IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 43, No. 3, June, 2026, pp. 1419-1433
Journal homepage: http://iieta.org/journals/ts

Screen Image Understanding, Interface Element Detection, and Intelligent Analysis of Visual Operational Behaviors for Adult Digital Literacy Education

Linping Han, Lin Chen*

School of Continuing Education, Baoding Open University, Baoding 071000, China

Corresponding Author Email: chen_lin770617@163.com

https://doi.org/10.18280/ts.430326

Received: 5 November 2025
Revised: 2 May 2026
Accepted: 21 May 2026
Available online: 30 June 2026

Keywords:
screen image understanding, interface element detection, hierarchical semantic graph, operational intention reasoning, digital literacy assessment

ABSTRACT

Accurate and objective digital literacy assessment is essential for advancing high-quality adult digital education. Existing methods rely primarily on questionnaires and manual evaluations and have several limitations. Moreover, it is difficult for current intelligent graphical user interface understanding techniques to interpret human cognition and quantitatively assess digital literacy. A four-tiered intelligent analytical framework was constructed, encompassing perception, modeling, reasoning, and evaluation, for fine-grained parsing of adult screen operation behaviors and automated quantitative assessment of digital literacy. At the perception tier, guided by regularities in adult interaction behavior, an interface element detection mechanism with adaptive heatmap guidance was established. A temporal decay accumulation strategy was employed to reinforce feature learning in high-frequency interaction regions, significantly enhancing detection accuracy for small-scale interface elements. At the modeling tier, a hierarchical screen semantic graph was constructed, in which a graph attention network was leveraged to mine spatial affiliation relationships and functional semantic attributes among interface elements, facilitating a hierarchical transition from visual feature perception to interface functionality cognition. At the reasoning tier, dynamic state transition characteristics of interfaces were integrated with temporal contextual information to build a multi-level operational reasoning architecture, enabling precise operational behavior recognition and high-level intention inference. At the evaluation tier, a multidimensional quantitative literacy indicator system was developed, and a visual feedback overlay technique was incorporated to form a complete closed loop from behavioral analysis to instructional intervention. A dedicated screen image understanding dataset was constructed for adult digital literacy assessment scenarios. Systematic experimental results demonstrated that the proposed method achieved state-of-the-art performance in interface element detection, semantic role classification, and operational intention reasoning tasks. The Pearson correlation coefficient between the automated assessment scores and expert ratings reached 0.882. This approach effectively addresses the technical deficiencies of traditional assessment methods and provides reliable technical support and a feasible practical paradigm for intelligent adult digital literacy evaluation and fine-grained instructional intervention.

1. INTRODUCTION

Digital literacy constitutes a foundational competency for citizens engaged in digitized production and daily life within modern society, and also serves as a core educational objective within adult lifelong learning systems [1, 2]. As defined by the United Nations Educational, Scientific and Cultural Organization (UNESCO), digital literacy encompasses the integrated capacity of individuals to perform information acquisition, resource integration, interactive communication, and content creation through digital technologies, with its developmental level directly determining adult occupational competitiveness and the degree of societal digital participation [3, 4]. Given that graphical user interfaces have become the predominant interaction medium for diverse software applications and intelligent devices, adult digital behaviors are invariably executed through interface operations, rendering interface operational behavior analysis a fundamental basis for quantitatively assessing digital literacy and identifying cognitive deficiencies [5, 6]. Current adult digital literacy assessment systems remain predominantly reliant on traditional subjective evaluation approaches, primarily employing methods such as questionnaires, standardized testing, and manual observation-based scoring, which are generally characterized by pronounced evaluator subjectivity, coarse-grained behavioral analysis, an inability to capture micro-level operational deviations and cognitive deficits, and significant latency in assessment outcomes that precludes real-time support for personalized instructional optimization [7, 8]. In recent years, substantial progress has been achieved in the field of automated graphical user interface understanding through computer vision techniques, with numerous cutting-

edge studies successively demonstrating foundational functionalities, including structured interface element parsing, screen operation recognition, and intelligent interpretation of interface behaviors, thereby establishing a technical basis for intelligent analysis of screen operation data [9, 10]. Nevertheless, the existing technical paradigm remains predominantly oriented toward machine-to-machine intelligent interaction and software automated testing scenarios, and has not yet been adapted to the core assessment requirements of adult education, being incapable of achieving deep cognitive interpretation or precise quantification of human operational behaviors. A dedicated intelligent analytical framework, specifically tailored to adult operational characteristics and educational evaluation contexts, is therefore urgently required [11, 12].

Although current graphical user interface visual understanding techniques have achieved mature capabilities in interface parsing and superficial behavior recognition, a fundamental misalignment exists between their technical paradigm and the core requirements of adult digital literacy assessment, with systematic research deficiencies evident across the perceptual, structural, and reasoning dimensions. At the visual perception tier, the training data for existing interface detection models are predominantly derived from standardized operational data produced by professional developers or artificially synthesized simulation data, with model optimization objectives focused on detection accuracy in general-purpose scenarios, while the unique interaction behavioral regularities of adult learners are entirely disregarded [13, 14]. In comparison with professional users, adult learners exhibit an overall slower operational tempo, with cursor movement trajectories characterized by pronounced hesitation and jitter, and click landing points demonstrating significant inter-individual variability. These distinctive behavioral characteristics result in substantial performance degradation of existing models when applied to real adult operational data, with the missed detection of small-sized interactive controls being particularly prominent, thereby failing to satisfy the fundamental prerequisite for fine-grained behavioral analysis [15, 16]. At the interface structural modeling tier, most existing studies organize detected interface elements into one-dimensional flattened sequences or simplistic tree structures, which only superficially represent the spatial nesting hierarchies of elements without being capable of mining the inherent functional semantic systems of interfaces. The core of digital literacy assessment lies in evaluating learners' cognitive understanding of interface functionalities and their ability to execute precise operations based on the task-functional attributes of elements. However, existing structural representation methods lack semantic role classification and functional association modeling, thus precluding the deep progression from visual element recognition to interface functionality cognition [14, 17]. At the behavioral cognitive reasoning tier, existing techniques are only capable of classifying and recognizing superficial interaction actions, being able to determine the basic operation types executed by users, yet unable to infer the high-level cognitive intentions underlying those operations [18, 19]. Learners' digital literacy levels are manifested in task logic planning and problem-solving capabilities, which necessitate comprehensive judgments that integrate interface functional contexts, pre- and post-operation interface state transitions, and temporal causal logic of sequential operations. To date, no cognitive reasoning mechanism incorporating multi-dimensional fusion has been established in existing research, rendering it incapable of parsing learners' deep-seated thought processes or supporting high-precision, traceable quantitative digital literacy assessment [20, 21].

To address the aforementioned technical limitations and research gaps, a four-tiered intelligent analytical framework for screen image understanding and operational behavior analysis—encompassing perception, modeling, reasoning, and evaluation—is constructed, thereby achieving technical innovation and systematic refinement for adult digital literacy assessment scenarios. By incorporating the distinctive interaction behavioral characteristics of adult learners, an interface element detection network with adaptive heatmap guidance is designed, which effectively enhances the detection performance of small-scale interface elements within complex operational contexts. A hierarchical screen semantic graph is constructed through the graph attention mechanism, enabling joint representation of interface spatial structures and functional semantics. A multi-stage operational intention reasoning pipeline is established, integrating interface state transitions and temporal contextual information to accomplish inference of learners' cognitive behaviors. Simultaneously, a multidimensional quantitative digital literacy evaluation system and a visual feedback mechanism are developed, and a dedicated adult screen operation dataset is independently constructed for model validation, forming a complete technical closed loop from screen data acquisition to intelligent educational intervention.

The organizational structure of the subsequent chapters is clearly delineated. Chapter 2 elaborates in detail on the overall architecture of the proposed four-tiered framework and the technical principles of each core module and describes the construction pipeline of the self-developed dataset, experimental parameter configurations, and specific experimental protocols. Chapter 3 systematically presents the experimental results, with the effectiveness of each module and the superiority of the proposed approach being validated through comparative and ablation experiments. Chapter 4 summarizes the research contributions, analyzes the limitations of the current methodology, and provides perspectives on future research directions.

2. METHODOLOGY

To address the critical challenges in adult digital literacy assessment scenarios—namely, poor adaptability of interface detection, absence of structured interface semantic representation, and insufficient cognitive reasoning capabilities for operations—a four-tiered end-to-end visual intelligence analytical architecture, comprising perception, modeling, reasoning, and evaluation, is established to enable the precise transformation of screen operation visual data into quantitative digital literacy indicators. The proposed approach takes as fundamental inputs the continuous screen recording video sequences and synchronized mouse trajectory data of adult learners, accepting time-series screen image frames at a resolution of 1920 × 1080, and progressively accomplishes visual feature extraction, structural semantic modeling, cognitive behavioral reasoning, and quantitative literacy evaluation through hierarchically organized functional modules. At the perception tier, focused on the differentiated interaction behavioral characteristics of adult users, an adaptive interface element detection mechanism is constructed,

in which feature learning in high-frequency interaction regions is reinforced, thereby effectively enhancing the detection accuracy and robustness of multi-scale interface elements within complex real-world operational contexts. At the modeling tier, relying on the interface element information acquired through detection, the spatial affiliation relationships and functional semantic attributes among interface units are mined, enabling the structured organization and high-order semantic representation of discrete visual elements, which compensates for the deficiency of existing methods in interface functionality cognition. At the reasoning tier, dynamic interface state transition characteristics are integrated with temporal contextual regularities of sequential operations to establish a complete reasoning logic that ascends from basic operation recognition to high-level behavioral intention inference, thereby achieving deep interpretation of learners' cognitive logic in digital operations. At the evaluation tier, the visual detection, structural modeling, and behavioral reasoning results from the preceding modules are synthesized to construct a multidimensional quantitative digital literacy evaluation system and a visual feedback mechanism, forming a closed-loop technical architecture that encompasses data acquisition, intelligent analysis, literacy assessment, and instructional intervention, and providing systematic algorithmic support for automated, fine-grained, and objective adult digital literacy assessment.

