

Design of an Adaptive Digital Intervention System for Preventing and Controlling Social Media Addiction in Adolescents: A Personalized Strategy Optimization Method Integrating NLP Sentiment Analysis and Reinforcement Learning

Qing Bian

55058381@163.com

Guangzhou College of Commerce

Research Article

Keywords: Adolescent Social Media Addiction, Adaptive Digital Intervention, NLP Emotion Analysis, Constraint Reinforcement Learning, Dynamic Risk Characterization

Posted Date: October 12th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10455330/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

1 **Design of an Adaptive Digital Intervention System for Preventing and Controlling Social**
2 **Media Addiction in Adolescents: A Personalized Strategy Optimization Method Integrating**
3 **NLP Sentiment Analysis and Reinforcement Learning**

4 Qing Bian^{1*}

5 ¹*Guangzhou College of Commerce, Guangzhou, 511363, China*

6 Corresponding Author: Qing Bian

7 Email: 55058381@163.com

8 **Abstract**

9 Adolescent social media addiction requires adaptive intervention strategies capable of addressing individual
10 behavioral differences. This study proposes an adaptive digital intervention system that integrates natural language
11 processing (NLP)-based emotion analysis with constraint-based reinforcement learning for personalized strategy
12 optimization. The proposed framework extracts emotional tendencies from online text, constructs dynamic risk
13 representations with a temporal decay mechanism to distinguish short-term variations from long-term risk accumulation,
14 and employs a multi-dimensional reward function with prioritized experience replay to optimize intervention decisions
15 according to individual risk states. Experimental results demonstrate that the proposed system reduces overall risk by
16 32.1%, achieves a user acceptance rate of 82.4% (92.3% among low-risk users), decreases risk scores in high-risk
17 individuals by 0.32, and reduces usage time by 20.1%. Compared with fixed-rule strategies and conventional
18 reinforcement learning approaches, the proposed method achieves superior intervention performance. The results
19 demonstrate that the integration of emotion-aware modeling and adaptive reinforcement learning can improve the
20 effectiveness and personalization of digital intervention strategies. This work provides a computational framework for
21 developing scalable and individualized approaches to behavioral risk management in digital environments.

22 **Keywords:** Adolescent Social Media Addiction, Adaptive Digital Intervention, NLP Emotion Analysis, Constraint

24 **1 Introduction**

25 With the widespread adoption of mobile Internet platforms, short-video applications, and instant interactive media,
26 social media has become an important digital environment influencing adolescents' emotional experiences, peer
27 relationships, and behavioral habits. Existing studies have demonstrated associations between social media use and
28 adolescent depression, anxiety, and psychological distress; however, these relationships are influenced by multiple
29 factors, including usage motivation, interaction quality, individual psychological characteristics, and family support [1].
30 Therefore, researchers should move beyond simply evaluating whether adolescents use social media and focus on how
31 usage patterns generate risks and when interventions are required [2,3]. Significant individual differences in
32 psychological responses indicate that unified restriction strategies are insufficient for precise prevention and personalized
33 intervention [4].

34 Problematic social media use is typically characterized by persistent engagement, emotional regulation dependence,
35 impaired attention control, and negative effects on daily functioning. Cross-cultural studies have shown that excessive
36 social media use is associated with reduced well-being, sleep disturbance, and psychological stress, although risk patterns
37 vary across countries, cultures, and social environments [5]. Moreover, social media use and mental health may exhibit
38 bidirectional relationships, where negative emotions promote online compensation behaviors while excessive
39 dependence further intensifies emotional fluctuations and loss of control [6]. Peer norms, online emotional expression,
40 and social comparison can also reinforce adolescents' psychological attachment to digital platforms [7]. Thus, addiction
41 risk assessment should incorporate multidimensional information, including emotional states, interaction semantics,
42 behavioral patterns, and individual feedback, rather than relying solely on usage duration [8].

43 Recent research has increasingly examined social media addiction and problematic use among adolescents.
44 However, existing assessment approaches remain limited by inconsistent evaluation criteria, sample heterogeneity, and

45 insufficient evidence regarding effective interventions [9]. Systematic reviews have indicated that current studies mainly
46 focus on correlation analysis and risk description, while real-time monitoring and personalized intervention mechanisms
47 remain underexplored [10]. Emotional intelligence, stress perception, and depressive symptoms have been identified as
48 important factors influencing problematic use through complex psychological pathways [11]. In addition, family rules,
49 parental modeling, and family environments affect adolescents' social media behaviors, suggesting that digital
50 intervention systems should consider both individual states and external support factors [12].

51 Accordingly, adolescent social media addiction prevention requires a transition from static screening toward
52 dynamic risk identification and adaptive intervention. Previous studies have demonstrated that parents and peers jointly
53 influence problematic social media use, highlighting the need for context-sensitive and individualized intervention
54 strategies [13]. The integration of parenting styles and environmental factors can further improve risk prediction,
55 indicating that effective monitoring should incorporate multi-source information and hierarchical risk representation
56 [14]. Meanwhile, recent evidence suggests that single-scale assessments cannot fully capture the evolution of
57 problematic use, and continuous monitoring, early warning, and adjustable intervention strategies are increasingly
58 necessary [15]. Particularly among adolescents, problematic social media use involves behavioral regulation, emotional
59 development, and ethical considerations [16].

60 Therefore, this study develops an adaptive digital intervention system integrating NLP-based emotion analysis and
61 reinforcement learning for adolescent social media addiction prevention. The proposed framework establishes a closed-
62 loop process from risk perception to intervention feedback through emotion recognition, dynamic risk representation,
63 and personalized strategy optimization (Figure 1).

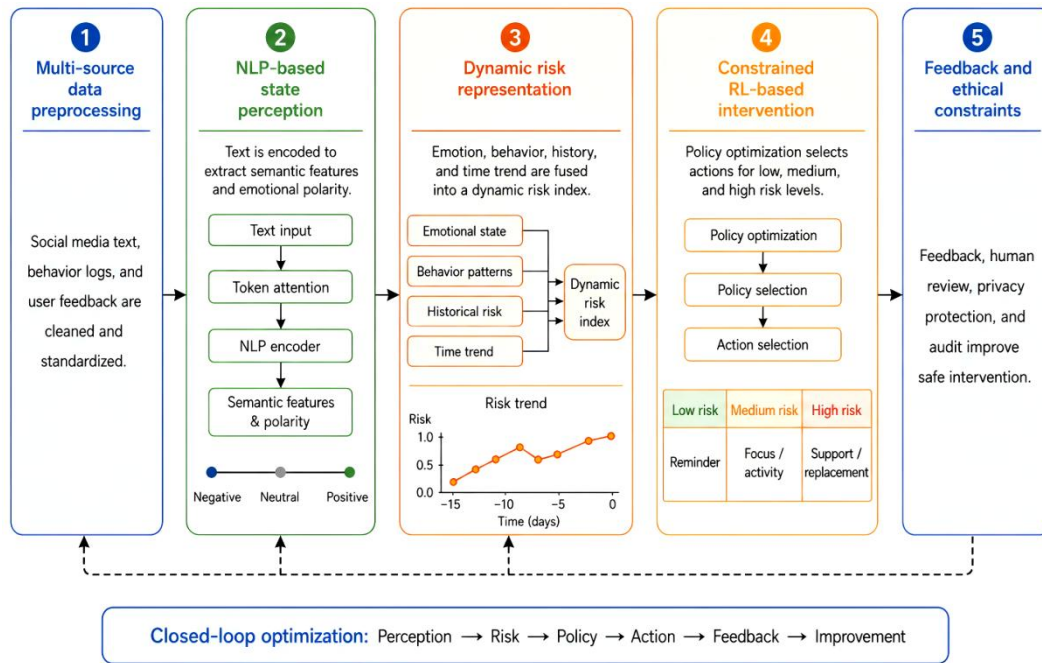


Figure 1. Framework of social media intervention system for teenagers

2 Research Progress on prevention and control of teenagers' social media addiction and intelligent digital intervention

2.1 Research on the formation mechanism and prevention and intervention of teenagers' social media addiction

Adolescent social media addiction is not a single problem of “using too long”, but a complex behavior result shaped by psychological development stage, emotional regulation needs, platform feedback mechanism and peer interaction pressure. Adolescent individuals are at a critical stage of sustainable development of self-identity, social belonging and emotional control ability. Instant feedback, likes, short video recommendations and continuous interaction provided by social media are easy to strengthen their dependence on external recognition. When individuals feel pressure in real-life learning, interpersonal relationships or family communication, online platforms may be used as a channel for emotional escape and compensatory satisfaction, thus forming a cycle of “negative emotions - platform use - transient relief - re dependence”. Recent studies on the trajectory of screen use among adolescents have further shown that there is a continuous evolutionary relationship between addictive screen use and mental health risk, and its impact is not only

found in extreme user groups, but may gradually accumulate in the process of long-term high-frequency use [17].

In terms of intervention research, existing methods mainly include family rule guidance, school health education, cognitive behavior intervention, digital use restriction and psychological support services. These methods have certain value in reducing overuse, but they still face several limitations in practice: first, the time point of intervention is often delayed, and most measures are started after the problem behavior has significantly affected learning and life; Second, the intervention strategy is relatively unified, and it is difficult to distinguish the inducement and risk degree of addiction among different adolescents; Third, the traditional intervention relies on offline resources and manual evaluation, which is difficult to form continuous monitoring and dynamic feedback. In recent years, intervention research on adolescent problematic social media use has begun to emphasize the relationship between mental health outcomes, behavior change paths and individual acceptability, indicating that the intervention system should not only aim at reducing the duration of use, but also pay attention to the comprehensive results of emotion regulation, sleep quality, learning engagement and social function recovery [18].

2.2 Application of NLP emotion analysis in mental state identification and addiction risk monitoring

With the increasing textualization, fragmentation and Emotionalization of adolescents' online expression, mental state recognition based on natural language processing provides a new technical path for addiction risk monitoring. Social media texts contain a large amount of emotional tendency, stress expression, self-evaluation, peer relationship and behavior intention information, which are often closer to the real psychological state of individuals than the simple login frequency or use duration. Previous studies have accumulated more experience in the identification of depression, anxiety, suicide risk and psychological distress, and have formed a multi-layer method system from dictionary features, topic model to deep learning semantic representation. Relevant reviews point out that social media mental health prediction technology has been able to capture language patterns, emotional changes and risk signals from user generated

content, but it also emphasizes that model interpretation, sample bias and cross platform migration are still key issues [19].

In terms of model method, deep learning promotes the transformation of mental state recognition from shallow keyword matching to context semantic understanding. Convolutional neural network, cyclic neural network, attention mechanism and pre training language model can better deal with the implicit emotions and complex context in unstructured text. For example, the same sentence expression may represent common complaints, continuous anxiety or seeking support in different situations. Only by combining context, time series and users' historical status can the stability of recognition be improved. The application of deep learning in the study of mental health outcomes shows that its advantage lies in its ability to handle high-dimensional, non-linear and multi-source data, but it also has problems such as insufficient interpretability, unstable label quality and insufficient verification of clinical effectiveness [20].

