IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 3 August 2026, accepted 1 September 2026, date of publication 4 September 2026, date of current version 11 September 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3731057

RESEARCH ARTICLE

Toward Behavioral Signatures of User Experience: An Exploratory Comparative Analysis of Passthrough XR and Desktop Learning Environments

YIYIN GU[1], YIQUN ZHANG[2], MIGUEL A. FEIJOO-GARCIA[2], ALEJANDRA J. MAGANA[2], (Associate Member, IEEE), BEDRICH BENES[1], (Senior Member, IEEE), AND VOICU POPESCU[1], (Member, IEEE)

[1]Department of Computer Science, Purdue University, West Lafayette, IN 47907, USA
[2]Department of Computer and Information Technology, Purdue University, West Lafayette, IN 47907, USA

Corresponding author: Yiyin Gu (gu251@purdue.edu)

This work was supported by the U.S. National Science Foundation under Grants 2212200, 2219842, 2309564, 2318657, and 2417510.

ABSTRACT Extended Reality (XR) systems create new opportunities for educational interaction, but evaluating user experience during immersive learning remains challenging because standard questionnaires are retrospective and can interrupt the learning flow. We report a pilot-scale exploratory between-subjects study (N = 34) using a Convolutional Neural Network learning module in two conditions: (1) a passthrough XR environment on Meta Quest 3 and (2) a traditional desktop interface. The interface comparison provides context for the learning task, while the primary contribution is a bounded test of whether consumer-grade XR tracking can produce candidate movement-based indicators associated with subjective user experience. We collected subjective measures (SUS, NASA-TLX, CSAT, NPS) and objective behavioral streams sampled at 10 Hz (XR head/hand tracking, desktop webcam-based head/attention proxy, and mouse tracking). Between-condition differences were descriptive rather than statistically significant: for example, SUS scores were higher in XR (80.88 vs. 75.59) but did not reach significance ($p = 0.360$), and Visual Attention Focus differed by only 0.9% (XR: 54.9% vs. Desktop: 54.0%, $p > 0.99$, $d = 0.041$). Within the XR condition (n = 17), only one behavioral association survived Bonferroni, Holm, and FDR correction across 25 tests: Head Angular Jerk correlated with SUS usability scores ($r = -0.725$, $p_{\mathrm{Bonferroni}} = 0.025$). Other behavioral associations, kinematic characterizations, and predictive-modeling results are treated as exploratory signals or negative results requiring larger samples. These findings position movement fluency as a candidate passive indicator for this passthrough XR learning task, while emphasizing that replication across tasks, hardware, and interaction designs is required before general behavioral-signature claims can be made.

INDEX TERMS Extended reality, passive assessment, user experience, visualization.

I. INTRODUCTION

The increasing adoption of Extended Reality (XR) technologies in educational and professional training contexts represents a shift in human-computer interaction paradigms. Unlike traditional desktop interfaces that constrain users to 2D mouse-and-keyboard interactions, XR environments enable natural spatial interaction through hand gestures, head movements, and full-body engagement [23], [55]. This transition from indirect manipulation through peripheral devices

The associate editor coordinating the review of this manuscript and approving it for publication was Orazio Gambino.

to direct spatial interaction can support learning outcomes, engagement, and more intuitive interfaces for complex data visualization tasks when the task and implementation benefit from spatial interaction [13].

However, the educational technology community faces a critical methodological challenge: how can we effectively evaluate user experience in these immersive environments without disrupting the learning process itself? Traditional assessment approaches rely heavily on post-task questionnaires and standardized usability instruments, which provide valuable insights but suffer from several fundamental limitations [7]. First, they are inherently retrospective, capturing user impressions after the experience has concluded rather than during natural task performance. Second, they are intrusive, requiring explicit interruption of the learning flow to gather feedback. Third, they may be subject to recall bias and social desirability effects, particularly when users interact with novel technologies [34].

The emergence of consumer-grade XR hardware with built-in tracking capabilities creates new opportunities for passive user experience assessment. Modern head-mounted displays like the Meta Quest 3 provide 27-degree-of-freedom hand tracking, 6-DOF head pose estimation, and real-time spatial understanding without requiring additional instrumentation [24], [38]. This behavioral data stream offers the potential to study candidate ''behavioral signatures'' that may correlate with traditional subjective measures of usability, cognitive load, and user satisfaction, while recognizing that consumer tracking quality limits fine-grained interpretation.

The theoretical foundation for such behavioral signatures emerges from embodied cognition research, which suggests that physical movements and spatial interactions can be associated with cognitive and affective states [53], [57]. In traditional desktop environments, passive interaction traces such as cursor movement, dwell time, and click-stream behavior can provide indirect evidence of attention or difficulty. XR environments expand the range of observable behaviors to include head movement smoothness, hand posture, spatial exploration behaviors, and ergonomic comfort indicators.

Educational applications present useful test cases for passive XR assessment because many involve spatial inspection, manipulation, and attention allocation. Neural network visualization, for example, can benefit from spatial layouts that help users inspect layered structures and relationships [21]. However, these same tasks present cognitive challenges that may appear in movement patterns: jerky head movements might indicate uncertainty or interaction friction, tense hand postures could signal discomfort, and limited spatial exploration may suggest disengagement [45]. These interpretations are candidate explanations rather than validated diagnostic labels.

This paper examines two related research questions with different priorities. First, as contextual background, we compare user experience between a passthrough XR interface and a traditional desktop interface for the same learning module. Second, as the primary contribution, we test whether passive behavioral metrics from XR head and hand tracking show exploratory associations with established subjective measures of workload, usability, and satisfaction.

To address these research questions, we conducted a controlled between-subjects pilot experiment ($N = 34$) comparing a learning module on Convolutional Neural Networks (CNNs) delivered through two distinct interfaces: a passthrough XR environment implemented on Meta Quest 3 and a traditional desktop interface using standard monitor and mouse interaction. The CNN visualization domain was selected because it contains layered spatial relationships and abstract educational content that can be represented in both traditional 2D visualization and immersive 3D manipulation [21].

Our experimental design collected comprehensive data across multiple dimensions: subjective user experience through validated questionnaires (SUS, NASA-TLX, CSAT, NPS, and platform-specific Likert custom items), objective attention patterns through gaze tracking proxies, and detailed behavioral metrics derived from built-in XR tracking sensors. For the XR condition, we leveraged the full capabilities of consumer-grade hardware to extract five distinct behavioral indicators: visual attention focus, head rotational exploration patterns, movement smoothness through angular jerk analysis, hand comfort through resting position deviations, and ergonomic stress through arm elevation tracking.

The central contribution of this work is an exploratory test of whether passive behavioral metrics derived from consumer-grade XR hardware can function as candidate proxies for established subjective user experience measures. The study does not propose a new measurement theory or a finalized evaluation model; instead, it integrates established UX instruments, embodied-interaction theory, and consumer-grade behavioral logging to evaluate which associations appear robust in this pilot sample and which require larger-scale validation. The main empirical result is one corrected movement-fluency association, while the XR-vs-desktop comparison and kinematic characterizations provide context for future work.

II. RELATED WORK

This section reviews literature foundational to our work, establishing the context for evaluating extended reality (XR) visualization interfaces using accessible, objective behavioral measures. We first contrast XR and desktop visualization approaches, highlighting XR's potential for engagement alongside a need for more rigorous empirical comparison (Sec. II-A). We then ground our investigation in embodied cognition theory, which posits that physical interactions in XR should correlate with cognitive states (Sec. II-B). We review evaluation approaches ranging from traditional questionnaire-based methodologies to passive behavioral monitoring in 2D interfaces, establishing a baseline for behavioral analytics (Sec. II-C). Subsequently, we examine current research on objective engagement measurement in XR, identifying a critical gap: while sophisticated multimodal systems exist, they rely on expensive, specialized hardware that limits practical application (Sec. II-D). Finally, we motivate our approach by reviewing the growing potential of consumer-grade hardware for behavioral assessment, positioning our work as an investigation into whether accessible, low-cost sensors can provide meaningful correlations with established user experience metrics (Sec. II-E).

A. XR VS DESKTOP VISUALIZATION INTERFACES

Data visualization research has long investigated interface modalities for exploring complex datasets, with growing attention to extended reality (XR) technologies as alternatives to traditional desktop approaches [43], [50]. Recent comprehensive reviews provide converging evidence: Zhang et al.'s umbrella review analyzed 17 studies revealing that XR technologies boost user motivation and facilitate engagement [55], while Hamilton et al.'s systematic literature review established methodological frameworks for experimental design [23]. Taborda et al.'s recent review specifically addresses engagement and attention in XR environments, highlighting the evolution toward multimodal measurement approaches [49]. Recent meta-analyses suggest that XR benefits depend heavily on task characteristics, user expertise, and application domain [37].

Neural network visualization presents particular challenges for interface design, requiring users to comprehend high-dimensional parameter spaces, layer interactions, and training dynamics simultaneously. Traditional desktop approaches constrain exploration to 2D projections and sequential navigation, while XR promises spatial interaction with 3D network architectures [21]. However, empirical comparisons between XR and desktop neural network visualization remain limited, with most studies focusing on implementation feasibility rather than controlled user experience evaluation [22].

Educational VR research demonstrates engagement benefits across various domains [49], including complex visualization tasks such as spherical geometry [11]. The Cognitive Affective Model of Immersive Learning (CAMIL) provides theoretical grounding for XR's learning mechanisms [37], while embodied cognition theory suggests spatial interaction should enhance understanding of complex relationships [53], [57]. However, most educational VR studies rely on self-reported measures, potentially missing nuanced behavioral differences that emerge during natural task performance.

Taken together, this literature suggests that XR can support spatial learning, but it does not establish a general XR advantage for every learning task or user population. For this reason, our XR-vs-desktop comparison is used as contextual evidence for one CNN visualization module rather than as the paper's primary novelty claim.

B. EMBODIED COGNITION AND PHYSICAL ENGAGEMENT IN XR

Embodied cognition theory proposes that cognitive processes are grounded in bodily interactions with the environment [53], suggesting that physical movement and spatial interaction directly shape cognitive engagement. Recent theoretical advances demonstrate the synergy between embodied cognition and cognitive load theory for optimized interaction [57], establishing a framework that explains how natural physical interactions can reduce cognitive burden while enhancing engagement. This theoretical foundation suggests that behavioral indicators like hand movement patterns and spatial interaction should correlate with cognitive and affective states.

Meta-analyses have found positive effects for technology-based embodied interaction [54], and this theoretical prediction gains empirical support from recent work demonstrating how embodied interactions can optimize cognitive resource allocation [57]. Recent VR education applications have explored embodied design strategies [35] and interaction frameworks [33], though empirical validation using objective behavioral measures remains limited [39].

Our work tests these theoretical predictions by examining whether spatial interaction patterns and hand tension indicators correlate with subjective measures of workload and satisfaction in XR learning environments.

The gap for the present study is not whether embodied interaction can matter in principle, but whether coarse, readily available movement features from consumer passthrough XR hardware show interpretable relationships with standard UX instruments in an authentic learning task. We therefore treat movement features as candidate indicators whose meaning must be bounded by task context and validated empirically.

C. EVALUATION APPROACHES: FROM QUESTIONNAIRES TO PASSIVE MONITORING

Traditional visualization evaluation relies primarily on post-task questionnaires and performance metrics, using established instruments like the System Usability Scale (SUS) for cross-study comparison [8]. While these approaches provide validated benchmarks, they may miss behavioral patterns that emerge during natural task performance, particularly the spatial interaction affordances unique to XR interfaces [7]. Recent advances demonstrate the value of multimodal assessment combining subjective measures with objective behavioral data [6].

In contrast to post-hoc questionnaires, traditional desktop interfaces have established foundations for passive user experience monitoring through behavioral analytics. Web-based systems can track interaction traces including click patterns, scroll velocity, dwell time, and mouse movement trajectories to infer engagement and usability issues. E-learning platforms leverage these metrics to detect student confusion, predict dropout, and adapt content delivery based on interaction patterns. Click-stream analysis reveals navigation inefficiencies, while eye-tracking studies demonstrate correlations between gaze patterns and comprehension in desktop reading tasks [12].

Educational technology research has validated several passive indicators for desktop learning environments. Mouse movement hesitation correlates with cognitive load, prolonged cursor hovering indicates decision uncertainty, and rapid scrolling suggests content skimming rather than deep engagement. Such behavioral traces can support real-time adaptation in intelligent tutoring systems and provide continuous assessment without interrupting learning flow. However, desktop passive monitoring is limited to 2D cursor movements and discrete interface events, lacking the rich spatial and embodied interaction data available in XR environments.

