Proceedings of CONF-MLA 2026 Symposium: Learning and Decision Making in Multi Agent Software Systems
DOI: 10.54254/2755-2721/2026.BA38767

Cross-Game Generalization and Context Dependence of Cybersickness Prediction Using Consumer-Grade HMD Motion Data

Yuanhang Hu

School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, USA
huyua@oregonstate.edu

Abstract. Cybersickness remains a persistent barrier to sustained virtual reality (VR) use. Consumer-grade head-mounted displays (HMDs) provide motion data that can support symptom prediction, but most prior models have been evaluated within a single game or data environment, leaving cross-game generalization uncertain. This study reanalyzes HMD motion trajectories and Simulator Sickness Questionnaire (SSQ) responses from 60 complete participants across four commercial VR games in the publicly available Questset dataset. Fifteen session-level motion features were extracted, symptom change was represented as ΔSSQ, and Ridge Regression and Random Forest were evaluated under within-game and cross-game conditions. Cross-game error increased most clearly when Beat Saber was the target game under both models, whereas other transfer directions showed smaller penalties. Several feature–symptom correlation directions also changed across games, although none of the 60 exploratory correlations remained significant after False Discovery Rate (FDR) correction. These results suggest that HMD-based cybersickness prediction may be context-dependent and should not be assumed to transfer automatically to unseen VR content; larger within-subject, multi-game studies are needed to confirm this pattern.

Keywords: virtual reality, cybersickness, head motion, cross-game generalization, context dependence

1. Introduction

VR is increasingly used in gaming, training, education, and healthcare, yet cybersickness remains a barrier to sustained use. Symptoms such as nausea, dizziness, eye strain, and disorientation can interrupt an experience, and their severity is influenced by visual motion, locomotion, exposure duration, and individual differences [1, 2]. Consumer-grade HMDs continuously record head position and orientation, creating a low-cost source of behavioral data for cybersickness modeling. The present analysis uses features summarized over complete game sessions and therefore evaluates offline prediction and cross-game generalization rather than real-time early warning.

Prior work has shown that HMD-related signals can support cybersickness prediction. Islam et al. combined eye- and head-tracking, while Maneuvrier et al. showed that head-rotation patterns

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.BA38767

recorded during VR can contribute to cybersickness prediction [3, 4]. More recently, head-tracking features from a consumer-grade Quest 2 were used directly for prediction [5]. However, strong performance inside one task or dataset does not establish reliability in unseen content. Different VR tasks can elicit different symptom patterns, and cross-dataset testing can expose performance losses that are hidden by within-dataset validation [6, 7]. This motivates direct evaluation of whether motion-based prediction transfers across games.

Accordingly, the study asks three linked questions: whether cross-game training increases ΔSSQ prediction error relative to within-game validation; whether the direction and strength of feature–symptom correlations remain consistent across paired games; and whether any generalization gap depends on the target game or model. This study reanalyzes HMD motion trajectories and SSQ scores from the Questset dataset across four commercial VR games [8]. Ridge Regression and Random Forest are compared under within-game and cross-game conditions, and motion feature–symptom relationships are examined across games. The observed asymmetry in transfer error and the instability of several feature–symptom correlations suggest context dependence and motivate explicit cross-game validation rather than assuming that within-game performance will generalize automatically.

2. Literature review

2.1. Motion sickness measurement and contextual influences

The SSQ remains a standard subjective measure for simulator- and VR-related symptoms [9]. Reviews show that cybersickness is shaped by multiple interacting factors rather than a single cause, including visual motion, locomotion, exposure duration, display conditions, and user characteristics. These factors also influence how users move inside a VR task. Consequently, game content can affect both the distribution of HMD motion features and the distribution of symptom outcomes, making content a relevant consideration when evaluating predictive models.