2.1 Adaptive interface element detection network

To address the characteristics of adult learners' interface interaction behaviors that differ from those of professional users, as well as the issues of high miss-detection rates for small-scale interface elements and weak feature adaptability exhibited by conventional detection models in real adult operational scenarios, an adaptive interface element detection network is constructed based on YOLO11-Large. The architecture of the adaptive interface element detection network is illustrated in Figure 1. Four groups of multi-scale feature maps are extracted by the network backbone, corresponding to down-sampling rates of fourfold, eightfold, sixteenfold, and thirty-twofold, respectively, thereby covering multi-dimensional feature information of screen interfaces ranging from micro-level textures to global semantics. To address the problem of small-scale user interface feature vanishing within deep networks, a bidirectional feature pyramid structure is introduced at the feature fusion stage. Through a weighted bidirectional cross-scale connection mechanism, adaptive multi-level feature propagation and reconstruction are accomplished, enabling efficient fusion of shallow fine-grained textural features and deep high-level semantic features, and effectively enhancing the foundational perceptual capability of the model for miniature interactive controls. The reconstructed multi-scale feature representations are subsequently utilized for feature aggregation and detection inference, with the corresponding computational formulation given as follows:

$$\widetilde{F}_l = BiFPN(F_1, F_2, F_3, F_4) \tag{1}$$

where, F_1, F_2, F_3, and F_4 denote the four-layer raw multi-scale feature maps output by the backbone network, and $\widetilde{F}_l$ represents the enhanced feature maps obtained through bidirectional weighted feature fusion.

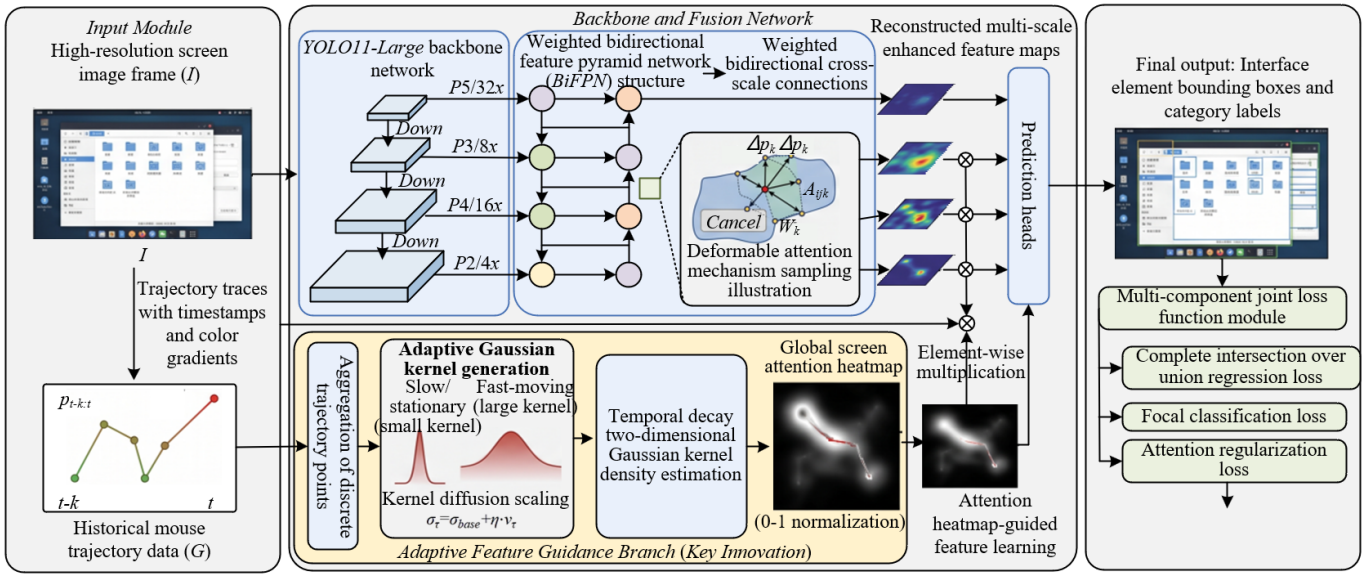


Figure 1. Architecture of the adaptive interface element detection network

To accommodate the visual characteristics of graphical interface elements—namely, their variable scales, irregular morphologies, and dense arrangements—a deformable attention mechanism is employed to accomplish adaptive aggregation of multi-scale features, thereby overcoming the limitations of fixed receptive field-based feature fusion. Global contextual information of images is extracted through global average pooling and a multi-layer perceptron, and dynamic fusion weights for features at each scale are generated via normalization processing, enabling differentiated weighted aggregation of features across hierarchical levels. During feature sampling, multi-point offset sampling is performed around reference positions, and learnable attention weights along with feature projection matrices are leveraged to adaptively capture the local structural characteristics of interface controls, precisely accommodating the morphological distribution features of irregular user interface elements. The overall computational process of multi-scale feature fusion and the deformable attention sampling formulation can be expressed as:

$$F_{fusion}=\sum_{l=1}^{4}\alpha_l \cdot DeformAttn(\widetilde{F}_l,P_l) \tag{2}$$

$$DeformAttn(\widetilde{F}_l,P_l)=\sum_{k=1}^{K}W_k\,\widetilde{F}_l(p+p_{ref}+\Delta p_{l,k})\cdot\Delta m_{l,k} \tag{3}$$

where, α_l denotes the dynamic fusion weight for each scale's enhanced features, adaptively generated from the global context of the image; P_l represents the positional encodings corresponding to features at each hierarchy; p denotes the feature query position, and p_{ref} is the local reference position; $\Delta p_{l,k}$ denotes the learnable sampling offsets, designed to accommodate irregular control morphologies; $\Delta m_{l,k}$ is the spatial attention weight scalar; W_k is the feature projection matrix, employed for dimensional alignment and feature refinement; and K is the number of sampling points per reference point. Based on the aggregated refined features, the network is capable of outputting bounding box coordinates, category labels, and interactivity confidence scores for interface elements, thereby comprehensively characterizing the visual compositional information of screen interfaces.

To align with the interaction behavioral regularities of adult learners—characterized by cursor hesitation, slow movement, and dispersed click landing points—a temporal decay heatmap guidance strategy is designed based on time-series mouse trajectory data, enabling behavior-adaptive optimization of the detection model. A fixed temporal sliding window is established to integrate cursor trajectory information across consecutive frames, through which a temporal decay factor is employed to attenuate the interfering effects of long-past trajectories while reinforcing the feature weights of interactions proximate to the current moment, and a global screen attention heatmap is generated via two-dimensional Gaussian kernel density estimation. The diffusion scale of the Gaussian kernel is dynamically adjusted by the real-time cursor movement speed: under slow or stationary conditions, the kernel size is contracted to focus on precise interaction regions, whereas under fast-moving conditions, the kernel size is expanded to accommodate the attentional dispersion characteristics of behavior, thereby faithfully reflecting real adult operational habits. Finally, the normalized attention heatmap is fused with the aggregated features through element-wise weighted fusion, guiding the model to preferentially learn feature representations from high-frequency user interaction regions while suppressing feature responses in ineffective background areas. The computational formulations for heatmap generation and feature guidance are given as follows:

$$A_{heat}(u,v)=\sum_{\tau=t-\Delta T}^{t}\gamma^{t-\tau}\cdot N\big((u,v);p_\tau^{cursor},\sigma_\tau^2 I_2\big) \tag{4}$$

$$\sigma_\tau=\sigma_{base}+\eta\cdot v_\tau \tag{5}$$

$$F_{guided}=F_{fusion}\odot Norm(A_{heat}) \tag{6}$$

where, ΔT denotes the temporal sliding window length; γ is the temporal decay factor used to regulate the weight attenuation rate of historical trajectories; u,v are the spatial coordinates of the screen image; p_τ^{cursor} is the cursor coordinate at time τ; N denotes the two-dimensional Gaussian distribution function; I_2 is the second-order identity matrix; σ_τ is the adaptive Gaussian kernel diffusion scale; σ_{base} is the base kernel size; η is the velocity adjustment coefficient; v_τ is the instantaneous cursor movement speed; $\odot$ denotes element-wise multiplication of features; and *Norm* denotes the min-max normalization operation.

A multi-component joint loss function is constructed for holistic optimization of the detection network, coordinating the multi-dimensional training objectives of interface element localization, classification, and attention feature constraints to accommodate the complex characteristics of adult interface detection tasks. The overall loss function is composed of bounding box regression loss, category classification loss, and attention regularization loss, with the specific formulation given as:

$$L_{det}=L_{bbox}+L_{cls}+\lambda\cdot L_{attn} \tag{7}$$

For bounding box regression, the complete intersection over union loss is adopted, which comprehensively accounts for the overlap area, center distance, and aspect ratio discrepancy between predicted and ground-truth boxes, thereby enhancing localization accuracy for densely arranged interface elements. For the classification task, the Focal Loss function is introduced to suppress the training weights of numerous simple background samples while focusing optimization on small-scale and rare controls, thereby alleviating the class imbalance problem among interface elements. The attention regularization loss, through constraining feature responses in non-interactive regions, aligns with the heatmap-guided feature learning mechanism and further reinforces the model's detection capability for effective interaction regions. The three loss components are jointly optimized through gradient coordination via fixed balancing weights, ensuring stability and precision of the network in adult screen interface detection tasks. The specific computational rules for each sub-loss are given as follows:

$$L_{CIoU}=1-IoU(B,B^{gt})+\frac{\rho^2(b,b^{gt})}{c^2}+\beta\cdot v \tag{8}$$

$$L_{cls}=-\sum_{i=1}^{N}\alpha_c\,(1-p_i(c))^{\kappa}\log p_i(c) \tag{9}$$

$$L_{attn}=\left\|(1-Norm(A_{heat}))\odot F_{fusion}\right\|_2 \tag{10}$$

where, λ denotes the attention loss balancing weight; B and B^{gt} represent the predicted box and the ground-truth annotation box, respectively; b and b^{gt} are the center coordinates of the corresponding boxes; ρ denotes the Euclidean distance function; c is the diagonal length of the minimum enclosing bounding box; v is the aspect ratio consistency measure; β is the aspect ratio trade-off coefficient; N is the total number of interface elements in a single frame; α_c is the class balancing weight; $p_i(c)$ is the model's predicted probability for the i-th element belonging to class c; κ is the focusing parameter, employed to suppress gradients from easily classified samples; and 1 denotes the all-ones matrix.

