Online-Journals.org

iJIM International Journal of Interactive Mobile Technologies

iJIM | eISSN: 1865-7923 | Vol. 19 No. 23 (2025) | OPEN ACCESS

https://doi.org/10.3991/ijim.v19i23.59253

PAPER

Development of a Mobile AR-Based English Vocabulary Learning Platform

Weifeng Deng[1], Lin Wang[1] (✉), Xue Deng[2]

[1]Hainan Vocational University of Science and Technology, Haikou, China

[2]Hebei Vocational University of Technology and Engineering, Xingtai, China

wanglintaichi@163.com

ABSTRACT

The integration of mobile computing and sensing technologies has propelled technology-enhanced language learning into a new era characterized by contextualization and immersion. A key challenge is the effective fusion of augmented reality (AR) and artificial intelligence (AI) to create a learning platform that not only fosters deep vocabulary acquisition but also facilitates flexible knowledge transfer. This study introduces an advanced AR-based English vocabulary learning platform that leverages semantic tracking and multimodal information fusion. This platform transitions from passive labeling to active contextualized content creation, enhancing cross-modal personalized interactions through a novel service modeling and matching module. Additionally, this research proposes an event-matching-driven transfer mode recognition method. This method, which uses learning interactions modeled as temporal event sequences and integrates textual semantic features, allows for a detailed analysis of vocabulary knowledge transfer across three dimensions: consolidation, contextual generalization, and conceptual abstraction. Technologically, this approach marks a progression from simple environmental labeling to sophisticated contextual understanding and response, establishing a comprehensive loop of intelligent environmental perception and adaptive content creation. Methodologically, it develops a new framework for analyzing transfer effects, shifting from broad effect assessments to detailed analyses of cognitive mechanisms, and setting a new standard for the design and evaluation of intelligent language learning systems.

KEYWORDS

augmented reality (AR), semantic tracking, mobile learning, English vocabulary learning, event matching, transfer effect analysis

1 INTRODUCTION

With the rapid development of mobile Internet [1–3] and sensing technology [4, 5], augmented Rrality (AR) [6, 7] has transitioned from desktop platforms to an immersive, contextualized new paradigm centered around smartphones. In this context, mobile-assisted language learning [8, 9] is undergoing a profound

Deng, W., Wang, L., Deng, X. (2025). Development of a Mobile AR-Based English Vocabulary Learning Platform. *International Journal of Interactive Mobile Technologies (iJIM)*, 19(23), pp. 68–82. https://doi.org/10.3991/ijim.v19i23.59253

Article submitted 2025-08-13. Revision uploaded 2025-10-07. Final acceptance 2025-10-17.

transformation from static content presentation to dynamic interactive experiences. In particular, AR technology, by seamlessly overlaying virtual information onto the real environment, creates an unprecedented "contextual fusion" experience for language learning, allowing learners to interact intuitively with virtual language symbols while perceiving the physical world. At the same time, breakthroughs in the field of AI [10, 11], especially the mature application of deep learning in computer vision and natural language processing, have enabled machines to "understand" the semantic content of visual scenes, thus providing the technological possibility for deeper, context-based adaptive learning. Vocabulary, as the cornerstone of language, highly depends on the contextualization and intelligent reproduction of its acquisition efficiency. Therefore, how to deeply integrate the immersive experience of AR with the semantic understanding capabilities of AI to construct a learning system that can perceive the environment, understand the user, and provide intelligent language scaffolding has become a cutting-edge exploration in the field of digital language learning.

Although many studies have confirmed the positive effects of AR technology on enhancing learning motivation and immersion, existing methods still have significant limitations in promoting deep, transferable language learning. Firstly, most AR educational applications, such as the early systems reviewed in references [12, 13], tend to focus on "what you see is what you get" surface-level interaction design. That is, they only overlay 3D models or text labels onto recognized objects, lacking in-depth semantic understanding of the scene and semantic-based interactive guidance, leading to a formalized and fragmented learning experience. Secondly, in the field of learning analytics, as pointed out in references [14, 15], many studies rely on delayed, test-score-based evaluation metrics, failing to effectively capture and quantify the dynamic cognitive construction and transfer behavior occurring during the learning process. For example, although the studies in references [16, 17] use learning logs, they are mainly used to predict final scores rather than performing fine-grained modeling of learners' behavior trajectories to identify different transfer patterns. The inadequacy in mining process-oriented behavior data makes it difficult to answer a key question: how does technological intervention step by step influence and facilitate knowledge transfer? These deficiencies point to a core issue: current research has not yet effectively bridged the gap between "environmental intelligent perception" and "deep learning mechanism analysis."

To address these shortcomings, the research in this paper consists of two organically integrated parts. The first part is the construction of an English vocabulary learning platform integrating AR reality and semantic tracking. We will tackle key technologies such as real-time environmental semantic understanding and service intelligent matching to build an intelligent system capable of upgrading from passive "labeling" of the environment to active "understanding" and "responding" to the environment. This system can not only recognize objects and label words but also infer scene relationships based on visual language models, dynamically generating matching sentences, tasks, and dialogues, thus achieving the elevation from "vocabulary presentation" to "contextualized language application." The second part is event-matching-driven transfer mode recognition. We innovatively model all learner interactions as multidimensional "events" and introduce event matching and sequence analysis algorithms to quantitatively identify and analyze vocabulary knowledge transfer patterns at three levels: micro, meso, and macro. This enables breakthroughs from macro effect evaluation to micro mechanism analysis. The core value of this research lies in its technological achievement in deeply integrating AR and semantic tracking, as well as in its methodological construction of a theoretical framework for

transfer effect analysis based on behavioral event sequence analysis, providing a new paradigm for the theoretical research and system design of technology-enhanced language learning with both technological innovation and scientific depth.

