LLM-Think-Alouds: Multimodal Player Experience Analysis for VR Playtesting

Erdem Murat, Yongqi Zhang, Liuchuan Yu, Siraj Sabah, and Lap-Fai Yu

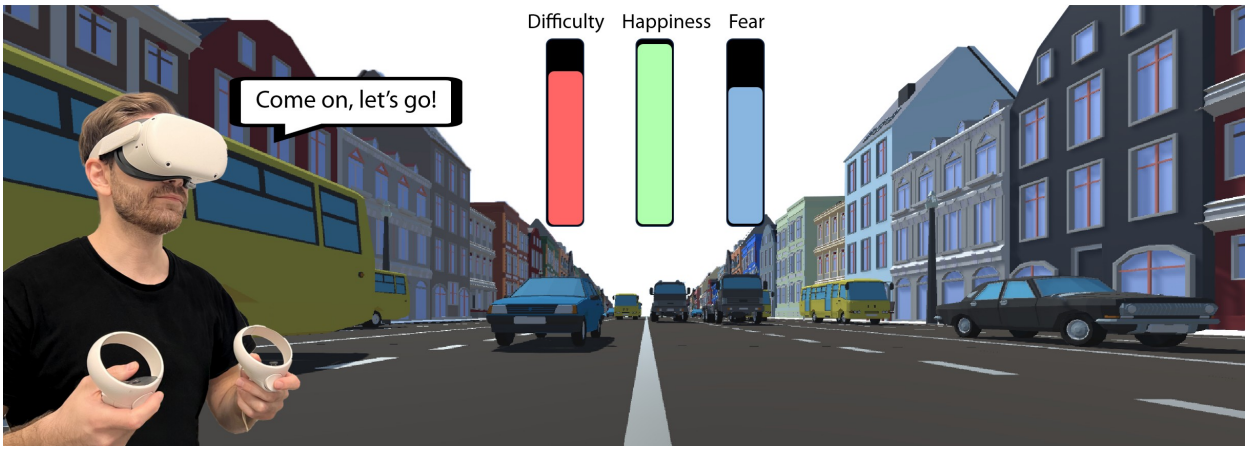

Figure 1: LLM-Think-Alouds combines synchronized first-person VR gameplay, think-aloud speech, and tone-related cues to produce temporally localized estimates of perceived difficulty and emotion for playtesting analysis.

Abstract—Virtual reality (VR) games can evoke strong and dynamic player responses due to their immersive nature. To design and refine these experiences, developers often rely on playtesting to understand how players' emotions and perceived difficulty evolve throughout gameplay. However, manually analyzing gameplay footage and think-aloud data is time-consuming and difficult to scale. We propose a large language model (LLM)-based approach for automated player experience analysis in VR games using synchronized gameplay video and player audio. Based on estimated rating trajectories, the approach supports gameplay analysis, comparison of player experiences, and deeper inspection of how experience unfolds over time. We evaluated the approach through a user study with three VR platformer games. Our findings suggest that LLMs can recover meaningful experiential signals, particularly perceived difficulty, while performance across emotional dimensions varies depending on game context and data expressiveness. Overall, the results demonstrate the potential of LLM-assisted methods as scalable tools for VR playtesting and game design analysis.

Index Terms—Virtual reality, large language models, video games, player emotion analysis, artificial intelligence

1 Introduction

Understanding player experience is a core goal of game design and evaluation [3, 10, 41]. In addition to performance or task completion, designers often care about how players *feel* during play: whether a level is difficult, surprising, exciting, or appropriately challenging. These moment-to-moment experiences influence engagement, perceived difficulty, and the overall quality of play [4, 18, 19]. Accordingly, playtesting and usability evaluation are widely used to examine how game mechanics, pacing, and presentation shape player perception and emotion [16, 17, 24].

These questions are particularly important in VR. Because VR places players inside an embodied, immersive environment, design issues such as difficulty balancing, pacing, comfort, and environmental feedback can feel especially immediate and intense [39]. This creates both an opportunity and a challenge: VR can evoke strong experiences, but poorly tuned mechanics can also amplify frustration, confusion, or discomfort. Prior work has noted that VR titles often face criticism related to gameplay design, balancing, and overall player experience, in part because VR-specific design knowledge is still maturing [14, 22, 32]. For this reason, tools that help designers understand *how players experience VR gameplay* are especially valuable.

A common way to study player experience is the *think-aloud* method, in which players verbalize their thoughts while playing [21, 25, 37]. Think-aloud data can reveal confusion, expectations, strategy shifts, and emotionally salient moments that may not be obvious from gameplay footage alone [8]. However, extracting useful insight from such sessions is labor intensive. Researchers or designers must review gameplay recordings, align player speech with events, interpret tone and context, and then summarize the experience across time. This manual process does not scale well, particularly when rapid iteration is needed during playtesting.

Recent advances in foundation models suggest a possible alternative. Multimodal models can process and reason over language together with visual context, making them a promising candidate for analyzing think-aloud gameplay sessions [36]. At the same time, their use for player-experience analysis should be treated cautiously. These models are not transparent psychological instruments, and it cannot be assumed a priori that they reliably infer player emotion from gameplay and speech alone. Rather than taking such capability for granted, we study whether a multimodal model can serve as a useful *designer-facing analysis tool* for summarizing think-aloud VR playtesting data.

In this work, we present and evaluate a domain-specific multimodal VR playtesting workflow built on existing foundation-model capabilities rather than a new model architecture. The workflow com-

• *Erdem Murat, Yongqi Zhang, Liuchuan Yu, Siraj Sabah, and Lap-Fai Yu are with George Mason University. E-mail: {emurat | yzhang59 | lyu20 | ssabah | craigyu}@gmu.edu.*

Manuscript received xx xxx. 202x; accepted xx xxx. 202x. Date of Publication xx xxx. 202x; date of current version xx xxx. 202x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.202x.xxxxxxx

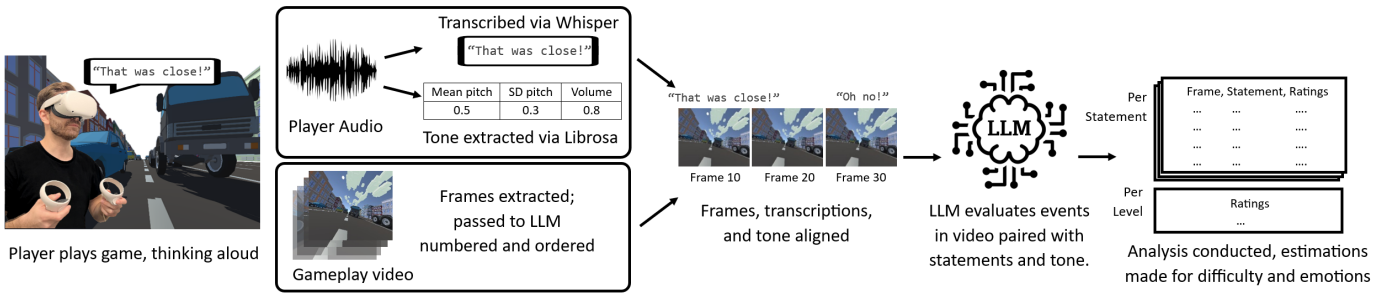

Figure 2: Overview of the proposed pipeline. During VR playtesting, think-aloud audio and first-person gameplay footage are recorded. Speech is transcribed, tone-related cues are extracted, and the resulting statements are temporally aligned with gameplay frames. A multimodal foundation model then produces structured per-statement and per-level estimates of perceived difficulty and emotion.

bines player point-of-view gameplay, think-aloud transcripts, and prosodic cues to estimate perceived difficulty and Ekman-style emotions, while also supporting designer-facing visualization and comparison of playtesting sessions over time.

We investigate whether synchronized multimodal VR playtesting data contains meaningful experiential signals that can support playtesting analysis. We do not claim superiority over transcript-only, video-only, or other analysis pipelines. Participant self-report remains essential; we frame the workflow as a comparative-analysis aid that complements feedback by providing temporally localized representations for inspection and navigation.

