Integrating Cognitive Strategies and Motor Dynamics to Analyze VR Learning Efficiency

Xiang Li[1] **, Fan Sun**[1] **, Hanlin Hu**[2] **, Geping Liu**[1,*]
[1]Faculty of Education, Southwest University, Chongqing, China
[2]China West Normal University, Nanchong, Sichuan, China
[*]Corresponding author: liugp@swu.edu.cn

Abstract

Learner performance in Virtual Reality (VR) often varies substantially across individuals, while conventional assessments provide limited insight into the links between cognitive strategies and motor execution. Focusing on the wire-connection phase of a VR-based Wheatstone bridge experiment, this study proposes a task-driven dual-path framework grounded in Perception-Action Coupling theory. The framework combines PAC-informed visual sampling markers with implicit spatiotemporal dynamics from raw controller trajectories. Using data from 98 participants, the hybrid model achieved 71.03% accuracy, showing a modest advantage over single-path baselines. Error-pattern analysis further suggests a possible cognitive-motor decoupling pattern, in which strategic planning and motor execution may contribute differently to learning efficiency. This study provides a task-specific computational perspective for analyzing learner states in immersive procedural tasks.

Keywords: Perception–Action Coupling; Multimodal Learning Analytics; Dual-path Hybrid Framework; Virtual Reality

Introduction

Virtual Reality (VR) technologies are increasingly prevalent in education and training, offering unique advantages in facilitating immersive experiences and procedural skill acquisition (Wu, Yu, & Gu, 2020). Empirical studies have demonstrated that VR-based laboratories significantly enhance learner engagement, conceptual understanding, and skill acquisition, particularly in science and engineering education (Makransky & Petersen, 2021; Tao, Cukurova, & Song, 2025). However, despite these affordances, learners exhibit significant individual variability in performance and learning efficiency when interacting with VR environments (Lawson & Mayer, 2024). Methodological limitations persist in current VR learning analytics. First, in terms of feature engineering, most studies rely on hand-crafted, explicit statistical metrics, such as average fixation duration or total path length, evaluated via traditional classifiers like SVM or Random Forest (Prevezanou et al., 2024; Lam et al., 2022). While interpretable, these aggregated metrics often oversimplify complex behaviors, resulting in the loss of rich micro-temporal patterns embedded in high-frequency sensor data. Second, deep learning architectures such as LSTM, CNN, and Transformers have become mainstream for action recognition (Khan et al., 2025; Kuo, Chen, & Kuo, 2022), their black-box nature poses challenges in educational contexts. Although end-to-end models can accurately predict outcomes (Sirajudeen et al., 2024), they struggle to explicate the cognitive mechanisms behind inefficiency. In addition, existing Multimodal Learning Analytics research often suffers from a dimensional split: studies typically focus either on the cognitive layer via eye-tracking or the behavioral layer via controller trajectories, rarely integrating explicit cognitive strategies and implicit motor execution within a unified computational framework (Xie, Yang, Zhang, Chen, & Li, 2025; Chango, Lara, Cerezo, & Romero, 2022).

Theoretically, *Perception-Action Coupling (PAC)* theory provides a principled framework for linking visual information sampling with goal-directed action (J. J. Gibson, 1978; J. Gibson, 2014). PAC views cognition as emerging from the continuous coordination between perception and action. In immersive procedural tasks, efficient performance depends on learners' ability to selectively sample task-relevant information and translate it into spatially precise motor execution. High-efficiency learners are therefore expected to show more focused visual search and more economical movement patterns, whereas inefficient performance may involve fragmented attention, excessive exploration, and repetitive corrective movements. However, PAC has rarely been operationalized as a computational framework for quantifying fine-grained differences in learning efficiency in immersive VR tasks. This study investigates learning efficiency in a VR-based Wheatstone bridge experiment. Rather than analyzing the task holistically, we employ a task-driven approach to identify a critical cognitive bottleneck: the *wire-connection task step*. This step not only exhibits the highest error rate across learners, but also imposes task demands that are closely related to perception-action coupling. During wire connection, learners must visually locate relevant terminals, maintain attention on task-relevant *Areas of Interest (AOIs)*, judge spatial correspondences, and translate visual information into precise hand movements within a short period. Compared with observation- or reading-dominated steps, this phase therefore provides a task-constrained setting for examining how visual information sampling and motor execution jointly shape learning efficiency. Focusing on this bottleneck, we analyze multimodal behavioral data to characterize individual differences in visual strategies and motor dynamics (Kelso, 1995; Aloimonos, Weiss, & Bandyopadhyay, 1988).

