Data-driven Personalized Family Music Therapy: a Reinforcement Learning Recommendation Model Based on Family Member Behavior and Feedback

Jinhui Zhang[1,2], Mohd Syafiq Md Salleh[1], Peifeng Yang[2], Ying Ding[2], Xin Wang[2], Liangliang Dong[2], Hongyan Lv[2], Na Lv[2], Jianxin Wu[2], Jinglei Zhao[2]

[1]Faculty of Education and Liberal Studies, City University Malaysia, Kuala Lumpur, Malaysia
[2]DongYing Vocational College, DongYing, China

Corresponding author: Jinglei Zhao (e-mail: edwinhere@163.com).

ABSTRACT This study proposes a data-driven personalized family music therapy recommendation framework to address the limitations of conventional recommendation systems in modeling emotional interactions among multiple family members. The framework employs a reinforcement learning approach tailored to family environments, where emotional states evolve through continuous interpersonal interactions and feedback. A relationship-aware family state representation is developed by integrating behavioral observations and physiological responses, enabling the modeling of both individual emotional states and emotional contagion among family members. To support adaptive intervention, the action space incorporates music selection, spatial delivery strategies, and intentional silence decisions, allowing the system to balance active intervention with natural emotional self-regulation. In addition, a multi-channel reward mechanism is designed by combining physiological recovery, behavioral relaxation, and environmental stability indicators, while a temporal credit assignment strategy addresses delayed therapeutic effects. Experimental evaluation on a multimodal dataset covering four representative family stress scenarios demonstrates that the proposed framework achieves 89.7% accuracy in family emotion representation, 86.3% consistency with expert intervention judgments, and an 82.3% positive emotion conversion rate after continuous intervention. These findings suggest that reinforcement learning can provide an effective foundation for transforming music recommendation systems from individual entertainment tools into proactive family emotional support systems.

INDEX TERMS Music Therapy Recommendation; Multi-member Emotion Coupling; Reinforcement Learning Modeling; Family State Representation; Physiological Feedback Reward

I. INTRODUCTION

THE interdisciplinary research of music recommendation systems and affective computing has developed a technical path that drives playback strategies based on user emotional states. However, existing work has always taken individuals as the basic unit of modeling, without addressing the dynamic coupling relationship representation between the behavior and emotional feedback of multiple members in family scenarios [1-2]. This deficiency means that music recommendation can only passively respond to single playback commands when facing family co-existence situations, and cannot actively perceive and regulate the group emotional states such as tension, fatigue or alienation that are contagious among family members [3-4]. The importance of solving this problem is reflected in two aspects: at the theoretical level, it promotes the extension of reinforcement learning recommendation models from single-agent Markov decision processes to multi-agent relationship perception decision frameworks, and is a necessary extension of the social situation understanding ability of sequential recommendation agents [5]. At the application level, family mental health intervention has long relied on professional home visits or remote consultations, which severely limits the coverage and real-time performance. A music recommendation model with proactive emotion tuning capabilities will provide a new technical carrier for universal family emotional health management [6-7]. The transfer of reinforcement learning recommendation models to family music therapy scenarios may suggest that three intertwined challenges constitute the most fundamental obstacles. Moreover, the multi-subject coupling problem in family state representation could indicate that this barrier remains the most critical theoretical

concern. Furthermore, the significant evidence may suggest that family emotions are not a simple sum of individual emotions, but an emergent phenomenon that is continuously and dynamically generated through interactive behaviors such as dialogue, eye contact, and spatial distance. In light of these findings, the key challenge of encoding the independent emotional state of each member alongside the intensity of emotional contagion in a unified mathematical structure could demonstrate that this theoretical obstacle has not yet been overcome in this field [8-9]. Findings show obstacle leads to restraint dilemma in action design. However, the significant results may suggest that the action space of traditional recommendation models could indicate an implicit tendency toward excess. Additionally, the key evidence appears to demonstrate that the core wisdom of family music therapy is precisely the timely silence, because the data could suggest that playing music at an inappropriate time may interrupt the

self-regulating rhythm of family emotions [10-11]. How to enable an agent whose basic function is to output actions to learn "active silence" is not merely an engineering problem of expanding the action space: silence itself has no observable immediate benefits, and policy networks have difficulty distinguishing between "natural recovery from restraint" and "opportunity loss from missed intervention" from single-step rewards. This makes the representation of the silence policy and the design of the optimization objective both face fundamental difficulties. The delayed and complex nature of healing signals further deepens the complexity of modeling. The physiological and psychological effects of music usually only gradually appear tens of minutes after playback and are scattered across heterogeneous modalities such as heart rate, respiration, facial expression, and sound pressure. These signals are often contradictory to each other, and existing reward function design methods cannot complete time attribution and multimodal weight allocation [12-13]. The three challenges are not isolated from each other: state representation determines whether the model can perceive the involvement of family emotions, action restraint determines whether the model interferes with the family's self-healing process, and reward signals determine the optimization direction of the model. They collectively reveal a fundamental fact: the three basic components of state, action, and reward in existing reinforcement learning recommendation models have not been defined for family emotional healing tasks. Researchers demonstrate that multiple technical approaches could address these challenges. However, the significant findings may suggest that traditional music recommendation based on collaborative filtering and content matching relies entirely on explicit clicks and favorites, resulting in a fundamental misalignment between its optimization goals and the physiological and psychological indicators required for emotional healing. Moreover, the evidence may indicate that this method fails to perceive or regulate the emotional state of the family [14-15]. In light of these results, emotion-based music recommendation could suggest that incorporating individual facial expressions or physiological signals into the recommendation decision presents important limitations, given that the modeling object remains a single user, failing to address the coupling relationship of multiple emotional interactions. Approach fails instant emotion matching for long-term healing tasks [16-17]. Furthermore, the significant results may indicate that multi-agent reinforcement learning has shown some potential in multi-person scenario modeling, but the key evidence could suggest that its reward function design still revolves around the coordination and competition of individual behaviors. Given that the findings demonstrate that a reward definition oriented towards group emotional stability remains absent, the results might suggest that this approach fails to address the issue of safe exploration in highly sensitive application contexts [18-19]. Notwithstanding these results, the evidence could indicate that existing methods, limited by the single-user modeling paradigm and the assumption of instant feedback, appear to fail in solving the core problem. Methods fail dynamic coupling representation across family members. This is precisely the technical hurdle that

this paper must overcome to shift music recommendation from individual entertainment to family healing. This paper aims to address the problem of music recommendation under the coupled emotional state of multiple family members. To this end, a reinforcement learning recommendation model for the family context is proposed, with its technical approach consisting of three progressively designed core components: At the state representation level, explicit interaction behaviors and implicit physiological emotional feedback are fused into a relationship-aware family state tensor, and a graph neural network is introduced to assign emotional contagion weights to each member relationship edge, enabling the model to represent the dynamics of coupled emotions. Furthermore, this paper extends the song list action space to an intervention decision tuple containing silent options and spatial placement so that the model can proactively control its actions. In addition, this paper constructs a reward function as a long-range composite signal of heart rate variability stability and family sound pressure level decrease and proposes a qualification trace mechanism to ascribe the healing effect to the correct time point. On a dataset constructed on four typical family stress scenarios, the proposed model reaches twin goals of representing the coupled emotional state of the family and learning corresponding silent actions. The experimental results show an obvious improvement of positive emotion conversion rate after 30 minutes of continuous intervention on a real dataset. These results indicate that by reconstructing the three parts of state, action and reward, this paper endow the reinforcement learning recommendation model with the ability to represent the dynamic relationships of emotions among multiple family members. This development paves the technological path for music recommendation systems to shift from individual entertainment tools to a medium for

proactive intervention in family emotions.

II. MODEL CONSTRUCTION FOR THE HOME ENVIRONMENT

A. PROBLEM FORMALIZATION

The sequential intervention of family music therapy is formalized as a Markov decision process of a six-tuple $M = (S, A, P, R, \gamma, \lambda)$. The state space S, action space A, state transition function P, reward function R, discount factor γ and qualification trace parameter λ jointly define the solution boundary of the decision problem. At each discrete time t, the agent observes the family state $S_t \in S$ and outputs the intervention action $a_t \in A$ according to the policy $\pi(a_t \mid S_t)$. The environment then returns a scalar reward r_{t+1} and pushes the system to the successor state S_{t+1}. The optimization objective is to maximize the expected cumulative reward $\mathbb{E}_\pi \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$. The state transition function P is not explicitly parameterized. This paper adopts a model-free reinforcement learning setting. The agent directly optimizes the policy by relying only on the interaction sampling with the real family environment. Implicitly, the transition dynamics are absorbed by the measured behavior and physiological data stream [20]. Reward function R is built as a composite signal of multimodal physiological, behavioral and environment channel, and adds a qualification trace mechanism to solve the long-term delay of therapeutic effect, and guide the strategy to connect music intervention and emotional enhancement in tens of minutes across a time interval. This optimization target is controlled by two structured constraints: state representation should contain the information of intensity of emotional contagion from members, and action decision making should contain the information of silence to protect the family's self-healing rhythm of emotion. These two constraints indicate the two basic principles of constructing state space and action space separately, and the construction is layer by layer to satisfy them one by one. The state of the family, consisting of N members, is folded into a third-order tensor $S_t \in \mathbb{R}^{N \times N \times D}$ with member state embedding dimension D at time t. The diagonal element $S_t[i, i, :]$ describes the independent emotional representation of the i-th family member generated by gating the fusion of implicit physiological signals and explicit behavioral characteristics of the member reflecting the state emotional representation of the i-th family member. The off-diagonal element $S_t[i, j, :]$ characterizes the interaction representation of member i with member j, reflecting the emotional contagion relationship between members. A flattened state containing only diagonal elements and no off-diagonal elements cannot distinguish between the fundamentally different family states of "two members are each tense" and "one member's tension is contagious to another member through an argument." Therefore, the state space must be constructed by simultaneously filling both types of elements. The construction of the off-diagonal elements is rooted in the propagation of graph neural networks on the family relationship graph. The family is abstracted as a fully connected directed graph, with each member corresponding to a node. The directed edge between member i and member j carries the emotional contagion weight α_{ij}^t, as shown in Fig. 1. Note: Nodes represent family members, and directed edges represent the direction of emotional contagion. There are two directed edges with opposite directions between each pair of members to reflect the asymmetry of contagion. The weights α_{ij}^t of each edge are dynamically calculated by the model at each time step and are not statically labeled in the graph. To update the infection weights of each edge at each time step, the graph neural network layer receives the node hidden states h_i^{t-1}, h_j^{t-1} and edge history representation u_{ij}^{t-1} from the previous time step at each time step. The concatenated vector is then linearly transformed by the parameter W_α and the bias b_α, and activated by the sigmoid function to generate the infection weights at the current time step.