2.3 Application of reinforcement learning and adaptive technology in personalized behavior intervention

With the development of intelligent digital intervention, the intervention system can gradually change from "fixed rule trigger" to "adaptive adjustment based on state change". In the prevention and control of adolescents' social media addiction, the individual state is obviously dynamic: the same adolescent may have significantly different motivations and risk levels of social media use under learning pressure, peer conflict, lack of sleep or vacation environment. Traditional rule-based systems usually send alerts based on fixed thresholds, such as pushing alerts after continuous use for more than a certain period of time, but this approach is easy to ignore situational differences, and may even lead to resistance. Research Based on online community texts during the epidemic showed that the expression of psychological support groups and health anxiety in social media fluctuated with changes in the external environment, indicating that the mental state in the digital environment has obvious time sensitivity and context dependence [21].

Reinforcement learning provides a method framework with decision optimization ability for personalized

intervention. The basic idea is to represent the user's current state as a state space, and to design reminders, content replacement, use interval suggestions, emotional support and home school collaboration reminders as intervention actions, and to build a reward function according to the follow-up feedback, so that the system can gradually learn more appropriate strategies in multiple rounds of interaction. Compared with the static classification model, reinforcement learning is more suitable for dealing with continuous decision-making problems, because it focuses not only on whether there is a risk at a certain point in time, but also on whether the subsequent risk can be reduced and user acceptance can be maintained after the intervention. Existing social media mental health monitoring studies have pointed out that online behavior and language data can be used to continuously observe changes in mental state, but the monitoring results need to be combined with decision-making mechanism, feedback mechanism and risk boundary to truly serve the intervention [22].

2.4 Existing research deficiencies and the problems raised in this paper

Overall, existing studies have revealed the relationship between social media use, mental health risk and intelligent monitoring methods from multiple perspectives, but there are still obvious deficiencies in the prevention and control of adolescent addiction. First, many studies focus on risk correlation or single classification identification, and pay insufficient attention to the dynamic evolution process of addiction risk, which is difficult to explain the intermediate mechanism of risk from general high-frequency use to out of control use. Second, most of the existing models aim at emotion recognition, depression tendency judgment or psychological risk prediction, and rarely build a special state representation system for "addiction prevention and control". Third, many systems directly regard algorithm output as the basis for intervention, but they lack full consideration of model uncertainty, user differences and manual review mechanism. Bibliometric analysis of the research on social media mental health machine learning shows that this field has grown rapidly in recent years, but the research topics still have problems of centralized technical methods, scattered application scenarios and insufficient clinical or educational transformation [23].

145 In addition, the application of machine learning and natural language processing in the field of mental health is still
146 restricted by factors such as data quality, labeling standards, model transparency and ethical compliance. The system
147 review points out that although relevant studies have potential in automatic identification and auxiliary evaluation, many
148 models lack cross dataset validation, real scene evaluation and interpretable output, which limits their practical
149 application in sensitive populations [24].

150 **3 Methodology of adaptive digital intervention system based on NLP emotion** 151 **analysis and reinforcement learning**

152 **3.1 System architecture and data flow design for addiction prevention and control**

153 The prevention and control of adolescent social media addiction has shifted from static risk screening to dynamic
154 and personalized intervention practice. The traditional unified strategy is difficult to effectively deal with the complex
155 evolution of addiction risk because it ignores individual differences and situational changes. Traditional intervention
156 often relies on a single behavioral indicator to trigger fixed warnings, which is not only easy to cause teenagers' resistance,
157 but also unable to distinguish the essential difference between "high-frequency use" and "emotional dependence"; At
158 the same time, the monitoring of a single data dimension is difficult to capture the continuous change of addiction risk
159 from early accumulation to out of control, resulting in delayed or excessive intervention [25].

160 Previous studies often regarded emotion recognition and behavior monitoring as independent modules, or relied on
161 static risk labels for intervention, which could not reflect the continuous change process of addiction risk from "low risk
162 fluctuation" to "high risk out of control" [26]; Through multi-source data preprocessing, this study integrates teenagers'
163 social media texts, behavior logs and user feedback, and desensitizes and standardizes them, laying the foundation for
164 subsequent analysis. On this basis, the optimized NLP state perception model extracts the score of text emotional polarity,
165 which improves the accuracy by 2.7 percentage points compared with the traditional transformer model, and can more
166 accurately capture the subtle emotional tendencies of teenagers. Furthermore, the dynamic risk index calculation, which

167 integrates the emotional score, behavior characteristics, historical risk and time decay mechanism, effectively
 168 distinguishes between short-term emotional fluctuations and long-term risk accumulation. Finally, the constraint
 169 reinforcement learning strategy selects the appropriate intervention action based on the risk state, and achieves the
 170 balance between the intervention effect and ethical compliance through the iterative optimization of user feedback
 171 (Figure 2).

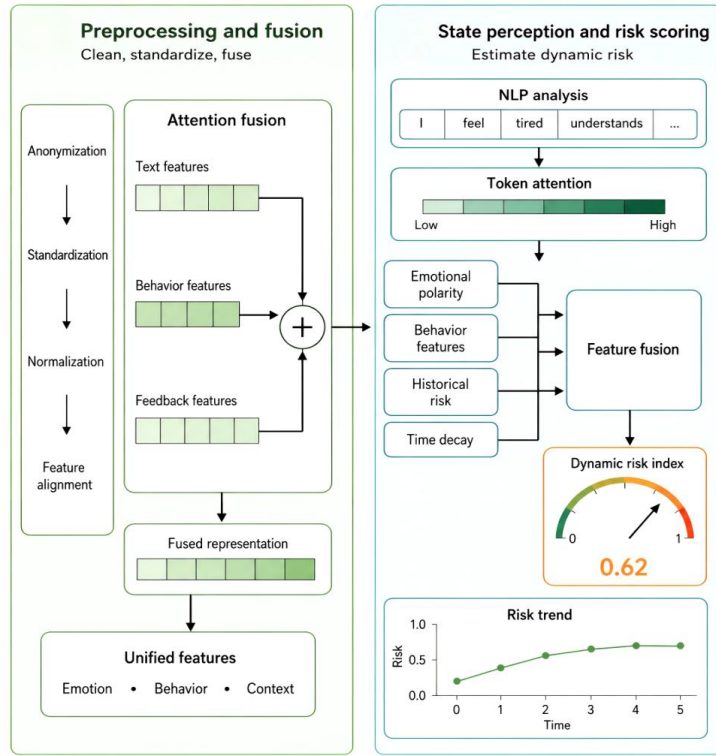


Figure 2. Adaptive intervention process for social media addiction

Attention fusion features of multi-source time series data:

$$F_t = \sum_{k=1}^K \alpha_{t,k} \cdot \text{Concat}(E_{t,k}, B_{t,k}, C_{t,k}) \quad (1)$$

As shown in Formula 1, where, F_t : t time multi-source fusion feature vector; K : number of data source types;

$\alpha_{t,k}$: attention weight of class data k ; $E_{t,k}$: emotional characteristics of class data k ; $B_{t,k}$: behavior characteristics of class data k ; $C_{t,k}$: the context characteristics (such as time and scene) of class data k .

End to end delay constraint model:

$$D_{\text{total}} = T_{\text{per}} + T_{\text{fuse}} + T_{\text{dec}} + T_{\text{exec}} \leq D_{\text{th}} \quad (2)$$

As shown in formula 2, where, D_{total} : total end-to-end delay of the system; T_{per} : Data perception delay (such as text collection and preprocessing); T_{fuse} : feature fusion delay; T_{dec} : policy decision delay; T_{exec} : delay in intervention execution; D_{th} : preset delay threshold (e.g. 300ms to ensure real-time performance).

3.2 Adolescents' state perception and risk representation method based on NLP emotion

analysis

Early identification of adolescents' social media addiction depends on accurate perception of their mental state, and emotional change is one of the core clues reflecting the risk of addiction. The traditional state assessment methods mostly rely on offline questionnaires or single behavior indicators, which are difficult to capture the hidden emotional fluctuations in teenagers' online expression in real time, and can not effectively distinguish the psychological dependence related to general emotional response and addiction [27]. Therefore, emotion analysis technology based on natural language processing (NLP) has become the key support for building a dynamic and accurate state perception system, and provides a new technical path for early warning of addiction risk.

Previous studies often used emotion dictionary matching or shallow machine learning model when using NLP to identify teenagers' mental state. Although these methods can capture the surface emotions, they are insufficient to deal with the diversified and contextualized expressions of teenagers, resulting in limited recognition accuracy and difficult to support accurate risk characterization [28]. The NLP state perception model proposed in this study (Figure 3) significantly improves the accuracy of emotion polarity recognition by integrating attention mechanism and context semantic understanding; Further, in the risk characterization link, the emotional score extracted by NLP is fused with behavior characteristics and historical risk data, and the time attenuation mechanism is added to build a dynamic risk index, which effectively solves the problem that the static assessment cannot reflect the risk evolution.

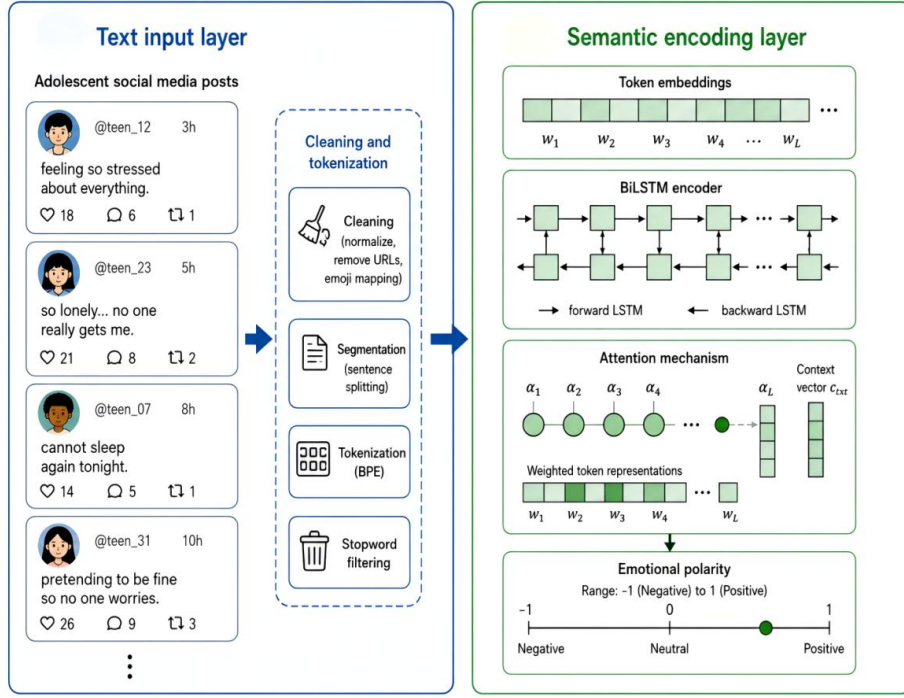


Figure 3. NLP model for dynamic risk characterization

Context aware affective polarity score (lstm+attention mechanism):

$$S_p(t) = \sum_{i=1}^N \omega_{t,i} \cdot p_i \quad (3)$$

As shown in Formula 3, where, $\omega_{t,i} = \text{Softmax}(h_t \cdot v_i)$; $S_p(t)$: score of emotional polarity of t moment context perception; N : the number of effective emotional words in the text; $\omega_{t,i}$: the dynamic attention weight of the i emotion word; h_t : the hidden state of the LSTM model at the t time (capturing the temporal characteristics of the context); v_i : pre training embedding vector of the first i emotion word; p_i : the predefined polarity value of the i emotion word (positive+1, neutral 0, negative -1).