To guide this investigation, we structure the paper around three research questions:

- **RQ1:** To what extent do model-generated *per-level* estimates align with players' self-reported ratings of perceived difficulty and emotion across VR games?
- **RQ2:** Is alignment stronger for constructs that are more clearly elicited by the game design and more explicitly expressed in player think-aloud speech?
- **RQ3:** Can model-generated *per-statement* estimates support interpretable visualizations for designer-facing analysis of gameplay experience?

We evaluate the approach through a user study spanning three VR platformer games. Using recorded gameplay, think-aloud audio, and post-play self-reports, we examine where the model shows meaningful agreement with player ratings, where it struggles, and how its outputs can be used to visualize and compare gameplay experience. Our results provide a proof of concept for automated analysis of think-aloud VR playtesting data, while also highlighting important limitations, including uneven performance across emotional categories and the need for careful interpretation of sparse or low-variance signals.

In summary, this paper makes the following contributions:

- We introduce a multimodal foundation-model pipeline for analyzing VR think-aloud playtesting sessions via synchronized gameplay, speech transcripts, and tone-related audio cues.
- We present a designer-facing method for estimating, visualizing, and comparing perceived difficulty and emotion trajectories across gameplay using interpolation and temporal alignment.
- We provide a user-study evaluation across three VR games that examines when model-generated estimates align with player self-reports and identifies important limitations of this approach for emotion analysis in VR playtesting.

2 Related Work

2.1 Playtesting and Think-Aloud in Game Research

Playtesting is a core part of game user research because it helps designers assess whether a game is producing the intended player experience and identify where revisions are needed [9, 11, 16]. Among playtesting methods, think-aloud is especially valuable because it captures players' moment-to-moment interpretations, expectations, and frustrations during play [12, 27]. Compared with post-play questionnaires or external observation alone, think-aloud can reveal confusion, strategy shifts, perceived fairness, and emotionally salient moments as they occur [8, 37]. Prior work has also paired think-aloud with physiological data and gameplay recordings, highlighting both the richness of such data and the substantial manual effort required to align and interpret it [37]. HieraVisVR complements these efforts with a hierarchical visual-analytics workflow for motion-centric VR playtesting [42]. Our work builds on this literature by treating think-aloud as a source of multimodal feedback that can be systematically analyzed during VR playtesting.

2.2 Automated Emotion and Player-State Inference

Prior work has explored automated ways of inferring player emotion and experience from gameplay-related signals, often using physiological sensing, facial-expression analysis, or behavioral traces [33, 35, 40]. These approaches show that player experience can be studied automatically, but they often depend on specialized sensing setups, feature engineering, or signals that may be difficult to capture reliably in practical playtesting settings. In VR, this challenge is amplified because gameplay experience is shaped by embodied and design-specific factors such as locomotion, immersion, comfort, and control, which can substantially affect challenge, engagement, and affect [2, 15, 30, 38]. Our work complements prior sensor and behavior-based methods by examining whether synchronized gameplay footage, think-aloud speech, and tone-related audio cues can serve as multimodal evidence for estimating perceived difficulty and emotion in a designer-facing VR playtesting setting.

2.3 Foundation Models for Game Analysis & Evaluation

Foundation models have recently been used in game research for tasks such as content generation, agent behavior, and evaluation [36]. Some prior work has explored these models as evaluative or assistive tools, for example by assessing player responses in serious games or supporting structured design tasks such as reward design [1, 20]. However, existing work has only begun to explore their role in affective game interactions and does not directly address multimodal analysis of think-aloud VR playtesting sessions [23, 36]. In contrast, we investigate whether a multimodal foundation model can integrate synchronized gameplay visuals, think-aloud speech, and tone-related audio cues to help structure and interpret VR playtesting data, and we evaluate this use case against player self-reports of perceived difficulty and emotion.

3 Overview

Our proposed pipeline overview for analyzing think-aloud VR playtesting sessions is summarized in Figure 2. During gameplay, players verbalize their thoughts using the think-aloud protocol while their first-person gameplay video and audio are recorded. We then transform these recordings into synchronized multimodal inputs that can be analyzed in a structured way.

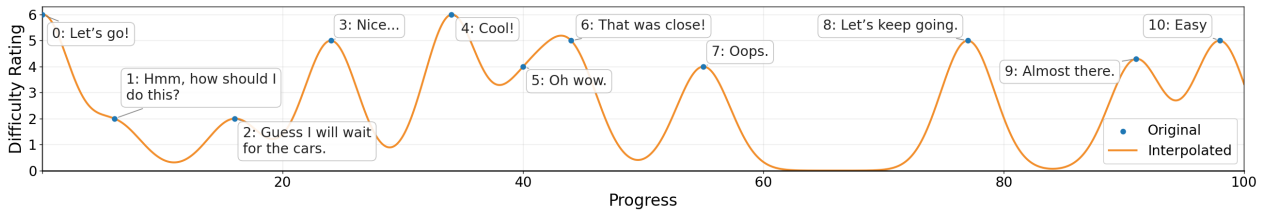

Figure 3: Difficulty estimates for a player over game progress. The blue dots are difficulty ratings estimated by the LLM from the player's statements, which are annotated at the corresponding samples. The orange curve shows the RBF-interpolated difficulty trajectory.

First, the think-aloud audio is transcribed using Whisper [28], and tone-related cues are extracted using Librosa [26]. We define a *statement* as a sentence or short coherent thought expressed during gameplay. Each statement is then temporally aligned with the corresponding gameplay frames and audio-derived cues, producing grounded segments that pair what the player said, how they said it, and what was happening in the game at that moment.

These aligned visual, linguistic, and audio-derived cues are then provided to a multimodal foundation model, which outputs estimated ratings for perceived difficulty and Ekman-style emotions. The model produces two forms of output. *Per-statement* estimates local gameplay moments and support fine-grained analysis throughout a session. *Per-level* estimates summarize the overall experience of a level and are later compared against participant self-reports.

To support designer-facing analysis, the per-statement outputs are also used to construct continuous trajectories of perceived difficulty and emotion over time. These trajectories are visualized through interpolation and compared across sessions using temporal alignment methods. Implementation details are provided in Section 4.

4 Technical Approach

Our technical approach consists of two components. First, a multimodal model maps aligned gameplay segments to structured estimates of perceived difficulty and discrete emotions. Second, the per-statement estimates are post-processed for visualization and exploratory trajectory comparison. This creates a deliberate asymmetry among the research questions: RQ1 and RQ2 evaluate per-level estimates against participant self-reports, whereas RQ3 examines whether per-statement estimates can support designer-facing visualization and comparative-analysis workflows. The interpolation and DTW views are therefore exploratory interpretive tools, not independently validated measurements of continuous "true" emotion.

4.1 Multimodal Model Inference

4.1.1 Prompting and Input Representation

We implemented inference through a Python pipeline using the ChatGPT API and GPT-4o-2024-08-06 [29], selected for its multimodal capabilities. All samples used the same fixed prompt and structured output schema, with only the game context and synchronized sample content varying; no fine-tuning was performed. The model was instructed to assess VR player experience and output estimates of perceived difficulty and Ekman-style emotions.

Each estimate was generated from three synchronized inputs: transcript content, gameplay frames, and prosodic descriptors. Using Librosa [26], we extracted mean pitch, pitch standard deviation, and volume, normalized each feature to a 0–1 scale, and included them in the prompt alongside the aligned statement and gameplay frames. The statement served as the primary evidence, while the model was also instructed to consider visible gameplay context.

Statements were segmented from sentence boundaries and short pauses in speech, producing brief verbalizations aligned across gameplay, audio, and transcription streams. After pipeline setup, transcription, prosodic-feature extraction, and timestamp alignment were fully automated and required no manual curation.

4.1.2 Structured Output Schema

The model was instructed to return machine-readable outputs so that the results could be parsed and stored automatically. For each analyzed statement, it produced ratings on a 0–10 scale for perceived difficulty and six Ekman-style emotion dimensions: happiness, surprise, fear, sadness, disgust, and anger [13]. It was also instructed to produce an overall summary for the level or gameplay segment under analysis. We selected Ekman-style emotion categories for their simplicity and interpretability as a structured output space for this exploratory study. However, they are not intended to fully capture the richness of VR gameplay experience.

