such as hands, voice, gaze, or tangible inputs, to support flexible and expressive interaction with virtual environments [19]. Usually, to complement affordances, e.g., hands provide continuous spatial manipulation and embodied control, whereas voice enables efficient high-level commands that reduce manual workload and interaction complexity [72]. Such combinations often mirror real-world creative workflows, where users naturally shift between physical manipulation and higher-level verbal reasoning [53]. In immersive 3D settings, prior systems pair voice with hand gestures to improve efficiency. For example, by combining spoken mode switching with manual manipulation in CAD modeling [23] or by letting users request tools without navigating hierarchical menus [20]. However, multimodal designs can also introduce inconsistent mappings, ambiguous command timing, or increased cognitive load when modalities are poorly coordinated [16, 64], so they must balance embodied interaction with deliberate control during complex spatial tasks [30, 68]. These tensions motivate our use of multimodal guidance that aligns with spatial reasoning processes and is triggered by users' embodied actions.

3 Spatial Visual Guides

When designing **Spatial Visual Guides** (SVG), our goal was to align visual guidance with behaviours that externalize three spatial thinking factors: orientation, visualization, and relations [29, 32, 52]. Users trigger each guide by a corresponding SHA that physically manifests an ongoing spatial reasoning process [52]. By coupling guide activation with naturalistic actions, SVG deliver context-sensitive assistance without requiring users to navigate interface menus or recall explicit commands. Figure 2 makes this mapping explicit by pairing each SHA with the guide it triggers and the spatial-reasoning process that guide is intended to support.

We also aim to keep SVG visually minimal and ephemeral, consistent with recommendations for reducing visual clutter and cognitive load in immersive interfaces [4]. Rather than relying on persistent overlays or manually activated tools, users activate one guide at a time using behavioural-timed triggering to provide contextual feedback only when relevant reasoning actions occur [7, 14, 21, 59]. The detection thresholds (angular velocity, grasp strength, proximity, and pause duration) are based on past work and iterative pilot testing to prevent error activation during incidental motion. Next, we describe the four SVG guides. Figure 2 illustrates the SHA and their corresponding visual guides.

ReferenceFrame (Orientation). Orientation reasoning requires users to mentally rotate objects while keeping track of their principal reference axes during manipulation, since objects' own axes turn with it and users can lose their orientation bearings mid-rotation [29, 32, 58]. The *ReferenceFrame* guide externalizes this rotational structure by rendering a local coordinate frame (see Figure 1) aligned with the object's principal axes to provide a coordinate frame. Users trigger the *ReferenceFrame* guide by doing an orientation SHA, which is detected when the angular velocity exceeds $\omega_{\text{on}} = 20°/\text{s}$. Once activated, the guide updates at 60 Hz using principal component analysis to compute the object's principal axes. Exponential smoothing is applied to reduce visual jitter during rapid hand motion. The guide retracts when angular velocity drops below 8°/s. By making the object's orientation visible during manipulation, *ReferenceFrame* helps users maintain consistent alignment and reduces unnecessary corrective rotations.

ConnectionIndicator (Visualization). 3D Assembly tasks require users to infer compatible surfaces or connectors before attempting

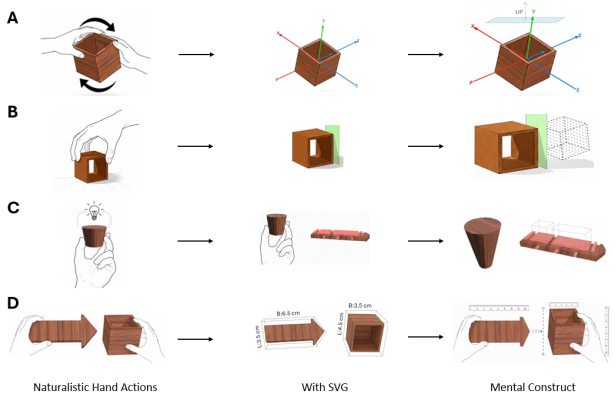


Figure 2: Mapping from SHA to SVG. First column: Naturalistic hand actions—(A) rotation triggers *ReferenceFrame*, (B) grab triggers *ConnectionIndicator*, (C) mental-image triggers *Blinker*, and (D) comparison triggers *Ruler*. Second column: Visual guides. Third column: Mental constructs.

placement. The *ConnectionIndicator* guide supports spatial visualization during the search for valid relationships between parts, as without explicit cues, this reasoning leads to exploratory trial-and-error contact attempts [8, 11]. This guide highlights predefined snap points on the held object where a connection can be made (see Figure 1). It does not indicate which pieces mate with them, so users must still find the matching part. Richer encodings such as directional arrows, volumetric pointers, or animated affordance cues are possible. However, we chose a plain highlight to keep the guide minimal and low-occlusion. Users trigger the guide by doing a grabbing SHA, which is detected when the user grasps an object with pinch or grab strength ≥ 0.8. By externalizing structural compatibility information, *ConnectionIndicator* reduces the need for users to simulate attachment possibilities mentally and can decrease invalid placement attempts.

Blinker (Visualization). Mennie et al. [42] found that people make look-ahead fixations toward objects they will manipulate, on average, roughly 3 seconds before acting, and that these anticipatory behaviours serve a planning role. In other words, users pause their movements to consider where an object might attach before committing to an action. The *Blinker* guide supports spatial visualization via a mental-image SHAs during these planning and anticipatory stops, as these pauses reflect internal spatial-visualization processes in which users evaluate alternative configurations [15, 62]. When a user holds an object with minimal motion (translational < 1.5 cm and rotational $< 10°$) for at least three seconds, we display the *Blinker* guide as flat placeholders blinking at 0.5 Hz at every candidate placement location on the base object (see Figure 1). The placeholders do not convey the shape or identity of the piece that belongs at each location. They let users evaluate alternatives without having to hold multiple imagined configurations in working memory.

Ruler (Relation). During assembly tasks, users judge relative sizes, distances, and component alignment. These relational comparisons often rely on approximate perceptual estimation, which can introduce uncertainty during manipulation [6, 61]. The *Ruler* guide supports spatial relational reasoning by helping users compare object dimensions and spatial relationships using a comparison SHA. When users hold two objects and approach them within 8 cm with a relative velocity below 2 cm/s, the *Ruler* guide appears around both objects. It renders bounding boxes, dimensional labels, and live positional differences ($\Delta x, \Delta y$) between objects. By allowing users to inspect differences in size and alignment, the *Ruler* guide externalizes geometric relationships that would otherwise require mental estimation. The guides differ in what they reveal: *ReferenceFrame* and *Ruler* present target-independent geometric information, whereas *ConnectionIndicator* and *Blinker* indicate where an attachment or placement is possible without identifying which piece belongs there or in what orientation, so selecting and orienting the correct part remains the user's reasoning process.

3.1 Proposed User Interfaces

We implemented the SVG in two interfaces that differ in invocation method but have identical guide geometries, spatial computations, update rates, and detection thresholds, as guidance invocation can shape how users utilize assistance in their problem-solving: continuous automatic assistance supports fluid interaction but risks unintended activations, whereas deliberate invocation gives users more control over when guidance appears [33, 34].

Hand-SVG. This interface activates guides automatically when the system detects a corresponding SHA. This unimodal design aims to preserve the flow of interaction by eliminating explicit command recall or interface navigation. Prior research suggests that automated assistance can maintain task continuity when trigger conditions align with natural user behaviour [31, 45]. By coupling guide activation with spontaneous spatial actions, users receive guidance precisely when spatial reasoning is already occurring.

