Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

IVEM: A Unified Evaluation Framework for AI-Driven Immersive Visualisation

Rani Adam[1]

rani.adam@westernsydney.edu.au

Western Sydney University

Daniel R. Catchpoole[2]

Children's Hospital at Westmead

Simeon J. Simoff[1]

Western Sydney University

Zhonglin Qu[1]

Western Sydney University

Paul J. Kennedy[3]

University of Technology Sydney

Quang Vinh Nguyen[1]

Western Sydney University

Research Article

Keywords: Immersive Analytics, Virtual Reality, Evaluation Framework, Visualisation Metrics, AI-Driven Visualisation

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10597829/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

IVEM: A Unified Evaluation Framework for AI-Driven Immersive Visualisation

Rani Adam[1]*, Daniel R. Catchpoole[2], Simeon J. Simoff[1], Zhonglin Qu[1], Paul J. Kennedy[3], Quang Vinh Nguyen[1]

[1] School of Computer, Data and Mathematical Sciences, Western Sydney University, Australia

[2] The Tumour Bank, Children's Cancer Research Unit, Kids Research, The Children's Hospital at Westmead; The University of Sydney; University of Technology Sydney, Australia

[3] Australian Artificial Intelligence Institute, Faculty of Engineering and Information Technology, University of Technology Sydney, Australia

** Correspondence: rani.adam@westernsydney.edu.au*

Abstract

Artificial intelligence (AI) is transforming immersive visualisation by enabling natural-language interaction, adaptive visual encodings, and analytical reasoning within virtual reality (VR). Yet evaluation remains fragmented: existing methods assess rendering performance, interaction efficiency, visualisation accuracy, or explainability in isolation, which makes objective comparison difficult and limits reproducibility. This paper introduces the Immersive Visualisation Evaluation Model (IVEM), a unified evaluation framework for AI-assisted immersive visualisation. IVEM organises evaluation into three complementary dimensions (Rendering Quality, AI Decision Quality, and Explanation Quality) and defines six quantitative metrics: Scene Stability Index, Update Latency, AI Encoding Accuracy, Intent Recognition Accuracy, Encoding Consistency Score, and Visualisation Reasoning Score, each with a formal definition, measurement protocol, and evidence-informed interpretation thresholds. We validate IVEM using MediVerse, an AI-powered voice-driven VR analytics platform, across two paediatric oncology cohorts, and further demonstrate cross-platform use through a matched pilot comparison against ChatGPT with Data Analysis (GPT-5.5), a state-of-the-art non-immersive general-purpose large-language-model data-analysis assistant, run on identical data and queries. Both systems achieve high intent-recognition and encoding accuracy, but they diverge sharply on the immersive dimension: GPT-5.5 is roughly an order of magnitude slower to render and cannot provide spatial scene stability. IVEM surfaces diagnostic insights that conventional measures cannot reveal, including limited decision reproducibility and incomplete explanation reasoning despite strong rendering and visualisation accuracy. The framework thereby establishes a foundation for objective benchmarking and systematic improvement of AI-driven immersive visualisation systems.

Keywords: Immersive Analytics, Virtual Reality, Evaluation Framework, Visualisation Metrics, AI-Driven Visualisation.

1. Introduction

Artificial intelligence (AI) is fundamentally transforming immersive analytics by enabling natural-language interaction, intelligent visualisation recommendation, adaptive rendering, and explainable analytical reasoning within virtual reality (VR) environments. Advances in large language models (LLMs), speech recognition, and AI-assisted visualisation now allow users to explore increasingly complex datasets through conversational interaction rather than traditional programming or query languages. Consequently, immersive analytics is evolving from a purely interactive visualisation paradigm toward intelligent analytical environments in which AI actively assists users in data exploration, interpretation, and decision making (Wang et al. 2025) (Shen et al. 2022).

These advances have expanded the capabilities of immersive visualisation across diverse domains, including healthcare, engineering, manufacturing, education, scientific visualisation, and digital twins (Cordeil et al. 2019). Modern immersive analytics platforms are expected not only to render high-dimensional data efficiently but also to interpret user intent, recommend appropriate visual encodings, generate analytical explanations, and maintain coherent interaction throughout dynamic exploratory workflows. As AI assumes greater responsibility for analytical decision making, evaluating these systems becomes substantially more complex than measuring rendering performance or interaction efficiency alone.

Existing evaluation methodologies were largely developed for conventional visualisation systems, where performance is assessed using rendering speed, task completion time, interaction efficiency, graphical perception, or user satisfaction (Islam et al. 2024). While valuable, these measures evaluate only individual aspects of immersive systems and do not adequately capture the behaviour of AI-assisted visualisation (Wang et al. 2025).

The emergence of AI-driven immersive visualisation introduces several new evaluation challenges. First, rendering performance must preserve users' spatial understanding during continuous visual updates rather than simply maintaining high graphical performance (Bhattacharya, Prinz, and Mathew 2025; König et al. 2021).

Second, AI systems must correctly interpret user intent and consistently recommend visualisations that accurately represent analytical tasks (Shen et al. 2022; Hu et al. 2019). Third, explanations generated by AI should communicate not only what is displayed but also why a particular visualisation was selected and how it supports analytical reasoning (Nauta et al. 2023; Hoffman et al. 2018). Existing research has investigated these topics independently through immersive analytics evaluation (Whitlock, Smart, and Szafir 2020), visualisation recommendation (Hu et al. 2019; Dibia and Demiralp 2019), and explainable AI (XAI) (Vilone and Longo 2021; Nauta et al. 2023), yet no unified framework integrates these complementary dimensions into a single quantitative evaluation methodology.

Without common evaluation criteria, objective comparison between immersive visualisation systems remains difficult, results are hard to reproduce, and developers have limited capability to identify which components of an analytical pipeline require optimisation (Podo, Prenkaj, and Velardi 2024). As conversational AI and adaptive visualisation become increasingly common within immersive environments, standardised evaluation methodologies will be essential for ensuring reliable, trustworthy, and reproducible immersive analytics systems.

To address this challenge, this paper introduces the Immersive Visualisation Evaluation Model (IVEM), a unified evaluation framework designed specifically for AI-driven immersive visualisation. Rather than proposing another immersive analytics platform, IVEM provides a reusable methodology for evaluating such platforms. The framework organises evaluation into three complementary dimensions (Rendering Quality, AI Decision Quality, and Explanation Quality) and defines six quantitative metrics that collectively characterise the technical performance, analytical correctness, and reasoning capability of AI-assisted immersive visualisation. These six metrics are complementary rather than interchangeable: each isolates a single property of the system, and together they span its technical behaviour, its AI-assisted analytical decisions, and the human-facing transparency of those decisions, so that a weakness in any one facet is surfaced rather than masked by strength in another. Each metric is formally defined through mathematical formulations, operational measurement protocols, and evidence-based interpretation thresholds, enabling reproducible evaluation across diverse systems. The framework is guided throughout by three principles: reproducibility, so that identical procedures can be applied to different systems; generalisability, so that it can evaluate platforms beyond the one used here; and diagnostic capability, so that each metric independently identifies a specific strength or optimisation opportunity.

The applicability of IVEM is demonstrated using MediVerse, an AI-powered voice-driven VR analytics platform for biomedical data exploration (Adam et al. 2025b), whose edge and cloud architecture builds on prior work on hybrid edge-cloud frameworks for managing and compressing large-scale omics data (Adam et al. 2025a; Adam et al. 2024). MediVerse is a representative validation environment because it integrates conversational interaction, AI-assisted visualisation recommendation, immersive rendering, and analytical explanation within a single system. Using two paediatric oncology datasets, the framework evaluates representative analytical workflows and identifies strengths and limitations that conventional evaluation methodologies cannot readily reveal. We further apply IVEM across systems, comparing MediVerse against ChatGPT with Data Analysis (GPT-5.5), a state-of-the-art non-immersive general-purpose LLM data-analysis assistant, on identical data and queries, to help evaluate the value of the immersive interface. A world-class model is chosen deliberately, so that any immersive advantage IVEM reports reflects a genuine capability gap rather than the weakness of the baseline. The main contributions of this paper are as follows:

1. A unified evaluation framework for AI-driven immersive visualisation that integrates rendering quality, AI decision quality, and explanation quality into a single quantitative methodology.
2. Six formally defined evaluation metrics, each supported by mathematical formulations, operational measurement procedures, and evidence-informed interpretation thresholds that enable reproducible evaluation across immersive analytics systems.
3. Experimental validation of IVEM using the MediVerse platform, together with a cross-platform comparison against ChatGPT with Data Analysis (GPT-5.5), a state-of-the-art non-immersive general-purpose LLM, demonstrating diagnostic insights that conventional evaluation methods cannot capture.