In our formulation, these outputs are used at two granularities. *Per-statement* estimates describe local gameplay moments and support fine-grained visualization of perceived difficulty and emotion over time. *Per-level* estimates summarize the overall experience of a level and are later compared directly against participant self-reports. This distinction is important because the quantitative evaluation is conducted at the same granularity as the available human ratings.

To encourage stable outputs, we used a decoding temperature of 0.1. The prompt also constrained the output format to reduce extra explanatory text and improve consistency across runs.

4.1.3 Development-Time Sanity Checks

We conducted development-time sanity checks to verify correct parsing, temporal alignment, and structured-output generation, and to confirm that the model appeared responsive to the supplied statement, visual, and prosodic inputs. We also performed simple controlled tests in which linguistic and visual content were held constant while prosodic cues varied. These checks were used solely for pipeline debugging and prompt refinement and are not treated as formal validation; evaluation against participant self-reports is reported in Section 6.

4.2 Emotion Rating Interpolation

The LLM produced emotion ratings only for frames with player statements. To produce an emotion trajectory throughout the gameplay (i.e. across all the frames), we interpolate the LLM-produced emotion ratings at frames $\{x_i\}$ using radial basis functions (RBFs) [7], as follows:

$$y(x) = \sum_{i=1}^{N} w_i \varphi(||x - x_i||), \tag{1}$$

where $y(x)$ is the interpolated emotion rating at the frame x; φ is the Gaussian kernel function; and w_i is the weight of the basis function associated with the frame x_i. In total, there are N radial basis functions corresponding to the N frames with GPT-produced emotion ratings. The weights $\{w_i\}$ are computed using the linear least squares matrix method with regularization:

$$\min_{\mathbf{w}} ||\Phi \mathbf{w} - \mathbf{y}^*||^2 + \lambda ||\mathbf{w}||^2, \tag{2}$$

where $\mathbf{y}^*$ denotes the vector of known GPT-generated emotion ratings at the frames $\{x_i\}$; Φ is the matrix of basis functions formulated using the Gaussian kernel function; $\mathbf{w}$ is the vector of weights associated with the basis functions; and λ is the regularization parameter to penalize large weights, which was set as 0.1.

We selected Gaussian RBF interpolation to form smooth trajectories from sparse statement-level estimates while avoiding abrupt discontinuities between neighboring observations. The interpolated curve is an exploratory visualization aid rather than a reconstruction of a ground-truth continuous emotional signal. Figure 3 shows how the resulting trajectory can place estimated difficulty trends and relevant statements in a shared temporal view.

4.3 Emotion Trajectory Comparison

Based on dynamic time warping (DTW) [34], our approach compares players' emotion trajectories by calculating trajectory similarities. We analyze trajectories from different players completing the same levels and use player positions and emotion values to calculate costs. The DTW-based emotion trajectory similarity is given by:

$$\mathrm{DTW}(\mathbf{y},\mathbf{y}') = \min_{\pi} \left[\sum_{(i,j)\in\pi} d(y_i, y'_j)^2 \right]^{\frac{1}{2}} W_\mathrm{P} W_\mathrm{M}, \tag{3}$$

where $\mathbf{y}$ and $\mathbf{y}'$ denote the two players' emotion trajectories being compared; and π represents a warping path that aligns the two trajectories. The distance cost d for each pair of compared emotion ratings, y_i and y'_j at frames x_i and x'_j, is calculated as their absolute difference:

$$d(y_i, y'_j) = |y_i - y'_j|. \tag{4}$$

In addition, for a more reasonably aligned comparison between the emotion trajectories over the gameplays of the two players, we introduce the period constraint W_P and the moment constraint W_M, as follows.

4.3.1 Period Constraint

The period constraint is defined to favor comparison of emotions within a pre-defined progression stage. In our illustrative game of *Why did the Chicken cross the Road?*, the progression within each lane refers to a period. This setting allows our approach to favor comparing the emotions of the players as they progressed through the same lane or nearby lanes. The period cost is defined as follows:

$$W_\mathrm{P} = 1 - e^{-T_\mathrm{P}(||p_1 - p_2|| + \sigma_\mathrm{P})}, \tag{5}$$

where p_1 and p_2 are the indexes of the periods to which the interpolated emotion ratings y_i and y'_j compared belong on their respective trajectories. Essentially, this constraint prioritizes comparisons within the same or nearby period. We set constants T_P as 1.1 and σ_P as 0.2 empirically in our experiments.

4.3.2 Moment Constraint

We define the moment constraint to prioritize comparing emotions that correspond to the same specific gameplay moment. For example, in our illustrative game, we define a moment constraint to specifically compare the emotions of the players when they died in their gameplays. Similarly, other moment constraints can be defined. Suppose there are multiple moment constraints. The formulation for the k-th moment constraint is:

$$W_{\mathrm{M}_k} = 1 - e^{-T_\mathrm{M}(\delta + \sigma_\mathrm{M})}. \tag{6}$$

We set T_M as 0.1 and σ_M as 0.2 empirically. These values were set to strike a balance between constraint and precision. δ refers to the sum of the distances between the frame being compared and the moment frame of the two players:

$$\delta = |x_i - \hat{x}_k| + |x'_j - \hat{x}'_k|. \tag{7}$$

Thus, compared frames x_i and x'_j closer to their respective moment frames $\hat{x}_k$ and $\hat{x}'_k$ receive greater preference.

Multiple moment constraints could be applied. The overall moment constraint W_M is calculated as the following product:

$$W_\mathrm{M} = \prod_k W_{\mathrm{M}_k}^h. \tag{8}$$

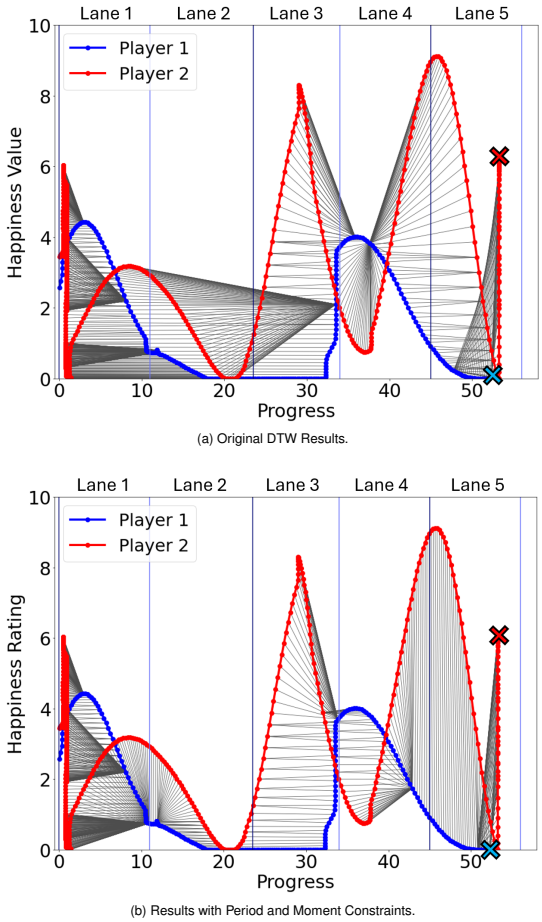

Figure 4: RBF-interpolated happiness ratings of two players with DTW matchings (grey). Player in-game death moments are marked with 'X'. (a) Original DTW. (b) Modified DTW with alignments within periods (lanes), and moments (death).

Here, each moment constraint W_{M_k} is raised to the power of h, defined as:

$$h = H(\alpha - |x_i - \hat{x}_k|)H(\alpha - |x'_j - \hat{x}'_k|), \tag{9}$$

$$H(s) = \begin{cases} 0, & \text{if } s < 0 \\ 1, & \text{otherwise.} \end{cases} \tag{10}$$

Accordingly, if a compared frame x_i lies farther than a threshold α from the moment frame $\hat{x}_k$, the Heaviside step function H returns 0, making $h = 0$ and thus $W_{\mathrm{M}_k}^h = 1$, which increases the cost in (3). In effect, this encourages DTW to align frames that are close to the specified moment frames.

Using the DTW and constraint formulations above, we compute a warping between the interpolated emotion trajectories of two players and obtain their trajectory similarity by solving (3).