Recent advances in webcam-based monitoring extend desktop capabilities through facial expression analysis and head pose estimation. The Beam SDK enables gaze approximation using standard cameras [4], while emotion recognition systems detect frustration and engagement through facial coding [20]. These approaches bridge traditional desktop analytics with physiological monitoring, though they remain constrained to observing rather than tracking full-body embodied interaction.

This prior work establishes that passive behavioral monitoring is feasible, but it also highlights a modality gap: desktop systems capture cursor and coarse head/face behavior, whereas XR systems capture spatial head and hand movement. Direct numerical comparisons across these sensing modalities therefore require caution, and within-XR associations may be more interpretable than cross-platform behavioral equivalence.

D. OBJECTIVE ENGAGEMENT MEASUREMENT IN XR

Objective engagement measurement in XR has evolved from single-modality approaches toward comprehensive multimodal frameworks. Foundational work in multimodal analytics established computational approaches for measuring complex interactive tasks using multiple sensor streams [6]. Recent advances demonstrate the integration of multiple sensing modalities for comprehensive engagement assessment, combining physiological sensors with behavioral tracking approaches.

VR-specific research has explored the relationship between cognitive load and engagement through facial behavior analysis tools like OpenFace [2] and eye-tracking technologies [1], [40]. Empirical studies demonstrate measurable effects of cognitive load on engagement in VR environments [9], with recent work showing how immersive visualization technologies affect cognitive load, motivation, usability, and embodiment [52]. Advanced approaches now enable cognitive load classification of mixed reality interactions using multimodal sensor signals [28], representing significant progress toward real-time engagement assessment.

However, existing multimodal approaches typically require specialized equipment (EEG, wearable sensors, high-precision eye trackers) that limits adoption in educational settings. These research-grade systems present significant costs and technical barriers, including complex setup procedures, long preparation times, and limited flexibility for non-manufacturer applications [56]. The PLUME framework exemplifies sophisticated capabilities but focuses on post-hoc analysis rather than real-time assessment [29]. Recent work demonstrated cognitive load detection using Varjo VR-3 eye tracking (\$3,000+), achieving 84% classification accuracy but addressing only single cognitive dimension rather than comprehensive user experience [41].

The practical gap is that many educational deployments cannot assume research-grade sensing. A useful intermediate step is to test whether lower-cost, built-in tracking streams can produce interpretable candidate indicators before investing in larger predictive models or real-time adaptive systems.

E. CONSUMER HARDWARE POTENTIAL FOR BEHAVIORAL ASSESSMENT

Commercial head-mounted displays now offer sophisticated 27-DOF hand tracking, creating new possibilities for understanding user behavior during interactive activities [24], [38]. Recent research highlights progress in gesture precision [51], surface touch input [16], and gesture taxonomies for different age groups [3]. Ergonomics research reveals potential differences in immersive environments compared to desktop interfaces [47], though hand ergonomics as engagement indicators represents an emerging area, particularly using commercial hardware in educational contexts.

Webcam-based approaches offer advantages including universal hardware compatibility and reduced setup complexity. The Beam SDK provides head pose estimation and gaze approximation using standard webcams [4], offering trade-offs between precision and accessibility compared to dedicated systems [12]. However, accessibility remains a key challenge for widespread adoption of multimodal engagement measurement.

The gap between sophisticated research systems and practical deployment needs motivates our exploration of consumer hardware capabilities. Rather than developing a comprehensive passive monitoring framework, we investigate whether basic behavioral indicators–attention patterns, hand tension, and movement patterns–captured through accessible hardware can correlate with established usability measures and reveal candidate user experience patterns. This gap analysis leads to a deliberately narrow research question: which, if any, consumer-tracked movement features show corrected associations with subjective UX in a passthrough XR learning task?

F. RESEARCH QUESTIONS AND EXPLORATORY HYPOTHESES

Synthesizing the literature reviewed above, we formulated hypotheses before data collection to avoid post-hoc selection bias. In the revised analysis narrative, H1 provides contextual interface-comparison expectations, while H2 represents the primary exploratory candidate-indicator question. Because the study has a small sample (N = 34; n = 17 per condition), all H1 effect-size patterns and all H2 associations not surviving correction are interpreted as descriptive or hypothesis-generating.

1) H1: CONTEXTUAL XR VS DESKTOP COMPARISON

Based on meta-analytic evidence that XR technologies enhance motivation and engagement [49], [55], and embodied cognition theory predicting that natural spatial interaction reduces cognitive overhead [57], we hypothesized:

H1a Enhanced Usability: XR will show higher system usability, measured by System Usability Scale (SUS) [8]. Prior work suggests that spatial interaction affordances improve perceived usability for visualization tasks [11].

H1b Reduced Cognitive Load: XR will demonstrate lower cognitive workload, measured by NASA Task Load Index (NASA-TLX) [25]. The CAMIL framework predicts that embodied interaction can reduce extraneous cognitive load [37].

H1c Higher Engagement: XR will exhibit higher visual attention focus, measured by percentage of time participants' gaze falls within the visualization area using head pose tracking as a proxy for visual attention. Immersive environments are theorized to increase attentional capture through environmental presence [49].

H1d Greater Satisfaction: XR will yield higher user satisfaction, measured by Customer Satisfaction (CSAT) ratings for experience quality and Net Promoter Score (NPS) for recommendation likelihood. Educational VR studies report higher satisfaction and recommendation rates [43].

2) H2: PRIMARY XR CANDIDATE INDICATOR HYPOTHESES

These five hypotheses represent our core theoretical predictions, grounded in embodied cognition theory [53] and motor control research:

H2a Movement Fluency: Motor control literature demonstrates that movement jerk reflects cognitive uncertainty and motor planning difficulty [17], [26]. We hypothesize that jerkier head movements indicate user confusion, resulting in *lower usability ratings*, measured by SUS scores.

H2b Hand Comfort: Occupational therapy research establishes that hand posture deviations from ergonomic ideals correlate with stress and reduced task performance [31]. We hypothesize that hand positions farther from medical ideal indicate discomfort, resulting in *lower interaction satisfaction*, measured by CSAT interaction ratings.

H2c Spatial Engagement: VR exploration research shows that head-orientation coverage can reflect how users inspect virtual environments [48]. We hypothesize that higher head rotational exploration may indicate curiosity or information seeking, resulting in *higher engagement*, measured by satisfaction scores, while recognizing that exploration can also reflect confusion in some task phases.

H2d Arm Ergonomics: Ergonomics research demonstrates that sustained arm elevation produces rapid fatigue and degraded task performance [10], [14]. We hypothesize that higher arm ergonomic scores correlate with *higher perceived physical demand*, measured by NASA-TLX.

H2e Attention & Effort: Sustained visual attention can indicate either effortless engagement or effortful concentration [30]. We hypothesize that higher Visual Attention Focus may indicate cognitive effort, correlating with *higher mental workload*, measured by NASA-TLX.

III. METHODOLOGY

A. PARTICIPANTS

We recruited N = 34 participants from the university community through campus-wide announcements. Participants were randomly assigned to either the XR condition (n = 17) or desktop condition (n = 17), with assignment stratified by self-reported technology experience to ensure balanced groups. Inclusion criteria required normal or corrected-to-normal vision, basic computer proficiency, and less than 10 hours of prior VR experience to minimize expertise confounds.

Demographics analysis revealed no significant differences between conditions. The XR group had a mean age of 24.2 years (SD = 4.5, range: 19-36) while the desktop group averaged 20.1 years (SD = 2.4, range: 18-27), with no significant age difference (Mann-Whitney U, $p = 0.091$). Gender distribution was balanced: XR condition included 10 males (58.8%), 6 females (35.3%), and 1 preferring not to disclose (5.9%); desktop condition comprised 11 males (64.7%) and 6 females (35.3%).

B. APPARATUS AND SYSTEM

The study employed two distinct hardware configurations optimized for each interface modality. The XR condition utilized a Meta Quest 3 headset with passthrough mode enabled, providing mixed reality visualization while maintaining real-world awareness. Participants remained seated at a standard desk, with virtual content anchored to the physical environment. The desktop condition used standardized 24-inch monitors (1920 × 1080 resolution) with standard mouse and keyboard input, representing typical desktop workstation configurations.

The educational content consisted of an interactive Convolutional Neural Network (CNN) visualization module implemented in Unity for both XR and desktop conditions. Both implementations presented identical educational material through four interactive sections: (1) exploring data completeness effects on model performance, (2) comparing balanced versus imbalanced dataset impacts, (3) identifying pattern variations in MNIST handwritten digits, and (4) analyzing classification accuracy across different data distributions. The CNN architecture visualization displayed

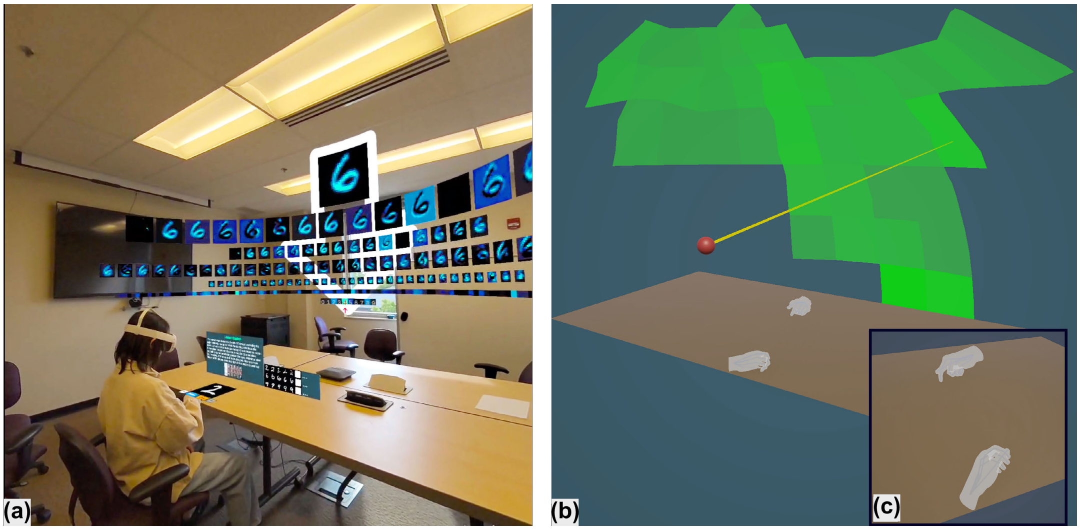


FIGURE 1. **Passive behavioral assessment framework for XR learning environments. (a) Participant learning CNN visualization through passthrough XR on Meta Quest 3, demonstrating natural hand-based interaction with virtual content while maintaining real-world awareness. (b) Behavioral tracking visualization showing head movement patterns overlaid on spatial grid, with opacity indicating attention distribution during learning tasks. (c) Hand tracking data capturing writing-on-notepad gesture as also visible in (a), demonstrating how consumer-grade sensors assess natural interaction behaviors and ergonomic comfort without additional instrumentation.**

network layers, activation patterns, and weight distributions using appropriate interaction metaphors for each platform.

In the XR condition, participants viewed a multi-layer CNN architecture rendered as a 3D spatial diagram on the curved virtual display. Individual network layers (input, convolutional, pooling, fully connected, output) were displayed as stacked rectangular volumes, with activation heatmaps color-coded on each layer's surface. Participants navigated between the four learning sections using virtual buttons anchored to the desk surface, operated through direct hand-touch gestures detected by Quest 3's hand tracking. Each section presented a specific data scenario (e.g., 50% vs 100% training data) and invited participants to observe changes in layer activations and classification accuracy. The desktop condition replicated this content using standard mouse-click interactions for navigation and 2D projections of the same CNN architecture. Since the XR application was developed for a prior study [21] and the present work focuses on behavioral assessment methodology rather than application design, we refer readers to that work for full implementation details.

C. XR IMPLEMENTATION AND MIXED REALITY DESIGN

The XR implementation leveraged Meta Quest 3's passthrough capabilities to create a mixed reality environment combining virtual content with real-world awareness. We use the term "passthrough XR" rather than "see-through AR" to distinguish video-based passthrough (camera capture with digital compositing) from optical see-through displays (transparent optics with digital overlay). This distinction matters because passthrough introduces latency and potential visual artifacts, but enables precise digital-physical alignment and consistent lighting between virtual and real content. We chose passthrough over fully immersive VR to support essential educational activities: participants needed to take handwritten notes on physical paper, reference printed materials, and maintain awareness of the real-world learning environment including the instructor. Passthrough mode was chosen for several advantages: (1) it dramatically reduces cybersickness by maintaining visual-vestibular coherence [18], (2) it preserves spatial awareness allowing users to see their physical desk and hands, (3) it enables haptic grounding when users touch their real desk while interacting with virtual UI elements, and (4) it reduces isolation common in fully virtual environments.

