Personalized Vocabulary Memory Curve Optimization Model Based on Deep Reinforcement Learning

Jun Wang
School of Foreign Languages, Inner Mongolia Minzu University, Tongliao 028000, Inner Mongolia Autonomous Region, China
Corresponding author: Jun Wang (e-mail: bgxd9805@163.com)

ABSTRACT Stable vocabulary retention is essential in university English learning, but individual cognitive differences often produce unstable memory curves. This paper proposes a personalized memory optimization model based on deep reinforcement learning to construct adaptive review strategies. An LSTM network encodes time-series behavioral data, including accuracy, reaction time, and review time, into a 128-dimensional latent state vector that captures dynamic memory characteristics. Memory strength is mapped into a four-level discrete state space, including newly learned, weakened, consolidation, and stable states, while five optional review timings form the action space. Proximal Policy Optimization learns the optimal review policy by optimizing a reward function that balances retention gain against learning-load penalty. The closed-loop process continuously incorporates learning feedback. Experiments show the highest accuracy of 89.2%–93.1% with low review frequency of 4.3–5.8 times. Long-term retention is also improved, with the average vocabulary retention rate on day 60 not lower than 78.3% and the average memory fluctuation coefficient not higher than 0.18. The proposed method alleviates retention instability caused by cognitive differences and variable learning pace, providing a personalized sequential-decision framework for adaptive learning systems.

INDEX TERMS deep reinforcement learning, LSTM, proximal policy optimization, memory curve, adaptive review

I. INTRODUCTION

THE core challenge of college English learning lies in efficiently mastering a large number of academic and interdisciplinary vocabulary, which requires not only memorization but also precise use in reading, writing, and expression. Similarly, in the industry, the core challenge for new engineers is efficiently mastering a large number of specialized technical terms and complex procedural knowledge. This specialized knowledge requires not just retention, but also precise application in machine operation, quality control, and technical documentation [1,2]. Unlike the basic stage, college students' cognitive maturity and learning goals are highly differentiated, and the memory process is affected by individual knowledge background, review rhythm, and psychological state. Precisely optimizing the memory curve is not only related to the systematic accumulation of language skills, but also directly affects the formation of academic competitiveness and cross-cultural expression ability, and has important theoretical value and practical significance. The "perception-decisionfeedback" closed-loop optimization framework constructed by this method can be extended to personalized skills training and knowledge maintenance systems in industry and other fields due to its strong adaptability to individual differences and dynamic changes.

Current vocabulary learning systems still have significant modeling blind spots in dealing with individual differences. Most platforms rely on fixed periods or empirical formulas to set review nodes, ignoring the nonlinear change characteristics exhibited by learners in the process of knowledge absorption [3,4]. For example, the same word may be forgotten at vastly different rates in different students. This difference is reflected not only in overall speed but also in multiple dimensions, such as fluctuation trends, recovery capacity, and critical response [5,6]. Ideal learning scheduling should maintain memory strength while minimizing cognitive input, but current methods lack the ability to optimize this dynamic balance. The lack of a feedback loop also limits the model's adaptability, making it difficult for the system to adjust its prediction logic in real-time based on newly generated behavioral data, causing the strategy to lag behind the actual learning process.

Early research on memory patterns focused on group average behavior, and the Ebbinghaus forgetting curve laid the foundation for the relationship between time intervals and memory retention rates [7,8]. Subsequently, Piotrowski

M proposed a systematic content editing scheme based on the SuperMemo interval repetition algorithm and strictly following the curriculum standards to address the problem of weak historical knowledge among Polish students, in order to optimize learning efficiency and knowledge coherence. This scheme was based on theoretical exploration and experimental verification, and was further improved through practical editing experience [9]. Although these methods have made progress in specific scenarios, overall, they generally lack the ability to deeply represent high-dimensional temporal behavioral data, making it impossible to continuously track and dynamically control memory states. Moreover, they do not formalize the learning strategy optimization problem as a sequential decision task, which restricts the adaptive performance of the system in real-world complex environments.

This research aims to build an intelligent learning system that can precisely perceive, dynamically respond to, and continuously optimize individual vocabulary memory traces. The core goal is to overcome the limitations of existing models that simplify the memory process and establish a personalized review scheduling mechanism that combines representational depth and decision intelligence. The innovations are reflected in four aspects: first, an LSTM network is used to perform end-to-end encoding of the learner's historical answer accuracy, reaction time, and review time series, generating a 128-dimensional time-series latent state vector, fully exploring the deep dynamic patterns in the behavioral data, and realizing a high-dimensional representation of the memory evolution process; second, a memory state discretization scheme based on semantic division is proposed, which divides the continuous memory strength into four categories: new learning, weakening, consolidation, and stability, constructs a state space with a clear structure and semantically interpretable, and provides a reliable environment foundation for strategy learning; third, an action space with five specific time options is designed, giving the intelligent agent fine-grained control over the timing of review, and combining the clipping mechanism of the PPO algorithm to ensure the stability of strategy updates and improve learning efficiency and convergence speed; fourth, a compound reward function that takes into account both memory retention gain and learning load penalty is constructed to guide the model to achieve a balance between long-term memory maintenance and cognitive resource conservation, and a closed-loop feedback mechanism is formed through periodic parameter fine-tuning to enhance the system's adaptability. The framework's capacity for full-chain automated optimization addresses core challenges in sectors defined by complex, sequential transformations, such as both the industry and intelligent education. By orchestrating the entire journey from initial input to a customized final product—whether a bespoke a personalized learning trajectory—it significantly enhances the level of recommendation personalization. Furthermore, this cross-disciplinary success provides a robust new modeling basis for tackling sequential decision problems within intelligent education systems.

II. CONSTRUCTION OF PERSONALIZED MEMORY SCHEDULING ARCHITECTURE BASED ON DEEP REINFORCEMENT LEARNING

Figure 1 illustrates the personalized vocabulary memory optimization model based on deep reinforcement learning. Starting with multi-dimensional behavioral data input, the system extracts 128-dimensional temporal features using an LSTM network and attention mechanism to encode memory dynamics. The features are mapped to four discrete states: new learning, weakening, consolidation, and stability through the state classifier. The policy network generates review actions at five time scales based on the state, obtains reward signals that balance memory gain and cognitive load through environmental interaction, and drives the PPO algorithm to optimize the strategy. An online feedback loop continuously monitors prediction errors and triggers model fine-tuning, forming a complete adaptive optimization loop that precisely regulates the personalized vocabulary memory curve.