2 CONSTRUCTION OF AN ENGLISH VOCABULARY LEARNING PLATFORM INTEGRATING AR AND SEMANTIC TRACKING

This section aims to explain the system architecture and the implementation of core modules of the platform. The core of the platform lies in constructing an intelligent system that can perceive the semantics of the physical environment and provide personalized language learning content based on this. We first introduce the overall architecture of the platform and then delve into the two key modules: environmental semantic understanding and service integration interaction.

The platform adopts a client-server layered architecture, aiming to build a highly cohesive, low-coupling intelligent learning system by following the principle of separation of concerns. The architecture is divided into three core layers from top to bottom: the perception and presentation layer, the semantic and decision layer, and the data and service layer. The topmost perception and presentation layer resides on the mobile terminal. Its core responsibility is to construct an immersive learning context. It captures real-time video streams and inertial measurement unit (IMU) data through the device's camera and sensors and utilizes the AR engine to implement simultaneous localization and mapping (SLAM), accurately registering virtual information onto the physical space. Meanwhile, this layer serves as a multi-modal data collection and interaction interface, recording fine-grained interaction events such as user touches, voice inputs, and gaze points, providing raw data for subsequent learning behavior analysis. The semantic and decision layer, located on the server side, acts as the "intelligent brain" of the system. It receives the environmental data and user behavior sequences uploaded from the frontend and performs deep environmental semantic analysis and user intent recognition through cascaded AI models, thus achieving a leap from low-level pixel data to high-level semantic concepts. The bottom data and service layer plays the role of the system's "knowledge base" and "toolbox." It persistently stores structured user profiles, domain knowledge graphs, and learning resources, and integrates third-party services through standardized interfaces, providing continuous knowledge support and functional extensions for the upper-level semantic decision-making. This hierarchical design ensures that the system maintains a smooth experience on the client side while handling high computational loads from AI tasks and provides clear boundaries and flexibility for functional iteration and expansion.

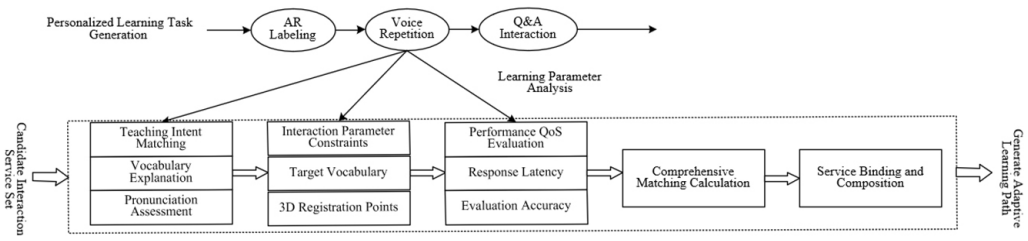

Fig. 1. Service matching and adaptive learning path generation process

Figure 1 details how the platform converts dynamic learning scenarios into personalized learning paths. The entire process begins with the personalized learning

task generation module, which generates a task sequence consisting of "AR labeling → voice repetition → question-answer interaction," based on the real-time perceived AR environment and the user's status. The system then analyzes the learning parameters of the task and extracts key semantic elements. Based on these parameters, the system performs comprehensive filtering and evaluation through three levels: teaching intent matching, interaction parameter constraints, and service effectiveness evaluation, from the candidate interaction service pool. Finally, through a comprehensive matching degree calculation and service binding and combination, it outputs an executable adaptive learning path that is fully suited to the current environment and user, thus achieving an intelligent feedback loop from environmental perception to precise teaching intervention.

The environmental semantic understanding module adopted by the platform is designed with the core idea of achieving a leap from low-level environmental perception to high-level semantic cognition. The construction of this module follows a hierarchical information processing pipeline. The process starts with visual entity detection and recognition, where we use an advanced model based on open vocabulary object detection, overcoming the limitation of the traditional model's predefined number of categories. After acquiring the object boundary boxes and basic labels, the process proceeds to the semantic labeling and relationship extraction stage. In this stage, a powerful visual language model is introduced, with its core ability to understand the overall context of the image and the interaction relationships between entities. It can not only generate more descriptive aliases and synonyms for detected entities but also perform relationship reasoning, outputting semantically rich descriptions such as "The cat is sleeping on the sofa" or "The cup is full of coffee."

However, a truly intelligent learning system must go beyond static understanding of a single image frame and be able to integrate spatiotemporal context and user status for dynamic decision-making. Therefore, the top layer of this module is learning scenario modeling. As an information fusion hub, it deeply integrates the static scene semantics generated in the previous two stages with the dynamic user profile and interaction history. Specifically, the model analyzes the user's historical knowledge mastery, real-time interaction behavior, and the spatiotemporal attributes of the learning scenario to construct a dynamically evolving probabilistic graphical model. This model can estimate which vocabulary items are cognitive blind spots for the user and which semantic relationships are most valuable for teaching in the current complex scenario.