Drawing on the PAC framework and *Active Vision* theory, we propose the following hypotheses **H1: Goal-directedness of Visual Sampling.** High-efficiency learners will exhibit more goal-directed visual sampling strategies. Their visual attention is expected to be more strictly focused on *Areas of Interest (AOIs)* directly relevant to the current operational goal. **H2: Economy of Perception-Action Coordination.**

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

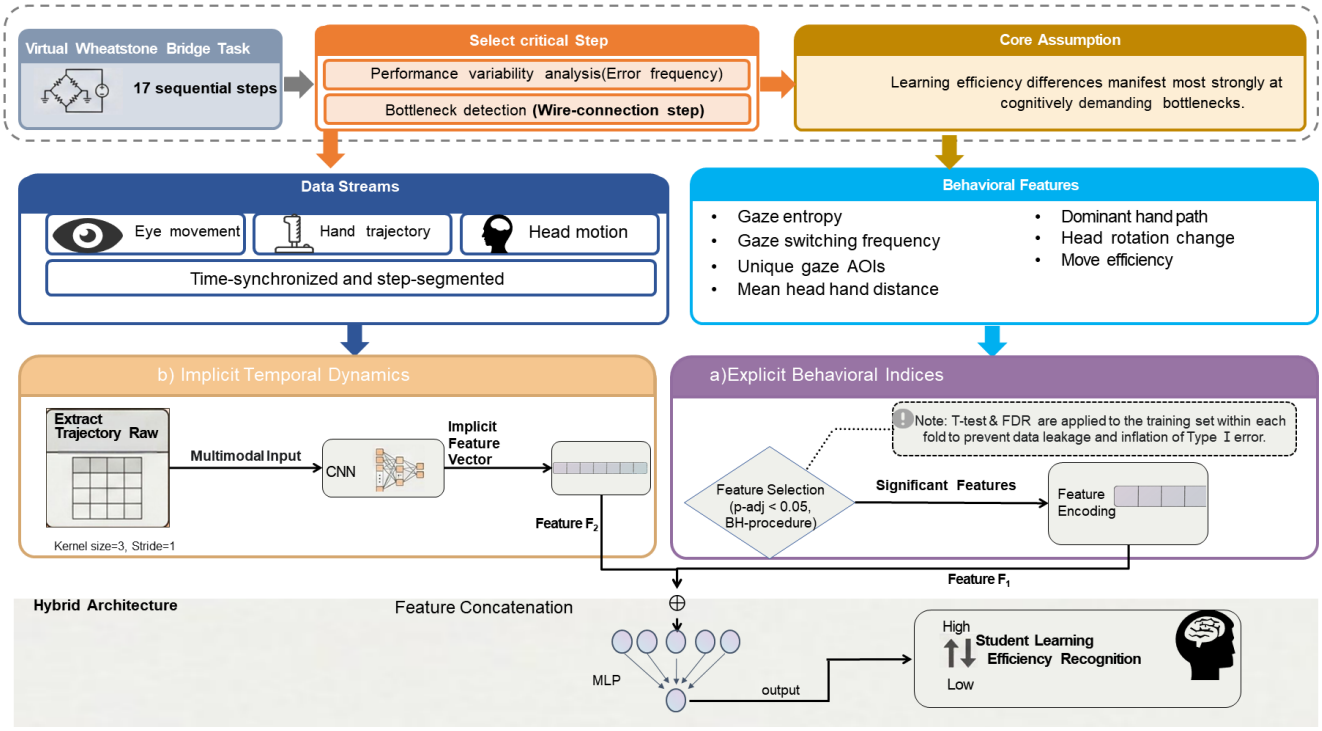

Figure 1: Overview of the proposed task-driven dual-path hybrid modeling framework. The framework first identifies a cognitively demanding bottleneck step based on task-level performance variability. Multimodal behavioral data are then synchronized and segmented at the step level. Explicit behavioral indices and implicit temporal motor dynamics are extracted through two complementary paths and integrated within a hybrid architecture to model individual differences in learning efficiency.

Compared to low-efficiency learners, high performers will demonstrate shorter hand trajectories and fewer redundant movements during the wire-connection phase, reflecting superior motor execution efficiency and coordination. **H3: Complementarity of Behavioral Features.** Learning efficiency involves complex temporal dependencies and dynamic interaction patterns. We hypothesize that explicit strategic features and implicit motor dynamics are complementary; relying on a single behavioral layer is insufficient to fully characterize learner states. In summary, this study contributes to VR learning analytics in three ways. **First**, it adopts a task-driven bottleneck analysis to focus multimodal behavioral modeling on a high-load procedural step, rather than treating the entire VR task as a homogeneous learning process. **Second**, it provides a PAC-informed mapping between task demands and behavioral indicators, linking visual information sampling and motor execution to the wire-connection phase. **Third**, it evaluates a dual-path hybrid model that combines interpretable strategic markers with implicit trajectory dynamics. The goal is not to claim a new fusion mechanism, but to examine whether these two behavioral representations provide complementary evidence for modeling learning efficiency.