A. LSTM FEATURE ENCODING OF LEARNER BEHAVIOR SEQUENCES

1) Sequential Modeling of Multimodal Time Series Data and LSTM Encoding Structure Design

To achieve a high-dimensional representation of learners' memory dynamics, the system continuously collects multi-source behavioral data streams during the vocabulary learning process, including accuracy rates on daily tests, individual reaction times, and the timestamps and frequency of active review behaviors. This data is organized into equally spaced sequences aligned with time steps, forming a structured time series input matrix. The system employs a fixed-length sliding time window to extract historical behavioral sequences, with the LSTM input sequence spanning 14 consecutive learning days (time steps). Each time step contains the day's average accuracy, average response time, and a binary flag indicating whether a review occurred. Each time step corresponds to a feature vector containing multiple observation dimensions. Accuracy rates reflect explicit knowledge mastery; reaction times serve as a proxy for implicit cognitive load; review frequency and its intervals reflect learners' self-regulation strategies. This sequence is fed into a three-layer stacked LSTM network, each containing 128 memory units. The input eigenvector explicitly contains the time interval since the last review to ensure that the model can accurately perceive the temporal dynamics of memory decay. A gating mechanism is used to extract temporal dependency features layer by layer. The input, forget, and output gates work together to selectively retain or suppress the transmission paths of historical information, effectively alleviating the vanishing gradient problem of traditional recurrent networks when processing long sequences. The network gradually advances in the time dimension, capturing

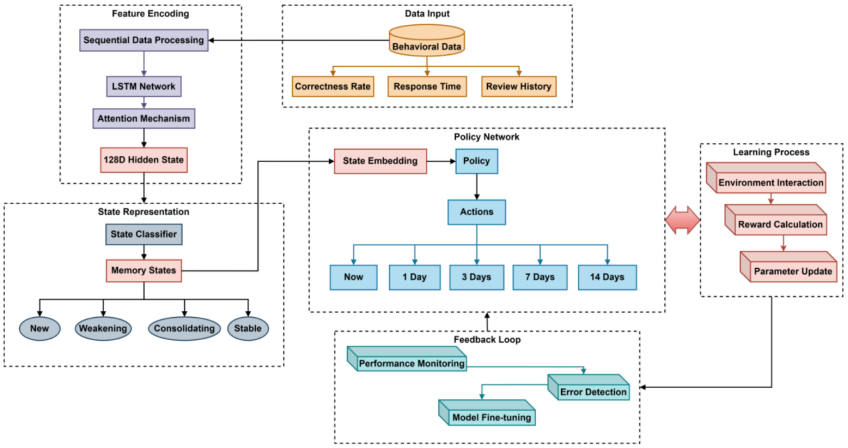

FIGURE 1. Personalized vocabulary memory optimization model

the memory fluctuation patterns of individuals at different stages, such as the rapid decline after short-term memory enhancement or the gradual consolidation trend caused by multiple repetitions. Finally, at the end of the sequence, the hidden state vector output by the LSTM is extracted as a 128-dimensional dense representation. This vector integrates the complete memory evolution trajectory from the initial learning to the current moment and serves as the state input of the subsequent decision module. The encoding process can be formally expressed as follows:

$$h_t = \mathrm{LSTM}(x_t, h_{t-1}) \tag{1}$$

x_t represents the input feature vector of the t-th time step, and h_t is the hidden state at the corresponding moment, with a dimension of 128, which carries the accumulated memory dynamic information up to that moment. This state vector is not a static snapshot, but a dynamically evolving cognitive representation that can sensitively respond to changes in the learning rhythm, such as the sudden drop in memory strength after continuous errors or the potential risk of forgetting due to long-term lack of review.

2) Memory Semantic Decoupling and State Sensitivity Enhancement of the Hidden State Vector

To improve the latent state vector's ability to discriminate changes in memory strength, the LSTM network optimizes the objective function through end-to-end backpropagation during training, aligning its latent space distribution with the subsequent memory state classification task. To further enhance its ability to identify critical memory states, a temporal attention mechanism is applied to weightedly aggregate the hidden state of the LSTM at each time step, focusing the model on the most discriminative historical nodes. This attention weight is calculated from the similarity between the learnable context vector and the hidden state at each time step, ensuring that the model prioritizes historical behaviors that have the greatest impact on the current memory strength when generating the final representation. The final output 128-dimensional vector not only encodes the macro trend of memory, but also embeds the fine-grained response capability to key turning points. This high-dimensional latent state, as the core representation of the memory system, directly serves the subsequent division of state space and strategic decision. Its continuity and richness provide a solid data foundation for personalized scheduling. The weighted aggregation processes can be expressed as:

$$s = \sum_{t=1}^{T} \alpha_t h_t \tag{2}$$

$$\alpha_t = \frac{\exp\left(v^\top \tanh(W h_t)\right)}{\sum_{k=1}^{T} \exp\left(v^\top \tanh(W h_k)\right)} \tag{3}$$

s is the final output state vector; α_t is the attention weight of the t-th time step; W is the projection matrix; $v^\top$ is the learnable attention context vector. This mechanism strengthens the model's ability to perceive critical points of memory and improves the decision utility of state encoding.

TABLE 1. LSTM feature encoding training hyperparameters

Parameter Category	Parameter Name	Value	Description
Regularization	Inter-layer Dropout Rate	0.3	Mitigates overfitting during training
Optimizer Configuration	Initial Learning Rate	0.001	Controls step size of parameter updates
Optimizer Configuration	β_1	0.9	Exponential decay rate for first moment
Optimizer Configuration	β_2	0.999	Exponential decay rate for second moment
Optimizer Configuration	ϵ	1e-8	Small constant for numerical stability
Training Settings	Batch Size	64	Number of sequences per training iteration
Training Settings	Maximum Epochs	100	Upper limit of training iterations
Training Settings	Clipping Threshold	1	Prevents gradient explosion

Table 1 lists the key training hyperparameters in the LSTM feature encoding module, including regularization,

optimizer configuration, and training process settings. The dropout rate is set to 0.3 to enhance model generalization. The Adam optimizer uses the standard momentum parameter, and the learning rate is initialized to 0.001, balancing convergence speed and stability. The gradient clipping threshold is set to 1.0 to prevent gradient explosion during training. The batch size is 64, balancing computational efficiency and gradient estimation accuracy. The maximum number of training epochs is 100, and an early stopping mechanism is used to prevent overfitting. All parameters are set based on deep learning practices to ensure robust training and reproducibility of the model in modeling learner behavior sequences, providing high-quality state representation support for subsequent decision modules.