3 VOCABULARY TRANSFER PATTERN RECOGNITION AND ANALYSIS BASED ON EVENT-SEMANTIC INTERACTION

To explore the transfer mechanisms of vocabulary knowledge in depth, this section constructs an analysis model that integrates learning behavior events and textual semantic features. The model aims to go beyond traditional accuracy and response time metrics by calculating the evolution of behavior event sequences and their interaction with deep semantic spaces, systematically recognizing and quantifying three typical transfer patterns from micro to macro levels.

3.1 Knowledge consolidation

"Knowledge consolidation," as the initial stage of transfer, is analyzed based on the intersection of cognitive psychology and educational data mining. We propose

that learners' mastery of a single vocabulary item is not completed instantaneously but is an internalization process reflected in the time dimension and composed of a series of explicit interaction events. Each interaction with the target vocabulary can be regarded as a "learning event" containing rich metadata. To extract continuous cognitive state changes from these discrete events, we use a sequential deep learning model as the core theoretical tool. Through the sequence modeling capabilities of long short-term memory (LSTM) or Transformer, we can encode multiple interactions with the same vocabulary item into a dynamic, computable state evolution sequence.

After obtaining the serialized representation of the knowledge state, the focus of the analysis shifts to precisely defining and classifying different consolidation patterns through multi-dimensional metrics and unsupervised learning. We define three intercorrelated computational metrics to characterize this process: help-seeking behavior decay rate, reflecting the increase in learner independence as they transition from external support to autonomous recall; response time convergence curve, reflecting the reduced cognitive load and increased automation of information retrieval and processing; and interaction accuracy improvement trajectory, directly representing the enhancement in the accuracy of knowledge application. These three metrics collectively form a multi-dimensional feature space that describes the intensity of "knowledge consolidation." Specifically, let α be the asymptote of help-seeking behavior, representing the final stable, minimum help-seeking probability after sufficient learning. Let β be the initial help-seeking probability, representing the tendency to seek help when first encountering the vocabulary. k is the decay coefficient, directly reflecting the rate of decay in help-seeking behavior. t represents time or attempt number, and the probability of help-seeking behavior observed at time point t, $P(t)$, can be calculated using the following formula:

$$P(t) = \alpha + (\beta - \alpha) \times e^{-kt} \tag{1}$$

Let RT_{min} be the minimum reaction time, representing the lowest possible response time reflecting the processing speed after complete knowledge consolidation. RT_{max} is the initial reaction time. c is the convergence index, reflecting the speed at which response time decreases and approaches RT_{min}. The average response time $RT(t)$ at the t-th attempt can be calculated using the following formula:

$$RT(t) = RT_{min} + (RT_{max} - RT_{min}) \times t^{-c} \tag{2}$$

Let y be the learning ceiling, representing the theoretical upper limit of accuracy. r is the learning rate, reflecting the speed at which accuracy improves from the initial level to the ceiling level. τ is the inflection point, representing the moment of fastest learning speed. The accuracy at time point or attempt number t can be calculated using the following formula:

$$Accuracy(t) = \frac{\gamma}{1 + e^{-r(t-\tau)}} \tag{3}$$

On this basis, we apply unsupervised clustering algorithms to mine this high-dimensional behavior feature space. The basic principle is to automatically and data-driven identify universal consolidation patterns by measuring the similarity of different learner sequence features. Examples of these patterns include the "rapid mastery type," which characterizes efficient learning, the "steady improvement type," which reflects steady progress, and the "difficult retention type," which indicates learning obstacles. This data mining-based pattern discovery not only verifies

the theoretical existence of knowledge consolidation but also provides accurate and actionable scientific evidence for implementing personalized teaching interventions.

3.2 Contextual generalization

The core cognitive principle of contextual generalization lies in whether the learner can transcend the "context binding" of vocabulary in a specific physical scene and abstract it into a semantic concept that can be flexibly applied in different contexts. Traditional transfer analysis struggles to quantify the key variable of "different contexts," while the innovation of this study is the introduction of a sentence-level semantic vector model, which provides a computable definition of "context." We use advanced sentence embedding technologies such as Sentence-BERT to map the overall contextual descriptions of different AR scenes into a high-dimensional semantic vector space. In this space, the cosine or Euclidean distance between vectors is precisely quantified as the semantic similarity between scenes. Specifically, the vectorized representation of context can be achieved through the following process:

a) Scene text description generation: For a given AR scene S_i, we use the environmental semantic understanding module in the platform to automatically generate a general natural language description D_i. This description should go beyond a simple listing of objects and should, as much as possible, include spatial and functional relationships between the objects.

b) Sentence embedding vector extraction: The generated scene description D_i is input into the pre-trained Sentence-BERT model. SBERT is designed to generate semantically meaningful sentence embeddings and can map variable-length sentences into a consistent high-dimensional vector space. The semantic vector representation of scene S_i is given by:

$$V_i' = SBERT(D_i) \tag{4}$$

After obtaining the vector representations V_i' and V_j' for two scenes S_i and S_j, we quantify their semantic similarity by calculating the cosine similarity in the high-dimensional space. Let n be the dimension of the vector. V_{ik} represents the component of vector V in the k-th dimension, and the formula for calculating the cosine similarity is as follows:

$$SIM(S_i, S_j) = \frac{V_i' \cdot V_j'}{\left\|V_i'\right\|\left\|V_j'\right\|} = \frac{\sum_{k-1}^{n} V_{ik} V_{jk}}{\sqrt{\sum_{k-1}^{n} V_{ik}^{\ 2}} \sqrt{\sum_{k-1}^{n} V_{jk}^{\ 2}}} \tag{5}$$

The result of the cosine similarity calculation is a value between [–1, 1], but in semantic similarity applications, the value typically lies in the range [0, 1], since vectors usually reside in the positive quadrant. A value close to 1 indicates that the semantics of the two scenes are highly similar. A value close to 0 indicates that the semantics of the two scenes are nearly unrelated. A value close to –1 indicates that the two scenes are semantically opposite.

