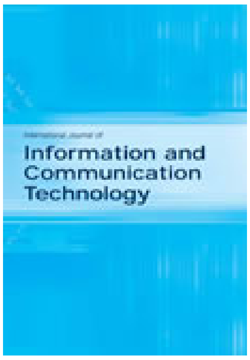
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Value-oriented meta-adaptive reinforcement learning for optimising emotional intervention

Bing Lin

DOI: 10.1504/IJICT.2025.10075265

Article History:

Received:	28 September 2025
Last revised:	29 October 2025
Accepted:	30 October 2025
Published online:	12 January 2026

Value-oriented meta-adaptive reinforcement learning for optimising emotional intervention

Bing Lin

Faculty of Teacher Education,
Zhangzhou City Vocational College,
Zhangzhou, 363000, China
Email: lyt110603@163.com

Abstract: Emotional intervention plays a crucial role in mental health support, yet traditional approaches often lack the dynamic adaptability to individual states and contextual changes. To address these limitations, this study proposes a value-guided meta-adaptive reinforcement learning framework. By integrating meta-learning with deep reinforcement learning, this approach enables intervention strategies to rapidly adapt to users' real-time emotional states and long-term needs. We design an attention-based meta-policy network to extract shared representations across users and introduce a value function to quantify long-term psychological benefits. Furthermore, the framework employs proximal policy optimisation for policy training and dynamically adjusts hyperparameters through a meta-adaptive mechanism to handle non-stationary intervention environments. Experiments on simulated and real-world user datasets demonstrate that the proposed method achieves approximately 22% higher emotional improvement rates and 33% faster convergence speed compared to the best baseline.

Keywords: meta-adaptive reinforcement learning; affective computing; personalised intervention; proximal policy optimisation; PPO.

Reference to this paper should be made as follows: Lin, B. (2025) 'Value-oriented meta-adaptive reinforcement learning for optimising emotional intervention', *Int. J. Information and Communication Technology*, Vol. 26, No. 50, pp.35–51.

Biographical notes: Bing Lin is an Associate Professor in the Faculty of Teacher Education at Zhangzhou City Vocational College, China. She received her Bachelor's in Education in 1993 and Master's in Educational Management in 2006 both from Fujian Normal University, China. Her research interests include mental health education, and meta-adaptive reinforcement learning.

1 Introduction

Mental health has emerged as a significant global public health challenge, and timely, effective psychological and emotional interventions are crucial for alleviating symptoms such as stress, anxiety, and depression. While traditional interventions like cognitive behavioural therapy have proven effective, they often face limitations in accessibility, high costs, and a 'one-size-fits-all' approach, making it difficult to address the heterogeneous and time-varying needs individuals exhibit within their dynamically changing life contexts (Wenzel, 2017). In recent years, with the proliferation of mobile health technologies, computationally driven personalised affective interventions have demonstrated immense potential. The core challenge lies in constructing an intelligent system capable of automatically learning and dynamically optimising intervention strategies. Such strategies must not only respond to the user's current state but also focus on promoting long-term mental health benefits (Ng and Weisz, 2016).

Computational approaches for personalised emotional interventions have emerged as a significant research direction in health informatics. Traditionally, rule-based systems and statistical analyses have been employed to deliver static intervention content based on users' self-reported data, yet these methods lack dynamic adaptability (Nye et al., 2023). The just-in-time adaptive interventions (JITAIs) paradigm has come about because of improvements in mobile sensing and passive data collection technologies. Its goal is to give effective interventions at the right time. Nahum-Shani et al. (2016) methodically delineated the design principles of JITAIs, establishing a theoretical framework for the development of adaptive intervention systems. Researchers subsequently utilised machine learning techniques, including clustering and classification algorithms, to ascertain user states from previous data or to forecast intervention time. For example, Mohr et al. (2017) used logistic regression models to predict times when depressed people would feel low, which led to interventions. However, most of these approaches focus on state recognition or short-term prediction rather than sequential decision optimisation, failing to fully account for the potential long-term cumulative effects of interventions.

The introduction of reinforcement learning into this field aims to directly optimise sequential decision problems and achieve personalised intervention strategies. Early studies modelled the intervention problem as a contextual multi-armed bandit, balancing exploration (trying new interventions) and exploitation (selecting the currently optimal intervention) to optimise immediate gains. For instance, Gönül et al. (2021) employed Thompson sampling to select notification types that maximise immediate user engagement. However, the contextual multi-armed bandit (CBM) (Cannelli et al., 2023) can only handle instantaneous rewards and cannot plan for long-term objectives. To address this, deep reinforcement learning (DRL) methods such as deep Q-networks (DQNs) (Barto, 2021) and policy gradient algorithms (Koo et al., 2010) have been applied to more complex intervention scenarios. Their advantage lies in capturing long-term value through value function approximations. Yang et al. (2024) demonstrated DRL potential for designing treatment plans for multiple chronic disease patients in simulated environments. Nevertheless, standard DRL methods typically require extensive interaction data with the environment to converge, which is neither cost-effective nor ethically feasible in intervention studies involving real users. Furthermore, strategies learned from one user cohort often struggle to generalise directly to new users, presenting a 'cold start' problem.

In recent years, the framework of meta-learning, or 'learning to learn' has offered a promising path to address data efficiency and rapid adaptation (Hospedales et al., 2021). Its core idea is to extract shared knowledge from a series of related tasks, enabling rapid adaptation to new tasks with minimal samples. Model-agnostic meta-learning (MAML) algorithms have garnered significant attention for their flexibility and have been applied in healthcare. For instance, Singh and Malhotra (2023) explored using MAML to rapidly personalise digital intervention strategies for different patients. In affective computing, preliminary attempts have also been made to apply meta-learning for cross-subject adaptation of emotion recognition models, though these efforts primarily focus on state recognition (perception) rather than decision-making (intervention) (Zhang et al., 2022). A more critical limitation lies in existing research combining meta-learning with RL for interventions. Most studies implicitly optimise algorithmically predefined reward signals, failing to explicitly adopt user-centric 'values' grounded in psychological theory and aligned with long-term well-being as the core objective (Kazdin, 2017). Defining, quantifying, and effectively integrating such long-term value into meta-reinforcement learning frameworks remains an open challenge.