VoiceHand-SVG. This interface introduces multimodal invocation to enable deliberate control over guidance activation. A guide is triggered only when two conditions are met simultaneously: the system detects the corresponding SHA and the user speaks a voice command associated with that spatial operation (*rotate*, *compare*, *grab*, or *imagine*). We selected these four words from the verbs participants used to describe the corresponding actions in prior SHA analysis [52] and confirmed them in a pilot study. *imagine* names the mental-imagery act that the planning pause externalizes and was retained because it is short and phonetically distinct from the other three. Voice acts as an explicit confirmation signal layered on top of embodied interaction. While the user continues manipulating objects with their hands, speaking the command indicates an intentional request for guidance. This multimodal gating mechanism reduces unintended activations and allows users to request assistance when they perceive a need for support.

Prior work suggests that verbalizing spatial intent can assist reasoning by externalizing cognitive processes [20]. The design also draws inspiration from think-aloud protocols, where articulating thoughts can support reflection during complex tasks [52]. By combining voice with hand-based interaction, the system enables

users to regulate guidance activation while maintaining embodied manipulation of objects.

3.2 Implementation

We implemented SVG in Unity (C#) using Unity 2022.3.46f1 with the Meta XR All-in-one SDK v57 on a Windows 11 workstation (RTX 4090, Intel i9-13900K). SHA were detected using the Quest 3 hand-tracking API, which combines Unity XR Hands for joint-pose tracking and Meta XR Core for device integration. We used these signals to compute hand motion, pinch strength, and object-interaction states that were required to trigger the four SVG guides.

For the **VoiceHand-SVG** interface, we implemented voice commands using a controlled Wizard-of-Oz (WoZ) procedure. During the study, a researcher pressed one of four mapped keys (Q/W/E/R) only after verifying that the participant had spoken the corresponding spatial keyword while manipulating the associated object with natural hand actions. WoZ ensured consistent detection timing across participants while keeping guide behaviour and visual feedback constant across both interfaces.Because a researcher confirmed each command, the procedure eliminated recognition errors and latency. The **VoiceHand-SVG** results represent an idealized upper bound for a deployed recognizer (Section 8). A guide was triggered only when the correct keyword accompanied its SHA, so a forgotten or mismatched word produced no activation. We did not systematically log such attempts, but participants' reports of recall effort are examined in the thematic analysis.

4 User Study

We conducted a controlled user study to evaluate the effect of SVG on spatial reasoning, task performance, and user experience during a 3D assembly task in VR. The study was approved by the Dalhousie University Research Ethics Board (REB #2023-6542).

4.1 Research Questions

Our research questions (RQ) are motivated by prior work on embodied cognition, gesture-supported reasoning, and multimodal interaction in immersive environments [2, 10, 62]. Our goal is to isolate the effect of behaviour-timed SVG, and to compare two invocation strategies for the same guides, not to outperform a specific prior guidance system. Thus, rather than evaluating spatial ability itself, we examine how different interaction mechanisms shape users' spatial reasoning strategies and behaviours during 3D assembling tasks in VR. Our RQ are:

RQ1 - Spatial Thinking: *How do SVG influence the ways users reason about spatial relationships during 3D assembly tasks in VR?* Prior research shows that spatial reasoning is frequently externalized through physical gestures and manipulation behaviours [2, 62]. We investigate whether coupling guidance with spatial hand actions changes how users describe and apply reasoning strategies such as mental rotation, relational comparison, and placement planning.

RQ2 - User Behaviour: *How do users integrate SVG guidance into their assembly strategies when guidance is tied to embodied interactions?* Embodied interaction research suggests that physical actions can reveal underlying reasoning processes during spatial tasks [52]. By examining guide usage patterns and interaction behaviours, we aim to understand how users incorporate behaviour-timed guidance into their assembly workflow.

RQ3 - Performance and User Experience: *Do the SVG interfaces improve task performance, and how do unimodal (**Hand-SVG**) and multimodal (**VoiceHand-SVG**) activation differ in perceived workload, usability, and preference?* This research question establishes that users can utilize our guides as designed and characterizes their performance trade-offs compared to not using visual guides.

4.2 Methodology

4.2.1 Participants. We recruited 24 participants (11 female, 13 male) aged between 18 and 31 years (M = 22.92, SD = 3.53). Twenty-three participants were right-handed, while one was left-handed. Participants reported varied experiences with spatial assembly activities such as Lego or puzzle construction. Nineteen participants identified themselves as beginners, one as intermediate, and four as experienced. Regarding prior exposure to 3D design tools, seventeen participants had no previous experience, six reported one to two years of experience, and one reported more than three years. Participants also reported varying levels of VR experience: eight had never used VR, seven had used VR fewer than five times, and nine had used VR more than five times. Eighteen participants had no prior experience assembling 3D objects in VR environments.

4.2.2 Evaluated Interaction Techniques. Following established evaluation practices in 3D user interface research [7, 10, 36, 59, 60], we evaluate **Hand-SVG** and **VoiceHand-SVG**, against a no-guidance baseline (**No Visual Feedback**).

No Visual Feedback (Baseline). Participants assembled objects using natural hand-grabbing and rotation without visual guidance, allowing us to test whether the guides help users assemble 3D objects in VR. We intentionally adopt this baseline to isolate the effect of SVG guidance on spatial reasoning during assembly, a common approach when evaluating new interaction assistance mechanisms [18, 70]. Moreover, direct comparison with existing VR guidance systems would introduce confounding factors because prior techniques differ substantially in task domain, invocation mechanisms, and visual persistence, see subsection 2.2.

Hand-SVG (Automatic SVG activation) and VoiceHand-SVG (Voice + SHA activation). These are the two SVG interfaces introduced in Section 3 that use show visual guides but differ in their activation methods: **Hand-SVG** triggers guidance automatically when the system detects a corresponding SHA, whereas **VoiceHand-SVG** requires the user to pair the SHA with a single-word voice command. By comparing **Hand-SVG** and **VoiceHand-SVG**, we can isolate the effect of guidance invocation on 3D assembly tasks.

4.2.3 Experimental Design. We used a within-subject design with two independent factors: *Interface* (**Hand-SVG**, **VoiceHand-SVG**, **No Visual Feedback**) and *Task Difficulty* (**Easy**, **Hard**). Each participant completed six assembly trials (3×2 conditions). Interface order was counterbalanced using a 6×6 Latin square to mitigate sequence and learning effects. Also, the initial positions and orientations of the pieces were randomized for every condition to reduce direct motor learning between trials. Participants assembled two versions of a train model: a 16-piece model for the easy condition

and a 24-piece model for the hard condition, as shown in Figure 3. The hard version subdivides several components into smaller parts, increasing the number of placement/rotation/attachment operations while keeping the overall target geometry constant. Difficulty was manipulated through part count only. Symmetry, the number of candidate connections per piece, and structural branching were held constant. Both difficulties yielded the same final object configuration, ensuring that differences in performance reflected spatial reasoning rather than geometric differences. Participants were never shown a picture or model of the finished train (Figure 3 is for the reader). They were asked to build a final object from the pieces and had to infer which parts connect and in what orientation. Therefore, the task is goal-directed with a single valid final configuration whose solution was never displayed. A snap counts as correct when a piece attaches to its intended mating piece at the intended connection. An attempted connection between non-mating pieces counts as an error. Both SVG interfaces provide support as soon as the triggering SHA is detected (**Hand-SVG**) or requested (**VoiceHand-SVG**); neither reserves an exploration period before guidance becomes available.