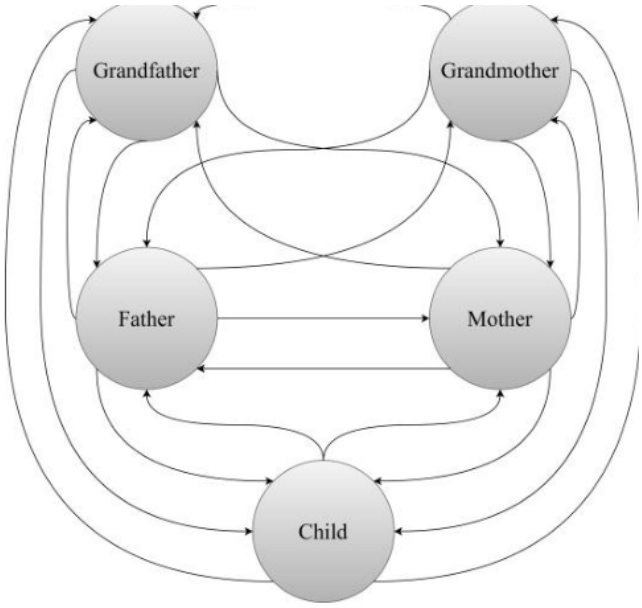

FIGURE 1. Emotional Contagion Diagram among Family Members

$$\alpha_{ij}^t = \sigma\left(W_\alpha^\top \left[h_i^{t-1} \| h_j^{t-1} \| u_{ij}^{t-1}\right] + b_\alpha\right) \tag{1}$$

In the formula, the contagion weight α_{ij}^t captures the emotional influence exerted by member i on member j at time t, and its value is updated time-by-time with the dynamic evolution of interactions and physiological states among members. This weight is further multiplied by the learnable relation basis vector $v_{ij} \in \mathbb{R}^D$ to form the interaction representation $S_t[i, j, :] = \alpha_{ij}^t v_{ij}$. The basis vector is constrained by the unit norm $\|v_{ij}\|_2 = 1$, making the magnitude of the interaction representation exactly equal to the contagion strength $\|S_t[i, j, :]\|_2 = \alpha_{ij}^t$, and the direction is determined by the stable relation semantics between member pairs encoded by the relation basis vector. This decoupled parameterization decouples the dynamic adjustment of the contagion strength from the long-term learning of relation

semantics into two independent optimization channels. The diagonal independent emotional representation and the off-diagonal interaction representation together constitute the complete family state tensor, thus directly realizing the multi-agent coupling constraint in mathematical structure. The state tensor provides the agent with the basis for perceiving the dynamics of family emotions, and the intervention action a_t output by the agent is designed as a hierarchical tuple (ξ_t, m_t, p_t), integrating discrete silence decision-making, music selection, and continuous spatial projection parameters. The silence variable $\xi_t \in \{0, 1\}$ is a structural switch in the action space: when $\xi_t = 0$, the music selection variable m_t is forced to a null value, the spatial projection intensity vector p_t is set to zero, and music playback is completely suppressed; when $\xi_t = 1$, m_t is selected from a music library containing K candidate music tracks, and each component $p_t^{(i)}$ of $p_t \in [0, 1]^N$ represents the sound pressure level proportion facing member i, and is subject to the normalization constraint $\sum_{i=1}^{N} p_t^{(i)} = 1$. The structural significance of distinguishing between the silence and non-silence modes lies in the fact that family emotions have a self-regulating rhythm, and applying music intervention at an inappropriate time may interrupt this spontaneous recovery process. Therefore, it is equally important for the agent to learn when not to intervene and when to intervene. The silent variable directly encodes this "active silence" ability as a legitimate option in the action space, enabling policy learning to inherently balance output and restraint, thereby satisfying the restraint constraint of the action space [21]. After an action is performed, the environment returns a reward signal to guide strategy optimization. However, the reward signal in the home music therapy scenario has two tricky properties. First, there is the multimodal contradiction. The therapeutic effect is distributed in heterogeneous modalities such as heart rate, breathing, and sound pressure. The feedback directions of each channel may conflict with each other at the same time [22]. Second, there is the delayed arrival. The physiological and

psychological effects of music usually appear gradually after tens of minutes of playback. The instantaneous reward cannot be directly attributed to the contribution of the historical action [23]. In order to deal with the multimodal contradiction, the instantaneous reward r_{t+1} is a weighted fusion of the physiological channel r_{t+1}^{phy}, the behavioral channel r_{t+1}^{beh}, and the environmental channel r_{t+1}^{env} with adjustable weights.

$$r_{t+1} = w_{\text{phy}} r_{t+1}^{\text{phy}} + w_{\text{beh}} r_{t+1}^{\text{beh}} + w_{\text{env}} r_{t+1}^{\text{env}} \tag{2}$$

The weights in the formula satisfy the normalization constraint $w_{\text{phy}} + w_{\text{beh}} + w_{\text{env}} = 1$. The physiological channel r_{t+1}^{phy} aggregates the individual rate of change of the Root Mean Square of the Successive Differences (RMSSD) of each member's heart rate variability index. This channel provides positive incentives when the RMSSD of all members improves synchronously. The behavioral channel r_{t+1}^{beh} uses the sharp drop in the overall sound pressure level of the family as a proxy signal for the group's emotional relief. The environmental channel r_{t+1}^{env} penalizes excessively high volume output and frequent track switching to suppress redundant interventions. The three channels jointly characterize the healing progress from different modalities, and the weight fusion mechanism allows the reward function to make trade-offs based on adjustable weights when there are modal conflicts. To deal with the delayed arrival attribute, a qualification trace vector e_t is introduced to complete the time backtracking of credit allocation. The qualification trace dimension is the same as the policy parameter θ, and its update uses the product $\gamma\lambda$ of the discount factor γ and the qualification trace parameter λ as the decay rate, recursively inheriting the historical trace and superimposing the logarithmic gradient of the current policy.

$$e_t = \gamma\lambda e_{t-1} + \nabla_\theta \log \pi(a_t \mid S_t) \tag{3}$$

The eligibility trace tracks the contribution of each historical action in an exponentially decaying manner, with the decay rate determined by $\gamma\lambda$. Credit allocation also requires quantifying the prediction bias at each step; for this purpose, a parameterized state value function V_w is introduced, with parameters w, which estimates the expected return using the state tensor as input and output. At time t, the value function estimates the current state S_t as $V_w(S_t)$ and the subsequent state S_{t+1} as $V_w(S_{t+1})$, which, together with the immediate reward r_{t+1}, constitute the temporal difference error.

$$\delta_t = r_{t+1} + \gamma V_w(S_{t+1}) - V_w(S_t) \tag{4}$$

This error could indicate that the deviation between the current value prediction and the sampled reward may suggest meaningful variation in the learning signal. Moreover, the significant update direction of the policy parameter θ appears to emerge from multiplying the temporal difference error by the eligibility trace, adjusted along $\theta \leftarrow \theta + \eta\delta_t e_t$, where η is the learning rate. Furthermore, the eligibility trace may demonstrate that credit allocation across the historical action sequence follows an

exponentially decaying distribution along the time axis. In light of these findings, the results could suggest that physiological improvements appearing tens of minutes after playback might indicate correct attribution to the music intervention action that triggered the improvement, thereby solving the temporal credit allocation problem of delayed rewards [24-25]. Eligibility trace distributes healing signal across historical actions, solving delayed reward attribution. Progressively Integrated Solution: A relation-aware state tensor, a hierarchical action space with silent options, and a qualification-track-driven multi-channel composite reward. State encoding gives a perception basis coupled with emotional dynamics for intervention decisions. A hierarchical action space endows the strategy with proactive optimization abilities. A qualification-track reward structure gives the correct gradient direction for long-term therapeutic effects. The three parts complete the formal process of the paradigm shift

from single-user instant recommendation logic to family-based proactive emotional healing logic.