The remainder of this paper is organised as follows. Section 2 reviews related work and identifies the need for a unified evaluation methodology. Section 3 introduces the IVEM framework and its design principles. Section 4 formally defines the six metrics and their measurement procedures. Section 5 describes the validation platform and the state-of-the-art non-immersive baseline. Section 6 presents the validation methodology, and Section 7 reports the experimental results. Section 8 discusses the findings, generalisability, and limitations, before Section 9 concludes the paper.

2. Related Work

Immersive analytics combines virtual and augmented reality with interactive visualisation to support intuitive exploration of complex multidimensional data, improving comprehension, collaboration, and decision making across scientific, engineering, and biomedical domains (König et al. 2021; Cherep et al. 2020; Whitlock, Smart, and Szafir 2020; Cordeil et al. 2019). Evaluation has traditionally relied on human-centred measures such as task completion time, interaction efficiency, cognitive workload, graphical perception, presence, and user preference (Whitlock, Smart, and Szafir 2020; Islam et al. 2024). These are effective for comparing interaction techniques where visualisations are relatively static, but as immersive systems increasingly use AI to generate visualisations, interpret natural-language queries, recommend encodings, and produce explanations, such measures capture only part of the behaviour of modern AI-driven systems (Wang et al. 2025).

In parallel, AI has automated visualisation design: machine-learning approaches such as VizML (Hu et al. 2019) and Data2Vis (Dibia and Demiralp 2019) predict or generate designs from data characteristics, and conversational LLM interfaces now let users create visualisations through natural language (Shen et al. 2022). Their evaluation, however, focuses on recommendation accuracy or user preference (Hu et al. 2019; Islam et al. 2024; Dibia and Demiralp 2019) and does not assess whether AI decisions are consistent across repeated tasks or justified through transparent reasoning (Podo, Prenkaj, and Velardi 2024). Explanation quality has instead been studied within explainable AI (XAI), whose frameworks assess completeness, faithfulness, interpretability, and user understanding (Vilone and Longo 2021; Nauta et al. 2023; Hoffman et al. 2018; Chatzimparmpas 2025); but these target predictive models and classifiers (Nauta et al. 2023; Hoffman et al. 2018) rather than visualisation decisions, and so do not judge why a particular encoding was chosen over alternatives.

Immersive analytics, AI-assisted visualisation, and XAI have thus each developed mature evaluation methodologies within their own communities, but these evolved independently and each measure only one facet: rendering and interaction (Islam et al. 2024; Whitlock, Smart, and Szafir 2020), AI prediction accuracy (Dibia and Demiralp 2019; Hu et al. 2019), or reasoning transparency (Vilone and Longo 2021; Nauta et al. 2023; Hoffman et al. 2018; Chatzimparmpas 2025). None evaluates rendering quality, AI decision quality, and explanation quality together, which makes objective cross-system comparison difficult and hinders reproducible benchmarking (Stauffert, Niebling, and Latoschik 2020; Whitlock, Smart, and Szafir 2020). Table 1 summarises this gap across the dimensions required for AI-driven immersive visualisation and motivates the IVEM framework presented next.

Evaluation Approach	Rendering Quality	AI Decision Quality	Explanation Quality	Unified Framework
Traditional Visualisation Evaluation (Islam et al. 2024; Qu et al. 2023)	✓	✗	✗	✗
Immersive Analytics Evaluation (König et al. 2021; Cherep et al. 2020; Whitlock, Smart, and Szafir 2020; Cordeil et al. 2019).	✓	✗	✗	✗
AI Visualisation Recommendation (Hu et al. 2019; Dibia and Demiralp 2019)	✗	✓	✗	✗
Explainable AI Evaluation (Vilone and Longo 2021; Nauta et al. 2023; Hoffman et al. 2018; Chatzimparmpas 2025)	✗	✗	✓	✗
IVEM (This Work)	✓	✓	✓	✓

Table 1. Comparison of AI-driven immersive visualisation evaluation approaches.

3. The IVEM Framework

IVEM organises evaluation into three complementary dimensions that together characterise AI-driven immersive visualisation (Figure 1). Unlike traditional approaches that report a single performance measure, IVEM separates technical system behaviour from AI-assisted analytical behaviour and explanation quality, enabling independent diagnosis of each component while preserving a unified overall methodology. The three dimensions, and the six metrics that quantify them, are:

i. **Rendering Quality** evaluates the technical behaviour of the display during dynamic interaction, specifically whether visualisations remain spatially stable and sufficiently responsive throughout continuous exploration. It is quantified by the Scene Stability Index (SSI), which measures spatial

stability during visual updates, and Update Latency (UL), which measures end-to-end system responsiveness.

ii. **AI Decision** Quality evaluates the correctness, reliability, and reproducibility of AI-assisted decisions, including whether the system interprets user intent, selects appropriate encodings, and behaves consistently for equivalent tasks. It is quantified by AI Encoding Accuracy (AEA), which evaluates whether the AI selects appropriate visualisations; Intent Recognition Accuracy (IRA), which assesses how accurately user intent is interpreted; and the Encoding Consistency Score (ECS), which evaluates the reproducibility of AI visualisation decisions across repeated queries.

iii. **Explanation Quality** evaluates the transparency of AI-assisted reasoning, namely whether explanations justify why encodings were selected, how they relate to the data, and how they support analytical goals. It is quantified by the Visualisation Reasoning Score (VRS), which measures how well the AI justifies its visualisation choices through clear, grounded explanations.

The six metrics are deliberately minimal and non-overlapping: each isolates a single property that no other metric captures, so removing any one would leave a corresponding blind spot. Excellence in one dimension cannot compensate for weakness in another. A system may render with perfect stability yet select inappropriate encodings, interpret intent accurately yet explain its choices only superficially, or select accurate encodings yet do so non-reproducibly. The six metrics span the technical, AI-analytical, and human-facing facets of immersive visualisation. This non-substitutable set provides the theoretical basis for reporting the metrics separately rather than collapsing them into a single aggregate score.

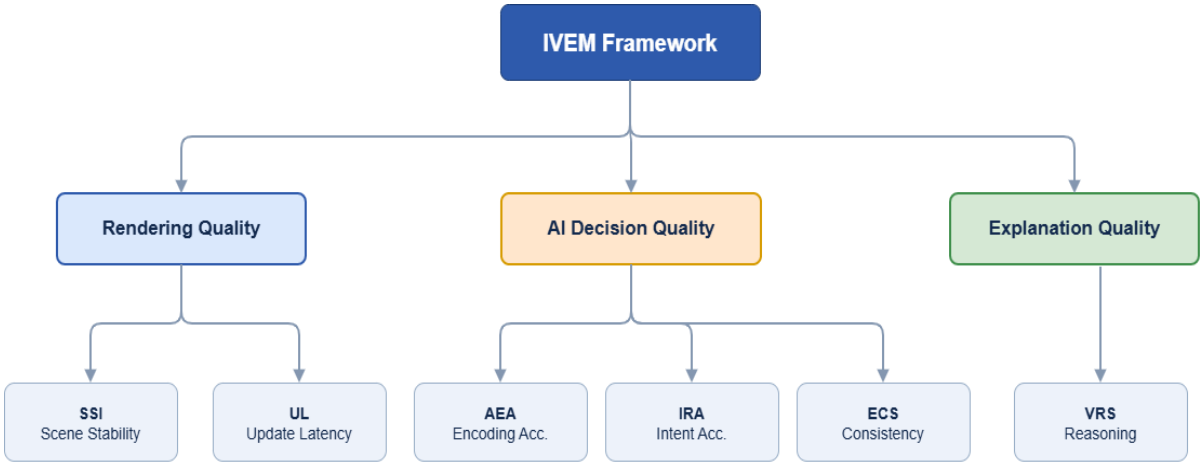


Figure 1. Overall architecture of IVEM, showing the three evaluation dimensions and the six quantitative metrics that collectively characterise AI-driven immersive visualisation systems.