2.2. HMD-based prediction and generalization

Research on automated prediction has increasingly used HMD-internal signals because they avoid the cost and complexity of additional physiological hardware. Islam et al. demonstrated that eye- and head-tracking information can contribute to cybersickness prediction, while Maneuvrier et al. linked automatically analyzed head rotations to cybersickness outcomes [3, 4]. Ramaseri-Chandra et al. further showed that head-tracking-only features from a Quest 2 can be modeled using conventional machine-learning methods [5]. The remaining issue is external validity. Jasper et al. showed that both individual and task characteristics account for substantial variability in cybersickness, Wang et al. reported task-dependent cybersickness responses, and Yang et al. found a clear drop between single-dataset and cross-dataset performance in VR motion-sickness detection [10]. Questset is particularly suitable for this problem because it contains paired commercial games, HMD trajectories, and repeated SSQ measurements [8]. The present study therefore treats cross-game transfer as the main test of generalization rather than using within-game accuracy alone. A model can perform well under familiar content while relying on motion patterns specific to that content, making cross-game evaluation a stronger test of external validity.

DOI: 10.54254/2755-2721/2026.BA38767

3. Methods

3.1. Dataset, participants, and game assignment

Questset was collected at the University of Padova between June 2023 and January 2024 and includes HMD/controller motion, network traffic, participant surveys, and SSQ responses [8]. The original study recruited 70 adults; 60 completed both assigned games, while 10 withdrew because of pronounced cybersickness. The primary analysis uses the 60 complete participants so that each person has paired sessions. Participants were divided into two groups of 30: Group 1 played Beat Saber and Cooking Simulator VR, whereas Group 2 played Medal of Honor: Above and Beyond and Forklift Simulator 2019. Game order was counterbalanced within each group.

Table 1 summarizes this paired-game design. Cross-game tests are restricted to Beat↔Cooking within Group 1 and Medal↔Forklift within Group 2; no direct transfer is estimated between the two participant groups. The original experiment used a Meta Quest 2 connected to Windows 11, and OculusMonitor recorded HMD position and quaternion orientation at approximately 60 Hz [8].

Table 1. Participant groups and paired VR games in Questset

Group	N	Fast game	Slow game	Order	Transfer range
Group 1	30	Beat Saber	Cooking Simulator VR	fast→slow / slow→fast	Beat ↔ Cooking
Group 2	30	Medal of Honor: Above and Beyond	Forklift Simulator 2019	fast→slow / slow→fast	Medal ↔ Forklift

3.2. Procedure, outcome construction, and feature engineering

Each session began with an SSQ baseline, followed by game training and 10–20 minutes of recorded gameplay. SSQ2 was collected after Game 1, participants rested for 15 minutes and completed SSQ3, and SSQ4 followed Game 2 [8]. Figure 1 shows the complete sequence and the two outcome windows. The first-session change score was $\Delta SSQ_1 = SSQ2 - SSQ1$, and the second-session score was $\Delta SSQ_2 = SSQ4 - SSQ3$. SSQ items were coded from 0 (None) to 3 (Severe), and total scores followed the weighted SSQ procedure [9]. A larger ΔSSQ indicates a larger post-game increase in symptoms.

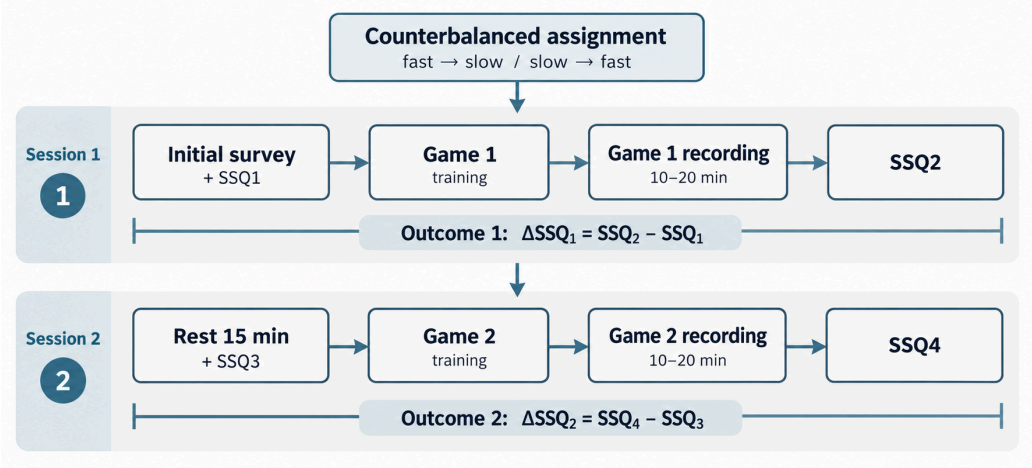


Figure 1. Questset experimental procedure and ΔSSQ outcome windows [8]