B. MULTIDIMENSIONAL DISCRETIZATION MODELING OF THE MEMORY STATE SPACE

1) Construction of Memory State Classification Mechanism Based on Two-Dimensional Criterion

To achieve a semantically interpretable representation of learner memory strength, the system uses a two-dimensional joint criterion to structurally classify memory states into four discrete states: "newly learned", "weakened", "consolidation", and "stable". The classification relies primarily on two core variables: the first is the time span since the most recent effective review, denoted as d (in days), reflecting the duration of exposure to forgetting; the second is an estimated forgetting probability derived from historical behavioral encoding, denoted as p_f. This value is generated by nonlinearly mapping the 128-dimensional latent state vector output by the previous module and represents the strength of the propensity for the knowledge item to be forgotten at the current moment. Specifically, the p_f is obtained by nonlinearly mapping the 128-dimensional LSTM hidden states using a single-layer fully connected network followed by a Sigmoid activation function. Its supervision signal originates from the binary accuracy labels (correct = 0, incorrect = 1) of the day's vocabulary test, and end-to-end training is performed using mean squared error (MSE) loss. To avoid label leakage, only behavioral data from the current and historical time steps are used during training; future labels are not visible. These two variables together constitute a two-dimensional criterion space, and deterministic state classification is achieved by setting segmentation thresholds. When d is less than or equal to 2 and p_f is greater than 0.7, the state is classified as "newly learned", indicating that the vocabulary is in the initial memorization stage and has not yet undergone sufficient consolidation, and the cognitive representation is fragile. When d is greater than 2 and p_f continues to rise to above 0.8, the state is classified as "weakened", indicating that the memory trace is rapidly decaying and on the verge of critical forgetting. If d is in the range [3, 10] and p_f significantly decreases below 0.3, the state is labeled as "consolidation", reflecting that the memory structure has tended to strengthen after timely intervention. When $d > 10$ and $p_f < 0.2$, the system enters a "stable" state, indicating that the word has formed a long-term memory representation and the risk of forgetting it in the short term is extremely low. For samples not falling into the above four categories (e.g., $d > 10$ but $0.2 \leq p_f < 0.3$), the system classifies them into the fifth category, "Transition," indicating that the memory is at a vague or unstable boundary and requires careful scheduling. Although semantically independent, this state still participates in policy decision-making as part of the discrete state space, ensuring that the environment is defined for all inputs. The above state division threshold is based on the time dynamics of the classical forgetting curve and calibrated by the cluster analysis of memory retention behavior on the training set to ensure that each state has distinguishable statistical characteristics in terms of time span and forgetting trend. This classification process is not based on a fixed schedule but dynamically responds to the individual's learning trajectory, ensuring that the state classification remains synchronized with the actual cognitive state. It is worth emphasizing that although p_f is an endogenous estimator of the model, its effectiveness can be verified by the performance of downstream tasks: if p_f cannot accurately reflect the true forgetting tendency, the state space constructed based on it will be invalid, leading to a significant decline in policy performance. In our ablation experiment, we removed p_f (using only the time interval d to divide the states), and the results showed that the retention rate decreased by 12.7% and the volatility coefficient increased by 0.11 on day 60, proving that p_f provides irreplaceable discriminative information.

Specifically, the K-means++ algorithm is used to cluster the two-dimensional samples composed of d and p_f on the training set. The optimal number of clusters is determined by combining the elbow rule and the silhouette coefficient (mean = 0.63), which is consistent with the preset semantic state. The final threshold is the median of the boundaries of each cluster. It should be emphasized that the state function only provides structured observations; the true "memory understanding" is reflected in how the PPO policy network dynamically selects actions based on this state—the policy is learned end-to-end through interaction with the environment, rather than executing preset rules.

Figure 2 shows the distribution of memory states in the two-dimensional criterion space, with the horizontal axis representing the number of days since the last review, and the vertical axis representing the probability of forgetting. Each dot represents a learner's memory state at a specific moment, with colors used to delineate different memory state regions, including "newly learned", "weakening", "consolidation", and "stable". As can be seen, when the review interval is short and the forgetting probability is high, the sample points are concentrated in the "newly learned" region. This is consistent with cognitive psychology expectations.

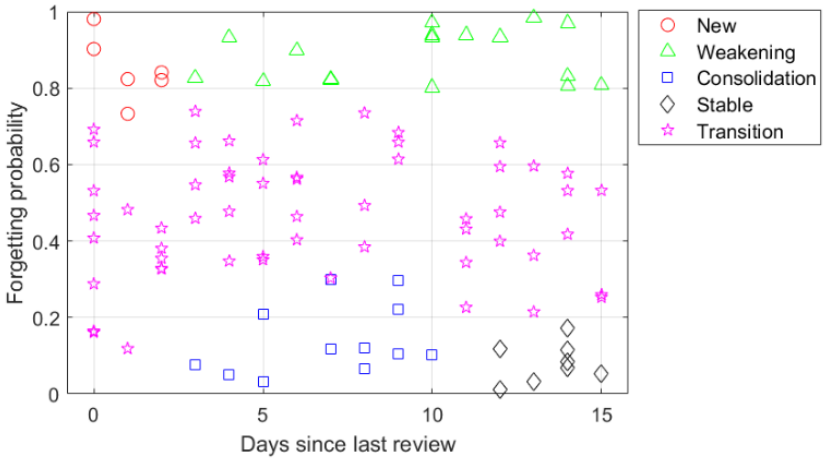

FIGURE 2. Distribution of memory states in the two-dimensional criterion space

2) Discretization Space Construction and Decision Support Mechanism Driven by State Semantics

It should be clarified that while the state space is discrete to facilitate efficient policy exploration in RL, it is meaningfully distilled from the LSTM-encoded latent features. Specifically, p_f acts as a soft-bridge that compresses high-dimensional cognitive dynamics into interpretable indicators. Although the final state s_t involves threshold-based discretization (which is non-differentiable), this design provides the RL agent with a highly structured observation space, effectively reducing the search complexity. To maintain learning consistency, the LSTM encoder and the PPO policy are trained in a decoupled yet synergistic manner: the encoder minimizes the p_f prediction error to ensure representational accuracy, while the policy maximizes the long-term reward based on these high-fidelity observations.

After completing state discrimination, the system maps the continuous memory evolution process into a discrete state space of a finite state Markov process, with each state carrying specific cognitive semantics and intervention logic. This discretization strategy is not a simple dimensionality reduction, but rather a structural abstraction of memory dynamics through semantic clustering, enabling the decision model to reason at a high-level semantic level. To enhance the robustness of state division, the p_f change trend within the sliding time window is applied as an auxiliary criterion to avoid state misjudgment due to single measurement noise. Specifically, the slope of the forgetting probability at the past three time points is calculated:

$$\Delta p_f = \frac{1}{2}\sum_{i=1}^{2}\left(p_f^{(t-i+1)} - p_f^{(t-i)}\right) \tag{4}$$

$p_f^{(t-i)}$ represents the estimated forgetting probability i days ago. This dynamic correction mechanism improves the temporal consistency of state recognition. In addition, each type of state is preset with a corresponding state activation threshold and minimum residence time to prevent the state from oscillating frequently in adjacent time steps and ensure the stability of the decision environment. The final state space constructed not only has clear semantic boundaries, but also embeds the ability to make forward-looking judgments on the direction of memory evolution. This discrete state, as the core observation input of the reinforcement learning environment, provides a structured decision context for the policy network, enabling it to select differentiated intervention strategies based on the cognitive characteristics of different states. For example, the "weakened" state triggers a high-priority review action, while the "stable" state guides the system to extend the intervention interval, thereby achieving refined scheduling control based on cognitive semantics. The state transition process satisfies:

$$s_t = f(d_t, p_f^t, \Delta p_f) \tag{5}$$

f is the preset classification function, and s_t represents the state category at time t. This mechanism ensures that the state space retains key dynamic information while also having the semantic interpretability and structural stability required for policy learning. The state transition is dynamically calculated by the environment according to the real-time behavior of the learner. The strategy network selects actions by observing the current state, so as to adapt to the state transition law indirectly, rather than explicitly modeling the transition probability matrix.

