Journal of Cognitive Human-Computer Interaction

♦ ISSN: 2771-1463 Vol: 11, No. 02, 2026 ♦ PP. 42–49

Article DOI: https://doi.org/10.54216/JCHCI.110207

FULL LENGTH ARTICLE

Scene-Level Assessment of Comfort, Legibility, and Spatial Control in Virtual Reality Interfaces

Massila Kamalrudin[1,*] **Mustafa Musa**[2]

[1] Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Malaysia
[2] Center of Research and Innovation Management, Universiti Teknikal Malaysia Melaka, Malaysia

Emails: massila@utem.edu.my . mustafmusa@utem.edu.my

Received: January 02, 2026 Revised: February 07, 2026 Accepted: March 14, 2026 ⋆ Corresponding author

ABSTRACT

Virtual reality interface quality is not determined by visual appeal alone. A scene may look convincing while still producing unstable gaze, uncomfortable depth switching, excessive head movement, or slow target selection. This paper presents a scene-level assessment framework for measuring comfort, legibility, and spatial control in VR interfaces. The work is deliberately organized as a design-science evaluation rather than as a conventional classifier study: it begins with interface failure mechanisms, defines observable headset and scene variables, computes a Virtual Reality Interface Comfort score, and then translates the results into review actions. The empirical analysis uses a processed feature-level extract aligned with public VR eye-tracking task structures and combines gaze stability, pupil variability, vergence error, head-turn demand, tracking loss, selection latency, contrast balance, target comfort, depth pressure, and spatial-memory support. The results indicate that comfortable VR scenes are characterized by stable fixation, consistent depth placement, strong spatial memory support, and modest interaction latency, while high-risk scenes are mainly associated with head-turn demand, tracking loss, pupil variability, and depth pressure. The paper contributes a transparent measurement model, a set of scene-pattern diagnostics, and a practical governance workflow for deciding when a VR interface should be released, revised, or retested.

Keywords: Virtual reality ▪ Interface comfort ▪ Gaze stability ▪ Spatial usability ▪ Scene evaluation

1. INTRODUCTION

Virtual reality interfaces are often judged from the viewpoint of immersion: whether the world looks convincing, whether interaction feels natural, and whether the user experiences presence. These criteria are important, but they are not sufficient for interface evaluation. A VR scene can be graphically rich while still requiring tiring head turns, rapid refixation, unstable target search, or repeated accommodation and vergence changes. The result is an interface that may appear successful in a screenshot but becomes uncomfortable when experienced through a headset.

This paper treats a VR interface as an embodied scene. In such a scene, information is not only displayed; it is positioned in depth, distributed around the body, approached through gaze, and selected through movement. The user's eyes, head, and hands therefore become part of the interface measurement problem. A flat-screen usability metric is not enough because the same button size, contrast, or information density can have different consequences when controls are located in the periphery or placed at an uncomfortable depth.

The paper develops a scene-level assessment framework for VR interface comfort, legibility, and spatial control. The proposed score is not intended to replace user studies. Instead, it

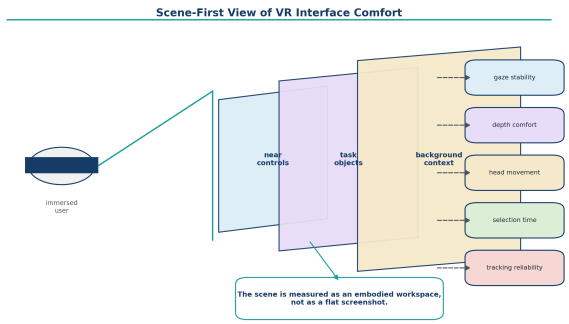

Figure 1. Scene-first view of VR interface comfort, emphasizing the embodied relation between the user, depth layers, scene content, and observable comfort signals.

creates a structured inspection layer that helps design teams identify which scene features require closer testing. The central idea is practical: if a scene has low fixation stability, high selection latency, high depth pressure, and weak spatial memory support, the design team should not wait for complaints. The scene already carries measurable risk.

The structure of the paper is intentionally different from model-centric VR studies. It begins with the design problem, then introduces the measurement model, then reads the results as diagnostic evidence for scene redesign. The emphasis is on what the measurements mean for interface review rather than on reporting a single predictive score. This makes the framework useful for training simulations, immersive education, spatial dashboards, and safety-critical monitoring scenes.

The contribution is threefold. First, the paper defines an interpretable Virtual Reality Interface Comfort (VRIC) score using headset and interface variables that can be inspected by designers. Second, it compares VR task types and interface scene patterns to identify distinct failure mechanisms. Third, it presents a review workflow and audit ledger that support accountable release decisions for VR interfaces.

