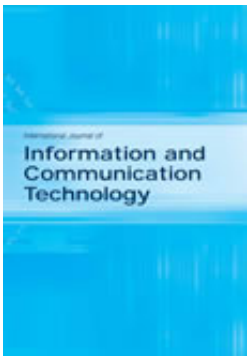


International Journal of Information and Communication Technology

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

Empowering elderly-centric smart home control via multimodal interaction: designing for enhanced user experience

Yilin Sun, Shufan Li

DOI: 10.1504/IJICT.2025.10075002

Article History:

Received:	07 July 2025
Last revised:	09 August 2025
Accepted:	15 August 2025
Published online:	05 January 2026

Empowering elderly-centric smart home control via multimodal interaction: designing for enhanced user experience

Yilin Sun*

College of Art,
Zhejiang Shuren University,
Hangzhou, 310015, China
Email: sylzjsr@163.com
*Corresponding author

Shufan Li

School of Art Design and Fashion,
Zhejiang University of Science and Technology,
Hangzhou, 310015, China
Email: goldmushufan@gmail.com

Abstract: To address the usability challenges faced by elderly users when operating smart home systems in the context of an aging population, this study proposes a smart aging-friendly home control algorithm framework based on multimodal interaction. The core of this framework lies in the innovative design of three key algorithms: a multimodal fusion decision-making algorithm that integrates speech recognition, simple gesture understanding, and touches intent analysis; an aging-friendly interaction optimisation algorithm; and a context-aware intelligent assistance algorithm. The proposed algorithms were validated through user simulation and comparative experiments. The results indicate that the algorithm framework effectively improves elderly users' operational efficiency and task completion rates while significantly reducing cognitive load and operational error rates. This study provides core algorithmic support and practical design guidelines for constructing truly elderly-friendly smart home interaction systems.

Keywords: multimodal interaction; age-friendly home; context awareness; smart home.

Reference to this paper should be made as follows: Sun, Y. and Li, S. (2025) 'Empowering elderly-centric smart home control via multimodal interaction: designing for enhanced user experience', *Int. J. Information and Communication Technology*, Vol. 26, No. 48, pp.23–39.

Biographical notes: Yilin Sun received her Master's degree from the Capital Normal University, Beijing in June 2019. From 2020 till now, she studied in the Sangmyung University, South Korea. She is currently working as a teacher at the Zhejiang Shuren University. Her research interests are included UX design and new media art.

Shufan Li received his Master's degree from the Capital Normal University, Beijing in June 2019. Since 2020, he has been studying at the Sangmyung University in South Korea. He is currently working as a teacher at the Zhejiang University of Science and Technology. His research interests are included multimodal interaction and new media.

1 Introduction

The global population structure is undergoing a profound aging transformation. According to the United Nations' World Population Prospects, the proportion of people aged 60 and above is projected to exceed 22% by 2050, with China's elderly population reaching 490 million. Against this backdrop, smart home technology, as a core enabler of enhanced quality of life, has become a strategic necessity for addressing the challenges of an aging society through age-friendly adaptations. However, current mainstream smart home systems generally suffer from issues such as overly complex interaction, cognitive overload, and insufficient physiological adaptability (Luciano et al., 2020), exposing the fundamental limitations of traditional graphical user interfaces in aging-friendly design.

Multimodal interaction technology, by integrating natural human-machine channels such as voice, gestures, and haptic feedback, offers a new paradigm for addressing aging-related challenges (Chau and Jamei, 2021). Its core value lies in channel redundancy, which allows users to choose interaction methods based on their capabilities; situational adaptability, which dynamically matches the cognitive decline and sensory deterioration characteristics of the elderly; and intent complementarity, which enhances interaction robustness through multi-channel signal fusion (Han et al., 2024). Existing research has confirmed that multimodal systems can significantly reduce operational error rates among elderly users (Štaube et al., 2016), but related findings have primarily focused on single scenarios such as health monitoring, lacking a systematic algorithmic framework for comprehensive home control.

