Quantifying Cognitive Development in Tower Crane Operators through VR Training: A Longitudinal Study Using the SEEV Model

Furui Man[1], Junyu Chen[1] and Hung-Lin Chi[1]

[1]Department of Building and Real Estate, Faculty of Construction and Environment, The Hong Kong Polytechnic University, Hong Kong
freeman-furui.man@connect.polyu.hk, junyu.chen@polyu.edu.hk, hung-lin.chi@polyu.edu.hk

Abstract –

This study proposes a quantitative framework for assessing the cognitive development of tower crane operators by integrating high-fidelity Virtual Reality (VR) simulation with the SEEV (Salience, Effort, Expectancy, Value) attention model. Forty participants completed a longitudinal training protocol incorporating a strict restart-on-error mechanism to simulate safety-critical operations. Eye-tracking data were analyzed via multiple linear regression to estimate the relative weightings of cognitive factors across three training sessions. Results demonstrate a significant increase in model predictive power (R^2 rising from 0.46 to 0.66 while the P-value is reducing from 0.05 to less than 0.01), confirming a transition trend from stochastic search to systematic strategies. Interpretation of regression coefficients reveals a restructuring of attentional priorities: operators consistently suppressed bottom-up, stimulus-driven Salience while emphasizing top-down factors, specifically Value and Expectancy. Additionally, the growing negative weighting of Effort suggests the internalization of efficiency-oriented strategies aimed at minimizing biomechanical gaze costs. The diminishing influence of Salience and Effort empirically validates the reduction to an Expectancy-Value (EV) driven pattern as a hallmark of advancement toward expertise. This specific attentional weighting profile offers a potential quantitative metric for evaluating trainees based on cognitive strategy maturity rather than relying solely on traditional performance outcomes such as task duration and error rates.

Keywords –

Virtual Reality, Crane Operation Training, SEEV Model, Eye Tracking, Cognitive Assessment

1 Introduction

Serving as primary material-handling equipment in high-rise and modular construction projects, tower cranes are essential to construction operations but represent a major source of safety risk. Tower crane operation is a safety-critical task in which tolerance for error is minimal, as even brief lapses in operator attention can lead to catastrophic structural failure or loss of life. Crane-related incidents, predominantly electrocution and struck-by-load events, account for a significant proportion of construction fatalities [1]. Given these stakes, the construction industry emphasizes training approaches that foster and evaluate operator competency.

Virtual Reality (VR) has emerged as a transformative instructional tool in construction safety training. Unlike traditional apprenticeship-based approaches, VR offers an immersive environment where trainees can practice complex maneuvers and experience the consequences of errors without physical hazards. Studies by Sacks et al. [2] demonstrated that immersive training enhances hazard recognition and spatial understanding more effectively than classroom instruction. More recently, the application of VR in construction safety training has expanded substantially, with systematic literature reviews confirming its effectiveness in improving learner engagement and knowledge retention [3].

Despite VR's adoption, current assessment methodologies remain predominantly outcome-oriented, resulting in a limited understanding of operators' cognitive development. Conventional evaluations typically rely on performance metrics, including task completion time, collision frequency, and trajectory accuracy, which describe observable outcomes but provide little insight into how operators perceived the dynamically evolving visual information. As noted by Yoon and Park [4], simply demonstrating the effectiveness of VR-based training is insufficient; there is a pressing need to understand the underlying "black box" of cognitive development that influences this effectiveness. In particular, research has yet to quantitatively characterize visual attention patterns that differentiate novice and expert operators.

To bridge this gap, this study uses the Salience, Effort,

Expectancy, and Value (SEEV) model [5] to quantify visual attention. Originally from aviation, SEEV frames attention as a competition between bottom-up (Salience, Effort) and top-down (Expectancy, Value) factors. By decomposing gaze behavior, we move beyond binary pass/fail outcomes to assess cognitive maturity. We present a longitudinal framework integrating eye-tracking within a high-fidelity VR simulation. Unlike snapshot studies, we track participants across three sessions using a strict restart-on-error protocol. The objective is to estimate SEEV weighting coefficients as novices gain expertise, establishing a quantitative benchmark for operator competency.