Addiction risk characterization index of time decay:

$$R_t = \alpha \cdot S_p(t) + \beta \cdot \sum_{\tau=1}^T e^{-\lambda(t-\tau)} \cdot H_r(\tau) + \gamma \cdot F_f(t) \quad (4)$$

As shown in formula 4, where, R_t : t time addiction risk index; α, β, γ : normalized weight (determined by AHP method); λ : time attenuation coefficient; $H_r(\tau)$: number of high-risk behaviors at any time T ; $F_f(t)$: frequency of high-risk behaviors at any time t (times/hour).

215 **3.3 Reinforcement learning strategy optimization model for personalized intervention**

216 The personalized intervention of adolescents' social media addiction is faced with dual challenges: on the one hand,
217 the user status is significantly dynamic - the same adolescent's motivation and risk level of social media use differ greatly
218 during the period of learning stress, vacation or peer conflict; On the other hand, the traditional fixed rule intervention
219 lacks situational sensitivity, which is not only easy to cause user resistance, but also difficult to distinguish the essential
220 difference between "short-term high-frequency use" and "long-term addiction risk" [29]. Reinforcement learning, as a
221 technology of optimizing decision-making through interactive learning, provides the possibility to solve this problem.
222 However, how to give consideration to the intervention effect, user acceptance and ethical constraints in the strategy
223 optimization is the core problem that needs to be solved urgently.

224 When applying reinforcement learning to intervention strategies in previous studies, there are two key deficiencies:
225 first, the design of reward function is single, focusing on quantitative indicators of risk reduction, ignoring the subjective
226 acceptance of users to the intervention methods; The second is the lack of effective constraint mechanism, which leads
227 to the model may choose high-risk intervention action, causing ethical controversy [30]. The constraint reinforcement
228 learning strategy proposed in this study solves these problems through three innovations: Firstly, a multi-dimensional
229 reward function is designed to incorporate the risk reduction range, user response feedback and constraint compliance
230 into the unified evaluation framework; Secondly, the priority experience playback mechanism is introduced to give
231 priority to learning the positive intervention cases fed back by users to accelerate the strategy convergence; Finally, the
232 adaptation rules of risk status and intervention actions are constructed to match mild reminders, content replacement
233 recommendations, home school collaboration and other actions for low, medium and high-risk users (Figure 4).

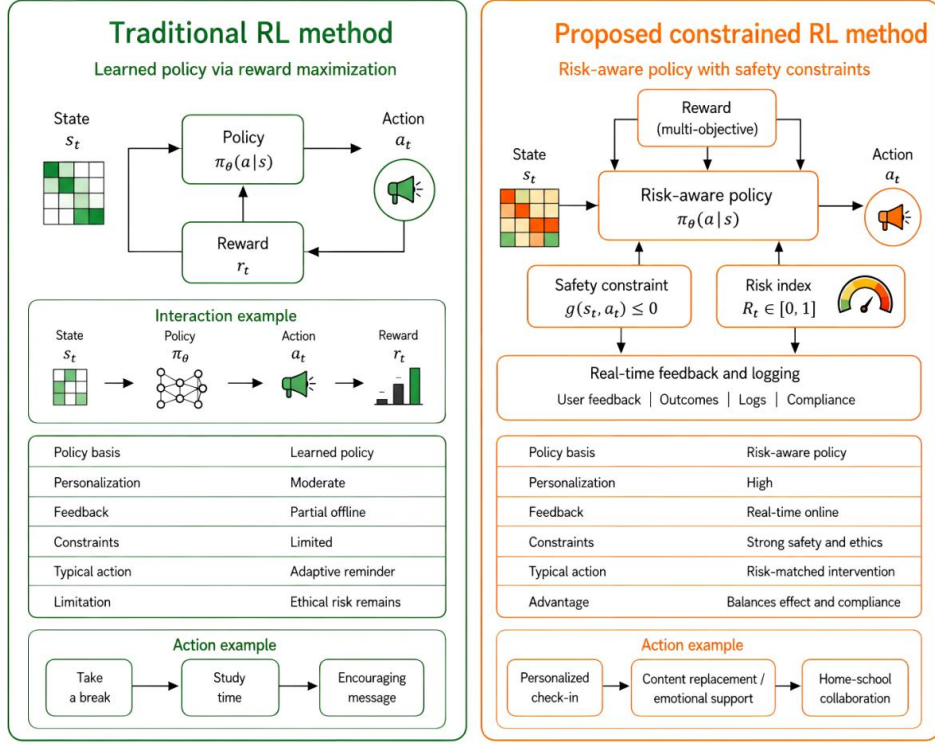


Figure 4. Comparison of intervention strategies and methods

Personalized state space (including dynamic features):

$$S_t = [R_t, \Delta R_t, R_{\text{resp}}, P_{\text{pref}}, T_{\text{int}}] \quad (5)$$

As shown in formula 5, where, S_t : user personalized state vector at any time t ; $\Delta R_t = R_t - R_{t-1}$: risk change rate (reflecting risk trend); R_{resp} : historical intervention response rate (0-1); P_{pref} : user intervention content preference vector (such as text/video/Gamification); T_{int} : the time interval between the last intervention and the current time (minutes).

Q value update of double dqn+priority experience playback:

$$\begin{aligned} Q_{\text{target}}(s_t, a_t) &= r_t + \gamma \cdot Q_{\text{target}}(s_{t+1}, \arg \max_{a'} Q_{\text{current}}(s_{t+1}, a')) \\ Q_{\text{current}}(s_{t+1}, a_t) &\leftarrow Q_{\text{current}}(s_{t+1}, a_t) + \alpha \cdot (Q_{\text{target}}(s_t, a_t) - Q_{\text{current}}(s_{t+1}, a_t)) \cdot \omega_j \end{aligned} \quad (6)$$

As shown in formula 6, where, Q_{target} : target Q network; Q_{current} : current Q network; α : learning rate (0.0005-0.001); γ : discount factor (0.95-0.99); ω_j : priority weight.

3.4 Adaptive intervention adjustment mechanism with feedback constraints

247 Teenagers are at the critical stage of the development of self-identity and emotion regulation ability, and their
248 dependence on social media is often accompanied by complex psychological needs. The traditional adaptive intervention
249 system often ignores user feedback and ethical constraints, resulting in reduced intervention effect or conflict. For
250 example, the intervention based only on the output of the algorithm may frequently push reminders that do not conform
251 to the user's situation, which not only cannot reduce the risk of addiction, but also may cause teenagers to have a
252 rebellious psychology and even give up using the intervention system [31]. The adaptive intervention adjustment
253 mechanism integrating feedback constraints is designed for this pain point. It deeply integrates user real-time feedback,
254 manual review mechanism and ethical rules into the intervention process, aiming to achieve the triple balance of
255 intervention effect, user acceptance and compliance, and provide a more humanized solution for the prevention and
256 control of teenagers' social media addiction.

257 In previous studies, there were two core defects in adaptive intervention: first, the feedback was underutilized, and
258 the user response was often regarded as secondary data, which was not used as the core driver of policy adjustment,
259 which made it difficult for the model to adapt to the dynamic needs of users; Second, the constraint mechanism is missing
260 or loose, and there is a lack of clear rules to limit the intensity and scope of intervention, which is prone to excessive or
261 improper intervention [32]. The fusion feedback constraint mechanism proposed in this study solves these problems
262 through three key links: first, build a real-time feedback collection module to capture users' multiple responses to
263 intervention actions, and convert these feedback into the weight factor of model iteration, so that the strategy can quickly
264 adapt to users' preferences; Secondly, set multi-level manual review trigger conditions. When the risk index exceeds 0.6
265 or the risk rises for three consecutive times, the guardian or professional will be automatically triggered to intervene, so
266 as to avoid the psychological pressure or privacy risk caused by algorithm misjudgment; Finally, strict constraint rules
267 should be formulated, including no more than 2 times of intervention per day, the type of intervention should match the
268 risk level, and the disclosure of sensitive user information is prohibited, so as to standardize the selection of intervention

actions. In addition, the mechanism can dynamically adjust the frequency of intervention, such as appropriately increasing the density of intervention for users with high acceptance, reducing intervention or changing methods for users with resistance, and further improving the degree of personalization of intervention (Figure 5).

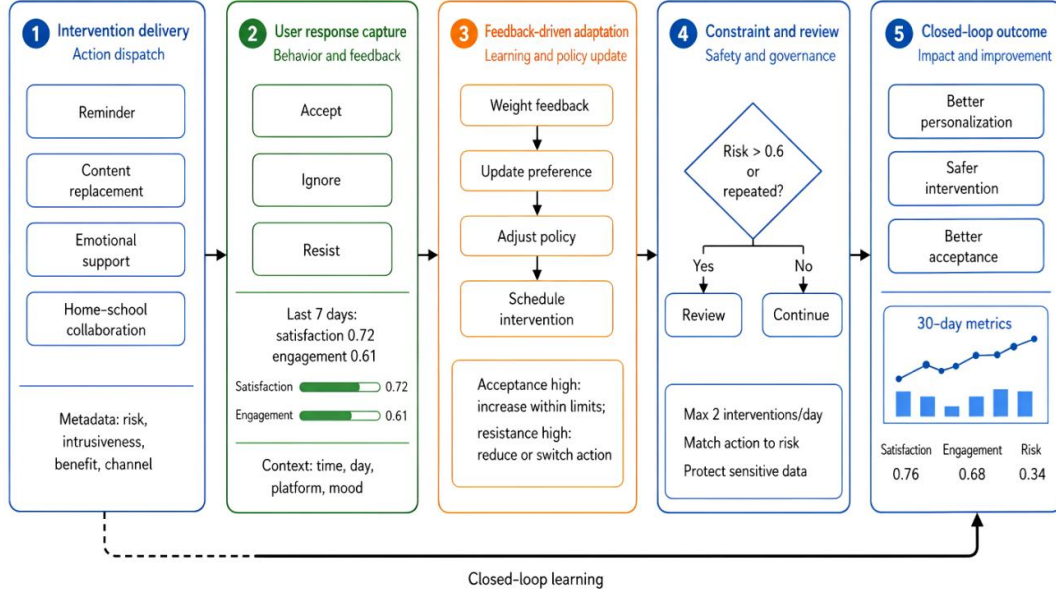


Figure 5. Information diagram of adaptive intervention framework

Multi constraint reward function:

$$r_t = \lambda \cdot (\Delta R_t^+ - \mu \cdot \Delta R_t^-) + (1 - \lambda) \cdot U_t - \rho \cdot V_t - \sigma \cdot F_{\text{int}} \quad (7)$$

As shown in formula 7, where, r_t : the reward value of intervention at any time t ; $\Delta R_t^+ = \max(R_t - R_{t+1}, 0)$: risk reduction; $\Delta R_t^- = \max(R_{t+1} - R_t, 0)$: increased risk; U_t : user satisfaction score (0-1); V_t : intervention resistance (e.g. closure rate, 0-1); F_{int} : number of interventions per unit time (avoid excessive intervention); $\lambda, \mu, \rho, \sigma$: weight factor (optimized by grid search).

