

A Multi-Context Usability and User Experience Evaluation of a Configurable Virtual Reality Environment for Medical Imaging Education

Alina Duca

Faculty of Computer Science
“Alexandru Ioan Cuza” University
Iași, Romania
alina_duca@yahoo.com

Adrian Iftene

Faculty of Computer Science
“Alexandru Ioan Cuza” University
Iași Romania
adiftene@gmail.com

ABSTRACT

Virtual Reality (VR) is increasingly used in medical education, but adapting an immersive learning environment to a new subject typically requires programming skills that medical educators do not have. Configurable environments promise to remove this barrier: a medical educator retargets the system to a new domain by authoring content rather than by writing code. Such systems are usually evaluated in a single context and for a single role, so the surface the educator actually uses is rarely evaluated at all. We report a usability and user experience evaluation of a VR environment for medical imaging education that spans an immersive VR module, a desktop web client, and a tumour detector with visual explanations. Twenty-eight participants—86% from medicine, veterinary medicine, and the life sciences, and 68% with no prior VR experience—rated seven modules across both interaction contexts and both the learner and the author role. Every module scored at or above 4.0 on a five-point scale except the automated guided corridor (3.71), the only element that takes locomotion out of the user’s hands, and the only one anybody rated negatively. Perceived usefulness ran above interaction quality for both AI-bearing features, small but consistent in direction. Retargeting the environment from brain to cardiac imaging confirmed that adapting it to a new domain is an authoring task rather than a development task. We draw three design implications for configurable immersive learning environments.

AUTHOR KEYWORDS

Usability evaluation; multi-context evaluation; virtual reality; human-AI interaction; explainable AI.

DOI: 10.37789/icusi.2026.18

INTRODUCTION

A recurring ambition in immersive learning research is to make an environment *configurable*, so that a domain expert who cannot program—typically a medical teacher—can re-target it to a new subject by supplying content instead of editing code. The argument is economic: authoring a lesson is cheap, whereas building a virtual environment is not, so configurability is what turns a research prototype into something an institution can actually reuse. The authoring-tool literature has made this case for Augmented Reality (AR) and VR broadly [2, 13, 22], and it inherits a longer tradition of end-user development [16, 24]. The claim is attractive, but the way such systems are evaluated tends to leave it untested in three specific ways.

First, evaluations are usually *single-role*. They recruit learners and measure what learners feel, which is reasonable, but a configurable system by definition has a second class of user—the author—whose interface is a different artefact with different demands. If nobody evaluates the authoring surface, the configurability claim rests on the fact that the developers themselves could use it. *Second*, evaluations are usually *single-context*. Immersive learning systems are rarely only immersive; they typically pair a headset experience with a desktop surface for administration, upload, or review. Reporting a single aggregate usability score across a system that spans a head-mounted display and a browser hides exactly the comparison that a designer needs. *Third*, evaluations are usually run on the wrong people. Immersive learning systems for medicine are frequently tested on samples recruited from the discipline that builds them rather than from the discipline that will use them, and a participant who is comfortable with a head-mounted display is precisely the participant least likely to notice where the interface asks too much.

This paper takes those three gaps as its evaluation design. We study a configurable VR environment for medical imaging education—a research prototype designed and implemented by the authors, whose intended users are students of medicine and allied health disciplines in the learner role and their educators in the author role—and contribute the evaluation rather than the system: a study with 28 participants drawn deliberately from medicine, veterinary medicine and the life sciences rather than from computer science, rating seven modules across two interaction contexts and two user roles, followed by an end-to-end retargeting exercise in which the brain tumour environment is turned into a cardiac imaging environment through the authoring interface alone. Our contributions are:

- (1) A **multi-context, dual-role usability and UX evaluation** ($n = 28$) of a configurable immersive medical learning environment, reported at the level of individual modules and separately for the immersive and desktop contexts, with a cohort that is domain-expert rather than technical.
- (2) The observation of a small but consistent **usefulness / interaction-quality gap**: for both AI-bearing features, participants endorsed the value of the feature more readily than the quality of the interaction that delivered it.
- (3) An **end-to-end verification** of the configurability claim: retargeting the environment to a new medical domain—the declared goal of authoring-based systems, not a novelty of ours—was completed through

the web form alone, without touching the Unity project.

- (4) Three **design implications** for configurable immersive learning environments, concerning onboarding locomotion, the separation of usefulness from usability in AI-infused interfaces, and the confidence gap on the authoring surface.

RELATED WORK

Immersive learning and its evaluation

Immersive VR reliably improves engagement, spatial understanding, and self-reported motivation in health-sciences education relative to flat displays [7, 9, 14, 21, 25], and systematic reviews of immersive applications in higher education report broadly positive but methodologically uneven evidence [26]. The more careful findings are the qualified ones. Makransky et al. showed that adding immersion to a science simulation increased presence yet *decreased* learning, because the immersive condition also increased extraneous cognitive load [19]; the CAMIL model subsequently formalised the point that immersion acts on learning only indirectly, through affective and cognitive mediators [18]. This matters for how we read our own results: enthusiasm for an immersive feature is not evidence that the feature teaches, and we are careful below to report what participants judged rather than what they learned. Locomotion is the sub-topic most relevant to our strongest negative finding. Comparative studies of VR locomotion techniques consistently find that the technique mediates comfort, control and sickness [4, 5], and that motion the user does not initiate is a well-known contributor to cybersickness [27]. Our study offers a small naturalistic confirmation from a cohort that had, in the main, never worn a headset.