Based on the precise matching of cross-scene learning events, the focus of the analysis shifts to the quantification and modeling of the "transfer effect." We propose that true generalization is reflected in the learner's improvement in cognitive

processing efficiency and accuracy when encountering known vocabulary or concepts in a new scene. To this end, we construct a transfer effect calculation function, which primarily compares the event feature differences when the same learner processes semantically matched vocabulary in the "source scene" and the "target scene." The specific expression of this function is as follows:

$$TE_{simple} = \frac{Perf(E_{tar}) - Perf(E_{sou})}{Perf(E_{sou})} \tag{6}$$

In this function, the core variables are defined as: E_{sou} represents the learner's interaction event when first encountering the target vocabulary in the source scene, which establishes the baseline performance level; E_{tar} represents the learner's interaction event when first processing the same vocabulary or highly related vocabulary in the target scene without prior practice. The core of the function is $Perf(E)$, which is a composite performance score extracted from a single interaction event. This score is a multi-dimensional metric that reflects the accuracy, efficiency, and autonomy of cognitive processing. Its recommended calculation method is:

$$Perf(E) = \frac{Accuracy(E) \times \log(1 + \frac{1}{Time_{RES}(T)})}{1 + Score_{Helpseek}(E)} \tag{7}$$

In this calculation, the accuracy of the event is represented by $Accuracy(E)$, the reaction time of the event by $Time_{RES}$, and the intensity of help-seeking behavior by $ScoreHelpseek$. The numerator combines accuracy and the reciprocal of response time through a logarithmic transformation, simultaneously rewarding answer correctness and processing speed, reflecting the fluency of cognitive processing. The denominator introduces a help-seeking behavior score to penalize reliance on external support, encouraging learner autonomy. Finally, the function calculates the difference between the target scene performance and the initial performance in the source scene and divides it by the initial performance itself to obtain a TE_{simple} value that represents the relative improvement. A value greater than zero indicates positive transfer, and the magnitude of the value directly quantifies the strength of the transfer effect.

This function synthesizes key metrics such as initial response accuracy and response efficiency and, through differential calculation with the learner's initial performance in the source scene, provides a purified transfer effect value. This value effectively removes the influence of the learner's prior level, directly reflecting the knowledge application gains or losses caused by context changes. Finally, by conducting distribution analysis or clustering analysis on the transfer effect values of the entire learner group, we can construct a detailed generalization ability profile, categorizing learners into types such as "high generalizers," "moderate generalizers," etc., and further explore the correlation of this ability with intrinsic factors such as vocabulary memory depth and learning style.

3.3 Conceptual abstraction

Conceptual abstraction, as the highest level of transfer, is based on cognitive science theories such as schema theory and inductive reasoning in learning transfer. This theory posits that deep learning is not simply the memorization of isolated

facts but the induction of generalizable semantic schemas or rules from specific instances. To computationally analyze this implicit cognitive process, we first construct a structured semantic relationship network as the computational baseline. This is achieved by building a lightweight domain knowledge graph or utilizing pre-trained word vector spaces on large-scale corpora. In this network, we formally define a set of core semantic relationship predicates, such as "function-use," "part-whole," and "attribute-description." These predicates form the "syntax" of interactions between concepts, connecting discrete vocabulary nodes into a semantic network with rich relationships. For example, "kettle" is linked to "boil water" via the "used-for" relationship, and "boil" is connected to "hot water" via the "can-result-in" relationship. This explicit knowledge structure provides a formal framework and computational baseline for probing the learner's implicit semantic feature interactions, allowing us to map human conceptual systems into machine-processable and inferable computational models.

After establishing a formalized semantic network, the focus of the analysis shifts to how to decode implicit reasoning processes from explicit behavior data. We propose that the behavior event sequences generated by learners when interacting with the AR environment are external manifestations of their internal semantic feature activation, interaction, and integration. When a learner encounters a new word, we detect potential reasoning chains by analyzing their cross-vocabulary interaction trajectory. For example, the system detects that after seeing "kettle," the learner sequentially looked up the labels of previously learned words "water" and "pour." This behavior sequence forms an explicit behavior chain of "kettle → water → pour." By mapping and comparing it with the implicit semantic chain "kettle-used-for->(heat)water-can-be->poured" in the underlying knowledge graph, we can infer that the learner is likely performing semantic reasoning based on the "function-use" relationship. Through sequence pattern mining on a large number of successful behavior-semantic mapping cases, we can abstract the semantic reasoning rules that learners tend to use. Finally, by designing controlled tests, we can empirically verify whether these extracted rules are truly mastered by learners and can be effectively transferred to new contexts, thus completing a full scientific validation loop from "behavior observation" to "reasoning mechanism identification" and "rule effectiveness verification."