C. DYNAMIC REVIEW STRATEGY GENERATION BASED ON THE PPO ALGORITHM

1) Construction of a Fine-Grained Temporal Action Space and Implementation of the Policy Network Architecture

It should be clarified that this study adopts an offline simulated reinforcement learning paradigm based on historical behavior logs. Rather than relying solely on coarse discrete memory labels, the policy network operates on a hybrid state representation s_t, which is designed to internalize both dynamic memory trends and interpretable cognitive

stages. Specifically, s_t is constructed by concatenating the high-level semantic memory state derived from Formula (5)—providing categorical anchors for action semantics and reward allocation—with condensed features extracted from the LSTM hidden state h_t, which capture fine-grained, individualized memory dynamics. Through this hybrid formulation, the PPO policy can perform optimal review scheduling while preserving rich personalized information, maintaining interpretability, and satisfying the rigor of the Markov Decision Process (MDP). To precisely control review timing, the system constructs a discrete but temporally resolvable action space $\mathcal{A} =$ { Instant Review, 1 Day Review, 3 Days Review, 7 Days Review, and 14 Days Review } , covering typical intervention windows from immediate reinforcement to long-term consolidation. It should be noted that all actions are daily scheduling instructions. For example, "14-day review" means that no intervention will be given and the system will be scheduled to trigger 14 days later. During this period, the system will still process other words every day, which is in line with the event-driven sparse decision-making paradigm. This design ensures policy interpretability while enabling the agent to fine-grainedly adjust the pace of memory intervention. Each action corresponds to a preset delayed scheduling instruction. At each decision moment, the policy network outputs a probability distribution for action selection based on the current memory state. The policy function is implemented using a deep neural network. Its input is the discrete memory state generated by the preceding module. After being mapped to a low-dimensional vector by the embedding layer, it undergoes a nonlinear transformation through a two-layer fully connected network. The final output is a five-dimensional logits vector, which is then softmax-normalized to form the action selection policy $\pi_\theta(a_t \mid s_t)$. This policy network models the adaptive response mechanism of the individual's learning rhythm in a parameterized form. Its training process is achieved by maximizing the long-term cumulative reward. To improve the stability of policy updates, the PPO (Proximal Policy Optimization) algorithm framework is applied. A proxy objective function is constructed in each policy iteration to constrain the policy update amplitude and avoid performance fluctuations caused by excessive step size. Specifically, the policy ratio is defined as follows:

$$r_t(\theta) = \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)} \tag{6}$$

$\pi_\theta(a_t \mid s_t)$ is the action probability under the current policy parameter θ, and $\pi_{\theta_{\text{old}}}$ is the target policy fixed in the previous iteration. This ratio is used to measure the relative change in policy updates and serves as the input of the clipping mechanism. By setting the clipping interval, the deviation of the policy update from the old policy is limited, ensuring smooth convergence of the learning process. The experiments used a standard

PPO configuration: pruning range ϵ=0.2, discount factor γ=0.99, GAE parameter λ=0.95; after collecting 2048 steps of experience in each round, 10 policy updates were performed using a batch size of 64; both the policy and value networks are two-layer fully connected structures with 128 neurons per layer and ReLU activation.

Figure 3 illustrates the impact of different review strategies on memory retention and learning load. Figure 3(a) shows that the immediate review strategy achieves the highest memory retention rate, with the flattest memory decay curve, indicating that high-frequency and timely intervention is most effective in maintaining memory strength. However, this strategy is accompanied by the highest learning load. As shown in Figure 3(b), immediate review and review after 1 day lead to higher cognitive burdens, reflecting that frequent review requires learners to invest a lot of cognitive resources, which can easily lead to learning fatigue. In contrast, the review strategies after 7 days and 14 days, while having lower memory retention rates, also significantly reduce learning load, indicating that moderately extending the review interval can effectively reduce learners' cognitive pressure while sacrificing some memory effect, allowing learners to consolidate memory in a more relaxed state. This result reveals the inherent trade-off between memory retention and learning load: more frequent review is more conducive to memory, but also requires higher cognitive resources; while sparse review can reduce the load, but at the cost of faster memory decay.

2) Multi-objective Balanced Reward Function Design and Long-Term Optimization Mechanism

To guide the policy network to achieve a dynamic trade-off between memory retention and learning efficiency, a composite reward function is designed, which comprehensively considers the short-term benefits of the intervention and the long-term cognitive load. It should be emphasized that the reward function does not presuppose any superiority or inferiority of actions; its value is entirely generated by the LSTM's real-time prediction of individual memory dynamics. The reward signal is composed of a linear weighted memory retention gain term and a learning load penalty term, as follows:

$$R_t = \alpha \cdot G_t - \beta \cdot L_t \tag{7}$$

G_t represents the expected improvement in memory strength after the current action is executed, which is approximately estimated by the decrease in the probability of forgetting in the next time step after the action is executed, reflecting the positive impact of the intervention on the memory trajectory. L_t is the learning load cost, which is quantified according to the expected review frequency corresponding to the action. Immediate or short-term review corresponds to a higher penalty, while long-interval actions have a lower load. α and β are adjustable weight coefficients used to control the optimization preference between memory effect and learning load. In the experiment, $\alpha = \beta = 1.0$ was set, and grid search verified that this configuration

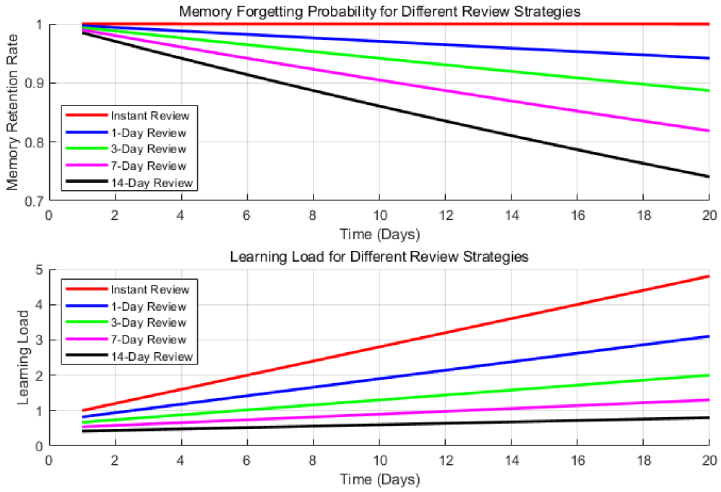

FIGURE 3. Effects of different review strategies on memory retention and learning load

achieved the best balance between memory gain and load. The reward function is generated in real-time with environmental feedback in each decision cycle, forming the driving signal for policy optimization. To capture the long-term memory evolution trend, the generalized advantage estimation (GAE) method is used to calculate the advantage function, and the state value $V_\theta(s_t)$ is fitted in combination with the value function network to improve the variance control ability of the policy gradient estimation. The value function is updated synchronously through the mean square error loss:

$$L_V = E_t\left[\left(V_\theta(s_t) - R_t^{\text{total}}\right)^2\right] \tag{8}$$