End-user configurability and authoring

End-user development seeks to let people who are not programmers adapt systems to their own purposes [16, 24], often by expressing the variable part of a system declaratively [20]. For immersive media specifically, Nebeling and Speicher catalogue the trouble with AR/VR authoring tools—the gulf between low-fidelity tools that non-experts can use and high-fidelity tools that require a developer [22]—and Ashtari et al. document how that gulf plays out in the practices of actual AR and VR creators [2]. The system we evaluate sits squarely in this tradition: a lesson is a structured record, resolved at runtime, and the authoring surface is a web form. What this literature rarely does, however, is put the authoring surface in front of the domain experts it is built for and ask them to rate it, which is one of the things we do here.

Explanations and trust in AI-assisted interfaces

The environment we evaluate does not only present authored content: it also runs a learned model over medical images and shows the learner a label, a confidence score, and a visual explanation. That places it inside the literature on human–AI interaction. Displayed confidence and explanations change how much users rely on an AI, and not always in the direction of better joint performance [3, 31]; cognitive forcing functions can reduce overreliance, at a cost in user satisfaction [6]. The Guidelines for Human-AI Interaction make the relevant prescription explicit: make clear how well the system can do what it can do, and make clear why the system did what it did [1]. Grad-CAM [28] is the standard way of discharging

the second obligation and is what the system we study uses, following a broader move towards explanation in medical image analysis [10, 17]. Our results speak to the first obligation from an unusual angle: participants rated the *usefulness* of the AI feature above the *quality of the interaction* that delivered it, which is a pattern any evaluator of an AI-infused interface should expect to have to disentangle.

Evaluation method

Finally, the study is a questionnaire-based evaluation conducted around a demonstrated rather than a performed interaction. The tradition of critically comparing usability evaluation methods [11], and of asking what a given protocol actually measures [12], informs how we bound our claims in Section *Limitations and future work*. Following the long-standing distinction between perceived usefulness and perceived ease of use [8, 30], we elicit two constructs separately for the AI-bearing features: *interaction quality*, by which we mean how easy and pleasant a feature is to operate—close to perceived ease of use, though rated here for a demonstrated interaction—and *perceived usefulness*, the value a participant attaches to what the feature does. Module-level evaluations commonly report a single overall score per module; we keep the two constructs apart, and we report distributions alongside means because the data are ordinal.

THE EVALUATED SYSTEM

The evaluated environment is a research prototype designed and implemented by the authors, and this study is its first evaluation with users. In the following, we summarize it only to the extent necessary to facilitate understanding of the evaluation, and we omit implementation details that have no bearing on what participants rated.

Two contexts. The system (Fig. 1) presents two interaction surfaces. The *immersive VR module*, built in Unity and used with a head-mounted display, contains an automated guided corridor that carries the user into the environment, a main menu, an embodied conversational tutor, a quiz that shows an MRI slice and asks for the tumour type, a lecture module, an anatomical *brain components* scene with selectable labelled structures, and a *brain tumours* scene in which a volumetric tumour model—generated directly from a NIfTI volume—can be sliced open with a clipping plane. The *desktop web client* provides an AI detection page, where an MRI or CT image is uploaded and returned with a bounding box, a predicted label, a confidence score, and a Grad-CAM overlay [28], and an environment configuration page.

The backend. A single cloud-hosted service hosts three modules. The *CNN module* performs detection and, on request, re-trains on a dataset uploaded by a teacher. The *VR configuration module* stores lesson configurations in Cloud Storage and serves them to the VR client, converting the uploaded `.fbx/.glb` models into Unity asset bundles. The *LLM module* drives the tutor, described next.

The conversational tutor. The tutor is operated by push-to-talk: the learner holds a controller button and speaks, in English or Romanian. The server transcribes the utterance, retrieves relevant passages from the teacher-supplied reference document—for the brain configuration, the NCCN Clinical Practice Guidelines for Central Nervous System Cancers—and a hosted large language model composes the answer, which is returned as text and as synthesised speech with lip

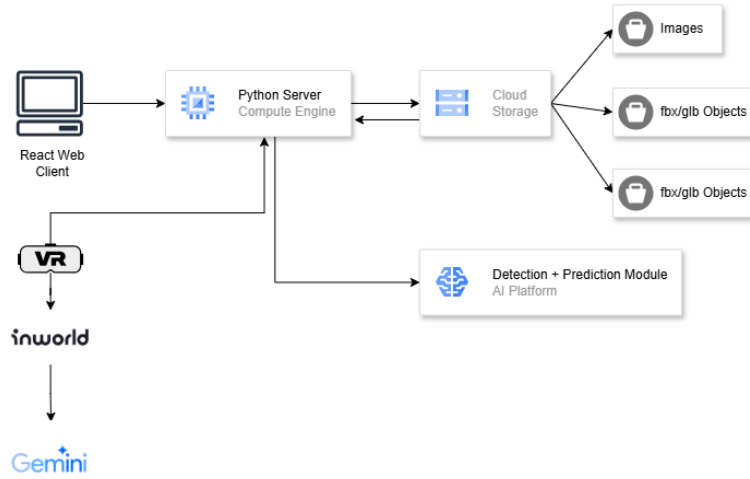


Figure 1: Platform architecture. A React web client and Unity VR client connect to a FastAPI server on a GPU instance. The server hosts the LLM, CNN detection, and VR modules and manages lesson assets in Cloud Storage. Learners use both clients, while authors use only the web client.

synchronisation on the embodied avatar. Two design decisions matter for the evaluation: the tutor is deliberately grounded in the teacher-supplied document rather than in an open-ended model, so that its answers stay within the authored lesson; and it answers within 3–4 s, after an earlier locally hosted variant proved unusable at a latency of roughly 20 s.