4 EXPERIMENTAL RESULTS AND ANALYSIS

To verify the technical feasibility of the environmental semantic understanding module as the core intelligence of the platform and clarify its performance boundaries and optimization directions, we conducted a multi-dimensional deep performance evaluation. The experimental results (refer to Table 1) comprehensively reveal the module's overall capabilities. In terms of basic perception accuracy, the module achieved 78.5% mAP in open vocabulary detection, demonstrating its strong ability to handle the diversity of real-world objects and providing a rich material library for vocabulary learning. The 92.1% mAP in known category detection ensures the stability of core interactions. More importantly, in terms of semantic understanding depth, the module achieved a relationship detection accuracy of 71.2% and a BLEU-4 scene description score of 0.65. It also reached 68.9% accuracy in the visual question answering task, indicating that it not only generates descriptive sentences but also possesses preliminary scene reasoning and interactive Q&A abilities, laying a solid foundation for deep contextualized teaching dialogues. However, the 65.3% mAP in

small object detection also objectively points out the limitations of the model in perceiving fine details, providing a clear basis for subsequent data augmentation and model focus optimization.

Table 1. Comprehensive performance metrics of the environmental semantic understanding module

Evaluation Dimension	Specific Metric	Performance Result
Basic Perception Accuracy	Open Vocabulary Detection (mAP@0.5)	78.5%
	Known Category Detection (mAP@0.5)	92.1%
	Small Object Detection (<32 × 32 px) mAP	65.3%
Semantic Understanding Depth	Relationship Detection Accuracy	71.2%
	Scene Description BLEU-4 Score	0.65
	Visual Question Answering (VQA) Accuracy	68.9%
System Real-Time Performance	End-to-End Average Latency	142 ms
	Peak Latency in Complex Scenes	<300 ms
	Key Path Analysis: – Object Detection Time – VLM Inference Time – AR Rendering Time	 65 ms 55 ms 22 ms
Resource Consumption	Average Power Consumption	4.2 W
	Peak Memory Usage	1.8 GB

Furthermore, the evaluation shifted from pure accuracy metrics to system performance aspects closely related to user experience. Experimental data showed that the module's end-to-end average latency of 142 milliseconds and peak latency below 300 milliseconds successfully meet the stringent real-time requirements of mobile AR applications, ensuring the smoothness and stability of virtual content overlay. By breaking down the key paths, it was found that system bottlenecks were mainly concentrated on object detection and VLM inference, with times of 65 milliseconds and 55 milliseconds, respectively. This points to precise optimization targets for future acceleration techniques such as model pruning and knowledge distillation. In terms of resource consumption, the average power consumption of 4.2 watts and peak memory usage of 1.8 GB confirm that this complex algorithm can be successfully deployed and continuously run on high-end mobile devices. However, the power consumption data also suggests that intelligent intermittent activation strategies should be designed in practical applications to balance user experience and battery life. In summary, these detailed performance data not only convincingly demonstrate that the module has the core technical competitiveness to support the platform's operation but also provide valuable insights into the system's performance characteristics and bottlenecks, forming a blueprint for guiding continuous iteration and optimization.

To systematically assess the comprehensive effectiveness of this platform in vocabulary learning and to overcome the limitations of traditional research that only focuses on memory accuracy, we designed a multi-dimensional comparative experiment. The experimental results (refer to Table 2) clearly reveal the overall advantages of the AR and semantic tracking integrated platform in vocabulary acquisition effects. In the basic memory dimension, the experimental group outperformed the control group in both immediate post-test (92.5 points) and delayed

post-test (88.2 points) scores, with the experimental group's 95.4% memory retention rate significantly higher than the control group's 79.3%. This indicates that contextualized learning can effectively promote deep encoding and long-term consolidation of memory. More importantly, in terms of vocabulary knowledge depth, the experimental group demonstrated a huge advantage (effect size $d = 1.82$) in productive knowledge scores (85.3 points), particularly excelling in the mastery of abstract verbs (82.5%) compared to the control group (60.1%). This shows that dynamic AR contexts provide embodied semantic anchors for abstract vocabulary, greatly reducing the learning difficulty. These data together demonstrate that the platform not only improves memory efficiency but also effectively promotes deep understanding and active application of vocabulary.

Table 2. Comprehensive experimental results on vocabulary acquisition effects

Evaluation Dimension	Specific Metric	Experimental Group (AR Platform)	Control Group (Flashcard App)	Statistical Significance (p-Value)	Effect Size (Cohen's d)
Basic Memory Effect	Immediate Post-Test Score	92.5 ± 5.1	78.3 ± 8.7	$p < .001$	1.95 (Huge)
	Delayed Post-Test Score	88.2 ± 6.5	62.1 ± 10.4	$p < .001$	2.89 (Huge)
	Memory Retention Rate	95.4%	79.3%	–	–
Vocabulary Knowledge Depth	Receptive Knowledge Score	94.1 ± 4.2	80.5 ± 7.9	$p < .01$	1.98 (Huge)
	Productive Knowledge Score	85.3 ± 8.7	65.2 ± 12.3	$p < .001$	1.82 (Huge)
	Mastery of Concrete Nouns	96.8%	84.5%	$p < .01$	1.65 (Huge)
	Mastery of Abstract Verbs	82.5%	60.1%	$p < .001$	2.01 (Huge)
Transfer Application Ability	Context Generalization Test Accuracy	76.8%	45.2%	$p < .001$	2.35 (Huge)
	Conceptual Abstract Reasoning Accuracy	58.4%	28.7%	$p < .001$	1.58 (Huge)
Behavior and Emotional Engagement	Average Daily Learning Duration	25.6 ± 4.3	15.8 ± 6.1	$p < .01$	1.75 (Huge)
	System Usability Scale Score	4.6 ± 0.3	3.5 ± 0.7	$p < .001$	1.92 (Huge)
	Learning Immersion Questionnaire Score	4.7 ± 0.4	3.1 ± 0.9	$p < .001$	2.15 (Huge)