4.2.4 Procedure. Participants completed the study in a quiet room with only the experimenter present. Upon arrival, they signed a consent form and completed a pre-study questionnaire capturing demographic information and prior VR experience. Participants then received a system overview and calibration of the VR headset. A 10-minute training session followed, in which participants practised the four SHA and experienced each interface condition. After training, participants completed six assembly trials corresponding to the experimental conditions. Each trial lasted up to five minutes. Participants completed NASA Task Load Index (NASA-TLX) [27] and System Usability Scale (SUS) [13] questionnaires after each interface condition. Five-minute breaks were provided between conditions to reduce fatigue and motion discomfort. After completing all conditions, participants participated in a semi-structured interview lasting approximately 15 minutes. The entire study session lasted for 90 minutes.

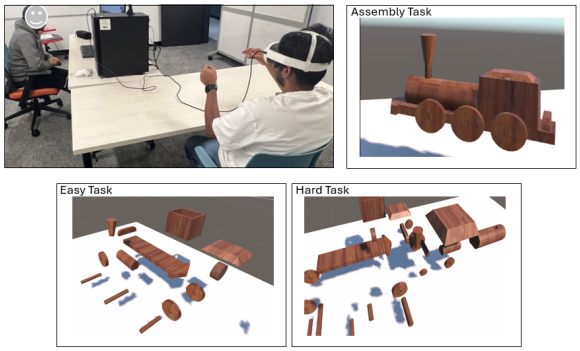


Figure 3: User study setup and task. Top left: Participant performing the study. Top right: Train model task. Bottom: Initial piece layouts for easy (left) and hard (right) conditions.

4.2.5 Evaluation Metrics. To address our RQs, we collected four categories of measures.

Table 1: ART-ANOVA results for per-guide activation counts (Grab, Rotate, Compare) and correct snapping, by effect (Interface, Difficulty, Interface × Difficulty).

Measure	Effect	F	p	η_p^2
Behavioural metrics (SVG-triggered interaction)				
Grab	UI	343.86	<0.001	0.833
	UI × Difficulty	4.28	0.042	0.058
Rotate	UI	39.53	<0.001	0.364
Compare	UI	73.21	<0.001	0.515
	Difficulty	17.07	<0.001	0.198
	UI × Difficulty	13.28	<0.001	0.161
Correct Snapping	Difficulty	10.30	0.002	0.130

Spatial Thinking Measures. To examine how participants reasoned about spatial relationships during 3D assembly, we conducted semi-structured interviews after each interface condition and at the end of the study, following the protocols used in prior work on spatial reasoning in design tasks [52]. We used two qualitative analyses: a) **Strategy Articulation**: Participants explained their placement decisions and tied individual guides to specific spatial thinking processes. We used this data to identify whether SVG supports specific reasoning processes. The first author coded these explanations and the second author reviewed each iteration, applying a deductive codebook of three spatial-thinking components: visualization (mental imagery), orientation (reference framing), and relations (comparative reasoning) [2, 52]. b) **Thematic Analysis**: Using Braun and Clarke's method [12], the first author led the initial coding, and the second author reviewed each iteration, contributing codes and surfacing gaps in interpretation. Following a consensus-based reflexive approach, codes and themes were developed and refined through iterative discussion.

Behavioural Metrics. For SVG conditions (**Hand-SVG**, **VoiceHand-SVG**), system logs captured guide usage: a) **Guide Activation Frequency**, number of SVG activations per trial; b) **Time-on-Guide**, the proportion of trial time during which at least one guide was visible; c) **Observed Assembly Strategies**, describing how participants incorporated guides, for example, exploratory versus targeted use.

Performance and Experience Measures. To confirm that the guides worked and to characterize their experiential trade-offs (RQ3), we also recorded task performance and standard subjective measures: **Task Completion (%)**, **Error Count**, and **Error Rate** from the system logs, together with the **NASA-TLX** for perceived workload, the **SUS** for usability, and **Preference Ranking** with open-ended responses collected after each interface.

4.2.6 Analysis. We analyzed each dependent variable with a model matched to its measurement structure. Task Completion was analyzed with a two-way repeated-measures ANOVA (RM-ANOVA) on Interface × Difficulty, with Bonferroni-corrected pairwise comparisons. Error Count, Error Rate, and the per-guide activation counts were not normally distributed and were analyzed with the Aligned Rank Transform ANOVA (ART-ANOVA) [67] on Interface × Difficulty. Time-on-Guide was compared between the two SVG

interfaces because its residuals departed from normality (Shapiro-Wilk, $p < .05$). We report an ART-ANOVA as the primary test and complement it with a paired comparison and a Bayes factor to quantify evidence for equivalence. NASA-TLX and SUS were administered once per interface condition and were therefore analyzed with a one-way RM-ANOVA on Interface (three levels). For all RM-ANOVAs, we assessed sphericity with Mauchly's test and applied the Greenhouse-Geisser correction when it was violated [25]; we report η^2 as the effect size and use $\alpha = .05$ throughout. We decomposed an Interface × Difficulty interaction into simple effects only where the interaction was statistically significant. A sensitivity analysis (G*Power 3.1, RM-ANOVA with three within-subject levels, $\alpha = .05$, $1 - \beta = .80$, $\rho = .5$) indicates that $N = 24$ can detect effects of $f \geq 0.27$ ($\eta_p^2 \approx .07$), well below the effects observed for our primary measures. We treat the omnibus tests on completion, errors, workload, usability, and total activations as confirmatory, with Bonferroni-corrected pairwise comparisons. The per-guide breakdowns and difficulty-driven shifts are exploratory and interpreted with caution, given the sample size.

5 Results

5.1 User Behaviour

5.1.1 Feedback Usage by Activation Strategy. We analyzed SVG logs to characterize how participants used guidance under **Hand-SVG** and **VoiceHand-SVG**. At a higher level, **VoiceHand-SVG** substantially reduced the frequency of guide activations relative to **Hand-SVG** (grab: 83% reduction; rotate: 36% reduction; compare: 52% reduction), while mental imagery feedback showed no significant modality difference. This frequency gap was large as **Hand-SVG** generated 6.2× more grab activations (82.83 vs 12.85, $p < .001$) and 16× more comparison activations in hard tasks (34.96 vs 2.17, $p < .001$). The full per-guide activation breakdown is shown in the supplementary material. Beyond overall counts, the temporal profile of activations differed between interfaces: Figure 4A plots when guides were invoked as a function of assembly progress. Both interfaces concentrated activations early in assembly. **Hand-SVG** then sustained frequent activations later into the build, while **VoiceHand-SVG** tapered to sparse, targeted requests.