DOI: 10.54254/2755-2721/2026.BA38767

Trajectory preprocessing retained timestamps, three-dimensional HMD position, and quaternion orientation. Missing records and abnormal timestamp gaps were removed. Linear speed was calculated from frame-to-frame displacement over time; angular speed from relative quaternion rotation; and linear and angular acceleration from changes in the corresponding velocities. Frame-level values were then summarized into 15 session-level features. As detailed in Table 2, the feature set covers central tendency, variability, high-percentile motion intensity, positional dispersion, and session duration. Each usable session therefore produced one Participant × Game × 15 Features record paired with one ΔSSQ value.

Table 2. Fifteen session-level HMD motion features used in the analysis

Category	Feature	Definition / calculation
Linear speed	Mean	Mean linear speed within the session
Linear speed	SD	Standard deviation of linear speed within the session
Linear speed	P95	95th percentile of linear speed within the session
Linear acceleration	Mean	Mean linear acceleration within the session
Linear acceleration	P95	95th percentile of linear acceleration within the session
Angular speed	Mean	Mean angular speed within the session
Angular speed	SD	Standard deviation of angular speed within the session
Angular speed	P95	95th percentile of angular speed within the session
Angular acceleration	Mean	Mean angular acceleration within the session
Angular acceleration	P95	95th percentile of angular acceleration within the session
Position variability	X-position SD	Standard deviation of HeadPosX
Position variability	Y-position SD	Standard deviation of HeadPosY
Position variability	Z-position SD	Standard deviation of HeadPosZ
Position variability	Overall position dispersion	Overall dispersion of 3D head position
Exposure duration	Session duration	Duration of valid HMD motion recording for the session

3.3. Models, validation, and statistical analysis

Ridge Regression was used as a linear baseline and Random Forest as a nonlinear comparison. Within every outer training fold, features were standardized using parameters estimated only from the training participants. Ridge regularization strength α was tuned over the 10^{-3}–10^{3} range within each outer leave-one-participant-out (LOPO) training set. The inner loop used participant-wise leave-one-out cross-validation across the 29 available training participants, yielding 29 inner validation folds. The α value with the lowest average validation error across the inner folds was retained. Random Forest used 100 trees, min_samples_leaf=2, and random_state=42. The analysis was implemented in Python with pandas, NumPy, SciPy, and scikit-learn. All preprocessing and model-selection decisions were confined to each training fold, so the held-out participant and target-session label did not contribute to feature scaling or hyperparameter selection.

Validation followed participant-wise LOPO. For within-game prediction, one participant was held out, and the other 29 participants from the same game formed the training set. For cross-game prediction, the same target participant was held out in the target game, and that participant's record was also excluded from the source-game training data, preventing participant leakage. Figure 2 summarizes the pipeline from raw Questset records through preprocessing, feature construction,

DOI: 10.54254/2755-2721/2026.BA38767

model fitting, and the two evaluation settings. This design yields 30 out-of-sample predictions for each target game and model.