To address these challenges, this paper proposes a value-oriented meta-adaptive reinforcement learning framework. The core contributions of this research are threefold: first, we design a novel meta-adaptive reinforcement learning paradigm that deeply integrates meta-learning algorithms like MAML with proximal policy optimisation (PPO). This enables the central policy model to extract shared knowledge from diverse user groups and rapidly personalise intervention strategies for new users. Second, we introduce a specially designed value function that not only quantifies immediate emotional state feedback but, more importantly, incorporates long-term well-being assessment metrics grounded in psychological theory. This ensures the learned intervention strategies are genuinely 'value-oriented' committed to maximising users' lifetime psychological well-being. Finally, we conducted extensive experimental validation across simulated environments incorporating multidimensional emotional signals and real-world datasets. Results demonstrate that compared to existing reinforcement learning and static intervention baselines, our proposed method exhibits significant advantages in intervention effectiveness, strategy adaptability for new users, and overall sample efficiency.

2 Related theoretical research

2.1 Computational emotion intervention

The computationalisation of emotional interventions aims to leverage data-driven models and algorithms to provide quantifiable, scalable, and personalised solutions for mental health support and promotion (Ramdoss et al., 2012). The emergence of this field relies heavily on the rapid advancement of mobile computing and sensing technologies, enabling continuous, passive collection of multimodal data across diverse contexts. This includes geolocation, physical activity, and communication patterns captured via smartphones, alongside physiological signals such as heart rate variability monitored through wearable devices. This granular data establishes the foundation for constructing dynamic computational models of user psychological states, transcending traditional evaluation methods reliant on discrete self-report questionnaires and enabling near-real-time perception of individual emotional shifts.

Early computational intervention methods mostly utilised rule-based static logic (Partala and Surakka, 2004). These systems usually included preset criteria based on clinical knowledge. For example, they might send an encouraging message or suggest a relaxation exercise when they noticed a big drop in social activity or when self-reported emotional ratings fell below a certain level. These systems had fixed intervention logic that could not change based on long-term user feedback, even though they were able to automate some tasks. They had a hard time adjusting to how different people were and how complicated state evolution was. To fix this problem, the JITAIs framework was suggested. Its core principle involves delivering the most appropriate intervention type at the most effective moment to maximise intervention efficacy. The theoretical framework of JITAIs emphasises dynamic modelling of users' internal states (e.g., emotions, stress), external environments (e.g., location, social settings), and the historical effectiveness of interventions themselves. This marks a significant shift in affective intervention research from static approaches toward adaptive systems.

With the growth of available data, machine learning methods have naturally been introduced to enhance JITAIs' decision-making capabilities. Early research primarily focused on using supervised learning models, such as logistic regression and support vector machines, to learn from historical data and predict the optimal timing for interventions or users' short-term responses. For instance, some studies employed classification models to predict users' stressful events or moments of low mood as signals to trigger interventions. These approaches significantly improved predictive accuracy for the 'when to intervene' question. However, they essentially decomposed intervention decisions into a series of independent prediction tasks, failing to treat intervention as a continuous decision-making process – that is, they ignored the potential long-term impact of current interventions on users' future states. Reinforcement learning naturally models personalised interventions as a sequential decision problem. Its objective is to directly learn a strategy that maximises long-term cumulative rewards through interaction with the environment, theoretically offering a way to overcome the short-sightedness of previous approaches.

Nevertheless, the overall development of computational emotional interventions still faces numerous challenges. These include ensuring model reliability in scenarios with small sample sizes and guaranteeing that learned strategies are not only effective but also ethically sound.

2.2 Application of reinforcement learning in personalised interventions

Reinforcement learning, owing to its inherent advantages in sequential decision-making problems, has emerged as the core technological paradigm for achieving personalised interventions. This framework formalises the intervention process as an interaction between an agent (the intervention system) and an environment (the user): at each decision point, the agent selects an intervention action a_t based on the current user state s_t, after which the environment transitions to a new state s_{t+1} and generates an immediate reward r_t. Its ultimate goal is to learn a policy $\pi(a|s)$ that maximises the long-term cumulative reward.

$$\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right] \tag{1}$$

where γ represents the discount factor. This formalisation enables RL to transcend immediate benefits and directly optimise long-term impacts on user health outcomes, thereby addressing the 'myopia' issue inherent in traditional approaches.

The CBM model was the main focus of early RL applications. CBM is a simpler version of RL that finds the best balance between exploration and exploitation to get the most immediate rewards. These strategies are straightforward to use and do not take up a lot of computer power. They work best in situations when treatments have quick benefits but weak long-term consequences. The main problem with CBM models is that they cannot simulate state transitions and long-term returns, which makes it hard to deal with possible delayed or sequential effects of treatments. The advantages of a solitary cognitive restructuring exercise may only become apparent after many days.

To capture this long-term dependency, research has progressively shifted toward employing the full Markov decision process (MDP) framework and DRL algorithms. Value-based methods, such as DQN (Mnih et al., 2015), formulate policies by approximating the optimal action-value function $Q^*(s, a)$ through neural networks. In contrast, policy-based methods like PPO (Gu et al., 2021) directly parameterise and optimise the policy action $\pi_\theta(a|s)$, often demonstrating advantages when handling continuous action spaces or requiring more stable training. PPO ensures training stability by limiting the step size of policy updates, with its objective function typically formulated as:

$$L^{CLIP}(\theta) = \mathbb{E}\left[\min\left(r_t(\theta)A_t, clip\left(r_t(\theta), 1-\grave{o}, 1+\grave{o}\right)A_t\right)\right] \tag{2}$$

where $r_t(\theta)$ represents the policy probability ratio, and A_t denotes the value function estimator. Such algorithms can learn more complex, state-dependent stochastic policies – for instance, adaptively adjusting the intensity and type of intervention content based on the user's current stress level and historical responses.