R_t^{total} is the multi-step reward estimation. The entire training process alternately optimizes the policy network and the value function network on batch data to ensure that the policy update responds to immediate feedback while taking into account the long-term memory maintenance goal. Through this mechanism, the model gradually learns to prioritize short-term and medium-term review when the memory is about to decline, and actively extend the interval when the memory is stable, forming an adaptive scheduling behavior that conforms to cognitive laws. The final policy output not only responds to the current state, but also has the ability to make forward-looking judgments on the memory evolution path, achieving true dynamic optimization.

III. MULTI-DIMENSIONAL PERFORMANCE COMPARISON OF PERSONALIZED MEMORY SCHEDULING MODELS

A. EXPERIMENTAL DATA

This study was approved by the Ethics Committee of Inner Mongolia Minzu University. The data for this study is derived from real user behavior logs on the university English learning platform, covering 12 consecutive weeks of vocabulary learning. The experimental population consists of 827 non-English major undergraduates who are included in the study with informed consent. All data is strictly segregated by learner identity. Each learner's data is used for training in the first 8 weeks, for validation in weeks 9–10, and for testing in weeks 11–12, ensuring no cross-individual leakage and strictly adhering to chronological order to prevent future information from influencing current state estimations. Their English proficiency profiles conform to the typical normal distribution of College English Test Band 4 scores. Data collection covers key behavioral indicators throughout the learning process: binary answer results on daily vocabulary tests, millisecond-accurate reaction times, timestamps and interaction frequency of active review behaviors, and the distribution of interval durations across tasks. To ensure data quality, a three-stage cleaning process is implemented: first, records with consistently abnormally short reaction times and excessively long intervals are removed; second, a sliding window smoothing method is used to address short-term behavioral fluctuations; finally, cross-validation is used to eliminate system errors. After cleaning, the valid sample consists of 56,842 learning session records, with an average of 68.7 complete learning cycles per learner. All sensitive information is anonymized to comply with educational data ethics standards. The dataset's high temporal resolution and multimodal nature provide a robust observational foundation for capturing the dynamic evolution of individual memory.

B. VISUALIZATION OF LSTM HIDDEN STATE EVOLUTION AND MEMORY DYNAMICS OF ATTENTION REGULATION

To evaluate the effectiveness of LSTM hidden state representations in modeling learners' memory dynamics, individual behavior sequences are forward-encoded from a large dataset using a trained encoder, extracting 128-dimensional hidden state vectors corresponding to each day. Subsequently, a sliding time window is used to perform dynamic similarity analysis on consecutive hidden state sequences,

calculating the cosine distance between adjacent windows to quantify the evolutionary stability of the memory representation. Furthermore, the hidden state vectors are input into a linear classifier with frozen weights to predict their corresponding true memory state labels, and the classification accuracy is used to evaluate the discriminative power of the representation. The so-called "true labels" are automatically generated by the defined two-dimensional threshold rules and are used for internal consistency evaluation. Furthermore, t-distributed Stochastic Neighbor Embedding (t-SNE) is used to map the high-dimensional hidden state to a two-dimensional space. The clustering structure and temporal continuity of samples at different learning stages are observed to verify whether they form an evolutionary path consistent with cognitive laws.

Figure 4 shows the dynamic activation pattern of the 128-dimensional hidden state generated by the LSTM network encoding the learner's vocabulary memory behavior sequence over 14 time steps. The horizontal axis represents days 1 to 14 of the learning process, while the vertical axis corresponds to the unit dimensions of the LSTM hidden layer (1 to 128). Color represents the activation strength of each unit, with a range of [-1.5, 1.5]. On days 3, 7, and 12 (corresponding to review events), significant red regions of high activation appear within specific dimensional intervals, indicating that these neurons selectively respond to memory reconsolidation. Days 3, 7, and 12 in the heatmap correspond to high-frequency time points in the test set where the policy actually triggers review. The activation pattern is calculated based on the average of the state vectors of all learners who performed review on that day. Activation is particularly strong on day 7, reflecting optimal memory recall. However, during the non-review period from days 9 to 11, overall activation levels decrease, indicating that the system is identifying a weakening memory trend.

Figure 5 systematically illustrates the key internal signals generated by the LSTM encoder as it processes the learner's behavioral sequence. Figure 5(a) shows the distribution of attention weights over the 14-day learning cycle, with the horizontal axis representing time steps (Days 1–14) and the vertical axis representing the normalized attention weights. The data used is the explicit values actually output by the model: the weight on Day 3 is 0.152, corresponding to the first review; it peaks at 0.283 on Day 7, reflecting significant memory reconsolidation triggered by this review; it reaches 0.221 on Day 12, indicating a strong response to reactivation. The weights on the remaining non-review days are mostly below 0.12, indicating that the model effectively identifies and focuses on high-impact intervention nodes. Figure 5(b) shows the evolution of the memory strength index, this index is constructed through a linear combination of the first two principal components of the LSTM hidden states and is only used to visualize the evolution trend of the representation; it is not used as model output or evaluation criteria. The horizontal axis also represents time steps, and the vertical axis represents the memory strength index value. The MSI (memory strength index) jumps from 0.43 to 0.71 on the 7th day, following the peak of attention, demonstrating the model's ability to predict memory enhancement effects. It drops back to 0.52 on the 9th day, reflecting a weakening trend caused by delayed consolidation. However, it rises again to 0.78 after review on the 12th day and continues to increase to 0.85, indicating that the system recognizes the formation of stable memories.

C. MEMORY RETENTION AND LEARNING EFFICIENCY

To comprehensively evaluate the model's adaptability and stability at different learning thresholds, the experimental design uses vocabulary recognition accuracy on day 30 as the core indicator of memory retention, combined with average number of reviews to measure learning efficiency. Learners are divided into three groups according to their initial mastery rate: those below 60% are the low-starting group; those between 60% and 80% are the medium group; those above 80% are the high-starting group. All groups studied the same vocabulary list (320 words), and there were no significant differences in total number of reviews and learning coverage, ensuring equivalent learning intensity among groups. While maintaining consistent training cycles, the proposed PPO+LSTM model is run alongside the currently popular fixed-interval method, traditional PPO, LSTM-QL (Long Short-Term Memory-Q-Learning), SM-2 (SuperMemo-2) algorithm, and the Leitner system. The mean performance of each model is recorded for each group. All baselines use only historical answer records (timestamps and correctness) as input, without accessing the forgetting probability estimated internally by this paper's model, ensuring fairness in comparisons under their respective standard information conditions. The trade-off between improving long-term memory retention and optimizing review burden is analyzed for each method, with a focus on the adaptability of personalized modeling to learners with different cognitive foundations.