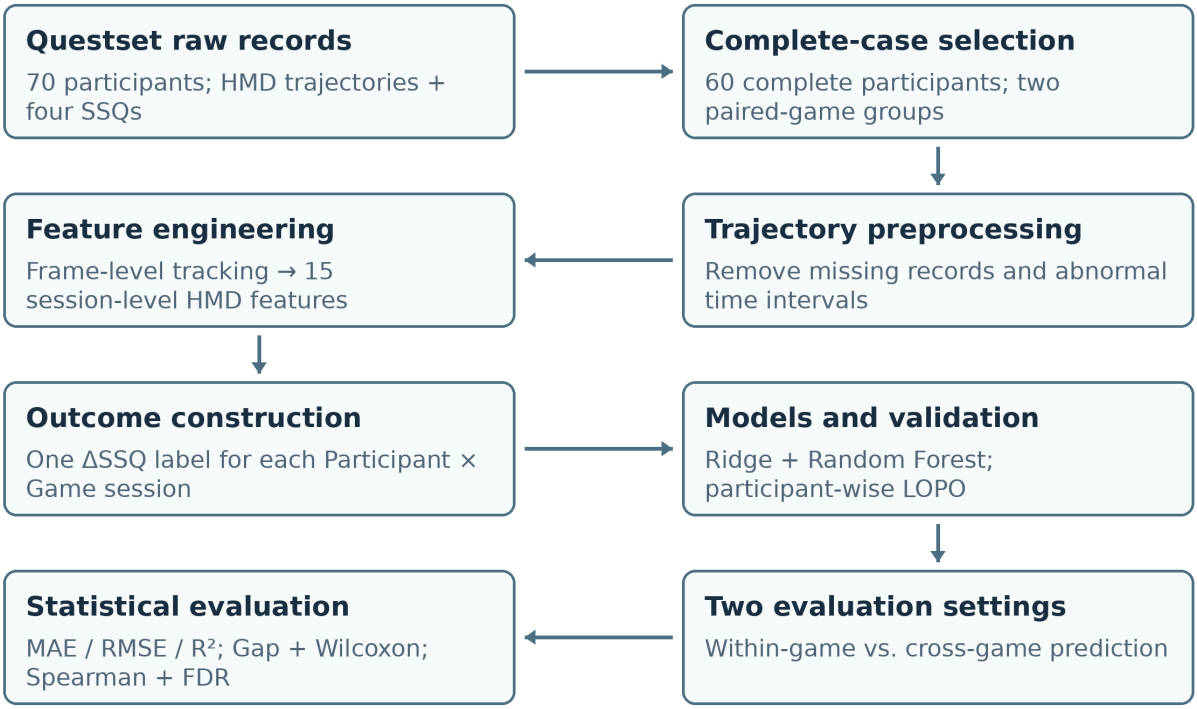


Figure 2. Secondary-analysis pipeline from Questset raw records to cross-game evaluation (photo credit: original)

Mean absolute error (MAE) was the primary metric, with Root Mean Square Error (RMSE) and Coefficient of Determination (R^2) also reported. R^2 was computed relative to a mean-prediction baseline based on the mean ΔSSQ of the outer training fold only. In cross-game settings, this training-fold mean can differ from the target-game outcome distribution; consequently, part of a negative R^2 can arise from source–target baseline mismatch in addition to weak model fit. Negative R^2 should therefore not be interpreted as model failure alone. Generalization Gap was defined as cross-game MAE minus within-game MAE for the same target game. Paired Wilcoxon signed-rank tests compared participant-level absolute errors between the two settings; raw p-values are reported, with a Bonferroni threshold across eight comparisons used as a sensitivity check. Spearman ρ described feature–ΔSSQ associations within each game, and Benjamini–Hochberg FDR correction was applied across the 60 exploratory feature tests. Both adjustments are sensitivity checks rather than substitutes for effect magnitudes.

4. Results

4.1. Descriptive results

ΔSSQ distributions differed across games (Table 3). Beat Saber had a mean change near zero, whereas Cooking Simulator showed the highest mean increase; the two Group 2 games were more similar. Because ΔSSQ can be negative, negative values indicate that post-game SSQ was below the relevant pre-game baseline rather than representing "negative sickness." These descriptive

DOI: 10.54254/2755-2721/2026.BA38767

differences are not interpreted as causal game effects because game order and participant-group membership were not experimentally isolated in the secondary analysis.

Table 3. Descriptive statistics of ΔSSQ by game

Game	N	Mean ΔSSQ	Median	SD
Beat Saber	30	-0.12	0.00	12.82
Cooking Simulator	30	24.06	13.09	29.05
Medal of Honor	30	11.59	11.22	15.60
Forklift Simulator	30	13.84	9.35	19.45

4.2. Cross-game prediction performance

Table 4 reports MAE, RMSE, and R^2 for all 16 train→test conditions; for compactness, Cooking Simulator is abbreviated as Cook in the table. The largest degradation occurred when Beat Saber was the target. Ridge MAE increased from 8.86 for Beat→Beat to 27.46 for Cooking→Beat, while Random Forest increased from 10.34 to 26.29. Of the 16 conditions, only one yielded a positive R^2 (0.06); all others were negative. Under the training-mean baseline used here, negative cross-game R^2 values can be amplified when the source and target ΔSSQ distributions differ, so they reflect distribution mismatch as well as predictive fit. The current 15 session-level features therefore provide limited absolute individual-level prediction, and the main evidence is the relative change between within-game and cross-game conditions.