4. IVEM Metric Definitions

This section formally defines the six IVEM metrics used to evaluate AI-driven immersive visualisation systems. For each metric, we present its mathematical formulation, measurement procedure, the property it captures, and evidence-informed interpretation thresholds. Together, these metrics quantify the three IVEM dimensions, namely Rendering Quality, AI Decision Quality, and Explanation Quality, and provide a reproducible basis for evaluating and comparing AI-assisted immersive visualisation systems. Table 2 summarises the complete metric suite.

Metric	Dimension	What it captures	Formula (summary)
Scene Stability Index (SSI)	Rendering	Spatial stability during in-place updates	1 − mean normalised object displacement
Update Latency (UL)	Rendering	End-to-end query-to-render responsiveness	mean($t_v - t^d$) over updates
AI Encoding Accuracy (AEA)	AI Decision	Whether the chosen encoding fits the task	mean s_i, $s_i \in \{0, 0.5, 1\}$
Intent Recognition Accuracy (IRA)	AI Decision	Whether NL intent is correctly interpreted	N_correct / N_queries
Encoding Consistency Score (ECS)	AI Decision	Reproducibility of choices across runs	mean C_i, $C_i \in \{0, 1\}$

Metric	Dimension	What it captures	Formula (summary)
Visualisation Reasoning Score (VRS)	Explanation	Grounded justification of the choice	mean fraction of 4 reasoning parts

Table 2. Summary of the IVEM metrics and evaluation criteria.

Throughout this section, the quality thresholds are presented as general-purpose defaults derived from the cited literature, not as universal constants. The metric definitions and measurement procedures are fixed but the boundaries between excellent, acceptable, and poor are expected to be calibrated to the precision demands, safety criticality, and interaction style of the target application domain.

4.1 Rendering Quality Metrics

Rendering quality metrics evaluate the perceptual stability and responsiveness of the immersive visualisation system. These metrics capture whether visualisations maintain spatial consistency during updates and whether the system responds within perceptual thresholds that preserve user comfort and task focus.

Scene Stability Index (SSI): SSI measures how much objects move when updating visual properties (such as colour, size, or filtering) within the same visualisation. Higher values indicate predictable positioning where objects maintain their spatial locations during property updates rather than being randomly repositioned, thereby preserving user spatial awareness in VR, as formally defined below (Eq. 1):

$$SSI = 1 - \frac{1}{N \cdot d_{\max}} \sum_{i=1}^{N} \| p_i' - p_i \|_2 \qquad \textit{Eq. 1}$$

where N = number of objects, p_i and p_i' = 3D position vectors before and after the update, $d_{max} = \sqrt{(\Delta x^2 + \Delta y^2 + \Delta z^2)}$ the scene bounding-box diagonal, and $\|\cdot\|_2$ denotes the Euclidean norm.

What It Captures	Quality Thresholds
● Support for user cognitive mapping in VR ● Scene coherence and spatial continuity ● Displacement magnitude across all objects	● ≥ 0.90 → Excellent ● 0.85 ≤ SSI < 0.90→ Acceptable ● < 0.85 → Poor

Thresholds Justification: No prior work defines explicit numeric thresholds for object displacement in immersive data visualisation, so we derive them here from established VR spatial-cognition research rather than adopting an existing benchmark. Navigation studies show that spatial consistency, maintaining stable relationships between locations and objects, is critical for cognitive-map formation (Bhattacharya, Prinz, and Mathew 2025; König et al. 2021; Cherep et al. 2020), and discontinuous movement such as teleportation degrades spatial awareness relative to continuous motion. Applying this principle to visualisation objects, we hypothesise that holding object positions during property updates preserves users' spatial mental models. The conservative thresholds (0.85–0.89 acceptable, ≥ 0.90 excellent) ensure most objects retain their positions during updates, consistent with perceptual principles under which larger displacements become increasingly salient and disruptive.

Update Latency (UL): UL quantifies end-to-end system responsiveness from data-change detection to completed visual rendering. It measures the total delay users experience during dynamic updates, as formally defined below (Eq. 2):

$$UL = \frac{1}{M} \sum_{i=1}^{M} (t_i^v - t_i^d) \qquad \textit{Eq. 2}$$

where M = number of update events, t_i^d = timestamp when the data changes, and t_{iv} = timestamp when the visualisation completes.

What It Captures	Quality Thresholds
● System responsiveness ● Real-time performance ● Interaction fluidity	● < 1000 ms → Excellent ● 1000 ms < UL ≤ 4000 ms→ Acceptable ● > 4000 ms → Disruptive

Thresholds Justification: Motion-to-photon thresholds (< 20 ms) are critical for head-tracked rendering to prevent cybersickness (Stauffert, Niebling, and Latoschik 2020), but Update Latency covers a fundamentally different interaction model (speech recognition, natural-language understanding, database execution, and rendering) where higher latencies are inherently acceptable. Established human-computer interaction (HCI) principles hold that response times below 1 second maintain flow of thought, while delays up to 10 seconds preserve task focus (Miller 1968). Because prolonged delays disrupt spatial presence in VR, a more conservative bound is warranted: recent work on VR conversational agents shows that latencies above 4 seconds significantly degrade experience (Maslych et al. 2025), supporting the adopted thresholds of < 1000 ms (excellent) and 1000–4000 ms (acceptable).

4.2 AI Decision Quality Metrics

AI decision quality metrics evaluate the correctness, reliability, and consistency of AI-driven visualisation selections. These metrics assess whether the system accurately interprets user intent, selects appropriate visual encodings, and maintains reproducible decision-making across repeated queries.

AI Encoding Accuracy (AEA): AEA evaluates the correctness of AI-selected visual encodings through expert judgement. It determines whether the AI chooses appropriate visualisations that accurately represent data semantics and support analytic tasks, as formally defined below (Eq. 3):

$$AEA = \frac{1}{M} \sum_{i=1}^{M} s_i \text{ where } s_i \in \{0, 0.5, 1\} \qquad \textit{Eq. 3}$$

where M = number of AI decisions evaluated and s_i = 0 (incorrect), 0.5 (partially appropriate), or 1 (optimal), assigned by expert judgement against the rubric (a panel of K ≥ 3 independent raters with reported inter-rater agreement is recommended for reliable scoring); AEA is the mean of the M scores.

What It Captures	Quality Thresholds
● Encoding appropriateness ● AI decision quality ● Trust in AI selections	● ≥ 0.90 → Highly accurate ● 0.85 ≤ AEA < 0.90 → Acceptable ● < 0.85 → Poor

Thresholds Justification: The thresholds of ≥ 0.90 (highly accurate) and 0.85–0.89 (acceptable) are informed by VizML (Hu et al. 2019), which reported roughly 70–95% neural-network accuracy across visualisation design-choice tasks. We adopt the upper end because MediVerse operates in an immersive clinical setting where an incorrect selection is immediately visible and disrupts the workflow, a more demanding standard than general-purpose recommendation. As Podo et al. (Podo, Prenkaj, and Velardi 2024) note, the absence of standardised benchmarks means field-wide thresholds do not yet exist; these therefore represent the first operationalised accuracy criteria for AI-driven visualisation selection in immersive environments.

Intent Recognition Accuracy (IRA): IRA measures the accuracy of natural-language query interpretation by the AI. It determines whether user intent is correctly understood and translated into appropriate visualisation operations, as formally defined below (Eq. 4):

$$IRA = \frac{N_{correct}}{N_{queries}} \qquad \textit{Eq. 4}$$

where N_queries is the total number of evaluated natural-language queries and N_correct is the number for which the AI-generated visualisation correctly reflects the user's intended analytical intent (correct = the AI generates the intended response; incorrect = the AI misunderstands and generates the wrong response).