B. CONSTRUCTION OF MULTI-MEMBER FAMILY STATE REPRESENTATION

Construction of the family state tensor $S_t \in \mathbb{R}^{N\times N\times D}$. This paper obtains and encodes two kinds of types of raw data streams. The first type of data stream captures the explicit interactive behaviors of the members. They are synchronously collected by a microphone array and a camera deployed in the family environment. The microphone array records the time series of sound pressure levels from each sound source direction at a sampling rate of 16 kHz. The mixed audio is decomposed into independent speech streams for each member using a sound source separation algorithm. Then, the Mel-frequency cepstral coefficients and fundamental frequency perturbation parameters are extracted using a 2-second sliding window to form the acoustic behavior vector $b_i^{\mathrm{aud},t} \in \mathbb{R}^{d_a}$ for member i at time t. The camera acquires panoramic video at a rate of 30 frames per second. The head orientation angle, torso tilt angle, and hand movement speed of each member are extracted using a skeletal keypoint detection model. These three are concatenated to form the visual behavior vector $b_i^{\mathrm{vis},t} \in \mathbb{R}^{d_v}$. Acoustic and visual behavior vectors characterize the explicit interaction features of members from different modalities. After being mapped to a unified dimension d_b via a fully connected layer, they are added to obtain the explicit behavior representation $b_i^t \in \mathbb{R}^{d_b}$ of member i. The second type of data stream is the implicit physiological and emotional feedback of members, acquired by a wrist-worn photoplethysmography sensor worn by each member at a sampling rate of 64 Hz. The raw photoplethysmography pulse wave signal is band-pass filtered to remove respiratory drift and high-frequency noise. The beat interval is calculated using a 60-second sliding window, and then the time-domain index RMSSD of heart rate variability, the frequency-domain index LF/HF ratio, and the amplitude of respiratory sinus arrhythmia are extracted. These three are concatenated together to obtain the physiological and emotional vector $p_i^t \in \mathbb{R}^{d_p}$ of member i at time t. Since the two streams of data are acquired synchronously, the behavioral representation and physiological vector are strictly aligned in terms of

timestamps at each instant with regard to a common observational basis at time t. Even though explicit behavioral representations and implicit physiological emotion vectors are time-stamp aligned on the time axis, their rhythm of change is quite different. Behavioral signals change abruptly in response to interactive events–for example, a sudden verbal conflict can be audible in the acoustic behavioral vector on the millisecond timescale. In contrast, the physiology is governed by the slower regulation of the autonomic nervous system, and heart rate variability can change tens of seconds later [26-27]. This misalignment in response rhythms means that directly concatenating the b_i^t and p_i^t cannot reflect the dynamic coupling relationship between the two. At the fusion time, an adaptive trade-off needs to be made based on the instantaneous confidence of the two types of signals. The gated fusion unit is introduced for this purpose. It receives the behavioral representation b_i^t and the physiological vector p_i^t, calculates the cross-modal similarity through the parameter matrix $W_g \in \mathbb{R}^{d_b\times d_p}$, and then generates the gated vector g_i^t using the sigmoid function.

$$g_i^t = \sigma\left(W_g\left[b_i^t \| p_i^t\right] + b_g\right) \tag{5}$$

Each component of the gating vector takes a value between 0 and 1, and its magnitude directly reflects the correlation between behavioral and physiological signals in the corresponding feature dimension. The gating vector is then used to perform element-wise weighting on the linear transformation results of the two types of representations to generate member-independent emotional representations.

$$S_t[i,i,:] = g_i^t \odot U_b b_i^t + (1-g_i^t)\odot U_p p_i^t \tag{6}$$

In the formula, $U_b, U_p \in \mathbb{R}^{D\times d_b}$ is a learnable up-dimensional matrix that maps behavior and physiological representations from their respective feature spaces to the embedding dimension D of the tensor. $\odot$ represents element-wise multiplication, and b_g is the bias vector. The gating mechanism plays a key role here: when a certain dimension of the behavior signal is highly correlated with the corresponding dimension of the physiological signal, the gating component approaches 1, and the fusion result tends to trust the instantaneous and observable behavior representation; when the correlation between the two is weak, the gating component approaches 0, and the fusion result tends to trust the physiological representation that reflects the true state of the autonomic nervous system but is lagging behind [28]. This adaptive mechanism enables the generation of diagonal elements to dynamically balance the confidence of the two heterogeneous signals at the fusion time, avoiding rashly favoring one side when the behavior and physiological signals have not yet established a reliable association. The diagonal element $S_t[i,i,:]$ only represents the emotional state of each member. The emotional contagion and dynamic relationships between members need to be captured through off-diagonal elements. The construction of the off-diagonal element $S_t[i,i,:]$ relies on the propagation process of the graph

neural network on the family relationship graph. Its input is the hidden state of the diagonal element after being mapped by the node encoder. The hidden state h_i^t of node v_i is obtained by mapping the independent emotional representation of the node through the node encoder ϕ_{node}, that is, $h_i^t = \phi_{\mathrm{node}}(S_t[i,i,:])$. The node encoder consists of two fully connected layers and a LayerNorm normalization layer, which compresses the gated and fused independent emotional representations into a latent space representation suitable for graph propagation. In the dynamic evolution of emotional contagion, the interaction strength at the current moment is not only determined by the current observation.

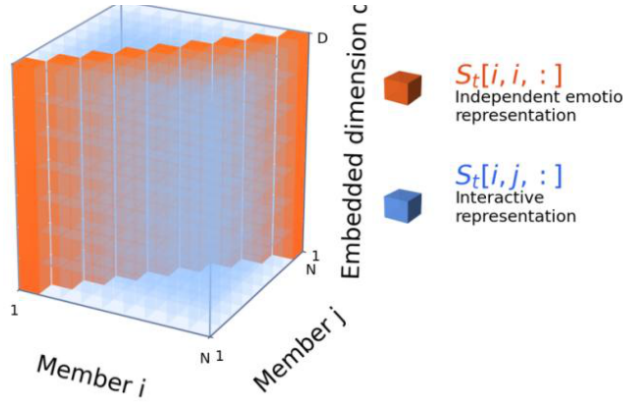

FIGURE 2. Structural decomposition of the family emotional coupling state tensor

The relationship patterns accumulated between members in the medium and long term also constitute an important prior. This prior is maintained through the edge history representation u_{ij}^{t-1}. u_{ij}^{t-1} is the exponential moving average of the interaction representations of member i with member j in the historical time steps, with a decay coefficient $\beta \in (0,1)$. The recursive formula is $u_{ij}^{t-1} = \beta u_{ij}^{t-2} + (1-\beta) S_{t-1}[i,j,:]$. The interaction influence at earlier times decays exponentially, and the decay rate is controlled by β. In each time step, the graph neural network layer concatenates the hidden state h_i^{t-1} of the sending node, the hidden state h_j^{t-1} of the receiving node, and the edge historical representation u_{ij}^{t-1}, and then performs a linear mapping and sigmoid activation to output the current sentiment contagion weight α_{ij}^t. This weight is then multiplied by the relation basis vector $v_{ij} \in \mathbb{R}^D$ to form the interaction representation $S_t[i,j,:] = \alpha_{ij}^t v_{ij}$. The relation basis vector consists of trainable parameters independently assigned to each pair of ordered member pairs (i,j). After the parameters are updated, explicit normalization is performed to maintain the unit norm. From the extraction of hidden states to the recursion of historical edges, and then to the calculation of contagion weights and the synthesis of interaction representations, this complete computational path injects the contagion intensity obtained from the propagation of the graph neural network into the off-diagonal elements, so that the interaction representation between any pair of members in the state tensor simultaneously carries the independent emotional signal from gating fusion and the coupled contagion signal from graph propagation, thus forming a complete tensor structure with both individual specificity and relational coupling, as shown in Fig2.

Fig2 shows the three-dimensional decomposition of this tensor. This architecture allows two of these capabilities that are impossible in single-user models. First, because each parameterization of the diagonal and off-diagonal elements is separately parameterized, the model can return two different types of answers to two different queries in the same forward propagation call: querying $S_t[i,i,:]$ yields "What is member is mood now?", and querying $S_t[i,i,:]$ to answer "How strong is the emotional influence of member i on member j?". The orthogonality of their indices in the tensor space avoids the confusion of conflating one's own state with interpersonal influence, as seen in single-user models. Second, the modulus of the off-diagonal element $S_t[i,i,:]$ is exactly equal to the emotional contagion weight α_{ij}^t, while v_{ij} is constrained by the unit norm. This design ensures that the strength and direction of the interaction representation correspond to the quantized value of the contagion weight and the semantics of the stable relationship between member pairs, respectively, facilitating independent gating and visual interpretation of the contagion strength in subsequent graph propagation. The dashed slices in the embedding dimension correspond to the information aggregation performed in parallel on multiple feature channels by the graph neural network. Each slice $S_t[:,:,d]$ is essentially an $N \times N$ membership matrix, and the diagonal and off-diagonal fillings are completed in the two regions of this matrix respectively. After filling the diagonal and off-diagonal lines, each slice $S_t[:,:,d]$ of the state tensor S_t corresponds to the d-th channel of the embedding dimension. The diagonal lines carry the components of the independent emotional representations of members in channel d, while the off-diagonal lines carry the components of the interaction representations between members in that channel. This tensor as a whole serves as the state input to the Markov decision process and is passed to the policy network and the value network. The third-order structure of the tensor enables the graph neural network to simultaneously aggregate information along the member dimension and the embedding dimension during forward propagation. The update of each interaction edge e_{ij} depends not only on the current representations of the sending and receiving nodes but also on the medium- to long-term trend of emotional contagion between members, which is preserved through the edge history representation u_{ij}^{t-1}. This allows for the simultaneous capture of short-term interaction events and long-term relationship patterns in a single forward computation. This state representation construction process unifies multi-source heterogeneous data into a structured tensor form, providing a complete relationship-aware state input for subsequent action space decision-making and qualification trace reward allocation.