Table 4. Prediction performance for all within-game and cross-game conditions

Group	Model	Train→test	MAE	RMSE	R^2
G1	Ridge	Beat→Beat	8.86	13.17	-0.09
G1	Ridge	Cook→Cook	21.43	28.96	-0.03
G1	Ridge	Beat→Cook	24.69	36.70	-0.65
G1	Ridge	Cook→Beat	27.46	30.71	-4.93
G1	Random Forest	Beat→Beat	10.34	14.26	-0.28
G1	Random Forest	Cook→Cook	21.56	28.67	-0.01
G1	Random Forest	Beat→Cook	26.41	39.40	-0.90
G1	Random Forest	Cook→Beat	26.29	30.52	-4.86
G2	Ridge	Medal→Medal	12.96	16.66	-0.18
G2	Ridge	Forklift→Forklift	14.47	19.89	-0.08
G2	Ridge	Medal→Forklift	14.70	20.65	-0.17
G2	Ridge	Forklift→Medal	15.58	18.66	-0.48
G2	Random Forest	Medal→Medal	11.25	14.87	0.06
G2	Random Forest	Forklift→Forklift	14.22	19.86	-0.08
G2	Random Forest	Medal→Forklift	18.90	24.09	-0.59
G2	Random Forest	Forklift→Medal	17.23	20.35	-0.76

The first direct summary of the transfer penalty is provided in Table 5. When Beat Saber was the target, the Generalization Gap was 18.60 for Ridge and 15.95 for Random Forest, with paired Wilcoxon $p < .001$ for both. Using the Bonferroni sensitivity threshold of .00625, both Beat-target

DOI: 10.54254/2755-2721/2026.BA38767

effects remain below the threshold, whereas the Random Forest Forklift→Medal result ($p = .016$) does not. Figure 3 visualizes the same pattern: transfer penalties are large for Beat Saber but much smaller for the other target games. Thus, cross-game degradation is directional rather than uniform.

The smaller gaps for the remaining targets further illustrate this asymmetry. When Cooking Simulator was the target, the gaps were 3.26 for Ridge and 4.85 for Random Forest; when Medal of Honor was the target, the corresponding gaps were 2.61 and 5.98. Forklift Simulator showed the smallest Ridge penalty (0.24), although Random Forest still produced a gap of 4.69. Thus, switching the training game did not impose a fixed penalty of similar size across conditions. Instead, the magnitude depended jointly on the target game and model, with the Beat Saber target standing out as the clearest case of degraded transfer under both model families.

Table 5. Generalization gap and paired Wilcoxon test by target game

Target game	Model	Within MAE	Cross MAE	Gap	p
Beat Saber	Ridge	8.86	27.46	18.60	<.001
Beat Saber	Random Forest	10.34	26.29	15.95	<.001
Cooking Simulator	Ridge	21.43	24.69	3.26	.428
Cooking Simulator	Random Forest	21.56	26.41	4.85	.329
Medal of Honor	Ridge	12.96	15.58	2.61	.177
Medal of Honor	Random Forest	11.25	17.23	5.98	.016
Forklift Simulator	Ridge	14.47	14.70	0.24	.670
Forklift Simulator	Random Forest	14.22	18.90	4.69	.058