Method

To operationalize PAC theory and information sampling into empirical analysis, this study constructed a hierarchical computational framework, as illustrated in Figure 1. Grounded in PAC and embodied cognition, we assume that complex VR learning tasks often contain task-specific phases in which learner performance is strongly constrained by the coordination between visual information sampling and goal-directed action (Gottlieb, Oudeyer, Lopes, & Baranes, 2013; Vinton et al., 2024; Xiong et al., 2025). Rather than treating the VR task as a single monolithic performance outcome, we decompose the learning process into task-specific **cognitive bottlenecks**.

In this study, a cognitive bottleneck refers to a task segment in which learners must coordinate visual search, action planning, and motor execution before completing a critical operational decision. It is not treated as a directly measured subjective psychological state, but as an operational construct defined by task analysis and the observed distribution of behavioral performance (Araújo, Davids, & Hristovski, 2006). Specifically, a step was identified as a cognitive bottleneck when it met two criteria: (1) it showed the highest error rate among the learner population, indicating a strong constraint on task progression; and (2) it required learners to coordinate visual information sampling, spatial judgment, and manual

execution within a short period. Centering on this bottleneck, we analyzed learner behavior from two complementary perspectives: explicit, theory-informed behavioral indices and implicit, high-resolution spatiotemporal dynamics.

Experimental Basis

This study was conducted in a dedicated Virtual Reality and motion capture laboratory measuring approximately 10 m^2. The core experimental apparatus was the HTC VIVE Pro Eye VR system, which includes a head-mounted display with binocular eye-tracking, two handheld controllers, and two spatial positioning base stations, as shown in Figure 2.c. Leveraging the Lighthouse laser positioning technology of the base stations, the system enables real-time spatial tracking of the HMD and controllers across six degrees of freedom.

The experimental task was a VR-based Wheatstone bridge resistance measurement experiment. It was designed as a procedural skill task consisting of 17 sequential steps, as illustrated in Figure 2.b. The system imposed strict logical constraints, requiring learners to complete the current step before proceeding to the next phase. This task was selected because it requires not only precise manual interaction but also the coordination of multiple cognitive resources, including target localization, numerical reading, spatial judgment, and state evaluation. We recruited 112 university student volunteers through campus advertisements. All participants completed a pre-test questionnaire on the Wheatstone bridge experiment before the session, and the results indicated no significant individual differences in prior knowledge. Written informed consent was obtained from all participants.

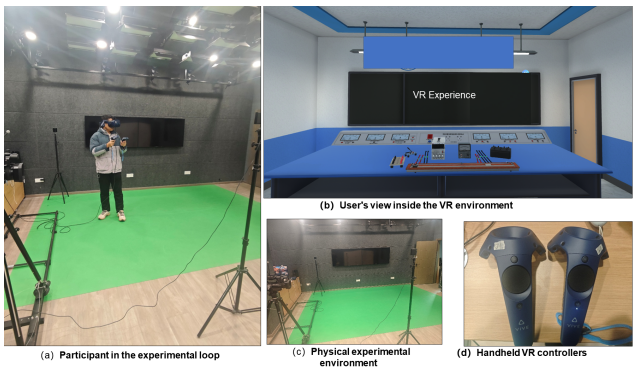

Figure 2: VR experimental setup, including the laboratory space, VR equipment, and participant interaction scene.

Data Processing

Raw multimodal data were recorded as continuous time series (Ochoa et al., 2022). Based on PAC theory, we define *learning efficiency* as the extent to which a learner coordinates visual information sampling with goal-directed manual actions during the VR task (Hayhoe & Ballard, 2005). Accordingly, learning efficiency was operationalized through behavioral indicators related to visual information search, temporal coordination, and motor planning efficiency.

We segmented the continuous data stream into multiple independent *Action Units*. Each Action Unit is defined as a complete process of goal-directed behavior driven by a clear operational intent (Zacks, Speer, Swallow, Braver, & Reynolds, 2007). To minimize subjective bias in manual segmentation, the boundaries of Action Units were determined using the event logs automatically generated by the VR system. Whenever a learner triggered the interactive feedback for a specific operation, the system recorded the corresponding timestamp. Following the planning-priority principle in PAC theory, each Action Unit was extracted from 3 s before the action-trigger timestamp to capture the perceptual and motor preparation preceding the operation (Cisek & Kalaska, 2010).