Even though DRL has a lot of potential, it has a lot of problems when it comes to real-world intervention applications. The main problem is sample efficiency: DRL usually needs a lot of interaction data to work, which is hard to get when you have to follow ethical rules and keep costs low when using real users. Second, there is a conflict between safety and exploration; in sensitive areas like mental health, exploring without knowing what you're doing could put you in danger.

2.3 Meta learning and application in healthcare

Meta-learning, or 'learning to learn', aims to design models capable of extracting shared knowledge or experience from a series of related tasks. This enables rapid adaptation to new tasks with minimal samples or interactions. This paradigm provides powerful theoretical tools for addressing the widespread challenges of data scarcity and model generalisation in real-world machine learning scenarios. Among various meta-learning algorithms, MAML (Finn et al., 2018) has garnered significant attention for its versatility and simplicity. MAML aims to discover an initial set of model parameters that enables excellent performance on a new task T_i after just one or a few gradient updates using minimal data. Its core optimisation objective can be formalised as:

$$\min_{\theta} \sum_{T_i \sim p(T)} L_{T_i}\left(f_{\theta_i'}\right) \tag{3}$$

$$\theta_i' = \theta - \alpha \nabla_{\theta} L_{T_i}\left(f_{\theta}\right) \tag{4}$$

where θ represents the shared initial parameters sought by the meta-learner, α denotes the inner-layer learning rate, and L_{T_i} signifies the task's loss function. By conducting meta-training across diverse task distributions, MAML endows the model with an innate ability to rapidly adapt to new tasks.

In the healthcare field, meta-learning's ability to quickly adjust is quite useful. Medical data often displays 'small-sample' traits, indicating that labelled data for particular diseases or individual patients is severely restricted, yet extensive data is available across various diseases or patients – this situation is ideally suited to meta-learning's 'multi-task' framework. For example, in the diagnosis of medical images, researchers used the MAML framework to consider classification jobs as discrete tasks during meta-training. These tasks were based on different illness categories or datasets from different medical centres. The resulting model quickly reached a high level of diagnostic accuracy with just a few new disease-type picture slices. This greatly reduced the need for massive amounts of labelled data and opened up new ways to help with the identification of rare diseases.

However, applying meta-learning to healthcare, particularly clinical decision support, still faces significant challenges. The first is defining and aligning task distributions – ensuring sufficient similarity between meta-training tasks and novel meta-testing tasks to guarantee effective knowledge transfer. Second is the challenge of model interpretability and safety. The inherent complexity of meta-learning models makes their decision-making logic harder to trace and validate, potentially limiting their clinical adoption in healthcare settings where tolerance for error is extremely low.

3 Methodology

3.1 Problem formulation

This paper formalises the personalised emotional intervention problem as a partially observable Markov decision process (POMDP) (Littman, 2009), a framework that effectively captures the uncertainty and partial observability inherent in the intervention process. A POMDP can be defined by a tuple (S, A, O, T, Ω, R, γ), where each element holds specific meaning within the context of emotional intervention. The state space S represents the user's actual mental health state – a latent variable inaccessible directly, potentially encompassing dimensions such as emotional state, cognitive patterns, physiological arousal levels, and environmental context. Since the complete state cannot be directly observed, the agent (i.e., the intervention system) can only infer the user's state through information in the observation space O. These observations typically originate from mobile device sensor data, user-reported mood scores, interaction logs, etc. Their relationship with the true state is determined by the observation function $\Omega(o|s, a)$, which defines the probability of observing o after executing action a in state s.

The action space A represents all intervention options the system can execute, such as sending specific types of messages, adjusting intervention frequency or intensity, or even choosing not to intervene at a given moment. The state transition function $T(s'|s, a)$ describes the dynamic changes in the user's state under the influence of intervention actions, i.e., the probability of transitioning from the current state s to a new state s′ after executing action a. This captures both the randomness of intervention effects and the complexity of user state evolution.

Within the POMDP framework, the agent must maintain a belief state $b_t(s)$ – a probability distribution over the state space S—representing the confidence in the current true state s given the observed history and action sequence. The belief state updates according to Bayesian rules:

$$b_{t+1}(s') \propto \Omega\left(o_{t+1} \middle| s', a_t\right) \sum_{s \in S} T\left(s' \middle| s, a_t\right) b_t(s) \tag{5}$$

Strategy $\pi(a|b)$ is a function that maps the current belief state to a probability distribution over the action space.

The reward function $R(s, a)$ is central to the model design, quantifying the immediate payoff from executing action a in state s. In this study, we designed it to be value-oriented, meaning the reward reflects not only short-term emotional improvement (e.g., reduced self-reported negative emotions) but also includes proxy measures of long-term psychological well-being (e.g., enhanced resilience, improved social functioning). The objective of the POMDP is to find an optimal policy π^* that maximises the expected cumulative discounted reward:

$$\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R\left(s_t, a_t\right)\right] \tag{6}$$

where $\gamma \in [0, 1)$ is the discount factor used to balance the importance of immediate versus future returns.