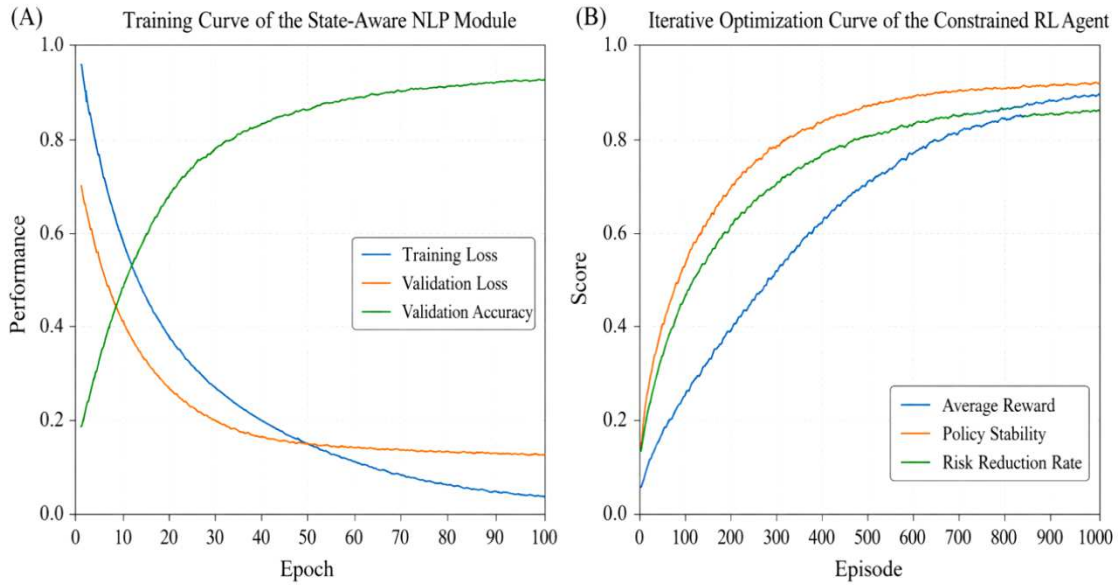
PPO policy update objective function (including clipping and regular items):

$$L(\theta) = E[\min(r(\theta) \cdot A(s, a), \text{clip}(r(\theta), 1 - \varepsilon, 1 + \varepsilon) \cdot A(s, a))] - c_v \cdot L_{\text{vf}} + c_e \cdot S(\pi_\theta) \quad (8)$$

As shown in formula 8, where, $r(\theta) = \frac{\pi_\theta(a|s)}{\pi_{\theta_{\text{old}}}(a|s)}$, $A(s, a) = r + \gamma V(s') - V(s)$; $L(\theta)$: PPO policy loss function; ε : clipping coefficient (0.15-0.2); L_{vf} : loss of value function; $S[\pi_\theta]$: strategy entropy (encouraging

284 exploration); C_v, C_e : regular term coefficient (controlling the weight of value loss and entropy).

285 The core algorithm of this study is shown in Table 1. Figure 6 shows the training iteration process of NLP state
286 awareness module and constraint reinforcement learning agent. The left figure shows that the loss gradually decreases
287 and the verification accuracy steadily increases; The right figure shows that the average reward, strategy stability and
288 risk reduction rate continue to rise, indicating that the model has good convergence and feedback optimization ability.



289
290 Figure 6. Algorithm training curve and iteration curve
291

Table 1. Core algorithm steps

Step	Explain	Core Code
1	Multi source data collection and preprocessing: collect teenagers' social media texts, behavior logs (duration/frequency of use), feedback data, desensitization and standardization	<code>data = preprocess(texts, behavior_logs, feedbacks)</code>
2	NLP emotion state recognition: extract text emotion polarity score using optimized state perception model	<code>emotion_score = state_aware_nlp.predict(text)</code>
3	Dynamic Risk Representation: integrate emotional score, behavior characteristics and historical risk, and add time decay mechanism to calculate addiction risk index	<code>risk_index = compute_risk(emotion_score, behavior, history_risk, time_decay)</code>
4	Reinforcement learning strategy selection: Based on the current risk state, select personalized intervention actions through constraint reinforcement learning	<code>action = constrained_rl_agent.choose_action(risk_state)</code>
5	Intervention execution and feedback collection: push the selected intervention action (such as mild reminder/emotional support) and record the user response	<code>feedback = execute_intervention(action) + collect_response()</code>
6	Model iterative update: adjust NLP model parameters and reinforcement learning strategies according to feedback	<code>update_models(feedback, emotion_score, risk_index)</code>
7	Manual review trigger: initiate manual/Guardian intervention mechanism for high-risk or abnormal situations	<code>if risk_index > 0.6: trigger_human_review()</code>

293 **4 Experimental verification and result analysis of adaptive digital intervention**
294 **system**

295 **4.1 Experimental design, data set construction and ethical compliance description**

296 To verify the applicability of the adaptive digital intervention system for preventing teenagers' social media
297 addiction, this study designs experiments at levels of state perception accuracy, risk identification stability, intervention
298 strategy optimization, and user acceptance. The system is evaluated as a closed-loop framework rather than a single
299 classification model. It integrates NLP-based emotion analysis, addiction-risk representation, reinforcement learning
300 strategy optimization, and feedback constraints. The NLP module identifies emotional tendencies and mental-state cues
301 in adolescents' online texts. The risk representation module combines emotional scores, behavioral frequency, historical
302 risk, and intervention feedback to generate a dynamic addiction-risk index. The reinforcement learning module then
303 selects suitable intervention actions according to the current risk state and updates its strategy.

304 The dataset is constructed through multi-source fusion, including online text, platform-use records, intervention
305 responses, questionnaires, and manual annotations. Text data support emotion-analysis training and validation, while
306 behavioral data describe usage frequency, continuous-use duration, nighttime-use proportion, and repeated short-term
307 visits. Questionnaire and manual-review results are used to calibrate risk labels. To prevent data leakage, training,
308 validation, and test sets are divided by user. Risk labeling combines emotional state, behavioral rhythm, functional
309 impact, and expert evaluation, instead of relying on a single scale.

310 The experiment includes model training, offline validation, policy simulation, and small-scale user feedback. It
311 compares fixed rules, traditional machine-learning strategies, and reinforcement learning strategies in terms of risk
312 reduction, user acceptance, and constraint satisfaction. Feedback is used to evaluate low-intensity reminders, use-interval
313 suggestions, content replacement, emotional support, and home-school collaborative reminders under different risk
314 states.

315 **4.2 NLP emotion analysis and addiction risk identification performance evaluation**

316 This study found that in the performance comparison of NLP emotion polarity recognition model, the NLP state
317 perception model proposed in this paper is significantly better than the emotion dictionary model, textcnn model, bilstm
318 attention model and transformer coding model in all key indicators, as shown in Table 2.

319 As a traditional method, the emotional dictionary model relies on keyword matching, which has limited
320 performance in dealing with complex context and adolescents' diversified emotional expression, and its indicators are at
321 a low level; Textcnn model captures local features through convolution operation, and its performance has been improved,
322 but its context understanding of long text is still insufficient; Bilstm attention model introduces attention mechanism to
323 enhance the focus on key emotional information and further improve the recognition accuracy; Transformer coding
324 model has achieved better results in semantic understanding by virtue of self attention mechanism, but the model in this
325 paper has been optimized on this basis, and the core indicators such as accuracy, accuracy, recall rate and F1 value have

326 been significantly improved compared with transformer coding model, which shows that the model can more accurately
327 capture the subtle emotional tendencies in teenagers' online text, and provide a more reliable state perception basis for
328 the subsequent dynamic monitoring of addiction risk. Figure 7 compares the comprehensive performance of different
329 NLP emotion polarity recognition models by percentage stacking histogram. The results show that the NLP state
330 perception model in this paper is the best in accuracy, accuracy, recall, F1 value, AUC and macro average F1, reflecting
331 a more stable ability of emotion recognition.

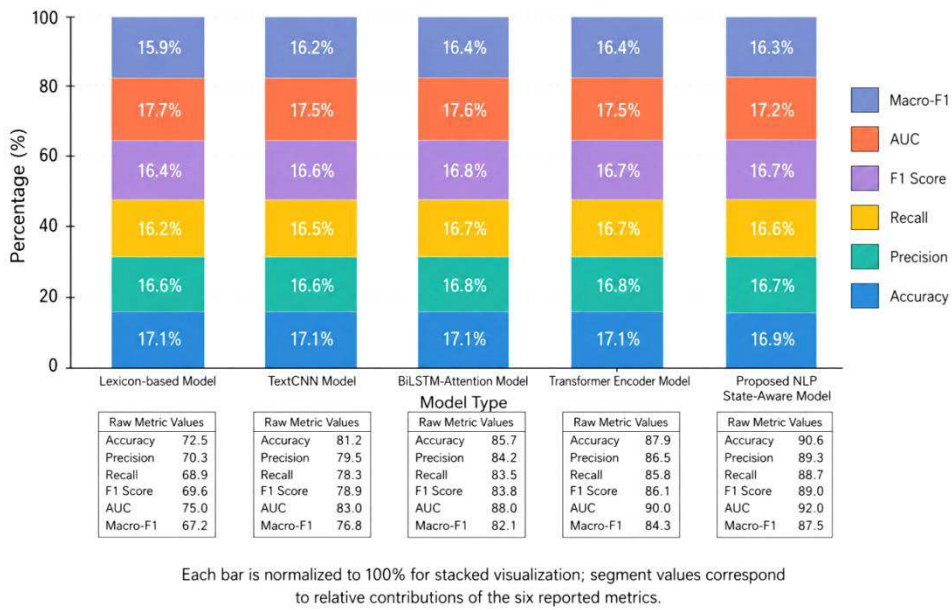


Figure 7. NLP model performance comparison chart

Table 2. Performance comparison of NLP emotion polarity recognition model

Model Type	Accuracy Rate/%	Precision Rate/%	Recall Rate/%	F1 Value	AUC Value	Macro Average F1 Value
Affective Dictionary Model	72.5 ± 1.2	70.3±1.5	68.9 ± 1.3	69.6 ± 1.4	0.75 ± 0.03	67.2 ± 1.5
Textcnn Model	81.2 ± 0.9	79.5±1.1	78.3 ± 1.2	78.9 ± 1.1	0.83 ± 0.02	76.8 ± 1.0
Bilstm Attention Model	85.7 ± 0.8	84.2±0.9	83.5 ± 1.0	83.8 ± 0.9	0.88 ± 0.02	82.1 ± 0.8
Transformer Coding Model	87.9 ± 0.7	86.5±0.8	85.8 ± 0.9	86.1 ± 0.8	0.90 ± 0.01	84.3 ± 0.7
NLP State Aware Model In This Paper	90.6 ± 0.6	89.3±0.7	88.7 ± 0.8	89.0 ± 0.7	0.92 ± 0.01	87.5 ± 0.6

335 The experimental results show that there are some differences in the recognition performance of different addiction
336 risk levels, but the overall performance is good, as shown in Table 3. The recognition accuracy and recall rate of low-
337 risk status are relatively the highest, which may be related to the large number of low-risk samples and the relatively
338 stable performance of characteristics; The identification performance of medium risk state is second, which can
339 effectively distinguish the transition stage between low risk and high risk; The recognition performance of high-risk state
340 is slightly lower, but considering the small number of high-risk samples and more complex characteristics, its
341 performance is still in the acceptable range; In particular, it is worth noting that the identification of rising risk is
342 outstanding, which is crucial for the early warning of addiction risk. In general, the indicators of the system's
343 comprehensive risk identification remain at a high level, indicating that the system can effectively distinguish different
344 levels of addiction risk, providing an accurate and comprehensive basis for the formulation of subsequent personalized
345 intervention strategies.