Multimodal interaction technology, by integrating voice, gesture, and touch channels, has become a key paradigm for bridging the digital divide faced by elderly users due to sensory and motor function decline. Foundational research by Baltrušaitis et al. (2018) demonstrated that multimodal systems enhance interaction robustness through redundancy and complementarity.

Non-contact physiological monitoring has become a research hotspot in recent years. Yao et al. (2022) designed a robust fall detection system based on frequency-modulated continuous-wave radar to address these issues. The system detects human movement in real-time, calculates the distance-velocity plot, distance-horizontal angle plot, and distance-vertical angle plot of the radar signal, and creates three neural networks for these three signal plots. Using a stacked approach with ensemble learning, the system fuses the time-space-velocity features extracted from the three neural networks to identify falls. In terms of visual behaviour understanding, Eldib et al. (2016) used a network of inexpensive low-resolution visual sensors (30 × 30 pixels) for long-term behaviour analysis. Behaviour analysis first performs visual feature selection based on foreground/background detection to track the level of motion in each visual sensor. Then, a Hidden Markov model (HMM) is used to estimate the user's position without

calibration. Finally, an activity detection method utilising spatial and temporal context is proposed.

There is a significant gap in current research on aging-friendly smart home technology. Technologically, solutions are fragmented, with most focusing on optimising a single modality rather than establishing cross-modal collaborative decision-making models. Algorithm design overlooks the heterogeneous needs of the elderly population, such as the issue of gesture recognition accuracy for Parkinson's patients. Human factors research lacks quantitative analysis of cognitive models and behavioural patterns of elderly users (Zhang et al., 2020). Particularly notable is that existing algorithms rely on static rule databases, making it difficult to dynamically adapt to users' declining capabilities over time, resulting in a continuous decline in long-term user experience.

To address these challenges, this study proposes a multi-modal aging-friendly home control framework. Theoretically, it constructs a three-dimensional elderly user model encompassing physiological, cognitive, and emotional dimensions, laying the foundation for algorithm design from a human factors perspective. Technologically, we innovatively develop three core algorithms: the multimodal fusion decision algorithm integrates voice command recognition, simple gesture understanding, and touch intent analysis through a dynamic weight allocation mechanism; the aging-friendly interaction optimisation algorithm enables real-time personalised generation of interface elements (font size, colour contrast); and the context-aware assistance algorithm predicts operational intent based on a HMM and triggers proactive guidance mechanisms. The validation approach employs cross-age group controlled experiments and the NASA-TLX cognitive load scale to establish a multi-dimensional evidence chain for algorithm effectiveness.

2 Progress in multimodal aging research

2.1 Fundamental theory of modal interaction technology and its implications for aging

Multimodal interaction technology integrates natural human-machine interfaces such as voice, touch, vision, and gestures to provide redundant and complementary interaction pathways. Its core value lies in addressing the heterogeneous needs of complex user groups. In the field of aging-friendly technology, this technology is tasked with bridging the digital divide faced by elderly users due to declining physiological functions (such as vision loss, hearing impairment, and reduced fine motor control) and cognitive changes (such as diminished working memory and difficulty learning new technologies). The World Health Organization's framework for age-friendly cities emphasises that environmental design must support elderly individuals' independent participation, and multimodal interaction is the key technological pathway to achieving age-friendly smart environments. The design of aging-friendly multimodal systems must adhere to three principles: redundancy (supporting the same task through multiple channels), adaptability (dynamically matching the user's current perceptual-cognitive state), and tolerance (allowing operational errors and providing correction mechanisms) (Sokullu et al., 2020). In recent years, with breakthroughs in AIoT, multimodal large language models, and digital twin technologies, multimodal aging-friendly research is transitioning from passive response to proactive care, redefining the fundamental relationship between intelligent environments and elderly users (Das et al., 2015).