To ensure data quality, raw spatial trajectories were resampled to a uniform frequency of 60 Hz using linear interpolation. A Savitzky-Golay filter with a window length of 5 and a second-order polynomial was applied to suppress high-frequency sensor noise. All spatial coordinates were standardized using Z-score normalization and transformed into a task-centric coordinate system, so that the subsequent analysis focused on functional movement dynamics rather than global position differences.

Dual-path Behavioral Feature Characterization

To clarify how PAC theory informed feature construction, we treated the wire-connection phase as a task-specific instance of perception-action coupling. In this phase, learners must visually locate relevant terminals, maintain attention on task-relevant *Areas of Interest (AOIs)*, judge spatial correspondences, and transform visual information into precise hand actions. Therefore, gaze-based features were used to characterize the information-sampling component of PAC, whereas controller- and head-motion features were used to characterize action execution, visuomotor coordination, and spatial reorientation cost. These measures were not intended as generic behavioral descriptors, but as task-constrained indicators of the perceptual and motor demands imposed by the wire-connection step.

We categorize behavioral features into two complementary paths. The **Strategic Layer** characterizes how learners actively sample information and allocate attention. We extracted three key metrics: (1) *Gaze Switch Frequency*, reflecting the information sampling rate and exploratory search intensity (Hayhoe & Ballard, 2005); (2) *Gaze Entropy*, characterizing the randomness versus goal-directedness of visual search (Horowitz & Wolfe, 1998); and (3) *Unique AOI Count*, measuring the breadth of exploratory information sampling (Hills, Todd, Lazer, Redish, & Couzin, 2015).

The **Execution Layer** captures movement efficiency and spatial coordination patterns. This path includes four metrics: (4) *Hand Trajectory Length*, indicating action planning efficiency and motion economy (Shadmehr & Krakauer, 2008); (5) *Mean Head-Hand Distance*, reflecting the maintenance of visuomotor coordination (Flash & Hogan, 1985); (6) *Movement Efficiency*, defined as the displacement-to-path ratio and

used to identify operational hesitation or redundant corrections (Wulf & Lewthwaite, 2016); and (7) *Total Head Rotation*, representing the motor cost of spatial reorientation (Wilson, 2002).

This dual-path characterization preserves theoretical interpretability while providing a multidimensional foundation for capturing both explicit strategic shifts and implicit spatiotemporal dynamics.

Hybrid Model Architecture

The proposed hybrid model, grounded in PAC theory, is illustrated in Figure 1. After Action Unit segmentation, the model processes data through two parallel streams. Given the relatively limited final sample size (N = 98), we adopted several strategies to reduce overfitting and improve the stability of model evaluation.

Explicit Path Selection: The seven extracted behavioral features were used as candidate explicit indicators. To reduce the influence of noisy variables, we applied the Benjamini-Hochberg procedure to control the false discovery rate during significance testing. This step was used to obtain a compact explicit feature vector and to examine whether theory-informed behavioral indices carried discriminative information about learning efficiency.

Implicit Path Modeling: In parallel, the implicit path used a 1D-CNN to model raw multimodal trajectory data. To increase training density under a modest sample size, data augmentation was performed using a sliding window with 50% overlap on the 3-second action segments. To further reduce overfitting risk, the 1D-CNN architecture was kept shallow with three convolutional layers. A Global Average Pooling layer was used to map local spatiotemporal patterns into an implicit representation vector while limiting the total number of parameters.

Fusion and Inference: Finally, we adopted a feature-level fusion strategy. The explicit strategic vector and the implicit dynamic vector were concatenated to construct a unified representation of visual strategy and motor dynamics. This fused vector was then input into an MLP classifier composed of fully connected layers. Dropout layers ($p = 0.5$) and L2 regularization were used during training to improve stability. Model performance was evaluated using subject-wise 10-fold stratified cross-validation, ensuring that segments from the same participant did not appear in both the training and testing sets. This design allowed us to examine whether explicit behavioral indices and implicit trajectory dynamics provided complementary information for modeling learning efficiency, without claiming that the fusion architecture itself constitutes a new cognitive mechanism.

Experiments

This section evaluates the proposed computational analysis framework through descriptive statistics, group-level behavioral comparisons, and model ablation analyses. First, we report the descriptive statistics used for cognitive bottleneck identification and learning efficiency grouping. Second, we examine group-level differences in explicit behavioral features between high- and low-efficiency learners. Finally, we compare the predictive performance of explicit-only, implicit-only, and hybrid models to assess whether the two behavioral representations provide complementary information.