Figure 4 illustrates this process for two players' happiness trajectories and shows the effect of the period and moment constraints. In both subfigures, interpolated points are paired by DTW across the two trajectories.

In Figure 4(b), the period constraint favors comparisons within the same "period," which in this example corresponds to the traffic lane occupied by the player. The lanes are shown as columns. The moment constraint is applied to the "moment of death" (getting hit by a car), corresponding to the end of each plot, so that points near the trajectory endpoints are preferentially matched.

Figure 5: Screenshots of our illustrative VR game: *Why did the Chicken cross the Road?*. (Top) First-person POV of player. (Bottom) The player starts from the left and runs across lanes with traffic to reach the end of the level.

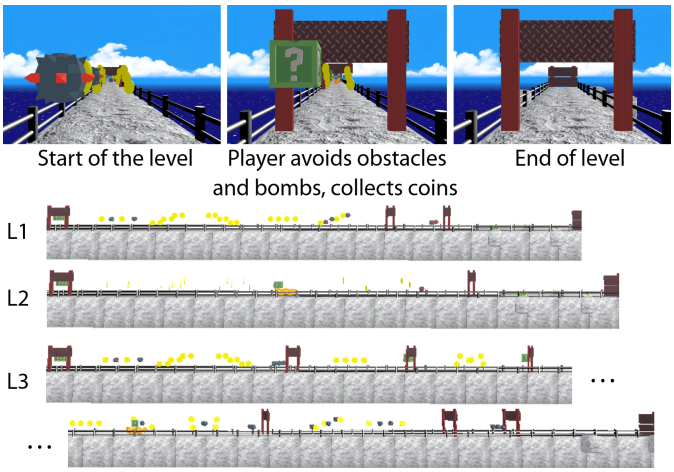

Figure 6: Top: Game 2's scenes from the player's point of view. Bottom: The three levels L1, L2 and L3 of the game.

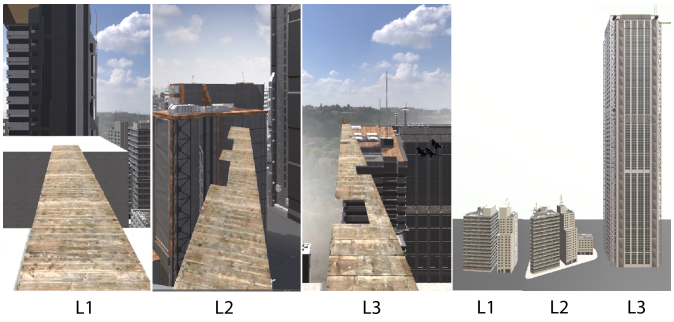

Figure 7: Left: Game 3's scenes from the player's point of view for levels L1, L2, and L3. Right: Buildings used for different levels, with the most complex level, L3, on a high-rise building.

5 User Study

To evaluate the approach, we conducted a user study using three VR platformer games with varying mechanics and experiential profiles.

VR Game 1: Why did the Chicken cross the Road?

The first game, *Why did the Chicken cross the Road?*, was the primary validation game in our study. It is a custom VR cross-the-road platformer in which players must cross multiple lanes of traffic without being hit by vehicles. Figure 5 shows example views of the game. It was inspired by cross-the-road designs such as *Frogger* and *Crossy Road*, giving it a familiar structure in which timing, hazard avoidance, and lane-by-lane progression naturally support perceived-difficulty variation.

For this study, we selected three levels from a larger parameterized design based on prior user-study data collected in earlier work. These prior ratings were used only to choose representative levels with differing perceived difficulty; they were not provided to the model during inference. Level 1 (L1) served as the easiest baseline, with adjustable parameters kept at their default settings. Level 2 (L2) increased car speed by 40%, while Level 3 (L3) increased traffic density by 47% without changing the other parameters.

In an earlier study with 57 players, these levels received average perceived difficulty ratings of 2.8, 3.7, and 6.9 on a 0–10 scale. We therefore used them as a controlled test case with an expected difficulty gradient, allowing us to assess if the model could recover player-specific relative difficulty patterns across levels.

VR Game 2: Test your Reflexes

The second game, *Test your Reflexes*, was designed to provide a less stressful but more physically demanding experience than Game 1. Figure 6 shows example views of the game. The game was influenced by movement-based Kinect-style exergames, which helped motivate its emphasis on full-body motion, physically demanding interaction, and event-driven obstacle avoidance.

Levels are organized into sequential *chunks*, each containing a specific arrangement of objects and hazards. While the forward progression speed is fixed, players can move side-to-side and up-and-down to position themselves relative to objects on the path. This structure creates repeated moments of collision avoidance, collection, and reaction to unexpected events, making the game useful for examining how the model behaves in a more movement-intensive and event-driven setting than Game 1.

From an analysis perspective, Game 2 also provides meaningful anchors for trajectory comparison. Events such as obstacle hits, coin collections, and bomb hits can serve as salient moments for alignment, while the chunk-based layout offers a natural notion of progression for comparing different parts of a level.

VR Game 3: Test the Heights

The third game, *Test the Heights*, was designed to evoke stronger fear and surprise than the other two games. Inspired by height-based VR experiences such as *Ritchie's Plank*, it places the player on a narrow plank at the edge of a tall building and requires them to walk to the end. Figure 7 shows example views of the game.

To create levels with differing experiential intensity, we manipulated several factors, including building height, plank length, plank layout, plank instability, gaps in the walkway, and crow attacks. Each level incorporated more of these difficulty-inducing features than the previous one. This made Game 3 a useful test case for examining whether the model could capture fear-related and event-driven reactions more clearly than in the other games.

From an analysis perspective, Game 3 also offers natural anchors for trajectory comparison. Because its levels are relatively short and organized around salient events, moment-based constraints are more appropriate than coarse progression-based constraints alone. Examples include stepping onto a shaky plank, crossing a gap, or encountering a crow attack. These events provide interpretable reference points for comparing how different players responded to the same fear-inducing design elements.

Table 1: Linear mixed-effects regression results relating model estimates to participant self-reports, reported as p-value and fixed-effect slope (β). A positive and statistically significant slope indicates that higher model estimates are associated with higher participant ratings after accounting for repeated measures within participants. Sparse low-variance dimensions should be interpreted in conjunction with Table 2. **Bold** indicates positive and statistically significant association ($p < 0.05$, $\beta > 0$). NA indicates that an association estimate was not available.

	Difficulty	Happiness	Surprise	Fear	Sadness	Disgust	Anger
Game 1	< 0.001, **1.170**	**0.026, 0.410**	0.150, 0.300	**0.003, 0.440**	0.700, 0.095	0.660, -0.230	0.400, -0.310
Game 2	**0.003, 0.730**	0.490, 0.100	**0.020, 1.000**	0.580, -0.200	**0.005, 1.500**	< 0.001, **2.300**	**0.002, 1.300**
Game 3	**0.040, 0.616**	0.817, -0.058	**0.041, 1.425**	**0.025, 0.725**	**0.004, 8.254**	NA	NA

Table 2: Median participant self-report ratings for perceived difficulty and emotions (0–10 scale) across the three games. **Bold** identifies dimensions with a positive and statistically significant association in Table 1.

	Difficulty	Happiness	Surprise	Fear	Sad	Disgust	Anger
Game 1	**4**	**5**	3	**2**	0	0	1
Game 2	**4**	5	**3**	1	**0**	**0**	**2**
Game 3	**4**	5	**5**	**3**	**0**	0	0

5.1 Apparatus and Data Capture

The study games were run on a Meta Quest 2 headset connected via Link to a desktop PC with an Intel i7 CPU, 32 GB RAM, and an NVIDIA RTX 2070 GPU. All gameplay recordings and derived data were stored locally on this machine, and the Quest 2 microphone was used to record think-aloud audio. All three games ran consistently above 90 FPS, helping preserve comfort and reduce the risk of simulator sickness in VR [5].

We also developed a Unity-based replayer to reconstruct recorded sessions in 3D and overlay derived annotations such as player statements and estimated experiential signals.

