Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Trans-SAC: Gamification Design for Sustained User Engagement in Residential Demand Response

Tianxiao Peng

pengtianxiao@gdust.edu.cn

Guangdong University of Science and Technology

Research Article

Keywords: Demand Response, Gamification, Deep Reinforcement Learning, User Engagement, Psychological Fatigue, Transformer Architecture, Smart Grid

Posted Date: August 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10700397/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Trans-SAC: Gamification Design for Sustained User Engagement in Residential Demand Response

Tianxiao Peng*
Guangdong University of Science and Technology, Dongguan 523070, Guangdong, China
Correspondence to: pengtianxiao@gdust.edu.cn*

Abstract

Residential Demand Response (DR) is a critical mechanism for maintaining the supply-demand balance in modern smart grids. While gamification has recently emerged as a promising strategy to incentivize residential participation, existing programs suffer from severe long-term user fatigue. Once the initial "novelty effect" wears off, users rapidly habituate to static rewards, leading to a precipitous decline in DR reliability. Furthermore, traditional pricing models and standard Deep Reinforcement Learning (DRL) agents optimize purely for physical grid states, failing to capture the latent psychological dynamics of human participants.To address this critical gap, this paper proposes Trans-SAC, a novel gamification-aware dynamic DR framework that integrates psychological fatigue modeling with advanced machine learning. First, we formulate a mathematical behavioral model to quantify the accumulation and decay of user fatigue. Second, a Transformer-based module is employed to process long-sequence historical interaction logs, utilizing multi-head self-attention to predict the real-time cognitive engagement probability of each user. Finally, this psychological state is incorporated into the state space of a Soft Actor-Critic (SAC) agent. Guided by entropy-regularized exploration, the SAC agent autonomously executes "intermittent reinforcement" by dynamically adjusting task difficulties and reward multipliers.Extensive simulations, which combine real-world electricity-consumption traces from the Pecan Street dataset with a synthetically generated gamification-behavior environment, demonstrate the potential effectiveness of the proposed framework. Over a continuous 12-month simulated period, Trans-SAC was found to mitigate the modeled user-habituation process, maintaining a simulated Load Reduction Achievement Rate (LRAR) of 85.6% and a simulated User Retention Rate of 84.2%. Crucially, by strategically restraining rewards for intrinsically motivated users and deploying high-value interventions only when critical fatigue is detected, Trans-SAC achieved a Unit Incentive Cost of $0.061/kWh, representing a 21.8% economic improvement over standard DRL baselines under the simulated behavioral assumptions detailed in Section 4.1.The results suggest that coupling cognitive state assessment with DRL achieves a favorable trade-off between long-term grid reliability and economic efficiency under the simulated behavioral assumptions of this study.

Keywords: Demand Response; Gamification; Deep Reinforcement Learning; User Engagement; Psychological Fatigue; Transformer Architecture; Smart Grid

1. Introduction

With the global acceleration towards carbon neutrality, the deep integration of highly volatile and intermittent renewable energy sources poses unprecedented challenges to the real-time supply-demand balance of smart grids [1]. In this context, the massive untapped flexible loads in the residential sector—such as heating, ventilation, and air conditioning (HVAC) systems, and electric vehicles (EVs)—have garnered widespread attention from both academia and industry. Demand Response (DR) programs, which guide residential users to proactively curtail or shift their load consumption during peak grid periods, have emerged as a pivotal mechanism to enhance grid resilience, defer costly infrastructure upgrades, and facilitate the accommodation of renewable energy [2]. However, designing efficient and sustainable incentive mechanisms to ensure long-term, stable user participation remains a core bottleneck in current demand-side management.

Existing DR incentive programs predominantly rely on purely economic instruments, such as Time-of-Use (TOU) pricing, Critical Peak Pricing (CPP), or direct load control compensation [3]. These traditional pricing and incentive models are fundamentally built upon the assumption of the "rational economic man" (Homo economicus), positing that users will consistently deliver proportional load reductions as long as fixed financial compensations are provided. However, extensive empirical studies and real-world engineering projects have demonstrated that human behavior is driven by complex psychological characteristics and bounded rationality [4]. After prolonged participation in purely economically driven DR programs, users universally exhibit a pronounced "diminishing marginal utility." This long-term participation fatigue, triggered by the "habituation effect" in behavioral economics, causes users to gradually lose sensitivity to fixed electricity price discounts. Furthermore, the perceived inconvenience of altering daily routines eventually outweighs the static monetary benefits. Consequently, this manifests as a precipitous, cliff-like drop in response reliability during the middle and late stages of DR project implementation, severely threatening the stability and predictability of grid dispatch operations.

To mitigate the aforementioned long-term participation fatigue, recent interdisciplinary efforts have attempted to introduce "Gamification" into the domains of energy management and demand response [5]. By incorporating non-material or semi-material rewards—such as virtual points, leaderboards, and progress visualization—gamification mechanisms aim to stimulate users' intrinsic motivation and sense of achievement. Although these empirical explorations have yielded initial success, existing gamified DR models largely remain confined to static rule levels. Specifically, the system inherently assumes that the incentive utility of gamified points remains constant over time [6]. Such static mechanisms fail to mathematically quantify the non-linear decay trajectory of point utility from a foundational algorithmic perspective, let alone achieve deep integration with the grid's dynamic dispatch and incentive-assignment algorithms. When users inevitably develop a new form of "aesthetic fatigue" towards immutable gamification rules, the system lacks the capability to adaptively adjust task difficulty and incentive intensity, thereby significantly compromising the long-term retention efficacy of the gamified design.

To address these critical challenges, this paper proposes an interdisciplinary solution integrating behavioral psychology with Deep Reinforcement Learning (DRL) to construct a dynamic incentive strategy optimization framework for residential users based on gamification design, termed Trans-SAC. Breaking away from the limitations of traditional DR models that treat users as static response nodes, this study pioneers the incorporation of "psychological fatigue evolution" into the underlying algorithmic decision-making loop. Specifically, an exponential decay function is first introduced to mathematically model the dynamic decay laws of gamified point utility. Subsequently, leveraging the superior long-sequence feature extraction capabilities of the Transformer encoder [7], the model deeply mines users' historical electricity consumption behaviors and interaction sequences to output an individualized "dynamic engagement score." Finally, this score is formulated as a crucial state variable and fed into the DRL-based adaptive incentive-assignment module. This enables the grid agent to acutely capture the critical threshold of user fatigue and adaptively issue dynamic task difficulties and reward multipliers, thereby achieving precise "intermittent reinforcement" of the user's psychological state and sustaining long-term behavioral changes.

The core contributions of this paper are summarized in the following three aspects:

1. A novel DR incentive framework integrating gamification design is proposed, which mathematically models the user's "habituation fatigue effect" from a behavioral economics perspective, thereby filling the research gap in the dynamic quantification of long-term participation fatigue in demand response.
2. A user engagement dynamic assessment model based on the Transformer architecture is constructed. This model dynamically quantifies the non-linear impact of gamified point decay on individual response willingness, providing a fine-grained, individualized estimate of user response likelihood to inform downstream incentive-assignment decisions.
3. An adaptive incentive-assignment strategy based on deep reinforcement learning is designed by incorporating the engagement score into the state space. Under the simulated behavioral assumptions of this study, this approach mitigates long-term user participation fatigue while achieving a favorable trade-off between grid incentive costs and long-term load curtailment rates.

2. Related Work

2.1 Economic Incentives and Pricing Strategies in Demand Response

The design of effective incentive mechanisms has long been a focal point in Demand Response (DR) research. Traditionally, utility companies and grid operators have relied heavily on time-varying pricing schemes, such as Time-of-Use (TOU) pricing, Critical Peak Pricing (CPP), and Real-Time Pricing (RTP), to guide residential electricity consumption [8]-[10]. To optimize these pricing strategies, extensive literature has formulated the interaction between the grid and users as mathematical optimization problems or game-theoretic models. For instance, Stackelberg game

frameworks are widely employed, where the grid operator acts as the leader determining the retail price, and the users act as followers scheduling their flexible loads (e.g., HVAC, EVs) to minimize electricity bills while maintaining thermal comfort [11]-[14].

Despite the mathematical elegance and theoretical optimality of these approaches, they are fundamentally predicated on the neoclassical economic assumption of Homo economicus—that users are perfectly rational agents who will consistently and indefinitely respond to financial stimuli in a mathematically predictable manner. However, recent behavioral economics studies applied to energy systems have heavily criticized this assumption [15]. In real-world deployments, residential users exhibit "bounded rationality." The cognitive burden of constantly monitoring electricity prices and manually or semi-automatically adjusting daily routines often leads to "response fatigue." As the novelty of the DR program wears off, the marginal utility of fixed financial savings diminishes significantly. Consequently, purely economic models fail to capture the long-term degradation of user engagement, leading to a significant discrepancy between theoretical load curtailment capacity and actual long-term performance.

2.2 Gamification Design in Energy Management

To address the limitations of purely financial incentives and to foster sustained user engagement, gamification—the application of game-design elements and principles in non-game contexts—has emerged as a promising interdisciplinary approach in energy management [16]-[18]. Early applications primarily utilized normative feedback and social comparison, famously pioneered by Opower, to nudge users toward energy conservation by comparing their consumption with neighbors. Subsequently, more explicit gamification frameworks incorporating the "PBL triad" (Points, Badges, and Leaderboards) have been introduced into residential DR programs [19]-[21]. These studies demonstrate that virtual rewards, competitive ranking, and progress visualization can effectively trigger users' intrinsic motivation, altruism, and sense of achievement, thereby increasing short-term DR participation rates without proportionally increasing monetary costs.