The mixed reality interface employed a spatial layout with two distinct zones. The primary interaction zone anchored virtual UI controls to the detected physical desk surface, providing passive haptic feedback. Users could see both their real hands and virtual interface elements simultaneously, reducing cognitive load from context switching. The visualization zone featured content on a curved virtual display positioned 2 meters away, spanning 180° horizontally as shown in Figure 1(A). This mixed reality approach maintained immersion while preserving real-world grounding.

Our hand tracking implementation utilized Quest 3's four infrared cameras to achieve markerless tracking of both hands simultaneously. The system tracked 26 skeletal joint positions per hand (4 joints per finger, 3 for the thumb, plus palm and wrist), creating a 27-DOF representation when including wrist position [24], as shown in Figure 1(C). We logged and analyzed the behavioral streams at 10 Hz.

Consumer markerless hand tracking introduces measurement uncertainty that is important for interpreting the

TABLE 1. Conceptual motivation and interpretation limits for the behavioral indicators.

Metric	Basis	Expected UX relation	Sensor source	Interpretation limit
Head Angular Jerk	Motor-control work links movement smoothness to planning fluency and corrective action.	Smoother head movement may accompany easier spatial navigation and higher usability.	XR head pose.	Does not directly measure usability; may also reflect task strategy or individual movement style.
Resting Hand Position Deviation	Ergonomics and occupational-therapy work define neutral hand postures for seated tasks.	Larger deviation may indicate discomfort or non-neutral interaction posture.	XR hand skeleton.	A standardized neutral posture cannot represent every participant's comfortable resting posture.
Head Rotational Exploration	Visual search and exploration literature links head/gaze coverage to information seeking.	Exploration may reflect engagement, but excessive exploration may also reflect confusion.	XR head pose.	Direction is ambiguous without task context and phase information.
Arm Ergonomic Score	Sustained arm elevation is associated with fatigue and physical demand.	Greater elevation-time may correspond to higher perceived physical workload.	XR wrist/hand pose.	Seated desk use constrains arm range; low scores may reflect task layout rather than comfort.
Visual Attention Focus	Eye-tracking literature links attention allocation to effort and task focus.	More focus may indicate engagement or effortful concentration.	XR head ray; desktop webcam/head proxy.	Head direction is a coarse gaze proxy and cross-platform sensing fidelity differs.

hand-derived metrics. Recent robotic characterization of Meta Quest Pro and Quest 3 markerless hand tracking reported that, in best-performing conditions, Quest 3 showed an average positional error of 1.73 cm and nontrivial jitter/latency behavior [19]. Robot-driven evaluation of Quest 3 inside-out tracking likewise shows that accuracy depends on motion axis and tracking conditions [36]. These limitations do not invalidate coarse session-level movement summaries, but they constrain claims about fine hand posture, small-amplitude finger motion, and latency-sensitive hand kinematics.

D. BEHAVIORAL METRICS: THEORETICAL JUSTIFICATION

Our selection of five specific behavioral metrics was grounded in established theoretical frameworks and prior empirical research, determined before data collection to avoid post-hoc selection bias.

We map these behavioral indicators to CAMIL cautiously. CAMIL treats presence and agency as central affordances that influence factors such as interest, intrinsic motivation, self-efficacy, embodiment, cognitive load, and self-regulation [37]. Head Angular Jerk is therefore treated as a candidate signal related to cognitive load or interaction friction because corrective head movements may accompany uncertainty during spatial inspection. Head Rotational Exploration and Visual Attention Focus may contextualize attention allocation, agency, and self-regulation during the task, but their direction is task-dependent: broad exploration can reflect curiosity, search difficulty, or disorientation. Resting Hand Position Deviation and Arm Ergonomic Score are treated as comfort and strain indicators that may influence embodiment and intrinsic motivation through ease of interaction and willingness to continue using the system. These mappings are theoretical motivations for metric selection rather than claims that the metrics directly measure CAMIL constructs.

Head Angular Jerk measures movement smoothness through triple differentiation of head pose data [26], based on motor control literature showing that jerky movements can reflect corrective action, cognitive uncertainty, or motor planning difficulty [17]. Smoother head movements may accompany confident spatial navigation and lower interaction friction, while jerkier patterns may indicate confusion or discomfort with interface controls. This metric operationalizes an embodied-cognition prediction that cognitive and affective states can be reflected in motor behavior, but it is not a direct measure of usability. Formally, given the head orientation vector $\boldsymbol{\theta}(t) = (\theta_{\text{yaw}}, \theta_{\text{pitch}}, \theta_{\text{roll}})$ over time, angular jerk is computed as the mean magnitude of the third derivative:

$$J_{\text{head}} = \frac{1}{T}\int_0^T \left\| \frac{d^3\boldsymbol{\theta}}{dt^3} \right\| dt \tag{1}$$

where $\|\cdot\|$ denotes the Euclidean norm of the component-wise third derivative and T is the session duration. In practice, derivatives are computed via finite differences on the discrete 10 Hz tracking signal after temporal smoothing.

Resting Hand Position Deviation quantifies deviation from a neutral ergonomic posture based on occupational therapy research on hand positioning for sustained computer work [31]. Relaxed hand postures can be associated with reduced strain, while tense or awkward positions may indicate discomfort. This metric is therefore an ergonomic proxy for embodied comfort rather than a personalized comfort model. We compute the mean angular deviation between the observed wrist-to-fingertip vector $\mathbf{v}_{\text{obs}}(t)$ and the reference resting vector $\mathbf{v}_{\text{ref}}$ during detected rest periods:

$$D_{\text{hand}} = \frac{1}{N_{\text{rest}}}\sum_{i=1}^{N_{\text{rest}}} \arccos\left(\frac{\mathbf{v}_{\text{obs}}(t_i)\cdot\mathbf{v}_{\text{ref}}}{\|\mathbf{v}_{\text{obs}}(t_i)\|\,\|\mathbf{v}_{\text{ref}}\|}\right) \tag{2}$$

where the reference vector $\mathbf{v}_{\text{ref}}$ represents the neutral hand posture defined in Section III-D1.

Head Rotational Exploration measures spatial search behavior through spherical binning analysis [48]. Active visual exploration can indicate engagement and learning-oriented behavior, though excessive exploration may signal disorientation, search difficulty, or interface confusion. This metric is therefore interpreted as context-dependent rather than inherently positive. The viewable sphere is discretized into $10^\circ \times 10^\circ$ bins in yaw (ϕ) and pitch (ψ). Exploration percentage is:

$$E_{\text{head}} = \frac{|\{(i,j) : \exists\, t \text{ s.t. } \phi(t) \in B_i,\, \psi(t) \in B_j\}|}{N_{\text{bins}}} \times 100\% \quad (3)$$

where B_i, B_j denote the i-th yaw and j-th pitch bins, and N_{bins} is the total number of bins in the accessible viewing range.

Arm Ergonomic Score captures prolonged arm elevation through cumulative elevation-time analysis [10], based on ergonomics research showing that sustained arm elevation can lead to fatigue and degraded performance [14]. This physiological constraint is relevant in XR environments where users interact with virtual content in 3D space, but the seated desk setup used here may also constrain the range of arm movement. The score integrates hand elevation above shoulder level over time:

$$S_{\text{ergo}} = \int_0^T \max(0,\ h_{\text{hand}}(t) - h_{\text{shoulder}})\, dt \quad (4)$$

where $h_{\text{hand}}(t)$ is the vertical position of the hand (wrist joint) and h_{shoulder} is the estimated shoulder height.

Visual Attention Focus employs 3D ray casting from the camera position to approximate coarse attention behavior, building on eye-tracking literature linking visual attention and cognitive engagement [30]. Because the consumer XR device used here lacks integrated eye tracking, head orientation is a proxy for broad attentional allocation rather than fine-grained fixation. A ray $\mathbf{r}(t) = \mathbf{p}(t) + \lambda \hat{\mathbf{d}}(t)$ is cast from the head position $\mathbf{p}(t)$ along the forward direction $\hat{\mathbf{d}}(t)$ at each frame. Focus percentage is:

$$F_{\text{attn}} = \frac{|\{t : \mathbf{r}(t) \cap \mathcal{V} \neq \emptyset\}|}{N_{\text{frames}}} \times 100\% \quad (5)$$

where $\mathcal{V}$ is the bounding volume of the visualization content and N_{frames} is the total number of tracked frames.

Head-Gaze Proxy Limitations: Using head orientation as a gaze proxy introduces systematic measurement error. Eye-tracking research shows that eye gaze can deviate from head orientation by up to $\pm 15^\circ$ [1], and in immersive environments, users frequently make eye-only saccades without corresponding head movements, particularly for targets within the central $\pm 20^\circ$ of the visual field. This means our Visual Attention Focus metric captures *coarse attentional allocation* rather than fine-grained fixation patterns. The $10^\circ \times 10^\circ$ binning granularity for Head Rotational Exploration partially accounts for this limitation, as it is larger than typical eye-only saccade amplitudes.

1) RESTING HAND POSITION: REFERENCE POSTURE AND DETECTION

The reference resting hand posture $\mathbf{v}_{\text{ref}}$ was defined based on occupational therapy guidelines for neutral hand positioning during seated desk work [31]: forearms pronated and resting on the desk surface, wrists in approximately 10–15° extension, and fingers in a relaxed semi-flexed position. This corresponds to the wrist-to-metacarpophalangeal (MCP) joint vector oriented approximately 15° below the horizontal plane relative to the desk surface.

Rest periods were detected using a velocity threshold: frames where hand velocity fell below 0.02 m/s for at least 0.5 seconds were classified as rest. This threshold was determined from pilot data to distinguish intentional resting from brief movement pauses.

We acknowledge that comfortable resting positions may vary across individuals due to differences in hand anatomy, habitual posture, and desk-chair ergonomics. Our metric captures deviation from a standardized medical reference, not from each participant's personal comfort position. While a personalized baseline calibration could improve sensitivity, we opted for the standardized reference to enable cross-participant comparison. Intra-user variability within a session was partially addressed by averaging across all detected rest periods, though systematic postural drift over the session was not explicitly modeled.

These five metrics were specifically chosen to capture complementary dimensions of embodied interaction: motor control (jerk), ergonomic comfort (hand position, arm elevation), spatial cognition (exploration), and visual attention (focus). This theoretical framework guided our selection before data analysis, ensuring that the candidate behavioral indicators reflect established psycho-physical relationships rather than data-driven correlations.

E. EXPERIMENTAL DESIGN

We employed a between-subjects design comparing interface type (XR vs. Desktop) to minimize learning effects and fatigue. We collected both subjective questionnaire data and objective behavioral metrics from tracking sensors.

F. PROCEDURE

The experimental procedure followed a standardized protocol across both conditions. After providing informed consent, participants completed demographic questionnaires assessing age, gender, technology experience, and prior VR exposure. A brief tutorial (2 minutes) familiarized participants with interface-specific controls without revealing study content. Participants then engaged in self-paced exploration of the CNN visualization module.

Post-task assessment included a comprehensive battery of validated questionnaires. The System Usability Scale (SUS) assessed perceived usability through 10 standardized items [8]. NASA Task Load Index (NASA-TLX) measured subjective workload across six dimensions: mental demand, physical demand, temporal demand, performance, effort, and frustration [25]. Customer Satisfaction (CSAT) ratings evaluated experience, visual, and interaction satisfaction on 5-point Likert scales [42]. Net Promoter Score (NPS) assessed recommendation likelihood on a 0-10 scale [44]. XR participants additionally completed the Simulator Sickness Questionnaire (SSQ) to assess cybersickness symptoms [32].

G. DATA COLLECTION

1) SUBJECTIVE MEASURES

We administered established validated instruments to capture subjective user experience dimensions. The System Usability Scale provided a standardized usability metric enabling cross-study comparison. NASA-TLX captured perceived workload across multiple dimensions, critical for understanding cognitive demands. CSAT and NPS metrics assessed overall satisfaction and advocacy potential. XR participants additionally completed the Simulator Sickness Questionnaire (SSQ) to assess cybersickness symptoms. SSQ inclusion was essential because cybersickness directly impacts usability and cognitive performance in XR environments—even mild symptoms can confound behavioral metrics and subjective ratings. Furthermore, documenting low cybersickness helps interpret the candidate behavioral indicators within a tolerable passthrough XR experience.

Beyond standardized instruments, we designed custom Likert-scale questions (1 = strongly disagree, 5 = strongly agree) to capture platform-specific experiences that validated instruments might miss.

XR-specific questions assessed five dimensions: headset comfort (“The headset was uncomfortable to wear”), display preferences (“I liked having a large display for the visualizations”), technology adoption intentions (“I would consider buying an XR headset for learning activities”), usage willingness (“I would use an XR headset if it was provided for learning”), and passthrough integration (“I liked being able to see the real world while using the interface”).

Desktop-specific questions focused on interface appropriateness for screen constraints (“The visualization was too complex for the small screen”).