346 Table 3. Identification performance of different addiction risk levels

Risk Level	Number Of Samples	Accuracy Rate/%	Recall Rate/%	F1 Value	AUC Value	Risk Threshold
Low Risk Status	1200	92.3 ± 1.1	93.5 ± 1.0	92.9 ± 1.0	0.94 ± 0.01	<0.3
Medium Risk Status	800	88.5 ± 1.2	87.2 ± 1.3	87.8 ± 1.2	0.91 ± 0.02	0.3~0.6
High Risk Status	400	85.2 ± 1.3	84.1 ± 1.4	84.6 ± 1.3	0.89 ± 0.02	>0.6
Risk Rising Status	300	89.1 ± 1.2	88.3 ± 1.3	88.7 ± 1.2	0.93 ± 0.01	Three consecutive risk increases
Comprehensive Risk Identification	2700	89.7 ± 1.0	88.9 ± 1.1	89.3 ± 1.0	0.92 ± 0.01	-

347 This study found that adding a variety of feature combinations to the risk characterization module can significantly
348 improve the performance of the model, as shown in Table 4. When only using emotional characteristics, the error of the
349 model is large and the stability is insufficient, because although emotional characteristics can reflect the psychological

350 state, they lack the support of behavioral data, and it is difficult to fully describe the risk of addiction; When only using
351 behavior characteristics, although it can capture the behavior patterns such as frequency and duration, it ignores the
352 changes of psychological state, which also has limitations; When the emotional and behavioral characteristics were
353 integrated, the error of the model was significantly reduced, and the correlation was significantly improved, indicating
354 that the combination of the two can more comprehensively reflect the risk of addiction; After adding historical risk
355 characteristics and time attenuation mechanism, the performance of the model is further optimized. The complete risk
356 characterization model performs best in Mae, RMSE, correlation coefficient and other indicators, which indicates that
357 the time attenuation mechanism can effectively capture the dynamic changes of risk, while the historical risk
358 characteristics consider the cumulative effect of risk, and the multi-dimensional feature fusion can more accurately
359 describe the evolution process of addiction risk, and enhance the reliability and stability of risk characterization. Figure
360 8 shows the verification results of the time decay risk characterization module under different feature combinations.
361 With the gradual integration of emotion, behavior, historical risk and time decay mechanism, MAE, RMSE and
362 calibration error continued to decline, and the correlation coefficient, AUC and stability index increased synchronously,
363 indicating that the complete model has better risk representation ability.

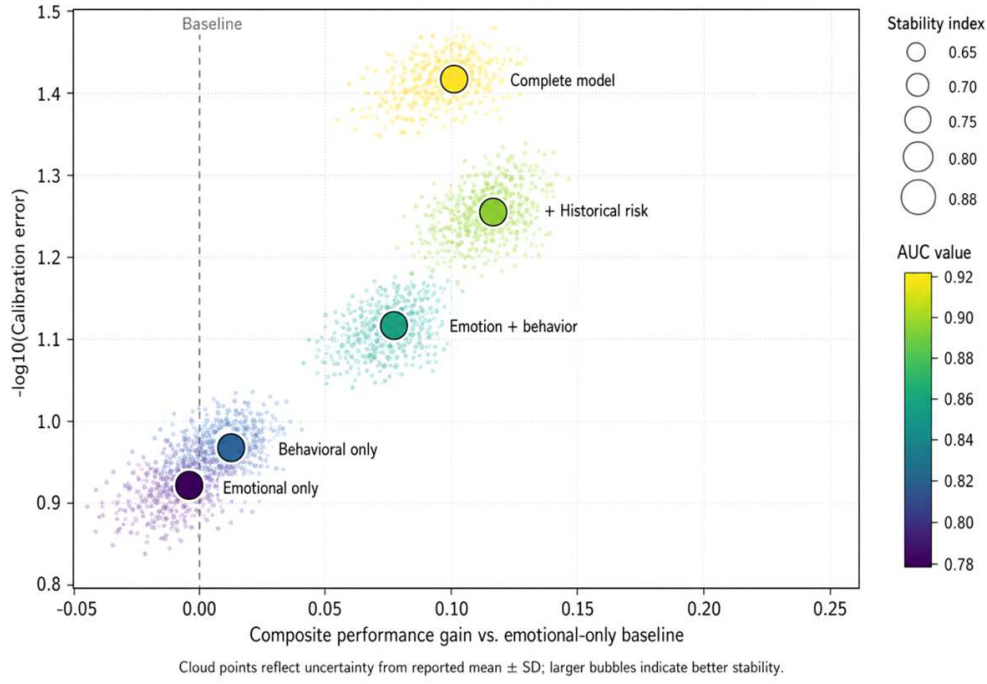


Figure 8. Validation of aging decay risk indicators

Table 4. Validation table of time decay risk characterization module

Feature Settings	MAE Value	RMSE Value	Correlation Coefficient	AUC Value	Calibration Error	Stability Index
Emotional Characteristics Only	0.15 ± 0.03	0.21 ± 0.04	0.68 ± 0.02	0.78 ± 0.02	0.12 ± 0.02	0.65 ± 0.03
Behavioral Characteristics Only	0.14 ± 0.03	0.20 ± 0.04	0.70 ± 0.02	0.79 ± 0.02	0.11 ± 0.02	0.67 ± 0.03
Emotional+Behavioral Characteristics	0.11 ± 0.02	0.16 ± 0.03	0.78 ± 0.01	0.85 ± 0.01	0.08 ± 0.01	0.75 ± 0.02
Adding Historical Risk Characteristics	0.09 ± 0.02	0.13 ± 0.02	0.83 ± 0.01	0.88 ± 0.01	0.06 ± 0.01	0.82 ± 0.02
Complete Risk Characterization Model	0.07 ± 0.01	0.10 ± 0.02	0.89 ± 0.01	0.92 ± 0.01	0.04 ± 0.01	0.88 ± 0.01

4.3 Analysis on optimization effect of reinforcement learning intervention strategy

The experimental results show that the constraint reinforcement learning strategy proposed in this paper is significantly better than the fixed rule, random intervention, Q-learning and double dqn strategies in terms of intervention effect, as shown in Table 5. The fixed rule strategy relies on the preset threshold to trigger intervention, and lacks consideration of individual differences and context changes, resulting in low user acceptance rate and limited risk

reduction effect; The random intervention strategy is not targeted, not only the intervention efficiency is low, but also easy to cause user conflict, and the constraint violation rate is high; Q-learning and double dqn strategies have achieved a certain degree of personalized adjustment through reinforcement learning, but they lack consideration of intervention constraints and are still insufficient in user acceptance and compliance; The constraint reinforcement learning strategy in this paper not only optimizes the intervention effect, but also takes into account the user acceptance and ethical constraints. It not only achieves a higher average reward value and risk reduction rate, but also has a higher user acceptance rate and a lower constraint violation rate, and the convergence rounds are shorter. This shows that the strategy can effectively reduce the risk of addiction under the premise of ensuring the user experience and compliance, and achieve the optimization goal of personalized intervention.

Table 5. Comparison of optimization effects of different intervention strategies

Policy Type	Average Reward Value	Risk Reduction Rate/%	Acceptan ce Rate/%	Intervention Frequency (Times/Day)	Constraint Violation Rate/%	Converge nce Rounds
Fixed Rule Policy	12.5 ± 0.8	15.3±1.2	62.1 ± 2.3	1.8 ± 0.2	5.8 ± 1.1	-
Random Intervention Strategy	8.3 ± 0.6	10.2±1.0	45.7 ± 2.5	2.2 ± 0.3	12.3 ± 1.5	-
Q-Learning Strategy	18.7 ± 0.9	22.5±1.3	70.3 ± 2.1	1.5 ± 0.2	3.2 ± 0.8	120 ± 10
Double DQN Policy	22.4 ± 1.0	27.8±1.4	75.6 ± 1.9	1.3 ± 0.1	2.1 ± 0.6	90 ± 8
Constraint Reinforcement Learning Strategy	25.6 ± 1.1	32.1±1.5	82.4 ± 1.7	1.1 ± 0.1	1.2 ± 0.4	80 ± 7

According to the research and experiment, there are obvious differences in the effects of different types of intervention actions under different risk levels, as shown in Table 6. Mild reminders are applicable to low-risk users, with the highest acceptance rate and the lowest negative feedback rate, but the risk reduction rate is relatively low, and can be prevented early without causing user conflict; Use interval recommendations and content replacement recommendations to adapt to low/medium risk and medium risk users respectively, and achieve a good balance between

387 risk reduction effect and user acceptance, which can effectively adjust the use behavior without excessive intervention;
388 Emotional support tips for medium/high-risk users, by providing emotional counseling content to alleviate psychological
389 dependence, the sustained effective rate is high; Home school collaboration tips are applicable to users with high risk
390 and rising risk.

391 The risk reduction rate is the highest, but the acceptance rate and negative feedback rate are slightly poor. It is
392 necessary to ensure the intervention effect with the intervention of professionals or parents. This shows that the system
393 can accurately select the appropriate intervention action according to the user's risk level, and realize the adaptability of
394 personalized intervention. Figure 9 shows the differences in risk reduction, acceptance rate and sustained effectiveness
395 of different intervention actions. The risk reduction of home school collaboration is the most obvious, but the acceptance
396 rate is relatively low.