5.1.2 Time-on-Guide. Activation counts describe how often guidance was requested, but not how long it remained available. To distinguish reduced activation from reduced guidance, we computed time-on-guide, the proportion of each trial during which at least one guide was visible. Although **VoiceHand-SVG** reduced activation frequency by 5-6× relative to **Hand-SVG**, two interfaces kept a guide visible for a similar time of the trial (**Hand-SVG** 72% easy, 77% hard; **VoiceHand-SVG** 79% easy, 77% hard). An ART-ANOVA on time-on-guide found no significant difference between **Hand-SVG** and **VoiceHand-SVG**, and a paired comparison was likewise non-significant ($t(23) = -1.31$, $p = .61$); a complementary Bayes factor favoured the null hypothesis of no difference ($BF_{01} = 2.2$). Figure 4B shows the per-guide breakdown of time-on-guide for two interfaces. In other words, the two SVG interfaces delivered an equitable amount of cumulative guidance through different temporal patterns as **Hand-SVG** produced many brief activations consistent with continuous verification, whereas **VoiceHand-SVG** produced fewer but more sustained activations consistent with deliberate, preview-then-execute use. This indicates that the smaller number of **VoiceHand-SVG** activations reflects a difference in how guidance was invoked rather than a reduction in the guidance available, addressing the alternative explanation that fewer activations mean less assistance or higher interaction overhead.

5.1.3 Difficulty-Driven Behavioural Shifts. Difficulty significantly increased comparison-guide use, with **Hand-SVG** showing a 126% increase in hard trials versus **VoiceHand-SVG**'s 67% increase. The Interface × Difficulty interaction was significant for both comparison ($F = 13.28$, $p = .001$, $\eta^2 = 0.16$) and grab feedback ($F = 4.28$, $p = .042$, $\eta^2 = 0.06$), indicating that continuous automatic activation led to stronger reliance on guides under higher complexity, as shown in Figure 5.

5.1.4 Snapping Outcomes. Despite fewer activations, **VoiceHand-SVG** maintained snapping outcomes comparable to **Hand-SVG** (completion %: **Hand-SVG** 53.35 vs **VoiceHand-SVG** 51.06, $p = .509$). Notably, in hard trials, **VoiceHand-SVG**'s error rate increased less than **Hand-SVG** (10.6% vs 14.3%, $p = .010$).

5.2 Order and Learning Effects

We mitigated order effects via a 6×6 Latin-square counterbalancing of trial order and randomized initial piece layouts and orientations in each trial to reduce direct transfer of motor sequences. Yet, each participant completed six timed trials (three interfaces × two difficulties), allowing them to practice the task.

Rather than treating learning as a confounding variable, we modeled it to test whether SVG affected performance beyond the effects of repetition. We fit a mixed-effects model with participant as a random intercept and TrialNumber as a covariate (see supplementary material). TrialNumber was significant ($\beta = 5.379$, $p < .001$), confirming a positive learning slope across repeated trials. Crucially, both SVG-enabled interfaces remained significantly higher than the no-guidance after controlling for TrialNumber (**Hand-SVG**: $\beta = 18.437$, $p < .001$; **VoiceHand-SVG**: $\beta = 16.146$, $p < .001$), indicating that the observed interface effects were not explained by trial order alone.

We then tested whether the rate of improvement differed by interface using an interaction model. The interface × TrialNumber interaction showed that learning trajectories were not uniform across conditions: **VoiceHand-SVG** exhibited a significantly steeper improvement slope than **No Visual Feedback** ($\beta = 5.732$, $p = .006$), and **Hand-SVG** showed a similar but weaker trend ($\beta = 3.738$, $p = .071$), as shown in Figure 4C.

5.3 Performance and User Experience

Task Performance. Both SVG interfaces improved completion over **No Visual Feedback** (**Hand-SVG**: 53.35%, **VoiceHand-SVG**: 51.06%, **No Visual Feedback**: 34.85%; all $p < .001$). The two SVG interfaces differed in where the gain appeared: **Hand-SVG** was strongest on easy trials, whereas **VoiceHand-SVG** held steadier across difficulty (a 7.0-point easy-to-hard drop, versus 15.8 for **Hand-SVG**). Errors showed a complementary pattern; an ART-ANOVA found **VoiceHand-SVG** produced fewer errors than both

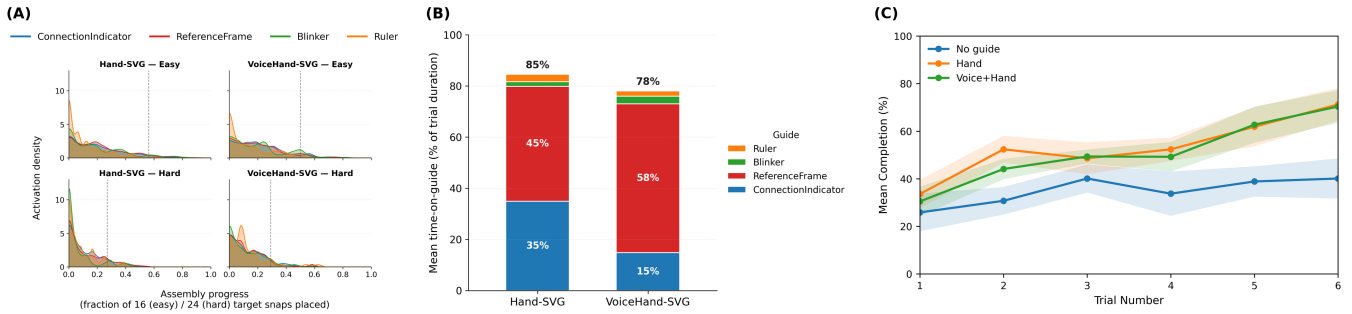


Figure 4: Behavioural results across interface conditions. (A) Guide-activation density over assembly progress for HAND-SVG and VOICEHAND-SVG in easy and hard tasks. (B) Mean time-on-guide (proportion of trial time with at least one guide visible). (C) Mean completion across the six repeated trials by interface

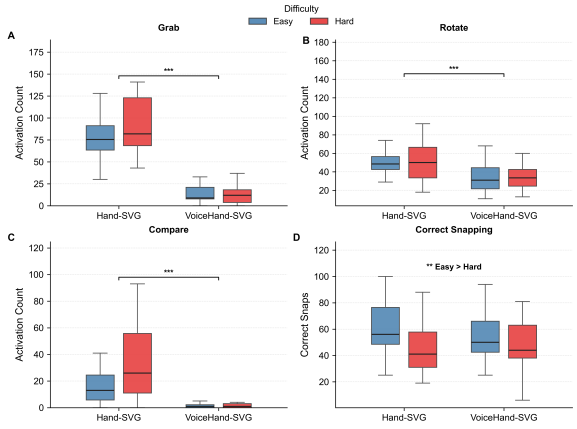


Figure 5: Per-guide SVG activation behaviour: (A) Grab activations, (B) Rotate activations, (C) Compare activations, and (D) Correct snapping.

NO VISUAL FEEDBACK and **HAND-SVG**, with its advantage concentrated in hard trials. In short, **HAND-SVG** maximized how much participants completed, while **VOICEHAND-SVG** protected accuracy under load.

User Experience. Both SVG interfaces also lowered perceived effort. Workload (NASA-TLX) differed across interfaces ($F(2, 46) = 38.4$, $p < .001$), with **VOICEHAND-SVG** the least demanding, and usability (SUS) followed the same ordering (**HAND-SVG** 82.4, **VOICEHAND-SVG** 76.8, **NO VISUAL FEEDBACK** 58.2; $p < .001$). Preference was split ($\chi^2(2) = 10.75$, $p = .005$): 14 of 24 participants preferred **HAND-SVG** for its immediacy, 9 preferred **VOICEHAND-SVG** for control, and 1 preferred **NO VISUAL FEEDBACK**. The two interfaces suited different priorities rather than one dominating. Full statistics and the task-performance figure appear in the supplementary material.