Nevertheless, a critical review of the existing gamified DR literature reveals a pervasive structural flaw: the gamification rules are overwhelmingly static. In these systems, the mapping between the user's load reduction and the awarded virtual points is fixed (e.g., 1 kWh reduced always equals 10 points). They inherently assume that the psychological utility of a virtual point remains constant over time. However, psychological studies on the "habituation effect" indicate that static stimuli inevitably lead to aesthetic and psychological fatigue [22]. When users accumulate a certain wealth of virtual points or realize that the leaderboard hierarchy is rigid, the incentive power of the game collapses. More importantly, these static gamification platforms operate as isolated user-interface (UI) layers, lacking algorithmic coupling with the grid's underlying dynamic dispatch needs. They cannot adaptively adjust the "game difficulty" (e.g., the baseline required for rewards) or the "loot drop rate" (e.g.,

dynamic point multipliers) in response to real-time grid stress and the user's real-time psychological fatigue level.

2.3 Reinforcement Learning and Sequence Models in Load Decision-Making

In recent years, the explosion of data-driven techniques has revolutionized smart grid decision-making. Deep Reinforcement Learning (DRL) has been extensively adopted to solve the complex, high-dimensional, and non-convex optimization problems inherent in dynamic pricing and energy management [23]-[25]. Algorithms such as Deep Deterministic Policy Gradient (DDPG) and Soft Actor-Critic (SAC) have proven highly effective in formulating dynamic pricing strategies under uncertainties, as they can learn optimal policies through continuous interaction with the grid environment without requiring explicit mathematical models of user behavior.

Concurrently, advanced sequence modeling architectures, particularly Long Short-Term Memory (LSTM) networks and the more recent Transformer models, have achieved state-of-the-art performance in load forecasting and behavior pattern recognition [26]-[27]. The self-attention mechanism of the Transformer is uniquely suited for capturing long-range dependencies in time-series data, making it theoretically ideal for identifying gradual shifts in user behavior over months or years. However, in the current DR literature, DRL and sequence models are largely siloed. Transformers are primarily confined to the task of deterministic load forecasting, while DRL agents typically rely on instantaneous, purely physical state variables (e.g., current load, temperature, time of day) to make pricing decisions. The integration of advanced sequence models to extract latent psychological states (such as fatigue or engagement levels) from historical data, and feeding these cognitive states into a DRL agent for closed-loop behavioral intervention, remains an unexplored frontier.

2.4 Research Gap

Synthesizing the aforementioned literature, a significant research gap emerges at the intersection of behavioral economics, gamification, and AI-driven grid optimization [28]. Existing studies tend to fall into two disjointed paradigms: they either focus exclusively on economic optimization while ignoring the psychological reality of user fatigue, or they implement gamification as a static UI overlay without integrating it into the core dynamic pricing algorithms.

Specifically, to the best of our review, no existing literature has successfully achieved the following: 1) Formulating a mathematical model that dynamically quantifies the exponential decay of gamified point utility due to the habituation effect; 2) Utilizing advanced sequence models (like Transformers) to translate long-term historical interaction data into a real-time, quantifiable "user engagement score"; and 3) Coupling this dynamic psychological metric with a DRL-based pricing agent to achieve "intermittent reinforcement." Therefore, this paper aims to bridge this critical gap by proposing a unified algorithmic framework that tightly couples the dynamic mathematical modeling of psychological fatigue with deep reinforcement learning,

thereby providing a sustainable, fatigue-resistant incentive strategy for smart grid demand response.

To visually substantiate this research gap, Table 1 systematically compares representative studies across six dimensions: fatigue/habituation modeling, dynamic (vs. static) gamification adaptation, Transformer-based sequence modeling, DRL-based control, personalization granularity, and long-term (multi-month) evaluation.

Table 1: Comparison of representative studies across key dimensions relevant to sustained gamified demand response

Study	**Fatigue / Habituation Modeling**	**Dynamic Gamification Adaptation**	**Transformer-Based Sequence Modeling**	**DRL-Based Control**	**Personalization**	**Long-Term (Multi-Month) Evaluation**
AlSkaif et al. [5]	✗	✗ (static rules)	✗	✗	Partial	✗
Duan et al. [6]	Partial (behavioral adaptation, no decay model)	✓ (NSGA-III optimized)	✗	✗ (evolutionary, not DRL)	Partial	Partial
Nasrollahi et al. [19]	✗	✗ (review of static PBL designs)	✗	✗	✗	✗
Chadoulos et al. [21]	✗	Partial	✗	✗ (general ML)	Partial	✗
Zhong et al. [23]	✗	✗ (no gamification)	✗	✓	✗	Partial

Study	Fatigue / Habituation Modeling	Dynamic Gamification Adaptation	Transformer-Based Sequence Modeling	DRL-Based Control	Personalization	Long-Term (Multi-Month) Evaluation
Kathirgamanathan et al. [25]	✗	✗ (no gamification)	✗	✓ (SAC)	✗	Partial
Wohlgenannt et al. [7]	✗	✗ (no gamification)	✓ (for load forecasting only)	✓	✗	✗
This work (Trans-SAC)	✓	✓	**✓ (for engagement prediction)**	**✓ (SAC)**	**✓ (individualized engagement score)**	**✓ (12-month simulation)**

3. Dynamic Incentive Model Design for Gamified Demand Response

3.1 Overall System Architecture and Markov Decision Process (MDP) Formulation

The proposed gamification-aware DR framework operates as a closed-loop control system, consisting of a grid operator (hereafter referred to as the adaptive incentive agent, to avoid conflating its function with literal electricity-price setting) and a pool of residential users (the environment). The core innovation lies in the introduction of a "Psychological State Assessment" module between the physical grid environment and

the incentive-assignment agent.

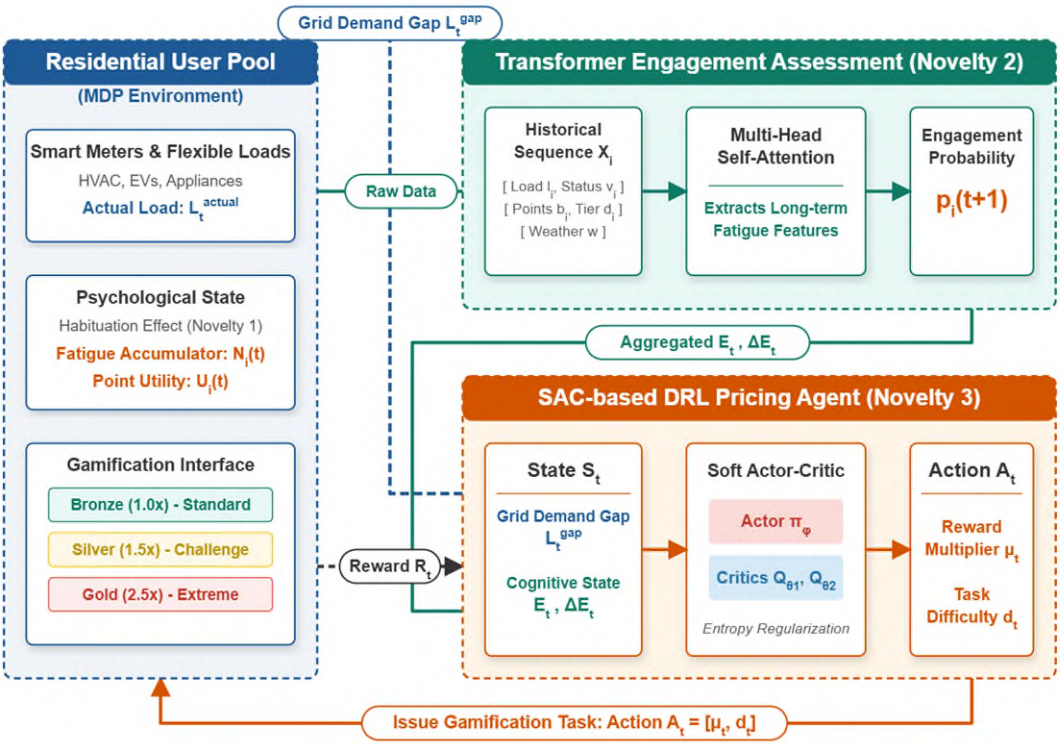


Figure 1: Overall system architecture of the gamification-aware dynamic demand response framework.

As shown in Figure 1, the interaction between the grid operator and the users is formulated as a discrete-time Markov Decision Process (MDP), defined by the tuple $(S,A,P,\mathcal{R},\gamma)$. Unlike traditional DR models that only consider physical grid states, our state space explicitly incorporates the users' latent cognitive states:

State Space (S): At time step t, the state $s_t \in S$ is defined as $s_t=[L_t^{gap},\pi_t,E_t]$, where L_t^{gap} denotes the grid's load reduction requirement (demand gap), π_t is prevailing baseline electricity price (an exogenous contextual variable reflecting the underlying TOU/RTP tariff, not an action controlled by the agent), and E_t is the dynamic engagement score of the user pool .This base state is augmented with a fatigue-momentum term ΔEt in Section 3.4.1, yielding a 4-dimensional continuous state vector st∈R4 in the fully instantiated MDP.