Two roles. A *learner* uses the VR module and the detection page. An *author* (a teacher) uses the configuration page: they name the lesson, select which modules will appear, and attach 3D models, documents, and videos. Saving produces a six-digit environment identifier, which a learner types into the VR client at launch; the client then resolves the configuration and builds the scene at runtime. No Unity rebuild and no redeployment are involved, which is the claim the retargeting exercise in Section *Retargeting the environment to a new domain* puts to the test.

The perception layer. Detection uses a YOLOv8-Nano classifier [15] fine-tuned from ImageNet weights, with Grad-CAM computed on the last convolutional block and thresholded to yield a bounding box. On the public brain tumour MRI benchmark [23] this configuration reaches 97.33% top-1 accuracy over four classes; it was selected from a comparison of six architectures, and it is the model the interface serves to learners. Figure 2 shows what a learner actually receives from the detection page: the uploaded slice, the predicted class drawn as a labelled bounding box, and the Grad-CAM overlay from which that box is derived. The explanation is what makes the output inspectable rather than merely asserted, and it is the feature whose usefulness participants rated highest in Section *Results*.

EVALUATION DESIGN

Research questions

RQ1 *How do domain learners, as opposed to technically trained participants, rate the usability and experience of each module—and does the immersive context fare differently from the desktop context?*

RQ2 *For the features that are backed by a learned model, does perceived usefulness track interaction quality, or do the two come apart?*

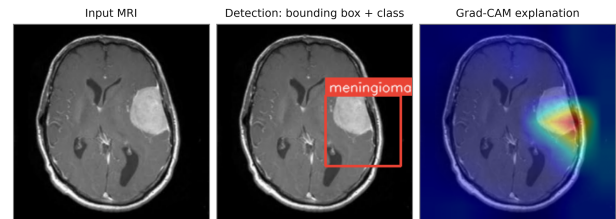


Figure 2: What the AI detection page returns for a meningioma case. The MRI slice (left) is classified by YOLOv8-Nano (97.33% accuracy) [23]; the predicted class is shown as a bounding box (centre), while the Grad-CAM overlay (right) highlights the region driving the decision, helping learners understand both *what* the model predicted and *why*.

RQ3 *Is the authoring surface legible to people who are not developers?*

RQ4 *Can the environment be retargeted to a new medical domain through the authoring interface alone?*

Participants

The environment’s target population is students of medicine and allied health disciplines, who meet it as learners, together with the educators who teach them, who meet it as authors. Twenty-eight participants took part. The cohort was recruited to be domain-representative rather than technically representative: 67.9% study medicine or worked in healthcare, 17.9% veterinary medicine, 7.1% biology and the life sciences, 3.6% bioengineering, and a single participant (3.6%) comes from a technical field. Eighty-six percent are therefore prospective domain users of the environment, and only one participant is a plausible developer of it.

Table 1 gives the profile. Two features of it drive the interpretation of our results. The cohort is *domain-strong*: mean self-rated familiarity with brain anatomy and brain tumours is 4.18 (SD 0.94) on a five-point scale, with 78.5% at 4 or 5. And it is *technology-naïve* with respect to the medium: 67.9% had never used a VR application, none used one more than rarely, no participant described themselves as very experienced with head-mounted displays, and 57.1% had never used any VR

Attribute	Level	%
Age	18–24	92.9
	25–34	7.1
Gender	Female	78.6
	Male	17.9
	Other	3.6
Field	Medicine / healthcare	67.9
	Veterinary medicine	17.9
	Biology / life sciences	7.1
	Technical or bioengineering	7.1
Prior VR use	None	67.9
	Some	32.1
VR frequency	Never	53.6
	Rarely	46.4
HMD experience	Inexperienced	60.7
	Neutral	28.6
	Somewhat experienced	10.7
Prior medical VR/mobile app	Never	57.1
	Mobile only	32.1
	Both	10.7

Table 1: Participants ($n = 28$). The cohort is domain-expert but VR-naïve, unlike typical immersive learning study populations.

or mobile application for medical education. This is the population the system is for, and it is not the population such systems are usually tested on.