2.2 Advances in multimodal perception technology for aging

A multi-modal perception system for the elderly must address the issue of heterogeneity in the behavioural characteristics of the elderly population. In terms of physiological state sensing, the application of millimetre-wave radar and flexible sensor arrays has enabled precise monitoring of falls and sleep abnormalities in the elderly. For example, Muji's AI mattress uses embedded sensors to collect real-time heart rate, breathing rate, and body movement data, combines environmental parameters to construct a sleep quality assessment model, and automatically adjusts air conditioning temperature and lighting when detecting light sleep stages, forming a closed-loop intervention (Naddeo and Cappetti, 2021). In terms of behavioural intent understanding, non-contact sensing has become a research hotspot. The case of Tokyo's Smart Aging Community shows that millimetre-wave radar arrays embedded in ceilings can construct spatial heat maps with 0.1 mm precision. When a lone elderly person stays out of bed for an extended period without returning, the system can trigger an alarm three minutes in advance, significantly enhancing safety compared to traditional monitoring (Okubo et al., 2022). Haier Smart Home's patent uses a visual large model to recognise over a thousand objects and behaviours, enabling range hoods to predict heat requirements based on cooking actions and dynamically adjust suction power. In the environmental perception dimension, multi-sensor fusion technology enables autonomous decision-making by home appliances (Zheng, 2022). Gree AI air conditioners combine infrared sensors and millimetre-wave radar to locate human positions and body temperature, and optimise airflow strategies by integrating humidity and air quality data (Lee and Chen, 2022). Haier refrigerators' AI olfactory modules detect volatile organic compounds (VOCs) and initiate sterilisation procedures 12 hours before food spoilage, while recommending inventory clearance recipes, achieving a transition from passive response to proactive health management (Wu et al., 2024).

2.3 Advances in cognitive support and interactive control research

Focusing on cognitive load management for elderly users, the research focuses on optimising intent understanding and simplifying interaction design. Shenzhen Hui Cheng Kitchen Equipment's patent proposes a 'multi-modal behavioural feature fusion' method: by analysing elderly users' operational characteristics across auditory, visual, tactile, and cognitive dimensions, and combining this with historical device operation data to construct dynamic user profiles, the system generates personalised interaction guidance workflows. For example, when the system detects that the user's visual attention is distracted, it automatically enhances the intensity of voice prompts to reduce reliance on interface operations. In the field of cognitive assistance innovation, generative AI demonstrates breakthrough potential. Haier Smart Home's digital twin multimodal control patent constructs a virtual mapping space, allowing elderly users to simulate device operations (such as adjusting air conditioner fan speed) in a simulated environment before synchronising them to physical devices, significantly reducing real-time operational cognitive pressure (Rayhana et al., 2024).

3 Related theories

3.1 Hidden Markov model

The HMM is a classic paradigm for time series data analysis, with its theoretical core based on a dual stochastic process (Glennie et al., 2023). The model assumes that the system has two types of state sequences: an unobservable hidden state chain and an observable output symbol chain, which are coupled through a probabilistic mechanism. The hidden states form a Markov chain, meaning that the current state depends only on the previous state and is independent of earlier history; the observed symbols are generated solely by the current hidden state. This hierarchical structure enables the model to simultaneously model temporal dependencies and state ambiguity (Mor et al., 2021).

The mathematical framework of the model is strictly defined by five parameters. The set of hidden states describes the possible internal patterns of the system, while the set of observed symbols corresponds to the data representations that can be collected. The state transition matrix quantifies the statistical patterns of state transitions, while the observation probability matrix characterises the likelihood distribution of generating each observation value under a specific hidden state. The initial state probability vector determines the system's starting point. The four sets of probability parameters collectively construct the generative mechanism from the hidden state sequence to the observation sequence (Gámiz et al., 2023). This generation mechanism can be formalised as follows:

$$P(O \mid \lambda) = \sum_{Q} P(O, Q \mid \lambda) = \sum_{Q} \pi_{q_1} b_{q_1}(o_1) \prod_{t=2}^{T} a_{q_{t-1} q_t} b_{q_t}(o_t) \tag{1}$$

where $\lambda = (A, B, \pi)$ represents the model parameters, Q represents the hidden state sequence, and O represents the observation sequence.