Comparable cross-platform questions enabled direct condition comparisons across six dimensions: visualization utility (“The visualizations were useful for understanding AI bias”), learning outcomes (“I learned something new about AI bias from this experience”), interface difficulty (“I found the interface difficult to use”), comprehension support (“The visualization helped me understand the AI model better”), complexity management (“I felt overwhelmed by the complexity of the visualization”), and prior familiarity (“I was already familiar with this type of content”).

2) OBJECTIVE MEASURES

For the XR condition, we logged 6-DOF head pose and 27-DOF hand joint data at 10 Hz, capturing complete skeletal hand representations including 26 joints per hand plus wrist position [38]. From this raw data, we extracted the five behavioral metrics described in the Feature Extraction section: (1) *Visual Attention Focus* using 3D ray casting, (2) *Head Rotational Exploration* through spherical binning, as shown in Figure 1(B), (3) *Head Angular Jerk* via triple differentiation, (4) *Resting Hand Position Deviations* based on medical standards, and (5) *Arm Ergonomic Score* measuring prolonged arm elevation.

For the desktop condition, we captured mouse coordinates and webcam-based head pose estimates at 10 Hz using the Beam SDK [4]. The *Visual Attention Focus* measured the percentage of time participants’ estimated gaze fell within the visualization area versus control regions. For XR, we used ray casting from head position to detect focus on virtual content.

Cross-Platform Measurement Caveat: The XR and desktop conditions employ fundamentally different sensing modalities with distinct precision characteristics, even though all behavioral streams were logged and analyzed at 10 Hz. XR behavioral data derives from Quest 3’s infrared-based inside-out tracking, which captures head pose and skeletal hand pose, whereas desktop head pose estimation relies on webcam-based computer vision (Beam SDK) with substantially lower spatial precision and no skeletal hand tracking. The qualitative difference in tracking fidelity (rather than sampling rate) limits the interpretability of direct cross-platform behavioral comparisons. The Visual Attention Focus comparison between conditions should be interpreted as an approximate indicator rather than a precise quantitative equivalence. The within-XR candidate-indicator correlations (RQ2) are not affected by this cross-platform limitation, as they rely exclusively on the XR tracking data.

H. EXPLORATORY SIGNAL PROCESSING

For exploratory characterization of movement quality, we implemented signal processing techniques that normalize for individual differences in movement duration and amplitude while extracting theoretically grounded features. These analyses describe movement patterns in the pilot sample and define measurement boundaries for future work; they are not used as confirmatory evidence that the current dataset supports multivariate behavioral prediction.

1) TEMPORAL SMOOTHING

All behavioral streams used in the analysis were sampled at 10 Hz. Raw tracking data was preprocessed using a 3-sample moving average filter, corresponding to a 0.3-second smoothing window, to reduce high-frequency sensor noise while preserving meaningful kinematic features. Derivatives for jerk-based metrics were computed after this smoothing step. This window size balances noise reduction against temporal resolution for human movement analysis [17].

2) MOVEMENT SEGMENTATION AND DIMENSIONLESS JERK

Traditional jerk metrics confound movement smoothness with duration and amplitude. We computed the Dimensionless Jerk metric, which normalizes the integrated squared-jerk by movement duration and amplitude:

$$\mathrm{DJ} = \sqrt{\frac{T^5}{2A^2} \int_0^T \left(\frac{d^3x}{dt^3}\right)^2 dt} \tag{6}$$

where T is movement duration and A is total path amplitude. Critically, this metric must be applied to *individual movements*, not entire recording sessions. We segmented continuous trajectories into discrete movements using velocity thresholds: movement onset was detected when velocity exceeded 5% of peak velocity, and offset when velocity dropped below this threshold. Movements shorter than 0.1 s or with path length less than 1 cm were excluded. DJ is dimensionless (normalized for duration and amplitude), with lower values indicating smoother movements. The minimum-jerk optimal trajectory [17] serves as the theoretical lower bound; under this formulation, it yields DJ $= \sqrt{360} \approx 19$. Higher observed values indicate less efficient, more corrective movements [26].

3) TWO-THIRDS POWER LAW ANALYSIS

Biological motion follows the two-thirds power law relating velocity (V) to path curvature (C): $V = k \cdot C^{-\beta}$, where $\beta \approx 1/3$ for natural human movement. We computed the power law exponent using principal component analysis on log-transformed velocity-curvature pairs, with deviation from the ideal $\beta = 1/3$ indicating non-biological motion characteristics [59].

4) CORRECTION SENSITIVITY

We computed the ratio of movement smoothness to acceleration magnitude as an indicator of fine motor control and corrective movements. Higher correction sensitivity suggests more controlled, intentional movements rather than ballistic motions.

5) MULTIMODAL FUSION FEATURES

Beyond unimodal analysis, we computed head-hand coordination metrics including: (1) cross-modal velocity correlation, (2) Phase Locking Value (PLV) measuring phase synchronization between head and hand movements, (3) spectral coherence in the frequency domain, and (4) a composite Coordination Index combining temporal synchronization measures. These fusion features capture emergent coordination patterns invisible to single-modality analysis [57].

I. DATA ANALYSIS

For between-groups comparisons (XR vs Desktop), we used Mann-Whitney U tests for non-parametric data and independent t-tests for normally distributed variables, confirmed through Shapiro-Wilk tests. Effect sizes were calculated using Cohen's d [46], with interpretations of small ($d = 0.2$), medium ($d = 0.5$), and large ($d = 0.8$). Given our small sample sizes (XR n = 17, Desktop n = 17), between-condition effect sizes are reported as descriptive context rather than as evidence of platform superiority when statistical tests are non-significant.

For correlation analyses within the XR condition (n = 17), we computed Pearson correlation coefficients between objective behavioral metrics and subjective measures. Correlation strength was interpreted using Cohen's conventions: small ($r = 0.1$), medium ($r = 0.3$), and large ($r = 0.5$) effects, with very large effects ($r > 0.7$) indicating strong relationships in this sample. Given multiple comparisons (5 behavioral metrics $\times$ 5 subjective measures $=$ 25 tests), we applied three correction methods to control family-wise error rate: Bonferroni correction ($\alpha_{\text{corrected}} = 0.05/25 = 0.002$) [15], Holm-Bonferroni step-down procedure [27], and Benjamini-Hochberg False Discovery Rate (FDR) control [5]. We report both uncorrected and corrected p-values, with primary conclusions based on correlations surviving Bonferroni correction. Statistical power calculations indicated our XR sample size could detect large effects ($r > 0.7$) with 80% power at $\alpha = 0.05$, but was underpowered for medium effects ($r = 0.3 - 0.5$). Therefore, uncorrected correlations and exploratory kinematic analyses are treated as hypothesis-generating. Missing data was handled through pairwise deletion.

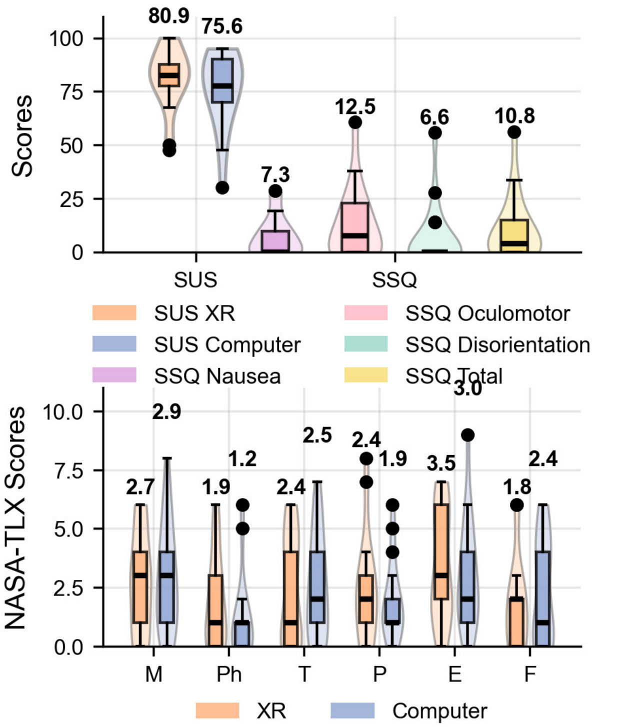


FIGURE 2. Comprehensive subjective assessment across XR and Desktop conditions. Violin plots show distributions for System Usability Scale (SUS), Simulator Sickness Questionnaire (SSQ, XR only), and NASA Task Load Index dimensions. XR demonstrates good usability (SUS M = 80.88 vs Desktop M = 75.59, d = 0.32) with low cybersickness scores (SSQ M = 10.78, SD = 15.36 on 0-48 scale). NASA-TLX reveals similar mental demand across conditions with a slightly lower frustration level in XR (d = −0.28). All comparisons are descriptive given the small sample; none reached statistical significance.

IV. RESULTS

A. RQ1: CONTEXTUAL XR AND DESKTOP COMPARISON

The between-subjects comparison (N = 34) provides context for interpreting the learning task and interface implementation. Because the study has n = 17 participants per condition, these results are not used to make broad claims about XR superiority.

1) SUBJECTIVE EXPERIENCE COMPARISON

The System Usability Scale assessment, as shown in Figure 2, showed XR participants (n = 17) achieved mean scores of 80.88 (SD = 14.74), placing the interface in the "Good" usability range, while desktop participants (n = 17) scored 75.59 (SD = 17.84) in the "OK" range [8]. This 5-point difference did not reach statistical significance (Mann-Whitney U = 171.5, $p = 0.360$). The small effect size (Cohen's $d = 0.32$) is noted for descriptive purposes; however, with n=17 per group, our study had approximately 26% power to detect a medium effect ($d = 0.50$) at $\alpha = 0.05$, meaning this comparison was substantially underpowered.

NASA-TLX workload dimensions revealed comparable cognitive demands between conditions. Mental demand showed minimal difference (XR: M = 2.71, SD = 2.17; Desktop: M = 2.94, SD = 2.30; $d = n - 0.11$, $p = 0.761$). Frustration levels were lower on average in XR (M = 1.76, SD = 1.92) compared to desktop (M = 2.35, SD = 2.23), but this difference did not reach significance ($d = -0.28$, $p = 0.417$). Physical demand, temporal demand, performance, and effort showed negligible differences between conditions (all $|d| < 0.30$).

Customer satisfaction metrics showed descriptive mean differences, as shown in Figure 4. Experience satisfaction: XR (M = 4.12, SD = 0.60) versus Desktop (M = 3.71, SD = 0.92), $d = 0.53$, $p = 0.181$. Visual satisfaction: XR (M=4.06, SD=0.56) versus Desktop (M = 3.76, SD = 0.83), $d = 0.42$, $p = 0.339$. Net Promoter Score: XR (M = 7.71, SD = 1.40) versus Desktop (M = 6.47, SD = 2.50), $d = 0.61$, $p = 0.164$. None of these comparisons reached statistical significance, and given that our sample provided approximately 26% power to detect medium effects, these effect sizes should be interpreted cautiously as descriptive trends requiring confirmation in adequately powered studies.

2) PLATFORM-SPECIFIC CUSTOM QUESTIONS

Analysis of custom questionnaire responses revealed important platform-specific insights, as detailed in Figure 3.

a: XR-SPECIFIC RESULTS

Participants reported general headset comfort (M = 2.06, SD = 0.90, where 1=strongly disagree with "uncomfortable") and strong appreciation for the large virtual display (M = 4.53, SD = 0.51). Purchase intentions were neutral (M = 2.29, SD = 1.21), but willingness to use if provided was high (M = 4.12, SD = 1.17). Participants highly valued real-world visibility through passthrough (M = 4.18, SD = 0.64).

b: DESKTOP-SPECIFIC RESULTS

Desktop users found the interface reasonably manageable for screen size constraints (M = 2.18, SD = 0.73), with most disagreeing that complexity was problematic for small screens.