5.2 Demographics

Thirty-one participants took part in our Institutional Review Board-approved study and provided informed consent prior to participation. After excluding two corrupted recordings and four participants who did not produce usable think-aloud speech, the final dataset included 25 participants recruited through undergraduate, graduate, and faculty/staff mailing lists.

Among the retained participants, 19 self-identified as male and 6 as female, with ages ranging from 18 to 31 years (mean = 21.4, SD = 3.35). Prior VR familiarity varied: 44% reported more than two years of VR experience, 36% reported none or barely any experience, and 72% reported never or rarely playing VR games (at most once per month). An additional 12% reported primarily using VR for projects or development.

5.3 Procedure

Each session lasted approximately 30–40 minutes. For each game, participants first completed a short demo level to become familiar with the mechanics and controls, then played three study levels and rated each one on a 0–10 scale for perceived difficulty, happiness, surprise, fear, sadness, disgust, and anger.

During gameplay, participants were asked to think aloud while their first-person gameplay footage and audio were recorded. Post-level ratings, audio, and first-person screenshots (captured at approximately 20 frames per second) were saved locally for analysis. We also recorded the positions and rotations of relevant in-game objects, including the player, so that sessions could be reconstructed in the Unity-based replayer.

The final dataset contained 75 game sessions from 25 participants. Because each game session comprised three study levels, the dataset yielded 225 level observations.

In addition to the per-level ratings, participants completed a post-game questionnaire about game enjoyment and perceived differences in difficulty among levels; Section 6.1 summarizes its content and findings. The complete questionnaire is available at https://form.jotform.com/241993164557063.

6 Experiments and Results

6.1 Validation Context and Manipulation Checks

Participants reported how enjoyable they found the three games. Game 3 received the highest number of "Strongly Agree" responses (36%). Game 1 was also rated positively overall, with 84% of participants selecting "Agree" or "Strongly Agree" and only 12% selecting "Disagree" or "Strongly Disagree." Game 2 received the weakest enjoyment ratings, with only 44% of participants agreeing that it was fun to play and 36% disagreeing or strongly disagreeing.

Participants also reported how clearly they perceived level-to-level difficulty variation. Game 1 showed the clearest perceived separation across levels: 80% of participants selected "Agree" or "Strongly Agree" to the statement that there was a big difference in difficulty between levels, compared with 56% for Game 3 and 36% for Game 2. This indicates that Game 1 provided the strongest level-to-level difficulty variation, while Games 2 and 3 were experienced as more internally consistent.

These responses help contextualize the alignment results that follow. Because Game 1 exhibited the clearest perceived difficulty differences across levels, it provides the strongest test of whether the model can recover relative difficulty patterns from playtesting data.

6.2 Per-Level Estimate Alignment (RQ1)

RQ1 asks whether model-generated *per-level* estimates align with participant self-reports across the three games. Our goal is not to reproduce exact numeric ratings, but to determine whether the model can recover *relative player-specific patterns* across levels. We therefore evaluate alignment using (1) *participant-aware association* between model estimates and participant ratings, and (2) *within-player comparative agreement* across levels. Signed delta (AI − User) and mean absolute error (MAE) are reported as secondary calibration measures.

To assess participant-aware association, we fit a linear mixed-effects model (LMM) for each game and rating dimension, using participant rating as the dependent variable, model estimate as a fixed effect, and participant as a random effect [31]. This accounts for repeated measures and participant-specific rating tendencies. Table 1 reports the fixed-effect slope (β) and associated p-value. The slope indicates how model estimates track participant self-reports after accounting for repeated observations; larger positive slopes indicate stronger participant-aware association. We interpret β as an association measure rather than a correlation coefficient.

Because the intended use case is comparative analysis, we also measured *within-player comparative agreement*. For each participant and rating dimension, we computed a pairwise order score across the three played levels by checking whether the model and participant agreed on the relative ordering of each level pair. A score of 1 indicates agreement on all three comparisons, 0.67 agreement on two, 0.33 agreement on one, and 0 no agreement.

Several prompted emotion dimensions showed restricted ranges,

which is important for interpreting the results. In particular, sadness and disgust showed strong floor effects across games (Table 2), with median participant ratings of 0. Under such low-variance conditions, statistically significant association, low absolute error, or moderate ordering agreement may be easier to obtain without indicating robust affective tracking. We therefore retain these dimensions for completeness, but treat them as low-incidence secondary outcomes rather than primary evidence of model capability.

Across the three games, the clearest and most consistent result was **perceived difficulty**. Difficulty showed positive and statistically significant association in all three games (Game 1: $p < 0.001$, $\beta = 1.170$; Game 2: $p = 0.003$, $\beta = 0.730$; Game 3: $p = 0.040$, $\beta = 0.616$), and it also achieved the strongest or near-strongest within-player comparative agreement in each game. For Game 1, difficulty reached a mean pairwise order score of 0.61 (median 0.67), with 48% exact three-level order matches, near-zero average bias (mean delta $= -0.13$), and moderate numerical error (MAE $= 1.97$). Game 2 showed the strongest player-wise comparative agreement overall (mean pairwise order score $= 0.67$; 71% exact order matches), despite mild underestimation (mean delta $= -0.95$, MAE $= 2.14$). Game 3 remained comparatively robust (mean pairwise order score $= 0.60$; 80% exact order matches), with slight overestimation on average (mean delta $= 0.40$, MAE $= 2.27$). Taken together, these results indicate that the model often preserved which levels felt relatively easier or harder for individual players, even when exact numeric calibration remained imperfect.

The emotional dimensions showed a more mixed pattern. Positive and statistically significant association was observed for selected dimensions: happiness and fear in Game 1, surprise and anger-related dimensions in Game 2, and surprise and fear in Game 3 (Table 1). However, these results were generally less stable than those for difficulty when examined through comparative agreement and calibration. Happiness was consistently underestimated and showed weak player-wise ordering performance, while surprise was notably underestimated in Game 3 despite a positive LMM slope. By contrast, some dimensions, such as fear in Game 3, showed more interpretable comparative structure. Overall, the emotion results provide selective evidence of alignment but are less consistent and less convincing than the difficulty results across association, ordering, and calibration metrics.

Overall, the results support a clear answer to RQ1. The model shows meaningful *per-level* alignment with participant self-reports, with the strongest evidence for *relative perceived difficulty* and more limited, dimension-dependent evidence for discrete emotions.

6.3 When Alignment Was Stronger or Weaker (RQ2)

RQ2 asks whether alignment was stronger for constructs that were more clearly elicited by the game design and more explicitly expressed in participant think-aloud speech. The results support this interpretation. Alignment was strongest when a target dimension was central to the structure of the game and exhibited enough variation in participant ratings to make comparison meaningful.

This pattern was clearest for **difficulty**. Game 1 provided the strongest test case because its levels were intentionally selected to span a broader range of challenge, and participants themselves most clearly reported difficulty differences across levels (Section 6.1). Consistent with this, Game 1 also produced the strongest overall difficulty alignment. Difficulty remained comparatively robust in Games 2 and 3 as well, but the evidence was most convincing when relative difficulty was especially salient in the game design.

A similar pattern appears for selected emotions. In **Game 3**, **fear** showed more interpretable alignment than in the other games, consistent with the design of that experience around height, instability, and threat. In **Games 2 and 3**, **surprise** showed more positive association than in Game 1, plausibly because these games contained more visually distinct or sudden gameplay events. By contrast, **happiness** was less stable across games despite consistently moderate participant ratings, suggesting that positive affect was harder to recover reliably from local multimodal cues than dimensions such as difficulty, fear, or surprise.

Alignment also depended on the available rating range. As shown in Table 2, **sadness** and **disgust** remained at floor across all three games, with median participant ratings of 0. Under such restricted ranges, apparently favorable association, low error, or moderate ordering performance can arise without demonstrating robust affective tracking. We therefore do not treat these dimensions as strong evidence that the model successfully recovers those emotions; rather, they indicate that low-incidence dimensions are poor targets for validation in short gameplay experiences that do not meaningfully elicit them.

Taken together, these findings support a clear answer to RQ2. Alignment was stronger when the target construct was clearly evoked by the game design, sufficiently variable across levels, and expressed clearly enough in participant speech and local gameplay context to provide usable cues for the model. Alignment weakened when the target dimension was diffuse, weakly elicited, or compressed near the bottom of the rating scale.