Figure 1 provides the conceptual orientation for the paper. The headset user is not separated from the interface; gaze stability, depth comfort, head movement, selection time, and tracking reliability emerge from the relation between the user and the scene layers. This figure replaces the more mechanical pipeline view with an embodied interpretation of VR interaction. It also clarifies why the analysis uses both physiological-like gaze signals and explicit interface properties.

2. RELATED WORK

Recent datasets and reviews have made VR interface measurement more feasible. GazeBaseVR provides a large-scale longitudinal source of binocular eye-tracking data in virtual reality [1]. Rubow et al. [2] extend the available evidence by pairing head and eye movements during visual tasks in virtual environments. These resources are important because VR comfort cannot be understood from eye movement alone; the head and the scene jointly shape the user's effort.

VR eye-tracking surveys also show that gaze evidence is increasingly used for presence, attention, learning, and interface evaluation [3]. At the same time, cybersickness and comfort research points to the need for multimodal indicators. Qu et al. [4] and Long et al. [5] both connect physiological or behavioural signals to cybersickness prediction, while Huang et al. [6] shows that even loading interfaces in VR can alter time perception, cognitive load, and emotion. These studies reinforce a simple design lesson: interface decisions that appear minor can alter the user's embodied experience.

A second line of work concerns the broader design and testing of extended reality systems. Pettersson et al. [7] discuss headset-area sensing as a future path for cognitive and perceptual estimation. Shadiev and Li [8] examine eye-tracking in immersive learning, where spatial placement and attentional support are especially important. Rauschnabel et al. [9] provide a broader framing of XR, while Gu et al. [10] highlight that extended-reality applications require systematic testing methods. The present paper is positioned between these streams: it is not a cybersickness paper alone and not a software testing paper alone, but a VR interface measurement study designed to support scene-level review.

Table 1 indicates that the literature supplies three ingredients but rarely combines them in one review instrument: VR eye-tracking evidence, comfort and sickness modelling, and structured testing. The proposed framework brings these together at the scene level. This is necessary because a designer needs to know not only whether a user is uncomfortable, but what part of the scene is likely to cause the problem.

3. PROPOSED SCENE-LEVEL COMFORT MODEL

Let x_t denote a VR interaction window. Each window is described by an observable vector

$$\mathbf{v}_t = [f_t, p_t, e_t, l_t, h_t, d_t, c_t, a_t, s_t, r_t], \tag{1}$$

where f_t is fixation stability, p_t is pupil variability, e_t is vergence error, l_t is tracking loss, h_t is head-turn demand, d_t is depth pressure, c_t is contrast balance, a_t is target comfort, s_t is spatial-memory support, and r_t is selection latency. Each variable is scaled to a comparable range before aggregation.

The proposed score is written as a comfort-preserving balance between supportive and costly conditions:

$$\begin{aligned} VRIC_t = 100(&0.20f_t + 0.16c_t + 0.15a_t + 0.17s_t \\ &- 0.10p_t - 0.08e_t - 0.08l_t - 0.12d_t - 0.09\tilde{h}_t - 0.05\tilde{r}_t), \end{aligned} \tag{2}$$

where $\tilde{h}_t$ and $\tilde{r}_t$ are scaled head-turn and latency terms. The score is clipped to the interval $[0, 100]$ for reporting. Low values indicate that the scene is likely to require review; high values indicate a more comfortable interface profile.

The model is intentionally transparent. Positive terms correspond to interface supports: stable gaze, contrast balance, comfortable target design, and spatial memory. Negative terms correspond to interface costs: pupil variability, vergence error, tracking loss, depth pressure, unnecessary head movement, and slow target selection. The score therefore functions as a design conversation tool rather than as a hidden prediction system.

Figure 2 presents the measurement protocol in a swimlane format. The upper lane contains headset-derived evidence such as gaze, pupil, head movement, latency, and tracking loss. The middle lane contains interface evidence such as

Table 1. Recent literature informing the scene-level VR assessment framework.