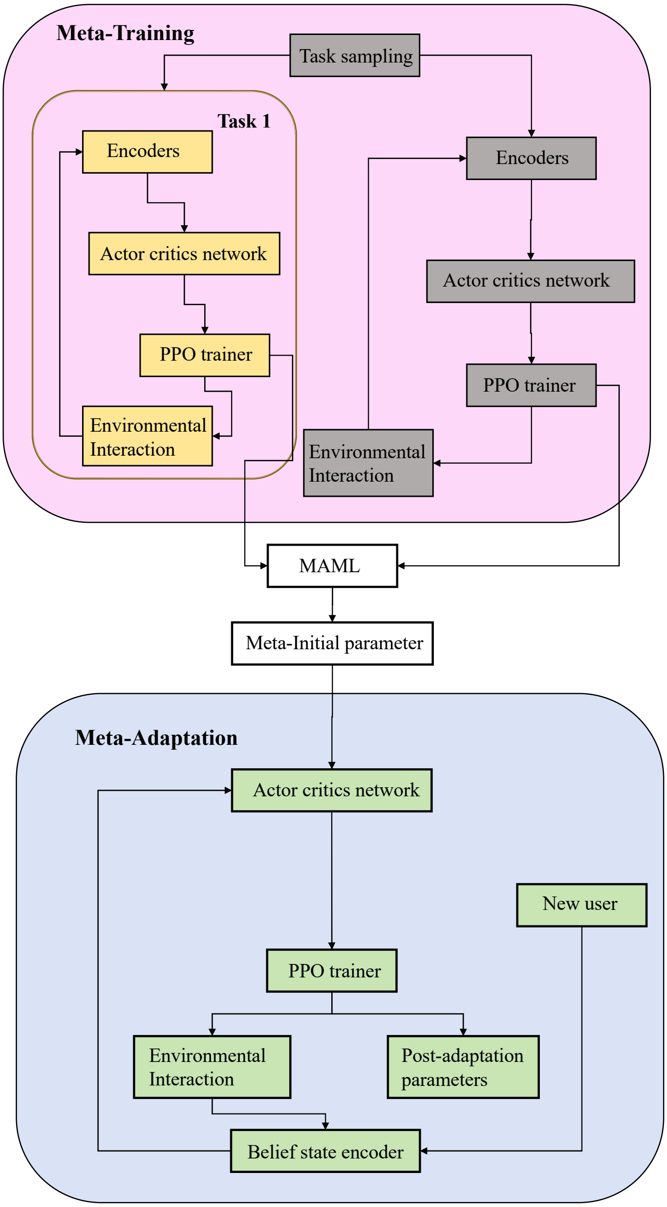
3.2 VG-MARL framework overview

The proposed value-guided meta-adaptive reinforcement learning (VG-MARL) framework is a hierarchical learning system whose core objective is to rapidly generate personalised, long-term-benefit-oriented emotional intervention strategies for unknown new users. As shown in Figure 1, the framework comprises two main phases: an offline meta-training phase and an online meta-adaptation phase. During the meta-training phase, the system leverages a source domain containing historical interaction data from multiple users. Through a MAML mechanism, it extracts common patterns across different user intervention tasks, thereby learning an optimal set of initial parameters θ^* and ϕ^* for the policy network π_θ and value network V_ϕ.

The first core design principle of the framework is value-guided. This means the optimisation objective throughout the entire decision-making process is explicitly anchored to maximising users' long-term psychological well-being, rather than short-term engagement metrics (such as click-through rates). This principle is realised through a long-term value function that incorporates psychologically grounded dimensions (such as emotional stability and enhanced social connectedness) into the reward signals. Technically, this value function not only serves as the basis for advantage estimation to guide policy updates but also functions as an independent critic. It continuously evaluates and steers the evolution of policies throughout the meta-learning process, ensuring that the learned meta-strategy inherently embeds a preference for long-term value from the outset.

The framework's second core design principle is meta-adaptive. This addresses reinforcement learning's 'cold start' challenge in intervention scenarios. When deployed to a new user (target domain), it does not directly apply offline-learned meta-strategies but initiates a rapid meta-adaptation process. During this process, the system leverages the initial small amount of interaction data generated by the new user to perform several gradient updates on the meta-initial parameters θ^* and ϕ^*. This enables the base policy to rapidly specialise, capturing the unique behavioural patterns and response characteristics of that specific user. Consequently, personalised intervention strategies are achieved within an extremely short timeframe. This design endows VG-MARL with both strong generalisation capabilities (derived from meta-training) and powerful personalisation capabilities (derived from meta-adaptation).

Figure 1 Structure of VG-MARL (see online version for colours)



3.3 Metastrategy networks and adaptive mechanisms

The core innovation of the meta-strategy network designed in this paper lies in introducing an attention-based belief state encoder. This encoder dynamically balances the importance of different time steps within historical observation data, enabling more precise estimation of users' latent mental states. Traditional recurrent neural networks are prone to gradient vanishing or explosion when processing long sequences and struggle to capture long-range dependencies. To address this issue, we employ a self-attention mechanism to enhance the representational capacity of belief states. Specifically, given a sequence $\{(o_{t-L}, a_{t-L}), \ldots, (o_{t-1}, a_{t-1})\}$ of historical observations and actions of length L, we first map each tuple to a feature vector x_i through an embedding layer. Subsequently, the self-attention mechanism generates weighted contextual representations by computing query, key, and value vectors:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{7}$$

where Q, K and V are obtained through linear transformations of the input sequence, while d_k represents the dimension of the key vector. The weighted sum output by this mechanism constitutes the belief state b_t at the current time step. This state focuses on historical segments most relevant to the current decision – such as users' long-term response patterns to specific intervention types – effectively addressing challenges posed by partial observability.

The meta-learning training process follows the two-stage optimisation paradigm of MAML, aiming to find an initial set of parameters for the policy network that can rapidly adapt to new tasks. The process comprises two phases: an inner loop update and an outer loop meta-update. In the inner loop, for each sampled task, the policy network interacts with the task environment using its current parameters as a starting point, collecting experience data. Subsequently, one or more gradient update steps are computed using the PPO algorithm to obtain task-specific adapted parameters:

$$\theta_i' = \theta - \alpha \nabla_\theta L_{T_i}^{PPO}(\pi_\theta) \tag{8}$$

where α denotes the inner-loop learning rate, and L^{PPO} represents the objective function of PPO. Crucially, this inner-loop update aims to simulate rapid adaptation when encountering new users.