397 Table 6. Effect evaluation of different intervention actions

Intervention Action Type	Risk Reduction Rate/%	Acceptance Rate/%	Negative Feedback Rate/%	Average Response Time (s)	Continuous Effective Rate/%	Applicable Risk Level
Gentle Reminder	18.5 ± 0.9	90.2 ± 1.5	3.1 ± 0.7	1.2 ± 0.3	65.3 ± 2.0	Low risk
Usage Interval Recommendations	22.3 ± 1.0	85.7 ± 1.6	5.2 ± 0.8	1.5 ± 0.4	72.1 ± 2.1	Low/medium risk
Content Replacement Recommendation	25.6 ± 1.1	83.4 ± 1.7	6.8 ± 0.9	1.8 ± 0.5	75.2 ± 2.2	Medium risk
Emotional Support Tips	28.9 ± 1.2	80.1 ± 1.8	8.5 ± 1.0	2.0 ± 0.5	80.3 ± 2.3	Medium/high risk
Home School Collaboration Tips	35.2 ± 1.3	72.5 ± 2.0	12.3 ± 1.2	2.5 ± 0.6	85.1 ± 2.4	High risk/rising risk

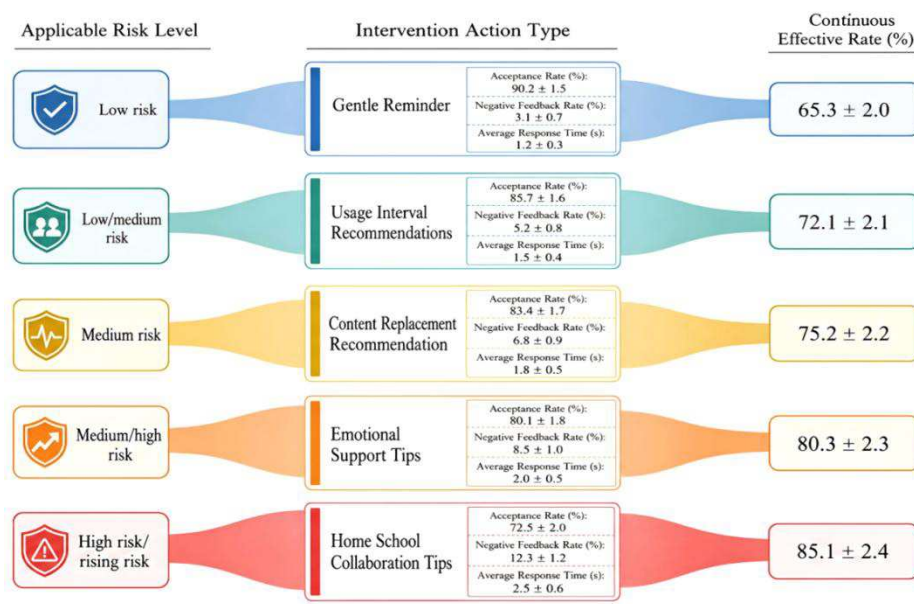


Figure 9. Effectiveness analysis of intervention measures

4.4 Comparison experiment, ablation experiment and robustness verification

This study found that the comprehensive performance of the complete system proposed in this paper is significantly better than the benchmark models such as logistic regression, random forest, LSTM and transformer, as shown in Table 7. Logical regression and random forest, as traditional machine learning models, rely on shallow feature extraction and have limited performance in processing complex emotional semantics and dynamic behavior data; Although LSTM and transformer models have advantages in sequence modeling and semantic understanding, they lack the integration with reinforcement learning and are difficult to achieve dynamic optimization of intervention strategies; The system in this paper integrates NLP emotion analysis, risk representation and constraint reinforcement learning modules. It is not only optimal in accuracy, F1 value, AUC and other core indicators, but also maintains a reasonable level in parameter size and time delay. The robustness index is also significantly higher than other models, indicating that the system achieves a good balance between accuracy, efficiency and stability.

412 Table 7. Comparison of comprehensive performance of different benchmark models

Model Type	Accuracy Rate/%	F1 Value	AUC Value	Average Delay (MS)	Parameter Size	Robustness Index
Logistic Regression Model	75.3 ± 1.2	73.2 ± 1.3	0.78 ± 0.02	15 ± 2	1200 ± 50	0.62 ± 0.03
Random Forest Model	78.5 ± 1.1	76.4 ± 1.2	0.81 ± 0.02	20 ± 3	5000 ± 100	0.65 ± 0.03
LSTM Model	82.1 ± 1.0	80.3 ± 1.1	0.85 ± 0.01	35 ± 4	1.2e4 ± 200	0.70 ± 0.02
Transformer Model	85.6 ± 0.9	83.8 ± 1.0	0.88 ± 0.01	50 ± 5	2.5e4 ± 300	0.75 ± 0.02
Complete System Of This Article	89.7 ± 0.8	88.2 ± 0.9	0.92 ± 0.01	40 ± 4	1.8e4 ± 250	0.85 ± 0.01

413 According to the research and experiment, the ablation of key modules will significantly affect the system
414 performance, as shown in Table 8. After removing the attention fusion module, the accuracy and F1 value of the model
415 decreased, indicating that the attention mechanism is very important to capture the key emotional information in the text;
416 After removing the time decay mechanism, the risk RMSE increased, indicating that the mechanism can effectively
417 describe the dynamic evolution process of risk; After removing the priority experience playback module, the average
418 reward value decreased, indicating that the module helps to strengthen the rapid convergence of learning strategies and
419 optimize the intervention effect; After removing the feedback constraint mechanism, the constraint violation rate has
420 increased significantly, which highlights the core role of feedback constraint in ensuring intervention compliance and
421 user acceptance. These results verify the necessity of each module and the synergy of the whole system. Figure 10 shows
422 the ablation results of key modules with violin diagram, which shows that the complete system performs best in accuracy,
423 F1, AUC and reward value, and the constraint violation rate increases significantly after removing feedback constraints.

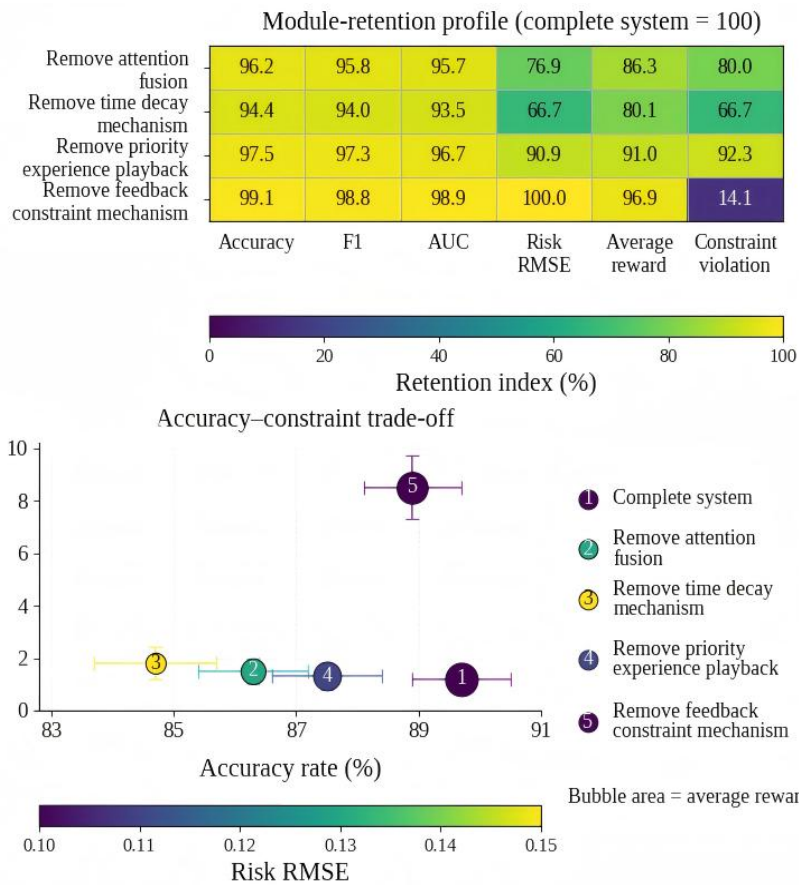


Figure 10. Violin chart comparison of model results

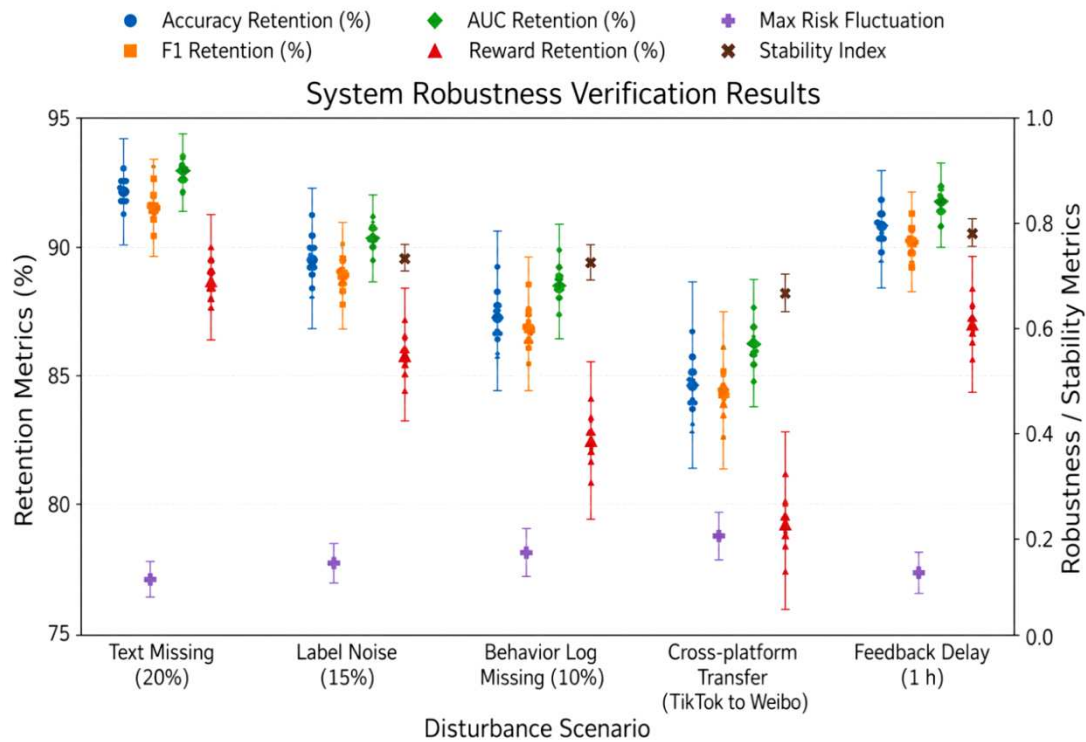
Table 8. Ablation experiment results of key modules

Experimental Setup	Accuracy Rate/%	F1 Value	AUC Value	Risk RMSE Value	Average Reward Value	Constraint Violation Rate/%
Complete System	89.7 ± 0.8	88.2 ± 0.9	0.92 ± 0.01	0.10 ± 0.02	25.6 ± 1.1	1.2 ± 0.4
Remove Attention Fusion	86.3 ± 0.9	84.5 ± 1.0	0.88 ± 0.01	0.13 ± 0.02	22.1 ± 1.0	1.5 ± 0.5
Remove Time Decay Mechanism	84.7 ± 1.0	82.9 ± 1.1	0.86 ± 0.02	0.15 ± 0.03	20.5 ± 0.9	1.8 ± 0.6
Remove Priority Experience Playback	87.5 ± 0.9	85.8 ± 1.0	0.89 ± 0.01	0.11 ± 0.02	23.3 ± 1.0	1.3 ± 0.5
Remove Feedback Constraint Mechanism	88.9 ± 0.8	87.1 ± 0.9	0.91 ± 0.01	0.10 ± 0.02	24.8 ± 1.1	8.5 ± 1.2

This study found that the system still maintained good robustness in a variety of disturbance scenarios, as shown in

Table 9. When the text is 20% missing, the accuracy rate is still more than 90%, indicating that the system has strong

429 adaptability to the lack of text data; In the face of 15% tag noise, the performance retention rate is also high, indicating
430 that the model has a good tolerance to noise; Under the scenario of 10% missing behavior log and 1 hour feedback delay,
431 the system performance is still in an acceptable range although it has declined; When migrating across platforms, the
432 performance is reduced, but it can still maintain a certain degree of effectiveness, indicating that the system has a certain
433 generalization ability. These results show that the system has strong stability and reliability in complex practical
434 application scenarios. Figure 11 shows the robustness performance of the system under different disturbance scenarios.
435 The results show that cross platform migration has the greatest impact, and the model still maintains high stability under
436 the condition of text missing and feedback delay.