2 Related Works

Given the risks associated with tower cranes [1], research identifies operator error as a primary accident cause, often stemming from misjudgements of dynamic visual information or spatial depth [2]. Historically, the construction industry has relied on apprenticeship-based approaches to develop this competency. However, this on-site training paradigm has come under scrutiny due to safety concerns, as noted by Sawhney et al. Consequently, there is an urgent industry imperative for training methodologies that facilitate the development and evaluation of novice operators' cognitive maturation without exposure to physical risk.

VR in construction broadly falls into two streams: safety hazard education and operational skill acquisition. In the domain of safety education, Sacks et al. [2] and Li et al. [3] established that immersive environments primarily enhance hazard recognition and user engagement by visualizing invisible risks [6]. Conversely, in operational skills training, while VR serves as a "safe-to-fail" testbed, competency assessment remains methodologically stagnant. Most frameworks assess discrete terminal outcomes (e.g., task duration, collisions) [4]. While these quantify the event, they lack the temporal resolution to explain error causality. Consequently, even studies acknowledging this "black box" [4] lack a standardized quantitative framework for monitoring continuous cognitive processes.

To bridge this gap, researchers have integrated eye-tracking to leverage visual attention metrics as indicators of cognitive resource allocation. In the context of hazard recognition, Hasanzadeh et al. [7] revealed that safety knowledge heavily influences visual search, while Jeelani et al. [8] showed experts exhibit shorter fixations and broader coverage. However, the application of eye-tracking in construction faces two notable limitations. First, the majority of existing studies focus on static hazard identification rather than continuous equipment operation [9]. Second, most studies adopt cross-sectional designs capturing performance at a single snapshot, leaving a gap in longitudinal research on how visual attention strategies evolve with expertise [10].

To quantitatively analyze this cognitive evolution, this study employs the SEEV model [5] as the theoretical framework to systematically decompose operator gaze behavior. Within this framework, Salience refers to the involuntary capture of attention by physically distinct stimuli, such as a joystick that contrasts sharply with the background color [11], while Effort represents the biomechanical cost of gaze redirection. These are counterbalanced by top-down cognitive factors: Expectancy, the anticipation of information change (bandwidth), and Value, the strategic importance of that information to the task goal. While validated in aviation [12], SEEV is rarely applied to vertical construction. By applying the SEEV model longitudinally, this study seeks to mathematically decompose operator gaze behavior, testing the hypothesis that effective training facilitates a structural transition from stochastic, salience-driven visual search to a systematic, model-driven scanning strategy governed predominantly by Expectancy and Value.

3 Methodology

3.1 Research Framework

This study adopted a quantitative research framework to examine the visual attention mechanisms of crane operators within a high-fidelity virtual environment. The framework integrated a restart-on-error training protocol with the SEEV model to interpret cohort-level gaze behaviors. Experimental data, including eye-tracking records and task performance metrics, were collected to fit a linear regression model. Employing a longitudinal design, this study tracked the same cohort of participants across three distinct training sessions. Within each session, gaze behaviours were analysed across three specific operational phases: the Preparation Phase, Hoisting Phase, and Landing Phase.

3.2 Virtual Training Scenario Development

3.2.1 Task Theme and Process Design

The virtual training scenario was constructed based on a real-world public housing construction project utilizing hybrid concrete construction methods. Within this scenario, participants were required to operate a tower crane to transport a steel wall formwork from a ground-level logistics zone to a designated installation point on a building's working surface.

To simulate the safety-critical nature of crane operations, the scenario incorporated a strict error-consequence mechanism. Any safety violation, e.g., collision and moving without a signal, resulted in immediate ter-

mination of the session, requiring the participant to restart the task from its initial state. This restart-on-error protocol was intended to enforce a safety-first cognitive set, ensuring that the Value parameter in the adopted cognitive model reflects authentic operational pressure. For analytical clarity, the operational flow was streamlined into three sequential phases:

1. Preparation Phase, which encompassed rigging checks and the initial hoisting of the load to a designated safe clearance height (3 meters above its initial resting position).
2. Lifting Phase, involving the horizontal and vertical transport of the suspended load across the virtual construction site environment.
3. Landing Phase, which focused on the precise alignment and placement of the formwork onto pre-installed guide pins at the target location.