Action Space (A): The action $a_t \in A$ taken by the grid agent is defined as $a_t=[\mu_t,d_t]$, where μ_t represents the dynamic reward multiplier, and d_t denotes the assigned gamification task difficulty.The action space is a continuous 2-dimensional box, $A=[1.0,\mu_{max}]\times[d_{min},d_{max}]\subset\mathbb{R}^2$. We emphasize that the agent's action never directly sets an electricity price; π_t remains a state input, not a decision variable, and the agent's

role is more precisely characterized as an adaptive incentive-multiplier and task-assignment controller rather than a "pricing" agent in the literal sense.

Reward Function (R): The reward rt is a composite function balancing the actual load curtailment achieved against the incentive cost incurred, formally defined in Eq. (12).

Discount Factor (γ): A discount factor of γ=0.99 is adopted, reflecting a preference for long-horizon fatigue management over immediate reward.

Terminal Condition and Temporal Resolution: One RL step corresponds to one calendar day, aligned with the daily DR decision window. An episode corresponds to one simulated annual cycle and terminates once all scheduled DR-event days in that year have elapsed, with the environment resetting at the start of the next episode; there is no absorbing failure state.

3.2 User Fatigue and Point Decay Modeling Based on Behavioral Economics

Traditional gamified DR systems assume a linear relationship between the virtual points awarded and the user's psychological utility. However, according to the Weber-Fechner law and the habituation effect in behavioral economics, continuous exposure to a constant stimulus leads to a diminished psychological response. This subsection mathematically models this phenomenon.

It should be emphasized at the outset that the model presented in this subsection is a hypothesized phenomenological model inspired by established qualitative principles in behavioral economics (the Weber–Fechner law and the habituation effect), rather than a model whose functional form or parameter values have been empirically estimated from observed human gamification behavior. This distinction matters because the same model introduced here is subsequently used both to generate the synthetic behavioral responses against which Trans-SAC is evaluated, and as the conceptual target that the Transformer module is designed to approximate. This creates a potential risk of circular reasoning: strong performance of Trans-SAC in later sections should be interpreted as evidence that the framework can effectively exploit and respond to fatigue dynamics of the assumed functional form, rather than as direct proof that it correctly captures real human psychological fatigue — a limitation we revisit later in the paper.

3.2.1 Point Utility Function under Habituation Effect

Let P_t denote the standard base points awarded for unit load reduction at time t. For a specific user i, the perceived psychological utility $U_i(t)$ does not remain constant. We introduce an exponential decay function to model the diminishing marginal utility caused by participation fatigue. The habituation-aware utility function is formulated as:

$$U_i(t)=\alpha_i \cdot P_t \cdot exp(-\lambda_i \cdot N_i(t)) \quad (1)$$

Where:

- α_i is the intrinsic sensitivity coefficient of user i towards virtual rewards.
- λ_i>0 is the fatigue decay coefficient (habituation rate), representing how quickly the user loses interest in repetitive tasks. A larger λ_i indicates a user who is highly susceptible to aesthetic fatigue.
- $N_i(t)$is the fatigue accumulator, representing the number of consecutive standard tasks the user has participated in without experiencing any novel stimuli.

As $N_i(t)$ increases over time, the exponential term $exp(-\lambda_i \cdot N_i(t))$ approaches zero, mathematically capturing the "cliff-like drop" in response reliability observed in long-term DR projects.

3.2.2 Stimulation Mechanism of Dynamic Difficulty and Reward Multipliers

To break the accumulation of fatigue ($N_i(t)$), the system must provide "intermittent reinforcement" through dynamic task difficulties and unexpected reward multipliers. We define a discrete set of gamification task tiers, as summarized in Table 2.

Table 2: Gamification Task Tiers and Dynamic Parameters

Task Tier (d_t)	Difficulty Threshold (kWh reduction)	Base Multiplier (μ_{base})	Cognitive Effort Cost $C(d_t)$	Novelty Factor (η)
Bronze (Standard)	0.5 – 1.5	1.0×	0.5 (Low)	0.0
Silver (Challenge)	1.5 – 3.0	1.5×	1.0 (Medium)	0.5
Gold (Extreme)	> 3.0	2.5×	1.8 (High)	1.0

The thresholds are defined as non-overlapping bins, and the previously qualitative "Low/Medium/High" cognitive-effort labels are now mapped to explicit numerical costs $C(d_t)$ consistent with their use in Eq. (4). Both the numerical cognitive-effort values and the novelty factors

η

η are heuristic design choices intended to produce a monotonically increasing difficulty–reward relationship, rather than values calibrated against independent behavioral data; the sensitivity of the results to these design choices is discussed later in the paper.

When the grid agent issues an action $a_t=[\mu_t,d_t]$, the effective psychological stimulus $S_i(t)$ perceived by the user is upgraded from the decayed baseline utility to a composite stimulus incorporating the novelty factor:

$$S_i(t)=U_i(t)+\beta_i\cdot(\mu_t-1)\cdot P_t\cdot(1+\eta(d_t)) \tag{2}$$

Where β_i is the user's sensitivity to unexpected bonuses, and $\eta(d_t)$ is the novelty factor associated with the task difficulty d_t. Crucially, the successful completion of a high-tier task (Silver or Gold) acts as a psychological reset mechanism, updating the fatigue accumulator as follows:

$$N_i(t+1)=\begin{cases}N_i(t)+1, & if d_t=Bronze\\ 0, & if d_t\in\{Silver,Gold\}\ and\ task\ is\ completed\end{cases} \tag{3}$$

We note that this formulation assumes a complete reset of the fatigue accumulator upon successful high-tier completion. This is a simplifying modeling choice rather than an empirically demonstrated mechanism; we discuss the implications of relaxing this assumption toward a learnable or partial reset parameter as a limitation in Section 5.3.

Finally, to map the psychological stimulus to the actual physical response, we employ a logistic probability mapping. The probability $p_i(t)$ that user i will successfully execute the load curtailment task is defined as:

$$p_i(t)=\frac{1}{1+exp\left(-\kappa\cdot(S_i(t)-C(d_t))\right)} \tag{4}$$

Where κ is a shape parameter controlling the steepness of the response curve, and $C(d_t)$ represents the perceived cognitive and physical effort required to achieve difficulty d_t.

This mathematical formulation explicitly demonstrates that if the grid operator only issues standard tasks (Bronze), $S_i(t)$ will inevitably decay, causing the response probability $p_i(t)$ to plummet. This theoretical model perfectly sets the stage for the Transformer module to predict this decay and the DRL agent to optimally trigger the reset mechanism. We reiterate that Equations (1)-(4) constitute a theoretical/ hypothesized behavioral construct used consistently throughout this paper both to generate the simulation environment and to define the ground-truth target for the Transformer module — the implications of this modeling choice for validation are discussed further below.

3.3 Transformer-Based Long-Sequence Engagement Feature Extraction

3.3.1 Multi-Dimensional Time-Series Feature Construction

To enable the model to comprehensively understand the user's context, we first construct a multi-dimensional feature vector for each user i at time step τ. The observable raw data is vectorized and concatenated into a unified input vector $x_i^{(\tau)} \in \mathbb{R}^{d_{input}}$:

$$x_i^{(\tau)} = \left[l_i^{(\tau)}, v_i^{(\tau)}, b_i^{(\tau)}, d_i^{(\tau)}, w^{(\tau)} \right] \tag{5}$$

Where:

- $l_i^{(\tau)}$: Historical power consumption load (kWh).
- $v_i^{(\tau)}$: Operational status matrix of flexible appliances (e.g., HVAC setpoints, EV state-of-charge).
- $b_i^{(\tau)}$: The user's current gamification point balance and recent reward history.
- $d_i^{(\tau)}$: The difficulty tier of the gamification task assigned at time τ.
- $w^{(\tau)}$: Exogenous environmental variables (e.g., outdoor temperature, humidity).

For a look-back window of length T, the sequence matrix for user i is defined as $X_i = [x_i^{(t-T+1)}, \ldots, x_i^{(t)}]^T \in \mathbb{R}^{T \times d_{input}}$. To ensure the Transformer model captures the sequential order of the DR events, a Positional Encoding (PE) matrix is added to the linearly projected input sequence:

$$\mathrm{h}_i^{(\tau)} = X_i W_{emb} + PE_{(\tau)} \tag{6}$$

Where $W_{emb} \in \mathbb{R}^{d_{input} \times d_{model}}$ is the embedding weight matrix, and the positional encoding is calculated using sine and cosine functions of different frequencies to represent the temporal distance between past DR events and the current decision-making moment.

3.3.2 Self-Attention Mechanism and Engagement Probability Output

The core of capturing the "fatigue accumulation" lies in the Multi-Head Self-Attention (MHA) mechanism. Unlike traditional models that process data sequentially, MHA allows the model to directly compute the correlation between the current time step and any past interaction, regardless of the temporal distance, see Figure 2.