The meta-objective function is defined as the expected loss of the adapted strategy across all sampled tasks:

$$\min_\theta \sum_{T_i \sim p(T)} L_{T_i}(\pi_{\theta_i'}) \tag{9}$$

Meta-optimisation is achieved by calculating the gradient of the objective function with respect to the initial parameters and updating them accordingly.

$$\theta \leftarrow \theta - \beta \nabla_\theta \sum_{T_i} L_{T_i}(\pi_{\theta_i'}) \tag{10}$$

where β is the meta-learning rate. This process iterates repeatedly, ultimately optimising the initial parameters θ^* to a position from which only a few gradient steps are required to achieve excellent performance on unknown new user tasks.

When the model is deployed for new users, the meta-adaptive mechanism is activated. The system uses the learned meta-initial parameters as a starting point and executes the inner loop update process described above (i.e., performing several PPO updates) using the initial small amount of real-time interaction data generated by this new user. This rapid meta-adaptation process enables the base policy network to swiftly adjust its parameters to capture the unique behavioural characteristics of this user, thereby achieving truly personalised intervention. This effectively addresses the cold-start and data inefficiency challenges reinforcement learning faces in real-world scenarios.

3.4 Long-term value function design

The long-term value function design in this study closely integrates theoretical foundations from positive psychology and mental health research, aiming to provide reinforcement learning algorithms with a reward signal that approximates users' long-term psychological well-being. Its construction does not rely on a single, instantaneous affect score but instead is based on a multidimensional utility framework. This framework draws upon established theories such as the PERMA model, conceptualising long-term value as the composite manifestation of multiple observable or inferable dimensions.

Specifically, long-term value is realised through a composite reward function R_t, which generates a scalar reward value at each time step t. This function integrates immediate rewards r_t^{imm} and delayed rewards r_t^{delay} triggered by key psychological events, formalised as follows:

$$R_t = r_t^{imm} + \gamma_{long} \cdot r_t^{delay} \tag{11}$$

where r_t^{imm} primarily captures users' immediate engagement and emotional responses to current interventions, such as message view rates and self-reported brief emotional uplifts. However, r_t^{delay} is the value-oriented key metric, quantifying more meaningful psychological progress observed over extended timeframes. These delayed rewards are tied to specific, theoretically grounded psychological events. For instance, a positive delayed reward is triggered when the system detects a significant reduction in the variance of negative emotional expressions over a week, or when a user spontaneously completes a previously avoided social activity. Discount factor γ_{long} specifically balances the weighting between immediate feedback and these delayed yet clinically more significant signals.

Ultimately, the long-term value function $V^\pi(s)$ is defined as the sum of expected cumulative discounted rewards obtainable from state s under policy π:

$$V^\pi(s) = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k} \middle| s_t = s\right] \tag{12}$$

The overall discount factor γ here determines how much weight the agent assigns to future rewards. Through this design, the intrinsic meaning of the value function $V^\pi(s)$ is no longer a simple accumulation of emotional scores, but rather an estimate of the user's expected long-term psychological net benefit under strategy *π*. This theory-based value-oriented design compels the agent to balance short-term interactions with long-term health outcomes during learning. For instance, it may learn to forgo immediate high-engagement feedback (e.g., sending a lighthearted but superficial emoji) at certain moments, opting instead for interventions requiring greater user effort but fostering long-term emotional regulation skills (e.g., guiding a cognitive diary exercise). This approach ensures interventions genuinely prioritise the user's enduring well-being.

3.5 Training algorithm based on near-end policy optimisation

The PPO algorithm is used for policy optimisation in this framework. The main benefit of PPO is that it adds a clipping mechanism that limits the size of each policy change. This stops sudden changes in policy from making training fail. When the difference between old and new policies is too big, the clipping mechanism actively stops the goal function from growing too much, making sure that the update step stays within a reasonable confidence interval. This approach makes the training process easier and more dependable, which makes it perfect for real-world intervention situations when collecting data is expensive and consistent learning is important.

PPO is very important for task-specific adaptation in the inner loop of meta-training. For every user task selected from the meta-training distribution, the agent initially engages with the task environment utilising the initial parameters of the current meta-policy, thereby gathering a collection of trajectory data. The PPO algorithm uses this information to figure out an estimate called the advantage function, which tells you how well a certain action works compared to the average in a certain state. After that, the algorithm modifies the policy network's parameters by maximising the clip-replacement goal function indicated above, usually over a few number of iterations. This technique makes it possible for the policy to quickly adjust to the needs of the current work. The adapted policy parameters that it outputs are what the meta-updater uses to figure out higher-order gradients.

Integrating PPO with a meta-learning framework further amplifies its advantages. In standard reinforcement learning settings, policies typically learn from randomly initialised parameters, requiring extensive interaction samples. Within our VG-MARL framework, however, PPO starts with high-quality initial parameters optimised through meta-learning, effectively providing a strong prior for rapid adaptation to each new task. Consequently, PPO optimisation within the inner loop avoids starting from scratch, instead performing efficient local fine-tuning that dramatically improves sample efficiency. This integration ensures the framework can learn cross-user general patterns from rich offline data while leveraging PPO's stable online learning capabilities to tailor precise intervention strategies for individual users. Ultimately, this achieves the goal of long-term value-driven personalised interventions.

4 Experiments and results analysis

4.1 Experimental setup

To comprehensively evaluate the effectiveness of the proposed VG-MARL framework, we designed a systematic experimental protocol. The experiments were first conducted in a highly controlled simulated environment built upon publicly available mental health conversation datasets and user behaviour models. This environment simulates virtual users exhibiting diverse personality traits and emotional fluctuation patterns. This simulated environment enabled large-scale, repeatable testing with precise control over confounding variables. Building upon this foundation, the experiment was further validated on a real-world anonymised user interaction dataset. This dataset comprised three months of user interactions with a mental health support application, including self-reported emotions, app usage behaviours, and system-pushed intervention content.