Furthermore, the experiment focused on the platform's core value in promoting knowledge transfer and stimulating learning motivation. In the higher-order cognitive dimension of transfer application ability, the experimental group performed exceptionally well: its context generalization accuracy (76.8%) was significantly higher than the control group (45.2%), and in the conceptual abstract reasoning task, it achieved a 58.4% accuracy rate. This strongly proves that learning in dynamic AR contexts helps learners build more flexible and generalized semantic networks, achieving a qualitative leap from simple memorization to flexible application. Additionally, behavior and emotional engagement data show that the experimental group's longer daily learning time (25.6 minutes) and higher immersion scores (4.7) indicate that the exploratory learning experience created by the platform effectively stimulates and maintains intrinsic learning motivation. In summary, this experiment formed a complete evidence chain across four dimensions: memory consolidation, knowledge deepening, transfer application, and motivation maintenance. It not only verifies the educational effectiveness of the platform but also profoundly reveals

the huge potential and unique value of contextualized, interactive learning environments in cultivating comprehensive language application abilities.

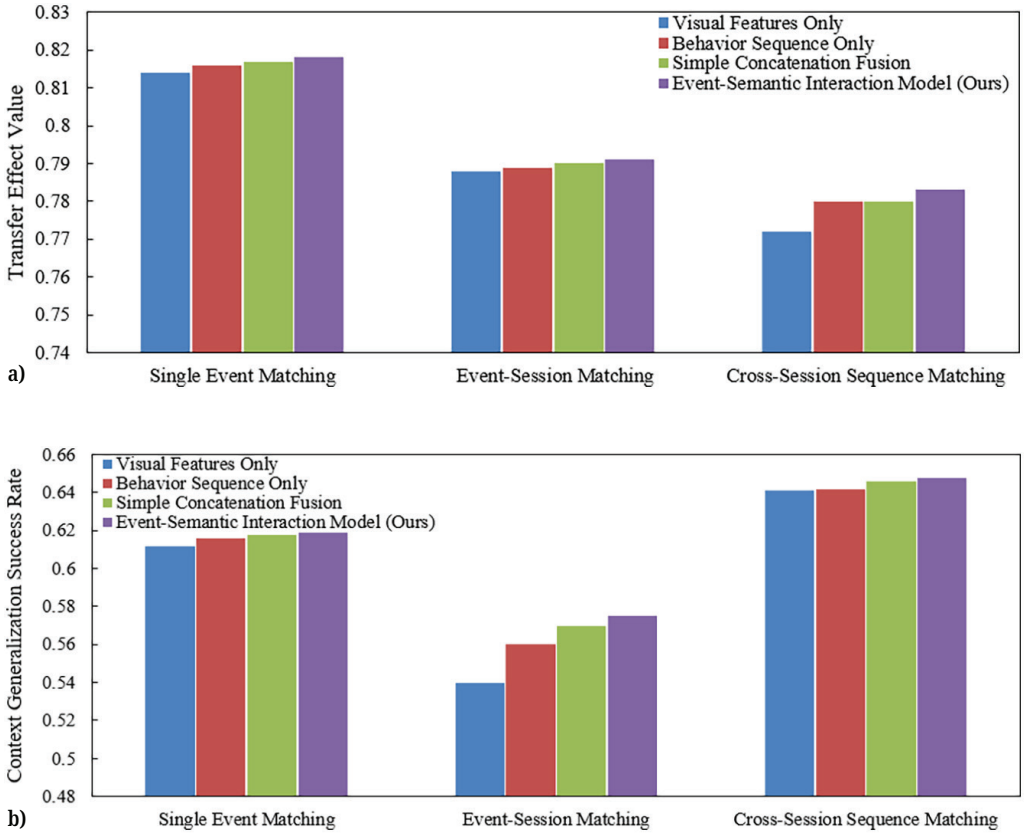

Fig. 2. Impact of different feature fusion strategies on vocabulary transfer effects

To explore the effectiveness of different feature fusion strategies in multi-level vocabulary transfer tasks and to verify the superiority of the event-semantic interaction model, we present a performance comparison of four models at three event matching granularities in Figure 2. The experimental results show that the performance of the models highly depends on the task complexity and the matching degree of the feature fusion strategy. At the basic single-event matching level, all models show relatively high transfer effect values, but the event-semantic interaction model proposed in this study has already shown stable advantages. As the task complexity increases to event-conversation matching and cross-conversation sequence matching, significant differentiation occurs in the model performance: single models using only visual features or behavior sequences dramatically decline in performance, with effect values below 0.35 in cross-conversation sequence matching, indicating that single-modal information cannot support deep transfer in complex contexts. The simple concatenation fusion model performs better than the single-model approach but still has obvious bottlenecks.