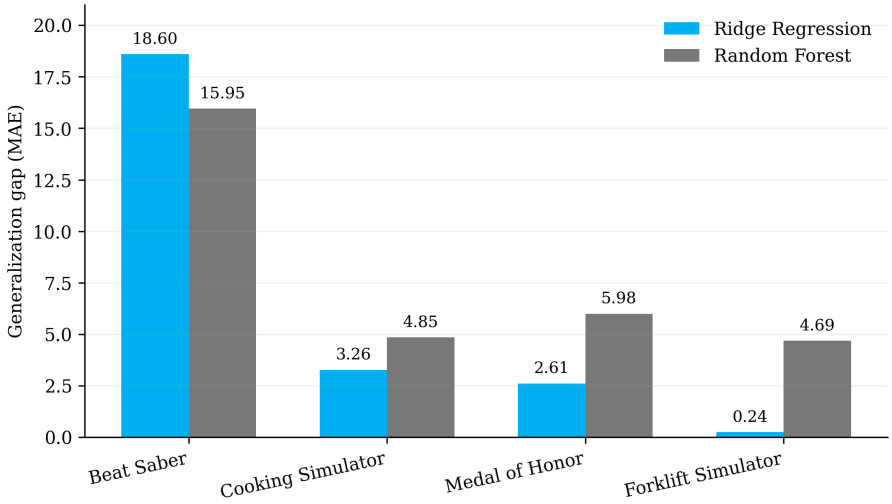


Figure 3. Generalization gap by target game and model (photo credit: original)

4.3. Feature stability and context dependence

Feature-level results show a second form of instability. In Group 1, none of the 15 features retained the same Spearman correlation direction across Beat Saber and Cooking Simulator. For example, linear-speed Standard Deviation (SD) changed from $\rho = .406$ to $\rho = -.450$, angular-speed SD from .365 to −.448, and angular-acceleration 95th Percentile (P95) from .213 to −.508. Group 2 showed

DOI: 10.54254/2755-2721/2026.BA38767

weaker changes; Y-position SD remained positive in both Medal of Honor (ρ = .255) and Forklift Simulator (ρ = .294). Figure 4 summarizes eight representative coefficients across the four games.

These sign changes are descriptive rather than definitive feature effects. Across 15 features × 4 games, 19 correlations had uncorrected $p < .05$, but none remained significant after Benjamini–Hochberg FDR correction. Together with the pronounced difference in ΔSSQ distributions between Beat Saber and Cooking Simulator, the results support context dependence as a plausible explanation: identical motion values may represent different behaviors or target distributions across games. However, the secondary dataset does not isolate task, visual motion, locomotion, or other game factors, so no specific factor can be identified as the cause of the observed transfer pattern.

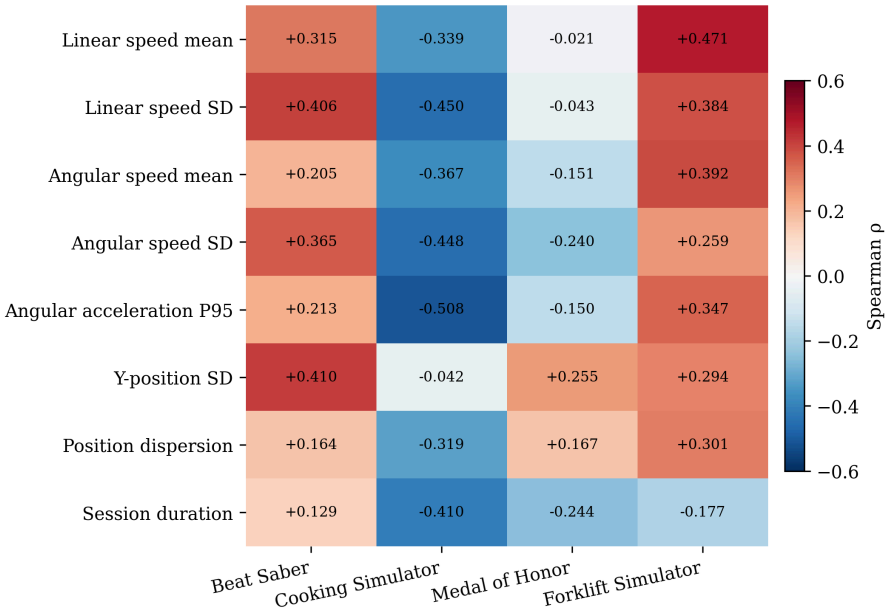


Figure 4. Spearman correlations between representative HMD features and ΔSSQ across games (photo credit: original)

The feature analysis and prediction analysis should also be interpreted as complementary rather than as a demonstrated causal chain. In Group 1, strong directional changes in several correlations occur alongside the largest transfer penalties, but the present design cannot establish that those correlation changes caused the higher MAE. Differences in ΔSSQ distribution, task semantics, movement repertoire, exposure characteristics, or unmeasured participant state could contribute simultaneously. The value of the combined evidence is therefore narrower: it shows that both model performance and the descriptive relationship between motion features and symptoms can change when the game context changes, which motivates explicit cross-game validation before assuming transportability.