C. INTERVENTION SPACE DESIGN FOR EMOTION REGULATION

The intervention action a_t has been formalized as a hierarchical tuple (ξ_t, m_t, p_t) in the problem formalization. This section describes the generation mechanism, constraints, and learning logic of the silent strategy in training for the three decision dimensions within this tuple. The three dimensions are not generated concurrently, but rather follow a hierarchical

dependency: first deciding whether to intervene, then selecting the intervention content and delivery method. This hierarchical structure is directly embedded in the output architecture of the policy network. The policy network re-

ceives the state tensor S_t as input. After three layers of graph convolution, it aggregates information along the member and embedding dimensions on the family relationship graph, outputting a full-graph pooled vector $z_t \in \mathbb{R}^{D_z}$. This pooled vector encodes a compressed representation of the family's overall emotional coupling state at the current moment, and is then branched and fed into the three decision heads. The silent decision head maps the pooling vector z_t to a scalar logit via a fully connected layer, and the silent variable $\xi_t \in \{0, 1\}$ is sampled and output by Gumbel-Softmax. The reparameterization property of Gumbel-Softmax allows the gradient of discrete sampling to be backpropagated to the policy network, thereby training the silent policy end-to-end. When $\xi_t = 0$, the subsequent two decision heads are bypassed, the track selection variable m_t is assigned a null value, the spatial projection intensity vector p_t is set to zero, and the complete output of the intervention action tuple is $(0, \varnothing, \mathbf{0})$. No music playback behavior occurs at this time step. When $\xi_t = 1$, the track selection head and the spatial projection head are activated sequentially, generating m_t and p_t respectively. The track selection head interacts with the pooling vector z_t and the embedding vector $c_k \in \mathbb{R}^{D_c}$ of each candidate track in the music library through attention, calculating the matching score for each track in the current home environment. The music library embedding matrix $C \in \mathbb{R}^{K \times D_c}$ is initialized by a pre-trained music emotion regression model before model training begins. This regression model predicts pleasure and arousal using audio features as input, and its penultimate hidden state is projected as the track embedding after principal component analysis. The matching score is normalized by softmax to form a K-dimensional probability vector, and m_t is sampled for category based on this probability. This attention mechanism ensures that track selection depends not only on the emotional attributes of the track itself, but also on the degree of interaction and matching between the global features of the current home environment and the track embedding. The spatial delivery head performs cross-attention calculations on the pooling vector z_t and the independent emotional representation $S_t[i, i, :]$ of each member, outputting N original weights. After softmax normalization, these weights form a spatial delivery intensity vector p_t, whose components $p_t^{(i)}$ satisfy $\sum_{i=1}^{N} p_t^{(i)} = 1$. Cross-attention uses member representations as queries and global pooling vectors as keys and values, so that the proportion of sound pressure level assigned to each member depends on the correlation strength between the member's current emotional state and the overall family state. Members whose emotional state is highly coupled with the overall family tension receive higher delivery weights, thus enabling music intervention to accurately target the members who most need emotional regulation [29-30]. Training a silent strategy faces a unique challenge: the correctness of a silent action cannot be verified immediately

within a single time step. Only after observing the spontaneous improvement of family emotions after several steps of non-intervention can the rationality of the silent decision be confirmed [31]. This delayed verification property makes the reward signal for the silent action rely on the eligibility trace mechanism for time backtracking. During training, when the agent performs a silent action and the family emotion index shows spontaneous improvement within tens of minutes thereafter, the temporal difference error δ_t assigns positive credit to the initial silent decision through the eligibility trace, thereby increasing the probability of outputting $\xi_t = 0$ in this state. Conversely, when the agent incorrectly outputs a play action during the self-regulation period of family emotions, resulting in an increase in sound pressure level or a deterioration in heart rate variability, negative credit is also retroactively assigned, suppressing the tendency of $\xi_t = 1$ in this state. The decay factor $\gamma\lambda$ of the eligibility trace determines the length of the historical window from which silent decision-making can obtain credit assignments. λ approaching 1 allows the model to capture long-term silent gains, but credit assignments spread across an excessively long historical sequence weaken the signal learning efficiency; λ approaching 0 limits the model's foresight. The specific value of this parameter falls under the scope of experimental configuration and will be given in the experimental section. The hierarchical tuple design may suggest that each of the three dimensions of the action space could demonstrate its specific function: the silence variable ξ_t controls the presence or absence of intervention, the music selection variable m_t determines the musical content of the intervention, and the spatial delivery vector p_t assigns the spatial orientation of the intervention. Moreover, the significant evidence might indicate that these three elements appear to work collaboratively within a hierarchical dependency relationship. Furthermore, the key findings could suggest that this structure enables the agent to either choose silence to await self-recovery or to choose targeted music delivery for proactive tuning. In light of these results, the data may indicate that the design appears to fully realize the restraint constraints of the action space. Design supports hierarchical dependency, enabling targeted agent decisions.

D. MULTI-CHANNEL COMPOSITE REWARD FUNCTION BASED ON QUALIFICATION TRAJECTORY

There are two temporally intertwined challenges in the design of reward function. The therapeutic effect of music is not uniformly spread during playback. Heart rate variability improvement, sound pressure level decrease and relaxation of facial expression all have different time constants. Feedback signals from different modalities at the same time point in opposite evaluation directions. Heart rate slowing is often accompanied by a momentary increase in sound pressure level and stable breathing occurs together with tense facial expressions [32-33]. The manifestation of the therapeutic effect lags behind the intervention that triggered it; physiological improvements appearing tens of minutes after playback need to be correctly attributed to the initial

music selection and delivery decision

across time intervals. The first challenge requires the reward function to make reasonable compromises when multimodal signals conflict; the second challenge requires the reward function to have a mechanism for distributing long-term effects forward along the time axis, corresponding to the multi-channel fusion structure and eligibility trace credit allocation discussed later in this section [34-35]. The immediate reward r_{t+1} is a weighted fusion of physiological, behavioral, and environmental channels using adjustable weights, the weighting form of which has been given in the problem formalization. The three channels capture aspects of healing progress from different modalities, and their calculation methods are explained one by one here. The physiological channel r_{t+1}^{phy} uses the heart rate variability time-domain index RMSSD for each member as the core observation variable. Let the RMSSD value of member i at time t be RMSSD_i^t, and its 60-second retrospective baseline be $\overline{\mathrm{RMSSD}}_i^t$. The rate of change for each person is defined as follows.

$$\Delta_i^t = \frac{\mathrm{RMSSD}_i^t - \overline{\mathrm{RMSSD}}_i^t}{\overline{\mathrm{RMSSD}}_i^t} \tag{7}$$

This rate of change, using a baseline reference, eliminates individual differences in resting heart rate variability among members, making the magnitude of improvement comparable across different members on a uniform scale. Physiological pathway rewards are further calculated as a weighted average of the rates of change across all members.

$$r_{t+1}^{\text{phy}} = \sum_{i=1}^{N} a_i^t \Delta_i^t \tag{8}$$

In the formula, the aggregation weight is allocated as $a_i^t = \exp(|\Delta_i^t|)/\sum_j \exp(|\Delta_j^t|)$ in the form of a softmax of the absolute value of the rate of change of each member. This weighting mechanism gives a higher weight to members who are currently deviating more from the baseline in the reward aggregation, so the model prioritizes the members with the most drastic emotional fluctuations, rather than distributing attention evenly among all members. When all members' RMSSD improves synchronously, each Δ_i^t is positive, and r_{t+1}^{phy} outputs a positive incentive; when the improvement rates among members are differentiated, positive and negative values cancel each other out in the weighted average, and the reward signal approaches zero, accurately reflecting the true state of the group's emotions, which have not yet formed a consistent direction of relief. The behavioral channel r_{t+1}^{beh} uses the sharp drop in the overall sound pressure level (SPL) of the household as a proxy signal for group emotional relief. The household SPL L_t could indicate that the A-weighted average SPL of each channel in the microphone array, measured in dBA, may suggest that this metric represents a key operational baseline. Moreover, the significant findings may suggest that downward changes in SPL appear to demonstrate a cooling of verbal conflict or a spread of silence, making it the most direct observation of family emotional relief at

the behavioral level. Furthermore, the evidence might indicate that the behavioral channel reward demonstrates that the first-order difference of the SPL, $\Delta L_t = L_t - L_{t-1}$, could reflect meaningful signal variation. In light of these results, the significant data may suggest that sign truncation and amplitude scaling appear to establish the processed reward measure reliably. Baseline reward shows r_{t+1}^{beh} derived after truncation and scaling.

$$r_{t+1}^{\text{beh}} = \frac{1}{L_{\text{ref}}} \cdot \max(-\Delta L_t, 0) \tag{9}$$

In the formula, L_{ref} is the reference sound pressure level, and rescales the reward amplitude to a range comparable to that of the physiological channel. In this channel, the output is positive only when the sound pressure level decreases, and zero (non-negative) when the sound pressure level increases. This truncation splits the roles of positive incentive and punishment: the behavioral channel is responsible for rewarding for soothing trends, and the environmental channel is responsible for the punishment for the rise in sound pressure levels, eliminating ambiguity arising from positive and negative gradient in a single channel. The environmental channel r_{t+1}^{env} exerts two suppression functions: punishing overly loud playback volume and punishing overly indiscriminate switching of tracks. These two forms of punishment reflect respectively the tendency to abuse the spatial projection intensity and the track selection variables as part of the intervention action. Let $V_t \in [0,1]$ be the current playback volume (i.e. the product of the magnitude of the vector p_t of spatial projection intensity multiplied by the system volume base value); take $\kappa_t \in \{0,1\}$, with $\kappa_t = 1$ if and only if $m_t \neq m_{t-1}$. The environmental channel reward is then given by:

$$r_{t+1}^{\text{env}} = -c_V V_t^2 - c_\kappa \kappa_t \tag{10}$$

In the formula, c_V and c_κ are penalty coefficients. The form of the volume penalty is quadratic, making the high volume range sharply suppressed, and the penalty in the low volume range nearly zero; the form of the track switching penalty is a constant, producing an equal negative signal for each switch; their superposition forms a continuous negative feedback against redundant intervention behaviour, forcing the model to learn moderation with respect to both the intervention intensity dimension and intervention frequency dimension. The weighted sum of the three channels yields an immediate reward r_{t+1}, but this feedback signal represents the multimodal evaluation only at the current instant and does not take into account the delayed healing effects. The improvement in the RMSSD signal of the physiological channel usually occurs several minutes or tens of minutes after the start of music playback; the improvement in the sound pressure level signal of the behavioral channel is faster but still exhibits several seconds or minutes of delay; and the volume and switching penalties in the environmental

channel offer nearly immediate feedback. The eligibility trace mechanism aims to establish a unified temporal credit allocation framework based on these differentiated delays. The recursive update formula for the eligibility trace vector e_t has been given in the problem formalization; its core function is

to record the participation traces of historical actions using an exponentially decaying chain. When the physiological channel detects a significant improvement in RMSSD at time $t+\tau$, the temporal difference error $\delta_{t+\tau}$ at that time is positive. After multiplying the eligibility trace with the policy gradient, positive credit is propagated back along the time axis to the music playback action before step τ, with the attenuation rate determined by $\gamma\lambda$. The sound pressure level decrease in the behavioral channel also propagates positive credit back to the playback action through the eligibility trace, but because the time constant of the behavioral response is shorter than that of the physiological response, its credit allocation is concentrated closer to the playback time. The penalty signal of the environmental channel has almost no delay, and the credit allocation of the eligibility trace for these two penalty terms is concentrated within a time window of one to two steps. The credit allocation of the three channels shares the same eligibility trace attenuation mechanism in mathematical form. The difference lies only in the value distribution of the temporal difference error δ_t of each channel at different time points. Its time delay hierarchy and credit propagation path are shown in Fig3. As shown in Fig3, the punishment signal in the environmental channel is almost instantaneous, the sound pressure response in the behavioral channel is intermediate, and the RMSSD improvement in the physiological channel is the most delayed. Together, these three constitute the signal hierarchy of the qualification trace across time attribution. The qualification trace propagates the improvement signal that appears later along the time axis backward. The attenuation rate $\gamma\lambda$ determines the depth of the historical window that credit allocation can cover, so that the physiological improvement that appears tens of minutes after playback can be correctly attributed to the initial intervention. The synergy between the multi-channel composite structure and the eligibility trace enables the reward function to possess both horizontal multimodal fusion capabilities and vertical long-range attribution capabilities. Horizontally, the physiological, behavioral, and environmental channels capture