3.2.2 Task Theme and Process Design

The simulation environment was developed using the Unity3D engine and deployed on a Windows-based workstation, with a Meta Quest Pro head-mounted display as the immersive interface. The Meta Quest Pro was selected due to its integrated, high-frequency eye-tracking capability, enabling precise capture of gaze data without external peripherals.

A Data Recorder module logged the simulation state at a sampling frequency of 50 Hz (i.e., 0.02 s intervals). To characterize participants' visual attention patterns, task-relevant scene elements were designated as Areas of Interest (AOIs), including the Load, Signaler, and Dashboard. At each sampling interval, a ray-casting method was employed to detect if the participant's gaze vector intersected the collider of any designated AOI.

In parallel, the system recorded operator control inputs and the kinematic state of the crane to delineate task phases temporally. The start and end timestamps for the Preparation, Lifting, and Landing phases were identified automatically using specific activity triggers, such as the activation of the hoist mechanism, clearance of the load from the ground-plane collider, and entry of the load into the predefined target proximity zone.

3.3 Experimental Implementation

3.3.1 Participants

A total of 40 university students (20 male, 20 female) with academic backgrounds in Construction and Environment were recruited for the study. To ensure that the analysis focused on novice cognitive processes, all participants were screened to confirm the absence of prior experience with either real-world or virtual crane operations.

3.3.2 Apparatus

To simulate the physical sensations associated with crane dynamics, including acceleration, braking, and wind-induced perturbations, participants were seated on a Motion Systems QS-S25 SPIDER 6-degree-of-freedom motion platform. Operator control inputs were captured using wireless VR controllers, which were mapped to the virtual crane's joystick interfaces to replicate standard operational controls. The simulation was executed on a desktop workstation equipped with an Intel Core i9-11900k CPU and an NVIDIA GeForce RTX 3060 GPU, ensuring a stable frame rate and low-latency rendering to mitigate simulation-induced discomfort. Visual immersion was delivered via a Meta Quest Pro head-mounted display.

3.3.3 Procedure

The experimental protocol followed a repeated-measures design. Initially, all participants underwent a Tutorial Session to familiarize themselves with the VR hardware and the basic controls of the tower crane, i.e., hoisting, trolley travel, and slewing. Following the induction, each participant completed three experimental training sessions on the same day continuously. During each training session, participants repeatedly performed a standardized lifting operation under enforcement of the restart-on-error protocol. Task repetition continued until the participant successfully transported the load from the designated origin to the target destination without any safety violations.

3.4 SEEV Model Implementation

The SEEV model, proposed by Wickens et al. [5], characterizes visual attention allocation as a weighted linear combination of four contributing factors. In this study, the probability of attending to a specific AOI, quantified here as Percentage Dwell Time (PDT), is expressed as:

$$PDT(AOI_i) = \beta_S \cdot S_i - \beta_E \cdot E_i + \beta_{Ex} \cdot Ex_i + \beta_V \cdot V_i + \epsilon \quad (1)$$

where S, E, Ex, and V represent Salience, Effort, Expectancy, and Value, respectively, and β represents the weighting coefficients derived from empirical data.

3.4.1 Salience

Salience (S) represents the magnitude of stimulus-driven attentional engagement induced by visual distinctiveness. In this study, Salience was quantified via the implementation of the Graph-Based Visual Saliency (GBVS) algorithm. Representative screenshots of the operator's natural view were captured during task execution and processed using the GBVS algorithm to generate a pixel-wise saliency map.

Raw salience scores were computed as the mean intensity of the top 30% of pixels within the bounding rectangle demarcated for each AOI, which empirically filters background noise to isolate the essential salience distribution [13]; to allow for cross-sample comparison, these values were normalized to the range $[0,1]$.