2.2 Hierarchical screen semantic graph

Interface element detection alone yields only discrete control entities and visual features within screen images, lacking a structured representation of the overall spatial topology and functional associations of the interface, and thus

cannot adequately support cognitive-level interpretation of user operational behaviors. To compensate for the limitations of flattened element representation, a hierarchical screen semantic graph model is constructed in this section, transforming scattered interface controls into a structured topological architecture endowed with spatial hierarchies and functional semantic associations. The detected interface elements are taken as fundamental nodes, and spatial nesting relationships along with local neighborhood association features are incorporated, enabling the screen interface to be elevated from a mere visual collection to a functionally structured system, thereby providing semantic-level support for subsequent operational intention reasoning and fine-grained digital literacy assessment. The construction and classification workflow of the hierarchical screen semantic graph is illustrated in Figure 2.

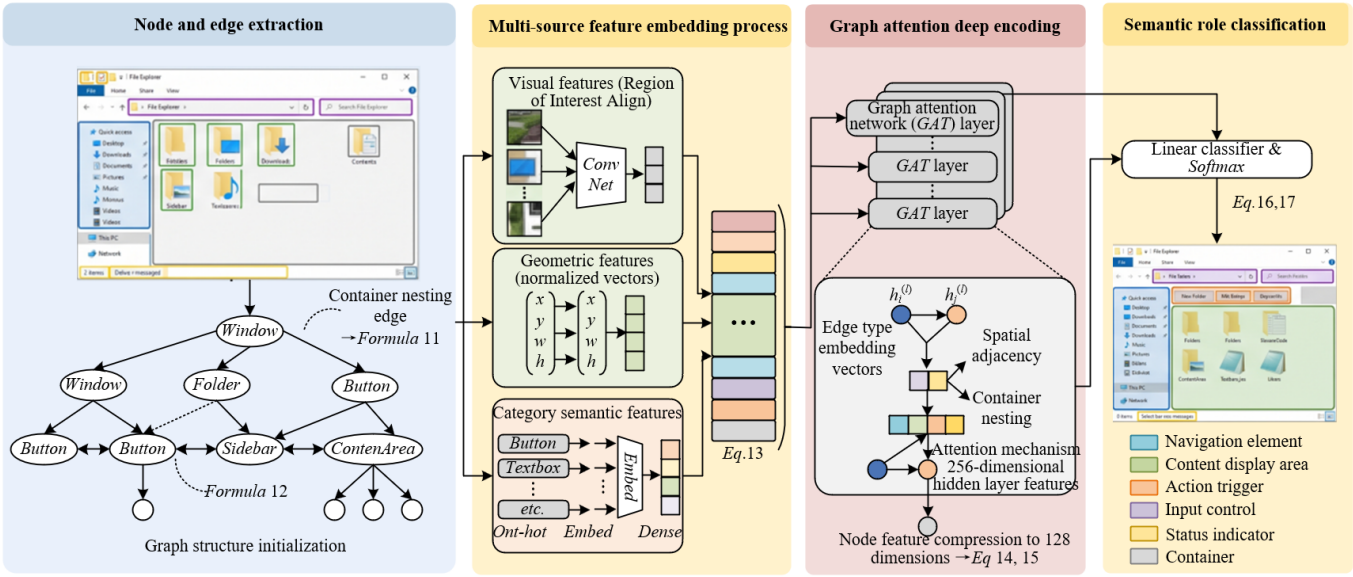


Figure 2. Construction and classification workflow of the hierarchical screen semantic graph

To construct a robust interface topological structure, a dual-edge association mechanism is designed by integrating the spatial distribution characteristics of interface elements, effectively circumventing the problem of structural hierarchy fragmentation caused by minor detection box errors. All detected interface controls are defined as a set of graph nodes, with each node carrying dedicated spatial coordinates, category attributes, and visual feature information. To address the prevalent container nesting structures within interfaces, a soft containment score is introduced to quantify the affiliation relationships between nodes, thereby mitigating the structural noise induced by hard-threshold decisions. When the computed containment degree between elements exceeds a preset threshold, a directed container association edge is established to characterize the hierarchical nesting relationships of the interface. The specific computational formulation is given as:

$$s_{ij}^{contain}=IoU(b_i,b_j)\cdot 1\begin{bmatrix}x_i\le x_j,y_i\le y_j,x_i+w_i\ge x_j\\+w_j,y_i+h_i\ge y_j+h_j\end{bmatrix} \tag{11}$$

For densely arranged sibling interface elements without nesting associations, a spatial proximity decision rule is further introduced to construct adjacency edges, where the degree of element association is measured through the relative ratio of node center distances to control dimensions, thereby comprehensively covering the topological association relationships of all elements within the interface. The corresponding spatial edge decision rule is formulated as follows:

$$e_{ij}^{spatial}=1\left[\frac{\|center_i-center_j\|_2}{\min(w_i,h_i)+min(w_j,h_j)}<\gamma\right] \tag{12}$$

where, x_i, y_i, w_i, and h_i denote the horizontal coordinate, vertical coordinate, width, and height of the bounding box of the i-th interface element, respectively; $center_i$ denotes the center coordinates of the element box; γ is the spatial association threshold, with a fixed value of 1.5, which is adapted to the typical arrangement regularities of conventional interface controls; and 1 denotes the indicator function, which outputs 1 when the condition is satisfied and 0 otherwise. By integrating the two types of association edges, a complete hierarchical screen semantic graph topological structure is formed.

To achieve multi-dimensional feature representation of graph nodes, three types of information—visual semantics, geometric structure, and category semantics—are fused to accomplish node feature initialization, thereby constructing high-dimensional and refined initial embedding features for nodes. Control-specific visual features are extracted from the fused guided feature maps through region feature alignment, characterizing the appearance and textural information of controls, while the geometric parameters of interface elements, including coordinates, width, and height, are normalized to encode the spatial structural attributes of interface controls. Additionally, vector embedding encoding is performed on element categories to supplement the inherent category semantic information of controls. After dimensionality transformation via a multi-layer perceptron, the three types of features are concatenated and fused to form a unified node feature of 256 dimensions. The overall initialization process is formulated as follows:

$$h_i^{(0)}=MLP\big(RoIAlign(F_{guided},b_i)\big)\oplus MLP([b_i;c_i])\oplus Embed(c_i) \tag{13}$$

where, $h_i^{(0)}$ denotes the initial feature vector of the i-th node; *RoIAlign* denotes the region feature alignment operation; b_i denotes the element bounding box features; c_i denotes the element category label; $\oplus$ denotes the vector concatenation operation; and *Embed* denotes the category embedding encoding function. The multi-source feature fusion approach simultaneously preserves the visual appearance, spatial structure, and semantic attributes of the interface, thereby providing comprehensive initial feature support for the graph model to mine deep functional associations.

A three-layer graph attention network is employed to accomplish deep encoding of the semantic graph, where adaptive attention mechanisms are leveraged to mine the differentiated semantic weights of distinct nodes and association edges, thereby achieving high-order representation of interface hierarchical semantics. Attention coefficients are dynamically computed based on the nodes' intrinsic features and neighborhood association features, while edge type embedding vectors are introduced to differentiate the semantic distinctions between container containment edges and spatial adjacency edges, enabling the model to precisely identify both the hierarchical structure and sibling-level association relationships of the interface. The iterative update rules for node features and the computational formulation for attention coefficients are expressed as:

$$h_i^{(l+1)}=\sigma\left(\sum_{j\in N(i)}\alpha_{ij}^{(l)}\,W^{(l)}h_j^{(l)}\right) \tag{14}$$

$$\alpha_{ij}=\frac{exp\left(LeakyReLU\left(a^T[Wh_i\oplus Wh_j\oplus\phi(e_{ij})]\right)\right)}{\sum_{k\in N(i)}exp\left(LeakyReLU\left(a^T[Wh_i\oplus Wh_k\oplus\phi(e_{ik})]\right)\right)} \tag{15}$$

where, $h_i^{(l)}$ denotes the feature vector of node i at the l-th network layer; σ denotes the non-linear activation function; $N(i)$ denotes the set of neighboring nodes of node i; $\alpha^{(l)}{}_{ij}$ denotes the layer-wise normalized attention coefficient; $W^{(l)}$ denotes the learnable weight matrix; a denotes the attention scoring vector; and $\phi(e_{ij})$ denotes the learnable edge type embedding features, employed to differentiate nested associations from spatial adjacency associations. The network is configured with an eight-head attention mechanism and 256-dimensional hidden layer features. Following iterative encoding, the node feature dimensions are compressed to 128 dimensions, thereby accomplishing deep condensation of structured semantic features.

Based on the encoded high-order node features, semantic role classification of interface elements is accomplished, enabling automatic partitioning of interface functional semantics. Interface elements are uniformly categorized into six functional roles: navigation elements, content display areas, action triggers, input controls, status indicators, and containers. The semantic role probability distribution for each node is output through a linear classifier and a softmax function, with the optimal label being selected as the final classification result. The specific computational formulation is given as:

$$\hat{r}_i=\underset{r\in R}{arg\,max}\,Softmax(W_r h_i^{(L)}+b_r) \tag{16}$$

To optimize the accuracy of semantic role classification, a node-level cross-entropy loss function is employed to supervise network training, constraining the deviation between node prediction labels and ground-truth functional labels on a per-node basis. The overall loss function is defined as follows:

$$L_{HSSG}=-\frac{1}{N}\sum_{i=1}^{N}\sum_{r\in \mathrm{R}}y_{i,r}\,log\,\hat{y}_{i,r} \tag{17}$$

where, R denotes the predefined set of semantic role labels; W_r and b_r denote the classification layer weight and bias parameters, respectively; $\hat{r}_i$ denotes the predicted semantic role of the node; $y_{i,r}$ denotes the one-hot label of the ground-truth semantic role; $\hat{y}_{i,r}$ denotes the model-predicted probability; and N denotes the total number of graph nodes. Through end-to-end training, the model is capable of precisely distinguishing the functional attributes of each interface element, thereby establishing a comprehensive interface functional semantic system and achieving effective transformation from screen visual information to functional cognition.