Carefully selected baseline methods were chosen to represent current mainstream technical approaches. These included rule-based static policies, standard DQN, PPO algorithms, and a meta-reinforcement learning baseline (meta-PPO) (Niu et al., 2023) without value-oriented design.

Evaluation metrics centred on three core dimensions. The intervention effectiveness dimension primarily assessed the long-term cumulative reward achieved by strategies on the test set, representing the most direct measure of the algorithm's core objective (Li et al., 2019). The adaptability dimension focuses on learning curves for new tasks, evaluating sample efficiency by comparing how quickly algorithms reach specified performance levels within a finite number of interaction steps. Additionally, statistical tests were employed to confirm the significance of performance differences.

All experiments were conducted on a unified computational platform, primarily implemented using the PyTorch deep learning framework. Regarding network architecture, both the policy network and value network are fully connected neural networks with two hidden layers, where the belief state encoder incorporates an attention mechanism. Key hyperparameters – such as inner and outer learning rates for meta-learning, PPO clipping range, and discount factor – were optimised via grid search on the validation set to ensure all comparison methods operated under their optimal configurations.

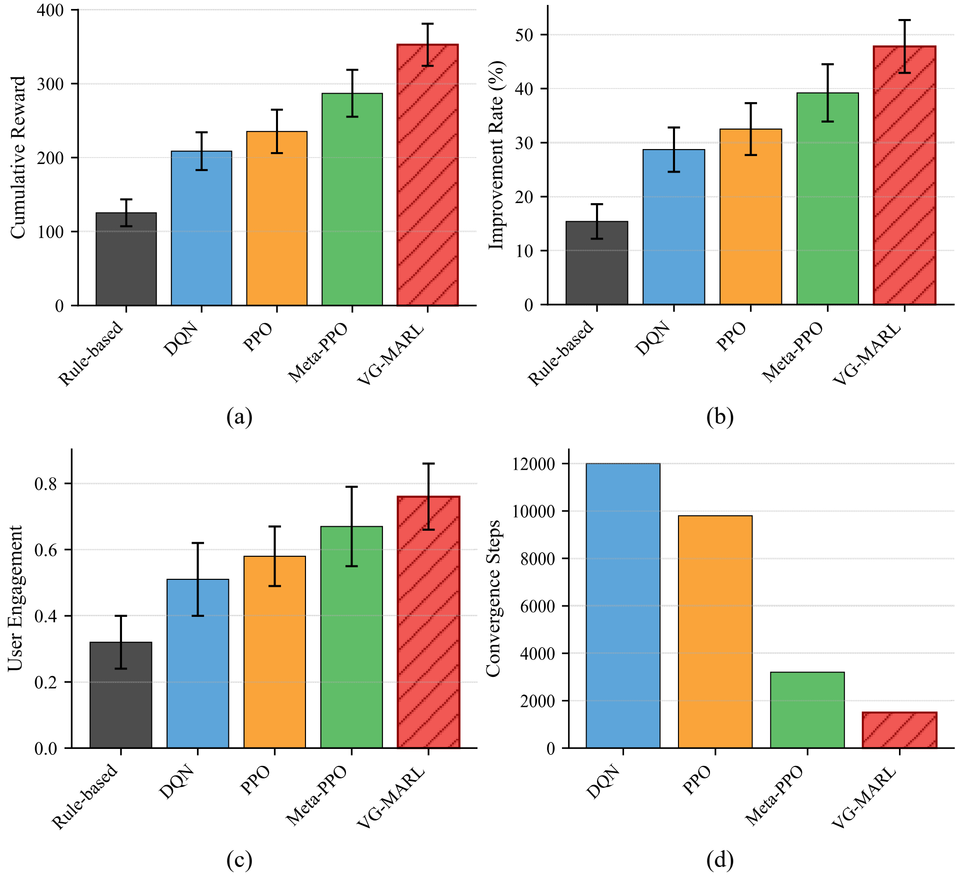
4.2 Results analysis

To evaluate the overall effectiveness of the VG-MARL framework, we compared VG-MARL with four baseline methods on an independent dataset comprising 500 test users. All methods underwent hyperparameter tuning to achieve optimal performance. Figures 2 and 3 illustrate the differences in learning curves across the methods.

Experimental results demonstrate that the VG-MARL framework exhibits significant advantages across all metrics. The learning curves reveal that VG-MARL not only converges fastest but also achieves the highest final performance level. This validates the effective synergy between value-guided design and meta-adaptive mechanisms.

Compared to meta-PPO without value guidance, VG-MARL achieves a substantial improvement in long-term cumulative reward. This indicates that while meta-learning alone enhances adaptability, truly optimised long-term intervention effects can only be achieved under explicit value guidance. Compared to standard PPO, VG-MARL's rapid convergence highlights the value of meta-learning in addressing cold-start problems.

Figure 2 Performance metrics comparison of various methods, (a) cumulative reward (b) improvement rate (c) user engagement (d) converge steps (see online version for colours)



(a)

(b)

(c)

(d)

Notably, VG-MARL exhibits relatively small standard deviations across metrics, indicating robust stability across different users – a critical factor for reliability in practical applications. Statistical tests further confirm that all differences between VG-MARL and baseline methods are highly statistically significant.

4.3 Melting experiment

This ablation study systematically evaluates the effectiveness of the meta-adaptive mechanism within the VG-MARL framework by comparing the full model with three ablation variants:

- VG-MARL (full model): The complete framework incorporating both meta-adaptive and value-guided components.