7 Discussion

7.1 What the Model Captures Most Reliably (RQ1)

Across the three games, the clearest and most consistent signal was *perceived difficulty*. This is meaningful because difficulty was both a central design variable in the study and a dimension that participants appeared willing and able to report comparatively across levels. Game 1 provided the strongest test case: its levels were intentionally selected to span a range of challenge, and the model showed positive participant-aware association together with the strongest within-player comparative agreement for this dimension. Taken together, these results suggest that the model often preserved which levels felt relatively easier or harder for a given player, even when exact numeric calibration remained imperfect.

This distinction is important for interpreting the contribution of the system. Our goal is not to reproduce participant self-reports exactly, but to determine whether the model can recover *player-specific comparative structure* from multimodal playtesting data. From this perspective, perceived difficulty is the dimension for which the method is currently most convincing. It showed the most stable combination of positive association, player-wise ordering performance, and relatively modest bias, making it the strongest evidence that the pipeline can extract practically useful experiential signal from think-aloud VR sessions.

The results for discrete emotions were more mixed. Rather than indicating a uniform failure, they suggest that performance depended on whether an emotion was both *meaningfully elicited by the game* and *expressed in think-aloud speech*. Fear in Game 3 and surprise in Games 2–3 fit this pattern more naturally than sadness or disgust, which were largely absent from participant reports. Happiness occupied an intermediate case: it was clearly relevant to the play experience, but the model did not consistently preserve player-specific patterns, suggesting that positive affect was harder to estimate from the available multimodal cues than difficulty.

Overall, these findings support a narrower but more defensible claim than full automated emotion understanding. The proposed pipeline appears most useful for estimating *relative perceived difficulty*, and more selectively useful for emotions that are strongly tied to game design and clearly expressed in local gameplay context. We therefore interpret the method as a designer-facing comparative analysis tool rather than as a replacement for participant self-report or a definitive measure of player emotion.

7.2 Alignment Variation across Games (RQ2)

Variation across games and rating dimensions appears to reflect a combination of elicitation strength, signal expressiveness, and rating range. Difficulty was comparatively robust because it was intentionally manipulated by level design and consistently reflected in participant reports. By contrast, emotional dimensions depended more on whether the game created moments that were both experientially salient and clearly expressible during think-aloud play.

Game-specific differences help explain this pattern. Game 1 manipulated traffic speed and density, making perceived difficulty a relatively stable target; its collision-based structure also plausibly supported some signal for fear, while surprise was less central. In contrast, Games 2

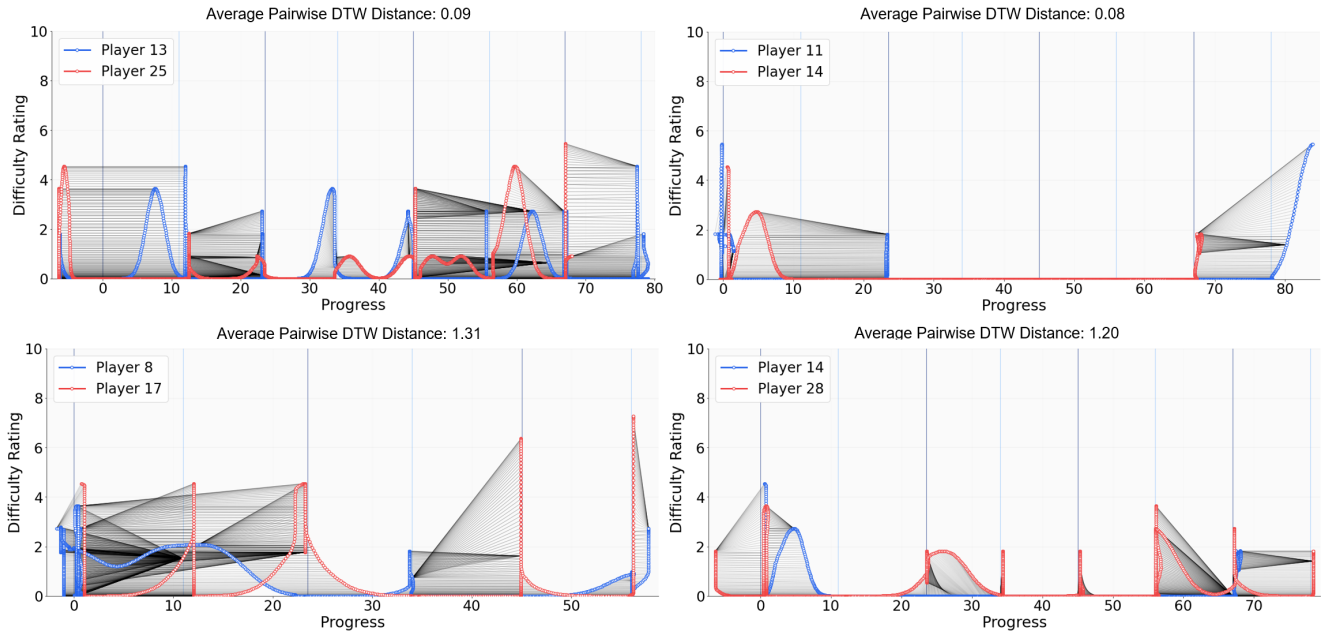

Figure 8: DTW-based comparisons of estimated difficulty trajectories for four player pairs in Level 1 of Game 1. Vertical lines indicate period (lane) boundaries, and the average pairwise DTW distance is shown above each plot, with smaller values indicating more similar trajectories. The top row shows more similar pairs (0.09, 0.08), while the bottom row shows more dissimilar pairs (1.31, 1.20). Some trajectories begin below 0 on the x-axis because the corresponding statements occurred before the first lane.

and 3 relied more on visually distinct or sudden events, making surprise a more natural target. Game 3 was also designed to be the most fear-inducing, so its stronger fear alignment is consistent with both the game design and the participant medians in Table 2. More generally, alignment was strongest when a target dimension was clearly cued by the game.

A key caveat concerns low-incidence negative emotions, especially sadness and disgust. As shown in Table 2, both exhibited strong floor effects across all three games, with median participant ratings of 0. Under such restricted ranges, favorable association, low error, or moderate ordering agreement can arise without demonstrating robust affective tracking. We therefore retain these dimensions for completeness, but do not treat them as primary evidence of model capability.

Happiness also warrants caution. Although it was relevant across all three games and participant medians were consistently moderate, the model showed weaker comparative agreement and substantial underestimation. This suggests that positive affect may be more diffuse and less directly grounded in local statement-level cues than dimensions such as difficulty, fear, or surprise.

Overall, these results suggest that model–participant alignment should be interpreted relative to what each game was designed to evoke and what the data can realistically express. A strong result for perceived difficulty does not imply equally strong performance for all emotions, and apparent success on sparse dimensions should not be overread. We therefore interpret the system not as a general emotion measurement instrument, but as a comparative aid that recovers some experiential signals more reliably than others.

7.3 Designer-Facing Trajectory Visualization (RQ3)

The trajectory visualizations are best understood as *interpretive tools* rather than additional model validation. Unlike RQ1 and RQ2, which assess agreement with self-reports, RQ3 explores whether sparse per-statement estimates can support designer-facing temporal comparison and analysis.

From a playtesting perspective, this temporal view is valuable because gameplay experience is rarely uniform across an entire session. Two players may report similar overall difficulty while arriving at that impression through very different moment-to-moment experiences. The interpolation and DTW views make such differences more visible by surfacing local spikes, sustained periods of tension, or relatively stable regions of play. In this sense, the visualization complements the per-level estimates by providing a more structured view of how an overall impression may have developed.

Figure 8 presents representative Game 1 comparisons, which we focus on because of space constraints. The constrained DTW formulation incorporates both progression structure and salient events, making comparisons more interpretable within similar regions of a level or around events such as failure. Games 2 and 3 have different gameplay structures, including the less discretely segmented experience of Game 3, so broader cross-game trajectory comparisons are an important direction for future analysis.