2.3 Visual operational behavior intention reasoning module

After structured modeling of interface spatial structures and functional semantics is accomplished through the hierarchical screen semantic graph, further cognitive-level interpretation of learners' dynamic interaction behaviors can be performed. Existing interface behavior analysis methods are only capable of recognizing and classifying basic interaction actions, without being able to incorporate interface functional attributes and operational temporal logic to infer the task objectives underlying user behaviors, and thus cannot adequately support the cognitive capability assessment required for digital literacy. To address this limitation, a multi-level visual operational behavior intention reasoning module is constructed, establishing a progressive reasoning pipeline composed of operation type recognition, operation target localization, and operation intention inference. Through the integration of spatiotemporal interface variation features, structured semantic information, and historical operation contexts, precise interpretation from surface-level behavioral observation to deep cognitive intention is achieved. The framework of the multi-level visual operational behavior intention reasoning architecture is illustrated in Figure 3.

Operation type recognition serves as the foundational tier of behavioral reasoning. To address the issue that single-frame images are incapable of distinguishing between similar interaction actions such as clicking and dragging, a long-short-term temporal difference mechanism is introduced to quantify the visual variation patterns of interfaces across consecutive frames. By computing the image difference features between adjacent frames and between frames with a five-frame interval, instantaneous operation actions and delayed interface feedback are captured separately, thereby effectively differentiating short-duration static interactions from sustained dynamic interactions. Consecutive three-frame images are selected as the fundamental input, and a lightweight 3D convolutional network is employed to extract spatiotemporal joint features. Furthermore, the cursor position heatmap is fused to reinforce feature responses in interaction regions, thereby achieving precise discrimination of six categories of basic interface operations. The overall recognition process is formulated as follows:

$$D_t^{short}=|I_t-I_{t-1}|,D_t^{long}=|I_t-I_{t-5}| \tag{18}$$

$$\hat{a}_t=\arg\max_{a\in A} p(a_t \mid I_{t-2:t},D_t^{short},D_t^{long},A_t^{cursor}) \tag{19}$$

where, I_t denotes the screen image at frame t; D_t^{short} denotes the short-term inter-frame difference feature, employed to characterize instantaneous interaction changes; D_t^{long} denotes the long-term inter-frame difference feature, employed to characterize delayed interface feedback; $I_{t-2:t}$ denotes the consecutive three-frame input image sequence; A denotes the predefined set of operation types; and $\hat{a}_t$ denotes the operation category at the current moment predicted by the model.

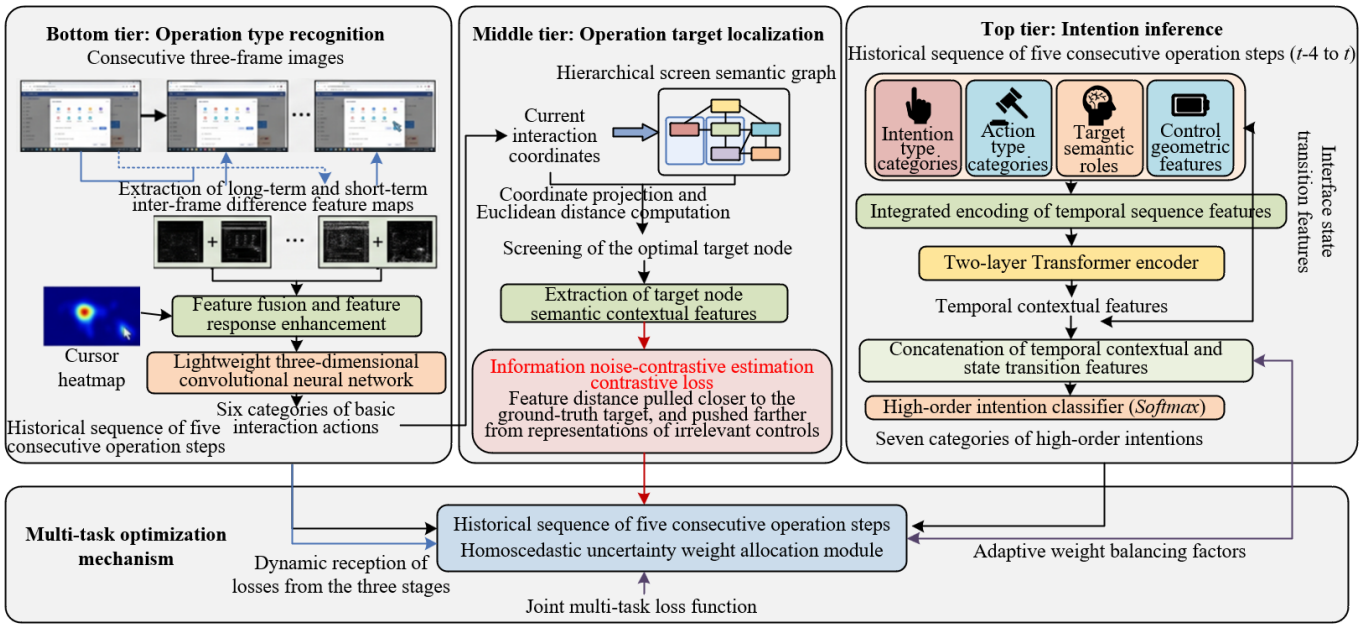


Figure 3. Framework of the multi-level visual operational behavior intention reasoning architecture

Upon clarification of the operation type, precise localization of the operation target is accomplished by integrating the structured semantic information of the interface. Different from traditional localization approaches that rely solely on geometric coordinate matching, the interaction coordinates are mapped into the node system of the hierarchical screen semantic graph, and the target control node corresponding to the current operation is identified by computing the Euclidean distance between the operation coordinates and the center positions of each interface element, thereby precisely matching the target object of the user interaction behavior. The screening rule for the target node is defined as:

$$v_{target}=\arg\min_{v_i\in V} \|p_t^{action}-center_i\|_2 \tag{20}$$

where, p_t^{action} denotes the interaction operation coordinates at time t; V denotes the set of all interface nodes in the semantic graph; and v_{target} denotes the optimally matched operation target node. By leveraging the topological constraints of the semantic graph, this localization approach effectively circumvents localization deviations caused by complex interface layouts and control occlusions. Upon completion of localization, the functional semantic roles of the target node and its neighborhood node features are simultaneously extracted, thereby providing interface semantic contextual support for subsequent high-level intention inference.

Operation intention inference constitutes the core component of behavioral cognitive analysis, through which high-level task objective reasoning is accomplished by integrating interface state transitions, individual operation attributes, and historical operational temporal logic. Visual difference features of the interface before and after operations are first extracted, where visual feedback information—including interface content updates and control state transitions—is captured via frame differencing between adjacent images, and fixed-dimensional state transition feature representations are generated through a lightweight convolutional network:

$$\Delta I_t=|I_{t+1}-I_{t-1}| \tag{21}$$

To mine the potential logical associations among sequential operations, a temporal sequence constructed from five consecutive historical operations is selected, and the action type, semantic role, and control geometric features of each operation are uniformly encoded to form a structured temporal feature vector:

$$E_\tau=Embed(a_\tau)\oplus Embed(r_{target,\tau})\oplus MLP(b_{target,\tau}) \tag{22}$$

Global modeling of the temporal sequence is performed through a two-layer Transformer encoder, capturing causal relationships and task logic among different operations, and temporal contextual features are output as follows:

$$h_{context}=Transformer(E_{t-K},\ldots,E_{t-1})\in \mathrm{R}^{256} \tag{23}$$

The temporal contextual features, interface state transition features, and target semantic role features are finally fused, and classification discrimination of seven categories of operation intentions is accomplished through a softmax function. The complete inference formulation is given as:

$$p(z_t \mid H_{t-K:t},\Delta I_t,r_{target})= Softmax\left(W_z[h_{context}\oplus f_{\Delta t}\oplus Embed(r_{target})]+b_z\right) \tag{24}$$

where, ΔI_t denotes the interface state transition feature map; E_τ denotes the single-step operation temporal embedding feature; $H_{t\text{-}K:t}$ denotes the set of historical operation sequences; K denotes the temporal context window length; $h_{context}$ denotes the global temporal contextual features; $f_{\Delta t}$ denotes the deep interface transition features; W_z and b_z denote the intention classification layer weight and bias parameters; and z_t denotes the predicted operation intention. The corresponding intention categories encompass core task scenarios of digital operations, including file management, format adjustment, information retrieval, and others, which comprehensively align with the task framework of adult digital literacy assessment.

To ensure training stability for multi-task collaborative optimization, a multi-task joint learning strategy is adopted, through which the three sub-tasks—operation recognition, target localization, and intention classification—are uniformly optimized. For the operation target localization task, the information noise-contrastive estimation contrastive loss is introduced, where the interaction position features are taken as anchors, the feature distance to the ground-truth target node is pulled closer, and the feature representations of irrelevant controls are pushed farther, thereby enhancing localization accuracy:

$$L_{target}=-\log\frac{exp(sim(e_{action},h_{target})/\tau)}{exp(sim(e_{action},h_{target})/\tau)+\sum_{v_{neg}\in V_{neg}}exp\,(sim(e_{action},h_{neg})/\tau)} \tag{25}$$

where, e_{action} denotes the interaction position feature vector; h_{target} denotes the ground-truth target node features; sim denotes the cosine similarity function; τ denotes the temperature coefficient, employed to scale the feature distribution and regulate the gradient update magnitude; and V_{neg} and h_{neg} denote the set of negative sample nodes and the corresponding negative sample features, respectively.