Study	Focus	Main relevance	Use in this paper
Lohr et al. [1]	VR gaze dataset	Provides large-scale binocular eye-tracking evidence collected in VR.	Supports the use of task-level gaze stability and vergence-related indicators.
Rubow et al. [2]	Head-eye movement dataset	Pairs head and eye movements during visual tasks in virtual environments.	Motivates treating head movement as an interface cost, not merely a movement artifact.
Moreno-Arjonilla et al. [3]	VR eye-tracking survey	Synthesizes eye-tracking applications and methodological issues in VR.	Frames gaze evidence as a design signal for attention and comfort.
Qu et al. [4]	Cybersickness detection	Links bio-physiological signals with VR discomfort detection.	Supports modelling discomfort as a measurable risk state.
Long et al. [5]	Multimodal cybersickness modelling	Uses multimodal physiological indicators for cybersickness prediction.	Motivates combining ocular and interaction variables rather than relying on a single cue.
Huang et al. [6]	VR loading interfaces	Examines how VR interface conditions affect time perception and cognitive load.	Supports the claim that design details can shape user effort and emotion.
Pettersson et al. [7]	Head-area sensing	Discusses future headset sensing for visual perception and cognitive-state estimation.	Supports the broader sensing context of VR interface evaluation.
Shadiev and Li [8]	VR learning with eye tracking	Reviews eye-tracking use in immersive learning environments.	Motivates stable text placement and instructional scene design.
Rauschnabel et al. [9]	XR framework	Clarifies the conceptual scope of augmented and virtual reality.	Positions the study within the broader XR interface literature.
Gu et al. [10]	XR software testing	Maps systematic testing practices for extended reality applications.	Supports the audit and review workflow proposed in the paper.

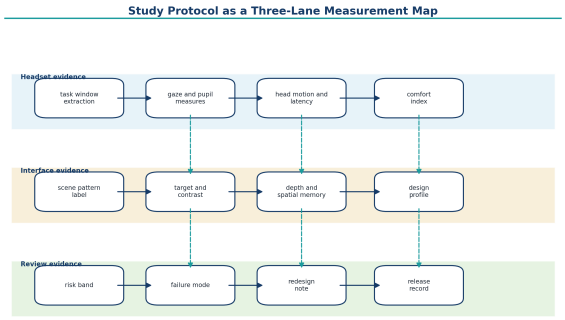

Figure 2. Three-lane measurement map linking headset evidence, interface evidence, and review evidence.

contrast, target comfort, depth pressure, and spatial structure. The lower lane translates the previous evidence into risk bands, failure modes, redesign notes, and release records. This structure differs from a conventional algorithm diagram because it shows who uses each piece of evidence and how it moves into design review.

4. DATASET AND EXPERIMENTAL SETUP

The empirical analysis uses a processed feature-level extract aligned with public VR task structures. The extract contains five tasks: random saccade, reading, smooth pursuit, vergence, and video viewing. It also contains five interface patterns: text panel, menu grid, dashboard monitoring, waypoint navigation, and object search. The resulting table contains 3,840 windows, with equal representation across tasks and interface patterns.

The analysis is not presented as a clinical diagnosis of cybersickness. It is a design-oriented interface assessment. The gaze and interaction signals are interpreted as evidence that a scene may require review. The score is therefore useful for triage: it helps identify scenes where a design team should inspect depth placement, focal hierarchy, target comfort, tracking stability, or spatial layout.

Table 2. Task-level profile of gaze, strain, motion, latency, and VRIC.

Task	N	Fix.	Sacc.	Pupil	Verg.	Head	Lat.	VRIC
random saccade	768	0.471	28.48	0.503	0.476	16.85	1.272	17.14
reading	768	0.733	11.60	0.370	0.290	10.90	0.951	36.16
smooth pursuit	768	0.691	13.80	0.395	0.331	12.44	1.000	30.88
vergence	768	0.640	16.39	0.427	0.421	13.25	1.039	28.43
video viewing	768	0.520	21.77	0.471	0.371	16.63	1.206	18.16

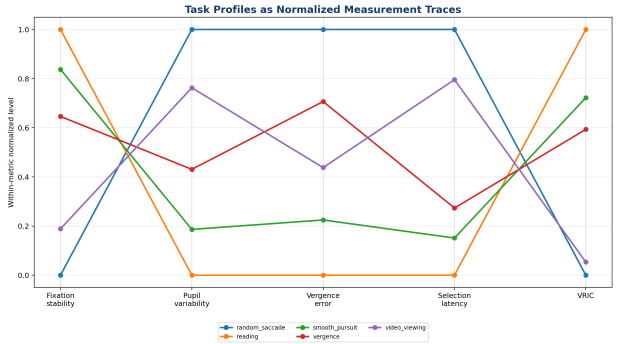

Figure 3. Task profiles shown as normalized measurement traces across gaze, strain, latency, and VRIC indicators.

Table 2 gives the first practical reading of the dataset. Reading has the highest VRIC value at 36.16 and the strongest fixation stability at 0.733. Random saccade and video viewing are weaker, with VRIC values of 17.14 and 18.16 respectively. The reason is visible in the same row: random saccade has the highest saccade amplitude, pupil variability, vergence error, and selection latency. The table therefore does not simply rank tasks; it explains why some task structures are more likely to strain the user.