Descriptive Statistics and Grouping

Cognitive Bottleneck Identification To identify critical cognitive bottlenecks, we analyzed error performance across the 17 sequential steps of the Wheatstone bridge experiment. After excluding 14 participants due to incomplete experimental procedures, a total of 98 participants were included in the subsequent analysis. Specifically, we calculated the average error rate for each step to characterize the relative difficulty encountered by learners during different operational phases.

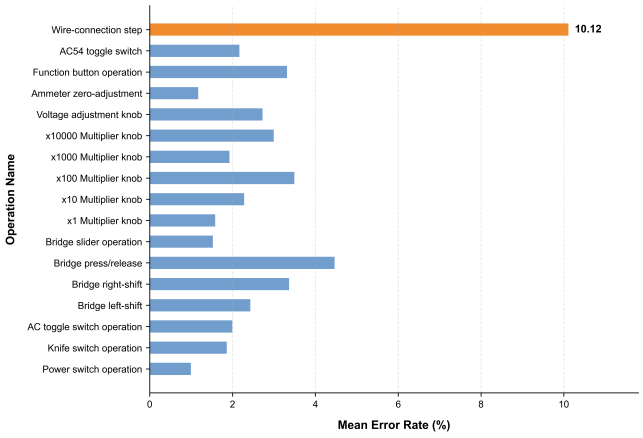

Figure 3: Average error rate across the 17 operational steps in the Wheatstone bridge experiment.

As shown in Figure 3, error rates varied across the 17 procedural steps. The *wire-connection phase* showed the highest average error rate, reflected in a prominent peak in attempt counts. Consistent with our task analysis, this phase also involved longer completion times and greater performance variability among participants. Compared with observation- or reading-dominated steps, wire connection requires learners to perform fine-grained manual manipulations in 3D space while continuously coordinating visual localization, spatial judgment, and action execution. Therefore, this phase was selected as the focal cognitive bottleneck for subsequent analyses. This choice allowed us to examine learner differences under a relatively constrained but high-load procedural context.

Learning Efficiency Grouping Figure 4 displays the distribution of total operation attempts for the 98 participants during the wire-connection phase. The distribution exhibits a certain degree of skewness, indicating substantial individual differences in operational efficiency at this critical step. Based on this distribution, we employed a **median split method** to group participants, thereby reducing the influence of extreme values. Specifically, participants were divided into

a High-Efficiency Group and a Low-Efficiency Group based on the median value.

It should be noted that this grouping serves as an operational definition for comparing behavioral patterns in subsequent analyses, rather than an absolute classification of learning ability. We acknowledge that median-based grouping simplifies a continuous distribution of learning efficiency. Therefore, the binary labels used here should be interpreted as an analytical contrast between relatively high- and low-efficiency behavioral profiles, rather than as discrete learner types. This operational choice was made to support interpretable group-level comparisons under a modest sample size and a skewed attempt-count distribution. Furthermore, pre-test results showed no significant differences in prior knowledge regarding the Wheatstone bridge between the two groups, suggesting that the behavioral differences observed subsequently are unlikely to be driven by prior knowledge.

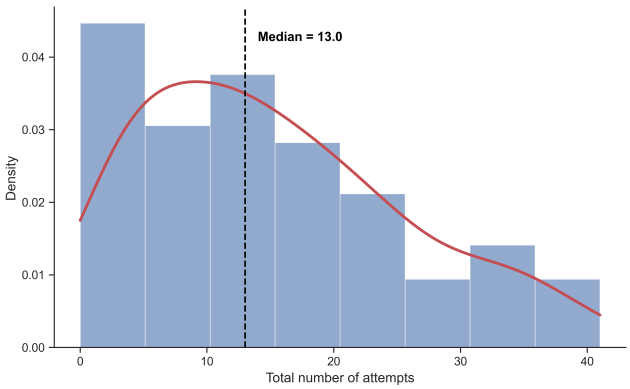

Figure 4: Distribution of total attempts with median-based learning efficiency grouping.

Behavioral Difference Analysis

To examine behavioral differences between groups, we conducted independent samples t-tests on the extracted features. The results are presented in Table 1.