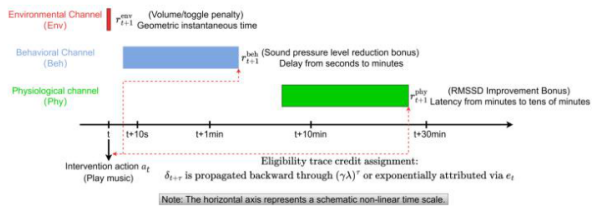

FIGURE 3. Time-series decomposition and credit allocation trajectory of multi-channel reward signals

aspects of healing progress from different modalities, while the weighted fusion mechanism compromises on w_{phy}, w_{beh}, and w_{env} when modal conflicts occur. Vertically, the eligibility trace passes delayed improvement signals forward along the time axis, and the decay factor $\gamma\lambda$ uniformly adjusts the depth of crediting channel of three channels for the model to capture the causal relationship between music intervention and physiological improvement over tens minutes intervals when updating policy parameters θ.

Therefore, the remaining three designs of state representation, action space and reward function have been constructed and integrated into an intact RL recommendation model in home environment.

III. EXPERIMENTS

A. DATASET CONSTRUCTION AND SCENARIO DESIGN

This study constructs a multimodal family emotional interaction simulation dataset to systematically evaluate the model's ability to perceive and regulate the dynamic coupling of emotions among multiple family members under controllable conditions. Moreover, the significant findings may suggest that recruiting a total of 12 families to participate in the simulation recording could provide the evidence needed to support this evaluation. Given that the data demonstrates the need to cover intergenerational interaction characteristics of three generations living together – grandparents, parents, and grandchildren – across four scenarios, the 12 families were recruited stratified according to member structure. Furthermore, the key results may indicate that each family consisting of 5 members – 2 grandparents, 2 parents or spouses, and 1 grandchild or teenager – could establish a total of 60 participants as a significant basis for analysis. Study shows recruitment ensured comparability across families. However, the important evidence may suggest that data collection could demonstrate a combination of script constraints and free interaction as the critical methodological approach. In light of the findings, each family appears to have received only the scenario background, member relationships, and core emotional goals before recording began, with no specific dialogue content, behavioral order, or emotional expression methods preset. Additionally, the results might indicate that members completing communication and responses autonomously based on their own family interaction habits could demonstrate the natural validity of the significant data collected. Notwithstanding the structured design, the evidence may suggest that this approach could establish that the data maintained a consistent direction of emotional evolution. Design preserves natural interaction dynamics. Thus, the key findings might indicate that real family interaction characteristics – such as language pauses, emotional delays, interruptions, and spatial movements – could demonstrate that the significant results remain ecologically valid. Since long-term emotional conflicts in real families involve continuous privacy exposure, cross-cycle relationship fluctuations, and uncontrollable environmental noise, it is difficult to

guarantee the uniformity of emotional coupling patterns among different families by directly conducting long-term natural data collection. Therefore, this study uses a controllable simulation scenario to construct a dataset, focusing on preserving the dynamic behavioral characteristics in the process of family emotional transmission and relationship coupling. The four scenarios correspond to four family interaction patterns with clear characteristics of emotional contagion. Argument Stalemate: This simulates a couple entering a state of avoidance and silence after a verbal conflict. Verbal interaction drops

sharply, heart rate variability remains low, and sound pressure level remains low. Emotional contagion is primarily characterized by avoidant alienation. Grandparents and grandchildren, though not directly involved in the argument, are indirectly affected by the spatial alienation and low-sound-pressure environment caused by the couple's cold war. Homework Tutoring: This simulates a parent's escalating verbal nagging and emotional contagion while tutoring their child. The tutor's anxiety spreads to the child through increased speech speed and pitch, causing the child to shift from calm to tension. Both parents' heart rate and sound pressure levels show a highly coupled, synchronous deterioration trend. Grandparents experience helplessness and blood pressure fluctuation because of the grandparents' experience of watching the intergenerational conflict. Afternoon Silence: It is like family members sit in silence for a long time in the afternoon and have little social participation. Other members give them occasional greetings but they are usually brief greetings and do not bring much emotional communication. The lack of interaction aggravates mood swing and diffuses among family members."Homecoming Fatigue" simulation shows three generations of family members come home one after another after a day of work and study. Then they finish the housework of the day such as making dinner and cleaning up, and stay in the same room after that. That is, they feel fatigue in all three generations with low arousal and low pleasure. The interaction between them is few and the sound pressure level is low. The three generations reinforce the fatigue of being in the same situation with each other instead of transmitting the fatigue in a unidirectional way. Each family group may suggest that continuous interactive scenarios across four scenarios could demonstrate a semi-structured script, with the significant results indicating that each recording lasted 30 to 40 minutes. Furthermore, the findings might indicate that annotators divided each complete continuous performance into independent segments based on the boundaries of the emotional contagion event. However, the evidence could suggest that each segment contained a complete cycle from the triggering to the relief of emotional contagion, lasting approximately 8 to 12 minutes, with an average of about 4 segments per recording. In light of these results, the significant data may demonstrate that a total of 192 valid segments are obtained across the four scenarios, totaling approximately 32.9 hours in duration. Scenario scripts constrain key interaction nodes but allow improvisation. Moreover, the evidence could indicate that the scenario scripts allowed members to improvise within the given contextual framework to ensure the natural dynamic range of behavior and physiological signals. Given that the findings demonstrate that all participating families signed informed consent forms before recording began, the key results might suggest that the data collection process is used solely for academic research. Additionally, the significant evidence may indicate that all audio and video data underwent anonymization before being used for model training, with the important findings suggesting that member names, home addresses, and other identifiable information are removed. Therefore, the results could demonstrate that the original data is not publicly disseminated. Data streams show three modalities included. Notwithstanding

these results, the significant findings may suggest that the final dataset includes three synchronous data streams: acoustic behavioral modality, visual behavioral modality, and physiological feedback modality. Furthermore, the evidence could indicate that the acoustic modality primarily records changes in speech rate, pitch fluctuations, and sound pressure level. The visual modality describes changes in head orientation, body movements, and spatial distance. The physiological modality reflects changes in the state of the autonomic nervous system. All three modalities are aligned using a unified timestamp and organized as a continuous sequence input within a 10-second time window. Each time window simultaneously retains the behavioral statistics within the current window and short-term dynamic changes from previous time windows, used to characterize the propagation trend and coupling strength of family emotions during continuous interaction. Data annotation is conducted on two levels. The first level is frame-by-frame behavioral annotation, where three trained annotators segmented and labeled the intensity of verbal interaction, changes in spatial distance, and significant emotional events within each segment. The Kappa coefficient for the three annotators is 0.82. The second level is intervention strategy annotation, independently completed by two professionals with psychological counseling qualifications. Annotators used 10-second time windows and, based on the trends in sound pressure level changes and heart rate variability within each window, determined whether the time window is an "appropriate intervention window" or a "suitable silence window." The appropriate intervention window may suggest that emotional indicators showed a continuous deterioration trend, the family's own rhythm is insufficient to recover within a reasonable time, and music intervention could disrupt the chain of deterioration. Moreover, the suitable silence window might indicate that the family's emotions are in a self-repairing phase or in a stable state, and the expected net benefit of external music intervention appears negative. Furthermore, the significant findings demonstrate that the Kappa coefficient for the two annotators is 0.79. In light of these results, the inconsistent time windows could indicate that discussion and agreement are necessary, serving as

TABLE 1. Data Composition and Labeling Statistics of Four Types of Family Stress Scenarios

Scene type	Number of segments	Total duration (hour)	Suitable Intervention Windows	Silence-Preferred Windows
Conflict Gridlock	48	8.1	726	1014
Homework Tutoring	52	8.7	1093	749
Afternoon Silence	46	7.9	681	1107
Post-Work Fatigue	46	8.2	739	989
Total	192	32.9	3239	3859

the key benchmark for subsequent silent action learning assessments. Labels show unified binary standard used. To ensure consistency of the labeling system during model training, the evidence may suggest that all time window labels are organized using a unified binary classification standard. Additionally, the "appropriate intervention window" could demonstrate that family emotions continue to deteriorate, negative emotions spread among members, and there appears to be a lack of self-recovery trend in the short term. Given that the results indicate the "appropriate silence window" corresponds to a state where family emotions are in a natural process of easing, external intervention might disrupt the existing recovery rhythm. Notwithstanding this observation, the significant data may suggest that scene-level emotion transformation labels are generated based on changes in the overall emotional state of the family, which are used to calculate the accuracy of subsequent state representation and the positive emotion conversion rate.