3.4.2 Effort

Effort (E) indexes the cognitive cost associated with attentional redirection. In this study, Effort was operationalized as the three-dimensional angular distance between the operator's neutral gaze vector (defined as forward-facing orientation aligned with the load or hook block) and the vector pointing to the centroid of the target AOI. Larger angular displacements correspond to greater biomechanical effort. All Effort values were normalized to a 0 to 1 scale.

3.4.3 Expectancy

Expectancy (Ex) reflects the operator's anticipation of information change (i.e., bandwidth) within a given AOI. Given the difficulty of directly measuring information bandwidth within a dynamic VR environment, Expectancy values were assigned using an expert elicitation procedure. Domain experts rated each AOI according to predefined categories of Expectancy and Value, as detailed in Table 1. These categories represent the expected frequency of state change using a discrete scale of $\{0, 0.1, 0.3, 0.5, 0.8, 1\}$.

3.4.4 Value

Value (V) represents the strategic relevance of an AOI to the successful completion of the current task goal, which was determined through expert assessment with the same discrete scale $\{0, 0.1, 0.3, 0.5, 0.8, 1\}$ and similar definitions shown in Table 2.

3.4.5 Quantification of Cognitive Parameters

The quantification procedures described in Sections 3.4.1 through 3.4.4 introduce a structured parameter set wherein the Salience (S), Effort (E), Expectancy (Ex), and Value (V) are treated as phase-dependent variables rather than static attributes of an AOI. Accordingly, for any given AOI i during operational phase p, the input vector to the cognitive model is defined as $X_{i,p} = [S_{i,p}, E_{i,p}, Ex_{i,p}, V_{i,p}]$. The specific locations and boundaries of the monitored AOIs within the simulation environment are illustrated in Figure 1, and the complete matrix of SEEV parameter assignments utilized in this study, derived from GBVS-based saliency estimation, biomechanical gaze effort calculations, and expert elicitation, is presented in Table 3.

Table 1. Ex Value Definitions

Ex Value	Frequency	Definition
0	Static	No state changes expected: the AOI contains static information.
0.1	Very Low	Negligible change frequency: state changes of the AOI are extremely rare.
0.3	Low	Infrequent changes: the AOI updates occasionally but remains largely stable.
0.5	Moderate	Average change frequency: the AOI exhibits a regular or typical update rate.
0.8	High	High change frequency: the AOI updates rapidly and requires frequent monitoring.
1	Maximum	Continuous or near-continuous state changes; maximum information bandwidth.

Table 2. V Value Definitions

V Value	Relevance	Definition
0	Irrelevant	The AOI has no strategic relevance to the current task objective.
0.1	Very Low	Minimal relevance: the AOI rarely contributes to task success.
0.3	Low	Supplemental relevance: useful but not necessary for completion.
0.5	Moderate	Average relevance: the AOI contributes to task performance but is not critical.
0.8	High	High strategic importance: the AOI is critical for optimizing performance and

avoiding errors.