Note: Avg is an abbreviation for Average.

Figure 6 compares the memory retention and learning efficiency of six vocabulary review strategies at different initial mastery levels, including vocabulary recognition accuracy and average number of reviews on day 30, reflecting long-term memory retention and cognitive load during learning. The results show that the PPO+LSTM model achieves the highest accuracy (89.2%-93.1%) across all three starting levels, and exhibits a superior gain trend with increasing initial mastery. This demonstrates that its scheduling strategy effectively leverages individuals' prior knowledge and dynamically optimizes the timing of subsequent interventions. In contrast, while the fixed-interval method achieves high-frequency reinforcement in the low-starting group, it fails to form lasting memory representations, resulting in limited improvement, reflecting a mismatch between rigidity and individual needs. Traditional methods such as PPO and LSTM-QL struggle to precisely capture critical points of memory decline due to their coarse state modeling or

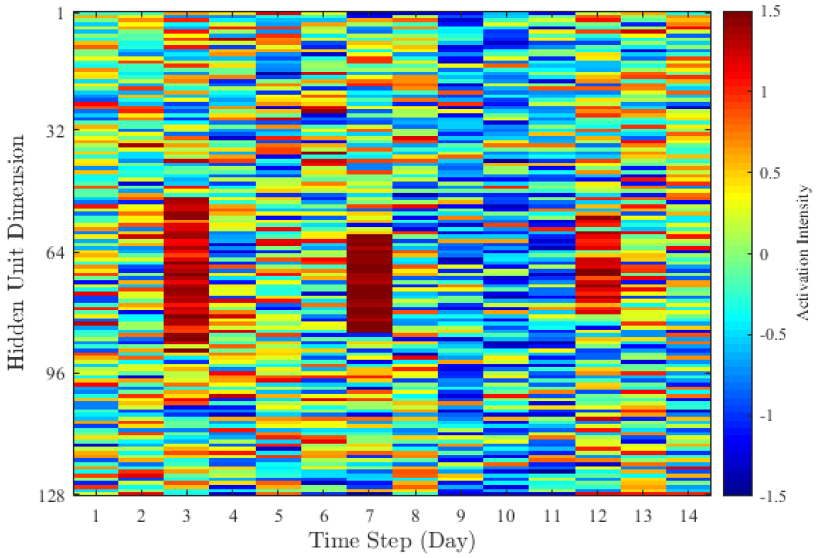

FIGURE 4. LSTM hidden state activation pattern

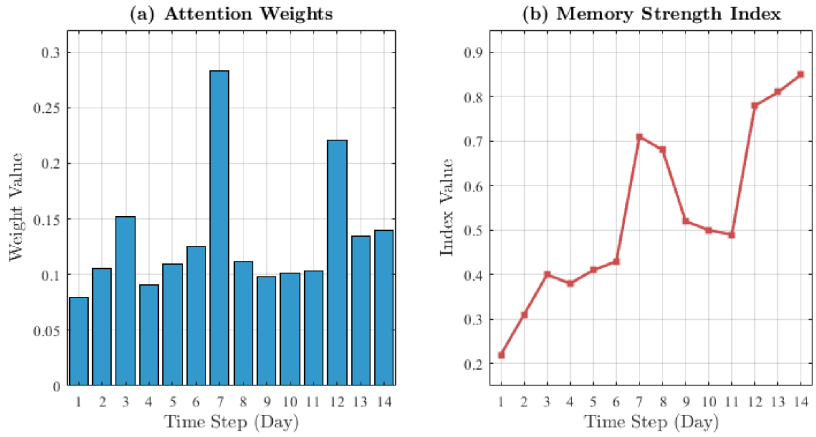

FIGURE 5. Evolution of key internal signals in the LSTM encoder

limited action space. The SM-2 and Leitner systems rely on heuristic rules and lack adaptability to fluctuations in learning pace. While maintaining the highest accuracy, the PPO+LSTM method requires significantly fewer review sessions (4.3-5.8 times) than the fixed-interval method, demonstrating that deep temporal modeling achieves a "small but precise" intervention. The overall trend suggests that the personalized scheduling mechanism integrating deep representations and reinforcement learning exhibits significant advantages in balancing learning efficiency and memory consolidation, particularly in those with medium or high starting points, demonstrating a stronger cognitive synergy.

D. COMPARISON OF LONG-TERM MEMORY STABILITY

To examine the long-term memory stability of different learning strategies, the baseline forgetting slopes of learners are quantified and ranked based on their initial memory decline trajectory without intervention. These groups are then divided into fast, medium, and slow groups, representing low, medium, and high memory stability, respectively. Within the same training cycle, PPO+LSTM is run alongside the fixed-interval method, traditional PPO, LSTM-QL, SM-2, and Leitner systems. Vocabulary retention on day 60 is recorded as a measure of long-term memory retention. A memory fluctuation coefficient, defined as the ratio of the standard deviation of the accuracy over the last ten tests to the mean, is applied to characterize the temporal stability of memory performance. Through dual-indicator cross-analysis, the continuous regulation ability of each model in different forgetting tendency groups is evaluated, focusing on whether the personalized scheduling mechanism can effectively suppress high-volatility or fast-forgetting individuals and promote slow forgetters to maintain low-load consolidation, thereby revealing the robustness and adaptation accuracy of the algorithm in long-term cognitive maintenance.

Table 2 shows the long-term memory stability of different vocabulary learning strategies across three baseline forget-