Dataset shows windows, labels, segments form evaluation foundation. The number of time windows, label distribution, and segment size in different scenarios could indicate that the data foundation for subsequent experimental evaluation appears comprehensive. Therefore, the key findings may suggest that the data composition and annotation statistics of the four types of scenarios demonstrate results shown in Table 1.

Note: Unlabeled time periods represent scene transitions, equipment calibration, and ambiguous intervals deemed unclassifiable by the annotator. As shown in Table 1, the number of suitable silence windows and appropriate intervention windows is unbalanced in the four types of scenarios. The number of suitable silence windows is larger than the number of appropriate intervention windows in conflict stalemate and afternoon silence scenarios. This means that the emotion of a family in these two types of scenarios has more space for self-regulation, and the agent has more to learn in terms of more repressive strategies. On the other hand, the number of appropriate intervention windows is larger than the number of suitable silence windows in homework tutoring scenarios. This means that more urgent intervention is needed in these scenarios. This distributional difference offers a natural comparative condition for the subsequent analysis of behavioral patterns of silent action learning.

B. EXPERIMENTAL SETUP AND HYPERPARAMETER CONFIGURATION

Model evaluation consists of five-fold cross-validation and an independent test. Two families out of 12 stratified families are randomly chosen as the independent test (32 segments in total). These 32 segments are only used for final evaluation and not used for training or hyperparameter optimization. The other 10 families are stratified into five folds (160 segments in total), and each fold is used as a validation set once while the other four folds are used as training set. Stratification is conducted based on family group and thus different scene segments from the same family not appear in both sets. The member embedding dimension D of the state tensor is 64. The explicit behavior representation dimension d_b and the implicit physiological vector dimension d_b are 32 and 32, respectively. The dimension-up matrices U_b and U_p in gated fusion unit map the above two dimensions to a 64-dimensional embedding space. The graph neural network takes three layers of GraphSAGE convolution layers with hidden dimension being 64.

The aggregation function agg is average pooling. The node encoder ϕ_{node} consists of two layers of fully connected layers followed by a LayerNorm normalization layer. This paper set the exponential moving average decay coefficient β representing the information of edge history to be 0.85 such that the model is sensitive to the intensity of infection in the short term and stable in the medium- to long term. The eligibility trace parameter λ is set to 0.9, and the discount factor γ is set to 0.99. Moreover, the findings may suggest that their product, $\gamma\lambda = 0.891$, could indicate the decay rate of long-range credit assignment in significant detail. Furthermore, the evidence may demonstrate that actions performed one time step ago retain approximately 89.1% of the trace weight at the current time, decaying to approximately 31.6% after ten time steps, and decreasing to approximately 10.0% after twenty time steps. Given that the results appear to support this decay pattern, the model could indicate that it effectively attributes physiological improvement signals within a window of approximately twenty steps to historical intervention actions. Results show λ affects credit assignment window. However, the key findings may suggest that the value of λ is used in a grid search on the validation set, ranging from 0.80 to 0.95 in 0.05 steps. Additionally, the significant evidence could demonstrate that $\lambda = 0.9$ achieved the optimal positive emotion conversion rate. Notwithstanding the important results, higher values might indicate that expanding the credit assignment window led to gradient signal diffusion and decreased learning efficiency. The weights of the three channels of the reward function are determined by grid search to be $w_{\mathrm{phy}} = 0.5$, $w_{\mathrm{beh}} = 0.3$, and $w_{\mathrm{env}} = 0.2$. The search range is $w_{\mathrm{phy}} \in [0.3, 0.7]$, $w_{\mathrm{beh}} \in [0.1, 0.4]$, and $w_{\mathrm{env}} \in [0.1, 0.3]$, with a step size of 0.05 for each channel, satisfying the

normalization constraint $w_{\text{phy}} + w_{\text{beh}} + w_{\text{env}} = 1$. The search results show that the model performance is stable when the physiological channel weights are in the range of 0.45 to 0.55, the behavioral channel weights fluctuate less in the range of 0.25 to 0.35, and the environmental channel weights exceeding 0.25 led to overly conservative silent action learning and insufficient playback frequency within the appropriate intervention window. The environmental channel penalty coefficients c_V and c_κ are set to 0.1 and 0.05 respectively. They are determined by grid search within $\{0.01, 0.05, 0.1, 0.2\}$. The current values make the quadratic term of the volume penalty significantly suppress the volume in the high volume area while the penalty in the low volume area is close to zero. The track switching penalty has a controllable cumulative effect when switching is infrequent. The music library size K is set to 200, and the track embedding dimension D_c is set to 64. The music library embedding matrix is initialized by a pre-trained music emotion regression model, which predicts pleasure and arousal using audio features as input. The penultimate hidden state of the model is projected to 64 dimensions using principal component analysis to serve as the music embedding vector. Both the policy network and the value network used the Adam optimizer with a learning rate η of 3×10^{-4} and a batch size of 32 time steps. The qualification trace vector is initialized to zero at the beginning of each scene. The study may suggest that the total number of training rounds could

demonstrate adequate coverage at 500, with all training segments traversed once per round. Furthermore, the positive emotion conversion rate might indicate the optimal model parameters, as the evidence shows that the validation set provides the key basis for selecting results for evaluation on the test set. Moreover, the significant early stopping mechanism could suggest that training terminates when the validation set metrics show no improvement for 30 consecutive rounds. In light of the findings, the results may indicate that the model appears to converge reliably between 320 and 380 rounds under actual operating conditions. All experiments are conducted on a server equipped with an NVIDIA A100 GPU and 128 GB of memory. The training time per round is approximately 47 seconds, and the model inference latency is approximately 18 milliseconds per step, meeting the real-time interaction requirements in a home setting.

IV. RESULTS ANALYSIS

A. VALIDATION OF THE EFFECTIVENESS OF RELATIONSHIP MODELING IN STATE REPRESENTATION

Emotional states in family music therapy scenarios exhibit significant multi-member coupling characteristics, making it difficult to fully characterize the continuously changing emotional transmission relationships within the family by relying solely on individual emotional signals. To verify the effectiveness of relation-aware state tensors in family emotion modeling, this paper removes key modules such

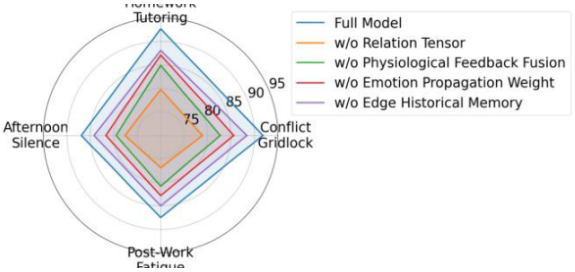

FIGURE 4. Comparison of state representation accuracy

as relation tensors, behavioral-physiological gating fusion, dynamic emotion transmission weights, and side-historical memory, and compares the state representation accuracy of different models in four types of family stress scenarios. The results are shown in Fig4. Note: The accuracy of state representation is calculated based on the family emotional coupling state classification task, and the classification labels are generated by manually labeled family overall emotional state. As shown in Fig4, the complete model achieved higher accuracy in state representation across all four scenarios than the degenerate models, reaching 91.8% and 92.7% accuracy in the conflict/stalemate and tutoring scenarios, respectively. Removing the relation tensor resulted in the most significant decrease in accuracy across all scenarios, indicating that inter-member relationship information significantly contributes to the representation of family emotional states. In contrast, while removing the behavior-physiology gating fusion resulted in a decrease in model accuracy, the overall reduction is relatively

limited, suggesting that the synergistic fusion of behavioral and physiological signals can further enhance state representation. Removing the dynamic emotion propagation weights led to a more significant performance drop in the tutoring scenario, indicating that dynamic relationship propagation plays a role in high-interaction scenarios. Removing the side-historical memory module resulted in a more significant decrease in accuracy in the afternoon silence and homecoming fatigue scenarios, suggesting that historical relationship information is helpful for state representation in low-interaction scenarios. Overall, each relation modeling module contributed to varying degrees of improvement in the representation of coupled family emotional states. Since behavioral and physiological signals in a home environment are easily affected by equipment errors, environmental interference, and fluctuations in the random behavior of family members, simply comparing the accuracy of state representation is insufficient to demonstrate the stability of the model under complex conditions. To further evaluate the impact of relationship modeling on the robustness of the model, this paper progressively adds random noise perturbations to the original behavioral and physiological data and tests the changes in the accuracy of state representation of different models under different noise levels. The results are shown in Fig5. As shown in Fig5, the accuracy of

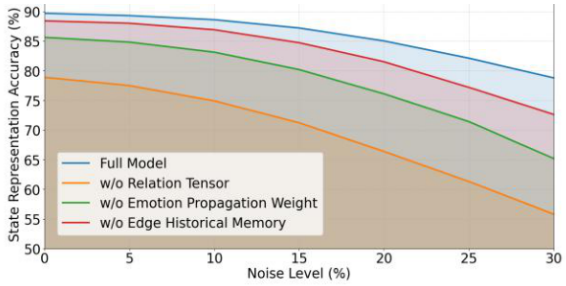

FIGURE 5. Noise robustness test

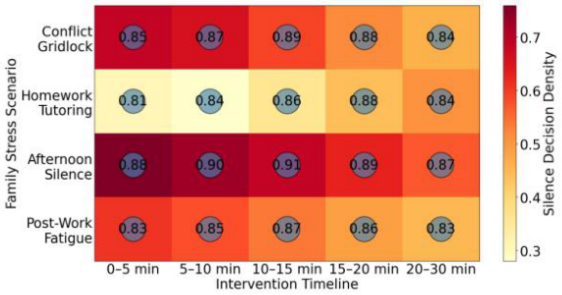

FIGURE 6. Silent decision distribution and expert consensus

state representation in all models decreases with increasing noise levels, but the magnitude of the decrease varies significantly. The complete model maintains high stability under low noise conditions; when the noise level increases from 0% to 10%, the accuracy only decreases from 89.7% to 88.6%. In contrast, the model after removing the relation tensor decreases more rapidly, with accuracy dropping to 55.8% under 30% noise conditions, indicating a significant weakening of state representation stability after the absence of relational structures. The model after removing the dynamic emotion propagation weights shows little difference from the complete model in low noise conditions, but the decrease gradually increases in the medium-to-high noise range, indicating that dynamic relation propagation plays a role in complex noise environments. After removing the side history memory module, the model's accuracy also decreases significantly in high noise conditions, indicating that historical relational information helps maintain the stability of state representation. Overall, the relation-aware state tensor not only improves the accuracy of family emotional coupling state representation but also exhibits better stability under noise perturbation conditions.