Table 1: Independent Samples t-Test Results for Behavioral Features (N = 98)

Feature Name	p-value	Cohen's d	High (M)	Low (M)
Gaze Switch Frequency	**0.034**	0.44	2.50	2.94
Gaze Entropy	**0.039**	0.42	0.98	1.13
Unique AOI Count	0.051	0.40	3.29	3.69
Hand Trajectory Length	0.091	0.35	0.30	0.46
Total Head Rotation	0.124	0.31	0.15	0.12
Movement Efficiency	0.646	0.09	0.72	0.73
Mean Head-Hand Dist.	0.775	0.06	0.56	0.53

Note: **Bold** indicates $p < .05$; underline indicates $0.05 \leq p < .10$ (trend-level).

The results from the **Cognitive Strategic Layer** provide support for Hypothesis **H1**. High-efficiency learners demonstrated a more structured visual search pattern, as indicated by significantly lower Gaze Entropy ($p = .039$, $d = 0.42$) and lower Gaze Switch Frequency ($p = .034$, $d = 0.44$). Unique AOI Count reached trend-level significance ($p = .051$, $d = 0.40$), suggesting a possible tendency for high-efficiency learners to attend to fewer and more task-relevant areas.

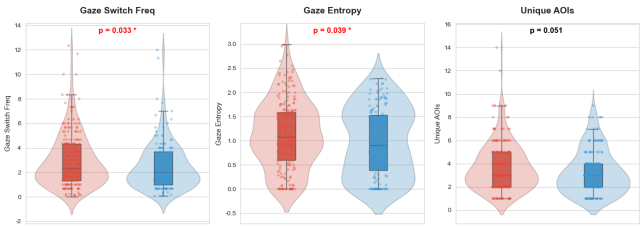

Figure 5: Visualization of goal-oriented visual strategies for high-efficiency learners.

The results from the **Action Execution Layer** provide only limited, trend-level support for Hypothesis **H2**. The high-efficiency group showed a trend toward shorter Hand Trajectory Length ($M_{high} = 0.30$ vs. $M_{low} = 0.46$, $p = .091$), whereas Movement Efficiency, Mean Head-Hand Distance, and Total Head Rotation did not reach statistical significance. Thus, the group-level statistical evidence for motor-execution differences was weaker than that for visual sampling differences.

Overall, these results suggest that high-efficiency learners differed more clearly in visual information sampling than in aggregated kinematic summaries. This pattern indicates that scalar movement features may not fully capture the temporal dynamics of motor execution, motivating the trajectory-based modeling approach presented in the subsequent section.

Ablation Study and Hybrid Performance

To address Hypothesis **H3**, we conducted an ablation study comparing the predictive performance of the explicit-only layer, the implicit-only layer, and the dual-path hybrid model on learning efficiency classification.

Performance Comparison As shown in Table 2, all models performed above the random chance level, suggesting that the behavioral data contained discriminative information for learning efficiency classification.

Table 2: Performance Comparison on Learning Efficiency Classification

Model	Accuracy (%)	F1-Score (%)
SVM (Explicit Baseline) (Cortes & Vapnik, 1995)	67.95	72.00
Random Forest (Breiman, 2001)	63.28	68.00
CNN (Implicit Baseline) (Kiranyaz, Ince, & Gabbouj, 2015)	69.23	76.28
LSTM (Hochreiter & Schmidhuber, 1997)	67.34	73.33
Proposed Hybrid Model	**71.03**	**77.45**

Note: Results are based on 10-fold cross-validation. **Bold** indicates the best performance.

Strategic Layer Baseline (SVM) The feature selection for the explicit SVM baseline was guided by the statistical results in Table 1. We excluded three trajectory-layer fea-

tures, namely Mean Head-Hand Distance, Movement Efficiency, and Total Head Rotation, because they did not reach trend-level significance ($p > .10$). Preliminary analyses indicated that including these weakly discriminative kinematic variables introduced noise and degraded the SVM's discriminative performance. Consequently, the SVM model used the four most robust markers, including two significant and two trend-level features, to establish a compact and theory-informed strategic baseline.

Trajectory Layer Baseline (CNN) The CNN model, learning directly from raw spatiotemporal trajectories without manual feature engineering, achieved an accuracy of 69.23%. The LSTM performed slightly worse than the CNN, with an accuracy of 67.34%. This pattern may be related to the short duration of the 3-second action segments and the modest sample size, which may limit the advantage of recurrent architectures. Although derived trajectory metrics showed limited significance in group-level statistical tests, the CNN results suggest that raw sensor streams may contain latent spatiotemporal patterns that are not fully captured by aggregated kinematic summaries.