5.4 Strategy Articulation

Across SVG-enabled interfaces, participants described structured reasoning strategies that externalized intermediate mental constructs, thereby supporting the mapping from each SVG to the spatial-reasoning process it targeted. For example, 14 participants described using *Ruler* as a relational scaffold to verify alignment before placing the object ("check align before committing" (P7); "fit in my head before snapping" (P12)), consistent with spatial-relation reasoning during assembly [24]. Participants also described *ReferenceFrame* as a "mental anchor" (P15) for maintaining consistent orientation across manipulations, and *ConnectionIndicator* as a way of reducing uncertainty about valid attachment affordances by "bringing distant pieces into play"(P10). 15 participants described *Blinker* as prompting mental rehearsal of placements ("I pictured it as a vehicle before I built it"(P19)). In contrast, when using **NO VISUAL FEEDBACK**, 19 participants described guess-check behaviour and repeated viewpoint switching ("I kept guessing until it fit"(P4)). These statements connect each guide to a distinct reasoning process, providing direct evidence that the guides supported the targeted spatial-thinking factors rather than offering undifferentiated assistance.

The two invocation strategies shaped how this support entered the reasoning process. **HAND-SVG** supported continuous monitoring and realignment, while **VOICEHAND-SVG** led to a preview-then-execute pattern, in which participants requested guidance selectively before committing. These differences align with embodied accounts of spatial reasoning in VR, where users distribute inference across perception, action, and external representations. SVG provided timely, minimal visual scaffolds that participants incorporated into their reasoning workflow rather than functioning as a persistent instructional overlay.

5.5 Thematic Analysis

Two researchers identified six codes and grouped them into three themes. These themes capture how participants experienced SVG as spatial support, how they perceived the trade-offs between **HAND-SVG** and **VOICEHAND-SVG**, and how their guide-use strategies changed with task difficulty.

Theme 1: SVG as Spatial Scaffolds for Comprehension. 16 participants explicitly linked SVGs to improved understanding of object relationships and reduced uncertainty during assembly. Participants described SVGs as making spatial constraints visible at the moment of need. For example, describing **Hand-SVG** as *"a 3D compass" (P13)*, especially during alignment or connection uncertainty, where **No Visual Feedback** led to frustration *"I wasted minutes trying impossible connections"(P17)*. These reports are consistent with SVGs functioning as scaffolds for reasoning strategies in the 3D assembly task.

Theme 2: Workflow Tradeoffs Between Automatic and On-demand Activation. Participants described **Hand-SVG** as low-friction and continuous *"Guides appeared instantly with my hand movements—no extra thinking"(P14)*. At the same time, **VoiceHand-SVG** provided targeted help *"Could request exact feedback when stuck"(P15)* but added recall overhead for some *"Remembering keywords distracted me from the puzzle"(P8)*. Participants who noted this recall cost valued deliberate triggering for keeping the scene uncluttered, a tradeoff also visible in the logs and interviews. This tradeoff mirrored the behavioural logs, with frequent guide usage under **Hand-SVG** vs selective usage under **VoiceHand-SVG**.

Theme 3: Adaptation Under Task Difficulty. Participants adjusted strategies as task difficulty increased. In easy trials, some reported internalizing patterns and reducing guide reliance *"After a few snaps, I internalized the patterns"(P11)*. In hard trials, **Hand-SVG** users relied on continuous feedback. In contrast, **VoiceHand-SVG** users reported needing to pause and articulate intent *"Had to think in words instead of just moving"(P18)*. These reports align with the observed Interface × Difficulty interactions in guide use and error outcomes.

6 Discussion

We investigated whether behaviour-timed spatial visual guidance (SVG), invoked through **Hand-SVG** or **VoiceHand-SVG**, supports spatial reasoning during 3D assembly in VR. Across performance, behavioural logs, workload, and qualitative reports, the results indicate that SVG influenced not only task outcomes but also how participants externalized and verified spatial relationships during assembly.

6.1 RQ1: Spatial Thinking

Our evidence for spatial thinking support is based on participants' articulated strategies, thematic analysis and behavioural logs, which show how each guide was used and when. The performance measures confirm that the guides helped with the task (RQ3) but do not measure reasoning. Therefore, we claim that the guides scaffolded specific reasoning behaviours during assembly. The mapping from guides to spatial-thinking factors rests on design rationale and on participants' own accounts of guide use. However, it was not validated against an independent measure of spatial ability. We treat this as a deliberate, human-centred stance, rather than inferring hidden cognitive states, we ask whether the guides connected to the reasoning behaviours users actually expressed and described during assembly [33, 55]. Three questions organize our findings: which reasoning behaviours changed, how they changed over the course of assembly, and how SVG enabled that change.

First, regarding which behaviours changed, participants across SVG-enabled conditions described using guides to maintain orientation anchors, compare spatial relationships, and rehearse placements before action (subsection 5.4). These behaviours align with the spatial processes the guides were designed to support: orientation, visualization, and relational reasoning [39, 52]. For instance, participants used *ReferenceFrame* as a stable orientation anchor *"mental anchor" (P15)*, *Ruler* for relational verification *"check align before committing" (P7)*, and *Blinker* for anticipatory visualization *"I pictured it as a vehicle before I built it" (P19)*. In contrast, **No Visual Feedback** elicited iterative guess–check behaviour and repeated viewpoint switching *"I kept guessing until it fit" (P4)*. Because each quote ties a specific reasoning act, this evidence speaks directly to the claim that the guides supported a particular spatial process, not merely that scaffolding improved outcomes in general.

Second, regarding how behaviour changed over assembly, guide use scaled with task demand. Comparison-guide use rose in hard trials, and far more under **Hand-SVG** than **VoiceHand-SVG** (a 126% versus 67% increase; interface × difficulty interaction $F = 13.28$, $p = .001$, $\eta^2 = 0.16$). Participants also shifted from internalized patterns during easy trials to a heavier reliance on guidance during hard trials. This is what distributed-cognition accounts predict: people offload more spatial reasoning onto external support as internal demand grows [33]. Our results extend that account to behaviour-timed VR guidance, and show that invocation style shapes the offloading. The same guides entered reasoning at different points, depending on both task demands and invocation styles.

Third, regarding how SVG enabled these changes, the guides let users inspect spatial relationships that would otherwise need to be mentally inferred [35, 56]: *ConnectionIndicator* to examine attachment compatibility, *Ruler* to inspect dimensional differences, and *Blinker* to preview placements without having to maintain multiple imagined configurations in working memory. This is consistent with embodied and extended cognition accounts, in which external representations redistribute cognitive work across perception, action, and the environment [33]. Importantly, invocation style shaped this support: **Hand-SVG** functioned as a continuous perceptual scaffold during manipulation, whereas **VoiceHand-SVG** positioned guidance as a deliberate verification step. Since the guides themselves were held constant across interfaces, this contrast isolates the effect of invocation strategy rather than guide content; the equivalent time-on-guide across the two interfaces (Figure 4B) reinforces that the difference lies in the timing of guidance, not its amount of information provided [48].