B. STRATEGY QUALITY ANALYSIS OF SILENT MOTION LEARNING

In family music therapy, continuous playback is not always the optimal intervention strategy. In some family emotional states, external music intervention may disrupt the self-healing rhythms that are forming among family members. Therefore, the model needs to learn not only when to play music but also when to remain silent. To verify the policy learning effect of the silence variable in the hierarchical action space, this paper further statistically analyzed the distribution of silence decisions in different family stress scenarios and different intervention stages, and compared its consistency with the expert-labeled appropriate silence window. The results are shown in Fig6. Note: The size of the bubble represents the time overlap rate between the model decision and the expert annotation in the silent window, and the number inside the bubble represents the corresponding consensus value. As shown in Fig6, there are significant differences in the distribution of silent decisions across different family scenarios. The argument/stalemate and afternoon silence scenarios exhibit a relatively high density of silent decisions, while the proportion of silence in the homework tutoring scenario is relatively low. This trend is largely consistent with the differences in the distribution of suitable silence windows and appropriate intervention windows for each scenario in Table 1. The expert consistency across different time periods may suggest that alignment remains notably stable, generally above 0.84 for most periods, with the findings reaching around 0.89 in the afternoon silence scenario, indicating that the model-generated silent strategies could demonstrate strong agreement with expert annotation results. Moreover, the significant evidence might indicate that the proportion of silence in the argument/stalemate and home-wake fatigue scenarios could reflect a gradual decrease with increasing intervention duration. Furthermore, the results may suggest that this pattern indicates a dynamic change in the distribution of silent and intervention decisions. In light of these key findings, the data could demonstrate that such temporal shifts appear meaningful across different stages. Findings show distribution shifts across periods as intervention duration increases. In contrast, the proportion of silence in the homework tutoring scenario is generally lower, reflecting that the model tends to proactively intervene in scenarios with high interaction and high emotional coupling. Overall, the results show that by introducing a silence variable, the model can form differentiated intervention and silence strategies based on different family emotional states,

rather than a fixed, uniform playback strategy. In all test scenarios, the average time window consistency rate between the model's silent decision-making and the expert-annotated appropriate silent window reached 86.3%, indicating that silent action learning can stably fit the family's emotional self-recovery rhythm.

C. CONTRIBUTION ABLATION OF MULTI-CHANNEL COMPOSITE REWARD FUNCTION

The reward signals in family music therapy are both heterogeneous and delayed, making it difficult to stably characterize the actual healing progress by relying on a single feedback source. To verify the independent contributions of the physiological, behavioral, and environmental channels in the overall reward structure, this paper removes the three reward channels respectively and normalizes the RMSSD

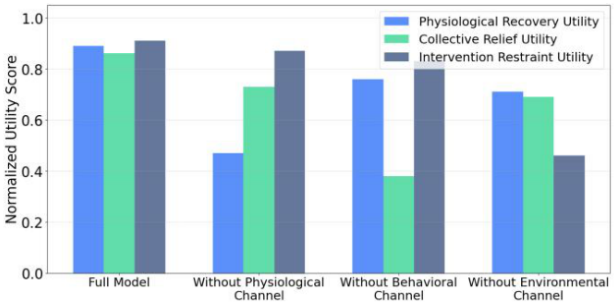

FIGURE 7. Results of ablation experiments using multi-channel reward functions

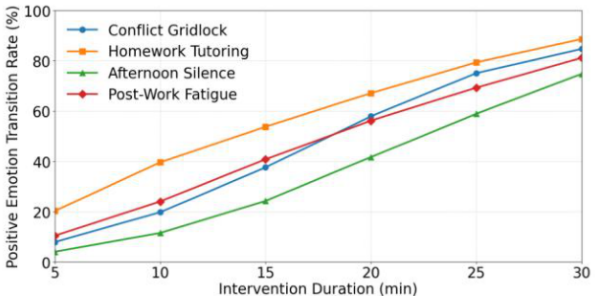

FIGURE 8. Dynamic changes in the positive emotion conversion rate with intervention duration

improvement, the reduction in family sound pressure level, and the environmental penalty term into three indicators: physiological recovery benefit, group soothing benefit, and intervention restraint benefit. The reward optimization results of different models are compared, and the results are shown in Fig 7. As shown in Fig 7, the complete three-channel model maintained the highest level in all three normalized benefit indicators, indicating that the multi-channel reward structure can simultaneously take into account the three optimization objectives of physiological recovery, group soothing, and intervention restraint. After removing the physiological channel, the physiological recovery benefit decreased most significantly, from 0.89 to 0.47, while the other two indicators showed relatively limited changes. After removing the behavioral channel, the group soothing benefit decreased from 0.86 to 0.38, becoming the lowest value among all models. After removing the environmental channel, the intervention restraint benefit decreased from 0.91 to 0.46, while the other two indicators also showed some degree of decline. Overall, the results indicate that the three reward channels correspond to different optimization focuses. The physiological channel mainly affects long-term recovery effects, the behavioral channel mainly affects family interaction soothing, and the environmental channel has a significant constraining effect on intervention stability. Together, they constitute a composite reward mechanism for family emotional healing.

D. ANALYSIS OF THE RELATIONSHIP BETWEEN THE POSITIVE EMOTION CONVERSION RATE AND THE DURATION OF INTERVENTION

To further evaluate the overall therapeutic effect of the model during continuous family emotional intervention, this paper statistically analyzed the dynamic changes in the positive emotion conversion rate over the duration of intervention in four types of family stress scenarios. The positive emotion conversion rate measures the proportion of the family's overall emotional state that transitions from a negative or low-arousal state to a stable and calm state; a higher value indicates a more significant moderating effect of the model on the family's emotional coupling state. Considering the differences in the rhythm of emotion propagation and the distribution of silent strategies in different scenarios, this paper uses 5-minute intervals to conduct phased statistical analysis of the positive emotion conversion rate during a 30-minute continuous intervention process, as shown in Fig 8. As shown in Fig8, the positive emotion conversion rate in all four scenarios continuously increased with the duration of intervention, but the growth rhythms differed significantly across scenarios. The homework tutoring scenario increased the fastest in the first 10 minutes, which means that the model can intervene in the family earlier in a highly interactive and coupled environment. For the afternoon silence scenario, the growth rate of the four scenarios was relatively slow in the early stage, and the conversion rate of the afternoon silence scenario began to increase rapidly after 15 minutes, indicating that the latency period of improving the level of emotion was relatively long; the argument/stalemate scenario increases slowly in the early stage, and after 20 minutes, the growth rate increases rapidly and reaches a relatively high level of 84.7%, which also has the characteristic of late-stage recovery; and the homecoming fatigue scenario increases in a relatively stable process, and the fluctuation of the curve is relatively small. Generally, the conversion rate curve of the four scenarios gradually became flat in the 25-30 minute range. The average positive emotion conversion rate of the four scenarios reached 82.3% at 30 minutes, indicating that the model can intervene in the family gradually to promote a gradual shift in the family's emotional state and enter a relatively stable recovery stage.

V. CONCLUSION

This paper tackles the issue of representing the coupled emotional states of multiple family members in a natural environment with traditional music recommendation models. This paper builds a reinforcement learning recommendation framework with

relationship-aware state tensor, silent action learning, and multi-channel composite reward mechanism. Graph neural network models the dynamic emotional propagation relationships among members. Eligibility traces allocate the time-based credit for long-term therapeutic feedback to help the model develop an emotion regulation strategy of intervention and restraint in continuous family interactions. The

experimental evaluation shows that the model owns stable state recognition ability and strategy consistency in different family stress scenarios. Relationship modeling structure helps to enhance the state robustness in more complex noisy environment. The learning of silent action is effectively synchronized with the emotional self-recovery rhythm of the family. The multi-channel reward mechanism achieves relatively balanced optimization effect among physiological recovery, group soothing, and intervention stability. Meanwhile, the emotional recovery curve of different scenarios in continuous intervention shows phased differences in accordance with the interaction pattern of family. There are still some limitations in our work. The available data are from scripted simulated family scenarios. Long-term behavioral drift and changes of member relationship in real-world family environment and cross-cycle emotional accumulation in real family environment have not been incorporated in modeling. The available reward signal is still based on physiological and behavioral indicators. Few semantic and emotional feedback are available. In the future, this paper will further incorporate long-term tracking data from real-world families, extend cross-room and multi-device collaborative perception, and combine more granular emotional semantic understanding with adaptive intervention strategy to support the development of family music therapy system towards long-term dynamic emotional support.

REFERENCES

[1] Z. Liu, W. Xu, W. Zhang, and Q. Jiang, "An emotion-based personalized music recommendation framework for emotion improvement," Information Processing & Management, vol. 60, no. 3, pp. 103256–103264, 2023, doi: https://doi.org/10.1016/j.ipm.2022.103256.

[2] R. De Prisco, A. Guarino, D. Malandrino, and R. Zaccagnino, "Induced emotion-based music recommendation through reinforcement learning," Applied Sciences, vol. 12, no. 21, pp. 11209–11215, 2022, doi: https://doi.org/10.3390/app122111209.

[3] S. I. Baek and Y. K. Lee, "A Continuous Music Recommendation Method Considering Emotional Change," Applied Sciences, vol. 15, no. 13, pp. 7222–7229, 2025, doi: https://doi.org/10.3390/app15137222.