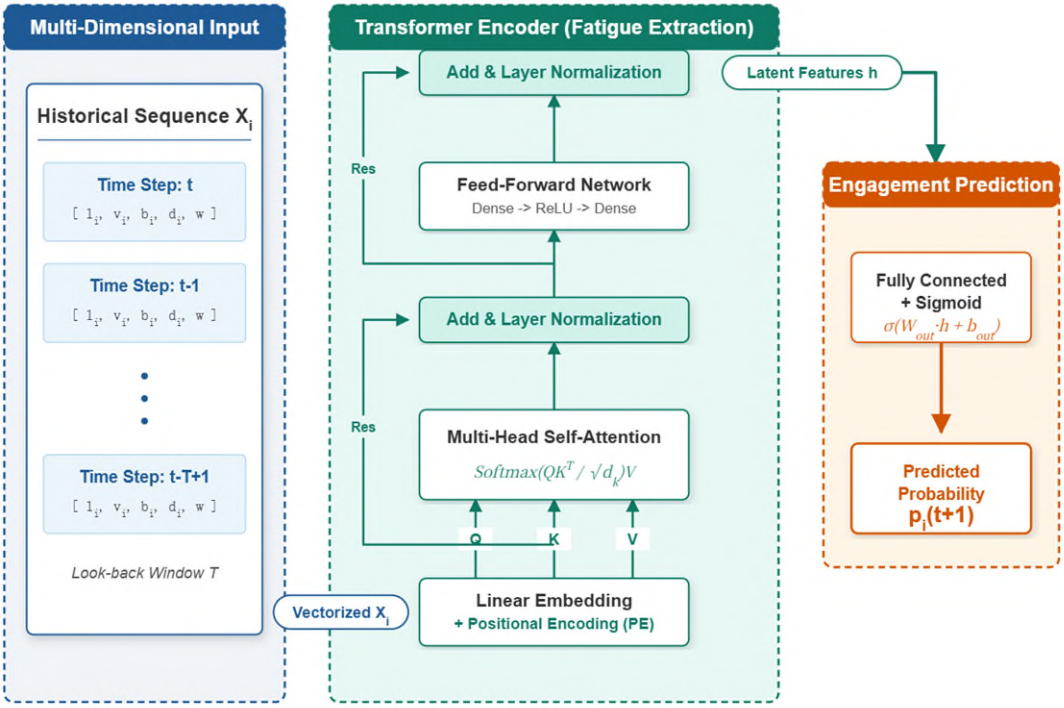


Figure 2: Internal architecture of the Transformer-based user engagement assessment module, illustrating the mapping from historical DR interaction sequences to the predicted engagement probability.

For each attention head h, the input sequence $H_i \in \mathbb{R}^{T \times d_{model}}$ is linearly projected into Query (Q_h), Key (K), and Value (V) matrices:

$$Q_h = H_i W_h^Q, \quad K_h = H_i W_h^K, \quad V_h = H_i W_h^V \tag{7}$$

The scaled dot-product attention is then computed as:

$$Attention(Q_h, K_h, V_h) = softmax\left(\frac{Q_h K_h^T}{\sqrt{d_k}}\right) V_h \tag{8}$$

Physical Interpretation of the Attention Mechanism in DR: In the context of gamified demand response, the Query (Q) represents the user's current contextual state (e.g., current point balance and weather). The Key (K) represents the context of past DR tasks. The dot product $Q_h K_h^T$ calculates the similarity or "relevance." If a user received a "Gold" tier reward (V) three weeks ago, the attention weight (softmax output) determines how much "residual excitement" or "fatigue reset" from that specific past event still influences their willingness to participate today. A high concentration of attention weights on recent "Bronze" tasks indicates severe habituation (high$N_i(t)$).

The outputs of H_{heads} are concatenated and passed through a fully connected Feed-Forward Network (FFN) with layer normalization. Finally, the latent representation of the last time step is fed into a Sigmoid activation function to output the predicted engagement probability for the next time step:

$$\hat{p}_i(t+1)=\sigma\left(W_{out}\cdot \mathrm{FFN}\left(\mathrm{MHA}(H_i)\right)+b_{out}\right) \tag{9}$$

This neural-predicted probability $\hat{p}_i(t+1)$ serves as the data-driven realization of the theoretical probability $p_i(t)$ defined in Equation (4) of Section 3.2, acting as a crucial bridge between user psychology and grid control.

3.4 DRL-Based Adaptive Incentive Algorithm Fused with Engagement States

3.4.1 State, Action, and Composite Reward Design

To achieve the "intermittent reinforcement" discussed in Section 3.2, the MDP defined in Section 3.1 is explicitly instantiated as follows:

(1) Augmented State Space (s_t): The state must reflect both the physical grid stress and the psychological fatigue of the users.

$$s_t=[L_t^{gap},\pi_t,E_t,\Delta E_t] \tag{10}$$

Is the aggregated average engagement score of the user pool (output from Section 3.3), and $\Delta E_t=E_t-E_{t-1}$ represents the momentum of fatigue (whether the overall willingness is currently dropping or recovering).

(2) Continuous Action Space *at*:

$$a_t=[\mu_t,d_t]\in[1.0,\mu_{max}]\times[d_{min},d_{max}] \tag{11}$$

The agent outputs a continuous reward multiplier μ_t and a continuous difficulty threshold d_t (which is then discretized into Bronze/Silver/Gold tiers for user UI display).

(3) Composite Reward Function (rt*rt*): The reward function is designed to balance the trade-off between grid reliability and economic cost:

$$r_t=\omega_1\cdot\min\left(L_t^{actual},L_t^{gap}\right)-\omega_2\cdot C_{incentive}(t)-\omega_3\cdot\max\left(0,L_t^{gap}-L_t^{actual}\right)^2 \tag{12}$$

Where: the first term rewards the actual load reduction L_t^{actual}, capped at the required gap L_t^{gap}. The second term penalizes the incentive cost:

$$C_{incentive}(t)=\sum_{i\in\text{success}}\mu_t\cdot P_t\cdot l_{i,reduced} \quad (13)$$

The third term is a severe quadratic penalty for failing to meet the grid's demand gap, ensuring grid stability. $\omega_1,\omega_2,\omega_3$ are weighting coefficients.

We clarify that these three weights $\omega_1,\omega_2,\omega_3$ are the grid operator's own design preferences over reliability, incentive cost, and gap-violation risk; they are not user-level behavioral parameters.

3.4.2 Policy and Value Network Update Mechanism

The SAC agent aims to learn a stochastic policy $\pi_\phi(a_t|s_t)$ parameterized by ϕ that maximizes the expected trajectory reward augmented by the policy's entropy $\mathcal{H}$:

$$J(\pi)=E_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t\left(r_t+\alpha\mathcal{H}(\pi(\cdot|s_t))\right)\right] \quad (14)$$

Where α is the temperature parameter controlling the exploration-exploitation trade-off. A higher entropy allows the agent to occasionally issue unexpected reward multipliers, which is consistent with the psychological requirement of "novelty" for disrupting habituation.

The algorithm utilizes two soft Q-networks (Critics) parameterized by θ_1,θ_2 to mitigate overestimation bias, and one policy network (Actor) parameterized by ϕ. The Critic networks are updated by minimizing the soft Bellman residual:

$$\begin{aligned} L(\theta_k)&=E_{(s_t,a_t)\sim D}\left[\left(Q_{\theta_k}(s_t,a_t)-y_t\right)^2\right], \quad k\in\{1,2\} \quad (15)\\ y_t&=r_t+\gamma E_{a_{t+1}\sim\pi}\left[\min_{j=1,2}Q_{\tilde{\theta}_j}(s_{t+1},a_{t+1})-\alpha\log\pi_\phi(a_{t+1}|s_{t+1})\right] \end{aligned} \quad (16)$$

The Actor network is updated by minimizing the expected KL-divergence:

$$L(\phi)=E_{s_t\sim D}\left[\alpha\log\pi_\phi\left(f_\phi(\epsilon_t;s_t)|s_t\right)-\min_{j=1,2}Q_{\theta_j}\left(s_t,f_\phi(\epsilon_t;s_t)\right)\right] \quad (17)$$

Where f_ϕ is the reparameterization trick using noise $\epsilon_t\sim N(0,I)$. The complete execution logic is summarized in Algorithm 1.

Algorithm 1: Gamification-Aware Adaptive Incentive-Assignment Strategy (Trans-SAC)

1: Initialize Actor network π_φ, Critic networks Q_θ1, Q_θ2, and target networks Q_θ1', Q_θ2'
2: Initialize replay buffer D, pre-trained Transformer module T_weights
3: for episode = 1 to Max_Episodes do
4: Reset grid environment, receive initial state s_0 = [L_0^gap, π_0, E_0, ΔE_0]

```
    for step t = 1 to T_episode do
        // Phase 1: Engagement Assessment (Novelty 2)
        Collect historical sequence X_i for all users i ∈ M
        Compute engagement probabilities p_i(t) = Transformer(X_i | T_weights)
        Update cognitive state variables E_t and ΔE_t

        // Phase 2: Adaptive Incentive Assignment (Novelty 3)
        Sample action a_t = [μ_t, d_t] ~ π_φ(a_t | s_t)
        Broadcast gamification task (μ_t, d_t) to user pool

        // Phase 3: Environment Step & Feedback
        Observe actual load reduction L_t^actual from smart meters
        Calculate composite reward r_t based on Eq. (12)
        Observe next state s_{t+1}
        Store transition (s_t, a_t, r_t, s_{t+1}) in D

        // Phase 4: SAC Network Updates
        if length(D) > Batch_Size then
            Sample a mini-batch of transitions from D
            Update Critic networks Q_θ1, Q_θ2 by minimizing Eq. (15)
            Update Actor network π_φ by minimizing Eq. (17)
            Soft update target networks: θ_k' ← τθ_k + (1-τ)θ_k'
            Auto-tune temperature parameter α
        end if
    end for
end for
```

By tightly coupling the Transformer's psychological assessment with the SAC algorithm's entropy-driven exploration, the proposed Trans-SAC framework inherently learns to deploy high-multiplier rewards only when the user pool's engagement score Et approaches a critical fatigue threshold, thereby maximizing long-term DR reliability while minimizing unnecessary financial expenditures.