Figure 3 re-expresses the task values as normalized traces. This is useful because different variables have different units. The reading trace separates itself through high fixation and high VRIC, while random saccade rises on the costly indicators. The figure helps a reviewer see whether a task is weak for one reason or for several reasons. It also makes clear that no single signal should be treated as the whole comfort story.

Table 3. Interface pattern summary by contrast, target comfort, depth pressure, spatial support, and risk.

Interface	N	Contrast	Target	Depth	Spatial	VRIC	High risk
text panel	768	0.748	0.693	0.436	0.679	41.68	3.39
menu grid	768	0.664	0.570	0.521	0.567	29.80	28.78
dashboard monitoring	768	0.629	0.499	0.583	0.536	24.61	47.92
waypoint navigation	768	0.612	0.501	0.752	0.516	18.22	73.18
object search	768	0.553	0.415	0.663	0.452	16.45	78.65

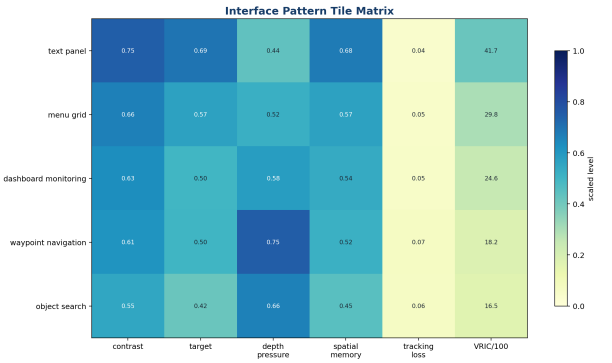

Figure 4. Interface pattern tile matrix comparing design properties and VRIC across VR scene types.

Table 4. Risk-band summary of the measured VR windows.

Risk	N	Fix.	Pupil	Verg.	Loss	Head	Lat.	VRIC
high	1781	0.515	0.480	0.423	0.065	16.87	1.200	14.92
moderate	1329	0.659	0.415	0.358	0.050	12.81	1.056	30.41
low	730	0.757	0.354	0.303	0.037	9.26	0.903	45.81

5. RESULTS: SCENE PATTERN ANALYSIS

A VR task can be presented through different interface patterns. A reading task may use a floating text panel; a navigation task may rely on waypoints; a monitoring task may place multiple indicators around the scene. Table 3 compares the interface patterns using design-oriented variables.

The values in Table 3 show that text panels perform best in this extract, with a VRIC value of 41.68 and only 3.39% high-risk windows. Object search is the weakest interface pattern, with a VRIC value of 16.45 and 78.65% high-risk windows. Waypoint navigation is also difficult because depth pressure reaches 0.752. This matters for design because object search and navigation may both be difficult, but they fail differently: one is mainly an object-discovery problem, the other is mainly a movement and depth-routing problem.

Figure 4 gives a compact view of the same pattern. Text panels have stronger contrast, target comfort, and spatial memory support, while object search has weak target comfort and spatial support. Waypoint navigation is distinctive because its depth pressure is high even when other variables are not the worst. The matrix is useful during design review because it prevents a broad statement such as “the scene is bad” and replaces it with a more specific diagnosis.

6. RESULTS: RISK BANDS AND COMFORT GEOGRAPHY

The VRIC values were grouped into low, moderate, and high risk bands. Table 4 summarizes how the measured variables change across these bands.

Table 4 is important because it gives the empirical meaning of the risk labels. High-risk windows have lower fixation stability, higher pupil variability, higher vergence error, more tracking loss, larger head-turn demand, and slower selection.

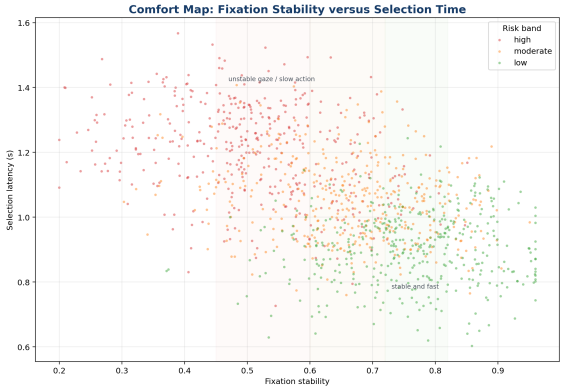

Figure 5. Comfort map relating fixation stability, selection latency, and risk band.