What It Captures	Quality Thresholds
● Query understanding ● Intent mapping reliability ● Natural-language interaction quality	● ≥ 0.95 → Excellent ● 0.90 ≤ IRA < 0.95 → Acceptable ● < 0.90 → Unreliable

Thresholds Justification: Intent-recognition errors propagate into incorrect filtering, aggregation, or chart selection (Shen et al. 2022). On intent-recognition datasets such as Snips, leading approaches report average slot-filling F1-scores in the high-80s to low-90s, with performance varying by query complexity (Coucke et al. 2018). Reported accuracy on the narrower natural-language-to-visualisation task is lower still: evaluated on the nvBench

benchmark, the ncNet neural translation model reaches 77.8% exact-match accuracy without a chart template and 79.6% with one, against a neural-network baseline of 65.0% (Luo et al. 2021). No prior study proposes numeric acceptability thresholds for this task, so we define them here, setting the acceptable floor at the upper end of reported general intent-recognition performance and reserving the excellent tier for systems exceeding it. Because a misread intent in a clinical setting silently displays the wrong patient data rather than producing a visibly wrong chart, we adopt deliberately conservative values (≥ 0.95 excellent, 0.90 ≤ IRA < 0.95 acceptable, < 0.90 unreliable), proposed as IVEM defaults and flagged for empirical validation in future work.

Encoding Consistency Score (ECS): ECS measures whether the AI makes the same visualisation choices when presented with identical or equivalent queries. It quantifies the stability of AI decision-making by tracking how consistently the system assigns visual properties (such as colour representing risk level or size representing age) across repeated queries or updates, as formally defined below (Eq. 5):

$$ECS = \frac{1}{M}\sum_{i=1}^{M} C_i \qquad \textit{Eq. 5}$$

where M = number of data attributes mapped to visual properties and $C_i = 1$ if the encoding for attribute i remains consistent across updates, 0 otherwise. An encoding is consistent if the same attribute maps to the same visual property (colour, size, position, shape) with the same semantic meaning.

What It Captures	Quality Thresholds
● AI decision reliability and reproducibility ● Stability of semantic mappings over time ● User ability to build reliable mental models ● Trust in AI-generated visualisation choices	● ≥ 0.95 → Excellent ● 0.85 ≤ ECS < 0.95→ Acceptable ● < 0.85 → Poor

Thresholds Justification: Consistency is fundamental to usability, letting users form accurate expectations of system behaviour (Nielsen 1994; Norman Donald 2013). In visualisation contexts, inconsistent mappings between attributes and encodings prevent users from learning stable semantic associations over time. Because no prior study defines a consistency threshold for AI-assisted visualisation, we define one here, adopting ≥ 0.95 for excellent consistency (predictable behaviour in at least 19 of 20 encoding decisions) and 0.85 as the acceptable bound, accommodating legitimate variability in ambiguous queries, while consistency below 0.85 indicates problematic non-determinism. As no established numerical benchmark exists, these thresholds are proposed as part of IVEM and flagged for empirical validation across multiple systems in future work (Podo, Prenkaj, and Velardi 2024).

4.3 Explanation Quality Metrics

Explanation quality metrics evaluate how well the AI justifies its visualisation decisions. These metrics assess whether explanations provide sufficient reasoning about data patterns, encoding choices, and analytical goals rather than merely describing visible content.

Visualisation Reasoning Score (VRS): VRS measures how well the AI explains why a specific chart type or encoding was chosen based on data characteristics and user intent, not merely content description. A complete explanation should include multiple reasoning components that justify the visualisation decision, as formally defined below (Eq. 6):

$$VRS = \frac{1}{M}\sum_{i=1}^{M}\left(\frac{1}{R_i}\sum_{k=1}^{R_i} r_{i,k}\right) \qquad \textit{Eq. 6}$$

where M denotes the number of visualisation explanations evaluated, R_i the number of required reasoning components for explanation i, and $r_{i,k} \in \{0, 1\}$ whether component k is present and correctly grounded. In this study $R_i = 4$: (1) data justification, why the selected data subset addresses the user's query; (2) encoding rationale, why the chosen encodings are appropriate for the data structure; (3) alternative consideration, why other visualisation types were not selected; and (4) context relevance, how the visualisation supports the user's analytical goal.

What It Captures	Quality Thresholds
• Grounding in data structure • Connection to user intent • Depth of visualisation justification	• 1.0 → Excellent (all 4 components) • 0.75 → Good (3 components) • 0.5 → Acceptable (2 components) • < 0.5 → Poor (≤ 1 component)

Thresholds Justification: Explanation quality depends on addressing multiple complementary reasoning dimensions rather than superficial description (Nauta et al. 2023; Hoffman et al. 2018). Complete explanations (VRS = 1.0) cover all four dimensions, letting users judge appropriateness from data, visual, comparative, and goal perspectives. Missing one dimension (VRS = 0.75) leaves critical gaps; two components (VRS = 0.5) is the acceptable minimum for basic evaluation, while single-component explanations (VRS < 0.5) provide only surface description without reasoning.

5. Validation Platform

The primary contribution of this work is the IVEM framework rather than any individual system. Accordingly, the platforms evaluated here serve as validation vehicles rather than as the primary objects of study. MediVerse exercises all six metrics through realistic analytical interactions and is therefore used as the primary validation platform. MediVerse (Adam et al. 2025b) is an AI-powered, voice-driven immersive VR system for biomedical analytics: users issue natural-language voice queries, which an AI model interprets, translates into schema-aware SQL, executes, and renders as interactive 2D and 3D visualisations inside a head-mounted display. Full implementation details appear in (Adam et al. 2025b); this section summarises the system briefly so that the evaluation in Sections 6 and 7 can be read against a concrete architecture.

MediVerse follows a layered architecture in which an edge layer on the VR headset (Meta Quest 2) captures voice through the head-mounted display microphone, transcribes it via a speech-to-text service, and transmits the text over TLS to an authenticated API gateway, while explanatory responses are vocalised through text-to-speech and 3D results are rendered in Unity3D. A cloud layer forms the intelligent core: it classifies each query as a simple retrieval or an analytical ("why") request, enriches it with schema metadata, generates schema-aware SQL using an LLM (Grok-1), executes it against a SQL Server backend, and then recommends a visualisation format, axis mappings, and styling. Because a central claim of IVEM is that it enables objective comparison across systems, MediVerse is evaluated not in isolation but against a strong, state-of-the-art non-immersive baseline. Testing IVEM against a general-purpose LLM model is deliberate: it ensures that any advantage attributed to the immersive interface reflects a genuine capability gap rather than the weakness of the baseline. To compare the two systems fairly, the same datasets and the same 50-query benchmark (see Section 6) are issued to each through its native interface, and each of the six IVEM metrics is measured with its defined protocol; holding data and queries constant isolates genuine differences between the systems from differences in workload.

The non-immersive baseline was ChatGPT with Data Analysis (GPT-5.5) (OpenAI 2026), a state-of-the-art general-purpose LLM model widely used for data analysis: the user uploads the same de-identified tabular files, issues the same natural-language questions, and the model interprets intent, writes and executes code, selects a chart type, renders a static two-dimensional image, and returns a written justification. It is therefore an AI-driven analytics system in the same broad class as MediVerse, sharing the natural-language, AI-encoding, and explanation dimensions that IVEM measures, but it produces flat 2D charts on a conventional display rather than navigable 3D scenes in a head-mounted display. Because it returns static images rather than a navigable scene, the Scene Stability Index, which quantifies object displacement during in-place updates, does not apply; the remaining five metrics map cleanly. Both systems are scored with the identical IVEM rubric by the same evaluator (Section 6.3), so that the reported values are directly comparable.

6. Initial Validation Methodology

The objective of this study is to demonstrate that IVEM provides a practical, reproducible, and quantitative methodology for evaluating AI-driven immersive visualisation systems. The evaluation verifies that the framework can systematically characterise rendering behaviour, AI-assisted decision making, and explanation quality within representative immersive workflows, and that identical procedures support cross-system comparison. The methodology follows three principles: reproducibility (identical procedures can be applied to different systems), generalisability (IVEM can evaluate platforms beyond MediVerse), and diagnostic capability (each metric independently identifies specific strengths and optimisation opportunities).