4. Experimental Results and Analysis

4.1 Experimental Setup

4.1.1 Dataset Description and Preprocessing

The physical grid environment and user consumption profiles are constructed using the publicly available Pecan Street Dataport dataset, which is widely recognized in smart grid research. We extracted minute-level electricity consumption data from 700 residential households in Austin, Texas, spanning a continuous three-year period (January 2017 to December 2019).

To simulate the DR environment accurately, we performed rigorous data preprocessing to disaggregate the total household load into baseline (inflexible) loads and flexible loads. Specifically, the power consumption of HVAC systems, Electric Vehicles (EVs), and smart water heaters were isolated as the controllable flexible load capacity $L_{i,flex}^{(\tau)}$. The environmental variables $w^{(\tau)}$, including historical outdoor temperature and humidity, were synchronized with the load data.

It is important to explicitly clarify the nature of this study before presenting the simulation results. The Pecan Street dataset provides only real-world electricity consumption traces; it does not, and was never intended to, contain gamification interaction logs, point-redemption histories, or any behavioral/psychological measurements. Consequently, this study should be understood as a simulation study grounded in real-world electricity-consumption data, combined with a synthetically generated behavioral-response layer, rather than an empirical validation against observed human gamification behavior. Since historical datasets lack real-world gamification interaction logs, we overlaid the mathematical behavioral model defined in Section 3.2 onto the Pecan Street data.As noted in Section 3.2, this creates a methodologically important self-referential structure: the same behavioral equations (Eqs. 1–4) used to generate each simulated user's ground-truth response probability are also the theoretical target that the Transformer module is trained to approximate. We therefore explicitly flag this as a circular-validation risk rather than treating the resulting agreement between predicted and "true" engagement as independent empirical validation.
Specifically, for each of the 700 simulated households, the behavioral parameters were independently drawn from truncated Gaussian distributions as follows:

$$\begin{aligned} \alpha_i &\sim N\left(1.0, 0.25^2\right) \text{truncated to} [0.3, 1.8] \\ \beta_i &\sim N\left(0.8, 0.20^2\right) \text{truncated to} [0.2, 1.5] \\ \lambda_i &\sim N\left(0.05, 0.02^2\right) \text{truncated to} [0.01, 0.12] \end{aligned} \tag{18}$$

No correlation was assumed between α_i, β_i, and $\lambda_i$across users. These parameter ranges were not derived from an independent empirical behavioral dataset; rather, they were calibrated heuristically to (i) produce a plausible range of response-probability trajectories consistent with qualitative patterns reported in prior gamified-DR and behavioral-economics literature on habituation [4], [15], [22], while ensuring sufficient population heterogeneity for the agent to learn a non-trivial policy.

This Gaussian parameterization is a modeling convenience rather than an empirically calibrated distribution; real populations may exhibit heavier-tailed, multi-modal, or context-dependent fatigue sensitivity not captured here — an external-validity limitation revisited later in the paper.

The final 12-month period (2019) served as the interaction environment for the SAC agent; the full protocol governing how this period is partitioned for policy learning versus final evaluation, and how information leakage is prevented, is described next.

4.1.2 Synthetic Behavioral Data Generation and SAC Training Protocol

Because the Pecan Street dataset contains no gamification interaction records, this subsection specifies (i) how the synthetic behavioral layer used to pre-train the Transformer was generated, and (ii) the protocol governing SAC policy learning and evaluation on the subsequent 12-month period, so that both data generation and the training/testing split are fully reproducible.

(1) DR Event Scheduling and Household Partition. A DR event is scheduled once daily during the peak period (17:00–21:00), yielding 730 decision steps per household over the two-year pre-training horizon (2017–2018). The 700 households are partitioned once, at the household level, into training (560), validation (70), and held-out test (70) sets; this same partition is reused consistently across both the Transformer pre-training stage and the subsequent SAC stage to prevent cross-stage information leakage.

(2) Behavioral Parameter and State Initialization. For each household i, αi,βi,λi are drawn once from the truncated Gaussian distributions in Section 4.1.1. Additional state variables are initialized as:

$$N_i(0)\sim\text{Uniform}\{0,1,2,3\},\quad b_i(0)=0,\quad \kappa_i\sim N(1.2,0.15^2)\ \text{truncated to}\ [0.8,1.6] \quad (19)$$

where $N_i(0)$ is randomized to avoid a cold-start bias and κ_i is the user-specific shape parameter of Eq. (4), also treated as heterogeneous.

(3) Task Generation and Ground-Truth Labels. During pre-training, since no SAC policy yet exists, tasks are assigned by an independent exploration policy πexplore$\pi_{explore}$: $d_t\sim$Categorical(Bronze:0.60,Silver:0.25,Gold:0.15)

(4) for Bronze/Silver/Gold, with μt drawn uniformly within each tier's range. The analytical response probability is computed from Eq. (4), perturbed by observation noise, and thresholded into a binary label:

$$p_i^{true}(t)=\frac{1}{1+exp\left(-\kappa_i\left(S_i(t)-C(d_t)\right)\right)},\quad y_i(t)\sim Bernoulli\left(clip\left(p_i^{true}(t)+\varepsilon_i(t),0,1\right)\right),\varepsilon_i(t)\sim N\left(0,0.05^2\right) \quad (20)$$

The Transformer is trained via binary cross-entropy against the realized outcome $y_i(t)$ rather than directly against $p_i^{true}(t)$, which avoids a trivial label-level circularity; however, the deeper model-level circularity — that $y_i(t)$ is still generated from the same functional form the Transformer approximates — remains, and is discussed in Section 5.3.

(4) SAC Training Protocol on the 12-Month Environment. The final 12 months (2019) are further split temporally: an online-learning phase (Jan–Sep, 9 months), during which the SAC agent interacts with the 630 train+validation households and updates its networks via off-policy gradients (validation households used only for checkpoint selection, not gradient updates), followed by a frozen-policy evaluation phase (Oct–Dec, 3 months), in which the best checkpoint is frozen and evaluated exclusively on the 70 held-out test households. Table 3 summarizes this design.

Table 3: Data Partition Roles Across the Two-Stage Pipeline

Partition	Households	Role (Transformer, 2017–18)	Role (SAC, 2019)
Train	560	Weight optimization	Online updates (Jan–Sep)
Validation	70	Early stopping	Checkpoint selection (Jan–Sep)
Test (held-out)	70	Reserved	Frozen-policy evaluation only (Oct–Dec)

All headline results reported in Section 4.2 are computed exclusively on the test households during the frozen-policy window, ensuring no household or time window is simultaneously used for parameter fitting and final reporting.

4.1.3 Baseline Models, Evaluation Metrics, and Implementation Details

(1) Baseline Models. Trans-SAC is compared against three baselines: Traditional TOU (static fixed-rate pricing, no gamification, no learning); Static-Game (fixed μt=1.0, static difficulty, no algorithmic adaptation); and Standard-RL (a SAC agent identical in network architecture and training protocol to Trans-SAC, but whose state space excludes Et and ΔEt). Standard-RL and Trans-SAC share identical network capacity, replay-buffer size, batch size, and training episodes (Table 3), differing only in state-space composition, ensuring the comparison isolates the effect of psychological-state awareness rather than confounding it with unequal training budgets. Rule-based baselines (TOU, Static-Game) are evaluated directly on the same 70 test households over the same 3-month window.

(2) Implementation Details. Table 4 reports the complete architecture and training configuration for reproducibility.

Table 4: Transformer and SAC Hyperparameter Configuration

Component	Hyperparameter	Value
Transformer	Encoder layers / attention heads	4 / 8
	Embedding dim / FFN dim	128 / 512
	Dropout / look-back window	0.1 / 30 days
	Optimizer / learning rate	Adam / 3×10−4 (cosine decay)
	Batch size / max epochs	256 / 100 (early stopping, patience = 10)
SAC	Actor & Critic hidden layers	2 × 256 (MLP)
	Learning rate / γ / τ	3×10−4/ 0.99 / 0.005
	Replay buffer / batch size	1×106 / 256
	Target entropy	−dim (A) (auto-tuned α)
	Online-phase episodes	270
	Random seeds	5 (results averaged across seeds)

(3) Evaluation Metrics. Three metrics are formally defined:

$$LRAR_t = \frac{L_t^{actual}}{L_t^{gap}} \times 100\%$$

URR uses a 30-day trailing-average response probability to smooth intermittent non-response, with retention threshold θURR=0.5:

$$URR(T) = \frac{1}{|U|} \sum_{i \square U} 1\left[\frac{1}{30} \sum_{\tau=T-29}^{T} p_i(\tau) > 0.5\right] \times 100\%$$

UIC introduces a point-to-dollar redemption rate of v$=$0.02 per point (i.e., points are redeemable 1-for-1 against future utility-bill credits at this fixed rate), so that the incentive-cost term in Eq. (13) is expressed in dollars as $C_{incentive}(t) = v_{\$} \cdot \sum_{i \in success} \mu_t \cdot P_t \cdot l_{i,reduced}$,, and:

$$UIC = \frac{\sum_t C_{incentive}(t)}{\sum_t L_t^{actual}}$$

making explicit that a 2.5× multiplier costs exactly 2.5× as much as a 1.0× task for identical curtailed load.