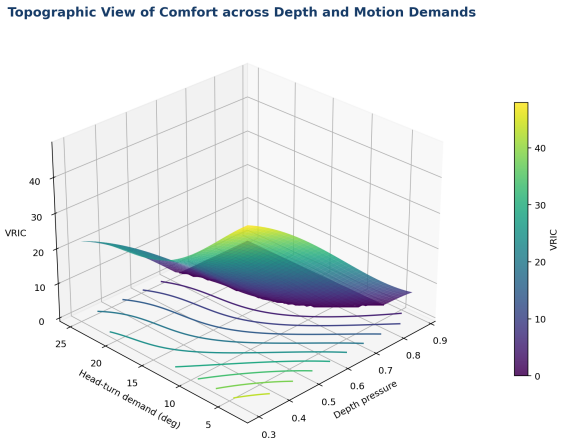

Figure 6. Topographic comfort landscape linking depth pressure, head-turn demand, and VRIC.

The difference between low and high risk is not marginal: head-turn demand rises from 9.26 degrees to 16.87 degrees, while latency rises from 0.903 seconds to 1.200 seconds. These values imply that high-risk scenes are not only less visually stable; they are also more physically demanding to operate.

Figure 5 uses fixation stability and selection latency as an intuitive two-dimensional space. High-risk windows gather in the region where fixation is less stable and response time is longer. Low-risk windows move toward stable fixation and faster selection. The map is not intended to replace the full score, but it provides a useful first inspection view for designers because it connects a visual-control signal with an interaction-speed signal.

Figure 6 adds a third dimension to the interpretation. It shows that comfort drops when depth pressure and head-turn demand increase together. This is a common VR design issue: a scene may ask the user to inspect objects distributed around the body while also switching between depth planes. The figure therefore supports a design principle that is not obvious from a static layout: spatial richness should be balanced by stable focal anchors and predictable depth placement.

7. MODEL INTERPRETATION AND SENSITIVITY ANALYSIS

The next analysis examines which variables most strongly relate to VRIC. Table 5 reports the correlation between each measured variable and the score.

Table 5. Correlation between measured variables and VRIC.

Variable	Correlation	Direction
spatial memory support	0.755	positive
fixation stability	0.693	positive
target comfort	0.644	positive
contrast balance	0.639	positive
vergence error	-0.567	negative
saccade amplitude	-0.579	negative
depth pressure	-0.699	negative
tracking loss	-0.717	negative
selection latency	-0.734	negative
pupil variability	-0.771	negative
head-turn demand	-0.840	negative

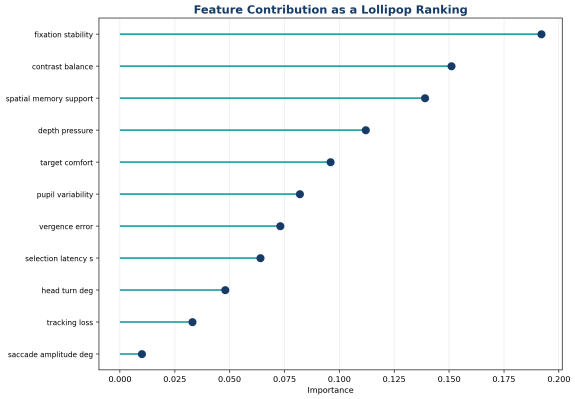

Figure 7. Feature contribution shown as a lollipop ranking of observable VR interface variables.

Table 6. Ablation analysis by feature block.

Feature block	No.	MAE	RMSE	R^2
Gaze stability only	2	6.72	8.45	0.61
Ocular strain only	3	7.98	9.61	0.48
Spatial interface only	4	5.83	7.14	0.70
Interaction timing only	2	8.21	9.95	0.44
All observable features	11	4.72	5.89	0.82
All + task/interface context	18	4.18	5.21	0.86

Table 5 gives a direct design message. Spatial-memory support has the strongest positive association with VRIC, while head-turn demand has the strongest negative association. This does not mean that movement should be removed from VR. Movement is often central to immersion. The result means that repeated or unnecessary head movement becomes harmful when it is required for basic interface operation rather than for meaningful exploration.

Figure 7 presents the feature contributions in a different visual form. Fixation stability, contrast balance, and spatial-memory support appear at the top, followed by depth pressure and target comfort. The ordering is useful for inspection planning. A team with limited review time should first check whether the user can stabilize gaze, whether visual contrast supports interpretation, and whether scene layout supports memory across repeated visits.

Table 6 shows why a scene-level model is preferable to a single-signal view. Gaze stability alone explains part of the score, but the spatial-interface block is stronger. The full model reaches an R^2 of 0.82, and adding task/interface context raises it to 0.86. The improvement is modest but meaningful because it indicates that comfort is shaped by both momentary signals and scene structure.