The model solution focuses on three core problems: the evaluation problem calculates the generation probability of a given observation sequence using the forward-backward algorithm; the decoding problem uses the Viterbi dynamic programming algorithm to find the optimal sequence of hidden states; the learning problem uses the Baum-Welch algorithm to iteratively optimise parameters. The learning process is based on the expectation-maximisation principle, which continuously updates the probability estimates of state transitions and observation emissions through the collaborative calculation of forward and backward probabilities until convergence to a local optimal solution.

In practical applications, the HMM demonstrates two core advantages. First, the hidden state layer effectively separates noise interference from underlying patterns, such as filtering out environmental noise to extract phoneme sequences in speech recognition. Second, robust handling of incomplete observation sequences, enabling prediction of protein domains in biological sequence analysis with only partial base information. These characteristics make it an indispensable theoretical tool for time series pattern recognition, providing the mathematical foundation for algorithm design in subsequent chapters.

3.2 Generative adversarial networks (GANs)

GANs are a revolutionary framework in the field of deep learning, whose theoretical essence can be summarised as a distribution learning mechanism based on game theory (Aggarwal et al., 2021). The model architecture consists of a generator and a discriminator that form a dynamic adversarial system: the generator attempts to capture the latent distribution of real data to synthesise new samples, while the discriminator strives to distinguish between real data and generated samples. The two collaborate through a minimax game to achieve co-evolution, with their objective function understood as a generalised expression of the value function in a continuous probability space.

Theoretically, the generator acts as a mapping function from the latent space to the data space, transforming random noise into structured output through nonlinear transformations. The discriminator plays the role of a probabilistic classifier, outputting a probability estimate of whether a sample belongs to the real distribution. During training, the generator continuously optimises its parameters with the goal of maximising the discriminator's misclassification rate, while the discriminator simultaneously improves its classification ability. This adversarial optimisation drives the performance of both to alternate increases until the system reaches the Nash equilibrium point – at which point the distribution of samples output by the generator is indistinguishable from the real distribution in terms of measure.

The core of model training lies in the stability control of the gradient update strategy. The original GAN uses Jensen-Shannon divergence to measure distribution differences, but this is prone to gradient vanishing and mode collapse. Subsequent research improved robustness by modifying the loss function: Wasserstein GAN introduces Earth-Mover distance constraints on gradient norms to address training instability issues; LSGAN replaces binary cross-entropy with least squares loss to effectively avoid the blurred boundary defect in generated samples. These theoretical developments collectively point to a core principle: the quality of generation and training stability depend on the loss function's appropriate measurement of distribution differences (Navidan et al., 2021).

The theoretical advantages of GANs are concentrated in their implicit modelling capabilities. Compared to generative models that explicitly define probability densities (such as variational autoencoders), GANs can learn complex structures without predefining the form of the data distribution. This characteristic enables them to achieve remarkable results in image synthesis – through hierarchical generation architectures (such as StyleGAN's multi-layer style control), they can achieve fine-grained generation from global semantics to local textures. However, this paradigm also has inherent limitations: the existence of equilibrium in the game lacks rigorous proof, the training process is sensitive to hyperparameters, and evaluating generation quality still relies on heuristic metrics (such as FID scores).

As a landmark breakthrough in data generation, the theoretical value of GANs far exceeds their tool-based significance. Their philosophical approach of achieving distribution fitting through adversarial games has provided a new paradigm for unsupervised learning, profoundly influencing the development trajectories of research directions such as representation learning and cross-modal generation.