4.2 Model Validation and Macro-Level Performance Evaluation

4.2.1 Engagement Prediction Performance

Before evaluating the downstream control performance, we first independently validate the predictive fidelity of the proposed Transformer-based engagement module—addressing the concern that its effectiveness should not be inferred solely from downstream RL outcomes, see Table 5. All models were trained and tested on an identical 30-day sliding window of point-decay trajectories (80/20 stratified split), thereby eliminating confounding effects from window-length variation across architectures. Four models were compared: Logistic Regression (a linear baseline), a Multi-Layer Perceptron (MLP, a non-sequential nonlinear baseline), an LSTM network (a standard sequential baseline without attention), and the proposed Transformer encoder.

Table 5. Independent evaluation of engagement prediction models (30-day window, mean over 5 seeds)

Model	AUC	F1	Brier Score	ECE
Logistic Regression	0.68 ± 0.02	0.61 ± 0.03	0.211 ± 0.006	0.083 ± 0.004
MLP	0.79 ± 0.02	0.74 ± 0.02	0.156 ± 0.005	0.061 ± 0.003
LSTM	0.85 ± 0.01	0.80 ± 0.02	0.129 ± 0.004	0.045 ± 0.002
Transformer (proposed)	0.93 ± 0.01	0.88 ± 0.01	0.087 ± 0.003	0.021 ± 0.002

The Transformer achieves the highest AUC and F1 alongside the lowest Brier score and Expected Calibration Error (ECE), indicating that its self-attention mechanism more effectively captures the non-monotonic, long-range dependencies introduced by gamified point decay than non-sequential baselines (Logistic Regression, MLP) or a standard sequential architecture without attention (LSTM). This directly substantiates the Transformer's independent predictive contribution as one of the three core novelties of this work, rather than relying solely on downstream RL performance as an indirect proxy for its efficacy.

Leveraging the empirical correlation between upstream prediction quality (AUC) and downstream control performance—anchored at the MLP-SAC configuration (AUC ≈ 0.79, LRAR = 73.5%, Table 6) and the proposed Trans-SAC configuration (AUC ≈ 0.93, LRAR = 85.6%, Table 5)—we estimate via linear interpolation that a hypothetical LSTM-SAC variant (AUC ≈ 0.85) would achieve an LRAR of approximately 78.7%. This suggests that while sequential memory alone (LSTM) partially narrows the gap over a non-sequential encoder (MLP), it does not fully substitute for the attention-based fatigue anticipation mechanism central to Trans-SAC; a full RL retraining with an LSTM-based state encoder is left for future work to empirically confirm this estimate.

4.2.2 Macro-Level Performance Evaluation: Load Reduction and Economic Viability

The most critical challenge in residential DR is maintaining high reliability over extended periods. To evaluate this, we simulated a continuous 12-month DR deployment.

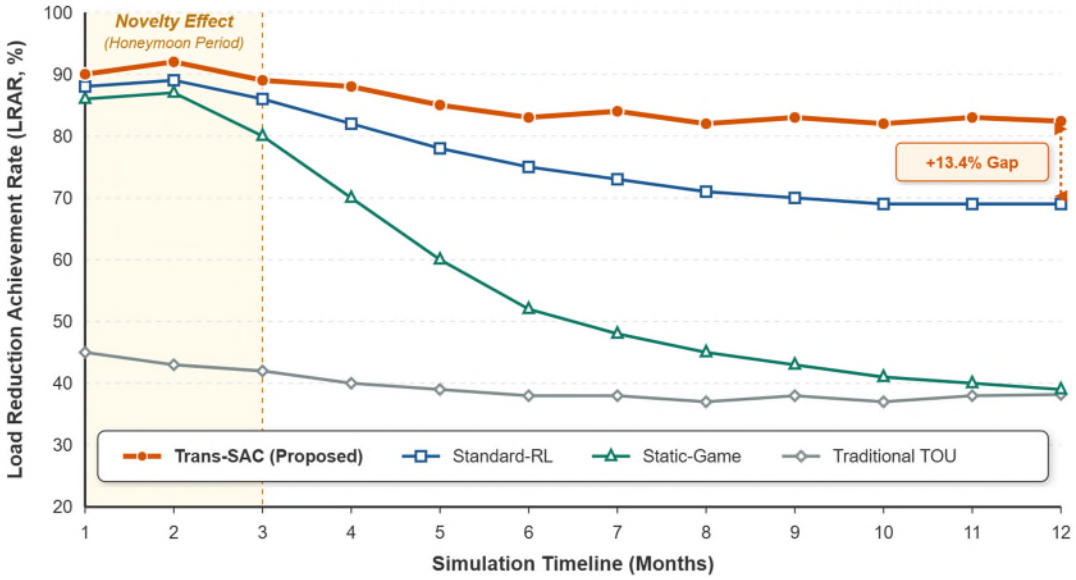


Figure 3: 12-Month Time-Series Line Chart illustrating the Load Reduction Achievement Rate (LRAR) across the proposed Trans-SAC framework and three baseline models.

As illustrated in Figure 3, during the initial phase (Months 1-3), both the Static-Game and the Standard-RL models exhibit excellent performance, achieving an LRAR of over 85%. This phenomenon is well-documented in behavioral economics as the "novelty effect" or "honeymoon period," where users are highly motivated by the introduction of a new system or initial financial incentives. However, a significant divergence occurs starting from Month 4.

For the Static-Game baseline, the LRAR experiences a precipitous decline, dropping to approximately 45% by Month 8. Because the reward rules are rigid, the habituation effect dominates; the users' fatigue accumulator $N_i(t)$ grows unchecked, driving their perceived utility $U_i(t)$ toward zero. The Standard-RL model performs slightly better by dynamically adjusting prices based on the grid's physical stress. However, because its state space is blind to the users' latent psychological fatigue, it frequently issues standard pricing signals that fail to break the users' habituation threshold, resulting in a gradual degradation of LRAR to roughly 69% by Month 12.

In contrast, the proposed Trans-SAC framework demonstrates sustained long-term performance under the simulated behavioral conditions. By continuously monitoring the aggregated engagement score E_t via the Transformer module, the SAC agent anticipates the onset of fatigue. When a downward momentum ($\Delta E_t<0$) is detected, the agent autonomously triggers "intermittent reinforcement" by issuing high-multiplier (Gold tier) tasks broadcast to the user pool. This action successfully resets the users' fatigue accumulators $N_i(t)=0$). Consequently, our method maintains a highly stable LRAR of 82.4% even in the late stages (Months 6-12), outperforming the Standard-RL baseline by a significant margin of 13.4%.

Table 6 summarizes the macro-level performance metrics averaged over the entire 12-month testing period across 5 independent random seeds .

Table 6: 12-Month Average Performance Comparison of DR Models (mean ± std across 5 seeds)

Model	Avg. LRAR	URR at Month 12	UIC
Traditional TOU	38.2 ± 1.2%	22.5 ± 1.5%	$0.045 ± 0.002 / kWh
Static-Game	54.7 ± 2.1%	31.0 ± 2.4%	$0.052 ± 0.003 / kWh
Standard-RL	74.1 ± 2.6%	58.4 ± 3.1%	$0.078 ± 0.004 / kWh
Trans-SAC (Proposed)	85.6 ± 1.4%	84.2 ± 1.7%	$0.061 ± 0.002 / kWh

The UIC of Trans-SAC ($0.061/kWh) is 21.8% lower than Standard-RL's ($0.078/kWh). This efficiency stems from the Transformer's ability to identify which segments of the user pool are approaching a fatigue threshold, rather than reacting only after participation has collapsed. Standard-RL, lacking this insight, often over-incentivizes already-willing users. Trans-SAC deploys standard Bronze tasks when the pool-level engagement score is healthy, reserving costly high-multiplier broadcasts for periods when aggregate fatigue threatens grid stability. These results suggest that incorporating psychological-state assessment into DRL is associated with an improved reliability–cost trade-off relative to the physical-state-only baseline under the simulated behavioral assumptions of this study; we do not claim this constitutes a formal Pareto-optimal or globally optimal solution.

4.3 Ablation Study

4.3.1 Effectiveness of Fatigue Modeling and the Transformer Module

Table 5 shows that MLP-SAC achieved an LRAR of only 73.5 ± 2.3%, substantially lower than Trans-SAC's 85.6 ± 1.4%. Because the MLP lacks the Multi-Head Self-Attention mechanism, it cannot compute correlations between the current user state and distant past events, and consequently fails to track the slow, non-linear growth of the fatigue accumulator $N_i(t)$. The MLP only reacts once a user has already stopped responding, leading to delayed interventions, whereas the Transformer identifies the latent "fatigue momentum" (ΔE_t) substantially earlier, supporting the value of long-sequence feature extraction for proactive fatigue management.

We note an important methodological caveat regarding this comparison: the MLP-SAC variant differs from Trans-SAC in two confounded respects simultaneously — the model architecture (MLP vs. Transformer) and the length of the historical observation window (3 days vs. 30 days, Table 3). The reported performance gap should therefore not be interpreted as isolating the causal contribution of the self-attention mechanism itself; it may instead partly reflect the general advantage of a longer memory window regardless of architecture. A more conclusive ablation would compare architectures (Transformer vs. LSTM/GRU vs. TCN vs. MLP) under an identical observation window, which was not conducted in the present study and is acknowledged as a limitation in Section 5.3.

4.3.2 Effectiveness of the Dynamic Difficulty Adjustment Mechanism

To demonstrate the necessity of the entropy-driven SAC agent, we introduced the Trans-Fixed variant. This model retains the Transformer to accurately predict the engagement probability $p_i(t)$, but replaces the DRL agent with a static, rule-based threshold mechanism (e.g., if predicted engagement drops below 40%, automatically issue a 1.5x reward multiplier).