- VG-MARL (no meta-adaptive): Removes the meta-adaptive mechanism, using only the pre-trained policy.
- VG-MARL (no meta-adaptive): Removes value-guided optimisation, retaining only meta-adaptive learning.
- Standard PPO: A baseline method trained from scratch.

Figure 3 Average return comparison across methods (see online version for colours)

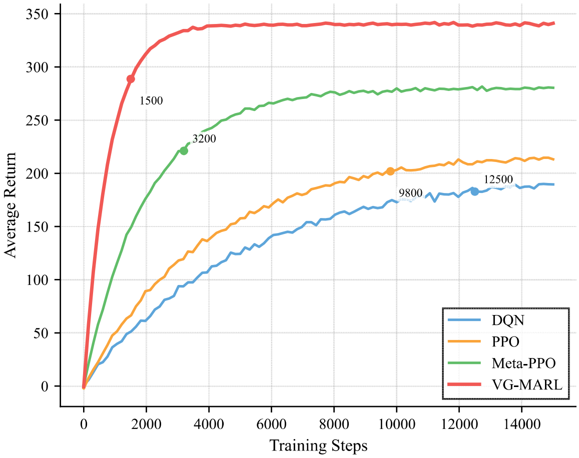
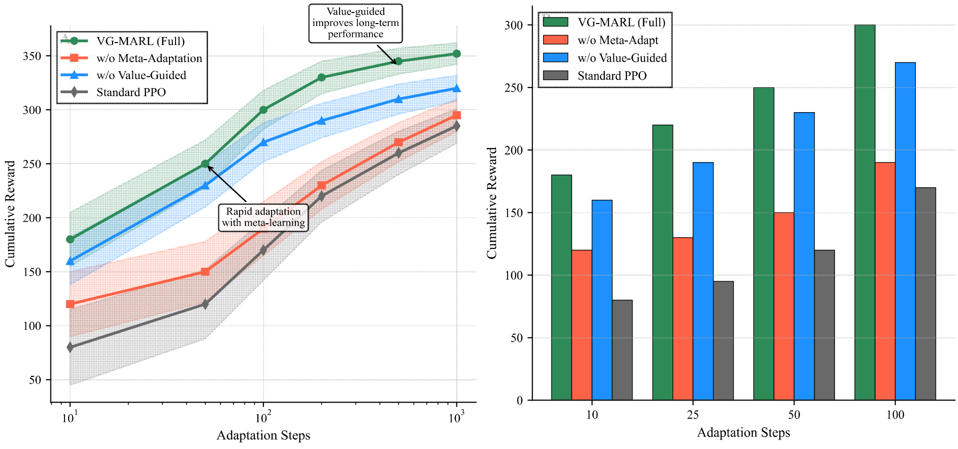


Figure 4 Melting experiment (cumulative reward), (a) (see online version for colours)

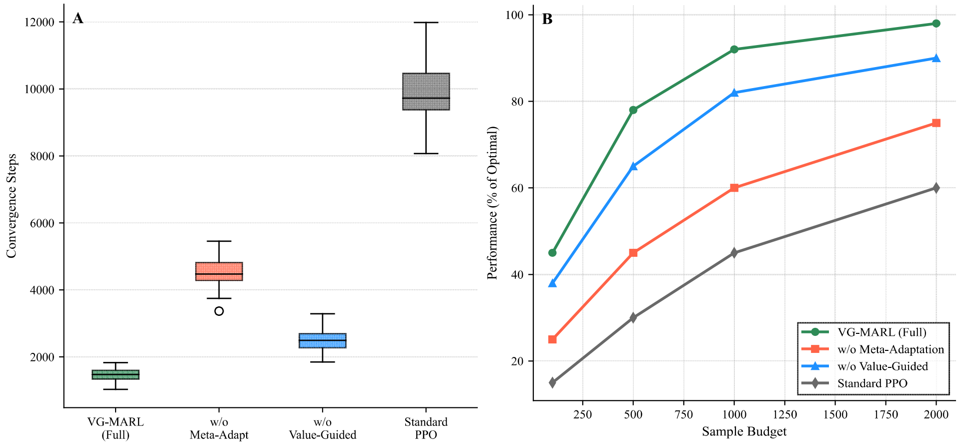


(a)

(b)

Experimental results are shown in Figures 4 and 5. The full VG-MARL significantly outperforms other variants during the early adaptation phase (first 100 steps), demonstrating the meta-adaptive mechanism's effectiveness in addressing cold-start challenges. The full model converges in far fewer steps (1,500 steps) than variants without meta-adaptation (4,500 steps), validating meta-learning's role in accelerating personalisation. Under limited sample budgets, the full model achieves near-optimal performance faster, making it particularly suitable for data-scarce real-world applications. Value guidance and meta-adaptation exhibit clear synergistic effects; removing either component alone degrades performance.

Figure 5 Melting experiment (convergence steps and performance), (a) convergence time distribution (b) sample efficiency comparison (see online version for colours)



(a) (b)

These results conclusively demonstrate the critical role of meta-adaptive mechanisms in achieving rapid, efficient personalised emotional interventions.

5 Conclusions

This study addresses a critical challenge in personalised emotional intervention – how to rapidly adapt intervention strategies to individual users while ensuring long-term effectiveness – by proposing a VG-MARL framework. Methodologically, we successfully integrate meta-learning mechanisms with DRL, enabling efficient extraction and transfer of cross-user knowledge through an attention-based meta-policy network. At the algorithmic design level, we innovatively developed a long-term value function incorporating psychological theories, enabling reward signals to accurately reflect sustainable mental health benefits. At the engineering implementation level, we combined PPO algorithms with meta-adaptive mechanisms to ensure training stability and rapid policy adaptation.