Protocol

The study ran in three stages. Participants first completed a pre-questionnaire capturing demographics and prior experience. They were then taken through the system by means of two structured video walkthroughs, one per role: a *learner* walkthrough covering account creation, MRI upload and classification with the Grad-CAM explanation, and then, in VR, the corridor, the main menu, a conversation with the tutor, four quiz items with correct and incorrect feedback, the lecture player, the brain components scene and the brain tumours scene; and an *author* walkthrough in which a teacher account configures a new environment—selecting modules, attaching a 3D model, documents and a video—receives the generated environment identifier, and then enters that identifier in the VR client and finds the environment rebuilt around the authored content. Participants then completed a post-questionnaire. Every module named in Section *The Evaluated System* was rated separately, on five-point scales anchored at 1 (poor / very difficult) and 5 (excellent / very easy), and both the authoring page and the detection page were rated in the same way. Two further items asked about the *usefulness* of the two AI-bearing features, phrased and scaled identically to the corresponding experience items, so that the two constructs could be compared directly. Free-text items collected suggestions and criticism. The full instrument is listed in the Appendix. We stress what this protocol does and does not deliver, because it bounds every claim we make. Participants judged an interaction that was *demonstrated* to them rather than one they *performed*. This is what made it possible to reach 28 domain participants instead of the small mixed-background samples that hands-on headset sessions realistically permit, and it is why the cohort in Table 1 looks like the intended user population. It also means that our “ease” ratings are anticipations of ease rather than measurements of it, that we have no interaction telemetry, no task-completion times, and—importantly, given the corridor result—no instrumented measure of presence or cybersickness [27, 29]. We return to this in Section *Limitations and future work*.

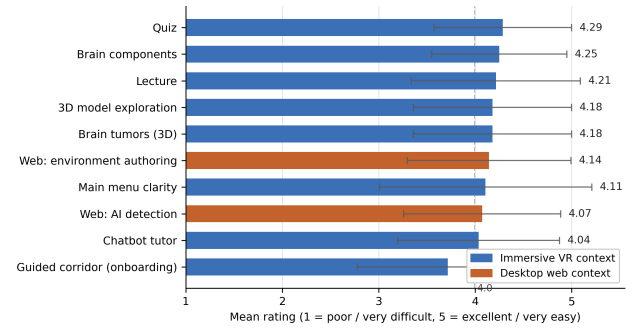


Figure 3: Mean module ratings ($n = 28$), sorted ascending by interaction context. Nine of ten items score at least 4.0; only the automated guided corridor scores lower.

Analysis

With $n = 28$ and ordinal responses, we report means with standard deviations *and* the underlying distributions, and we do not perform significance testing on the between-item differences. Where we describe a difference as consistent, we mean consistent in direction across items, not statistically established; this is stated again at each point where it matters. In particular, the learner-facing and author-facing items are reported side by side for completeness: we do not compare the two roles inferentially, since they measure different artefacts.

RESULTS

Module-level ratings across two contexts (RQ1)

Figure 3 reports every rated item. The headline is that the system is broadly well received by a cohort that had mostly never used VR: 82.1% found the VR functionalities easy to understand (14.3% “somewhat”, one participant “no”), and nine of the ten rated items sit at or above 4.0.

The interesting result is the one that does not. The **automated guided corridor**—a scripted camera move through hospital corridors, used as an onboarding device before the main menu—scores **3.71 (SD 0.94)**, the lowest item by a clear margin. Its distribution is the diagnostic part: half of the participants (14/28) sat on the neutral midpoint, and it is the *only* item in the entire instrument to attract a rating below the midpoint. Nothing else in the system was rated negatively by anyone. The corridor is also the only element in the system in which locomotion is taken out of the user’s hands. Every other VR module—the tumour scenes, the components scene, the quiz, the tutor—is something the user does; the corridor is something done to the user. That the single passively experienced element is the single worst-received one, in a cohort with almost no headset experience, is consistent with what the locomotion and cybersickness literature would predict [5, 27], and we treat it in Section *Discussion* as the clearest design signal in the study. Comparing the two contexts, we find no penalty for immersion. The web items (authoring 4.14, detection 4.07) sit in the same band as the VR items rather than above them, and the two highest-rated items in the study are both immersive: the quiz (4.29, SD 0.71) and the brain components scene (4.25, SD 0.70), which also have the tightest distributions in the study. The often-reported finding that a browser interface is “easier” than a headset does not reproduce here—plausibly because the protocol removed the physical cost of wearing the headset, which is precisely the kind of effect a demonstrated interaction cannot capture.

Item	1	2	3	4	5	M	SD
Guided corridor	0	1	14	5	8	3.71	0.94
Chatbot tutor	0	1	6	12	9	4.04	0.84
Web: AI detection	0	0	8	10	10	4.07	0.81
Main menu	1	1	6	6	14	4.11	1.10
Web: authoring	0	0	8	8	12	4.14	0.85
3D model exploration	0	0	7	9	12	4.18	0.82
Brain tumours	0	0	7	9	12	4.18	0.82
Lecture	0	2	2	12	12	4.21	0.88
Brain components	0	0	4	13	11	4.25	0.70
Quiz	0	0	4	12	12	4.29	0.71
<i>Usefulness items (asked separately)</i>							
3D tumour usefulness	0	0	8	4	16	4.29	0.90
AI detection usefulness	0	0	6	7	15	4.32	0.82

Table 2: Full response distributions ($n = 28$), with mean and standard deviation. The corridor is the only item rated below the midpoint; the main menu shows the widest dispersion, while the quiz and brain components scene show the narrowest.

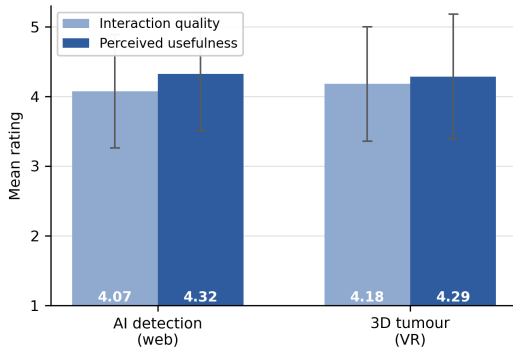


Figure 4: For both AI features, perceived usefulness exceeds interaction quality (+0.25 and +0.11). While not statistically significant, both gaps suggest that users accept the value faster than the interface.