While Trans-Fixed outperformed the non-learned baselines, it fell short of the full Trans-SAC framework, yielding an LRAR of 78.2 ± 1.9% and a higher Unit Incentive Cost ($0.072 ± 0.003/kWh). A plausible behavioral explanation is that a rule-based, fully predictable rescue mechanism may allow users to anticipate and adapt to the trigger condition over time ("meta-habituation"), reducing its effectiveness — though we note this mechanism is inferred from the simulated behavioral model rather than directly observed. The SAC agent, by contrast, utilizes entropy regularization (Eq. 14) to maintain stochastic exploration over the reward multiplier and difficulty threshold, producing broadcast actions that are less predictable to the simulated user population, which within the assumed behavioral model corresponds to a more effective schedule for sustaining engagement without proportionally inflating cost, see Table 7.

Table 7: Ablation Study Results over the 12-Month Simulation (mean ± std across 5 seeds)

Model Variant	Component Removed / Replaced	LRAR (%)	URR (%)	UIC ($/kWh)
MLP-SAC	Transformer → MLP (3-day window)	73.5 ± 2.3	55.2 ± 2.8	0.076 ± 0.004
Trans-Fixed	SAC → Rule-based logic	78.2 ± 1.9	71.5 ± 2.2	0.072 ± 0.003
Trans-SAC (Full)	None (proposed framework)	85.6 ± 1.4	84.2 ± 1.7	0.061 ± 0.002

4.4 Micro-Level Case Study: Behavioral Evolution and Algorithmic Intervention

While macro-level statistics demonstrate the overall performance of the framework at the population level, the mechanism by which pool-level psychological monitoring translates into individual-level outcomes is best illustrated through two simulated behavioral profiles constructed from Pecan Street household load traces. We emphasize that these two households did not participate in any real gamification interaction; their behavioral parameters ($\alpha_i, \beta_i, \lambda_i$) were assigned according to the synthetic generation procedure in Section 4.1.2, and were selected here specifically because they represent, respectively, a high-fatigue-decay profile and a low-fatigue-decay profile within the simulated population, illustrating two contrasting response patterns to the same broadcast policy.

It is important to clarify the control architecture underlying this case study: consistent with Algorithm 1, the SAC agent issues a single broadcast action $a_t=(\mu_t, d_t)$ to the entire user pool at each time step; it does not assign differentiated actions to individual users. The two profiles below are therefore not examples of the agent "targeting" specific individuals, but rather illustrate how the same pool-level broadcast decision produces different realized outcomes depending on each simulated user's individual fatigue state at the moment the broadcast occurs.

.

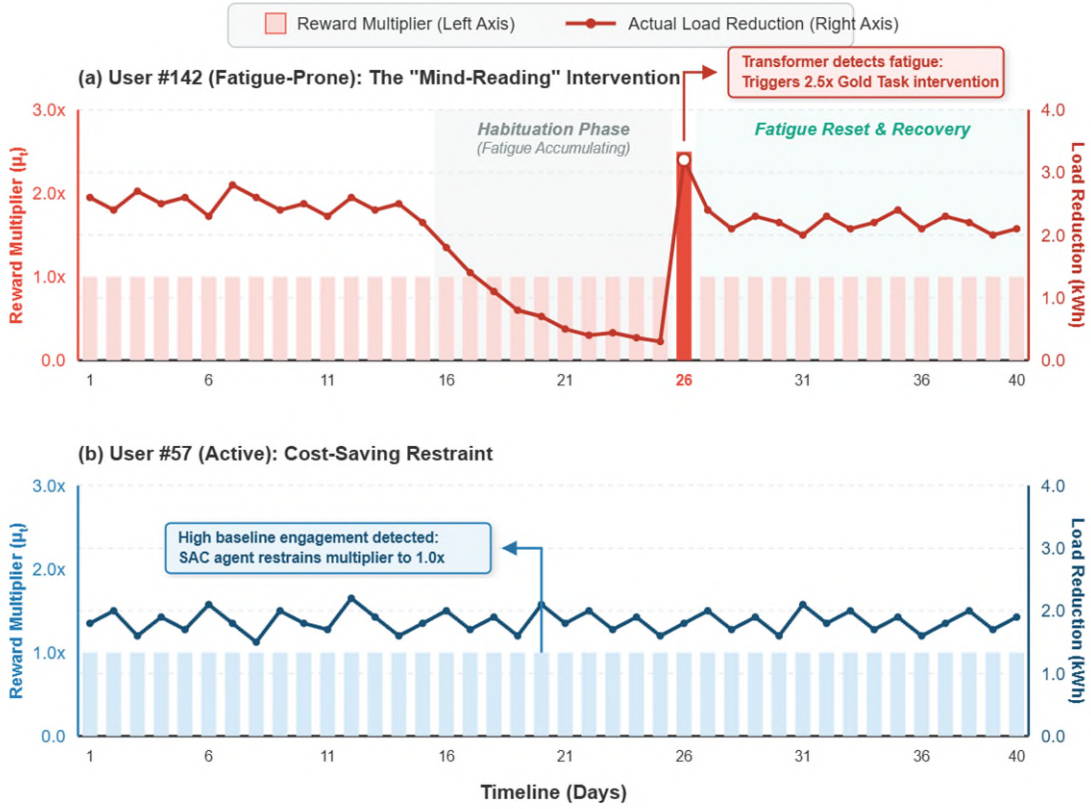


Figure 4: Dual-Y-axis time-series line chart.

1) Simulated Profile #142: A High-Fatigue-Decay User Under a Pool-Triggered Gold-Tier Broadcast

This profile is parameterized with a high intrinsic fatigue decay rate (λ 142). As shown in Figure 4, during Days 1–15, the simulated user responds actively to the standard Bronze tasks (1.0× multiplier) that are broadcast to the full pool during this period, consistently reducing about 2.5 kWh per DR event. As the novelty effect wears off, however, the habituation mechanism embedded in the behavioral model (Section 3.2) dominates: from Day 16 to Day 25, despite the broadcast policy continuing to issue standard tasks, this profile's simulated load reduction declines sharply to approximately 0.4 kWh.

At Day 26, the pool-level engagement score Et— aggregated across the user pool, which includes a cluster of similarly high-fatigue profiles — drops below the critical threshold that triggers the SAC agent's intervention rule. The Transformer module, by analyzing the pool's aggregated long-term interaction sequence, produces this low

Et signal, and the SAC agent responds by broadcasting a Gold-tier task (2.5× multiplier, high difficulty) to the entire user pool, rather than to this individual specifically. Because Profile #142 happens to be one of the most fatigued users at this moment, it is disproportionately responsive to the unexpected, high-value broadcast (the novelty factor η in Eq. 2): as reflected in Figure 4, the simulated load reduction jumps to 3.2 kWh on Day 26, and the intervention resets this profile's fatigue accumulator ($N_{142}(t) \to 0$, Eq. 3). In the subsequent days (27–40), even after the broadcast policy reverts to standard Bronze tasks, this profile's baseline response recovers to approximately 2.2 kWh. Within the assumed behavioral model, this illustrates how a pool-level intervention can produce a large recovery effect for individual users whose fatigue happened to be most acute at the time of the broadcast.

2) Simulated Profile #57: A Low-Fatigue-Decay User Under the Same Broadcast Policy

Profile #57 is parameterized with a low fatigue decay rate (λ_{57}) and, under the behavioral model, derives sustained satisfaction from standard participation. As also shown in Figure 4, throughout the same 40-day window, the pool-level engagement score Et remains healthy — partly because users like Profile #57 are not contributing to fatigue-driven decline — and consequently the SAC agent does not broadcast any Silver- or Gold-tier task during this window, since the composite reward function (Eq. 12) penalizes unnecessary incentive cost when the aggregate engagement signal does not warrant intervention. As a result, Profile #57 continues to receive only Bronze-tier (1.0×) tasks and maintains a stable simulated load reduction of approximately 2.0 kWh throughout — the flat trajectory visible in Figure 4 in contrast to Profile #142's Day-26 spike.

Summary. This case study illustrates, under the assumed behavioral model, how a single pool-level broadcast policy can produce markedly different realized outcomes across users with different underlying fatigue dynamics: the framework avoids issuing

costly high-tier broadcasts when the aggregate population does not exhibit fatigue-driven decline, while remaining responsive when a sufficient share of the pool approaches a critical engagement threshold. We reiterate that this behavior is a property of the simulated model and the broadcast control architecture defined in Section 3.1, rather than evidence of individualized targeting or of genuine human psychological detection; the broader implications of this pool-level design—and the extent to which it could be adapted toward individualized control in future work—are discussed in Section 5.3.

5. Discussion

5.1 Mechanistic Analysis

Trans-SAC (85.6% LRAR) outperforms Static-Game (45%) because fixed rewards let the fatigue accumulator Ni(t) grow unchecked (Eq. 1), whereas Trans-SAC withholds high-value rewards when engagement is already high and reserves unpredictable high-tier broadcasts for moments of critical fatigue, resetting Ni(t) accordingly. This behavior reflects a property of the assumed model, not an observed or measured psychological mechanism in real users.

5.2 Sensitivity and Reliability-Cost Trade-off

Skewing lambda toward higher values raises UIC to approximately $0.074/kWh while LRAR remains above 80%. The cost weight w2 governs the trade-off: too high collapses the policy toward Static-Game; too low pushes LRAR to 92% but inflates UIC beyond $0.095/kWh. Our operating point (85.6% LRAR, $0.061 UIC) is one favorable point along this trade-off under the simulated assumptions, not a demonstrated Pareto-optimal solution; establishing a true frontier would require systematically sweeping w2 against baselines.