1	Essential	Maximum relevance: the task cannot be successfully completed without this AOI.

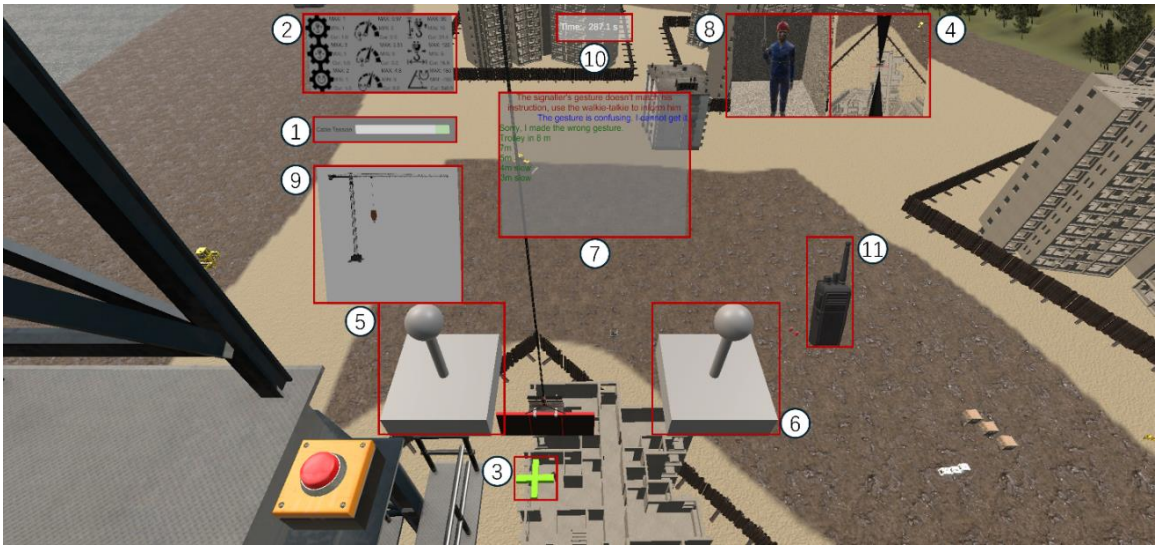

Figure 1. AOI Locations and boundaries

Table 3. Quantified SEEV Parameters

p	$AOI\ Index$	AOI	S	E	Ex	V
Preparation Phase (Prep)	1	Cable Tension Meter	1.000	0.649	0.0	0.1
	2	Components Meter	0.623	0.791	0.3	0.5
	4	Hook Top View	0.514	1.000	0.1	0.1
	5	Left Joystick	0.582	0.505	0.5	0.8
	6	Right Joystick	0.505	0.557	0.5	0.8
	7	Messages (UI)	0.443	0.302	1.0	1.0
	9	Crane Side View	0.663	0.522	0.3	0.1
	10	Timer (UI)	0.220	0.680	1.0	0.1
	11	Walkie Talkie	0.712	0.739	0.8	1.0
Lifting Phase (Lift)	1	Cable Tension Meter	1.000	0.649	0.8	0.8
	2	Components Meter	0.798	0.791	1.0	0.8
	4	Hook Top View	0.763	1.000	1.0	0.3
	5	Left Joystick	0.874	0.505	0.8	0.8
	6	Right Joystick	0.959	0.557	0.5	0.5
	9	Crane Side View	0.731	0.522	0.8	0.8
	10	Time (UI)	0.210	0.680	1.0	0.1
	11	Walkie Talkie	0.875	0.739	0.5	0.3
Landing Phase (Land)	1	Cable Tension Meter	0.598	0.649	0.1	0.1
	2	Components Meter	0.437	0.791	0.8	0.5
	3	Floor near the Destination	0.824	0.582	1.0	1.0
	4	Hook Top View	0.698	1.000	0.5	1.0
	5	Left Joystick	1.000	0.505	0.5	0.5
	6	Right Joystick	0.736	0.557	0.5	0.5
	7	Messages (UI)	0.487	0.302	1.0	1.0
	8	Signaller View	0.678	0.780	1.0	1.0
	9	Crane Side View	0.491	0.522	0.3	0.1
	10	Time (UI)	0.266	0.680	1.0	0.1

p	$AOI\ Index$	AOI	S	E	Ex	V
	11	Walkie Talkie	0.437	0.739	0.5	0.5

3.5 Model Fitting

To empirically estimate the weighting coefficients of the SEEV model, we implemented a hierarchical data processing framework. This framework transforms raw, high-frequency gaze logs into a structured regression dataset, enabling the systematic identification of session-specific visual attention strategies.

3.5.1 Data Aggregation and PDT Calculation

The primary unit of analysis was defined as a valid operational phase p completed by a participant k during an experimental session j. For each such instance, we extracted two temporal metrics: the total duration of the phase ($\sum_{k=1}^{N} T_{p,j,k}$) and the cumulative dwell time on each specific AOI i ($\sum_{k=1}^{N} t_{i,p,j,k}$).