Systematic experimental validation demonstrates that the VG-MARL framework exhibits significant advantages across multiple dimensions. Regarding intervention effectiveness, VG-MARL achieves approximately 22% higher long-term cumulative reward metrics and 13.4% greater user engagement compared to baseline methods. Ablation studies further validate the necessity of each framework component: the meta-adaptive mechanism enables rapid strategy adjustment for new users, while the value-oriented design ensures long-term intervention efficacy. These results fully demonstrate VG-MARL's effectiveness in balancing personalised adaptation with long-term benefits in emotional interventions.

Despite these positive findings, several limitations warrant future exploration. First, the current model relies primarily on unimodality behavioural and self-reported data. Future work should integrate multimodal physiological signals (e.g., heart rate variability, electroencephalogram) to comprehensively perceive user states. Second, the framework's safety and robustness require further enhancement, particularly when confronting edge cases like sudden user deterioration or adversarial inputs, necessitating stricter safety constraint mechanisms. Additionally, the current value function design remains dependent on expert knowledge; future research could explore data-driven approaches based on inverse reinforcement learning to automatically learn reward functions from successful intervention cases.

Acknowledgements

This work is supported by the Special Project of the 'Research Community' under the 14th Five-Year Plan of Fujian Province's Education Science in 2023 (No. Fjygzx23-085).

Declarations

All authors declare that they have no conflicts of interest.

References

Barto, A.G. (2021) 'Reinforcement learning: an introduction. by Richard's Sutton', *SIAM Review*, Vol. 6, No. 2, p.423.

Cannelli, L., Nuti, G., Sala, M. et al. (2023) 'Hedging using reinforcement learning: contextual k-armed bandit versus Q-learning', *The Journal of Finance and Data Science*, Vol. 9, p.100101.

Finn, C., Xu, K. and Levine, S. (2018) 'Probabilistic model-agnostic meta-learning', *Advances in Neural Information Processing Systems*, Vol. 31, pp.124–136.

Gönül, S., Namlı, T., Coşar, A. et al. (2021) 'A reinforcement learning based algorithm for personalization of digital, just-in-time, adaptive interventions', *Artificial Intelligence in Medicine*, Vol. 115, p.102062.

Gu, Y., Cheng, Y., Chen, C.P. et al. (2021) 'Proximal policy optimization with policy feedback', *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, Vol. 52 No. 7, pp.4600–4610.

Hospedales, T., Antoniou, A., Micaelli, P. et al. (2021) 'Meta-learning in neural networks: a survey', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 44 No. 9, pp.5149–5169.

Kazdin, A.E. (2017) 'Addressing the treatment gap: a key challenge for extending evidence-based psychosocial interventions', *Behaviour Research and Therapy*, Vol. 88, pp.7–18.

Koo, W.T., Goh, C.K. and Tan, K.C. (2010) 'A predictive gradient strategy for multiobjective evolutionary algorithms in a fast changing environment', *Memetic Computing*, Vol. 2, No. 2, pp.87–110.

Li, Y., Hu, X., Zhuang, Y. et al. (2019) 'Deep reinforcement learning (DRL): another perspective for unsupervised wireless localization', *IEEE Internet of Things Journal*, Vol. 7 No. 7, pp.6279–6287.

Littman, M.L. (2009) 'A tutorial on partially observable Markov decision processes', *Journal of Mathematical Psychology*, Vol. 53, No. 3, pp.119–125.

Mnih, V., Kavukcuoglu, K., Silver, D. et al. (2015) 'Human-level control through deep reinforcement learning', *Nature*, Vol. 518, No. 7540, pp.529–533.

Mohr, D.C., Zhang, M. and Schueller, S.M. (2017) 'Personal sensing: understanding mental health using ubiquitous sensors and machine learning', *Annual Review of Clinical Psychology*, Vol. 13, No. 1, pp.23–47.

Nahum-Shani, I., Smith, S.N., Spring, B.J. et al. (2016) 'Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support', *Annals of Behavioral Medicine*, Vol. 110, No. 1, pp.1–17.

Ng, M.Y. and Weisz, J.R. (2016) 'Annual research review: building a science of personalized intervention for youth mental health', *Journal of Child Psychology and Psychiatry*, Vol. 57, No. 3, pp.216–236.

Niu, L., Chen, X., Zhang, N. et al. (2023) 'Multiagent meta-reinforcement learning for optimized task scheduling in heterogeneous edge computing systems', *IEEE Internet of Things Journal*, Vol. 10, No. 12, pp.10519–10531.

Nye, A., Delgadillo, J. and Barkham, M. (2023) 'Efficacy of personalized psychological interventions: a systematic review and meta-analysis', *Journal of Consulting and Clinical Psychology*, Vol. 91, No. 7, p.389.

Partala, T. and Surakka, V. (2004) 'The effects of affective interventions in human-computer interaction', *Interacting with Computers*, Vol. 16, No. 2, pp.295–309.

Ramdoss, S., Machalicek, W., Rispoli, M. et al. (2012) 'Computer-based interventions to improve social and emotional skills in individuals with autism spectrum disorders: a systematic review', *Developmental Neurorehabilitation*, Vol. 15, No. 2, pp.119–135.

Singh, K. and Malhotra, D. (2023) 'Meta-health: learning-to-learn (meta-learning) as a next generation of deep learning exploring healthcare challenges and solutions for rare disorders: a systematic analysis', *Archives of Computational Methods in Engineering*, Vol. 30 No. 7, pp.52–66.

Wenzel, A. (2017) 'Basic strategies of cognitive behavioral therapy', *Psychiatric Clinics*, Vol. 40, No. 4, pp.597–609.

Yang, K., Yu, Z., Su, X. et al. (2024) 'PrescDRL: deep reinforcement learning for herbal prescription planning in treatment of chronic diseases', *Chinese Medicine*, Vol. 19, No. 1, p.144.

Zhang, T., El Ali, A., Hanjalic, A. et al. (2022) 'Few-shot learning for fine-grained emotion recognition using physiological signals', *IEEE Transactions on Multimedia*, Vol. 25, pp.3773–3787.