To address the training bias issue caused by magnitude discrepancies among multi-subtask losses, a homoscedastic uncertainty mechanism is introduced, through which the weight coefficients of each loss function are adaptively and dynamically adjusted, replacing the manual configuration of fixed weights. The overall joint loss function of the model is defined as:

$$L_{intent}=L_{action}+L_{target}+L_{intent\_cls} \tag{26}$$

$$L_{intent}=\frac{1}{2\sigma_1^2}L_{action}+\frac{1}{2\sigma_2^2}L_{target}+\frac{1}{2\sigma_3^2}L_{intent\_cls}+\log\sigma_1\,\sigma_2\sigma_3 \tag{27}$$

where, L_{action} , L_{target} , and L_{intent_cls} denote the operation recognition loss, the target localization contrastive loss, and the intention classification loss, respectively; and σ_1, σ_2, and σ_3 denote the learnable uncertainty parameters, through which the model adaptively balances the optimization weights of each subtask via backpropagation, thereby ensuring multi-task collaborative convergence and effectively enhancing the robustness and accuracy of behavioral intention reasoning in complex scenarios.

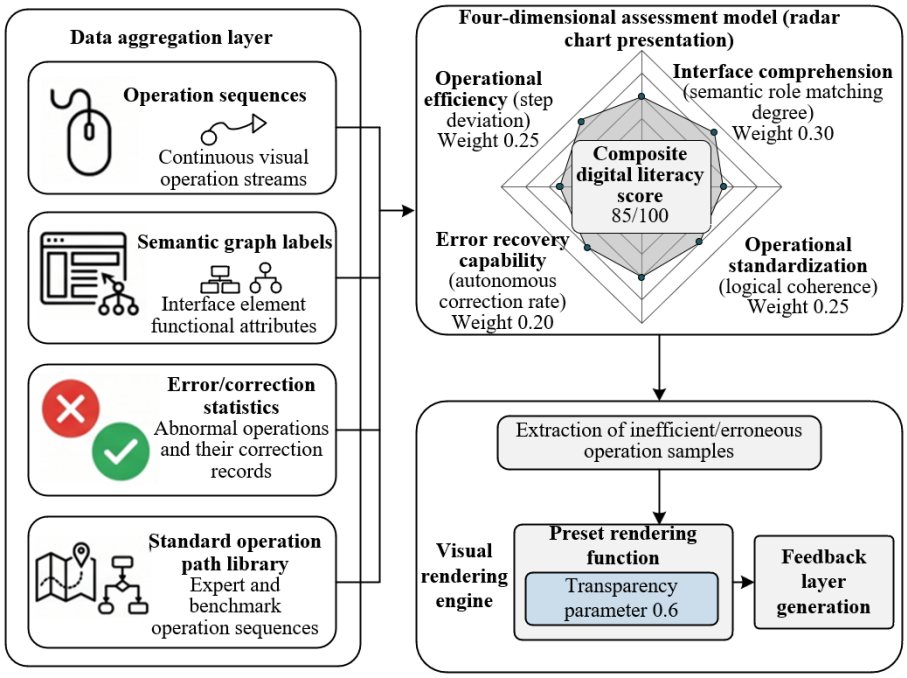


Figure 4. Four-dimensional quantitative digital literacy assessment and visual feedback system

2.4 Digital literacy visual assessment and feedback system

Based on the outputs of the aforementioned interface perception, semantic modeling, and behavioral reasoning modules, an end-to-end digital literacy visual assessment and feedback system is constructed, enabling quantitative evaluation and visual instructional guidance of adult digital operational behaviors. Existing digital literacy assessment systems predominantly rely on macro-level questionnaire scoring and manual subjective judgment, lacking the capability for quantitative decomposition of micro-level operational behaviors and deficiency traceability, and thus cannot precisely characterize learners' interface cognitive proficiency or operational shortcomings. By leveraging the

complete temporal operation sequences, interface semantic information, and behavioral intention data, a multi-dimensional quantitative assessment model is established, and visual feedback results are generated through image overlay rendering techniques, forming a complete technical closed loop from behavioral analysis, literacy scoring, to instructional intervention. The four-dimensional quantitative digital literacy assessment and visual feedback system is illustrated in Figure 4.

By leveraging the comparison relationship between learners' actual operation sequences and standardized task operation paths, a four-dimensional integrated quantitative digital literacy indicator system is established, through which comprehensive assessment is accomplished across four dimensions: operational efficiency, interface comprehension, operational standardization, and error recovery capability. Operational efficiency quantifies the degree of conciseness in task completion by learners, where the efficiency of task execution is characterized through the deviation between the actual number of operation steps and the standard path step count. The computational formula is given as follows:

$$Efficiency=1-\frac{|T-T^*|}{\max(T,T^*)} \tag{28}$$

where, T and T^* denote the length of the learner's actual operation sequence and the length of the task standard operation sequence, respectively. Interface comprehension, grounded in the functional semantic labels of the hierarchical screen semantic graph, measures the matching degree between the learner's operation targets and the standard semantic roles, directly reflecting the learner's cognitive proficiency regarding the interface functional structure. The computational formulation is as follows:

$$Understanding=\frac{1}{T}\sum_{t=1}^{T} 1\,[r_{target,t}=r^*_{target,t}] \tag{29}$$

where, $r_{target,t}$ and $r_{target,t}^*$ denote the actual target semantic role and the standard semantic role for the single operation at time, respectively. Operational standardization quantifies the degree of alignment between the overall operation flow and the software's predefined operational logic, where statistical evaluation is performed through temporal alignment comparison between operation behaviors and the standard operation specification library:

$$Normality=\frac{1}{T-1}\sum_{t=1}^{T-1} 1\,[(a_t,v_{target,t}) \text{ conforms to the predefined specification}] \tag{30}$$

Error recovery capability is employed to measure learners' autonomous correction and problem-solving abilities, where the ratio between the number of erroneous operations and the number of successfully corrected operations during the task process is statistically computed, with a small constant introduced to ensure computational stability:

$$Recovery=\frac{N_{corrected}}{N_{errors}+\epsilon} \tag{31}$$

where, N_{errors} denotes the total number of detected erroneous operations, $N_{corrected}$ denotes the number of successfully corrected erroneous operations, and ϵ is set to 10^{-6} to avoid computational anomalies caused by a zero denominator.

Based on the four individual evaluation metrics, a weighted fusion comprehensive digital literacy scoring model is constructed, where the weight for each dimension is assigned according to the competency composition characteristics of adult digital literacy, with particular emphasis placed on the central role of interface functionality cognition in literacy evaluation. The comprehensive score is computed as follows:

$$Score=\alpha\cdot Efficiency+\beta\cdot Understanding+\gamma\cdot Normality+\delta\cdot Recovery \tag{32}$$

The model weight parameters are set to α=0.25, β=0.30, γ=0.25, and δ=0.20, corresponding to the contribution weights of operational efficiency, interface comprehension, operational standardization, and error recovery capability, respectively. The highest weight is assigned to interface comprehension, which aligns with the evaluation logic in digital literacy assessment that prioritizes cognitive capability over operational form. The remaining weights are assigned in a balanced manner according to the literacy contributions of operational efficiency, operational standardization, and error correction capability, thereby ensuring the scientific rigor and rationality of the comprehensive score.

To enable the visual implementation of quantitative assessment results and instructional intervention, a screen visual feedback overlay technique is designed, through which abstract assessment data are transformed into intuitive on-screen visual prompts. All inefficient and erroneous operation samples are filtered by the system, and corresponding optimization prompts and standard operation guidelines are overlaid onto the original screen images. Multi-type visual elements are fused and overlaid through customized rendering functions, with the specific computational formulation given as:

$$I_{feedback}=I_{end}+\sum_{t\in E}\lambda_t\cdot Render(type_t,b_{target,t}) \tag{33}$$

where, I_{end} denotes the original final screen image; E denotes the index set corresponding to inefficient and erroneous operations; $type_t$ denotes the visual feedback type, which includes highlighting annotations, path guidance, and text prompts; $b_{target,t}$ denotes the interface element box corresponding to the anomalous operation; and λ_t denotes the visual overlay transparency parameter, with a fixed value of 0.6, which preserves the original interface information in its entirety while clearly presenting the learner's operational deficiencies and optimization solutions. This mechanism enables precise identification of learners' individualized operational shortcomings, providing intuitive and implementable technical support for targeted training and fine-grained instruction in adult digital literacy.

3. EXPERIMENTS

Through systematic comparative experiments, ablation experiments, and generalization experiments, the effectiveness and superiority of the proposed adaptive interface element detection network, hierarchical screen semantic graph, operation intention reasoning module, and

digital literacy assessment system were comprehensively validated. Multi-dimensional performance evaluation was conducted based on the self-constructed dedicated dataset, through which the technical gains of each core module were quantified. Through comparative analysis with mainstream state-of-the-art methods, the adaptive advantages and application value of the proposed approach in adult digital literacy intelligent assessment scenarios were substantiated.

3.1 Dataset construction

Existing publicly available graphical user interface understanding datasets are primarily oriented toward intelligent agent automatic interaction and software automated testing tasks, with annotation frameworks focused on visual interface element recognition, while lacking critical annotations for educational assessment dimensions—including adult operational behaviors, functional semantic roles, and high-level operational intentions—that are essential for training and validating quantitative digital literacy assessment models. To address this gap, a screen image understanding dataset tailored for adult digital literacy assessment, designated as DLSU, was constructed, thereby filling the research gap in graphical user interface behavioral cognition datasets within educational contexts. A total of 50 adult learners were recruited for data collection in this experiment. The age distribution of participants ranged from 25 to 55 years, with a mean age of 42.3 years and a standard deviation of 8.7 years. The digital literacy levels of the participants covered both beginner and intermediate tiers, and their occupational backgrounds encompassed multiple industries, including education, healthcare, manufacturing, and services, thereby faithfully reflecting the digital operational characteristics of typical adult users. All participants completed informed consent forms, ensuring the compliance and authenticity of the data collection process.