To derive a model that captures cohort-level visual attention strategies rather than individual behaviors, data were aggregated across the full participant sample ($N = 40$) for each distinct session and phase. Specifically, for a given session $j \in \{1, 2, 3\}$ and operational phase $p \in$ {Prep, Lift, Land}, the aggregate PDT for an AOI i was calculated as:

$$PDT_{i,p,j} = \frac{\sum_{k=1}^{N} t_{i,p,j,k}}{\sum_{k=1}^{N} T_{p,j,k}} \quad (2)$$

where $t_{i,p,j,k}$ denotes the fixation time allocated to AOI i by participant k during phase p in session j, and $T_{p,j,k}$ represents the total phase duration. This process yields a normalized measure of group-level visual attention allocation within a specific context of participant k, phase p, and session j, providing a reliable empirical basis for subsequent regression-based estimation of SEEV model parameters.

3.5.2 Regression Dataset Construction

Following aggregation, we constructed a structured dataset to facilitate multiple linear regression analysis. Each observation corresponds to a specific AOI within a given operational phase and experimental session. Each data point is characterized by its empirically observed visual attention allocation, measured as $PDT_{i,p,j}$ and its associated theoretical predictors ($S_{i,p}, E_{i,p}, Ex_{i,p}, V_{i,p}$).

Recognizing that visual attention strategies may evolve as operators gain develop proficiency, a single global model was not suitable for this study. Instead, the dataset was stratified by session, and separate regression models were estimated to obtain session-specific weighting coefficients. For each experimental session j, we fit the following linear model:

$$PDT_{i,p,j} = \beta_{S,j} \cdot S_{i,p} - \beta_{E,j} \cdot E_{i,p} + \beta_{Ex,j} \cdot Ex_{i,p} + \beta_{V,j} \cdot V_{i,p} + \epsilon_j \quad (3)$$

In this formulation, the SEEV parameters (S, E, Ex, V) serve as independent variables and are defined as phase-dependent but session-invariant, according to Table 1, while the aggregated PDT serves as the dependent variable. The error term ϵ_j captures unexplained variance at the session level. This approach yields a session-specific coefficient vector $\beta_j = [\beta_{S,j}, \beta_{E,j}, \beta_{Ex,j}, \beta_{V,j}]$, which quantifies the relative influence of Salience, Effort, Expectancy, and Value on the operators' visual attention during each stage of training. Model parameters were estimated using Ordinary Least Squares (OLS), minimizing the residual sum of squares.

4 Results

4.1 Model Fitting Overview

Multiple linear regression analyses were performed to derive the SEEV weighting coefficients (β) for each experimental session. Table 4 summarizes the resulting goodness-of-fit metrics and parameter estimates across the training timeline. Additionally, Figure 2 visualizes the correlation between model-predicted and empirically observed PDT, enabling a comparative assessment of model fit between the initial training stage (i.e., Session 1) and the final training stage (i.e., Session 3).

Table 4. Regression Model Performance and SEEV Coefficients Across Training Sessions

Metric	Parameter / Statistic	Session 1 (Novice)	Session 2 (Intermediate)	Session 3 (Proficient)
Goodness-of-Fit	Coefficient of Determination (R^2)	0.4594	0.5939	0.6641
	Pearson r	0.6778	0.7707	0.8149
	Mean Absolute Error (MAE)	0.0660	0.0658	0.0592

	Root Mean Square Error (RMSE)	0.0817	0.0766	0.0691
SEEV Coefficients	Salience (β_S)	0.0875	0.0816	0.0830
	Effort (β_E)	-0.0955	-0.1426	-0.1699
	Expectancy (β_{Ex})	0.1240	0.1313	0.1390
	Value (β_V)	0.1102	0.1492	0.1511
	Intercept (ϵ)	-0.0261	-0.0176	-0.0066

4.2 Evolution of Model Predictive Power

The regression diagnostics indicate a monotonic improvement in the SEEV model's explanatory power as the operators gained experience, as evidenced by significant increases in the coefficient of determination (R^2) and Pearson correlation coefficient (r) from Session 1 to Session 3. As visually corroborated by Figure 2, the tighter clustering of data points around the ideal fit line ($y = x$) compared to the Novice Group (Session 1) reflects a transition from dispersed, stochastic visual search behavior to a systematic, model-predictable strategy perfectly aligned with the SEEV framework..