6.1 Experimental Datasets

Two independent biomedical datasets were selected to exercise diverse analytical workflows while remaining representative of realistic immersive analytics:

- **Paediatric leukaemia (clinical).** A clinical dataset containing demographic attributes, treatment information, and disease characteristics. It primarily supports interactive filtering, categorical grouping, colour updates, and dynamic rendering stability, with representative tasks such as grouping patients by gender, recolouring by treatment response, and filtering clinical cohorts during exploration.
- **Paediatric rhabdomyosarcoma (multimodal).** A dataset integrating gene-expression measurements with clinical metadata. It exercises multidimensional exploration, automatic chart recommendation, analytical explanation, and visualisation reasoning, with representative tasks such as visualising gene-expression trends, filtering histological subtypes, and generating AI-supported explanations.

Using two substantially different datasets demonstrates that IVEM evaluates system behaviour independently of any particular analytical scenario. Both datasets were fully de-identified before use, and no personally identifiable information was uploaded to any external service; the evaluation concerns the quality of visualisation and reasoning rather than clinical inference about individual patients.

6.2 Benchmark Task Design

To ensure reproducibility, validation used a structured benchmark of 50 natural-language queries spanning the four query classes (descriptive, filtering, comparative, and diagnostic) across both cohorts. The complete query list is available at https://mediverse.life/evaluation-instrument. Table 3 summarises the composition of the instrument by query class, together with the analytical operations, datasets, and IVEM metrics each class exercises; the benchmark was designed so that each class exercises one or more IVEM metrics while the full set collectively covers all three evaluation dimensions. The individual metrics are computed over the benchmark or over representative subsets of it.

Query class	Queries	Analytical operations covered	Datasets	Primary IVEM metrics
Descriptive	16	Retrieval and grouping (e.g. show all patients; 3D scatter by attribute; colour or group by category)	Leukaemia, RMS	UL, IRA, SSI, AEA
Filtering	14	Cohort and subtype filtering; compound record-level filters	Leukaemia, RMS	IRA, SSI, ECS
Comparative	12	Comparison and trend analysis (e.g. compare treatment response; gene-expression trends)	Leukaemia, RMS	AEA, VRS, UL
Diagnostic	8	Explanatory “why” queries requiring analytical justification	Leukaemia, RMS	VRS
Total	50	All four classes across both cohorts	Leukaemia, RMS	SSI, UL, AEA, IRA, ECS, VRS

Table 3. Composition of the 50-query benchmark instrument by query class, showing the analytical operations, datasets, and IVEM metrics each class exercises across both paediatric cohorts.

6.3 Experimental Procedure

Each benchmark task followed an identical procedure: (1) execute the benchmark query on the system under test; (2) record rendering behaviour, response latency, AI-generated visualisations, and explanatory responses; (3) compute the rendering metrics (SSI and UL) automatically from logged positions and timestamps; (4) evaluate AI-assisted decision quality (AEA and IRA) using expert assessment; (5) execute repeated queries to determine Encoding Consistency (ECS); and (6) evaluate explanation completeness using the Visualisation Reasoning Score (VRS). Two metrics require expert judgement, AEA and VRS. In this initial validation, both were evaluated by the first author using the predefined IVEM scoring rubric: AEA according to the suitability of the selected visualisation for the data characteristics and analytical objective, and VRS according to whether explanations addressed the four required reasoning components (data justification, encoding rationale, alternative consideration, and contextual relevance). The objective of this study was to demonstrate the applicability of the proposed evaluation framework rather than to establish inter-rater reliability; inter-rater agreement is therefore not reported, and future work will include independent expert evaluation and formal assessment of agreement.

Because benchmark tasks, measurement procedures, scoring protocols, and thresholds are explicitly defined, the same procedure can be applied to different systems under comparable conditions, which supports the cross-system comparison set up in Section 5.

Encoding Consistency was measured by executing a twenty-query subset of the benchmark three times on each system. For the non-immersive baseline each repetition was issued in a new conversation, so that the model could not condition on its own earlier responses; without this isolation, repeated queries within one session inflate apparent consistency. An encoding was scored consistent only when all three repetitions produced the same chart type for the same query.

7. Experimental Results

This section reports the IVEM results. IVEM is first applied to MediVerse across the two paediatric oncology cohorts using the 50-query benchmark of Section 6.2 (composition summarised by query class in Table 3), reporting one measured value per metric from representative multi-turn interactions, and then extended across systems through the matched pilot against ChatGPT with Data Analysis. The objective is to demonstrate the diagnostic capability of IVEM rather than to maximise application performance; accordingly, each metric is measured under realistic conditions, and the six values together form a diagnostic profile rather than a single aggregate score.

Scene Stability Index (SSI). For a representative update with N = 247 patients in a 3D scatter plot, when the user requests "colour by treatment status," all 247 points keep their exact positions and only colours change (Figures 2–3). The total displacement is 0 metres, giving SSI = 1.0 (perfect stability, exceeding the ≥ 0.90 threshold) and confirming that MediVerse maintains spatial stability during property updates, enabling users to build stable spatial mental models of the data (Cherep et al. 2020; Bhattacharya, Prinz, and Mathew 2025).

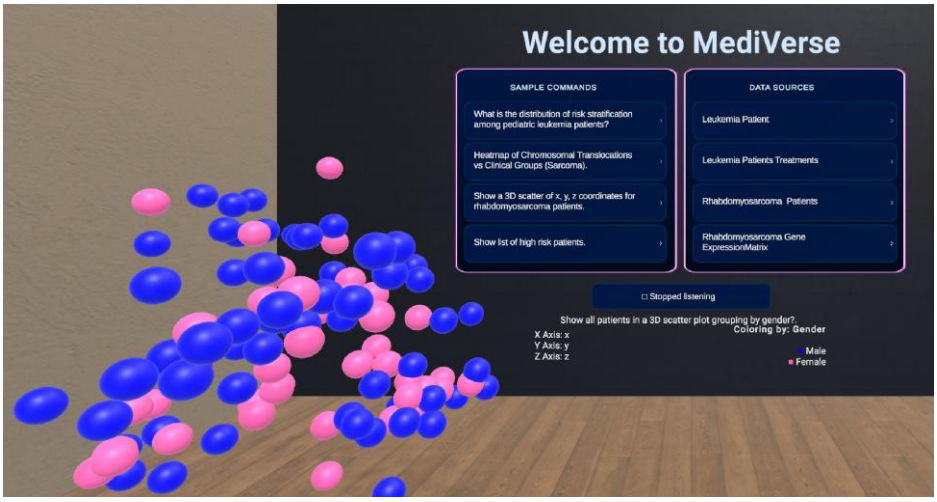


Figure 2. Immersive 3D scatter plot of paediatric leukaemia patients with gender-based colour encoding.

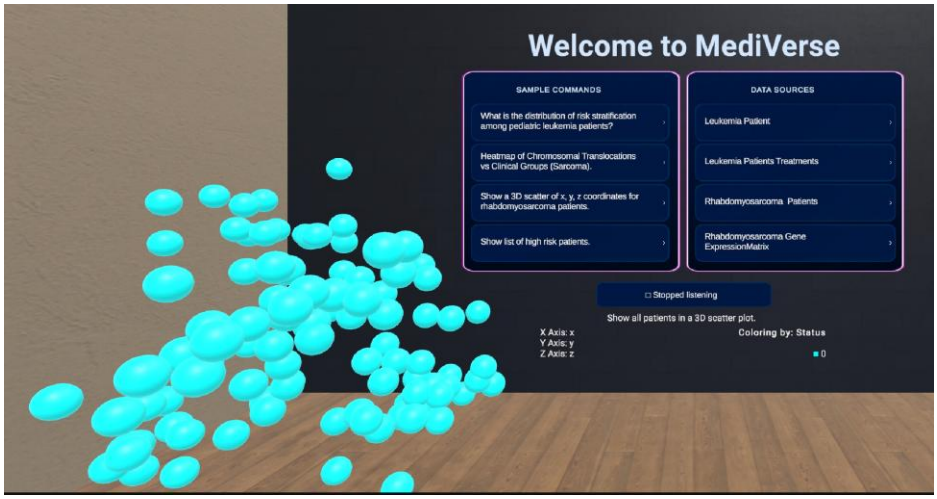


Figure 3. The same 3D scatter plot coloured by treatment status, illustrating perfect scene stability (SSI = 1.0).