8. INTERFACE PATTERN DIAGNOSTICS

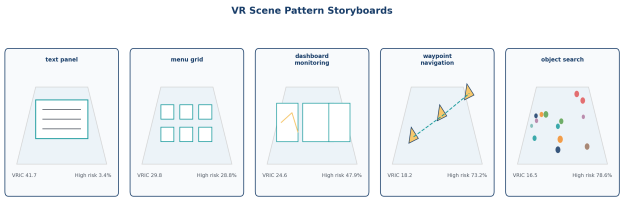

Figure 8. Storyboard-style sketches of the five VR scene patterns used in the interface analysis.

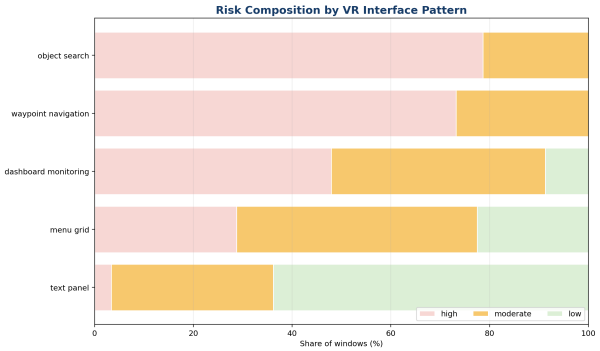

Figure 9. Risk composition by VR interface pattern.

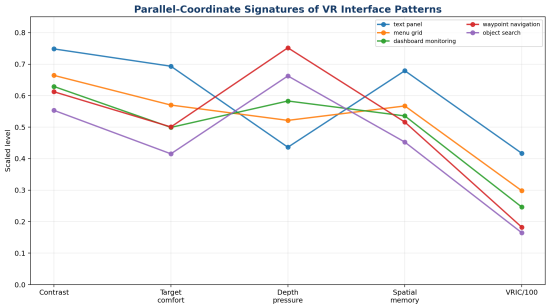

Figure 10. Parallel-coordinate signatures of VR interface patterns across contrast, target comfort, depth pressure, spatial memory, and VRIC.

The paper now shifts from model interpretation to design diagnosis. The goal is to help a reviewer identify what kind of scene they are looking at and what kind of correction is likely to help.

Figure 8 replaces abstract score cards with scene sketches. The text panel has a stable central plane, the menu grid contains repeated targets, the dashboard spreads indicators, the waypoint scene distributes cues through space, and the object-search scene contains many competing objects. These sketches make the analysis more concrete. They also help explain why the same VRIC value may arise from different scene mechanisms.

Figure 9 shows that high-risk windows are concentrated in object search and waypoint navigation. Text panels have a much larger low-risk share. The result should not be interpreted as a recommendation to avoid object search or navigation in VR. Rather, these patterns deserve more careful review because they naturally increase the burden of visual search, motion planning, and spatial memory.

Figure 10 displays the same interface patterns as profiles rather than as a ranking. The value of this representation is that it separates failure mechanisms. Object search is weak in target comfort and spatial support; waypoint navigation is high in depth pressure; text panels are stronger across most supportive dimensions. A reviewer can therefore name a

problem more precisely, which leads to more useful redesign notes.

9. DESIGN GOVERNANCE AND REVIEW WORKFLOW

A measurement score becomes useful only when it changes a design decision. The paper therefore proposes a studio-oriented review workflow rather than an automated approval rule.

Table 7. Scene-review checklist for VR comfort and usability.

Review item	Question for the reviewer	Evidence to inspect	Action if weak
Visual anchor	Does the scene provide stable points for gaze?	Fixation stability, saccade amplitude, and target salience.	Add stable anchors and reduce competing peripheral motion.
Depth placement	Are text and controls placed at a comfortable depth?	Vergence error and depth-pressure values.	Move critical content to a consistent depth plane.
Interaction path	Can the user operate controls without slow search?	Selection latency and target-comfort values.	Increase target spacing and reduce unnecessary alternatives.
Spatial layout	Can recurring tools be remembered across visits?	Spatial memory support and head-turn demand.	Fix tool locations and group recurring functions.
Tracking reliability	Are measurements and interactions stable enough?	Tracking loss and window-level anomalies.	Re-test sensor setup or reduce rapid target switching.
Release record	Is the design decision documented?	Version history, risk flags, and correction log.	Archive the review and schedule retest after change.

Table 7 translates the framework into reviewer questions. The checklist deliberately avoids vague wording such as "improve immersion." Instead, it asks whether the scene has visual anchors, comfortable depth placement, operable controls, stable spatial layout, and reliable tracking. These questions are directly linked to measurable evidence.