The task framework of the dataset comprised 12 typical digital operation tasks, categorized into four core application scenarios: document processing, web information retrieval, operating system configuration, and email management, which comprehensively cover the high-frequency operational scenarios of adult daily digital production and life. Each task was collaboratively calibrated with standard operation paths by three experts in the field of educational technology, providing an authoritative benchmark for subsequent path comparison and indicator computation in literacy assessment. Throughout the data collection process, screen recording was performed at a resolution of 1920×1080 and a frame rate of 30 frames per second, with mouse trajectory and keyboard interaction data being synchronously captured. The final dataset comprises 600 valid screen operation videos, with a cumulative video duration of approximately 120 hours, and a total of 430,000 valid annotated image frames. The annotation content encompasses four core label categories: user interface element bounding boxes and categories, interface semantic roles, operation behavior types, and high-level operation intentions, with a cumulative total of 152,000 groups of interface element box annotations.

To ensure the reliability of the annotation data, Krippendorff's Alpha coefficient was employed for inter-annotator agreement evaluation, with all annotation dimensions yielding coefficients exceeding 0.85, thereby demonstrating a high degree of consistency and trustworthiness in the annotation results. The dataset was randomly partitioned into training, validation, and test sets at a ratio of 7:1.5:1.5, ensuring balanced data distribution and satisfying the experimental requirements for model training, parameter tuning, and performance testing.

3.2 Experimental settings

All models in this study were implemented using the PyTorch 2.0 framework, with computations performed on four NVIDIA A100 40GB graphics processing units and the CUDA 11.8 computing environment. The detection network was initialized with YOLO11-Large as the backbone weights, and fine-tuned for a total of 30 epochs with a batch size of 16. The initial learning rate was set to 10^{-4}, with parameter updates performed using the AdamW optimizer and a weight decay coefficient of 5×10^{-4}. The learning rate was dynamically scheduled through a cosine annealing strategy. For the hierarchical screen semantic graph, the graph attention encoder was configured with a three-layer network structure, a hidden layer dimension of 256, 8 attention heads, and a dropout rate of 0.1 to effectively mitigate model overfitting. The Transformer encoder of the intention reasoning module comprised two network layers, with a hidden dimension of 512, a feed-forward network dimension of 2048, and 8 attention heads, balancing temporal modeling capability with computational efficiency.

Multi-dimensional quantitative metrics were employed for performance evaluation. For the interface element detection task, mAP@0.5, mAP@0.75, and mAP@0.5:0.95 were selected as core accuracy metrics, with the recall rate for small elements additionally introduced to characterize the detection performance of miniature controls. For the interface semantic role classification task, accuracy and macro-average F1-score were employed for evaluation. For the operation type recognition and operation intention reasoning tasks, recognition accuracy and macro-average accuracy were adopted as evaluation criteria, respectively. For the automated digital literacy assessment task, the Pearson correlation coefficient was employed to quantify the consistency between model-generated scores and expert manual scores, thereby validating the effectiveness of the assessment system. Comparative experiments were conducted with mainstream state-of-the-art methods in the field. For the interface element detection task, comparisons were made against OmniParser and the native YOLO11 baseline model. For the interface structural representation task, comparisons were made against Screen2AX and the flattened multilayer perceptron baseline model. For the operational behavior analysis task, comparisons were made against SeeAction and the GUI-Narrator algorithm. These comparative experiments comprehensively validated the performance advantages and scenario adaptability of the proposed approach relative to existing graphical user interface understanding techniques.

3.3 Experimental design and result analysis

3.3.1 Performance evaluation and ablation analysis of the adaptive detection module

This experiment was designed to validate the effectiveness of the multi-scale deformable attention fusion structure and the temporal decay heatmap guidance strategy in enhancing interface element detection performance. Through mainstream method comparisons and module ablation experiments, the technical gains of each innovative mechanism were quantified.

The experimental results are presented in Figure 5.

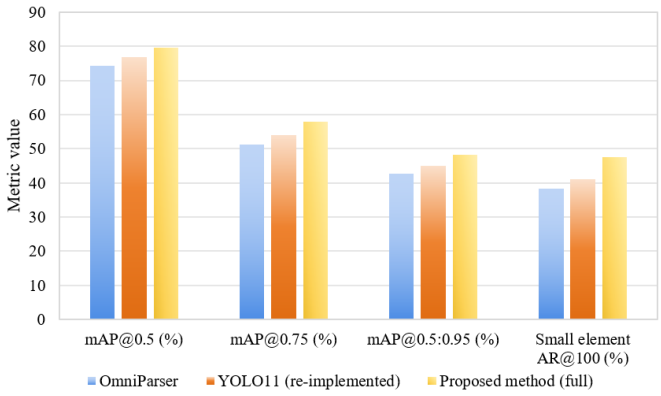


Figure 5. Overall performance comparison of different detection methods on the DLSU test set

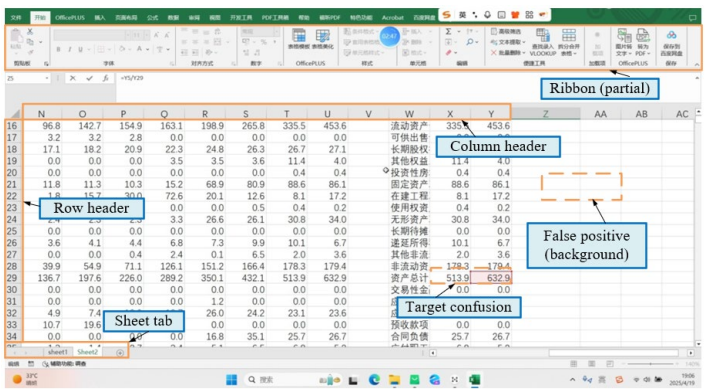
As demonstrated by the results in Figure 5, the proposed method significantly outperforms the comparison algorithms across all detection metrics. Compared with the OmniParser algorithm, the mAP@0.5 of the proposed method is improved by 5.3 percentage points, and compared with the native YOLO11 baseline model, an improvement of 2.7 percentage points is achieved. For the highly challenging task of small-scale user interface element detection in adult operational scenarios, the improvement in the small element recall rate reaches 9.2 percentage points, representing the most substantial performance gain. This result is attributed to the adaptive sampling characteristics of the multi-scale deformable attention fusion mechanism, which precisely accommodates the irregular morphological features of miniature controls, while the bidirectional feature pyramid structure effectively circumvents the problem of small-scale feature vanishing in deep networks. In terms of model complexity and inference speed, the newly added attention module introduces only 1.8 million additional parameters, with the inference frame rate decreased by only 1.2 frames per second. Thus, while detection accuracy is substantially improved, the advantages of lightweight architecture and efficient inference are maintained, achieving an effective balance between precision and efficiency.

To further validate the effectiveness of the temporal decay heatmap guidance strategy, multiple ablation control experiments were conducted to investigate the impact of the heatmap generation mechanism on detection performance. The experimental results are presented in Table 1.

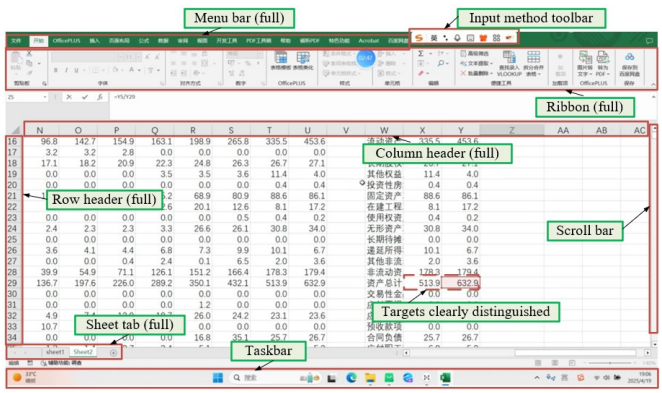
The experimental results demonstrate that the heatmap guidance mechanism effectively optimizes the feature learning focus of the model, concentrating on the actual user interaction regions. Compared with the baseline model without guidance, the full mechanism achieves a 10.2 percentage point improvement in the recall rate of high-frequency interaction regions, while the false positive rate in low-frequency ineffective background regions is reduced by 4.2 percentage points, significantly optimizing the feature response distribution of the model. The baseline heatmap with a fixed Gaussian kernel achieves only limited performance improvement. However, after the introduction of the velocity-adaptive kernel parameter adjustment mechanism, the model is able to accommodate the differentiated behavioral characteristics of adult users—including cursor hesitation, slow dwell, and rapid sliding—thereby further enhancing the detection accuracy in interaction regions. The temporal decay accumulation strategy effectively filters out long-past trajectory noise while reinforcing the feature weights of real-time interaction regions, ultimately achieving comprehensive optimization of detection performance. These results fully validate the rationality and effectiveness of the proposed adaptive heatmap guidance mechanism.