Update Latency (UL). Latency was measured across four query classes (Figure 4): descriptive queries at approximately 1100 ms, filtering at approximately 1400 ms, comparative at approximately 1740 ms, and diagnostic at approximately 1700 ms. Averaged across these classes, UL = 1485 ms, within the acceptable 1000–4000 ms threshold and consistent with HCI principles that delays below 1 second maintain flow of thought and up to 10 seconds preserve task focus (Miller 1968). These values were obtained under controlled conditions. MediVerse was evaluated on a Meta Quest 2 head-mounted display running the Unity3D client and communicating with the cloud backend over an authenticated TLS connection. The non-immersive baseline was ChatGPT with Data Analysis running GPT-5.5, accessed through Google Chrome 150.0.4078.83 on a Lenovo Yoga Slim 7 14ILL10 workstation with an Intel Core Ultra 5 226V processor (8 cores, 2.10 GHz) and 16 GB RAM, running Windows 11 Home (build 26200). Both systems were exercised over the same wireless network connection, measured at 78 Mbps downlink at the time of testing. All measurements were collected in August 2026. production latency may vary with query complexity, computational resources, network bandwidth, and concurrent load.

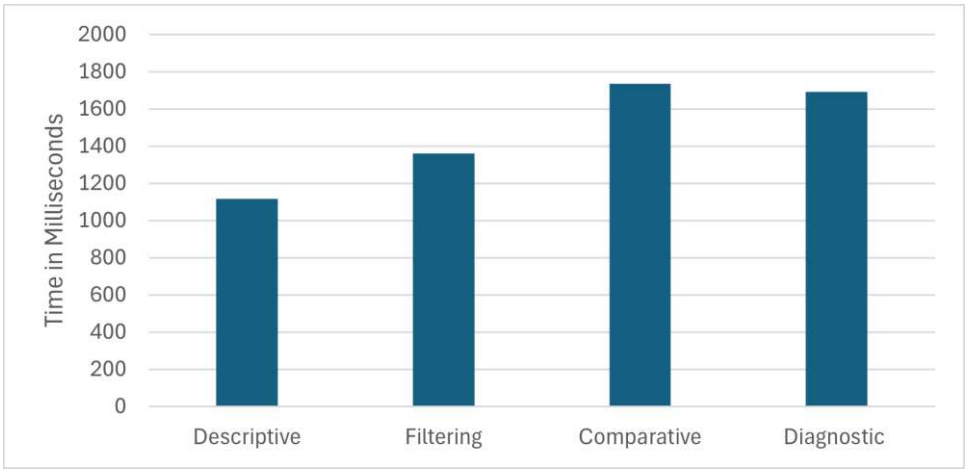


Figure 4. Update Latency by query type (milliseconds). Descriptive and filtering queries complete fastest, while comparative and diagnostic queries incur additional processing but remain within the acceptable 1000–4000 ms threshold.

AI Encoding Accuracy (AEA). For M = 50 representative encoding decisions, 44 were rated optimal (1.0) and 6 acceptable (0.5), giving AEA = 47/50 = 0.94, comfortably above the ≥ 0.90 threshold informed by prior work on ML-based visualisation recommendation, including VizML's approximately 89% accuracy against crowdsourced human consensus (Hu et al. 2019).

Intent Recognition Accuracy (IRA). For 50 representative natural-language queries, 48 were correctly interpreted and 2 misinterpreted, giving IRA = 48/50 = 0.96, indicating excellent natural-language understanding and exceeding the typical 87–93% F1-score range reported for natural-language understanding systems (Coucke et al. 2018).

Encoding Consistency Score (ECS). Twenty queries were each executed three times (60 total executions). Fourteen produced consistent encodings across all repetitions; six were inconsistent for underspecified queries such as "colour by treatment protocols" and "show all patient data." ECS = 14/20 = 0.70, below the acceptable 0.90 threshold, revealing non-reproducible visualisations for 30% of queries and the need for consistent default encoding rules or clarification mechanisms (Nielsen 1994) (Norman Donald 2013).

Visualisation Reasoning Score (VRS). Asked "why was a scatter plot selected to visualise all patients?" the system provided data justification, encoding rationale, and context relevance but did not explicitly consider alternative visualisation types (Figure 5). VRS = 3/4 = 0.75, indicating partial but structured reasoning (Nauta et al. 2023; Hoffman et al. 2018).

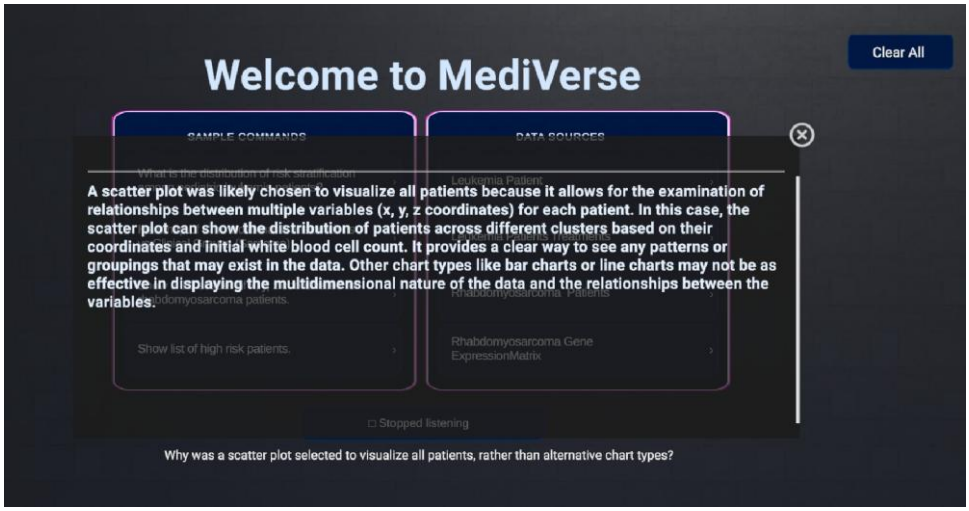


Figure 5. Example of a system-generated visualisation explanation in MediVerse, justifying the selection of a 3D scatter plot for exploring multi-dimensional patient data in response to a "why" query.

The same six metrics were then applied across systems. Following the cross-system setup of Section 5 and using the identical expert-judged rubric applied to MediVerse, the matched pilot compared MediVerse against the state-of-the-art general-purpose LLM model GPT-5.5 baseline (ChatGPT with Data Analysis) on identical data and queries. On this pilot, GPT-5.5 interpreted user intent and selected appropriate encodings without error, yielding IRA = 1.00 and AEA = 1.00, comparable to MediVerse's strong AI-decision scores (IRA = 0.96 and AEA = 0.94). On encoding consistency the ordering is more pronounced. Repeating twenty benchmark queries three times, each repetition issued in an independent session so that no prior response was visible to the model, GPT-5.5 returned an identical result on seventeen of twenty queries (ECS = 0.85), against MediVerse's 0.70. All three disagreements involved underspecified requests. "Show initial WBC counts across all patients" was rendered once as a line chart over patient index and twice as a histogram; "show sample counts by primary site" changed which axis carried the categorical variable; and "show high-risk patients only" was resolved once as the High stratum alone and twice as High together with Very High, returning different record sets under an unchanged encoding. These GPT-5.5 values are preliminary: they were obtained on a small matched pilot and, like the MediVerse scores, were assigned by the single evaluator, so independent multi-rater assessment is left for future work. Both systems produced structured explanations, and two differences are decisive. First, GPT-5.5's written explanations were on average more complete: it consistently and explicitly weighed rejected alternatives (for example, preferring a box plot over a bar chart, or a line chart over a scatter), yielding a mean Visualisation Reasoning Score of 0.92 against MediVerse's 0.75, whose explanations omitted the alternative-consideration component. Second, and more importantly for immersive analytics, GPT-5.5 was roughly an order of magnitude slower, a mean Update Latency of approximately 14 s versus MediVerse's 1.5 s, because each answer required code generation and execution, and it could not provide scene stability at all, since a static 2D image has no persistent spatial layout to preserve. GPT-5.5 also failed one query whose fields spanned a table it had not been given, which MediVerse resolved through its integrated backend schema. Table 4 summarises the comparison.