437
438 Figure 11. System robustness across disturbance scenarios
439

440

Table 9. Results of system robustness verification

Disturbance Scenario	Accuracy Retention Rate/%	F1 Retention Rate/%	AUC Retention Rate/%	Reward Retention Rate/%	Maximum Risk Fluctuation	Stability Index
Text Missing (20%)	92.3 ± 2.3	91.5±2.1	93.1 ± 1.8	88.7 ± 2.5	0.12 ± 0.03	0.82 ± 0.03
Label Noise (15%)	89.5 ± 2.5	88.7±2.3	90.2 ± 2.0	85.6 ± 2.7	0.15 ± 0.04	0.78 ± 0.03
Missing Behavior Log (10%)	87.1 ± 2.7	86.3±2.5	88.4 ± 2.2	82.3 ± 2.9	0.18 ± 0.05	0.75 ± 0.04
Cross Platform Migration (Dithering → Microblogging)	84.2 ± 3.0	83.5±2.8	85.7 ± 2.4	78.9 ± 3.1	0.21 ± 0.06	0.70 ± 0.04
Feedback Delay (1H)	90.6 ± 2.4	89.8±2.2	91.5 ± 1.9	86.8 ± 2.6	0.14 ± 0.04	0.80 ± 0.03

441 **4.5 Analysis of personalized intervention effect and user acceptance**

442 The research experiment shows that personalized intervention has significant effects on users with different risk
443 levels, as shown in Table 10. The acceptance rate of low-risk users was the highest, and the risk score and use duration
444 decreased slightly, indicating that moderate intervention can effectively prevent risk escalation; The risk score and use
445 duration of moderate risk users decreased significantly, and the acceptance rate also maintained a high level, indicating
446 that targeted intervention can effectively adjust behavior patterns; The risk reduction of high-risk users is the largest.
447 Although the acceptance rate is slightly low, it can still reach more than 78%, indicating that deep intervention can
448 effectively alleviate the symptoms of addiction; The intervention effect of users with increased risk is also significant,
449 and they are willing to continue to use it. Overall, the personalized intervention not only effectively reduced the risk of
450 addiction and the length of use, but also maintained a high degree of user acceptance and satisfaction, which verified the
451 practical application value of the system. Figure 12 visually compares the different characteristics of different risk users
452 in risk reduction, change of use duration, acceptance, satisfaction and willingness to continue using.

Personalized Intervention Effects and User Acceptance

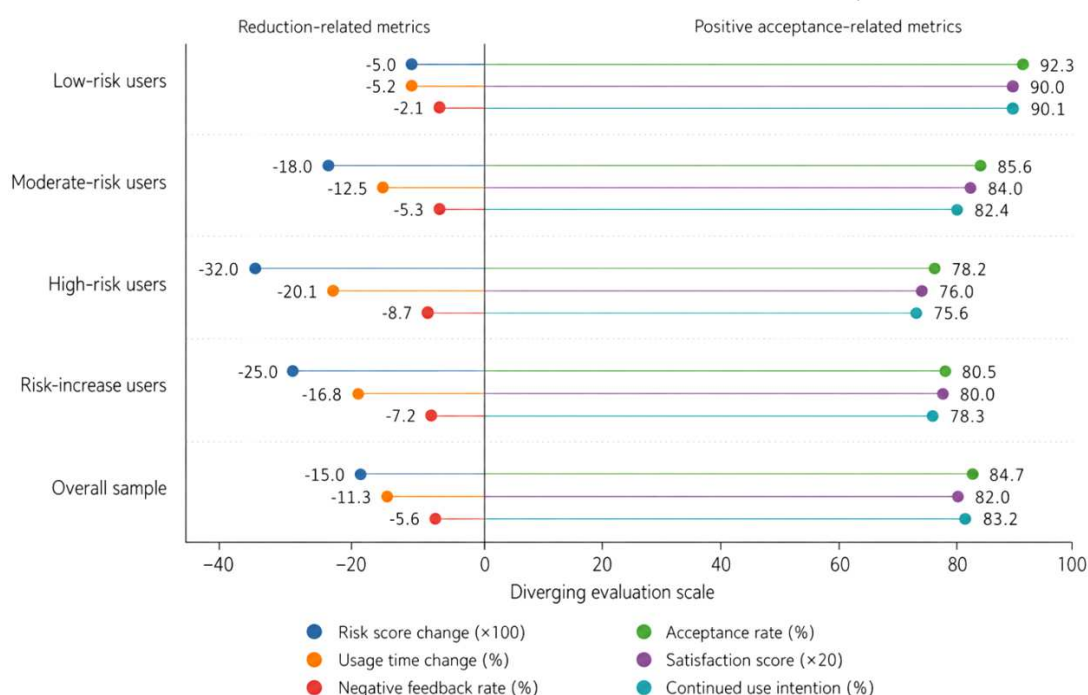


Figure 12. Visualization of personalized intervention effect

Table 10. Evaluation form of personalized intervention effect and user acceptance

Evaluation Object	Change In Risk Score	Usage Duration Change/%	Acceptance Rate/%	Satisfaction Score	Negative Feedback Rate/%	Willingness To Continue Using
Low Risk Users	-0.05 ± 0.02	-5.2 ± 1.8	92.3±1.5	4.5 ± 0.3	2.1 ± 0.7	90.1 ± 1.6
Medium Risk User	-0.18 ± 0.04	-12.5 ± 2.3	85.6±1.7	4.2 ± 0.4	5.3 ± 0.9	82.4 ± 1.8
High Risk Users	-0.32 ± 0.05	-20.1 ± 2.5	78.2±1.9	3.8 ± 0.5	8.7 ± 1.1	75.6 ± 2.0
Users With Increased Risk	-0.25 ± 0.04	-16.8 ± 2.4	80.5±1.8	4.0 ± 0.4	7.2 ± 1.0	78.3 ± 1.9
Population Sample	-0.15 ± 0.03	-11.3 ± 2.1	84.7 ± 1.6	4.1±0.4	5.6 ± 0.9	83.2 ± 1.7

5 Discussion

5.1 Effectiveness and application value of adaptive digital intervention system

The adaptive digital intervention system integrated NLP emotion analysis and reinforcement learning constructed in this study has shown significant technical effectiveness and practical application value in the prevention and control of teenagers' social media addiction (Table 11). In the aspect of state perception, the NLP state perception model is optimized by attention mechanism and context semantic understanding, with an accuracy rate of 90.6%, which is 2.7%

462 higher than the traditional transformer model, and can effectively capture the implicit emotional tendencies in teenagers'
463 online texts, such as subtle psychological clues such as implicit anxiety and loneliness expression. The dynamic risk
464 representation module integrates the emotional score, behavior characteristics, historical risk and time attenuation
465 mechanism.

466 The root mean square error of the complete model is reduced to 0.10, and the correlation coefficient is 0.89. It
467 realizes the continuous characterization of addiction risk from early accumulation to out of control, and solves the
468 problem that the static assessment cannot reflect the evolution of risk. In terms of personalized intervention strategies,
469 the risk reduction rate of constraint reinforcement learning strategy was 32.1%, the user acceptance rate was 82.4%,
470 which was 16.8% and 20.3% higher than the fixed rule strategy, and the constraint violation rate was only 1.2%, which
471 balanced the intervention effect and user experience.

472 The system adapts intervention actions for different risk levels: low-risk users use mild reminders, with an
473 acceptance rate of 92.3%; Medium risk users use content to replace recommendations, and the risk reduction rate is
474 25.6%; High risk users started home school collaboration, and the risk reduction rate reached 35.2%, realizing the full
475 cycle coverage from prevention to mitigation. The robustness test shows that the accuracy retention rate of the system is
476 over 89% in the disturbance scenarios such as 20% text missing and 15% label noise, and the accuracy retention rate
477 from dithering to microblogging is 84.2% when migrating across platforms, providing stability guarantee for the actual
478 deployment.

Type Of Study	Core Technology Methods	Risk Identification Dimension	Characteristics Of Intervention Strategies	Ethical Compliance Mechanism
Research On Static Risk Screening	Questionnaire/single behavior indicator	Static threshold judgment	No or unified warning	Lack of privacy protection and excessive intervention prevention and control
Fixed Rule Intervention System	Preset threshold trigger logic	Fixed behavior indicators	Unified intervention action	No personalized adaptation and user feedback mechanism
Traditional Machine Learning Intervention Model	Classification/regression model	Static risk label	Fixed strategy based on model output	Lack of constraint mechanism and manual review
Adaptive Digital Intervention System	NLP emotion analysis+constraint reinforcement learning	Dynamic multi feature fusion (emotion+behavior+historical risk+time decay)	Personalized adaptive strategy	Privacy desensitization, manual review, constraint rules

480

5.2 Privacy protection, ethical risks and algorithmic boundaries in adolescent scenarios

481

Teenagers, as a sensitive group, the system design should strictly follow the privacy protection and ethical norms,

482

and clarify the application boundary of the algorithm. In terms of privacy protection mechanism, data collection adopts

483

the principle of informed consent and guardian authorization, and all data are desensitized to avoid the disclosure of

484

identifiable individual information; Set trigger conditions for manual review, and start manual intervention when the risk

485

index exceeds 0.6 to prevent stigmatization or excessive intervention caused by algorithm misjudgment. The ablation

486

experiment results showed that after removing the feedback constraint mechanism, the constraint violation rate increased

487

from 1.2% to 8.5%, which verified the key role of constraint rules on ethical compliance. In terms of ethical risk

488

prevention and control, the system does not take the model output as the conclusion of psychological diagnosis, only

489

provides early warning and auxiliary support, and reserves the user's right to choose intervention actions, such as accept,

490

delay or reject; The home school collaboration prompt is only applicable to high-risk users and needs to be confirmed

491

by the guardian to avoid unauthorized intervention. The boundary of the algorithm is clearly defined as an “auxiliary

492 tool”, which does not replace professional psychological intervention and complements offline services to ensure the
493 safety and professionalism of the intervention.