The **main menu** deserves a note: its mean of 4.11 conceals the widest dispersion in the study (SD 1.10), with 14 participants rating it 5 and two rating it 1 or 2. A mean of 4.11 would suggest a menu that needs no work; the distribution says it works very well for most and fails outright for a few. This is the practical argument for reporting distributions rather than means in studies of this size, and Table 2 therefore gives the full response counts for every item, so that a reader can see where a mean is carried by consensus and where it is carried by an average of disagreement.

The interaction-quality gap (RQ2)

Two features are backed by the learned model: the web detection page and the 3D tumour visualisation with its clipping plane. For both, we asked separately about the experience of using the feature and about its usefulness (Fig. 4). For AI detection, interaction quality scores 4.07 (SD 0.81) while usefulness scores 4.32 (SD 0.82), a gap of +0.25. For the 3D tumour, exploration ease scores 4.18 (SD 0.82) and the usefulness of being able to see through the model scores 4.29 (SD 0.90), a gap of +0.11. Fifteen of 28 participants gave AI detection usefulness the maximum rating, and 16 of 28 did so for the see-through tumour—in each case, more than gave the corresponding experience item a 5. These gaps are small, and with ordinal data and $n = 28$ we make no claim that either is significant. What we do observe is that both run in the same direction, and that this direction is the one that should worry an evaluator: participants endorsed *what the*

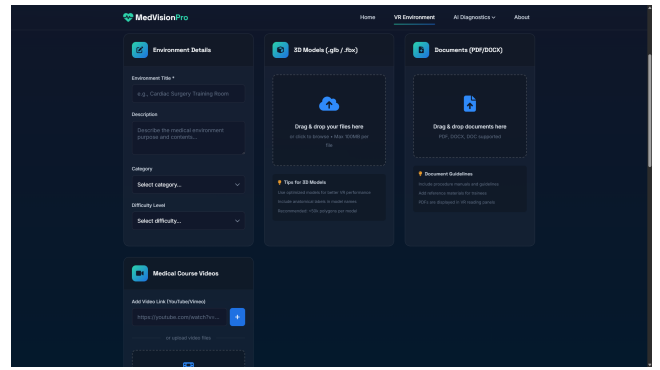


Figure 5: The authoring surface lets teachers configure modules and attach lesson assets, generating an identifier for VR access. Rated 4.14 (SD 0.85), with 28.6% neutral responses—the joint-highest in the study.

feature is for more readily than how it felt to use. A single blended usability score over this system would have been pulled upward by enthusiasm for the capability, and would have reported an interface in better shape than the one participants actually described. The separation of the two constructs is old advice [8], but it is advice that AI-infused interactive systems make newly urgent, because the capability on offer is exactly the thing most likely to generate enthusiasm.

The authoring surface (RQ3)

The authoring page (Fig. 5) scored 4.14 (SD 0.85). Twelve participants rated it 5 and none rated it below the midpoint, which we read as support for the basic claim: a configuration surface of this kind is legible to domain experts who are not developers, and 27 of our 28 participants are not developers. The qualification is in the shape of the distribution. Eight participants (28.6%) sat neutral—jointly the largest neutral band in the study, alongside the corridor. Nothing about the page was rated *badly*; a substantial minority simply declined to endorse it. Read together with the fact that these same participants rated the learner-facing modules confidently in both directions, we read the neutral band as uncertainty about the authoring task rather than dissatisfaction with the form: participants could see what the page did without being sure they could use it to build a lesson of their own. This is a familiar pattern in end-user development, where the barrier is typically not the syntax of the tool but the user’s confidence in their standing to use it [2, 16].

Overall, 78.6% of participants said the system would make learning tumours and medical imaging easier for students, 21.4% said “maybe”, and *no participant said no*.

Retargeting the environment to a new domain (RQ4)

The three preceding results concern how the system is experienced. This one concerns whether the configurability claim survives contact with a domain the environment was not built for. We retargeted the environment from brain tumours to cardiac imaging by the route the system prescribes. Acting as an author, we configured a cardiac environment through the web form of Fig. 5: a 3D heart model, cardiac reference documents for the tutor’s retrieval component, a lecture video, and a cardiac imaging dataset for the detector. The VR client then resolved the resulting six-digit identifier and rebuilt the scene around the authored content—the carousel now offering cardiac pathologies rather than brain tumours, the tutor

answering questions from the cardiac document, and the quiz drawing on cardiac images. The detector was retrained on the uploaded dataset from within the same interface, a step that completed in roughly nine minutes, after which the detection page returned a bounding box on an uploaded cardiac MRI with the label *hypertrophic cardiomyopathy* at 87% confidence. The point of the exercise is what it did *not* require. No Unity code was written, no scene was rebuilt in the editor, and no service was redeployed; the entire domain change was expressed as data through a web form that our participants rated 4.14 for legibility. Authoring-based retargeting is the declared purpose of configurable environments rather than a novelty of ours; what the exercise establishes is that *this* system delivers it end to end. On the terms in which configurability is usually claimed—that adapting to a new domain is an authoring task rather than a development task—the claim holds, and it holds for the same surface that the study in Section *Results* found to be usable by domain experts who are not programmers.