Hybrid Fusion The proposed hybrid model employs a feature-level fusion strategy. It concatenates the explicit strategic vector with the implicit dynamic vector extracted by the CNN to construct a unified representation of visual strategy and motor dynamics. This fused vector is input into an MLP classifier for inference. The hybrid model achieved the highest mean accuracy of 71.03%, but the gain over the CNN baseline was modest. We therefore interpret this result not as evidence of a large predictive advantage, but as evidence that explicit strategic features and implicit trajectory dynamics contain partially complementary information.

Cognitive-Motor Decoupling Analysis Because the present study did not compute a dedicated coupling coefficient between visual and motor streams, cognitive-motor decoupling is used here as an interpretive framework derived from complementary feature patterns and model error cases, rather than as a directly measured coupling index.

To further interpret the model comparison, we selected four baseline models covering distinct paradigms: **SVM and Random Forest** as traditional machine learning models using handcrafted features, and **CNN and LSTM** as deep learning models for raw temporal trajectories.

As shown in Table 2, the relative performance of the CNN and LSTM suggests that learning efficiency in the 3-second wire-connection segments may be associated more with local spatiotemporal patterns than with long-range recurrent dependencies. Under the current sample size, the LSTM may also be more vulnerable to temporal noise in short, high-frequency sensor streams.

The hybrid model achieved the highest mean accuracy, suggesting that visual strategy and motor dynamics may provide partially complementary information. By combining the interpretability of the strategic layer with the temporal sensitivity of the CNN path, the hybrid framework may reduce some blind spots of single-path classifiers.

Inspection of classification errors suggested two typical mismatch patterns:

Case 1: High Strategy, Low Motor. Some learners showed relatively structured visual sampling, such as low gaze entropy, but their trajectories still contained micro-adjustments or speed fluctuations. These cases may reflect learners who had formed a relatively clear strategy but had not yet achieved fluent motor execution.

Case 2: Low Strategy, High Motor. Conversely, some learners showed relatively fluent manual movements, but their visual sampling remained less stable or more exploratory. These cases may reflect learners whose motor execution appeared efficient while their task-relevant visual strategy was still developing.

These error patterns provide preliminary evidence for a possible mismatch between strategic visual sampling and motor execution during skill acquisition. Thus, the value of the hybrid model lies less in a large increase in classification accuracy than in its ability to combine two complementary behavioral views of the learning process.

Conclusion

This study examined how visual sampling strategies and motor execution dynamics are associated with learning efficiency in a VR-based Wheatstone bridge task. By focusing on the wire-connection phase as a cognitively demanding bottleneck, we analyzed learner behavior under a constrained procedural context requiring visual localization, spatial judgment, and manual execution. The results showed clearer group differences in visual information sampling than in aggregated motor-execution summaries: high-efficiency learners exhibited lower gaze entropy and lower gaze switching frequency, whereas scalar kinematic features provided only limited, trend-level evidence.

The dual-path hybrid model achieved 71.03% accuracy and showed a modest advantage over single-path baselines, suggesting that explicit strategic features and implicit trajectory dynamics provide partially complementary information. Error-pattern analysis further suggested a possible cognitive-motor mismatch, but this interpretation should be treated as preliminary because no dedicated coupling coefficient was computed. Several limitations remain: learning efficiency was operationalized using a median split of attempt counts, the study focused on a single procedural task and one bottleneck step, and the generalizability of the framework requires further validation. Future work should examine more diverse VR learning tasks and develop explicit indices for quantifying the temporal coupling between visual sampling and motor execution.

Acknowledgments

This work is supported by the National Natural Science Foundation of China project "Key Technologies for Deep Cognitive Tracking and Learning Intervention Models in VR Environments" under Grant No. 62277044.

References

Aloimonos, J., Weiss, I., & Bandyopadhyay, A. (1988). Active vision. *International journal of computer vision*, *1*(4), 333–356.

Araújo, D., Davids, K., & Hristovski, R. (2006). The ecological dynamics of decision making in sport. *Psychology of sport and exercise*, *7*(6), 653–676.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Chango, W., Lara, J. A., Cerezo, R., & Romero, C. (2022). A review on data fusion in multimodal learning analytics and educational data mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *12*(4), e1458.

Cisek, P., & Kalaska, J. F. (2010). Neural mechanisms for interacting with a world full of action choices. *Annual review of neuroscience*, *33*(1), 269–298.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273–297.

Flash, T., & Hogan, N. (1985). The coordination of arm movements: an experimentally confirmed mathematical model. *Journal of neuroscience*, *5*(7), 1688–1703.

Gibson, J. (2014). *The ecological approach to visual perception: classic edition.* Psychology Press.

Gibson, J. J. (1978). The ecological approach to the visual perception of pictures. *Leonardo*, *11*(3), 227–235.