494 **5.3 Limitations and practical deployment challenges of system popularization and** 495 **Application**

496 The system promotion is faced with the dual challenges of technology and reality. In terms of technical limitations,
497 the accuracy retention rate from dithering to microblogging during cross platform migration is 84.2%, indicating that
498 the adaptability of the model to the semantic features of different platforms still needs to be improved; The dynamic
499 changes of teenagers’ network language require the NLP model to be continuously updated, which increases the
500 maintenance cost; In addition, the system relies on multi-source data, and some platforms may restrict data access,
501 affecting model training and deployment. In terms of practical deployment challenges, there are differences in user
502 compliance. The willingness of high-risk users to continue using is only 75.6%. It is necessary to further optimize the
503 intervention methods to improve the long-term stickiness; Home school collaboration relies on the cooperation of parents
504 and schools. Some parents lack the ability or time to intervene and need to support training mechanisms; At the policy
505 level, it is necessary to comply with *the regulations on the protection of minors’ networks* and other regulations to ensure
506 the compliance of data processing and intervention. In the future, we need to improve the cross platform generalization
507 ability through transfer learning, and build a “online monitoring offline support” collaborative system combined with
508 offline psychological services to meet the above challenges.

509 **6 Conclusion and future research directions**

510 **6.1 Core conclusion**

511 This study aimed at the precise needs of teenagers’ social media addiction prevention and control, constructed an
512 adaptive digital intervention system integrating NLP emotion analysis and constraint reinforcement learning, and
513 realized the closed-loop design from risk dynamic perception to personalized strategy optimization. The system

514 integrates multi-source data, uses the optimized NLP state perception model to capture adolescents' implicit emotional
515 tendencies, and combines the dynamic risk representation module of time decay mechanism to effectively distinguish
516 short-term emotional fluctuations from long-term risk accumulation, which solves the problem that static assessment
517 cannot reflect the evolution of risk; The constraint reinforcement learning strategy realizes the precise adaptation of
518 intervention action and risk level through multi-dimensional reward function and priority experience playback
519 mechanism. The risk reduction rate is 32.1%, and the user acceptance rate is 82.4%, which is significantly higher than
520 the fixed rule strategy.

521 At the same time, the constraint violation rate is controlled within 1.2%, which balances the effect and ethical
522 requirements; The system effectively protects teenagers' privacy and avoids excessive intervention or stigmatization
523 through mechanisms such as data desensitization, manual review trigger and daily intervention limit; The experimental
524 results verify the practical effectiveness of the system. The acceptance rate of low-risk users is 92.3%, the risk score of
525 high-risk users is reduced by 0.32, and the length of use is reduced by 20.1%. It provides a technical paradigm and
526 practical path for the prevention and control of teenagers' social media addiction.

527 **6.2 Future outlook**

528 In the future, we can deepen the research from multiple dimensions: Firstly, according to the semantic
529 characteristics of different social media platforms, we introduce the optimization model of transfer learning technology
530 to improve the accuracy retention rate of cross platform migration and expand the scope of application of the system;
531 Secondly, integrate multimodal data such as images and voice to enrich the dimension of state perception and further
532 improve the comprehensiveness and accuracy of risk identification; In addition, combined with game design or social
533 incentive, it can enhance the continuous use intention of high-risk users and strengthen the long-term effect of
534 intervention; At the same time, it is connected with the school psychological counseling room, family monitoring system
535 and professional psychological institutions to establish a collaborative mechanism of "online monitoring and early

536 warning - offline intervention support” to solve complex psychological problems that cannot be covered by the algorithm;
537 Attention visualization, decision path tracking and other technologies can also be introduced to improve the transparency
538 of the model, so that users and guardians can understand the basis of intervention decision-making and enhance trust;
539 Finally, we will continue to follow up the updating of regulations such as *the regulations on the protection of minors’*
540 *networks*, optimize the data processing process and intervention boundary, ensure the compliance and sustainability of
541 the system, and further enhance the practicability and social value of the system.

542 **Author contributions**

543 Qing Bian: Conceptualization, Methodology, Formal analysis, Data Curation, Visualization, Writing—original draft,
544 Writing—review and editing.

545 **Funding**

546 This study did not receive additional funding sponsorship.

547 **Ethical approval**

548 **Informed consent**

549 All methods of this study adhere to the Declaration of Helsinki. This study was approved by Guangzhou College
550 of Commerce Ethics Committee (GCCEC-20251254). All individual participants included in this study have signed the
551 paper version of the informed consent form.

552 **Declaration of Interest**

553 The author declares that they have no conflict of interest.

554 **Data Statement**

555 The original data of this study can be obtained from the corresponding author.

556

557 **References**

- 558 [1] Keles B, McCrae N, Grealish A. A systematic review: The influence of social media on depression, anxiety and
559 psychological distress in adolescents. *Journal of Adolescence*, 2020, 79: 79-93.
- 560 [2] Odgers C L, Jensen M R. Annual Research Review: Adolescent mental health in the digital age: Facts, fears, and
561 future directions. *Journal of Child Psychology and Psychiatry*, 2020, 61(3): 336-348.
- 562 [3] Orben A. Teenagers, screens and social media: A narrative review of reviews and key studies. *Social Psychiatry and
563 Psychiatric Epidemiology*, 2020, 55(4): 407-414.
- 564 [4] Beyens I, Pouwels J L, van Driel I I, et al. The effect of social media on well-being differs from adolescent to
565 adolescent. *Scientific Reports*, 2020, 10: 10763.
- 566 [5] Boer M, van den Eijnden R J J M, Boniel-Nissim M, et al. Adolescents' intense and problematic social media use
567 and their well-being in 29 countries. *Journal of Adolescent Health*, 2020, 66(6S): S89-S99.
- 568 [6] Boer M, Stevens G W J M, Finkenauer C, et al. Social media use intensity, social media use problems, and mental
569 health among adolescents: Investigating directionality and mediating processes. *Computers in Human Behavior*,
570 2021, 116: 106645.
- 571 [7] Marino C, Gini G, Angelini F, et al. Social norms and e-motions in problematic social media use among adolescents.
572 *Addictive Behaviors Reports*, 2020, 11: 100250.
- 573 [8] Kross E, Verduyn P, Sheppes G, et al. Social media and well-being: Pitfalls, progress, and next steps. *Trends in
574 Cognitive Sciences*, 2021, 25(1): 55-66.
- 575 [9] Cheng C, Lau Y C, Chan L, et al. Prevalence of social media addiction across 32 nations: Meta-analysis with
576 subgroup analysis of classification schemes and cultural values. *Addictive Behaviors*, 2021, 117: 106845.
- 577 [10] Shannon H, Bush K, Villeneuve P J, et al. Problematic social media use in adolescents and young adults: Systematic
578 review and meta-analysis. *JMIR Mental Health*, 2022, 9(4): e33450.

579 [11] Arrivillaga C, Rey L, Extremera N. A mediated path from emotional intelligence to problematic social media use
580 in adolescents: The serial mediation of perceived stress and depressive symptoms. *Addictive Behaviors*, 2022, 124:
581 107095.

582 [12] Geurts S M, Koning I M, Vossen H G M, et al. Rules, role models or overall climate at home? Relative associations
583 of different family aspects with adolescents' problematic social media use. *Comprehensive Psychiatry*, 2022, 116:
584 152318.

585 [13] Leijse M M L, Koning I M, van den Eijnden R J J M. The influence of parents and peers on adolescents' problematic
586 social media use revealed. *Computers in Human Behavior*, 2023, 143: 107705.

587 [14] Geurts S M, Koning I M, van den Eijnden R J J M, et al. Predicting adolescents' problematic social media use from
588 profiles of internet-specific parenting practices and general parenting dimensions. *Journal of Youth and*
589 *Adolescence*, 2023, 52(9): 1829-1843.

590 [15] Casale S, Akbari M, Seydavi M, et al. Has the prevalence of problematic social media use increased over the past
591 seven years and since the start of the COVID-19 pandemic? A meta-analysis of the studies published since the
592 development of the Bergen Social Media Addiction Scale. *Addictive Behaviors*, 2023, 147: 107838.

593 [16] Montag C, Demetrovics Z, Elhai J D, et al. Problematic social media use in childhood and adolescence. *Addictive*
594 *Behaviors*, 2024, 153: 107980.

595 [17] Xiao Y, Meng Y, Brown T T, et al. Addictive screen use trajectories and suicidal behaviors, suicidal ideation, and
596 mental health in US youths. *JAMA*, 2025.

597 [18] Nagata J M, Hur J O, Talebloo J, et al. Problematic social media use interventions for mental health outcomes in
598 adolescents. *Current Psychiatry Reports*, 2025, 27(9): 491-499.

599 [19] Chancellor S, De Choudhury M. Methods in predictive techniques for mental health status on social media: A
600 critical review. *npj Digital Medicine*, 2020, 3: 43.

601 [20] Su C, Xu Z, Pathak J, et al. Deep learning in mental health outcome research: A scoping review. *Translational*
602 *Psychiatry*, 2020, 10: 116.

603 [21] Low D M, Rumker L, Talkar T, et al. Natural language processing reveals vulnerable mental health support groups
604 and heightened health anxiety on Reddit during COVID-19: Observational study. *Journal of Medical Internet*
605 *Research*, 2020, 22(10): e22635.

606 [22] Skaik R, Inkpen D. Using social media for mental health surveillance: A review. *ACM Computing Surveys*, 2021,
607 53(6): 1-31.

608 [23] Kim J, Lee D, Park E. Machine learning for mental health in social media: Bibliometric study. *Journal of Medical*
609 *Internet Research*, 2021, 23(3): e24870.

610 [24] Le Glaz A, Haralambous Y, Kim-Dufoir D H, et al. Machine learning and natural language processing in mental
611 health: Systematic review. *Journal of Medical Internet Research*, 2021, 23(5): e15708.

612 [25] Zhang T, Schoene A M, Ji S, et al. Natural language processing applied to mental illness detection: A narrative
613 review. *npj Digital Medicine*, 2022, 5: 46.

614 [26] Sedgwick R, Bittar A, Kalsi H, et al. Investigating online activity in UK adolescent mental health patients: A
615 feasibility study using a natural language processing approach for electronic health records. *BMJ Open*, 2023, 13(5):
616 e061640.

617 [27] Ibrahim S T, Li M, Patel J, et al. Utilizing natural language processing for precision prevention of mental health
618 disorders among youth: A systematic review. *Computers in Biology and Medicine*, 2025, 188: 109859.

619 [28] Gönül S, Namlı T, Coşar A, et al. A reinforcement learning based algorithm for personalization of digital, just-in-
620 time, adaptive interventions. *Artificial Intelligence in Medicine*, 2021, 115: 102062.

621 [29] Wang s, Zhang C, Kr öse B, et al. Optimizing adaptive notifications in mobile health intervention systems:
622 Reinforcement learning from a data-driven behavioral simulator. *Journal of Medical Systems*, 2021, 45: 12.

- 623 [30] Carpenter S M, Menictas M, Nahum-Shani I, et al. Developments in mobile health just-in-time adaptive
624 interventions for addiction science. *Current Addiction Reports*, 2020, 7(3): 280-290.
- 625 [31] Balaskas A, Schueller S M, Cox A L, et al. Ecological momentary interventions for mental health: A scoping review.
626 *PLoS ONE*, 2021, 16(3): e0248152.
- 627 [32] Wies B, Landers C, Ienca M. Digital mental health for young people: A scoping review of ethical promises and
628 challenges. *Frontiers in Digital Health*, 2021, 3: 697072.