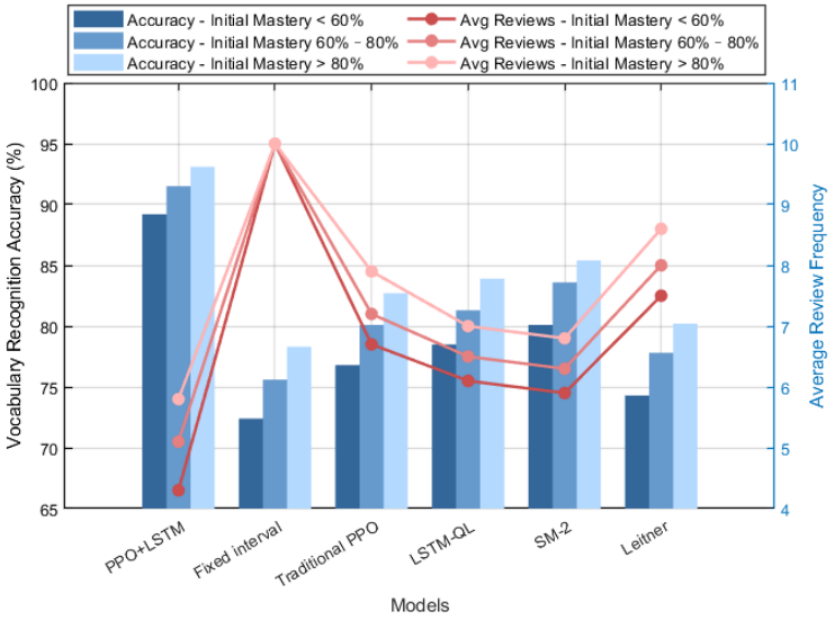

FIGURE 6. Recognition accuracy and average number of reviews

TABLE 2. Comparison of long-term memory stability

Model	Forgetting Group	Day-60 Retention (%)	Memory Fluctuation Coefficient
PPO+LSTM	Fast	78.3 ± 2.1	0.18 ± 0.03
PPO+LSTM	Medium	85.6 ± 1.8	0.14 ± 0.02
PPO+LSTM	Slow	91.2 ± 1.5	0.10 ± 0.02
Fixed interval	Fast	59.4 ± 3.6	0.32 ± 0.05
Fixed interval	Medium	68.7 ± 3.2	0.29 ± 0.04
Fixed interval	Slow	75.1 ± 2.9	0.25 ± 0.03
Traditional PPO	Fast	66.8 ± 3.0	0.27 ± 0.04
Traditional PPO	Medium	74.2 ± 2.6	0.23 ± 0.03
Traditional PPO	Slow	79.5 ± 2.3	0.20 ± 0.03
LSTM-QL	Fast	68.5 ± 2.8	0.25 ± 0.04
LSTM-QL	Medium	76.3 ± 2.4	0.21 ± 0.03
LSTM-QL	Slow	81.7 ± 2.1	0.18 ± 0.03
SM-2	Fast	65.2 ± 3.3	0.28 ± 0.05
SM-2	Medium	72.9 ± 3.0	0.24 ± 0.04
SM-2	Slow	78.4 ± 2.7	0.22 ± 0.03
Leitner	Fast	62.1 ± 3.5	0.30 ± 0.06
Leitner	Medium	70.3 ± 3.1	0.27 ± 0.05
Leitner	Slow	76.8 ± 2.8	0.23 ± 0.04

ting groups. PPO+LSTM achieves the highest retention rate (mean no less than 78.3%) and the lowest memory fluctuation coefficient (mean no more than 0.18) across all groups, showing significant advantages in the fast-forgetting group, demonstrating its ability to effectively suppress memory decline in this high-risk group. In contrast, the fixed-interval method exhibits the lowest retention rate and the highest fluctuation coefficient among fast-forgetters, reflecting a significant mismatch between the rigid review rhythm and individual cognitive dynamics. While traditional methods such as PPO and LSTM-QL exhibit some adaptability, they still exhibit memory trace oscillations during the long-term maintenance phase, indicating insufficient state modeling accuracy. Limited by heuristic rules, the SM-2 and Leitner systems perform reasonably well in the slow-forgetting group, but fail to fully utilize low-load windows for optimal resource allocation.

E. EFFICIENCY EVALUATION OF THE TRADE-OFF BETWEEN LEARNING LOAD AND MEMORY GAINS

To quantify the optimization efficiency of learning strategies under resource constraints, a multi-level load response evaluation framework is constructed using three scenarios: 20 minutes (basic load), 30 minutes (regular load), and 40 minutes (high load) of daily available learning time. Two efficiency metrics are used: the first is the memory improvement magnitude, defined as the net increase in accuracy per review (Δaccuracy/ reviews), reflecting the cognitive transformation efficiency of each intervention; the second is the time cost efficiency, defined as the vocabulary acquired per unit of cumulative learning time (number of words mastered/minute), measuring the intensive utilization

of time resources. Under the premise of controlling the consistency of the initial mastery level, each model runs a complete training cycle under different load conditions, records the dynamic learning trajectory, and calculates the mean and standard deviation of the indicators through multiple independent simulations.

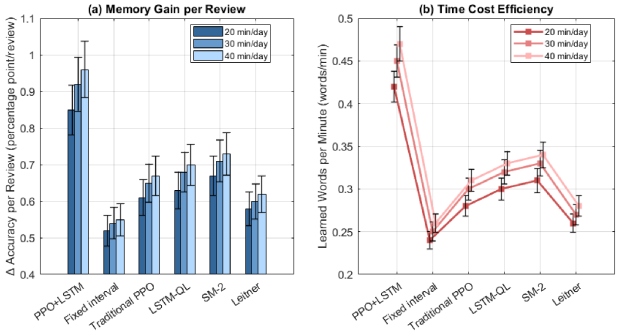

FIGURE 7. Resource utilization efficiency under different time loads

Figure 7 systematically evaluates the resource utilization efficiency of six vocabulary learning strategies under different time loads. The results show that the PPO+LSTM model significantly outperforms the comparison methods in both metrics, with an average memory improvement of 0.85%-0.96%/review and an average time cost efficiency of 0.42-0.47 words/minute. Its performance shows a positive trend with increasing available time, demonstrating that the model can effectively utilize incremental resources for better scheduling. In contrast, the fixed-interval method lacks state awareness and maintains high-frequency, inefficient review under high loads, failing to translate into proportional memory gains. Traditional PPO and LSTM-QL are limited by coarse state representation or insufficient discretization of the action space, resulting in limited optimization boundaries. SM-2 and Leitner systems rely on static rules and are difficult to adapt to individual rhythm changes. PPO+LSTM achieves near-saturation efficiency even under a baseline load, demonstrating that it achieves a "small but precise" intervention mechanism by precisely identifying critical points of memory. As learning resources become more available, the model can dynamically adjust its review distribution, prioritizing weak links and thus continuously improving marginal benefits.

F. DYNAMIC RESPONSE TO MEMORY STATE CHANGES

To examine the dynamic responsiveness of different learning systems to individual cognitive rhythms, a dynamic adaptability evaluation framework is developed, focusing on two key turning points: the mid-term (day 15) and the late-term (day 45) of learning. The mean absolute error (MAE) of memory state predictions is used to measure the model's recognition accuracy of the current cognitive level, reflecting the timeliness and accuracy of its perception capabilities. The labels are generated offline using threshold rules; because the four states have a monotonic order relationship in cognitive stability, encoding them as {0,1,2,3} and then using MAE to measure the order prediction bias is appropriate. Furthermore, the action synchronization rate metric, defined as the ratio of the number of action switches to the number of detected memory state changes, is applied to quantify the efficiency of the decision system's synchronization of state evolution. A synchronization rate close to 1 indicates that the strategy switching and cognitive state changes are basically aligned; a rate slightly higher than 1 indicates that the model introduces a small amount of maintenance review in a stable state, which is a controllable over-adjustment and reflects the effectiveness of the event-driven mechanism overall.