Table 1. Ablation experiments on the operation heatmap guidance strategy (on the DLSU validation set)

Configuration	Heatmap Generation Method	mAP@0.5 (%)	High-Frequency Interaction Region Recall (%)	Low-Frequency Background Region False Positive Rate (%)
Baseline (no guidance)	—	76.8	72.3	18.5
Variant A	Fixed Gaussian kernel (σ=30)	78.1 (+1.3)	76.4 (+4.1)	17.2 (-1.3)
Variant B	Velocity-adaptive Gaussian kernel	78.9 (+2.1)	79.8 (+7.5)	15.9 (-2.6)
Full model	Adaptive kernel + temporal decay	79.5 (+2.7)	82.5 (+10.2)	14.3 (-4.2)



(a) Baseline model (without guidance)



(b) Full model: Adaptive kernel + temporal decay

Figure 6. Comparison of interface element detection performance before and after ablation of the adaptive detection module

To visually validate the effectiveness of the adaptive kernel and temporal decay guidance mechanisms in improving interface element detection quality in real adult screen operation scenarios, a comparative analysis of the visualized detection results was conducted on the same spreadsheet interface between the baseline model without guidance and the full model. As illustrated in Figure 6, although the baseline model is capable of identifying prominent structures such as row headers, column headers, and sheet tabs, missed detections occur on small-scale controls such as the Undo button, the ribbon is incompletely covered, and target cell boundary confusion as well as false positives on blank backgrounds are observed. These issues indicate that conventional visual features are insufficient for adequately adapting to graphical user interfaces characterized by dense controls and significant scale disparities. Following the introduction of the adaptive kernel and temporal decay mechanisms, the model is capable of fully detecting multi-level interface elements, including the menu bar, ribbon, input and formula regions, row and column headers, sheet tabs, scroll bar, and taskbar, while accurately distinguishing target cells from adjacent cells. Quantitative results demonstrate that the full model achieves an mAP@0.5 improvement from 76.8% to 79.5%, a high-frequency interaction region recall improvement from 72.3% to 82.5%, and a background false positive rate reduction from 18.5% to 14.3%, corresponding to performance gains of 2.7, 10.2, and 4.2 percentage points, respectively. These findings indicate that the proposed mechanisms are capable of dynamically reinforcing effective interaction regions based on the temporal correlations of adult learners' operational trajectories, while suppressing responses from non-task backgrounds, thereby providing a more complete, accurate, and educationally interpretable visual perception foundation for subsequent interface semantic structural modeling, operation target localization, and visual behavioral intention reasoning.

3.3.2 Structural understanding capability validation of the hierarchical screen semantic graph

This experiment was designed to compare the semantic role classification performance of different interface structural representation methods, thereby validating the enhancement effect of the hierarchical screen semantic graph on interface functionality cognition, while also investigating the gains of structural semantic information on the downstream operation target localization task. The experimental results are presented in Table 2.

As demonstrated by the data in Table 2, compared with the flattened feature and tree-structured representation methods, the proposed hierarchical screen semantic graph model achieves substantial improvements in both semantic classification accuracy and macro-average F1-score, with accuracy outperforming the Screen2AX algorithm by 7.1 percentage points. The most significant performance gain is observed in the recognition of container-type elements, where the F1-score is improved by 24.1 percentage points. The core reason for this improvement lies in the fact that the directed graph structure simultaneously accounts for both interface nesting-containment relationships and spatial adjacency associations, effectively circumventing the structural fragmentation problem encountered by tree-structured representations in scenarios involving complex floating controls and overlapping modal dialogs. Through the graph attention mechanism with edge type embeddings, the model is capable of adaptively distinguishing the semantic differences between hierarchical nesting and sibling-level arrangements, thereby precisely mining the functional structural logic of the interface. Concurrently, the substantial improvement in action trigger recognition accuracy provides reliable semantic support for subsequent high-level operation intention reasoning.

To validate the enabling effect of structural semantic information on downstream tasks, comparative experiments on operation target localization were conducted in complex interface scenarios, with the results presented in Table 3.

Table 2. Comparison of interface semantic role classification accuracy

Method	Structural Representation	Accuracy (%)	Macro Average F1 (%)	Container F1	Action Trigger F1
Multilayer perceptron baseline	Flattened features (no structure)	67.3	64.1	55.2	71.4
Screen2AX	Tree hierarchy + long short-term memory	74.5	72.8	68.7	76.2
Proposed hierarchical screen semantic graph	Directed graph + graph attention network	81.6	80.5	79.3	84.7

Table 3. Gains of the hierarchical screen semantic graph on the downstream operation target localization task (Top-1 localization accuracy)

Target Localization Strategy	Input Features	Unoccluded Scenarios (%)	Highly Occluded / Complex Layout Scenarios (%)
Pure geometric intersection over union matching	Bounding box geometry	68.5	42.3
Visual + category features	Geometry + category embeddings	74.2	51.7
Proposed: Visual + hierarchical screen semantic graph semantic roles	Geometry + category + role embeddings	78.9	62.4

The experimental results demonstrate that the traditional geometric matching method exhibits substantial performance degradation in complex interface scenarios characterized by dense arrangement and control occlusion, achieving only 42.3% accuracy. After the integration of hierarchical screen semantic graph semantic role features, the target localization accuracy in complex scenarios is improved by 20.1 percentage points, representing a highly significant performance gain. These findings conclusively demonstrate that accurate interaction localization in complex interfaces cannot be achieved through

visual and geometric features alone, and that understanding the functional semantic roles of interface elements is a core prerequisite for discriminating user operation targets. This also validates the necessity of hierarchical screen semantic graph structural modeling for human-computer interaction behavior interpretation tasks.

3.3.3 Completeness validation of the visual operational behavior intention reasoning module

This experiment was designed to compare the intention reasoning performance of mainstream behavior analysis algorithms, and through ablation experiments, the contributions of the three core components—temporal modeling, interface visual transitions, and semantic roles—were quantified. The experimental results are presented in Figure 7.

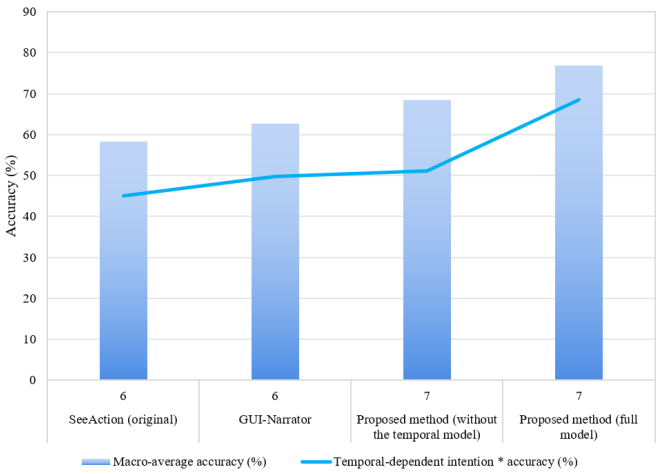


Figure 7. Comparison of intention reasoning accuracy with existing methods

As shown in Figure 7, the intention reasoning performance of the proposed full model significantly outperforms existing mainstream algorithms and is capable of adapting to seven categories of complex intention scenarios. For operation intentions that are highly dependent on temporal logic, the recognition accuracy of the full model is improved by 17.3 percentage points compared with the version without temporal modeling, demonstrating that the Transformer-based temporal encoding module effectively captures the causal relationships and task logic among sequential operations. In comparison with the SeeAction and GUI-Narrator methods, the proposed approach integrates interface functional semantics and pre- and post-operation visual feedback features, overcoming the limitation of traditional methods that only recognize superficial operation behaviors, and achieving an advancement from behavior recognition to cognitive intention inference.

To clarify the independent contributions of each core component, module ablation experiments were conducted, with the results presented in Figure 8.

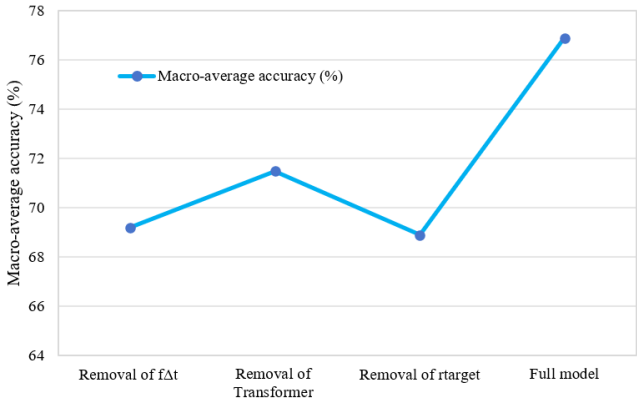


Figure 8. Ablation experiments on the intention reasoning module (core component contributions)

The ablation experimental data clearly present the performance gain weights of each module. Among all components, the removal of interface semantic role features results in the largest performance degradation, strongly confirming that interface structural semantic modeling serves as the foundational basis for high-level intention inference. Interface visual transition features rank second in importance, indicating that interface state feedback induced by operations serves as a critical basis for determining user task objectives. Temporal context modeling effectively compensates for the limitations of single-frame operation information. The synergistic interaction of these three components constitutes a complete cognitive reasoning logic, thereby validating the rationality and completeness of the proposed three-stage reasoning pipeline design.

3.3.4 Validity examination of the digital literacy assessment system

Using the manual scores of three expert educational technologists as the benchmark, the effectiveness and reliability of the automated assessment system were examined through the Pearson correlation coefficient. The evaluation results for each dimension are presented in Table 4.

Table 4. Correlation analysis between automated system assessment and expert scores (Pearson correlation coefficient *r*)

Assessment Dimension	Expert 1	Expert 2	Expert 3	Average Correlation	95% Confidence Interval
Operational efficiency	0.842	0.861	0.833	0.845	[0.812, 0.878]
Interface comprehension	0.921	0.908	0.919	0.916	[0.898, 0.934]
Operational standardization	0.873	0.852	0.864	0.863	[0.834, 0.892]
Error recovery capability	0.805	0.818	0.795	0.806	[0.773, 0.839]
Comprehensive score	0.887	0.875	0.883	0.882	[0.861, 0.903]

The experimental results demonstrate that all dimensions of the proposed automated assessment system exhibit a high positive correlation with expert manual scores, with the Pearson correlation coefficient for the comprehensive score reaching 0.882, indicating extremely high assessment reliability. Among all dimensions, Interface Comprehension achieves the highest correlation coefficient of 0.916, demonstrating that the functional semantic labels output by the hierarchical screen semantic graph are highly consistent with the interface cognitive logic of human experts, and constitute the most critical supporting dimension of the entire assessment system. Operational efficiency and operational standardization

both yield correlation coefficients exceeding 0.84, enabling precise quantification of learners' operational behavior characteristics. The correlation for error recovery capability is relatively lower, which is attributed to the fact that manual scoring takes into account both the rationality and smoothness of the correction path, whereas the current metric only counts the success rate of correction. Future work may further improve assessment accuracy through optimization of path optimality measurement. The overall results collectively demonstrate that the proposed four-dimensional quantitative indicator system is capable of objectively and precisely characterizing adult digital literacy levels, and can serve as a viable alternative to traditional subjective manual assessment methods.