5. Conclusion

This study evaluated whether cybersickness prediction based on consumer-grade HMD motion data generalizes across VR games. Using Questset data from 60 complete participants, Ridge Regression and Random Forest were tested under within-game and cross-game conditions, and 15 session-level motion features were examined in relation to ΔSSQ. Cross-game transfer produced the clearest error increase when Beat Saber was the target, whereas other directions showed smaller penalties.

DOI: 10.54254/2755-2721/2026.BA38767

Feature–symptom correlations also changed across games, but no individual correlation remained significant after FDR correction. These results indicate that within-game predictive relationships should not be assumed to transfer automatically to unseen content.

The evidence also has clear limits. Each game contributed only about 30 complete participants, 10 participants with pronounced cybersickness were excluded from the paired analysis because they withdrew, and the four games were not experienced by the same participants. In addition, the session-level features compress temporal dynamics and cannot support real-time early warning. Future work should use larger within-subject multi-game datasets, temporal features, external validation, and transfer or domain-adaptation methods. Under the current data, the principal contribution is therefore the identification of a directional cross-game generalization boundary rather than a claim of a deployment-ready prediction system. This distinction is practically important because deployment across diverse VR applications requires evidence of robustness beyond the content used for model development.

References

[1] Saredakis, D., Szpak, A., Birckhead, B., Keage, H. A. D., Rizzo, A., & Loetscher, T. (2020). Factors associated with virtual reality sickness in head-mounted displays: A systematic review and meta-analysis. *Frontiers in Human Neuroscience*, 14, 96.

[2] Caserman, P., Garcia-Agundez, A., Gámez Zerban, A., & Göbel, S. (2021). Cybersickness in current-generation virtual reality head-mounted displays: Systematic review and outlook. *Virtual Reality*, 25, 1153–1170.

[3] Islam, R., Desai, K., & Quarles, J. (2021). Cybersickness prediction from integrated HMD's sensors: A multimodal deep fusion approach using eye-tracking and head-tracking data. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)* (pp. 31–40).

[4] Maneuvrier, A., Nguyen, N.-D.-T., & Renaud, P. (2023). Predicting VR cybersickness and its impact on visuomotor performance using head rotations and field (in)dependence. *Frontiers in Virtual Reality*, 4, 1307925.

[5] Ramaseri-Chandra, A. N., Reza, H., & Pothana, P. (2025). Exploring the feasibility of head-tracking data for cybersickness prediction in virtual reality. *Electronics*, 14(3), 502.

[6] Wang, P., Li, M., Wu, J., Wang, L., & Liu, Y.-J. (2026). Cybersickness exploration for different VR tasks under variable rendering conditions. *International Journal of Human–Computer Interaction*, 42(10), 7414–7437.

[7] Yang, W., Hua, C., Tao, J., Dai, Z., Chen, Y., Yin, L., & Liu, J. (2026). Cross-dataset study on virtual reality motion sickness detection based on EEG. *Biomedical Signal Processing and Control*, 126(B), 110952.

[8] Baldoni, S., Battisti, F., Chiariotti, F., Mistrorigo, F., Shofi, A. B., Testolina, P., Traspadini, A., Zanella, A., & Zorzi, M. (2024). Questset: A VR dataset for network and quality of experience studies. In *Proceedings of the 15th ACM Multimedia Systems Conference* (pp. 408–414).

[9] Kennedy, R. S., Lane, N. E., Berbaum, K. S., & Lilienthal, M. G. (1993). Simulator Sickness Questionnaire: An enhanced method for quantifying simulator sickness. *The International Journal of Aviation Psychology*, 3(3), 203–220.

[10] Jasper, A., Sepich, N. C., Gilbert, S. B., Kelly, J. W., & Dorneich, M. C. (2023). Predicting cybersickness using individual and task characteristics. *Computers in Human Behavior*, 146, 107800.