What participants asked for

Three free-text items invited participants to say what should be improved, what they would like to see added, and anything else they wished to record. What they wrote is worth reporting in its own right, because the requests cluster in a way the rating scales do not reveal: with one exception, they ask for *more agency and more clinical realism*, not for a simpler interface. The largest cluster concerns **making the pathology dynamic rather than static**. Participants asked to simulate how a tumour might grow over time under different clinical scenarios, and to explore how different treatment options—surgery, radiation, chemotherapy—would change the outcome. The environment currently presents a tumour as a fixed object to be inspected; these participants, most of them medical students, want it to be a process they can intervene in. A related request was to correlate a tumour with the patient's symptoms, which asks the system to close the loop between the image and the presentation that the student will actually meet in a clinic. A second cluster concerns **study mechanics**: the ability to take quick notes inside the VR environment, progress tracking across sessions, procedure-based training, and enlarging the anatomical images inside the quiz module, which several participants found too small to read comfortably. A third asked for **multi-user collaboration**, so that a group of students, or a student and an instructor, could occupy the environment together. Finally, and consistent with the chatbot's position near the bottom of Table 2, participants asked for the tutor's answers to be able to **include images** rather than being text and speech alone. A tutor that can say “a meningioma looks like this” and show the slice is a different pedagogical instrument from one that can only describe it, and this is arguably the single clearest actionable request in the free-text data. We note the shape of this feedback rather than quantifying it: with 28 participants and open questions, counting mentions would give a spurious precision. But the direction is unambiguous, and it points away from simplification and towards depth—which is a comfortable result for a system whose modules all scored above 4.0, and an uncomfortable one for the assumption that VR-naïve users primarily want things made easier.

DISCUSSION

Do not take agency away during onboarding

The corridor was designed as a welcome. It was received as the worst part of the system, by the widest margin in the study, and it is the only element to have been rated negatively at all. It is also the only element in which locomotion is not the user's decision. For a cohort in which 67.9% had never worn a headset, the system's first act was to move them somewhere they had not chosen to go. The design implication is narrow and actionable: **the onboarding sequence of an immersive system is the worst possible place to spend the user's agency**. If a scripted transition is wanted for narrative reasons, it should be skippable, speed-adjustable, or replaced by a user-initiated teleport [5]. The general point is that passive transport is usually a designer's convenience—it guarantees the user arrives in the right place facing the right way—and that this convenience is paid for by the user, at the moment they are least equipped to absorb the cost.

Separate usefulness from usability, especially over AI

Both AI-bearing features were rated more useful than they were pleasant to use (Fig. 4). The effect is small, and we do not claim significance for it, but its direction is the one that matters for evaluation practice: had we collected a single blended usability score, the enthusiasm participants plainly felt for automated tumour detection would have been folded into it, and the resulting number would have described an interface in better condition than the one they actually reported. The recommendation is therefore methodological. **In systems whose headline feature is an AI capability, usefulness and interaction quality should be elicited as separate, identically scaled items**, and the difference between them reported. The distinction is old [8, 30]; what is new is the strength of the confound, because a capability that feels like magic is unusually good at borrowing credit for the interface that delivers it. It also bears on how such interfaces communicate: a learner who is predisposed to find automated detection valuable is a learner who will accept its output readily, which is exactly the condition under which calibrated confidence and honest explanation matter most [1, 6, 31].

A legible authoring surface is not yet a confident author

The authoring page was rated well (4.14) and never rated badly, and the retargeting exercise in Section *Retargeting the environment to a new domain* confirms that it does what it claims. Yet it also drew the joint-widest band of participants who declined to commit either way. We read this as the gap between *understanding* an authoring interface and *feeling entitled to use it*: a medical student can see that the form uploads a model and generates an identifier without believing that building a lesson is something they would be able, or permitted, to do. The implication is that **a configurable system needs to invest in authoring confidence, not only authoring legibility**. Worked examples, a cloneable starter environment, a preview of the resulting VR scene before publication, and a visible record of what other teachers have built would all address the neutral band more directly than further simplification of a form that participants already understood. This is the familiar territory of end-user development, where adoption tends to be limited less by the syntax of the tool than by the user's sense of standing [2, 16, 24].

What the protocol bought, and what it cost

The evaluation reached 28 participants, of whom 86% come from medicine, veterinary medicine, and the life sciences, and 68% had never used VR—a cohort that resembles the system’s intended users far more closely than the small, technically skewed samples that hands-on headset sessions typically afford. That reach was bought by demonstrating the interaction rather than having participants perform it, and the price was real: no telemetry, no completion times, no presence or cybersickness instrument, and ease ratings that are anticipations rather than measurements. The corridor finding is the sharpest illustration of the trade in both directions—a cohort this VR-naïve is exactly the right one to notice that passive transport is unpleasant, and a protocol without a sickness instrument is exactly the wrong one to say why. We would use a demonstrated-interaction protocol again to recruit the right people and to survey a system this broad, and we would not use it to make any claim about embodied experience.