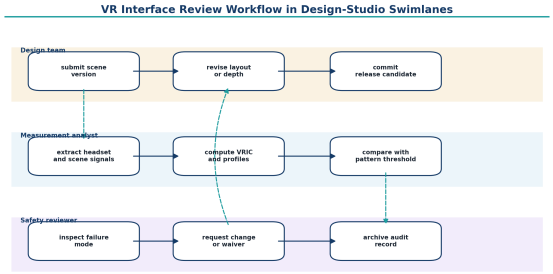

Figure 11. Design-studio swimlane workflow for VR scene review, redesign, threshold comparison, and release documentation.

Figure 11 divides responsibility across design, measurement, and safety-review lanes. The design team submits a scene version, the measurement analyst extracts headset and scene signals, and the safety reviewer inspects the failure mode before requesting change or accepting residual risk. This differs from a simple flowchart because it clarifies accountability. VR comfort review is not only a computation; it is a documented design decision.

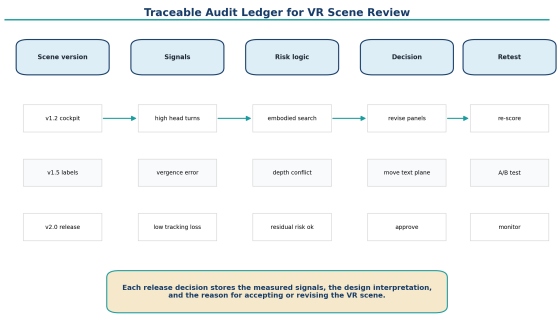

Figure 12. Traceable audit ledger for VR scene review and residual-risk documentation.

Figure 12 shows how the review can be recorded. A useful audit trail stores the scene version, measured signals, risk logic, decision, and retest plan. This is especially important in high-stakes VR training or monitoring environments. If a demanding scene is released despite moderate risk, the team should be able to explain why the risk was accepted and what evidence supported the decision.

10. DESIGN IMPLICATIONS FOR VR APPLICATIONS

The numerical results become more useful when they are read in relation to application scenarios. A cockpit, a medical simulator, an educational lesson, and a museum scene should not share the same acceptance rule.

Table 8. Scenario-level interpretation of the VR measurement results.

Scenario	Likely measurement pattern	Design risk	Recommended review action
Training cockpit	High head-turn demand with moderate latency	Users may miss peripheral information or become fatigued during repeated scans.	Group critical indicators in a stable forward arc and test peripheral alerts separately.
Medical monitoring scene	Strong object-search demand with high depth pressure	Users may inspect the wrong object or lose time during diagnostic search.	Separate essential objects from decorative detail and strengthen visual anchors.
Educational VR lesson	Good fixation but rising vergence error around panels	Learners may read slowly or disengage from text-heavy guidance.	Keep text panels at a consistent depth and reduce moving backgrounds behind labels.
Navigation tutorial	High waypoint demand and frequent head turns	Users may follow the route but lose spatial orientation.	Use fewer simultaneous cues, maintain landmarks, and reduce depth switching between signs.
Museum or tourism scene	Rich visual content with weak target comfort	Users may enjoy the space but struggle to locate interactive features.	Increase discoverability of actionable objects without removing environmental richness.

Table 9. Failure modes, measurements, and likely redesign decisions.

Failure mode	Measurement signal	How it appears to users	Design correction
Visual instability	Low fixation stability and high saccade amplitude	The user keeps searching, re-fixating, or missing the intended object.	Reduce competing motion, strengthen focal hierarchy, and simplify local visual context.
Depth conflict	High vergence error and high depth pressure	Text or targets feel uncomfortable even when they are legible.	Place high-frequency content on a stable depth plane and reduce unnecessary depth alternation.
Embodied search cost	High head-turn demand with weak spatial support	The user turns repeatedly and may forget where tools are located.	Reposition recurring tools, add landmarks, and reduce wide spatial distribution of controls.
Response friction	High selection latency with low target comfort	The target is visible, but the user needs too long to activate it.	Increase target size and spacing, improve controller mapping, and shorten the selection path.
Tracking fragility	Elevated tracking loss across windows	Interaction feels inconsistent or measurement becomes unreliable.	Re-test the scene under headset conditions and reduce rapid occlusion or target switching.

Table 8 illustrates the main practical point: VR review should be scenario-sensitive. A high head-turn value may be acceptable in exploration but risky in monitoring. A dense visual environment may be appropriate in a museum scene but harmful in a safety drill. The framework supplies evidence, while the scenario supplies the interpretation.

Table 9 helps the design team avoid one-size-fits-all redesign. A depth-conflict problem should not be fixed only by enlarging buttons, and a response-friction problem should not be fixed only by reducing visual detail. The failure mode identifies the mechanism; the mechanism determines the correction.