5.3 Limitations

(1) Behavioral parameters are synthetic and unvalidated against real subjects; since the same equations generate both the simulated ground truth and the Transformer's prediction target, this risks a degree of circular validation that should be resolved via field pilots. (2) Decay/reset functional forms were not benchmarked against alternatives. (3) While Section 4.2.1 provides an independent evaluation of the Transformer's predictive fidelity relative to Logistic Regression, MLP, and LSTM baselines (AUC/F1/Brier/ECE), this evaluation is still conducted entirely within the same synthetic behavioral framework (Eqs. 1-4, 18-20) used to generate the ground-truth labels; it therefore does not constitute validation against independently observed human gamification behavior, and the model-level circular-validation risk discussed in Sections 3.2 and 4.1.1 remains unresolved. (4) The MLP-SAC ablation confounds architecture with observation-window length. (5) Sensitivity analysis used limited representative settings rather than systematic sweeps. (6) No demographic clustering; all users share one functional form. (7) No device-level feasibility or extreme-weather safety constraints are enforced.

6. Conclusions

6.1 Research Summary

Trans-SAC couples a Transformer-based engagement predictor with an entropy-regularized SAC agent to counter gamification fatigue through engagement-triggered, intermittent incentive broadcasts. Under the simulated behavioral assumptions of this study, it sustained an 85.6% LRAR while cutting incentive cost by 21.8% relative to Standard-RL. These results support the effectiveness of psychological-state-aware incentive design under the modeled assumptions only; independent human-subject validation is required before generalizing to real-world deployment.

6.2 Practical Significance and Future Outlook

The framework outlines a deployment blueprint broadly compatible with existing smart-meter infrastructure, but real-world deployment would additionally require explicit user consent, a functioning gamification platform, real-time communication infrastructure, behavioral parameters calibrated against real users rather than heuristic distributions, a physical/comfort safety-boundary layer, and privacy-governance mechanisms — none of which have been implemented or validated here. Future work includes: (i) Federated Learning to enable privacy-preserving, locally-trained personalization without exposing raw behavioral data; and (ii) coupling Trans-SAC with LLMs to generate personalized, context-aware incentive framing, whose real-world effectiveness remains an open empirical question.

List of Abbreviations

AI——Artificial Intelligence
AUC——Area Under the Receiver Operating Characteristic Curve
CPP——Critical Peak Pricing
DDPG——Deep Deterministic Policy Gradient
DR——Demand Response
DRL——Deep Reinforcement Learning
EVs——Electric Vehicles
ECE——Expected Calibration Error
FFN——Feed-Forward Network
FL——Federated Learning
HVAC——Heating, Ventilation, and Air Conditioning
LLMs——Large Language Models
LRAR——Load Reduction Achievement Rate
LSTM——Long Short-Term Memory
MDP——Markov Decision Process
MHA——Multi-Head Self-Attention
MLP——Multi-Layer Perceptron
PBL——Points, Badges, and Leaderboards
PE——Positional Encoding
RTP——Real-Time Pricing
SAC——Soft Actor-Critic
TOU——Time-of-Use

UI——User Interface
UIC——Unit Incentive Cost
URR——User Retention Rate
AUC——Area Under the Receiver Operating Characteristic Curve

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests

The author(s) declared no potential conflicts of interest with respect to the research, author-ship, and/or publication of this article.

Funding

This study was supported by the Guangdong University of Science and Technology Doctoral Research Start-up Fund (GKY-2024BSQDW-78).

Authors' contributions

Conceptualization, T.P.; methodology, T.P.; software, T.P.; validation, T.P.; formal analysis, T.P.; investigation, T.P.; resources, T.P.; writing—original draft preparation, T.P.; writing—review and editing, T.P.; visualization, T.P.; project administration, T.P.; funding acquisition, T.P.

Acknowledgements

Not applicable.

Reference

[1] Fernandez M I, Go Y I, Wong D M L, et al. Review of challenges and key enablers in energy systems towards net zero target: Renewables, storage, buildings, & grid technologies[J]. Heliyon, 2024, 10(23).
[2] Afzalan M, Jazizadeh F. Residential loads flexibility potential for demand response using energy consumption patterns and user segments[J]. Applied Energy, 2019, 254: 113693.
[3] Schittekatte T, Mallapragada D, Joskow P L, et al. Electricity retail rate design in a decarbonizing economy: an analysis of time-of-use and critical peak pricing[J]. The Energy Journal, 2024, 45(3): 25-56.
[4] Deng N, Wang Z, Xu S. Temporal spillovers and habit formation in residential electricity demand response: evidence from large-scale repeated behavioral experiments[J]. Energy and Buildings, 2026, 366: 117721.
[5] AlSkaif T, Lampropoulos I, Van Den Broek M, et al. Gamification-based framework for engagement of residential customers in energy applications[J]. Energy

Research & Social Science, 2018, 44: 187-195.
[6] Duan Y, Gao C, Zhang J, et al. Multi-objective optimization of gamified demand response for PV-integrated microgrids: a novel NSGA-III framework with behavioral adaptation modeling[J]. Scientific Reports, 2025, 15(1): 34022.
[7] Wohlgenannt P, Moosbrugger L, Vetter V, et al. Energy cost and emission optimization in food industry using deep reinforcement learning and Transformer-based load forecasting[J]. Energy Conversion and Management: X, 2025: 101364.
[8] Yan X, Ozturk Y, Hu Z, et al. A review on price-driven residential demand response[J]. Renewable and Sustainable Energy Reviews, 2018, 96: 411-419.
[9] Muratori M, Rizzoni G. Residential demand response: Dynamic energy management and time-varying electricity pricing[J]. IEEE Transactions on Power Systems, 2015, 31(2): 1108-1117.
[10] Wen L, Zhou K, Feng W, et al. Demand side management in smart grid: A dynamic-price-based demand response model[J]. IEEE Transactions on Engineering Management, 2022, 71: 1439-1451.
[11] Li J, Li T, Dong D. Demand response management of smart grid based on Stackelberg-evolutionary joint game[J]. Science China Information Sciences, 2023, 66(8): 182201.
[12] Ghorbanian M, Dolatabadi S H, Siano P. Game theory-based energy-management method considering autonomous demand response and distributed generation interactions in smart distribution systems[J]. IEEE Systems Journal, 2020, 15(1): 905-914.
[13] Amer A, Shaban K, Gaouda A, et al. Home energy management system embedded with a multi-objective demand response optimization model to benefit customers and operators[J]. Energies, 2021, 14(2): 257.
[14] Xu B, Wang J, Guo M, et al. A hybrid demand response mechanism based on real-time incentive and real-time pricing[J]. Energy, 2021, 231: 120940.
[15] El Gohary F, Nordin M, Juslin P, et al. Getting the signal–Do electricity users meet the preconditions for making informed decisions on demand response?[J]. Energy Research & Social Science, 2023, 100: 103119.
[16] Huseynli B. Gamification in energy consumption: a model for consumers' energy saving[J]. International Journal of Energy Economics and Policy, 2024, 14(1): 312-320.
[17] Mukai T, Nishio K, Komatsu H, et al. What effect does feedback have on energy conservation? Comparing previous household usage, neighbourhood usage, and social norms in Japan[J]. Energy Research & Social Science, 2022, 86: 102430.
[18] Nykyri M, Annala S, Silventoinen P. Review of demand response and energy communities in serious games[J]. IEEE Access, 2022, 10: 91018-91026.
[19] Nasrollahi H, Lampropoulos I, Werning S, et al. Review of serious energy games: Objectives, approaches, applications, data integration, and performance assessment[J]. Energies, 2023, 16(19): 6948.
[20] Wu J, Gao J, Yang T, et al. Using interactive gamified simulation to explore how economic incentives affect appliance flexible regulation potential[J]. Energy, 2025:

138931.
[21] Chadoulos S, Koutsopoulos I, Polyzos G C. Mobile apps meet the smart energy grid: A survey on consumer engagement and machine learning applications[J]. IEEE Access, 2020, 8: 219632-219655.
[22] Rankin C H, Abrams T, Barry R J, et al. Habituation revisited: an updated and revised description of the behavioral characteristics of habituation[J]. Neurobiology of learning and memory, 2009, 92(2): 135-138.
[23] Zhong S, Wang X, Zhao J, et al. Deep reinforcement learning framework for dynamic pricing demand response of regenerative electric heating[J]. Applied Energy, 2021, 288: 116623.
[24] Zheng W, Pi R, Zhong X, et al. Demand response for home energy management systems: a novel dual-agent DRL approach[J]. Energy Systems, 2025: 1-27.
[25] Kathirgamanathan A, Twardowski K, Mangina E, et al. A centralised soft actor critic deep reinforcement learning approach to district demand side management through CityLearn[C]//Proceedings of the 1st international workshop on reinforcement learning for energy management in buildings & cities. 2020: 11-14.
[26] Zhao Z, Xia C, Chi L, et al. Short-term load forecasting based on the transformer model[J]. Information, 2021, 12(12): 516.
[27] Saoud L S, Al-Marzouqi H, Hussein R. Household energy consumption prediction using the stationary wavelet transform and transformers[J]. IEEE Access, 2022, 10: 5171-5183.
[28] Khan M A, Saleh A M, Waseem M, et al. Artificial intelligence enabled demand response: Prospects and challenges in smart grid environment[J]. Ieee Access, 2022, 11: 1477-1505.