At the same time, these visualizations inherit the uncertainties of the underlying per-statement estimates. They should therefore be used to guide closer human inspection rather than as standalone evidence that two players truly experienced identical or different emotions at specific moments. Their role is not to replace qualitative interpretation, but to make large collections of think-aloud playtesting data easier to navigate and compare.

7.4 Practical Implications for Playtesting Workflows

The replayer interface illustrates the intended practical role of the system as a *playtesting aid*. By overlaying estimated experiential signals and statement annotations onto replayed gameplay, it can help designers inspect where different players appear to respond differently to the same content. This is particularly useful for identifying candidate moments for closer review, such as segments associated with elevated estimated difficulty, abrupt affective changes, or recurring verbal reactions across participants. Figure 9 shows an example of this idea, where player paths, statements, and estimated happiness are visualized together to highlight how different players can experience the same level differently.

One practical advantage of this approach is that it connects abstract estimates back to concrete gameplay context. Rather than treating an estimated emotion value as meaningful in isolation, the replayer allows designers to examine the surrounding gameplay, spatial position, and player commentary together. This makes the outputs easier to interpret and more useful for iterative design tasks such as balancing difficulty, refining level pacing, or investigating where a mechanic may

be producing confusion or frustration.

Beyond interpretability, a key question is whether the pipeline is practical for real playtesting. In our experiments, about 60 minutes of gameplay and think-aloud audio was processed in roughly 80 minutes. This is not real-time and still requires human review, but it shows that multimodal analysis of think-aloud VR sessions can be partially automated at a workable offline pace.

From a workflow perspective, the strongest value of the approach is not that it replaces participant self-report or expert judgment, but that it can reduce some of the manual effort involved in reviewing recordings, organizing observations, and identifying sessions or moments that warrant closer inspection. In this sense, the system is best understood as a triage and comparative-analysis aid: it helps surface candidate patterns in perceived difficulty and emotion that a designer can then inspect more carefully through the corresponding gameplay, statements, and replay context.

This practical value should still be interpreted with appropriate caution. The current pipeline depends on cloud-based multimodal model inference, which introduces latency and may not be suitable for time-sensitive or privacy-constrained settings. Moreover, the system is most convincing for relative perceived difficulty and only selectively useful for game-congruent emotions, so its outputs are better treated as structured cues for designer attention than as authoritative labels of player psychology. Even with these caveats, the results suggest that the approach may reduce some of the repetitive analysis burden in think-aloud VR playtesting.

8 Limitations and Future Work

We present this work as a proof of concept for designer-facing analysis, not as a general-purpose emotion measurement system. The Ekman-style categories offered methodological simplicity and an interpretable structured output space, but they do not capture the full richness of VR gameplay experience. The floor effects for sadness and disgust reinforce that some discrete categories were weak fits for the studied games. More generally, alignment depends on whether a construct is meaningfully elicited by the game design and expressed in the available multimodal cues. Future work should therefore compare alternative representations, including valence–arousal models and game-specific user-experience constructs.

A first limitation concerns *validation granularity*. Our quantitative evaluation is strongest at the *per-level* scale, where model summaries can be compared directly against participant self-reports. By contrast, the per-statement outputs used for interpolation, DTW, and replay-based inspection do not have dense moment-by-moment human ground truth. These visualizations should therefore be treated as interpretive aids rather than independently validated measurements of emotion at specific moments.

A second limitation concerns *model dependence, privacy, and reproducibility*. The current pipeline relies on a proprietary cloud-based multimodal model, which introduces latency, limits transparency, and may be unsuitable for privacy-sensitive playtesting data. It also reduces exact reproducibility across model versions. Future work should examine whether similar results can be obtained with more controllable local or open-weight models.

Third, our study covers only three structured single-player VR games. While chosen for diverse experiential profiles, the method remains untested in open-ended or socially interactive settings like sandbox or multiplayer games. Extending it to these domains will likely require more flexible alignment and richer gameplay context representations.

Fourth, the approach inherits the limitations of the *think-aloud method* itself. Think-aloud can disrupt natural play [6], and participants who spoke very little were harder to analyze reliably because the model had less linguistic evidence to ground its estimates. Some states may also be weakly verbalized even when genuinely experienced, which likely contributed to weaker performance for dimensions such as happiness.

Future work should improve speed, reproducibility, and robustness and should compare the full workflow with simpler transcript-only and video-only baselines to isolate the contribution of each modality. It should also evaluate the trajectory and replay views formally with designers or playtesting practitioners, extend trajectory analysis across Games 2 and 3 and less discretely segmented experiences, and investigate aggregate or cumulative summaries across players. Such summaries could help identify gameplay segments that are consistently difficult or emotionally salient. Complementary signals such as movement traces or physiological data may further strengthen this comparative aid, whose role remains to support rather than replace self-report and expert interpretation.

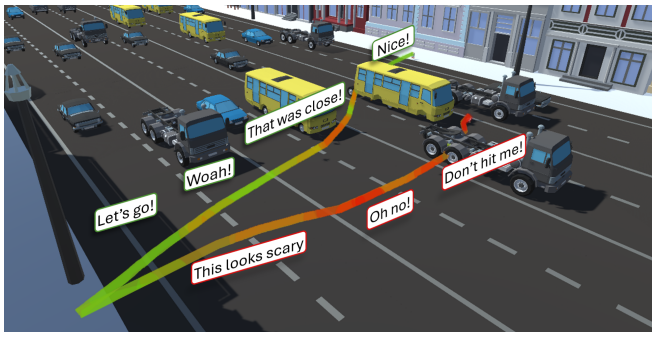

Figure 9: Player paths, statements, and changes in estimated happiness. Path color changes from red for lower estimated happiness to green for higher estimated happiness, while statement annotations provide corresponding verbal context.

9 Conclusion

We introduced an LLM-based framework for analyzing player experience in VR playtesting from synchronized gameplay footage and think-aloud speech. The framework supports both per-level estimation and DTW-based temporal comparison, enabling automated analysis of player ratings and experience trajectories. Evaluations across three VR games showed that the approach is particularly effective for perceived difficulty, while performance across emotional dimensions depends on how strongly those experiences are elicited and expressed during play. Overall, our findings demonstrate the promise of LLMs as a scalable comparative tool for VR playtesting, supporting more efficient inspection of player experience during game design.

Acknowledgments

This project was supported by National Science Foundation grants (award numbers 2418236, 2430673, and 2310210).

References

[1] I.-C. Baek, T.-H. Park, J.-H. Noh, C.-M. Bae, and K.-J. Kim. Chatpcg: Large language model-driven reward design for procedural content generation. In *2024 IEEE Conference on Games (CoG)*, pp. 1–4, 2024. doi: 10.1109/CoG60054.2024.10645619 2

[2] Z. Bai and A. F. Blackwell. Analytic review of usability evaluation in ismar. *Interact. Comput.*, 24(6):450–460, 2012. doi: 10.1016/j.intcom.2012.07.004 2

[3] R. Bernhaupt. *Evaluating User Experience in Games: Concepts and Methods*. Springer Publishing Company, Incorporated, 1st ed., 2010. 1

[4] R. Bernhaupt, A. Boldt, T. Mirlacher, D. Wilfinger, and M. Tscheligi. Using emotion in games: emotional flowers. In *Proc. ACE*, pp. 41–48, 2007. doi: 10.1145/1255047.1255056 1

[5] K. Boos, D. Chu, and E. Cuervo. Flashback: Immersive virtual reality on mobile devices via rendering memoization. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services*, pp. 291–304. Association for Computing Machinery, 2016. doi: 10.1145/2906388.2906418 6

[6] T. Boren. Thinking aloud: Reconciling theory and practice. *Professional Communication, IEEE Transactions on*, 43:261 – 278, 10 2000. doi: 10.1109/47.867942 9

[7] M. D. Buhmann. Radial basis functions. *Acta Numerica*, 9:1–38, 2000. doi: 10.1017/S0962492900000015 3

[8] S. Ching Yang. Reconceptualizing think-aloud methodology: refining the encoding and categorizing techniques via contextualized perspectives. *Computers in Human Behavior*, 19(1):95–115, 2003. doi: 10.1016/S0747-5632(02)00011-0 1, 2