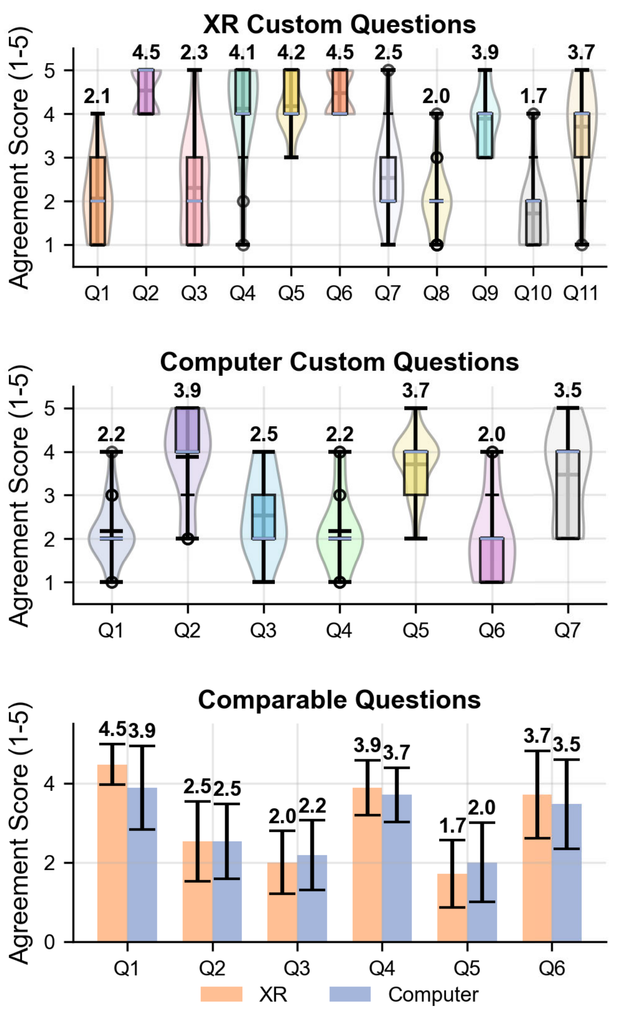


FIGURE 3. Platform-specific custom questionnaire responses comparing XR and Desktop conditions. These Likert-scale questions addressed unique aspects of each interface modality, including comfort, complexity, and usability concerns specific to headset-based versus traditional desktop interactions.

c: CROSS-PLATFORM COMPARISONS

XR participants found visualizations more useful (XR: M = 4.47, SD = 0.51 vs Desktop: M = 3.88, SD = 1.05), reaching statistical significance (p = 0.047). Both groups reported similar learning outcomes (XR: M = 2.53, SD = 1.01 vs Desktop: M = 2.53, SD = 0.94, p > 0.99). XR users found the interface less difficult (XR: M = 2.00, SD = 0.79 vs Desktop: M = 2.18, SD = 0.88, p = 0.544), reported greater visualization helpfulness (XR: M = 3.88, SD = 0.70 vs Desktop: M = 3.71, SD = 0.69, p = 0.462), felt less overwhelmed by complexity (XR: M = 1.71, SD = 0.85 vs Desktop: M = 2.00, SD = 1.00, p = 0.362), and had similar content familiarity (XR: M = 3.71, SD = 1.10 vs Desktop: M = 3.47, SD = 1.12, p = 0.543).

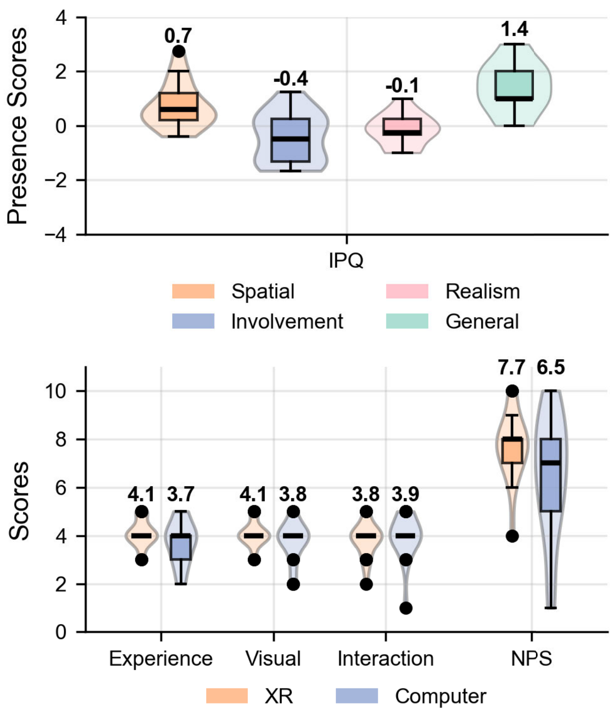


FIGURE 4. Satisfaction metrics comparison between XR and Desktop conditions. Mean values were higher for XR on Experience (M = 4.12 vs 3.71, d = 0.53), Visual Satisfaction (M = 4.06 vs 3.76, d = 0.42), and Net Promoter Score (M = 7.71 vs 6.47, d = 0.61), but none of these differences reached statistical significance. Effect sizes are reported for descriptive purposes given the study's limited power to detect medium effects (≈ 26% at n = 17 per group).

These platform-specific insights complement the standardized measures. Only one comparison (visualization usefulness) reached statistical significance. The remaining directional trends favoring XR require confirmation in larger samples before drawing conclusions about platform advantages.

3) OBJECTIVE ATTENTION COMPARISON

Analysis of the Visual Attention Focus revealed that XR participants maintained focus on educational content 54.9% of the time (SD = 25.8%) compared to 54.0% for desktop users (SD = 15.9%)—a small 0.9% difference. This difference was not statistically significant (Mann-Whitney U = 128.0, $p > 0.99$) and showed a negligible effect size (Cohen's $d = 0.041$). However, this comparison should be interpreted with caution given the different sensing modalities: XR attention was measured via ray casting from Quest 3 head pose, while desktop attention relied on webcam-based head pose estimation with substantially lower spatial precision (both pipelines sampled at 10 Hz).

B. RQ2: CANDIDATE BEHAVIORAL SIGNATURES IN XR

Our exploratory analysis examines correlations between passive behavioral metrics and subjective measures within the XR condition (n=17). These analyses investigate whether built-in tracking data shows meaningful associations with user experience measures.

1) FEATURE EXTRACTION

From the raw tracking data, we computed five objective behavioral metrics for the XR condition:

- Visual Attention Focus: Percentage of time focused on primary visualization using 3D ray casting from head position/rotation to detect gaze intersection with the visualization area (XR: M = 54.9%, SD = 25.8%; Desktop: M = 54.0%, SD = 15.9%)
- Head Rotational Exploration: Percentage of viewable sphere explored, calculated by dividing the sphere into $10° \times 10°$ bins and tracking unique bins visited (0-100%)
- Head Angular Jerk: Smoothness of head movement calculated through triple differentiation (position → velocity → acceleration → jerk), with lower values indicating smoother movement ($°/s^3$)
- Resting Hand Position Deviations: Average angular deviation from ideal resting hand position based on medical standards, with lower values indicating more comfortable hand postures (degrees)
- Arm Ergonomic Score: Cumulative elevation-time score measuring prolonged arm elevation when hands are held above shoulder level, with higher values indicating worse ergonomics (meter-seconds, i.e., height above shoulder integrated over time)

2) ADVANCED KINEMATIC FEATURES

Using the signal processing pipeline described in Section III-H, we computed normalized kinematic features to enable comparison across participants with varying movement styles:

a: DIMENSIONLESS JERK

Using movement segmentation, we detected an average of 258 (SD = 109) individual movements per session, with mean movement duration of 0.71 seconds (SD = 0.21). Head movement dimensionless jerk averaged 137.66 (SD = 50.30), while bilateral hand movements showed higher jerk values (M = 253.37, SD = 179.27), indicating that hand movements in XR involve more corrective adjustments than head movements. Both values substantially exceed the minimum-jerk lower bound (≈ 19 under our formulation), reflecting the learning and exploration nature of XR interaction rather than practiced, ballistic reaches.

b: TWO-THIRDS POWER LAW COMPLIANCE

The power law exponent β averaged 0.61 (SD = 0.05), deviating from the biological ideal of $\beta = 1/3$. The coefficient of determination ($R^2 = 0.57$, SD = 0.05) indicated moderate adherence to the power law model, with biological naturalness scores averaging 0.19 (SD = 0.13), suggesting that XR interaction movements differ from natural biological motion patterns.

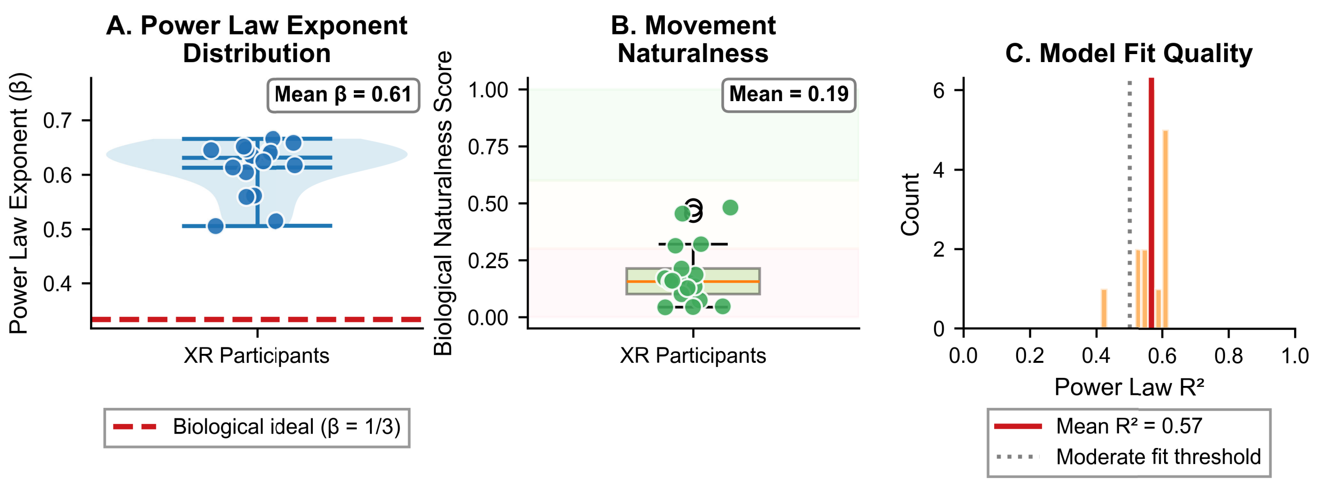


FIGURE 5. **Two-thirds power law analysis for XR participants reveals deviation from natural biological motion. (A) Distribution of power law exponents (β) showing substantial deviation from the biological ideal of $\beta = 1/3$ (dashed red line). Mean $\beta = 0.61$ represents approximately 83% deviation from natural kinematics, indicating XR hand movements follow different motor control strategies than biological reaching movements. (B) Biological naturalness scores (M = 0.19) confirm limited compliance with natural motion patterns. (C) Model fit quality (R^2) distribution showing moderate adherence to the power law relationship, though with non-biological exponents.**

c: CORRECTION SENSITIVITY

Hand movement correction sensitivity averaged 0.076 (SD = 0.118), indicating the ratio of smoothness to acceleration that characterizes intentional corrective movements. The high variability reflects individual differences in motor control strategies during XR interaction.

d: PATH LENGTH KINEMATICS

The path ratio (actual path length / straight-line distance) averaged 83.65 (SD = 40.66) for hand movements across the entire session, reflecting the cumulative exploration and manipulation behaviors during the learning task rather than single point-to-point reaches.

These normalized metrics provide an exploratory characterization of movement quality that is less dependent on individual movement speed and amplitude preferences. They are used to describe the XR interaction patterns in this sample rather than to establish validated UX predictors. Figure 5 visualizes the power law compliance and biological naturalness distributions, while Figure 6 provides a comprehensive view of the exploratory kinematic features.

3) CORRELATION ANALYSIS

We examined correlations between five objective behavioral metrics and subjective measures within the XR condition (n=17). Given multiple comparisons (5 metrics $\times$ 5 subjective measures = 25 tests), we applied rigorous multiple comparison corrections. At Bonferroni-corrected $\alpha = 0.002$, only correlations with $p_{\text{uncorrected}} < 0.002$ survive correction.

Significant correlation (surviving Bonferroni correction):

- Head Angular Jerk (mean absolute jerk) with SUS scores: $r = -0.725$, $p_{\text{uncorrected}} = 0.001$, $p_{\text{Bonferroni}} = 0.025$, significant after all correction methods

Additional corrected-matrix associations (not surviving correction):

- Head Correction Sensitivity with SUS scores: $r = 0.554$, $p_{\text{uncorrected}} = 0.021$, $p_{\text{Bonferroni}} = 0.522$
- Exploration Percentage with SUS scores: $r = -0.496$, $p_{\text{uncorrected}} = 0.043$, $p_{\text{Bonferroni}} = 1.000$
- Visual Attention Focus showed no significant correlations (all $p > 0.05$)

Of the 25 correlations tested, 3 were significant at uncorrected $\alpha = 0.05$, but only 1 (Head Angular Jerk $\leftrightarrow$ SUS) survived Bonferroni, Holm-Bonferroni, and FDR corrections. This rigorous approach reduces the probability of at least one Type I error from $1 - (1 - 0.05)^{25} \approx 72\%$ (when testing 25 correlations without correction) to the nominal 5% family-wise level. Figure 7 presents the correlation matrix with Bonferroni correction indicators. Figure 8 shows the same 25 tests as a volcano plot and the number surviving each correction method.