4.3 Optimization of Cognitive Control Parameters

The session-specific regression coefficients (β) reveal a distinct restructuring of the operators' attentional priorities. The persistently low influence of Salience (β_S) across all sessions suggests that that the safety-critical context compels operators to suppress visually salient but task-irrelevant stimuli early on.

In contrast, the top-down cognitive factors, i.e., Value (β_V) and Expectancy (β_{Ex}), exhibited a marked and monotonic increase through training. The increased weighting of Value and Expectancy, whichindicates a progressive shift toward goal-directed visual scanning, in which attentional allocation is increasingly driven by the strategic importance and anticipated information bandwidth of the AOIs rather than their perceptual salience. Such a shift is consistent with the theoretical foundations of the simplified Expectancy-Value (EV) model, which has been widely used to characterize expert attentional behavior.

Furthermore, the coefficient for Effort (β_E) demonstrated the largest relative change, increasing in magnitude from -0.0955 in Session 1 to -0.1699 in Session 3, implying heightened sensitivity to the physical cost of gaze redirection as proficiency develops. Two potential interpretations may explain this effect:

- Intrinsic Efficiency Strategy: Proficient operators may scan paths to minimize head rotation while maintaining situational awareness.
- Spatial Collinearity: Alternatively, this trend may reflect a spatial collinearity inherent in the task layout. High-value AOIs typically reside in the central (low-effort) field of view, penalizing peripheral glances due to a lack of necessity rather than deliberate avoidance.

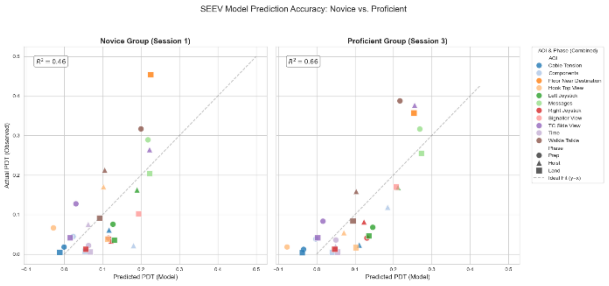

Figure 2. SEEV Model Prediction Accuracy

5 Conclusion

This study introduces a quantitative framework for assessing crane operators' visual attention mechanisms by integrating high-fidelity VR simulation with the SEEV model. Through a longitudinal restart-on-error training protocol, the proposed approach moves beyond traditional outcome-based performance metrics to capture the underlying cognitive development of novice operators. The model's improved goodness-of-fit (R^2 increasing from 0.46 to 0.66) confirms that proficiency drives a transition from stochastic visual search to a systematic, model-predictable scanning strategy.

Multiple linear regression reveals a marked restructuring of attentional priorities through VR training. Specifically, the persistent suppression of stimulus-driven Salience, coupled with the progressive amplification of goal-directed Value and Expectancy, indicates a decisive transition in visual attention allocation. This finding provides empirical support for the simplified EV model as a descriptor of expert attentional behavior [14], suggesting that proficient operators prioritize information based on strategic relevance and anticipated information rather than perceptual prominence. In addition, the increasingly negative weighting of Effort indicates that expert-like pattern involves the internalization of an efficiency-oriented attention strategy, whereby operators minimize biomechanical gaze costs while maintaining situational

awareness. Such a "maximum information, minimum effort" approach is likely critical for sustaining attention during complex and prolonged lifting operations.