6.2 RQ2: User Behaviour

The two activation strategies produced distinct ways of incorporating guidance into the assembly workflow. Under **Hand-SVG**, guidance appeared automatically as participants manipulated pieces, allowing them to check relationships while acting rather than pausing to request support. This matches perception-action coupling accounts, in which physical interaction structures spatial reasoning [61, 63]. Our logs extend those accounts to behaviour-timed VR guidance: automatic activation produced far denser checking than

deliberate invocation, with **Hand-SVG** triggering 6.2× more grab activations (82.83 vs. 12.85) and 16× more comparison activations in hard trials (34.96 vs. 2.17, both $p < .001$). This density is a double-edged result. Prior work warns that highly persistent overlays can compete with scene interpretation [28, 54]; our activation rates show that continuous, action-coupled guidance can reach that level, making visual load a design risk to manage rather than a benefit of automation.

In contrast, participants treated **VoiceHand-SVG** as a selective verification tool, invoking it primarily for confirmation when uncertain. Although it produced far fewer activations, snapping outcomes remained comparable, and errors decreased in difficult trials. Critically, the time-on-guide analysis shows this selectivity reduced how often guidance was invoked without reducing how long it remained available. Hence, the difference reflects invocation strategy rather than reduced assistance, mirroring multimodal findings where speech carries higher-level commands while gesture supports continuous manipulation [23, 49].

6.3 RQ3: Performance and User Experience

The performance and user experience measures confirm that the behaviour-timed guidance helped task outcomes, not only reasoning. Both SVG interfaces increased completion and reduced workload relative to **No Visual Feedback**, confirming prior 3D user interface work showing that well-matched assistance improves manipulation [10, 36]. We extend that work by separating two outcomes it usually treats together: **Hand-SVG** raised completion most, while **VoiceHand-SVG** lowered errors most in hard trials. Because both interfaces used the same guides, this split reflects invocation, not guide design, as continuous activation favored throughput and deliberate activation favored accuracy. Usability and preference echo this: **Hand-SVG** was rated most usable, **VoiceHand-SVG** least demanding, and preference divided between them. Neither strategy won outright; each suited a different priority. This motivates adaptive guidance that shifts between continuous and on-demand activation as demands change, rather than one fixed strategy.

7 Implications for Design and Research

Our findings suggest three directions for spatial guidance in immersive assembly. First, VR systems can **treat natural hand actions as signals of spatial reasoning**, interpreting SHA as lightweight queries that trigger guidance at moments of uncertainty rather than using hand tracking only as a manipulation channel [33, 61]. Second, designers can **balance continuous and on-demand guidance**. Because automatic (**Hand-SVG**) and voice-gated (**VoiceHand-SVG**) activation delivered comparable guidance, the choice between them concerns temporal control rather than how much assistance to provide [10, 36]. Third, guidance should function as **brief cognitive scaffolding**, offering short, context-sensitive cues for verifying, comparing, or previewing rather than persistent overlays [39, 52].

8 Limitations

Our study has five limitations. First, the controlled block-assembly task isolates spatial reasoning but does not capture real-world workflows such as CAD modeling, architecture, or collaborative design. Second, external validity is limited by the use of a single assembly object (a locomotive model) with difficulty operationalized by part count only; counterbalancing, randomized layouts, and explicit modeling of trial order mitigated practice effects, but participants may still have learned the model's structure, and our claims apply to goal-directed block assembly of this kind rather than to 3D assembly in general. Third, our evidence rests on observable hand actions and articulated strategies, not direct measurement of internal cognitive states; the guide-to-factor mapping was not validated against a spatial-ability measure. Fourth, **VoiceHand-SVG** used a Wizard-of-Oz procedure that removed recognition errors, latency, accent and noise sensitivity, false positives, and missed commands, so its results are an idealized upper bound; part of its lower workload and error rate may reflect error-free recognition rather than the invocation strategy itself. Finally, we compared SVG only against a no-guidance baseline, which isolates whether the guides help and how invocation matters, but not how SVG compares with adaptive visual-guidance techniques. We did not log false-positive activations, so the effect of over-guidance under **Hand-SVG** on trust and skill acquisition remains untested.

9 Conclusion

This paper introduced **Spatial Visual Guides** (SVG), behaviour-timed visual guidance triggered by spatial hand actions or voice commands when users' behaviour suggests spatial reasoning is occurring. In a study (N=24) comparing unimodal (**Hand-SVG**) and multimodal (**VoiceHand-SVG**) activation, SVG improved task performance over a no-guidance baseline, and participants tied individual guides to specific reasoning acts, such as stabilizing orientation, comparing relationships, and previewing placements, indicating the guides scaffolded the spatial processes they targeted rather than offering undifferentiated assistance. Holding the guides constant, automatic activation supported fluid embodied interaction, whereas voice activation encouraged deliberate, selective use while keeping guidance visible for a comparable share of the task. These findings position behaviour-timed, human-centered guidance as a promising direction for scaffolding spatial reasoning in immersive 3D assembly.

Acknowledgments

Research funded by NSERC Discovery Grant. Author Francisco Ortega would like to acknowledge the National Science Foundation (NSF) 2444132, NSF 2439474, NSF 2444132, NSF 2327569, NSF 2238313, and NSF 2223432.

References

[1] M. Abdullah, V. Nirmal, and M. Rahman. 2024. Comparing the Efficacy of Virtual Reality Training, Augmented Reality Instruction and Traditional Paper-Based Instruction Methods for Assembly Tasks. In *Human Factors in Design, Engineering, and Computing (AHFE Open Access, Vol. 159)*, Tareq Ahram and Waldemar Karwowski (Eds.). AHFE International, USA. doi:10.54941/ahfe1005670

[2] Martha W Alibali. 2005. Gesture in spatial cognition: Expressing, communicating, and thinking about spatial information. *Spatial cognition and computation* 5, 4 (2005), 307–331.

[3] Martha W Alibali, Robert C Spencer, Lucy Knox, and Sotaro Kita. 2011. Spontaneous gestures influence strategy choices in problem solving. *Psychological science* 22, 9 (2011), 1138–1144.

[4] Rahul Arora, Rubaiat Habib Kazi, Fraser Anderson, Tovi Grossman, Karan Singh, and George Fitzmaurice. 2017. Experimental Evaluation of Sketching on Surfaces in VR. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing*

Systems (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 5643–5654. doi:10.1145/3025453.3025474

[5] Mayra Donaji Barrera Machuca, Paul Asente, Wolfgang Stuerzlinger, Jingwan Lu, and Byungmoon Kim. 2018. Multiplanes: Assisted Freehand VR Sketching. In *Proceedings of the 2018 ACM Symposium on Spatial User Interaction* (Berlin, Germany) *(SUI '18)*. Association for Computing Machinery, New York, NY, USA, 36–47. doi:10.1145/3267782.3267786

[6] Mayra Donaji Barrera Machuca, Wolfgang Stuerzlinger, and Paul Asente. 2019. The Effect of Spatial Ability on Immersive 3D Drawing. In *Proceedings of the 2019 Conference on Creativity and Cognition* (San Diego, CA, USA) *(C&C '19)*. Association for Computing Machinery, New York, NY, USA, 173–186. doi:10.1145/3325480.3325489

[7] Mayra Donaji Barrera Machuca, Wolfgang Stuerzlinger, and Paul Asente. 2019. Smart3DGuides: Making Unconstrained Immersive 3D Drawing More Accurate. In *25th ACM Symposium on Virtual Reality Software and Technology* (Parramatta, NSW, Australia) *(VRST '19)*. Association for Computing Machinery, New York, NY, USA, Article 37, 13 pages. doi:10.1145/3359996.3364254

[8] Gökçe Elif Baykal, Ali Leylekoğlu, Sude Arslan, and Demet Ozer. 2023. Studying children's object interaction in virtual reality: a manipulative gesture taxonomy for vr hand tracking. In *Extended abstracts of the 2023 CHI conference on human factors in computing systems*. 1–7.