TABLE 3. Mean absolute error and action synchronization rate

Model	Phase	Memory State Prediction MAE	Action Synchronization Ratio
PPO+LSTM	Mid (Day 15)	0.13 ± 0.02	1.02 ± 0.02
PPO+LSTM	Late (Day 45)	0.15 ± 0.03	1.08 ± 0.03
Fixed interval	Mid (Day 15)	0.38 ± 0.05	0.21 ± 0.06
Fixed interval	Late (Day 45)	0.41 ± 0.06	0.19 ± 0.05
Traditional PPO	Mid (Day 15)	0.29 ± 0.04	0.76 ± 0.11
Traditional PPO	Late (Day 45)	0.33 ± 0.05	0.68 ± 0.10
LSTM-QL	Mid (Day 15)	0.26 ± 0.03	0.83 ± 0.12
LSTM-QL	Late (Day 45)	0.30 ± 0.04	0.75 ± 0.11
SM-2	Mid (Day 15)	0.31 ± 0.04	0.42 ± 0.09
SM-2	Late (Day 45)	0.35 ± 0.05	0.38 ± 0.08
Leitner	Mid (Day 15)	0.34 ± 0.05	0.33 ± 0.07
Leitner	Late (Day 45)	0.37 ± 0.06	0.30 ± 0.06

Table 3 shows the dynamic responsiveness of different learning systems to memory state changes in the midterm (day 15) and late-term (day 45) learning stages. The performance of the models' perception-decision coordination under conditions of abrupt learning rhythm changes is evaluated using two metrics: memory state prediction error (MAE) and action synchronization. PPO+LSTM maintains the lowest prediction error at all stages (mean MAE of 0.13-0.15), demonstrating that its LSTM encoder precisely captures non-stationary memory evolution. More importantly, its average action synchronization rate is 1.02–1.08, close to the ideal value of 1, indicating that the frequency of strategy switching basically matches the changes in cognitive state, with only a slight conservative response, and overall maintaining good synchronization. In contrast, the fixed-interval method and the Leitner system, while maintaining a stable review rhythm, exhibit high state prediction errors and delayed action adjustments, reflecting an "open-loop" control approach that is incapable of responding to real-world memory fluctuations. SM-2 and LSTM-QL, while incorporating empirical rules or temporal modeling, exhibit a disconnect between state recognition and action matching. Traditional PPO suffers from a coarse state representation, limiting its responsiveness. The PPO+LSTM model achieves a tight coupling between perception and decision through a closed-loop feedback mechanism, demonstrating excellent dynamic tracking and control sensitivity during cross-stage learning evolution, providing a reliable technical path for addressing uncertainty in real-world learning.

IV. CONCLUSION

This paper constructs a personalized vocabulary memory optimization model based on deep reinforcement learning. The core methodology is also applicable to industrial training challenges, enabling the construction of a personalized skill retention and practice model based on deep reinforcement learning for new operators in the industry. Through LSTM, the learner's multimodal behavior sequence is encoded into a high-dimensional latent state, achieving a fine-grained representation of memory dynamics. On this basis, four semantically interpretable discrete memory state spaces and multi-granularity review opportunity action sets are proposed. Combined with the PPO algorithm, a reward mechanism that takes into account both memory gain and cognitive load is designed to form a closed-loop decision framework. The system achieves online adaptation through periodic parameter fine-tuning, effectively addressing the time-varying nature of individual learning patterns. Experiments demonstrate that the model dynamically balances memory retention and learning efficiency under varying initial mastery levels, forgetting rates, and time constraints, significantly outperforming traditional rule-driven and static modeling paradigms. Compared to traditional methods, this model exhibits stronger cross-stage adaptability and state responsiveness, achieving both improved long-term memory stability and resource utilization. The study verifies the technical path of combining deep temporal modeling with reinforcement learning strategy optimization, which has good robustness and scalability in personalized language learning intervention. This robust and scalable methodology provides new methodological support not only for the regulation of cognitive processes in intelligent education systems but also for the personalized, adaptive training of personnel in complex industrial settings, such as enhancing skill mastery and process retention for operators in the industry.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

This study was approved by the Ethics Committee of Inner Mongolia Minzu University, and was performed in accordance with the principles of the Declaration of Helsinki. All eligible participants signed an informed consent form.

AUTHOR CONTRIBUTIONS

Jun Wang designed, collected and analyzed the data, and drafted the manuscript. Jun Wang conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Jun Wang participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] Y. Fu, L. Zhang, S. Zhao, et al., "Perceptions of non-English major college students on learning English vocabulary with gamified apps," International Journal of Emerging Technologies in Learning (iJET), vol. 16, no. 18, pp. 268-276, 2021, doi: 10.3991/ijet.v16i18.24125.

[2] R. Al-Jarf, "Learning vocabulary in the app store by EFL college students," Online Submission, vol. 5, no. 1, pp. 216-225, 2022, doi: 10.47191/ijsshr/v5-i1-30.

[3] Z. Pi, Q. Yu, Y. Zhang, et al., "Presenting points or rank: The impacts of leaderboard elements on English vocabulary learning through video lectures," Journal of Computer Assisted Learning, vol. 40, no. 1, pp. 104-117, 2024, doi: 10.1111/ jcal.12871.

[4] R. Gao, "The vocabulary teaching mode based on the theory of constructivism," Theory and Practice in Language Studies, vol. 11, no. 4, pp. 442-446, 2021, doi: 10.17507/tpls.1104.14.

[5] R. Al-Jarf, "Online vocabulary tasks for engaging and motivating EFL college students in distance learning during the pandemic and post-pandemic," International Journal of English Language Studies (IJELS), vol. 4, no. 1, pp. 14-24, 2022, doi: 10.32996/ijels.2022.4.1.2.

[6] J. Damanik I and V. Katemba C, "Netflix as a digital EFL learning aid for vocabulary improvement: College students' perspective," ETERNAL (English, Teaching, Learning, and Research Journal), vol. 7, no. 2, pp. 442-455, 2021.

[7] X. Li, "Investigating innovative approaches to intelligent English classroom teaching for medical and health profession specialization," Journal of Commercial Biotechnology, vol. 28, no. 3, pp. 53-60, 2023.

[8] J. Murre J M and G. Chessa A, "Why Ebbinghaus' savings method from 1885 is a very 'pure' measure of memory performance," Psychonomic bulletin & review, vol. 30, no. 1, pp. 303-307, 2023, doi: 10.3758/s13423-022-02172-3.

[9] M. Piotrowski, "Utrwalanie w nauczaniu historii przy pomocy algorytmu SuperMemo i jemu podobnych-propozycja metody redagowania treści historycznych," Kultura-Społeczeństwo-Edukacja, vol. (2), pp. 259-273, 2022, doi: 10.14746/kse.2022.22.15.