Limitations and future work

Beyond the protocol, three limitations bound these results. The sample is a convenience sample, young (92.9% aged 18–24) and skewed female (78.6%), so the ratings should not be read as population estimates. There is no control group and no comparison against conventional instruction: we can say participants expect the system to help, not that it does—a distinction the immersive-learning literature has shown to be more than academic [19]. And the differences we highlight between items are small and untested for significance, which is why we have deliberately reasoned from the shape of distributions rather than from the ordering of means. These limitations set the future work. The immediate next step is a hands-on evaluation in which participants—where possible, the same cohort—perform the interactions themselves rather than watch them demonstrated, instrumented with task-completion times and interaction telemetry, and paired with standardised presence and cybersickness instruments, so that the corridor result in particular can be checked against a measurement rather than inferred from a rating. Two further studies follow directly from the findings: a controlled comparison against conventional instruction, to test whether the learning benefit participants expect is real [19], and an authoring study in which practising medical educators build a complete lesson end to end, to test whether the confidence gap we observed on the authoring surface closes with use.

ETHICS AND PRIVACY STATEMENT

Participation was voluntary, anonymous, and unpaid; participants gave informed consent, no personal data beyond coarse demographics were collected, and all imaging data used in the system and in this study come from public, de-identified research datasets. The system is an educational aid and not a diagnostic device: its detector is trained on public benchmarks, its confidence scores are not clinically calibrated, and it must not be used to support clinical decision-making. Because the environment is configurable by teachers, a lesson can be retargeted to a medical domain in which the detector has not been validated; deployments should therefore make the provenance and the validated scope of the model visible to learners, so that an automated label is never mistaken for a diagnosis.

THE QUESTIONNAIRE

The **pre-questionnaire** collected: age bracket; gender; field of study or work; whether the participant had used VR applications before (none / some); how often they use VR (never / rarely / occasionally / often); self-rated experience with head-mounted displays (five-point scale); prior use of VR or mobile applications for medical education (never / mobile applications only / both); and self-rated familiarity with brain anatomy and brain tumours (five-point scale).

The **post-questionnaire** asked one experience item per module, each on a five-point scale anchored at 1 (poor / very difficult) and 5 (excellent / very easy): the guided corridor; the main menu; the chatbot tutor; the quiz; the lecture module; the brain components scene; the brain tumours scene; 3D model exploration; the web AI detection page; and the web environment-authoring page. Two usefulness items, on the same scale, asked how useful the participant considered automated tumour detection with a visual explanation, and how useful the ability to see through the 3D tumour model. Two categorical items asked whether the VR functionalities were easy to understand (yes / somewhat / no) and whether the system would make learning tumours and medical imaging easier for students (yes / maybe / no). Three free-text items asked what should be improved, what the participant would like to see added, and anything else they wished to record.

CONCLUSION

We evaluated a configurable immersive environment for medical imaging education across two interaction contexts and two user roles, with 28 participants drawn from the domain the system serves rather than from the discipline that builds such systems. The environment is well-received. Nine of ten rated modules score at or above 4.0 on a five-point scale, 82.1% of a cohort that was two-thirds VR-naïve found its functionality easy to understand, and not one participant rejected the proposition that it would help students learn. The exceptions are the informative part. The single worst-rated element is the automated corridor, the only one that moves the user without asking, which is a reminder that the onboarding sequence is the worst place in an immersive system to spend the user’s agency. The authoring page, although rated well and never rated badly, drew the study’s widest band of participants who would not commit either way, which suggests that a configuration surface can be perfectly legible and still leave a domain expert unsure that authoring is theirs to do. And ratings of usefulness ran consistently above ratings of interaction quality for both AI-backed features, by small margins, we do not overclaim, but this direction is a caution to anyone who would evaluate an AI-infused interface with a single blended score. Finally, retargeting the environment from brain tumours to cardiac imaging confirmed the property the system was built for: a new medical domain was expressed entirely as authored content through the same web form the study found usable, with no code written and no redeployment. Configurability, on this evidence, is a realistic goal for immersive learning environments—and the remaining work is less about making the authoring surface simpler than about making the domain experts who face it confident that it is theirs to use.