Practically, the derived weighting coefficients provide a quantitative cognitive benchmark for assessing operator competency in safety-critical operations. By evaluating the convergence of a trainee's specific attentional weights—particularly the ratio of top-down (Value, Expectancy) to bottom-up (Salience, Effort) factors—toward the expert-like norms, instructors can identify specific cognitive deficiencies, such as excessive reliance on visually salient distractors or inefficient gaze mechanics, that are not detectable through traditional performance logs (e.g., completion time). Moreover, this approach contributes to designing adaptive VR training systems capable of fostering robust attentional schemas required for safe and efficient crane operations in complex construction environments. However, these contributions must be considered alongside several limitations. Firstly, the sample consisted of 40 university students whose educational backgrounds do not fully align with the demographics of actual construction workers, and the absence of a true expert dataset limits the depth of the baseline comparison. Furthermore, the assignment of Expectancy and Value parameters relied heavily on expert elicitation, introducing potential evaluator bias that might influence the regression outcomes by skewing the theoretical weightings.

To address these limitations, future research must enhance model robustness and ecological validity. To mitigate evaluator bias in parameter assignment, studies should incorporate objective neurophysiological metrics and employ the Delphi method—a structured technique utilizing iterative, multi-round expert surveys to achieve a stabilized, anonymous consensus. Additionally, longitudinal transfer learning studies are vital to confirm whether the expert-like visual patterns acquired in VR successfully translate to real-world physical operations.

6 Acknowledgements

The authors would like to thank the Research Grants Council, Hong Kong, for its funding support under the General Research Fund (PolyU 15223322).

References

[1] R. Beaulieu and C. Freeman. *Crane-related deaths in construction and recommendations for their prevention*. The Center for Construction Research and Training (CPWR), Silver Spring, MD, 2009.

[2] R. Sacks, A. Perlman, and R. Barak. Construction safety training using immersive virtual reality. *Construction Management and Economics*, 31(9):1005–1017, 2013.

[3] X. Li, W. Yi, H. L. Chi, X. Wang, and A. P. Chan. A critical review of virtual and augmented reality (VR/AR) applications in construction safety. *Automation in Construction*, 86:150–162, 2018.

[4] J. W. Yoon and J. S. Park. Understanding VR-based construction safety training effectiveness: The role of telepresence, risk perception, and training satisfaction. *Applied Sciences*, 13(3):1709, 2023.

[5] C. D. Wickens, J. Goh, J. Helleberg, W. J. Horrey, and D. A. Talleur. Attentional models of multitask pilot performance using advanced display technology. *Human Factors*, 45(3):360–380, 2003.

[6] P. Wang, P. Wu, J. Wang, H. L. Chi, and X. Wang. A critical review of the use of virtual reality in construction engineering education and training. *International Journal of Environmental Research and Public Health*, 15(6):1204, 2018.

[7] S. Hasanzadeh, B. Esmaeili, and M. D. Dodd. Measuring the impacts of safety knowledge on construction workers' attentional allocation and hazard detection using remote eye-tracking technology. Journal of Management in Engineering, 33(5):04017024, 2017. DOI: 10.1061/(ASCE)ME.1943-5479.0000526.

[8] I. Jeelani, K. Han, and A. Albert. Automating and scaling personalized safety training using eye-tracking data. *Automation in Construction*, 93:63–77, 2018.

[9] B. Cheng, X. Luo, X. Mei, H. Chen, and J. Huang. A Systematic Review of Eye-Tracking Studies of Construction Safety. *Front. Neurosci*., 16:891725, 2022. DOI: 10.3389/fnins.2022.891725.

[10] P. Kasarskis, J. Stehwien, J. Hickox, A. Aretz, and C. Wickens. Comparison of expert and novice scan behaviors during VFR flight. *In Proceedings of the 11th International Symposium on Aviation Psychology*, Columbus, OH, USA, 2001.

[11] L. Itti and C. Koch. Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203, 2001.

[12] W. J. Horrey, C. D. Wickens, and K. P. Consalus. Modeling drivers' visual attention allocation while interacting with in-vehicle technologies. *Journal of Experimental Psychology: Applied*, 12(2):67–78, 2006. DOI:10.1037/1076-898X.12.2.67.

[13] J. Wu, W. Kang, H. Tang, Y. Hong, and Y. Yan. On the faithfulness of vision transformer explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10936–10945, 2024

[14] C. D. Wickens and J. S. McCarley. *Applied Attention Theory*. CRC Press, Boca Raton, FL, 2008