[4] M. Carvalho, N. Cera, and S. Silva, "The "ifs" and "hows" of the role of music on the implementation of emotional regulation strategies," Behavioral Sciences, vol. 12, no. 6, pp. 199–206, 2022, doi: https://doi.org/10.3390/bs12060199.

[5] M. Stratigi, E. Pitoura, and K. Stefanidis, "SQUIRREL: A framework for sequential group recommendations through reinforcement learning," Information Systems, vol. 112, no. 1, pp. 102128–102134, 2023, doi: https://doi.org/10.1016/j.is.2022.102128.

[6] K. Howland, K. Edvardsson, H. Lees, and L. Hooker, "Telehealth use in the well-child health setting. A systematic review of acceptability and effectiveness for families and practitioners," International journal of nursing studies advances, vol. 8, no. 1, pp. 100277–106282, 2025, doi: https://doi.org/10.1016/j.ijnsa.2024.100277.

[7] V. Peters, J. Bissonnette, D. Nadeau, A. Gauthier-Légaré, and M. A. Noël, "The impact of musicking on emotion regulation: A systematic review and meta-analysis," Psychology of Music, vol. 52, no. 5, pp. 548–568, 2024, doi: https://doi.org/10.1177/03057356231212362.

[8] B. Paley and N. J. Hajal, "Conceptualizing emotion regulation and coregulation as family-level phenomena," Clinical child and family psychology review, vol. 25, no. 1, pp. 19–43, 2022, doi: https://doi.org/10.1007/s10567-022-00378-4.

[9] A. Tran, K. H. Greenaway, J. Kostopoulos, S. T. O'Brien, and E. K. Kalokerinos, "Mapping interpersonal emotion regulation in everyday life," Affective Science, vol. 4, no. 4, pp. 672–683, 2023, doi: https://doi.org/10.1007/s42761-023-00223-z.

[10] S. S. Yap, F. T. Ramseyer, J. Fachner, C. Maidhof, W. Tschacher, and G. Tucek, "Dyadic nonverbal synchrony during pre and post music therapy interventions and its relationship to self-reported therapy readiness," Frontiers in human neuroscience, vol. 16, no. 1, pp. 912729–912734, 2022, doi: https://doi.org/10.3389/fnhum.2022.912729.

[11] C. Stolterfoth and E. Pfeifer, "Music therapy and silence–Silence in music therapy. A systematic review," The Arts in Psychotherapy, vol. 93, no. 1, pp. 102286–102293, 2025, doi: https://doi.org/10.1016/j.aip.2025.102286.

[12] Y. Rong, N. Chen, J. Dong, Q. Li, X. Yue, L. Hu, et al., "Expectations of immediate and delayed reward differentially affect cognitive task performance," NeuroImage, vol. 262, no. 1, pp. 119582–119587, 2022, doi: https://doi.org/10.1016/j.neuroimage.2022.119582.

[13] A. W. Awan, S. M. Usman, S. Khalid, A. Anwar, R. Alroobaea, S. Hussain, et al., "An ensemble learning method for emotion charting using multimodal physiological signals," Sensors, vol. 22, no. 23, pp. 9480–9488, 2022, doi: https://doi.org/10.3390/s22239480.

[14] X. Hu, F. Li, and R. Liu, "Detecting music-induced emotion based on acoustic analysis and physiological sensing: A multimodal approach," Applied Sciences, vol. 12, no. 18, pp. 9354–9361, 2022, doi: https://doi.org/10.3390/app12189354.

[15] P. Pichappan, "A Review of the Emotion-Induced Music Recommendation Systems," Journal of Digital Information Management, vol. 23, no. 2, pp. 112–133, 2025, doi: https://doi.org/10.6025/jdim/2025/23/2/112-133.

[16] X. Han, F. Chen, and J. Ban, "A GAI-based multi-scale convolution and attention mechanism model for music emotion recognition and recommendation from physiological data," Applied Soft Computing, vol. 164, no. 1, pp. 112034–112041, 2024, doi: https://doi.org/10.1016/j.asoc.2024.112034.

[17] Z. Ahmad and N. Khan, "A survey on physiological signal-based emotion recognition," Bioengineering, vol. 9, no. 11, pp. 688–693, 2022, doi: https://doi.org/10.3390/bioengineering9110688.

[18] A. Wong, T. Bäck, A. V. Kononova, and A. Plaat, "Deep multiagent reinforcement learning: challenges and directions," Artificial Intelligence Review, vol. 56, no. 6, pp. 5023–5056, 2023, doi: https://doi.org/10.1007/s10462-022-10299-x.

[19] Y. Miyashita and T. Sugawara, "Two-stage reward allocation with decay for multi-agent coordinated behavior for sequential cooperative task by using deep reinforcement learning," Autonomous Intelligent Systems, vol. 2, no. 1, pp. 10–17, 2022, doi: https://doi.org/10.1007/s43684-022-00029-z.

[20] B. Xiao, H. K. Lam, X. Su, Z. Wang, F. P. W. Lo, S. Chen, et al., "Sampled-data control through model-free reinforcement learning with effective experience replay," Journal of Automation and Intelligence, vol. 2, no. 1, pp. 20–30, 2023, doi: https://doi.org/10.1016/j.jai.2023.100018.

[21] X. Lou, Q. Yin, J. Zhang, C. Yu, Z. He, N. Cheng, et al., "Offline reinforcement learning with representations for actions," Information Sciences, vol. 610, no. 1, pp. 746–758, 2022, doi: https://doi.org/10.1016/j.ins.2022.08.019.

[22] I. Gordon, A. Tomashin, and O. Mayo, "A theory of flexible multimodal synchrony," Psychological review, vol. 132, no. 3, pp. 680–687, 2025, doi: https://doi.org/10.1037/rev0000495.

[23] S. Delleli, I. Ouergui, C. G. Ballmann, H. Messaoudi, K. Trabelsi, L. P. Ardigo, et al., "The effects of pre-task music on exercise performance and associated psycho-physiological responses: a systematic review with multilevel meta-analysis of controlled studies," Frontiers in Psychology, vol. 14, no. 1, pp. 1293783–1293791, 2023, doi: https://doi.org/10.3389/fpsyg.2023.1293783.

[24] R. Chen and Y. Tan, "Credit assignment with predictive contribution measurement in multi-agent reinforcement learning," Neural Networks, vol. 164, no. 1, pp. 681–690, 2023, doi: https://doi.org/10.1016/j.neunet.2023.05.021.

[25] D. Wang, J. Wang, L. Hu, and M. Zhao, "Event-based online learning control design with eligibility trace for discrete-time unknown nonlinear systems," Engineering Applications of Artificial Intelligence, vol. 123, no. 1, pp. 106240–106244, 2023, doi: https://doi.org/10.1016/j.engappai.2023.106240.

[26] A. Israelsson, A. Seiger, and P. Laukka, "Blended emotions can be accurately recognized from dynamic facial and vocal expressions," Journal of Nonverbal Behavior, vol. 47, no. 3, pp. 267–284, 2023, doi: https://doi.org/10.1007/s10919-023-00426-9.

[27] J. Luo, G. Zhang, Y. Su, Y. Lu, Y. Pang, Y. Wang, et al., "Quantitative analysis of heart rate variability parameter and mental stress index,"

Frontiers in Cardiovascular Medicine, vol. 9, no. 1, pp. 930745–930452, 2022, doi: https://doi.org/10.3389/fcvm.2022.930745.

[28] L. Gong, W. Chen, M. Li, and T. Zhang, "Emotion recognition from multiple physiological signals using intra-and inter-modality attention fusion network," Digital Signal Processing, vol. 144, no. 1, pp. 104278–104285, 2024, doi: https://doi.org/10.1016/j.dsp.2023.104278.

[29] E. Ghaleb, J. Niehues, and S. Asteriadis, "Joint modelling of audio-visual cues using attention mechanisms for emotion recognition," Multimedia Tools and Applications, vol. 82, no. 8, pp. 11239–11264, 2023, doi: https://doi.org/10.1007/s11042-022-13557-w.

[30] S. L. Jacobsen, G. Gattino, U. Holck, and J. Ø. Bøtker, "Music, families and interaction (MUFASA): a protocol article for an RCT study," BMC psychology, vol. 10, no. 1, pp. 252–259, 2022, doi: https://doi.org/10.1186/s40359-022-00957-8.

[31] Y. Shibata, A. Yoshimoto, K. Yamashiro, Y. Ikegaya, and N. Matsumoto, "Delayed reinforcement hinders subsequent extinction," Biochemical and biophysical research communications, vol. 591, no. 1, pp. 20–25, 2022, doi: https://doi.org/10.1016/j.bbrc.2021.12.101.

[32] B. Flater, A. Brean, and D. S. Quintana, "Music therapy and vagally mediated heart rate variability: A systematic review and narrative synthesis," International Journal of Psychophysiology, vol. 219, no. 1, pp. 113288–113294, 2026, doi: https://doi.org/10.1016/j.ijpsycho.2025.113288.

[33] A. Warmbrodt, R. Timmers, and R. Kirk, "The emotion trajectory of self-selected jazz music with lyrics: A psychophysiological perspective," Psychology of Music, vol. 50, no. 3, pp. 756–778, 2022, doi: https://doi.org/10.1177/03057356211024336.

[34] Q. Wang, C. Zhang, and B. Hu, "Dynamic programming with meta-reinforcement learning: a novel approach for multi-objective optimization," Complex & Intelligent Systems, vol. 10, no. 4, pp. 5743–5758, 2024, doi: https://doi.org/10.1007/s40747-024-01469-1.

[35] J. Tang, Z. Ma, K. Gan, J. Zhang, and Z. Yin, "Hierarchical multimodal-fusion of physiological signals for emotion recognition with scenario adaption and contrastive alignment," Information Fusion, vol. 103, no. 1, pp. 102129–102136, 2024, doi: https://doi.org/10.1016/j.inffus.2023.102129.