[9] George M Bodner and Roland B Guay. 1997. The Purdue visualization of rotations test. *The chemical educator* 2, 4 (1997), 1–17.

[10] Doug A Bowman, Ryan P McMahan, and Eric D Ragan. 2012. Questioning naturalism in 3D user interfaces. *Commun. ACM* 55, 9 (2012), 78–88.

[11] TJ Branoff and M Dobelis. 2012. The relationship between spatial visualization ability and students' ability to model 3D objects from engineering assembly drawings. *The Engineering Design Graphics Journal* 76, 3 (2012).

[12] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[13] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

[14] Lizhou Cao, Huadong Zhang, Chao Peng, and Jeffrey T Hansberger. 2023. Real-time multimodal interaction in virtual reality-a case study with a large virtual interface. *Multimedia Tools and Applications* 82, 16 (2023), 25427–25448.

[15] Senthil Chandrasegaran, Devarajan Ramanujan, and Niklas Elmqvist. 2018. How do sketching and non-sketching actions convey design intent?. In *Proceedings of the 2018 Designing Interactive Systems Conference*. 373–385.

[16] Fang Chen, Natalie Ruiz, Eric Choi, Julien Epps, M Asif Khawaja, Ronnie Taib, Bo Yin, and Yang Wang. 2013. Multimodal behavior and interaction as indicators of cognitive load. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 2, 4 (2013), 1–36.

[17] Roberto Colom, Mª José Contreras, Juan Botella, and José Santacreu. 2002. Vehicles of spatial ability. *Personality and Individual Differences* 32, 5 (2002), 903–912.

[18] Lea M Daling and Sabine J Schlittmeier. 2024. Effects of augmented reality-, virtual reality-, and mixed reality–based training on objective performance measures and subjective evaluations in manual assembly tasks: a scoping review. *Human factors* 66, 2 (2024), 589–626.

[19] Elias Dritsas, Maria Trigka, Christos Troussas, and Phivos Mylonas. 2025. Multimodal interaction, interfaces, and communication: a survey. *Multimodal Technologies and Interaction* 9, 1 (2025), 6.

[20] Ke Fang and Jing Wang. 2024. Interactive design with gesture and voice recognition in virtual teaching environments. *IEEE Access* 12 (2024), 4213–4224.

[21] Katherine Fennedy, Jeremy Hartmann, Quentin Roy, Simon Tangi Perrault, and Daniel Vogel. 2021. OctoPocus in VR: Using a dynamic guide for 3D mid-air gestures in virtual reality. *IEEE Transactions on Visualization and Computer Graphics* 27, 12 (2021), 4425–4438.

[22] George Fitzmaurice, Justin Matejka, Igor Mordatch, Azam Khan, and Gordon Kurtenbach. 2008. Safe 3D navigation. In *Proceedings of the 2008 Symposium on Interactive 3D Graphics and Games* (Redwood City, California) *(I3D '08)*. Association for Computing Machinery, New York, NY, USA, 7–15. doi:10.1145/1342250.1342252

[23] Markus Friedrich, Stefan Langer, and Fabian Frey. 2021. Combining Gesture and Voice Control for Mid-air Manipulation of CAD Models in VR Environments.. In *VISIGRAPP (2: HUCAPP)*. 119–127.

[24] Hugo C Gómez-Tone, Jorge Martin-Gutierrez, John Bustamante-Escapa, and Paola Bustamante-Escapa. 2021. Spatial skills and perceptions of space: Representing 2D drawings as 3D drawings inside immersive virtual reality. *Applied Sciences* 11, 4 (2021), 1475.

[25] Samuel W Greenhouse and Seymour Geisser. 1959. On methods in the analysis of profile data. *Psychometrika* 24, 2 (1959), 95–112.

[26] Tovi Grossman, Ravin Balakrishnan, Gordon Kurtenbach, George Fitzmaurice, Azam Khan, and Bill Buxton. 2002. Creating Principal 3D Curves with Digital Tape Drawing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Minneapolis, Minnesota, USA) *(CHI '02)*. Association for Computing Machinery, New York, NY, USA, 121–128. doi:10.1145/503376.503398

[27] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[28] Devamardeep Hayatpur, Seongkook Heo, Haijun Xia, Wolfgang Stuerzlinger, and Daniel Wigdor. 2019. Plane, Ray, and Point: Enabling Precise Spatial Manipulations with Shape Constraints. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology* (New Orleans, LA, USA) *(UIST '19)*. Association for Computing Machinery, New York, NY, USA, 1185–1195. doi:10.1145/3332165.3347916

[29] Mary Hegarty and Maria Kozhevnikov. 1999. Types of visual–spatial representations and mathematical problem solving. *Journal of educational psychology* 91, 4 (1999), 684.

[30] Juan David Hincapié-Ramos, Xiang Guo, Paymahn Moghadasian, and Pourang Irani. 2014. Consumed endurance: a metric to quantify arm fatigue of mid-air interactions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 1063–1072. doi:10.1145/2556288.2557130

[31] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 159–166.

[32] David Kirk, Tom Rodden, and Danaë Stanton Fraser. 2007. Turn it this way: grounding collaborative action with remote gestures. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 1039–1048.

[33] David Kirsh. 2010. Thinking with external representations. *AI & society* 25, 4 (2010), 441–454.

[34] David Kirsh and Paul Maglio. 1994. On distinguishing epistemic from pragmatic action. *Cognitive science* 18, 4 (1994), 513–549.

[35] Jill H Larkin and Herbert A Simon. 1987. Why a diagram is (sometimes) worth ten thousand words. *Cognitive science* 11, 1 (1987), 65–100.

[36] Joseph J LaViola Jr, Ernst Kruijff, Ryan P McMahan, Doug Bowman, and Ivan P Poupyrev. 2017. *3D user interfaces: theory and practice*. Addison-Wesley Professional.

[37] Kun-Hsi Liao. 2017. The abilities of understanding spatial relations, spatial orientation, and spatial visualization affect 3D product design performance: using carton box design as an example. *International Journal of Technology and Design Education* 27 (2017), 131–147.

[38] Jen-Shuo Liu, Portia Wang, Barbara Tversky, and Steven Feiner. 2022. Adaptive visual cues for guiding a bimanual unordered task in virtual reality. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 431–440.

[39] David F Lohman. 2014. Spatial abilities as traits, processes, and knowledge. In *Advances in the psychology of human intelligence*. Psychology Press, USA, 181–248.

[40] Mathias N Lystbæk, Ken Pfeuffer, Tobias Langlotz, Jens Emil Sloth Grønbæk, and Hans Gellersen. 2024. Spatial gaze markers: supporting effective task switching in augmented reality. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–11.

[41] Mark G McGee. 1979. Human spatial abilities: psychometric studies and environmental, genetic, hormonal, and neurological influences. *Psychological bulletin* 86, 5 (1979), 889.

[42] Neil Mennie, Mary Hayhoe, and Brian Sullivan. 2007. Look-ahead fixations: anticipatory eye movements in natural tasks. *Experimental brain research* 179, 3 (2007), 427–442.