Gottlieb, J., Oudeyer, P.-Y., Lopes, M., & Baranes, A. (2013). Information-seeking, curiosity, and attention: computational and neural mechanisms. *Trends in cognitive sciences*, *17*(11), 585–593.

Hayhoe, M., & Ballard, D. (2005). Eye movements in natural behavior. *Trends in cognitive sciences*, *9*(4), 188–194.

Hills, T. T., Todd, P. M., Lazer, D., Redish, A. D., & Couzin, I. D. (2015). Exploration versus exploitation in space, mind, and society. *Trends in cognitive sciences*, *19*(1), 46–54.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Horowitz, T. S., & Wolfe, J. M. (1998). Visual search has no memory. *Nature*, *394*(6693), 575–577.

Kelso, J. S. (1995). *Dynamic patterns: The self-organization of brain and behavior*. MIT press.

Khan, M. A., Asadi, H., Qazani, M. R. C., Arogbonlo, A., Pedrammehr, S., Anwar, A., ... others (2025). Enhancing cognitive workload classification using integrated lstm layers and cnns for fnirs data analysis. *Computers*, *14*(2), 73.

Kiranyaz, S., Ince, T., & Gabbouj, M. (2015). Real-time patient-specific ecg classification by 1-d convolutional neural networks. *IEEE transactions on biomedical engineering*, *63*(3), 664–675.

Kuo, R.-J., Chen, H.-J., & Kuo, Y.-H. (2022). The development of an eye movement-based deep learning system for laparoscopic surgical skills assessment. *Scientific Reports*, *12*(1), 11036.

Lam, K., Chen, J., Wang, Z., Iqbal, F. M., Darzi, A., Lo, B., ... Kinross, J. M. (2022). Machine learning for technical skill assessment in surgery: a systematic review. *NPJ digital medicine*, *5*(1), 24.

Lawson, A. P., & Mayer, R. E. (2024). Individual differences in executive function affect learning with immersive virtual reality. *Journal of Computer Assisted Learning*, *40*(3), 1068–1082.

Makransky, G., & Petersen, G. B. (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), 937–958.

Ochoa, X., Lang, C., Siemens, G., Wise, A., Gasevic, D., & Merceron, A. (2022). Multimodal learning analytics-rationale, process, examples, and direction. *The handbook of learning analytics*, *2*, 54–65.

Prevezanou, K., Seimenis, I., Karaiskos, P., Pikoulis, E., Lykoudis, P. M., & Loukas, C. (2024). Machine learning approaches for evaluating the progress of surgical training on a virtual reality simulator. *Applied Sciences*, *14*(21), 9677.

Shadmehr, R., & Krakauer, J. W. (2008). A computational neuroanatomy for motor control. *Experimental brain research*, *185*(3), 359–381.

Sirajudeen, N., Boal, M., Anastasiou, D., Xu, J., Stoyanov, D., Kelly, J., ... Francis, N. (2024). Deep learning prediction of error and skill in robotic prostatectomy suturing. *Surgical Endoscopy*, *38*(12), 7663–7671.

Tao, L., Cukurova, M., & Song, Y. (2025). Learning analytics in immersive virtual learning environments: a systematic literature review. *Smart Learning Environments*, *12*(1), 43.

Vinton, L. C., Preston, C., de la Rosa, S., Mackie, G., Tipper, S. P., & Barraclough, N. E. (2024). Four fundamental dimensions underlie the perception of human actions. *Attention, Perception, & Psychophysics*, *86*(2), 536–558.

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic bulletin & review*, *9*(4), 625–636.

Wu, B., Yu, X., & Gu, X. (2020). Effectiveness of immersive virtual reality using head-mounted displays on learning performance: A meta-analysis. *British journal of educational technology*, *51*(6), 1991–2005.

Wulf, G., & Lewthwaite, R. (2016). Optimizing performance through intrinsic motivation and attention for learning: The optimal theory of motor learning. *Psychonomic bulletin & review*, *23*(5), 1382–1414.

Xie, Y., Yang, L., Zhang, M., Chen, S., & Li, J. (2025). A review of multimodal interaction in remote education: Technologies, applications, and challenges. *Applied Sciences*, *15*(7), 3937.

Xiong, H., Xu, X., Wu, J., Hou, Y., Bohg, J., & Song, S. (2025). Vision in action: Learning active perception from human demonstrations. *arXiv preprint arXiv:2506.15666*.

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., & Reynolds, J. R. (2007). Event perception: a mind-brain perspective. *Psychological bulletin*, *133*(2), 273.