C. CYBERSICKNESS ASSESSMENT

Simulator Sickness Questionnaire results indicated good tolerability of the XR system in this sample (XR only, n=17). Total SSQ scores were very low (M=10.78, SD=15.36) on a 0-48 scale, indicating minimal cybersickness symptoms. The 16-item symptom analysis revealed that most symptoms were completely absent, with only a few participants reporting mild symptoms.

When asked directly about motion sickness status, the vast majority of participants reported experiencing no motion sickness. These results are consistent with the expected benefits of the passthrough approach: by maintaining visual connection to the real world and enabling natural proprioceptive feedback, passthrough mixed reality may mitigate several factors commonly associated with VR sickness.

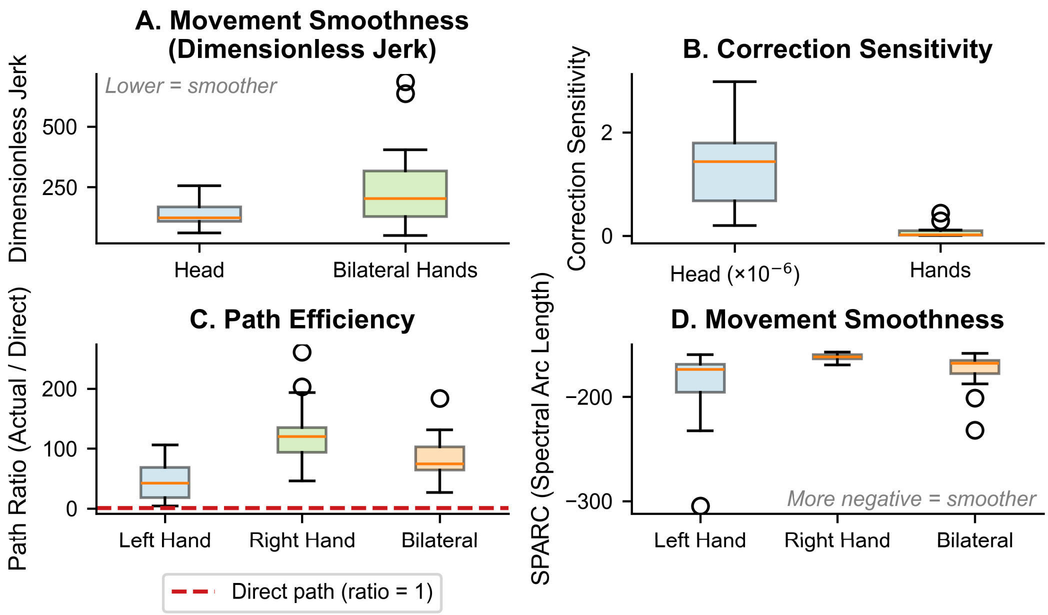


FIGURE 6. **Advanced kinematic features dashboard comparing head and hand movement metrics. (A) Dimensionless jerk distributions computed from individual movement segments—lower values indicate smoother, more optimal movements. Head movements (M = 137.66) show smoother kinematics than hand movements (M = 253.37). (B) Correction sensitivity indicating the ratio of smoothness to acceleration. (C) Path ratio comparing actual movement paths to direct trajectories across entire sessions. (D) SPARC (Spectral Arc Length) smoothness metric, where more negative values indicate smoother movements.**

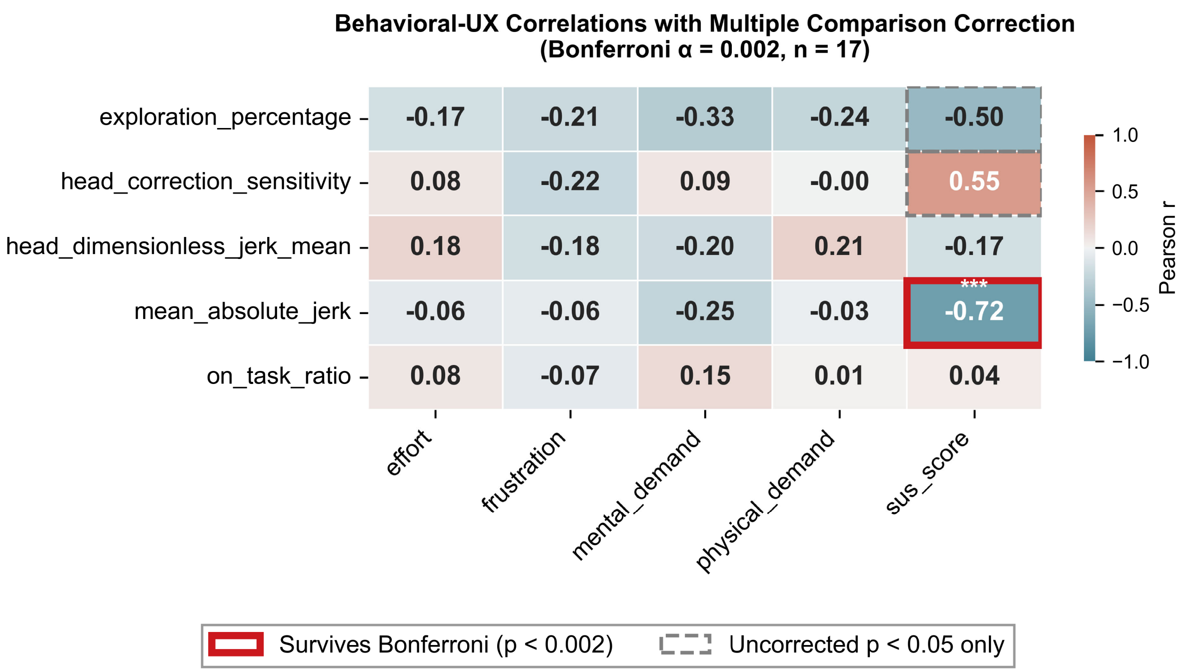


FIGURE 7. **Correlation matrix between behavioral metrics and UX measures with multiple comparison correction. Cell colors indicate correlation strength (blue=negative, red=positive). Thick red borders highlight correlations surviving Bonferroni correction ($\alpha = 0.002$); dashed gray borders indicate uncorrected $p < 0.05$ only. Only the Angular Jerk–SUS correlation ($r = -0.725$) survives rigorous correction.**

The combination of real-world visual anchoring, physical desk grounding, and visible real hands plausibly supports a coherent sensory experience during seated desk-based work. However, given the pilot-scale sample and the absence of

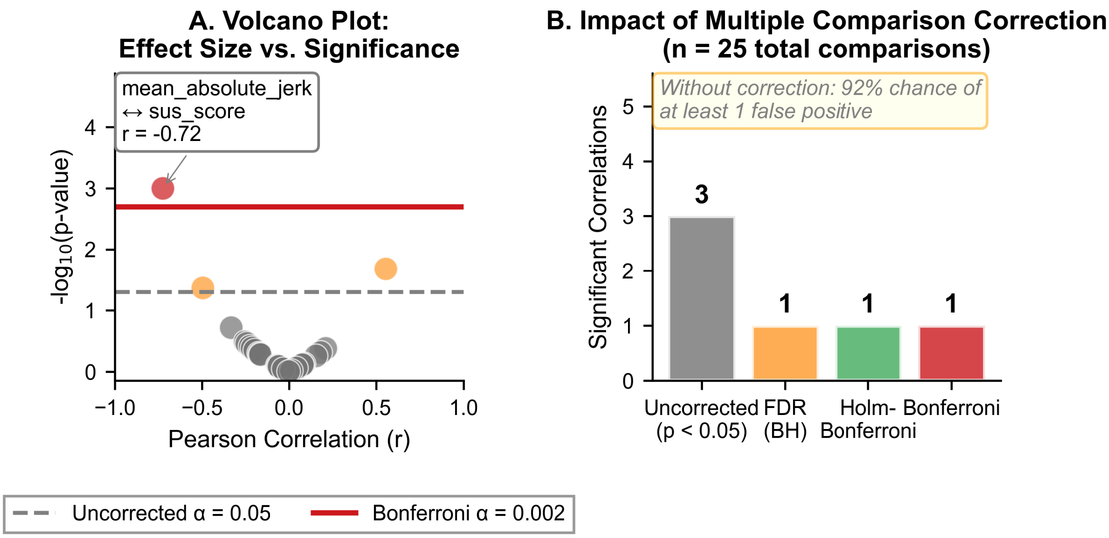


FIGURE 8. **Methodological robustness of correlation findings. (A) Volcano plot showing effect size (Pearson r) versus statistical significance ($-\log_{10}(p)$) for all 25 behavioral-UX correlations. Horizontal lines indicate uncorrected ($\alpha = 0.05$) and Bonferroni-corrected ($\alpha = 0.002$) significance thresholds. Only the Angular Jerk–SUS correlation (labeled) exceeds both thresholds. (B) Impact of multiple comparison correction: of 3 uncorrected significant correlations, only 1 survives all correction methods (Bonferroni, Holm, FDR), reducing the family-wise error rate from ≈ 72% to the nominal 5%.**

a fully immersive VR comparison condition, these results characterize the tolerability of this particular system rather than validate the passthrough approach in general. The SSQ subscale scores are presented in Figure 2.

D. SYSTEM USABILITY AND COMBINED ASSESSMENT

The System Usability Scale (SUS) assessment showed higher mean usability scores for XR (M=80.88, SD=14.74) than desktop (M=75.59, SD=17.84), though the difference was not statistically significant ($p = 0.360$) and the effect size was small ($d = 0.32$). The XR mean falls in the "Good" range (80–90) and the desktop mean in the "OK" range (70–80); these descriptive bands should not be over-interpreted given the small per-group sample (n = 17). Combined with low cybersickness scores, this indicates that the XR system was at least as usable as the desktop interface in our sample. Figure 2 shows the comprehensive assessment including SUS, SSQ, and NASA-TLX measurements, illustrating the distributions and comparative patterns across all traditional subjective measures.

E. MULTIMODAL FUSION ANALYSIS

Analysis of head-hand coordination revealed patterns of multimodal integration during XR interaction. The Coordination Index averaged 0.169 (SD=0.041), indicating low temporal synchronization between head and hand movements—participants tended to move head and hands independently rather than in coordinated sequences. The Phase Locking Value (PLV) averaged 0.326 (SD=0.067), suggesting weak-to-moderate phase synchronization between head and hand movement signals. Mean spectral coherence was low (M=0.019, SD=0.009), indicating that head and hand movements operated at different frequency components.

Notably, bilateral hand symmetry was high (M=0.958, SD=0.038), indicating that both hands coordinated similarly with head movements when engaged. This asymmetry pattern suggests that users primarily relied on one hand for interaction while maintaining head orientation for visual attention, consistent with the passthrough XR design that anchored UI controls to a physical desk surface. Figure 9 visualizes these coordination patterns.

F. PREDICTIVE MODELING BOUNDARY

We also tested whether multivariate machine learning models could predict UX scores from behavioral features using Leave-One-Out cross-validation. All tested models performed below a simple mean baseline (e.g., SUS: $R^2 = -0.123$, Mental Demand: $R^2 = -0.257$), indicating that the current n=17 XR sample is not sufficient for reliable multivariate prediction. We therefore do not use these models to support the manuscript's main claims. Detailed model settings and feature-importance outputs are moved to the supplementary material as a negative-result analysis documenting the sample-size boundary for future predictive work [58].