IVEM Metric	MediVerse (immersive VR)	ChatGPT with Data Analysis (non-immersive, pilot)
Scene Stability Index (SSI)	1.0	N/A (2D display)
Update Latency (UL)	1485 ms	~14,100 ms
AI Encoding Accuracy (AEA)	0.94	1.00
Intent Recognition Accuracy (IRA)	0.96	1.00
Encoding Consistency Score (ECS)	0.70	0.85
Visualisation Reasoning Score (VRS)	0.75	0.92

Table 4. Cross-platform comparison of IVEM results for MediVerse and ChatGPT with Data Analysis (GPT-5.5).

8. Discussion

The validation shows that a multi-dimensional evaluation framework can capture both strengths and optimisation opportunities in AI-assisted immersive visualisation, and can compare systems across the immersive and non-immersive divide. Contemporary immersive analytics increasingly incorporates conversational interfaces, automated visualisation recommendation, adaptive rendering, and explainable reasoning. These capabilities cannot be adequately characterised by usability, task-completion, or graphical-performance measures alone. Rather than replacing established methods, IVEM complements them with a structured methodology that diagnoses multiple dimensions of system behaviour simultaneously.

Positioned against existing evaluation frameworks, IVEM advances beyond approaches that each address only part of the AI-driven immersive visualisation problem. Traditional and immersive-analytics evaluation measures rendering performance, task completion, graphical perception, and user experience (Islam et al. 2024; Qu et al. 2023; Whitlock, Smart, and Szafir 2020); these remain essential for characterising the display and the human interaction but do not assess whether the AI selected an appropriate encoding or justified it. AI visualisation-recommendation research instead evaluates the accuracy of automated design choices, typically against human consensus as in VizML (Hu et al. 2019) and Data2Vis (Dibia and Demiralp 2019), yet reports a single accuracy figure rather than separating correctness, consistency, and reasoning, and disregards rendering stability and immersive responsiveness. Explainable-AI frameworks assess explanation completeness, faithfulness, and interpretability (Vilone and Longo 2021; Nauta et al. 2023; Hoffman et al. 2018; Chatzimparmpas 2025) , but target predictive models and classifiers rather than visualisation decisions, so they cannot judge whether a chosen encoding is warranted by the data. IVEM differs from all three in integrating rendering quality, AI decision quality, and explanation quality within a single quantitative methodology with evidence-based thresholds, and in producing a per-dimension diagnostic profile rather than one aggregate score. In doing so it also begins to address the absence of standardised, cross-system benchmarks that recent surveys identify as a central obstacle in AI-assisted visualisation (Podo, Prenkaj, and Velardi 2024).

The principal strength of IVEM is that it separates complementary aspects of behaviour instead of combining them into a single aggregate score, and applying it to two very different systems shows why this matters: neither platform dominates across all six metrics. On rendering quality, MediVerse maintained perfect spatial stability (SSI = 1.00) and responded in 1485 ms on average, whereas GPT-5.5 required approximately 14 s per answer and, by construction, offers no scene stability, since a static 2D image has no persistent spatial layout to preserve. On AI decision quality, the ordering reverses: on the matched pilot GPT-5.5 interpreted intent and selected encodings without error (IRA = AEA = 1.00), whereas MediVerse achieved strong but slightly lower scores (AEA = 0.94 and IRA = 0.96). Both systems also exhibited non-deterministic behaviour on repeated queries, with MediVerse showing greater variability (ECS = 0.70) than GPT-5.5 (ECS = 0.85). On explanation quality, the reversal is sharper still: GPT-5.5 achieved VRS = 0.92 by explicitly weighing rejected alternatives, whereas MediVerse achieved VRS = 0.75 because its explanations omitted the alternative-consideration component.

Read together, these two profiles provide diagnostic guidance that a single score would conceal. For MediVerse, IVEM identifies two specific and actionable limitations: inconsistent recommendations for ambiguous queries (ECS = 0.70), a deficit highlighted by the baseline score of 0.85, suggesting that more consistent default encoding rules or clarification prompts are needed; and incomplete reasoning (VRS = 0.75), indicating that explanations should explicitly describe alternative visualisation choices. For GPT-5.5, the diagnosis is partly the mirror image: stronger AI decision quality and more complete explanations, but substantially higher latency and no spatial dimension. Encoding consistency, however, remains a weakness shared by both systems rather than a point of separation. In each case, IVEM identifies which component of the analytical pipeline requires improvement, demonstrating the value of reporting a multidimensional evaluation profile rather than a single aggregate score.

The consistency results also point to a failure mode common to both systems rather than to either implementation. Every disagreement observed across repeated runs in both platforms arose on queries that were underspecified rather than on well-formed ones: requests that did not state whether a quantity should be shown per record or as a distribution, which axis should carry a categorical variable, or where the boundary of a qualitative category such as "high risk" lies. Neither system failed on queries with a single defensible reading. This suggests that low Encoding Consistency in AI-assisted visualisation is driven substantially by ambiguity in the request rather than by instability in the model, and that the remedy lies in clarification prompts and documented default encoding rules rather than in model selection alone. It also indicates that ECS is measuring something real and transferable, since it identifies the same category of weakness in two systems built on entirely different architectures.

This diagnostic capability enables meaningful comparison between immersive and non-immersive systems using a common evaluation methodology.

9. Conclusion and Future Work

Artificial intelligence is fundamentally transforming immersive analytics by enabling natural-language interaction, adaptive visualisation, and AI-assisted analytical reasoning within virtual and mixed reality environments. These advances have also exposed limitations in existing evaluation methodologies, which typically assess rendering performance, interaction quality, or AI behaviour independently, making objective comparison difficult and limiting reproducible benchmarking. This paper introduced the Immersive Visualisation Evaluation Model (IVEM), a unified evaluation framework specifically designed for AI-driven immersive visualisation. IVEM integrates three complementary dimensions (Rendering Quality, AI Decision Quality, and Explanation Quality) within a single quantitative framework and defines six formally specified metrics: SSI, UL, AEA, IRA, ECS, and VRS, together providing a structured methodology for evaluating the technical performance, analytical correctness, and transparency of AI-assisted immersive visualisation.

The framework was validated using the MediVerse voice-driven immersive analytics platform across representative biomedical workflows, and extended through a cross-platform comparison against ChatGPT with Data Analysis, a non-immersive general-purpose LLM assistant, on identical data and queries. The evaluation demonstrated that IVEM characterises complementary aspects of system behaviour while identifying diagnostic insights (limited decision reproducibility and incomplete explanation reasoning) that conventional methodologies cannot readily reveal, and that it fairly isolates the value of the immersive interface in cross-system comparison. Because IVEM evaluates observable behaviour rather than implementation details, it is applicable across virtual, augmented, mixed reality, and future AI-assisted visualisation platforms, providing a foundation for reproducible benchmarking, objective cross-system comparison, and community-wide evaluation standards.

Limitations also point to future work. Only one immersive platform was evaluated, and the cross-system run was a matched pilot rather than the final benchmark, so extending the comparison to the more comprehensive query set and to additional immersive platforms will strengthen generalisability. The two systems also rest on different underlying models, Grok-1 within MediVerse and GPT-5.5 for the baseline, so part of the observed difference in AI decision and explanation quality may reflect the language model rather than the interface; isolating these factors, for example by running MediVerse on a matched model, is left for future work. All scoring was carried out by a human evaluator against the IVEM rubric, so no language model was used to evaluate either system. The judgement-based metrics (AEA and VRS) were scored by a single evaluator, which was sufficient to demonstrate the rubric but does not establish inter-rater reliability; future work will therefore involve multiple independent experts, report agreement statistics such as Fleiss' κ (Fleiss 1971) or Krippendorff's (Krippendorff 1996). Because the benchmark tasks were drawn from biomedical workflows, applying IVEM in engineering, manufacturing, urban analytics, and collaborative settings will provide further evidence of versatility, and the proposed thresholds can be refined empirically as additional benchmark datasets accumulate. The framework also remains extensible to further dimensions such as collaboration quality, cognitive workload, and trust calibration as immersive analytics matures.