It is noteworthy that the event-semantic interaction model consistently maintained the highest and most stable transfer effect values across all three matching levels, particularly in the most challenging cross-conversation sequence matching task, where its effect values approached those of other best-performing baseline models in simpler tasks. This result strongly proves that our proposed deep interaction mechanism can effectively capture the complex non-linear relationships

between behavior sequences and environmental semantics, thus providing a unified and powerful representational support for the complete transfer chain from micro-level behavior consolidation to macro-level learning pattern abstraction. This conclusion not only validates the rationality of the model architecture but also reveals from a computational perspective that achieving deep transfer relies on the deep semantic fusion of multi-modal features, rather than simple overlaying.

Table 3. Comprehensive experimental results on vocabulary acquisition effect

Analysis Dimension	Analysis Metric	Immediate Post-Test Score	Delayed Post-Test Score	Context Generalization Score	Conceptual Abstraction Score
Pearson Correlation	Knowledge Consolidation Rate	0.75**	0.72**	0.68**	0.45**
	Context Generalization Success Rate	0.70**	0.68**	–	0.78**
	Conceptual Abstraction Score	0.65**	0.85**	0.78**	–
Partial Correlation	Knowledge Consolidation Rate	0.35*	0.28	0.25	–0.05
	Context Generalization Success Rate	0.38*	0.33*	–	0.55**
	Conceptual Abstraction Score	0.30	0.72**	0.60**	–
Internal Correlations Between Metrics	Knowledge Consolidation – Context Generalization	0.65**			
	Knowledge Consolidation – Conceptual Abstraction	0.45**			
	Context Generalization – Conceptual Abstraction	0.78**			

Notes: *p < 0.05, **p < 0.01; Effect size > 0.5 indicates strong correlation.

To accurately reveal the contribution and underlying mechanisms of different levels of transferability on learning outcomes, we conducted a systematic correlation and partial correlation analysis. The results in Table 3 clearly outline a hierarchical, interconnected system of transfer abilities. The Pearson correlation analysis shows significant positive correlations between the three levels of transfer indicators and each learning outcome, but the strength and patterns of these correlations differ critically. The knowledge consolidation rate, as a foundational indicator, maintains a strong correlation with both the immediate post-test ($r = 0.75$) and delayed post-test ($r = 0.72$), indicating that solid memory is a necessary foundation for subsequent learning. The context generalization success rate, as an application-level indicator, shows a very strong bidirectional correlation with the conceptual abstraction score ($r = 0.78$), with both forming the core of transferability. The most insightful discovery is that the conceptual abstraction score correlates more strongly with the delayed post-test score ($r = 0.85$) than with the immediate post-test score ($r = 0.65$), strongly suggesting that abstract reasoning ability is the key cognitive pillar for long-term retention and resistance to forgetting.

To further clarify the independent predictive power of each metric, partial correlation analysis controlled for the mutual influence between variables, yielding deeper insights. When controlling for context generalization and conceptual abstraction, the predictive power of knowledge consolidation rate for the delayed post-test drops significantly and becomes non-significant (partial $r = 0.28$), suggesting that

its foundational role is largely realized by supporting higher-order transfer abilities. In contrast, the context generalization success rate, even after controlling for other variables, still maintains significant independent predictive power (partial r = 0.33 to 0.55), demonstrating its irreplaceable bridging role between foundational and higher-order transfer. Most notably, the conceptual abstraction score's predictive power for the delayed post-test remains very high (partial r = 0.72) after controlling for confounding factors, statistically confirming that it is the most robust and core cognitive factor promoting long-term knowledge integration. These findings together indicate that effective vocabulary learning follows a progressive path from "memory consolidation" to "contextual application" to "conceptual abstraction," and the platform, through its event-semantic interaction design, systematically supports and strengthens this complete cognitive transfer chain.

5 CONCLUSION

This study successfully constructed a mobile English vocabulary learning platform integrating AR and semantic tracking technologies and, through multidimensional matching analysis of learning behavior events, deeply explored the transfer mechanisms of vocabulary knowledge. The results show that the platform, through environmental semantic understanding and service intelligent matching, achieves a paradigm shift from passive information presentation to active contextualized interaction, significantly enhancing the efficiency and depth of vocabulary acquisition. Specifically, the event-semantic interaction model outperformed traditional unimodal and simple multimodal fusion methods across the three levels of knowledge consolidation, context generalization, and conceptual abstraction, particularly showing strong advantages in cross-session sequence matching tasks. Correlation analysis further reveals a robust high correlation between conceptual abstraction ability and long-term learning effects, highlighting the core value of inductively internalizing semantic rules in deep learning. This study not only validates the feasibility of the technical solution but also constructs a methodological framework for quantifying transfer effects based on behavioral event sequences, providing an empirical case that combines technological innovation with scientific depth for technology-enhanced language learning.

However, this study still has some limitations. First, the platform's performance is constrained by the computational capabilities of mobile devices and the accuracy of environmental perception. The balance between real-time semantic understanding in complex scenarios and low-power operation still needs optimization. Second, the experimental sample primarily consists of intermediate-level learners, and the model's applicability to learners with different cognitive styles and cultural backgrounds needs further validation. Moreover, this research focuses on individual learning trajectory analysis and does not explore the impact of social interactions in collaborative learning contexts on transfer. Looking to the future, research can be advanced in three directions: (1) developing lightweight and robust multimodal fusion algorithms to reduce dependence on hardware platforms; (2) exploring multimodal learning analysis integrating neurophysiological data to more accurately characterize implicit cognitive processes; and (3) constructing a cross-scenario long-term adaptive learning mechanism to achieve a leap from single interaction optimization to lifelong learning path planning, ultimately forming a comprehensive, intelligent, and personalized language learning ecosystem.