V. DISCUSSION

A. MAIN FINDING: ONE CORRECTED MOVEMENT-FLUENCY ASSOCIATION

The most defensible empirical result is the within-XR association between Head Angular Jerk and SUS usability scores ($r = -0.725, p_{\text{Bonferroni}} = 0.025$). This finding suggests that,

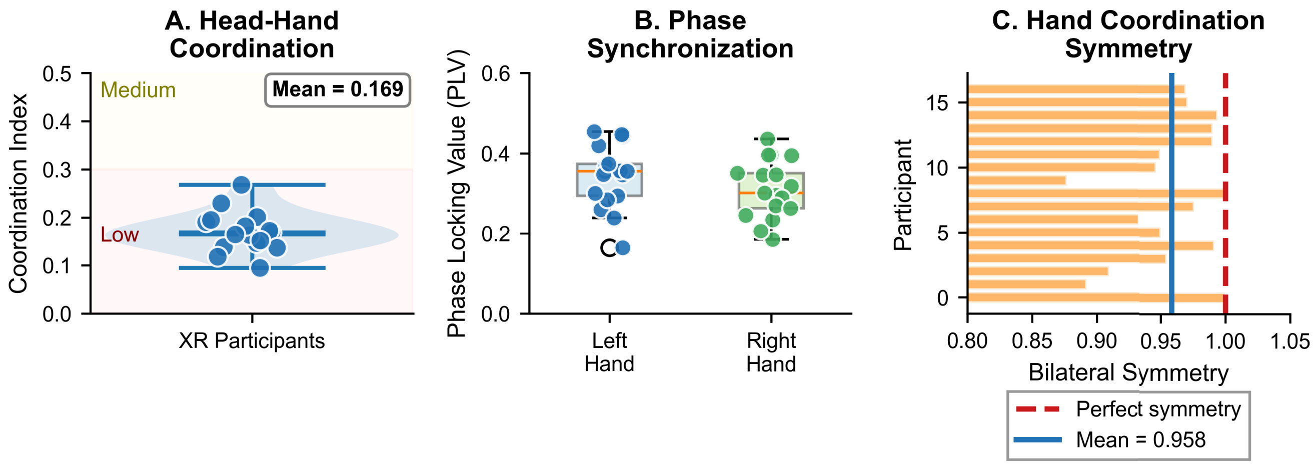


FIGURE 9. **Multimodal fusion analysis of head-hand coordination during XR interaction. (A) Coordination Index distribution showing low temporal synchronization (M=0.169), with interpretation zones (low<0.3, medium 0.3–0.6, high>0.6). (B) Phase Locking Value (PLV) comparison between left and right hand coordination with head movements, showing weak-to-moderate phase synchronization. (C) Bilateral hand symmetry scores indicating consistent coordination patterns across both hands (M=0.958).**

in this passthrough XR CNN visualization task, smoother head movement was associated with higher perceived usability. The result is consistent with embodied-interaction accounts in which cognitive uncertainty or interaction friction can appear in movement patterns, but it remains correlational. Head Angular Jerk should therefore be treated as a candidate passive indicator rather than as a validated general-purpose usability measure.

B. CONTEXTUAL INTERFACE COMPARISON

The XR-vs-desktop comparison did not provide statistically significant evidence that XR was superior on standard UX outcomes. SUS, satisfaction, NASA-TLX, and Visual Attention Focus results are best interpreted as descriptive context for the learning module. The one significant custom-question result, visualization usefulness ($p = 0.047$), suggests that participants perceived the XR visualization as useful in this implementation, but the small sample and multiple exploratory comparisons prevent broad platform claims. This narrower interpretation aligns the interface comparison with its role in the paper: it contextualizes the XR learning task rather than establishing a general XR advantage.

C. WHY FOUR BEHAVIORAL FEATURES DID NOT SURVIVE CORRECTION

Four of the five behavioral features in the corrected correlation matrix did not show statistically significant corrected associations with subjective UX measures. We interpret these null corrected findings as informative rather than as simple failures of measurement. First, Exploration Percentage has ambiguous psychological direction. Wider head-orientation coverage may reflect curiosity and active inspection, but it may also reflect visual search difficulty, disorientation, or uncertainty about where relevant content is located. Without task-phase annotations, the same exploration value can mix productive investigation with inefficient search.

Second, Head Correction Sensitivity is a derived kinematic feature that may capture individual movement strategy or local corrective adjustments rather than perceived usability in a stable way. Third, Head Dimensionless Jerk Mean normalizes movement smoothness for duration and amplitude, but this normalization may remove some of the absolute movement-fluency variation that relates to subjective usability in this task. Fourth, Visual Attention Focus used head direction as a coarse gaze proxy because the Quest 3 does not include integrated eye tracking. Head-based attention can miss eye-only saccades and central-field inspection, and the cross-platform comparison further involves different sensing modalities for XR and desktop.

Across all four non-surviving features, the XR sample size (n=17), conservative correction across 25 tests, construct ambiguity, and task-specific movement constraints limit the ability to detect anything except large and consistent effects. Hand-posture and arm-elevation summaries remain useful descriptive ergonomics variables, but consumer markerless hand-tracking error and the seated desk-anchored task make them inappropriate as confirmatory evidence in this dataset. For these reasons, the non-corrected metrics should be treated as candidates requiring larger samples, personalized baselines, task-phase labels, and more precise sensing before operational use.

D. METHODOLOGICAL BOUNDARY

The multiple-comparison correction is central to the interpretation of this study. Without correction, testing 25 correlations would produce a family-wise error probability of approximately $1 - (1 - 0.05)^{25} \approx 72\%$. After Bonferroni, Holm, and FDR correction, only the Angular Jerk–SUS correlation remains. This result narrows the paper's contribution: it identifies one robust candidate association and a set of secondary exploratory signals, not a complete behavioral-signature framework. Similarly, the predictive modeling analysis performed below baseline at n=17; detailed results are reported in the supplementary material to document this boundary without using failed models as support for the main claim.

E. TASK AND IMPLEMENTATION SPECIFICITY

The observed association may be specific to the current implementation: a seated, passthrough XR CNN visualization task on Meta Quest 3 with desk-anchored interaction. CNN visualization is spatial and layered, so head movement may be especially relevant to content inspection. Other learning contexts–text-heavy instruction, procedural skills, collaborative problem solving, or abstract reasoning–may produce different movement patterns and different relationships between behavioral metrics and UX. Hardware and interaction design also matter: a different headset, eye-tracking capability, hand-tracking quality, or content layout could change the meaning of the same metric. Future work should therefore validate candidate indicators across tasks, hardware platforms, and interaction designs before generalizing them.

F. IMPLICATIONS FOR BEHAVIORAL ASSESSMENT

The results suggest that movement fluency deserves further study as a passive UX indicator in passthrough XR learning environments. However, none of the behavioral metrics should be treated as a substitute for learning outcomes or validated questionnaires at this stage. A practical assessment pipeline should combine subjective UX instruments, behavioral monitoring, and separate learning-performance measures. Larger, preregistered studies are needed to determine whether movement fluency can support real-time adaptation or diagnostic feedback.

VI. LIMITATIONS

This study has several limitations. First, the sample is small (N = 34, n = 17 XR), making the study best understood as a pilot-scale exploratory investigation. The design is underpowered for small-to-medium effects and makes multivariate prediction inappropriate for confirmatory claims. Second, the task is a single CNN visualization module; movement patterns from this spatial learning task may not transfer to other educational domains. Third, the implementation is hardware- and design-specific: results may differ with other passthrough headsets, eye-tracking hardware, hand-tracking quality, seating arrangements, or virtual layout designs. Fourth, Quest 3 markerless hand tracking introduces centimeter-scale positional error, jitter, latency, and condition-dependent accuracy [19], [36]. These factors particularly constrain interpretation of hand-posture, arm-ergonomic, and fine-grained hand-kinematic metrics. Fifth, Visual Attention Focus relies on head direction as a coarse gaze proxy because Quest 3 lacks integrated eye tracking. Sixth, cross-platform behavioral comparisons are limited by sensing differences between XR tracking and webcam/mouse-based desktop tracking, even though all behavioral streams were analyzed at 10 Hz. Finally, the short session duration may not capture longer-term adaptation, fatigue, or learning effects.

VII. CONCLUSION

We presented a pilot-scale exploratory study of passive behavioral indicators for user experience in a passthrough XR learning environment. The XR-vs-desktop comparison provides contextual evidence for the CNN visualization task, but does not establish broad XR superiority. The primary result is a corrected within-XR association between Head Angular Jerk and SUS usability scores, suggesting that movement fluency is a promising candidate indicator for future validation in similar spatial learning tasks.

For designers of XR learning systems, the practical implication is not to replace validated questionnaires or learning assessments with passive tracking. Instead, designers can use movement fluency as a diagnostic signal to identify moments of possible interaction friction, pair behavioral logs with subjective UX and learning-outcome measures, and design passthrough interactions that preserve physical grounding, reduce cybersickness risk, and avoid sustained uncomfortable arm postures. Hand-derived metrics should be interpreted conservatively when using consumer markerless tracking, especially for small posture differences or latency-sensitive interactions.

The research agenda is therefore concrete: replicate the movement-fluency association with larger preregistered samples, test whether it generalizes across learning domains and headset platforms, add task-phase annotation to distinguish productive exploration from confusion, incorporate eye tracking or more accurate hand sensing when available, and validate candidate metrics against both subjective experience and independent learning outcomes. Until such validation exists, behavioral signatures in passthrough XR should be treated as candidate indicators for exploratory assessment rather than as deployable general-purpose UX measures.

ACKNOWLEDGMENT

Any opinions, findings, and conclusion or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] I. B. Adhanom, P. MacNeilage, and E. Folmer, "Eye tracking in virtual reality: A broad review of applications and challenges," *Virtual Reality*, vol. 27, no. 2, pp. 1481–1505, Jun. 2023.

[2] T. Baltrusaitis, P. Robinson, and L.-P. Morency, "OpenFace: An open source facial behavior analysis toolkit," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–10.

[3] G. E. Baykal, A. Leylekoğlu, S. Arslan, and D. Ozer, "Studying children's object interaction in virtual reality: A manipulative gesture taxonomy for VR hand tracking," in *Proc. Extended Abstr. CHI Conf. Human Factors Comput. Syst.*, New York, NY, USA, Apr. 2023, pp. 1–7.

[4] (2023). *Beam Eye Tracker SDK*. [Online]. Available: https://docs.beam.eyeware.tech/

[5] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *J. Roy. Stat. Soc. Ser. B, Stat. Methodology*, vol. 57, no. 1, pp. 289–300, Jan. 1995.

[6] P. Blikstein and M. Worsley, "Multimodal learning analytics and education data mining: Using computational technologies to measure complex learning tasks," *J. Learn. Analytics*, vol. 3, no. 2, pp. 220–238, Sep. 2016.

[7] D. A. Bowman, J. L. Gabbard, and D. Hix, "A survey of usability evaluation in virtual environments: Classification and comparison of methods," *Presence*, vol. 11, no. 4, pp. 404–424, Aug. 2002.

[8] J. Brooke, "Sus—A quick and dirty usability scale," in *Usability Evaluation in Industry*. New York, NY, USA: Taylor & Francis, 1996, pp. 189–194.

[9] J. A. Bueno-Vesga, X. Xu, and H. He, "The effects of cognitive load on engagement in a virtual reality learning environment," in *Proc. IEEE Virtual Reality 3D User Interfaces (VR)*, Mar. 2021, pp. 645–652.

[10] D. B. Chaffin, G. B. J. Andersson, and B. J. Martin, *Occupational Biomechanics*. Hoboken, NJ, USA: Wiley, 2006.

[11] A. Chrysanthakopoulou, T. Chrysikopoulos, L.-N. Arvanitopoulos, and K. Moustakas, "Beyond euclid: An educational VR journey into spherical geometry," in *Proc. IEEE Conf. Virtual Reality 3D User Interfaces Abstr. Workshops (VRW)*, Mar. 2024, pp. 1100–1101.

[12] V. Clay, P. König, and S. U. König, "Eye tracking in virtual reality," *J. Eye Movement Res.*, vol. 12, no. 1, p. 3, Apr. 2019.

[13] C. Dede, "Immersive interfaces for engagement and learning," *Science*, vol. 323, no. 5910, pp. 66–69, Jan. 2009.

[14] J. Dul and B. Weerdmeester, *Ergonomics for Beginners: A Quick Reference Guide*, 3rd ed., New York, NY, USA: Taylor & Francis, 2008.

[15] O. J. Dunn, "Multiple comparisons among means," *J. Amer. Stat. Assoc.*, vol. 56, no. 293, pp. 52–64, Mar. 1961.

[16] C. Dupré, C. Appert, S. Rey, H. Saidi, and E. Pietriga, "TriPad: Touch input in AR on ordinary surfaces with hand tracking only," in *Proc. CHI Conf. Human Factors Comput. Syst.*, New York, NY, USA, May 2024, pp. 1–18.

[17] T. Flash and N. Hogan, "The coordination of arm movements: An experimentally confirmed mathematical model," *J. Neurosci.*, vol. 5, no. 7, pp. 1688–1703, Jul. 1985.

[18] M. Gallagher and E. R. Ferrè, "Cybersickness: A multisensory integration perspective," *Multisensory Res.*, vol. 31, no. 7, pp. 645–674, 2018.

[19] E. Godden, W. Steedman, and M. K. X. J. Pan, "Robotic characterization of markerless hand-tracking on meta quest pro and quest 3 virtual reality headsets," *IEEE Trans. Vis. Comput. Graphics*, vol. 31, no. 5, pp. 3025–3034, May 2025.

[20] J. F. Grafsgaard, J. B. Wiggins, K. E. Boyer, E. N. Wiebe, and J. C. Lester, "Automatically recognizing facial expression: Predicting engagement and frustration," in *Proc. 6th Int. Conf. Educ. Data Mining*, 2013, pp. 43–50.

[21] Y. Gu, M. A. Feijoo-Garcia, Y. Zhang, A. J. Magana, B. Benes, and V. Popescu, "An XR environment for ai education: Design and first implementation," in *Proc. IEEE Conf. Virtual Reality 3D User Interfaces Abstr. Workshops (VRW)*, Mar. 2024, pp. 157–162.