[9] J. O. Choi, J. Forlizzi, M. Christel, R. Moeller, M. Bates, and J. Hammer. Playtesting with a purpose. In *Proc. CHI PLAY*, pp. 254–265, 2016. doi: 10.1145/2967934.2968103 2

[10] M. Croissant, G. Schofield, and C. McCall. Emotion design for video games: A framework for affective interactivity. *ACM Games*, 1(3), art. no. 19, 24 pp., oct 2023. doi: 10.1145/3624537 1

[11] H. Desurvire and M. S. El-Nasr. Methods for game user research: Studying player behavior to enhance game design. *IEEE Comput. Graph. Appl.*, 33(4):82–87, July 2013. doi: 10.1109/MCG.2013.61 2

[12] H. Desurvire and C. Wiberg. Game usability heuristics (play) for evaluating and designing better games: The next iteration. In A. A. Ozok and P. Zaphiris, eds., *Online Communities and Social Computing*, pp. 557–566. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009. 2

[13] P. Ekman and R. J. Davidson. *The Nature of emotion : fundamental questions / edited by Paul Ekman, Richard J. Davidson.* Oxford University Press, New York, 1994. 3

[14] R. Epp, D. Lin, and C.-P. Bezemer. An empirical study of trends of popular virtual reality games and their complaints. *IEEE Transactions on Games*, 13(3):275–286, 2021. doi: 10.1109/TG.2021.3057288 1

[15] J. Frommel, S. Sonntag, and M. Weber. Effects of controller-based locomotion on player experience in a virtual reality exploration game. In *Proceedings of the 12th International Conference on the Foundations of Digital Games*, FDG '17, art. no. 30, 6 pp. Association for Computing Machinery, New York, NY, USA, 2017. doi: 10.1145/3102071.3102082 2

[16] T. Fullerton. *Game Design Workshop: A Playcentric Approach to Creating Innovative Games*. A K Peters/CRC Press, New York, 5th ed., 2024. doi: 10.1201/9781003460268 1, 2

[17] B. Giżycka and G. J. Nalepa. Emotion in models meets emotion in design: Building true affective games. In *2018 IEEE Games, Entertainment, Media Conference (GEM)*, pp. 1–5, 2018. doi: 10.1109/GEM.2018.8516439 1

[18] G. Gonçalves, É. Mourão, L. Torok, D. Trevisan, E. Clua, and A. Montenegro. Understanding user experience with game controllers: A case study with an adaptive smart controller and a traditional gamepad. In N. Munekata, I. Kunita, and J. Hoshino, eds., *Entertainment Computing – ICEC 2017*, pp. 59–71. Springer International Publishing, Cham, 2017. 1

[19] C. Hodent. Cognitive psychology applied to user experience in video games. In N. Lee, ed., *Encyclopedia of Computer Graphics and Games*, pp. 1–6. Springer International Publishing, Cham, 2016. doi: 10.1007/978-3-319-08234-9_53-1 1

[20] A. Isaza-Giraldo, P. Bala, P. F. Campos, and L. Pereira. Prompt-gaming: A pilot study on llm-evaluating agent in a meaningful energy game. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI EA '24, art. no. 272, 12 pp. Association for Computing Machinery, New York, NY, USA, 2024. doi: 10.1145/3613905.3650774 2

[21] T. Knoll. 189the think-aloud protocol. In *Games User Research*. Oxford University Press, 01 2018. doi: 10.1093/oso/9780198794844.003.0012 1

[22] L. Lee, T. Braud, P. Zhou, L. Wang, D. Xu, Z. Lin et al. All one needs to know about metaverse: A complete survey on technological singularity, virtual ecosystem, and research agenda. *CoRR*, abs/2110.05352, 2021. 1

[23] A. Marincioni, M. Miltiadous, K. Zacharia, R. Heemskerk, G. Doukeris, M. Preuss et al. The effect of llm-based npc emotional states on player emotions: An analysis of interactive game play. In *2024 IEEE Conference on Games (CoG)*, pp. 1–6, 2024. doi: 10.1109/CoG60054.2024.10645631 2

[24] L. A. Martínez-Tejada, A. Puertas González, N. Yoshimura, and Y. Koike. Videogame design as a elicit tool for emotion recognition experiments. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 4320–4326, 2020. doi: 10.1109/SMC42975.2020.9283321 1

[25] J. May. Youtube gamers and think-aloud protocols: Introducing usability testing. *IEEE Transactions on Professional Communication*, 62(1):94–103, 2019. doi: 10.1109/TPC.2018.2867130 1

[26] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg et al. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, vol. 8, 2015. 3

[27] J. Nielsen. *Usability Engineering*. Morgan Kaufmann, San Francisco, CA, 1993. 2

[28] OpenAI. Whisper: A general-purpose speech recognition model. `https://openai.com/research/whisper`, 2022. Accessed: 08/30/2024. 3

[29] OpenAI. Gpt-4o: Openai's multimodal and reasoning-enhanced model, 2024. Accessed: 2024-08-29. 3

[30] X. Peng, J. Huang, A. Denisova, H. Chen, F. Tian, and H. Wang. A palette of deepened emotions: Exploring emotional challenge in virtual reality games. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pp. 1–13. Association for Computing Machinery, New York, NY, USA, 2020. doi: 10.1145/3313831.3376221 2

[31] J. C. Pinheiro and D. M. Bates. *Mixed-Effects Models in S and S-PLUS*. Springer, New York, NY, 2000. doi: 10.1007/b98882 6

[32] A. M. Research. Global virtual reality content creation market research report: By content type (videos, 360-degree photos and games), by component (software and services), by end user (real estate, travel & hospitality, media & entertainment, healthcare, gaming, automotive and others), by region (north america, europe, asia-pacific, middle east & africa, south america) - forecast till 2030. *Allied Market Research*, 2022. 1

[33] S. Roohi, E. D. Mekler, M. Tavast, T. Blomqvist, and P. Hämäläinen. Recognizing emotional expression in game streams. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, CHI PLAY '19, pp. 301–311. Association for Computing Machinery, New York, NY, USA, 2019. doi: 10.1145/3311350.3347197 2

[34] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, 1978. doi: 10.1109/TASSP.1978.1163055 4

[35] R. Somarathna, D. S. Elvitigala, Y. Yan, A. J. Quigley, and G. Mohammadi. Exploring user engagement in immersive virtual reality games through multimodal body movements. In *Proceedings of the 29th ACM Symposium on Virtual Reality Software and Technology*, VRST '23, art. no. 3, 8 pp. Association for Computing Machinery, New York, NY, USA, 2023. doi: 10.1145/3611659.3615687 2

[36] P. Sweetser. Large language models and video games: A preliminary scoping review. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces*, CUI '24, art. no. 45, 8 pp. Association for Computing Machinery, New York, NY, USA, 2024. doi: 10.1145/3640794.3665582 1, 2

[37] C. T. Tan, T. W. Leong, and S. Shen. Combining think-aloud and physiological data to understand video game experiences. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pp. 381–390. Association for Computing Machinery, New York, NY, USA, 2014. doi: 10.1145/2556288.2557326 1, 2

[38] C. T. Tan, T. W. Leong, S. Shen, C. Dubravs, and C. Si. Exploring gameplay experiences on the oculus rift. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play*, CHI PLAY '15, pp. 253–263. Association for Computing Machinery, New York, NY, USA, 2015. doi: 10.1145/2793107.2793117 2

[39] C. J. Wilson, A. Soranzo, et al. The use of virtual reality in psychology: A case study in visual perception. *Computational and mathematical methods in medicine*, 2015, 2015. 1

[40] W. Yang, M. Rifqi, C. Marsala, and A. Pinna. Physiological-based emotion detection and recognition in a video game context. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2018. doi: 10.1109/IJCNN.2018.8489125 2

[41] G. N. Yannakakis, K. Karpouzis, A. Paiva, and E. Hudlicka. Emotion in games. In *Proc. ACII*, pp. 497–497, 2011. 1

[42] Y. Zhang, E. Murat, L. Yu, H. Huang, M. Choi, C. Mousas et al. HieraVisVR: Hierarchical visual analytics for motion-centric VR playtesting. In *Proc. CHI*, art. no. 629, 19 pp., 2026. doi: 10.1145/3772318.3790591 2