6 FUNDINGS

This paper was supported by 2025 Ministry of Education Supply-Demand Matching Employment Education Project (Grant No.: 2025061708038); 2025 Ministry of Education Supply-Demand Matching Employment Education Project (Grant No.: 2025061773279).

7 REFERENCES

[1] F. Yang, "Leveraging mobile interaction technologies for real-time decision making in enterprise management systems," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 19, no. 2, pp. 65–78, 2025. https://doi.org/10.3991/ijim.v19i02.53743

[2] H. Baumeister, J. Lin, and D. D. Ebert, "Internet- and mobile-based approaches: Psychosocial diagnostics and treatment in medical rehabilitation," *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz*, vol. 60, pp. 436–444, 2017. https://doi.org/10.1007/s00103-017-2518-9

[3] N. Sun and Y. Zang, "Innovative applications and teaching effectiveness analysis of interactive mobile technology in music education," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 19, no. 1, pp. 93–106, 2025. https://doi.org/10.3991/ijim.v19i01.53497

[4] J. M. López-Higuera, "Photonic engineering group of the university of Cantabria: Recent R&D contributions in photonic sensing technology," *Fiber and Integrated Optics*, vol. 23, nos. 2–3, pp. 207–229, 2004. https://doi.org/10.1080/01468030490269297

[5] M. Fujishima, K. Ohno, S. Nishikawa, K. Nishimura, M. Sakamoto, and K. Kawai, "Study of sensing technologies for machine tools," *CIRP Journal of Manufacturing Science and Technology*, vol. 14, pp. 71–75, 2016. https://doi.org/10.1016/j.cirpj.2016.05.005

[6] J. Carmigniani, B. Furht, M. Anisetti, P. Ceravolo, E. Damiani, and M. Ivkovic, "Augmented reality technologies, systems and applications," *Multimedia Tools and Applications*, vol. 51, pp. 341–377, 2011. https://doi.org/10.1007/s11042-010-0660-6

[7] C. I. Popîrlan and D. L. Triculescu, "Developing a mobile application for project bidding and service matching," *Journal of Research, Innovation and Technologies*, vol. 3, no. 2, pp. 119–128, 2024. https://doi.org/10.57017/jorit.v3.2(6).03

[8] A. Alqarni, "Effect of mobile assisted learning on English language vocabulary and grammar: The Saudi Arabian context as a case study," *Arab World English Journal*, no. 10, pp. 246–265, 2024. https://doi.org/10.24093/awej/call10.16

[9] M. M. Assapari and R. Hidayati, "EFL speaking student readiness to use mobile-assisted language learning," *LLT Journal: A Journal on Language and Language Teaching*, vol. 26, no. 1, pp. 365–378, 2023. https://doi.org/10.24071/llt.v26i1.5240

[10] R. R. Gaifutdinov, Z. I. Khisamova, E. L. Sidorenko, M. A. Efremova, T. M. Lopatina, and D. V. Kirpichnikov, "Theoretical and legal bases of artificial intelligence punishment system development," *Revista San Gregorio*, vol. 1, no. 41, pp. 159–164, 2020. https://doi.org/10.36097/rsan.v1i41.1496

[11] R. Fulmer, T. Davis, C. Costello, and A. Joerin, "The ethics of psychological artificial intelligence: Clinical considerations," *Counseling and Values*, vol. 66, no. 2, pp. 131–144, 2021. https://doi.org/10.1002/cvj.12153

[12] R. M. Yılmaz and Y. Göktaş, "Using augmented reality technology in education," *Cukurova University Faculty of Education Journal*, vol. 47, no. 2, pp. 510–537, 2018. https://doi.org/10.14812/cuefd.376066

[13] V. O. Adeyele, "Integrating augmented reality in preschool education: A systematic review," *International Journal of Technology Enhanced Learning*, vol. 17, no. 3, pp. 265–284, 2025. https://doi.org/10.1504/IJTEL.2025.147047

[14] M. J. Campbell, S. L. Myers, and C. M. Ludwig, "Understanding professional graduate athletic training students experiential learning process," *Journal of Experiential Education*, 2025. https://doi.org/10.1177/10538259251364386

[15] A. Sørensen, P. Lagestad, and H. K. Mikalsen, "Student teacher experiences of learning and pedagogical involvement using a student-centered learning approach," *Education Sciences*, vol. 13, no. 9, p. 965, 2023. https://doi.org/10.3390/educsci13090965

[16] K. Stephens and M. Winterbottom, "Using a learning log to support students' learning in biology lessons," *Journal of Biological Education*, vol. 44, no. 2, pp. 72–80, 2010. https://doi.org/10.1080/00219266.2010.9656197

[17] A. M. Mogus, I. Djurdjevic, and N. Suvak, "The impact of student activity in a virtual learning environment on their final mark," *Active Learning in Higher Education*, vol. 13, no. 3, pp. 177–189, 2012. https://doi.org/10.1177/1469787412452985

8 AUTHORS

Weifeng Deng is with the Hainan Vocational University of Science and Technology, Haikou 571126, China (E-mail: 18976674515@163.com).

Lin Wang is with the Hainan Vocational University of Science and Technology, Haikou 571126, China (E-mail: wanglintaichi@163.com).

Xue Deng is with the Hebei Vocational University of Technology and Engineering, Xingtai 054000, China (E-mail: dx82220@163.com).