[43] Mark R Mine, Frederick P Brooks Jr, and Carlo H Sequin. 1997. Moving objects in space: exploiting proprioception in virtual-environment interaction. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*. 19–26.

[44] Andreea Muresan, Jess Mcintosh, and Kasper Hornbæk. 2023. Using Feedforward to Reveal Interaction Possibilities in Virtual Reality. *ACM Trans. Comput.-Hum. Interact.* 30, 6, Article 82 (Sept. 2023), 47 pages. doi:10.1145/3603623

[45] Don Norman. 2013. *The design of everyday things: Revised and expanded edition*. Basic books.

[46] Alfred Oti and Nathan Crilly. 2021. Immersive 3D sketching tools: Implications for visual thinking and communication. *Computers & Graphics* 94 (2021), 111–123.

[47] Sharon Oviatt. 1999. Ten myths of multimodal interaction. *Commun. ACM* 42, 11 (Nov. 1999), 74–81. doi:10.1145/319382.319398

[48] Sharon Oviatt. 1999. Ten myths of multimodal interaction. *Commun. ACM* 42, 11 (1999), 74–81.

[49] Sharon Oviatt, Rachel Coulston, and Rebecca Lunsford. 2004. When do we interact multimodally? Cognitive load and multimodal communication patterns. In *Proceedings of the 6th international conference on Multimodal interfaces*. 129–136.

[50] Siyou Pei, Alexander Chen, Jaewook Lee, and Yang Zhang. 2022. Hand Interfaces: Using Hands to Imitate Objects in AR/VR for Expressive Interactions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 429, 16 pages. doi:10.1145/3491102.3501898

[51] Wim TJL Pouw, Jacqueline A De Nooijer, Tamara Van Gog, Rolf A Zwaan, and Fred Paas. 2014. Toward a more embedded/extended perspective on the cognitive

function of gestures. *Frontiers in psychology* 5 (2014), 359.

[52] Prashant Rawat and Mayra Donaji Barrera Machuca. 2025. Spatial Hand Actions: Exploring the Hand Actions used to Represent Spatial Thinking for 3D Assembling Tasks. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–21.

[53] Florent Robert, Hui-Yin Wu, Lucile Sassatelli, Stephen Ramanoël, Auriane Gros, and Marco Winckler. 2023. An Integrated Framework for Understanding Multimodal Embodied Experiences in Interactive Virtual Reality. In *Proceedings of the 2023 ACM International Conference on Interactive Media Experiences* (Nantes, France) *(IMX '23)*. Association for Computing Machinery, New York, NY, USA, 14–26. doi:10.1145/3573381.3596150

[54] Francielly Rodrigues, Alexander Giovannelli, Leonardo Pavanatto, Haichao Miao, Jauvane C de Oliveira, and Doug A Bowman. 2023. Amp-it and wisdom: Improving 3D manipulation for high-precision tasks in virtual reality. In *2023 IEEE international symposium on mixed and augmented reality (ISMAR)*. IEEE, 303–311.

[55] Richard Rodriguez, Brian T. Sullivan, Mayra Donaji Barrera Machuca, Anil Ufuk Batmaz, Cyane Tornatzky, and Francisco R. Ortega. 2024. An Artists' Perspectives on Natural Interactions for Virtual Reality 3D Sketching. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 163, 20 pages. doi:10.1145/3613904.3642758

[56] Mike Scaife and Yvonne Rogers. 1996. External cognition: how do graphical representations work? *International journal of human-computer studies* 45, 2 (1996), 185–213.

[57] Arne Seeliger, Raphael P Weibel, and Stefan Feuerriegel. 2024. Context-adaptive visual cues for safe navigation in augmented reality using machine learning. *International Journal of Human–Computer Interaction* 40, 3 (2024), 761–781.

[58] Roger N Shepard and Jacqueline Metzler. 1971. Mental rotation of three-dimensional objects. *Science* 171, 3972 (1971), 701–703.

[59] Rumeysa Turkmen, Zeynep Ecem Gelmez, Anil Ufuk Batmaz, Wolfgang Stuerzlinger, Paul Asente, Mine Sarac, Ken Pfeuffer, and Mayra Donaji Barrera Machuca. 2024. EyeGuide & EyeConGuide: Gaze-based Visual Guides to Improve 3D Sketching Systems *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 178, 14 pages. doi:10.1145/3613904.3641947

[60] Rumeysa Türkmen, Ken Pfeuffer, Mayra Donaji Barrera Machuca, Anil Ufuk Batmaz, and Hans Gellersen. 2022. Exploring Discrete Drawing Guides to Assist Users in Accurate Mid-Air Sketching in VR. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 276, 7 pages. doi:10.1145/3491101.3519737

[61] Barbara Tversky. 2011. Spatial thought, social thought. *Spatial dimensions of social thought* 18 (2011), 17–38.

[62] Barbara Tversky. 2019. *Mind in motion: How action shapes thought*. Hachette UK.

[63] William H Warren Jr. 1990. The perception-action coupling. In *Sensory-Motor Organizations and Development in Infancy and Early Childhood: Proceedings of the NATO Advanced Research Workshop on Sensory-Motor Organizations and Development in Infancy and Early Childhood Chateu de Rosey, France*. Springer, 23–37.

[64] Johann Wentzel, Sasa Junuzovic, James Devine, John Porter, and Martez Mott. 2022. Understanding how people with limited mobility use multi-modal input. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–17.

[65] Christopher D Wickens. 2008. Multiple resources and mental workload. *Human factors* 50, 3 (2008), 449–455.

[66] Adam S Williams, Jason Garcia, and Francisco Ortega. 2020. Understanding multimodal user gesture and speech behavior for object manipulation in augmented reality using elicitation. *IEEE Transactions on Visualization and Computer Graphics* 26, 12 (2020), 3479–3489.

[67] Jacob O Wobbrock, Leah Findlater, Darren Gergle, and James J Higgins. 2011. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 143–146.

[68] Difeng Yu, Tilman Dingler, Eduardo Velloso, and Jorge Goncalves. 2024. Object Selection and Manipulation in VR Headsets: Research Challenges, Solutions, and Success Measurements. *ACM Comput. Surv.* 57, 4, Article 98 (Dec. 2024), 34 pages. doi:10.1145/3706417

[69] Ya-Ting Yue, Xiaolong Zhang, Yongliang Yang, Gang Ren, Yi-King Choi, and Wenping Wang. 2017. WireDraw: 3D Wire Sculpturing Guided with Mixed Reality. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 3693–3704. doi:10.1145/3025453.3025792

[70] Xiaotian Zhang, Weiping He, Mark Billinghurst, Yunfei Qin, Lingxiao Yang, Daisong Liu, and Zenglei Wang. 2024. Usability of visualizing position and orientation deviations for manual precise manipulation of objects in augmented reality. *Virtual Reality* 28, 3 (2024), 134.

[71] Yu Zhao and Bobby Bodenheimer. 2025. How Spatial Ability Affects Response to Gaze-Adaptive Cueing in Mixed Reality Spatial Navigation. In *2025 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 1638–1648.

[72] Xiaoyan Zhou, Adam Sinclair Williams, and Francisco Raul Ortega. 2022. Eliciting multimodal gesture+ speech interactions in a multi-object augmented reality environment. In *Proceedings of the 28th ACM Symposium on Virtual Reality Software and Technology*. 1–10.