[22] M. A. Feijoo-Garcia, Y. Zhang, Y. Gu, A. J. Magana, B. Benes, and V. Popescu, "Exploring extended reality (XR) in teaching AI: A comparative study of XR and desktop environments," in *Proc. 9th Int. Conf. Human-Comput. Interact. Theory Appl. (HUCAPP)*, 2025, pp. 472–482, doi: 10.5220/0013141000003912.

[23] D. Hamilton, J. McKechnie, E. Edgerton, and C. Wilson, "Immersive virtual reality as a pedagogical tool in education: A systematic literature review of quantitative learning outcomes and experimental design," *J. Comput. Educ.*, vol. 8, no. 1, pp. 1–32, Mar. 2021.

[24] S. Han, B. Liu, R. Cabezas, C. D. Twigg, P. Zhang, J. Petkau, T.-H. Yu, C.-J. Tai, M. Akbay, Z. Wang, A. Nitzan, G. Dong, Y. Ye, L. Tao, C. Wan, and R. Wang, "MEgATrack: Monochrome egocentric articulated hand-tracking for virtual reality," *ACM Trans. Graph.*, vol. 39, no. 4, pp. 87:1–87:13, Aug. 2020.

[25] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (task load index): Results of empirical and theoretical research," *Adv. Psychol.*, vol. 52, pp. 139–183, 1988.

[26] N. Hogan and D. Sternad, "Sensitivity of smoothness measures to movement duration, amplitude, and arrests," *J. Motor Behav.*, vol. 41, no. 6, pp. 529–534, Jan. 2009.

[27] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandin. J. Statist.*, vol. 6, no. 2, pp. 65–70, 1979.

[28] Y. Hou, Q. Xie, N. Zhang, and J. Lv, "Cognitive load classification of mixed reality human computer interaction tasks based on multimodal sensor signals," *Sci. Rep.*, vol. 15, no. 1, Apr. 2025, Art. no. 13732.

[29] C. Javerliat, S. Villenave, P. Raimbaud, and G. Lavoué, "PLUME: Record, replay, analyze and share user behavior in 6DoF XR experiences," *IEEE Trans. Vis. Comput. Graphics*, vol. 30, no. 5, pp. 2087–2097, May 2024.

[30] M. A. Just and P. A. Carpenter, "Eye fixations and cognitive processes," *Cognit. Psychol.*, vol. 8, no. 4, pp. 441–480, Oct. 1976.

[31] N. Kamakura, *Postures Movement Patterns Human Hand: A Framework for Understand. Hand Activity for Clinicians Engineers*, Irvine, CA, USA: BrownWalker Press, 2022.

[32] R. S. Kennedy, N. E. Lane, K. S. Berbaum, and M. G. Lilienthal, "Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness," *Int. J. Aviation Psychol.*, vol. 3, no. 3, pp. 203–220, Jul. 1993.

[33] S. Klingenberg, R. Bosse, R. E. Mayer, and G. Makransky, "Does embodiment in virtual reality boost learning transfer? Testing an immersion-interactivity framework," *Educ. Psychol. Rev.*, vol. 36, no. 4, Oct. 2024, Art. no. 116.

[34] D.-H. Kwak, X. Ma, and S. Kim, "When does social desirability become a problem? Detection and reduction of social desirability bias in information systems research," *Inf. Manage.*, vol. 58, no. 7, Nov. 2021, Art. no. 103500.

[35] X. Lin, R. Li, Z. Chen, and J. Xiong, "Design strategies for VR science and education games from an embodied cognition perspective: A literature-based meta-analysis," *Frontiers Psychol.*, vol. 14, Jan. 2024, Art. no. 1292110.

[36] J. Lopez-Xarbau, M. A. Rodriguez-Fernandez, M. Faundez-Zanuy, J. Calvo-Sanz, and J. J. Garcia-Tirado, "Robot-driven calibration and accuracy assessment of meta quest 3 inside-out tracking using a TECHMAN TM5–900 collaborative robot," *Sensors*, vol. 26, no. 8, p. 2285, Apr. 2026.

[37] G. Makransky and G. B. Petersen, "The cognitive affective model of immersive learning (CAMIL): A theoretical research-based model of learning in immersive virtual reality," *Educ. Psychol. Rev.*, vol. 33, no. 3, pp. 937–958, Sep. 2021.

[38] (2023). *Hand Tracking Developer Guide for Quest 3*. [Online]. Available: https://developers.meta.com/horizon/documentation/unity/unity-handtracking-overview/

[39] H.-H. Mira, R. Chaker, I. Maria, and H. Nady, "Review of research on the outcomes of embodied and collaborative learning in STEM in higher education with immersive technologies," *J. Comput. Higher Educ.*, vol. 37, no. 3, pp. 1161–1198, Sep. 2025.

[40] J. Moreno-Arjonilla, A. López-Ruiz, J. R. Jiménez-Pérez, J. E. Callejas-Aguilera, and J. M. Jurado, "Eye-tracking on virtual reality: A survey," *Virtual Reality*, vol. 28, no. 1, Feb. 2024, Art. no. 38.

[41] M. Nasri, M. Kosa, L. Chukoskie, M. Moghaddam, and C. Harteveld, "Exploring eye tracking to detect cognitive load in complex virtual reality training," in *Proc. IEEE Int. Symp. Mixed Augmented Reality Adjunct*, Oct. 2024, pp. 51–54.

[42] R. L. Oliver, "A cognitive model of the antecedents and consequences of satisfaction decisions," *J. Marketing Res.*, vol. 17, no. 4, pp. 460–469, Nov. 1980.

[43] J. Radianti, T. A. Majchrzak, J. Fromm, and I. Wohlgenannt, "A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda," *Comput. Educ.*, vol. 147, Apr. 2020, Art. no. 103778.

[44] F. F. Reichheld, "The one number you need to grow," *Harvard Bus. Rev.*, vol. 81, no. 12, pp. 46–54, Dec. 2003.

[45] R. Richer, V. Koch, L. Abel, F. Hauck, M. Kurz, V. Ringgold, V. Müller, A. Küderle, L. Schindler-Gmelch, B. M. Eskofier, and N. Rohleder, "Machine learning-based detection of acute psychosocial stress from body posture and movements," *Sci. Rep.*, vol. 14, no. 1, Apr. 2024, Art. no. 8251.

[46] R. Rosenthal, "Parametric measures of effect size," in *The Handbook of Research Synthesis*. New York, NY, USA: Russell Sage, 1994, pp. 231–244.

[47] A. K. Singh, "VR/AR in ergonomics and workspace design: A dual-perspective analysis of applications and implications," *Appl. Ergonom.*, vol. 129, Nov. 2025, Art. no. 104612.

[48] V. Sitzmann, A. Serrano, A. Pavel, M. Agrawala, D. Gutierrez, B. Masia, and G. Wetzstein, "Saliency in VR: How do people explore virtual environments?" *IEEE Trans. Vis. Comput. Graphics*, vol. 24, no. 4, pp. 1633–1642, Apr. 2018.

[49] C. L. Taborda, H. Nguyen, and P. Bourdot, "Engagement and attention in XR for learning: Literature review," in *Proc. Int. Conf. Virtual Reality Mixed Reality*, 2025, pp. 207–242.

[50] T. Tene, J. A. Marcatoma Tixi, M. D. L. Palacios Robalino, M. J. Mendoza Salazar, C. Vacacela Gomez, and S. Bellucci, "Integrating immersive technologies with STEM education: A systematic review," *Frontiers Educ.*, vol. 9, Jun. 2024, Art. no. 1410163.

[51] X. Wang, B. Lafreniere, and J. Zhao, "Exploring visualizations for precisely guiding bare hand gestures in virtual reality," in *Proc. CHI Conf. Human Factors Comput. Syst.*, New York, NY, USA, 2024, pp. 1–19.

[52] N. Wenk, J. Penalver-Andres, K. A. Buetler, T. Nef, R. M. Müri, and L. Marchal-Crespo, "Effect of immersive visualization technologies on cognitive load, motivation, usability, and embodiment," *Virtual Reality*, vol. 27, no. 1, pp. 307–331, Mar. 2023.

[53] M. Wilson, "Six views of embodied cognition," *Psychonomic Bull. Rev.*, vol. 9, no. 4, pp. 625–636, Dec. 2002.

[54] R.-L. Zhang, S.-Y. Yu, R.-F. Gao, L.-F. Wang, and X.-C. He, "How does technology-based embodied learning affect learning effectiveness?–based on a systematic literature review and meta-analytic approach," *Interact. Learn. Environments*, vol. 33, no. 6, pp. 4135–4158, Apr. 2025.

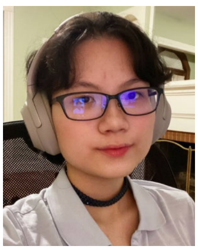
[55] Y. Zhang, M. A. Feijoo-Garcia, Y. Gu, V. Popescu, B. Benes, and A. J. Magana, "Virtual and augmented reality in science, technology, engineering, and mathematics (STEM) education: An umbrella review," *Information*, vol. 15, no. 9, p. 515, Aug. 2024.

[56] L. Zhu and J. Lv, "Review of studies on user research based on EEG and eye tracking," *Appl. Sci.*, vol. 13, no. 11, p. 6502, May 2023.

[57] L. Zou, Z. Zhang, M. Mavilidi, Y. Chen, F. Herold, K. Ouwehand, and F. Paas, "The synergy of embodied cognition and cognitive load theory for optimized learning," *Nature Human Behaviour*, vol. 9, no. 5, pp. 877–885, Mar. 2025.

[58] A. Vabalas, E. Gowen, E. Poliakoff, and A. J. Casson, "Machine learning algorithm validation with a limited sample size," *PLoS ONE*, vol. 14, no. 11, Nov. 2019, Art. no. e0224365.

[59] F. Lacquaniti, C. Terzuolo, and P. Viviani, "The law relating the kinematic and figural aspects of drawing movements," *Acta Psychologica*, vol. 54, nos. 1–3, pp. 115–130, Oct. 1983.



YIYIN GU is currently pursuing the Ph.D. degree in computer science with Purdue University, West Lafayette, IN, USA.

They are also a Graduate Research Assistant with the Computer Graphics and Visualization Laboratory (CGVLab), Department of Computer Science, Purdue University. Their research interests include extended reality (XR), virtual reality, and their applications in STEM education.

YIQUN ZHANG received the B.S. and M.S. degrees in computer and information technology from Purdue University, in 2021 and 2023, respectively, where she is currently pursuing the Ph.D. degree in technology with the School of Applied and Creative Computing.

Her primary research interest includes immersive technology-based education.

MIGUEL A. FEIJOO-GARCIA received the M.Sc. degree in information engineering and the M.Ed. degree in higher education teaching. He is currently pursuing the Ph.D. degree in computer and information technology with Purdue University, West Lafayette, IN, USA.

He was awarded a Fellowship by Purdue Polytechnic Institute to support his doctoral studies. He received the Best Poster Award at the International Conference on Computer-Human Interaction Research and Applications (CHIRA). His research interests include computer science education, learning analytics, and artificial intelligence applications in education.

ALEJANDRA J. MAGANA (Associate Member, IEEE) received the B.E. degree in information systems and the M.S. degree in technology from Tecnológico de Monterrey, Mexico, and the M.S. degree in educational technology and the Ph.D. degree in engineering education from Purdue University, West Lafayette, IN, USA.

She is currently the W. C. Furnas Professor in enterprise excellence with the Department of Computer and Information Technology and a Professor with the School of Engineering Education, Purdue University. Her research investigates how model-based cognition in STEM can be better supported through data science computation, modeling, and simulation. She received NSF CAREER Award, in 2015. She was inducted into Purdue University Teaching Academy, in 2022.

BEDRICH BENES (Senior Member, IEEE) received the M.S. and Ph.D. degrees in computer science from Czech Technical University in Prague, Prague, Czech Republic, in 1991 and 1998, respectively.

He is currently a Professor and the Associate Head of Computer Science with Purdue University, West Lafayette, IN, USA. He has published over 200 research papers in computer graphics. His research interests include procedural and inverse procedural modeling, simulation of natural phenomena, deep learning for geometry synthesis, and additive manufacturing. He is a fellow of European Association for Computer Graphics (Eurographics), a Purdue University Faculty Scholar, and the Editor-in-Chief of *Graphical Models*.

VOICU POPESCU (Member, IEEE) received the B.S. degree in computer science from the University of Cluj-Napoca, Romania, in 1995, and the Ph.D. degree in computer science from the University of North Carolina at Chapel Hill.

He is currently an Associate Professor in computer science with Purdue University. His research interests include virtual and augmented reality, computer graphics, and visualization.