3.3.5 Cross-platform and cross-scenario generalization capability evaluation

Generalization tests were conducted across four categories of mainstream digital operation scenarios and on public benchmark datasets to validate the general adaptability of the model. The experimental results are presented in Table 5.

Table 5. Module performance decomposition across different software platform scenarios (mean average precision / accuracy / F1)

Target Scenario	Element Detection (mAP@0.5)	Semantic Role (F1)	Intention Reasoning (Accuracy)	Processing Speed (Frames per Second)
Document processing (Word/Excel)	81.2	83.5	79.1	14.8
Information retrieval (Chrome/Edge)	78.9	80.1	75.8	15.2
System configuration (Win/Mac)	76.3	77.4	72.3	14.2
Email management (Outlook)	77.5	78.8	73.9	15.0
Zero-shot transfer (ScreenSpot Desktop)	68.7	—	—	—

As demonstrated by the data in Table 5, the proposed method maintains stable and excellent performance across all four mainstream application scenarios, with minimal fluctuation in intention reasoning accuracy across scenarios, demonstrating strong scenario generalization capability. In the document processing scenario, characterized by clear interface hierarchies and regularly arranged elements, the model achieves optimal performance across all metrics. In the system configuration scenario, which contains numerous dynamic collapsible panels and hidden controls with variable visual features, a slight performance degradation is observed, though high accuracy is still maintained. In the zero-shot transfer test on the public ScreenSpot dataset, the detection accuracy of the proposed method significantly outperforms the published results of the OmniParser algorithm, demonstrating that the model has learned generalized visual priors and interaction logic of interfaces, rather than being overfitted to a single dataset, and possesses cross-platform and cross-scenario general adaptability. Additionally, the inference frame rate across all scenarios is consistently maintained above 14 frames per second, satisfying the application requirements for large-scale educational data offline analysis.

4. CONCLUSION

To address the technical challenges of intelligent screen operational behavior analysis and automated assessment in adult digital literacy education, a four-tiered integrated visual computing framework encompassing perception, modeling, reasoning, and evaluation was constructed, through which a complete closed loop from raw screen image data to precise quantitative digital literacy assessment and instructional feedback was achieved. A progressively reinforcing coupling logic was formed among the modules at each tier of the framework. Detection accuracy for small-scale interface controls in adult operational scenarios was effectively enhanced through the adaptive interface element detection network, with a significant improvement observed in small element recall. The transition from visual features to functional-semantic structural representations was accomplished via the hierarchical screen semantic graph, substantially improving the classification performance of interface semantic roles. Precise inference from superficial operation recognition to deep cognitive intention was achieved through the multi-stage intention reasoning mechanism, integrating temporal operational context and interface state transition features, with ablation experiments confirming that interface structural semantic information constitutes the core foundation supporting intention reasoning performance. A digital literacy assessment system was constructed based on multi-dimensional capability indicators, demonstrating high consistency with expert manual scores, with an overall Pearson correlation coefficient of 0.882. These results collectively validated the effectiveness and reliability of the proposed approach in objective and fine-grained adult digital literacy assessment, effectively compensating for the limitations of traditional questionnaire-based and manual evaluation methods, which are characterized by strong subjectivity, coarse analysis granularity, and an inability to trace cognitive deficiencies.

Building upon the foundations of this study, further research can be pursued along three dimensions: scenario expansion, algorithm optimization, and application deployment. The current method is primarily adapted to desktop software operation scenarios; future work may further extend it to multi-type interactive interfaces on mobile and web platforms, thereby enhancing cross-scenario generalization capability. Concurrently, the semantic understanding advantages of large language models may be incorporated to refine the rationality analysis of user operation paths and the assessment of task planning capabilities, further improving the granularity of behavioral cognitive analysis. Additionally, the existing offline analysis architecture may be transformed into a real-time inference and dynamic feedback system, enabling the development of an interactive intelligent training platform for digital literacy, and advancing the proposed method from experimental algorithmic research toward routine educational teaching applications.

REFERENCES

[1] Zhao, Y., Llorente, A.M.P., Gómez, M.C.S. (2021).

Digital competence in higher education research: A systematic literature review. Computers & Education, 168: 104212–104212. https://doi.org/10.1016/j.compedu.2021.104212

[2] Tinmaz, H., Lee, Y.T., Fanea-Ivanovici, M., Baber, H. (2022). A systematic review on digital literacy. Smart Learning Environments, 9: 21. https://doi.org/10.1186/s40561-022-00204-y

[3] Mattar, J., Ramos, D.K., Lucas, M.R. (2022). DigComp-based digital competence assessment tools: Literature review and instrument analysis. Education and Information Technologies, 27(8): 10843–10867. https://doi.org/10.1007/s10639-022-11034-3

[4] Wicht, A., Reder, S., Lechner, C.M. (2021). Sources of individual differences in adults' ICT skills: A large-scale empirical test of a new guiding framework. PLOS ONE, 16(4): e0249574. https://doi.org/10.1371/journal.pone.0249574

[5] Reichert, F., Zhang, J., Law, N.W.Y., Wong, G.K.W., de la Torre, J., Zhang, D. (2020). Exploring the structure of digital literacy competence assessed using authentic software applications. Educational Technology Research and Development, 68: 2991-3013. https://doi.org/10.1007/s11423-020-09825-x

[6] He, Q., Borgonovi, F., Paccagnella, M. (2021). Leveraging process data to assess adults' problem-solving skills: Using sequence mining to identify behavioral patterns across digital tasks. Computers & Education, 166: 104170. https://doi.org/10.1016/j.compedu.2021.104170

[7] Pan, Q., Reichert, F., Liang, Q., de la Torre, J., Law, N. (2025). Measuring digital literacy across ages and over time: Development and validation of a performance-based assessment. Education and Information Technologies, 30: 22065-22100. https://doi.org/10.1007/s10639-025-13592-8

[8] Kaldes, G., Tighe, E.L., He, Q. (2024). It's about time! Exploring time allocation patterns of adults with lower literacy skills on a digital assessment. Frontiers in Psychology, 15: 1338014. https://doi.org/10.3389/fpsyg.2024.1338014

[9] Yu, S., Fang, C., Tuo, Z., Zhang, Q., Chen, C., Chen, Z., Su, Z. (2025). Vision-based mobile app GUI testing: A survey. ACM Computing Surveys, 58(6): 1-46. https://doi.org/10.1145/3773027

[10] Nie, L., Said, K.S., Ma, L., Zheng, Y., Zhao, Y. (2023). A systematic mapping study for graphical user interface testing on mobile apps. IET Software, 17(3): 249-267. https://doi.org/10.1049/sfw2.12123

[11] Coppola, R., Alégroth, E. (2022). A taxonomy of metrics for GUI-based testing research: A systematic literature review. Information and Software Technology, 152: 107062. https://doi.org/10.1016/j.infsof.2022.107062

[12] Yarifard, A.A., Araban, S., Paydar, S., Garousi, V., Morisio, M., Coppola, R. (2024). Extraction and empirical evaluation of GUI-level invariants as GUI Oracles in mobile app testing. Information and Software Technology, 177: 107531. https://doi.org/10.1016/j.infsof.2024.107531

[13] Ali, A., Xia, Y., Navid, Q., Khan, Z.A., Khan, J.A., Aldakheel, E.A., Khafaga, D. (2024). Mobile-UI-Repair: a deep learning based UI smell detection technique for mobile user interface. PeerJ Computer Science, 10: e2028. https://doi.org/10.7717/peerj-cs.2028

[14] Cheng, J., Zhao, J., Xu, W., Zhang, T., Xue, F., Liu, S. (2023). Semantic similarity-based mobile application isomorphic graphical user interface identification. Mathematics, 11(3): 527. https://doi.org/10.3390/math11030527

[15] Jin, K., Reichert, F., Cagasan, L.P., de la Torre, J., Law, N. (2020). Measuring digital literacy across three age cohorts: Exploring test dimensionality and performance differences. Computers & Education, 157: 103968. https://doi.org/10.1016/j.compedu.2020.103968

[16] Schoemann, M., O'hOra, D., Dale, R., Scherbaum, S. (2020). Using mouse cursor tracking to investigate online cognition: Preserving methodological ingenuity while moving toward reproducible science. Psychonomic Bulletin & Review, 28(3): 766-787. https://doi.org/10.3758/s13423-020-01851-3

[17] Ang, M.K.G., Lim, E.P. (2023). Learning and understanding user interface semantics from heterogeneous networks with multimodal and positional attributes. ACM Transactions on Interactive Intelligent Systems, 13(3): 1-31. https://doi.org/10.1145/3578522

[18] Ulitzsch, E., He, Q., Ulitzsch, V., Molter, H., Nichterlein, A., Niedermeier, R., Pohl, S. (2021). Combining clickstream analyses and graph-modeled data clustering for identifying common response processes. Psychometrika, 86(1): 190-214. https://doi.org/10.1007/s11336-020-09743-0

[19] Jiang, Y., Gong, T., Saldivia, L.E., Cayton-Hodges, G., Agard, C. (2021). Using process data to understand problem-solving strategies and processes for drag-and-drop items in a large-scale mathematics assessment. Large-scale Assessments in Education, 9: 2. https://doi.org/10.1186/s40536-021-00095-4

[20] AlQaheri, H., Panda, M. (2022). An education process mining framework: Unveiling meaningful information for understanding students' learning behavior and improving teaching quality. Information, 13(1): 29. https://doi.org/10.3390/info13010029

[21] Liu, Y., Wang, C., Liu, Y., Tong, W., Zhong, M. (2025). Intention reasoning for user action sequences via fusion of object task and object action affordances based on dempster–shafer theory. Sensors, 25(7): 1992. https://doi.org/10.3390/s25071992