References

- [1] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen,

- Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–13. doi:10.1145/3290605.3300233
- [2] Narges Ashtari, Andrea Bunt, Joanna McGrenere, Michael Nebeling, and Parmit K. Chilana. 2020. Creating Augmented and Virtual Reality Applications: Current Practices, Challenges, and Opportunities. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–13. doi:10.1145/3313831.3376722
- [3] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–16. doi:10.1145/3411764.3445717
- [4] Costas Boletis. 2017. The New Era of Virtual Reality Locomotion: A Systematic Literature Review of Techniques and a Proposed Typology. *Multimodal Technologies and Interaction* 1, 4 (2017), 24. doi:10.3390/mti1040024
- [5] Costas Boletis and Jarl Erik Cedergren. 2019. VR Locomotion in the New Era of Virtual Reality: An Empirical Comparison of Prevalent Techniques. *Advances in Human-Computer Interaction* 2019 (2019), 7420781. doi:10.1155/2019/7420781
- [6] Zana Bućinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21. doi:10.1145/3449287
- [7] George-Gabriel Constantinescu and Adrian Iftene. 2026. Intelligent Virtual Reality System for Decision Support in Social Anxiety Management through Progressive Exposure Scenarios. In *Proceedings of 18th International KES Conference on Intelligent Decision Technologies (KES-IDT-26)*, 17-19 June, 2026. Barcelona, Spain.
- [8] Fred D. Davis. 1989. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly* 13, 3 (1989), 319–340. doi:10.2307/249008
- [9] Alina Duca and Adrian Iftene. 2026. VR-MedTutor: A Multimodal Virtual Reality Platform with AI Agent-Assisted Brain Tumor Education. In *Computers, Software, and Applications (COMPASAC 2026)*, *Agentic AI: Innovative Software, Computing Paradigms, and Applications for Migrating from Reactive Models to Proactive Systems*, 7-10 July, 2026. IEEE, Madrid, Spain.
- [10] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut learning in Deep Neural Networks. *Nature Machine Intelligence* 2 (2020), 665–673. doi:10.1038/s42256-020-00257-z
- [11] Wayne D. Gray and Marilyn C. Salzman. 1998. Damaged Merchandise? A Review of Experiments That Compare Usability Evaluation Methods. *Human-Computer Interaction* 13, 3 (1998), 203–261. doi:10.1207/s15327051hci1303_2
- [12] Morten Hertzum, Kristin D. Hansen, and Hans H. K. Andersen. 2009. Scrutinising usability evaluation: does thinking aloud affect behaviour and mental workload? *Behaviour & Information Technology* 28, 2 (2009), 165–181. doi:10.1080/01449290701773842
- [13] Adrian Iftene and Diana Trandabăţ. 2018. Enhancing the Attractiveness of Learning through Augmented Reality. In *In Proceedings of International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2018, 3-5 September 2018*, Vol. 126. Procedia Computer Science, Belgrade, Serbia, 166–175.
- [14] Haowen Jiang, Sunitha Vimalasvaran, Jeremy King Wang, Kian Bee Lim, Sreenivasulu Reddy Mogali, and Lorainne Tudor Car. 2022. Virtual Reality in Medical Students' Education: Scoping Review. *JMIR Medical Education* 8, 1 (2022), e34860. doi:10.2196/34860
- [15] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. 2023. Ultralytics YOLO (version 8). <https://github.com/ultralytics/ultralytics>.
- [16] Henry Lieberman, Fabio Paternò, Markus Klann, and Volker Wulf. 2006. End-User Development: An Emerging Paradigm. In *End User Development*. Human-Computer Interaction Series, Vol. 9. Springer, Dordrecht, 1–8. doi:10.1007/1-4020-5386-X_1
- [17] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghahfoorian, Jeroen A. W. M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. 2017. A survey on deep learning in medical image analysis. *Medical Image Analysis* 42 (2017), 60–88. doi:10.1016/j.media.2017.07.005
- [18] Guido Makransky and Gustav B. Petersen. 2021. The Cognitive Affective Model of Immersive Learning (CAMIL): A Theoretical Research-Based Model of Learning in Immersive Virtual Reality. *Educational Psychology Review* 33, 3 (2021), 937–958. doi:10.1007/s10648-020-09586-2
- [19] Guido Makransky, Thomas S. Terkildsen, and Richard E. Mayer. 2019. Adding immersive virtual reality to a science lab simulation causes more presence but less learning. *Learning and Instruction* 60 (2019), 225–236. doi:10.1016/j.learninstruc.2017.12.007
- [20] Marjan Memik, Jan Heering, and Anthony M. Sloane. 2005. When and how to develop domain-specific languages. *Comput. Surveys* 37, 4 (2005), 316–344. doi:10.1145/1118890.1118892
- [21] Christian Moro, Zane Stromberga, Athanasios Raikos, and Allan Stirling. 2017. The effectiveness of virtual and augmented reality in health sciences and medical anatomy. *Anatomical Sciences Education* 10, 6 (2017), 549–559. doi:10.1002/ase.1696
- [22] Michael Nebeling and Maximilian Speicher. 2018. The Trouble with Augmented Reality/Virtual Reality Authoring Tools. In *2018 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. IEEE, Piscataway, NJ, USA, 333–337. doi:10.1109/ISMAR-Adjunct.2018.00098
- [23] Masoud Nickparvar. 2021. Brain Tumor MRI Dataset. Kaggle. <https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset>.
- [24] Fabio Paternò and Carmen Santoro. 2019. End-user development for personalizing applications, things, and robots. *International Journal of Human-Computer Studies* 131 (2019), 120–130. doi:10.1016/j.ijhcs.2019.06.002
- [25] Jack Pottle. 2019. Virtual reality and the transformation of medical education. *Future Healthcare Journal* 6, 3 (2019), 181–185. doi:10.7861/fhj.2019-0036
- [26] Jaziar Radianti, Tim A. Majchrzak, Jennifer Fromm, and Isabell Wohlgenannt. 2020. A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and a research agenda. *Computers & Education* 147 (2020), 103778. doi:10.1016/j.compedu.2019.103778
- [27] Lisa Rebenitsch and Charles Owen. 2016. Review on cybersickness in applications and visual displays. *Virtual Reality* 20, 2 (2016), 101–125. doi:10.1007/s10055-016-0285-9
- [28] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, Piscataway, NJ, USA, 618–626. doi:10.1109/ICCV.2017.74
- [29] Mel Slater. 2018. Immersion and the illusion of presence in virtual reality. *British Journal of Psychology* 109, 3 (2018), 431–433. doi:10.1111/bjop.12305
- [30] Viswanath Venkatesh, Michael G. Morris, Gordon B. Davis, and Fred D. Davis. 2003. User Acceptance of Information Technology: Toward a Unified View. *MIS Quarterly* 27, 3 (2003), 425–478. doi:10.2307/30036540
- [31] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-assisted Decision Making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. ACM, New York, NY, USA, 295–305. doi:10.1145/3351095.3372852