As immersive analytics continues to evolve from interactive visualisation toward intelligent analytical environments, we envision IVEM serving as a foundational evaluation methodology that supports objective benchmarking, systematic optimisation, and the development of more transparent, reliable, and trustworthy immersive visualisation systems.

Acknowledgements

The authors thank Western Sydney University and the Children's Cancer Institute for their support of this research.

Funding

This research was partially funded by School of Computer, Data and Mathematical Sciences research grant and ARC LIEF LE240100131.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

Rani Adam conceived the study, developed the IVEM framework, implemented the MediVerse platform, designed and conducted the experiments, analysed the results, and wrote the manuscript. Quang Vinh Nguyen, Zhonglin Qu, Daniel R. Catchpoole, Simeon J. Simoff, and Paul J. Kennedy contributed to the conceptual development, supervision, manuscript review, and revision.

Ethics Approval

This study used de-identified biomedical datasets obtained under the relevant institutional ethics approvals.

Consent to Participate

Not applicable.

Consent for Publication

Not applicable.

Data Availability

The datasets analysed during the current study are not publicly available due to ethical and privacy restrictions but may be available from the corresponding author on reasonable request and subject to the relevant approvals.

References

Adam, Rani, Daniel R Catchpoole, Simeon J Simoff, Zhonglin Qu, Paul J Kennedy, and Quang Vinh Nguyen. 2025a. 'Lossless compression with trie-based shared dictionary for omics data in edge–cloud frameworks', *Journal of Sensor and Actuator Networks*, 14: 41.

———. 2025b. "MediVerse: AI-Powered Interactive Voice-Driven Virtual Reality for Health Data Analytics." In *Pacific Rim International Conference on Artificial Intelligence*, 35-50. Springer.

Adam, Rani, Daniel R Catchpoole, Simeon S Simoff, Paul J Kennedy, and Quang Vinh Nguyen. 2024. 'Novel hybrid edge-cloud framework for efficient and sustainable omics data management', *Innovations in Digital Health, Diagnostics, and Biomarkers*, 4: 81-88.

Bhattacharya, Sharmistha, Lisa Marie Prinz, and Tintu Mathew. 2025. "What you see is what you map: Exploring the impact of viewpoint transitions on cognitive mapping in vr." In *International Conference on Human-Computer Interaction*, 145-65. Springer.

Chatzimparmpas, Angelos. 2025. 'Visual analytics for explainable and trustworthy artificial intelligence', *IEEE Computer Graphics and Applications*, 45: 100-11.

Cherep, Lucia A, Alex F Lim, Jonathan W Kelly, Devi Acharya, Alfredo Velasco, Emanuel Bustamante, Alec G Ostrander, and Stephen B Gilbert. 2020. 'Spatial cognitive implications of teleporting through virtual environments', *Journal of Experimental Psychology: Applied*, 26: 480.

Cordeil, Maxime, Andrew Cunningham, Benjamin Bach, Christophe Hurter, Bruce H Thomas, Kim Marriott, and Tim Dwyer. 2019. "IATK: An immersive analytics toolkit." In *2019 IEEE conference on virtual reality and 3D user interfaces (VR)*, 200-09. IEEE.

Coucke, Alice, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, and Thibaut Lavril. 2018. 'Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces', *arXiv preprint arXiv:1805.10190*.

Dibia, Victor, and Çağatay Demiralp. 2019. 'Data2vis: Automatic generation of data visualizations using sequence-to-sequence recurrent neural networks', *IEEE Computer Graphics and Applications*, 39: 33-46.

Fleiss, Joseph L. 1971. 'Measuring nominal scale agreement among many raters', *Psychological bulletin*, 76: 378.

Hoffman, Robert R, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. 'Metrics for explainable AI: Challenges and prospects', *arXiv preprint arXiv:1812.04608*.

Hu, Kevin, Michiel A Bakker, Stephen Li, Tim Kraska, and César Hidalgo. 2019. "Vizml: A machine learning approach to visualization recommendation." In *Proceedings of the 2019 CHI conference on human factors in computing systems*, 1-12.

Islam, Md Rafiqul, Shanjita Akter, Linta Islam, Imran Razzak, Xianzhi Wang, and Guandong Xu. 2024. 'Strategies for evaluating visual analytics systems: A systematic review and new perspectives', *Information Visualization*, 23: 84-101.

König, Sabine U, Ashima Keshava, Viviane Clay, Kirsten Rittershofer, Nicolas Kuske, and Peter König. 2021. 'Embodied spatial knowledge acquisition in immersive virtual reality: Comparison to map exploration', *Frontiers in Virtual Reality*, 2: 625548.

Krippendorff, Klaus. 1996. *Content Analysis: An Introduction to its Methodology* (TPB).
Luo, Yuyu, Nan Tang, Guoliang Li, Jiawei Tang, Chengliang Chai, and Xuedi Qin. 2021. 'Natural language to visualization by neural machine translation', *IEEE transactions on visualization and computer graphics*, 28: 217-26.
Maslych, Mykola, Mohammadreza Katebi, Christopher Lee, Yahya Hmaiti, Amirpouya Ghasemaghaei, Christian Pumarada, Janneese Palmer, Esteban Segarra Martinez, Marco Emporio, and Warren Snipes. 2025. "Mitigating response delays in free-form conversations with LLM-powered intelligent virtual agents." In *Proceedings of the 7th ACM Conference on Conversational User Interfaces*, 1-15.
Miller, Robert B. 1968. "Response time in man-computer conversational transactions." In *Proceedings of the December 9-11, 1968, fall joint computer conference, part I*, 267-77.
Nauta, Meike, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice Van Keulen, and Christin Seifert. 2023. 'From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai', *ACM Computing Surveys*, 55: 1-42.
Nielsen, Jakob. 1994. 'Usability engineering Morgan Kaufmann Publishers Inc', *Burlington, Massachusetts (San Francisco, California until 2008)*.
Norman Donald, A. 2013. *The design of everyday things* (MIT Press).
OpenAI. 2026. *Introducing GPT-5.5* (OpenAI).
Podo, Luca, Bardh Prenkaj, and Paola Velardi. 2024. 'Agnostic visual recommendation systems: Open challenges and future directions', *IEEE transactions on visualization and computer graphics*, 31: 1902-17.
Qu, Zhonglin, Simeon J Simoff, Paul J Kennedy, Daniel R Catchpoole, and Quang Vinh Nguyen. 2023. 'Visualisation for explainable machine learning in biomedical data analysis.' in, *Data Driven Science for Clinically Actionable Knowledge in Diseases* (CRC Press).
Shen, Leixian, Enya Shen, Yuyu Luo, Xiaocong Yang, Xuming Hu, Xiongshuai Zhang, Zhiwei Tai, and Jianmin Wang. 2022. 'Towards natural language interfaces for data visualization: A survey', *IEEE transactions on visualization and computer graphics*, 29: 3121-44.
Stauffert, Jan-Philipp, Florian Niebling, and Marc Erich Latoschik. 2020. 'Latency and cybersickness: Impact, causes, and measures. A review', *Frontiers in Virtual Reality*, 1: 582204.
Vilone, Giulia, and Luca Longo. 2021. 'Notions of explainability and evaluation approaches for explainable artificial intelligence', *Information Fusion*, 76: 89-106.
Wang, Chaoming, Valeria Garro, Veronica Sundstedt, Yan Hu, and Prashant Goswami. 2025. 'Immersive analytics meets artificial intelligence: A systematic review', *Computational Visual Media*, 12: 1-34.
Whitlock, Matt, Stephen Smart, and Danielle Albers Szafir. 2020. "Graphical perception for immersive analytics." In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, 616-25. IEEE.