11. DISCUSSION

The results support three observations. First, VR interface quality is multi-causal. Fixation stability, contrast, target comfort, spatial memory, depth pressure, tracking reliability, head movement, pupil variability, and latency all contribute to the final comfort profile. The largest negative correlation in the extract is head-turn demand, which suggests that unnecessary scanning is one of the most damaging scene properties.

Second, interface pattern is not cosmetic. Text panels, dashboards, menus, navigation cues, and object-search scenes place different demands on the user. A scene with strong contrast can still fail if it places tools too widely or requires repeated depth switching. This is why the paper presents scene sketches, risk composition, and parallel-coordinate signatures rather than a single score table.

Third, measurement should be embedded into the design workflow. The review process should not ask only whether a VR scene looks good. It should ask whether the user's gaze can stabilize, whether the depth layout is comfortable, whether the interaction path is efficient, and whether the decision to release the scene has been documented. This shifts VR evaluation from aesthetic inspection toward accountable scene design.

12. LIMITATIONS AND FUTURE WORK

The study has limitations. The included extract follows public VR task structures and uses feature-level design overlays, but it should be validated with raw headset logs and live users in fully implemented scenes. The results are therefore best interpreted as design-science evidence rather than as a clinical comfort threshold.

The VRIC score is interpretable but not universal. Different applications may require different weights. A surgical simulator, a museum tour, a classroom environment, and a VR game do not share the same tolerance for motion, depth complexity, or response friction. Future work should calibrate the score to application-specific risk and combine it with subjective presence, learning, and task-success measures.

A further limitation is that comfort and challenge can sometimes conflict. A demanding VR scene is not automatically poor design if the task is meant to train search, navigation, or stress management. Future work should distinguish harmful discomfort from productive task difficulty by connecting VRIC with learning outcomes and long-term user adaptation.

13. CONCLUSION

This paper presented a scene-level assessment framework for measuring comfort, legibility, and spatial control in virtual reality interfaces. The framework combines headset-derived evidence with interface design variables and translates them into a transparent VRIC score, scene diagnostics, and review workflow. The findings suggest that VR interface comfort is strongest when gaze can stabilize, depth is predictable, targets are comfortable, contrast is balanced, and scene layout supports spatial memory.

The broader message is that VR interfaces should be reviewed as embodied environments rather than as static screens. A design team should be able to say why a scene was released, why it was revised, and which signals supported the decision. The proposed framework offers one practical route toward that goal.

REFERENCES

[1] D. Lohr, S. Aziz, L. Friedman, and O. V. Komogortsev, "GazeBaseVR, a large-scale, longitudinal, binocular eye-tracking dataset collected in virtual reality," *Scientific Data*, vol. 10, article 177, 2023.

[2] C. Rubow, C.-H. Tsai, E. Brewer, C. Mattson, D. S. Brown, and H. Zhang, "A dataset of paired head and eye movements during visual tasks in virtual environments," *Scientific Data*, vol. 11, article 1328, 2024.

[3] 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, article 38, 2024.

[4] C. Qu, X. Che, S. Ma, and S. Zhu, "Bio-physiological-signals-based VR cybersickness detection," *CCF Transactions on Pervasive Computing and Interaction*, vol. 4, no. 3, pp. 268–284, 2022.

[5] Y. Long, T. Wang, X. Liu, Y. Li, and D. Tao, "Toward accurate cybersickness prediction in virtual reality: A multimodal physiological modeling approach," *Sensors*, vol. 25, no. 18, article 5828, 2025.

[6] Y.-T. Huang, C.-C. Hsu, and T.-H. Wang, "Effects of interactive loading interfaces for virtual reality game environments on time perception, cognitive load, and emotions," *Frontiers in Virtual Reality*, vol. 6, article 1540406, 2025.

[7] K. Pettersson, J. Tervonen, J. Heininen, and J. Mäntyjärvi, "Head-area sensing in virtual reality: Future visions for visual perception and cognitive state estimation," *Frontiers in Virtual Reality*, vol. 5, article 1423756, 2024.

[8] R. Shadiev and D. Li, "A review study on eye-tracking technology usage in immersive virtual reality learning environments," *Computers & Education*, vol. 196, article 104681, 2023.

[9] P. A. Rauschnabel, R. Felix, C. Hinsch, H. Shahab, and F. Alt, "What is XR? Towards a framework for augmented and virtual reality," *Computers in Human Behavior*, vol. 133, article 107289, 2022.

[10] R. Gu, J. M. Rojas, and D. Shin, "Software testing for extended reality applications: A systematic mapping study," *Automated Software Engineering*, vol. 32, article 56, 2025.