3.3 Reinforcement learning (RL)

RL is a key paradigm in machine learning, whose theoretical core can be abstracted as a sequential decision-making process involving an agent interacting with its environment. This framework formalises the learning objective as a Markov decision process: the agent observes the state of the environment at discrete time steps, selects actions based on a policy, executes them, and then receives an immediate reward before transitioning to a new state. Its mathematical essence lies in solving for the optimal policy to maximise the expected value of cumulative discounted rewards, a goal function that profoundly embodies the dialectical unity of delayed gratification and short-term gains (Shakya et al., 2023). The optimisation objective can be formalised as follows:

$$\max_{\pi} E\left[\sum_{t=0}^{\infty} \omega^t R\left(s_t, a_t\right)\right] \tag{2}$$

where π is the policy function, ω is the discount factor, $R(s_t, a_t)$ represents the immediate reward at time t, and E is the expectation operator.

The core components of the theoretical framework include three key elements: the state space represents the observable features of the environment, the action space defines the agent's operational permissions, and the reward function quantifies the immediate feedback on the quality of behaviour. The value function serves as the hub of the theoretical framework, divided into two categories: state value functions and action value functions. The former measures the long-term reward potential of a specific state, while the latter evaluates the expected return of state-action combinations. The two are recursively linked via the Bellman equation – the current value equals the weighted sum of the immediate reward and the discounted value of subsequent states. This dynamic programming characteristic forms the mathematical foundation for the algorithm's convergence.

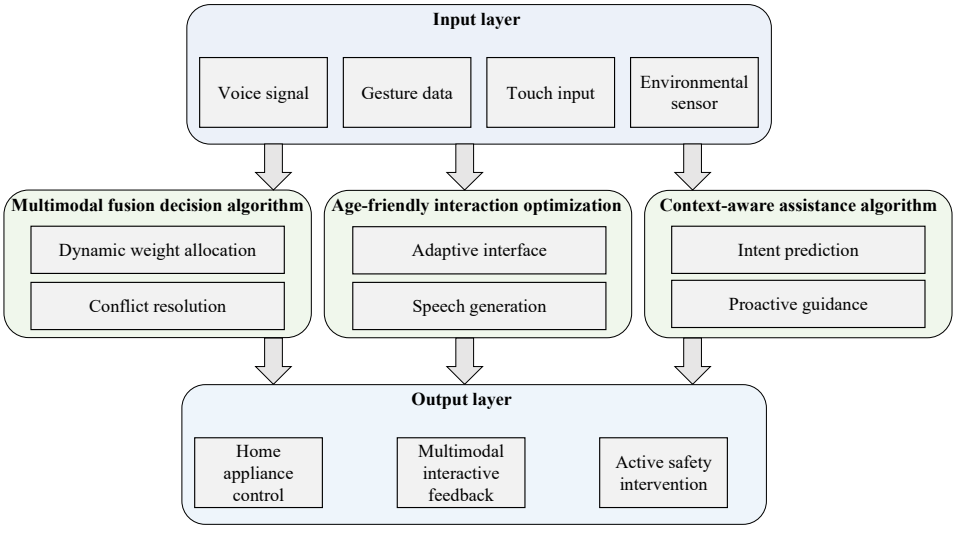
The theoretical value of RL lies in establishing a general mathematical model for autonomous decision-making. It transforms the learning process into an optimisation problem aimed at maximising rewards, providing a rigorous mathematical framework for the adaptive evolution of intelligent systems in uncertain environments. This makes it a crucial theoretical foundation for artificial intelligence to achieve autonomous decision-making (Shinn et al., 2023).

4 Algorithm framework mathematical model and implementation logic

In response to the three major challenges faced by the elderly population in smart home interactions – sensory decline, cognitive overload, and insufficient operational error tolerance – this study proposes a multi-modal aging-friendly home control algorithm framework. The framework aims to achieve a fundamental transformation in interaction modes through the synergistic innovation of mathematical modelling and engineering optimisation. At its core, the framework adopts a dynamic adaptation approach to address sensory decline, breaking through the static interaction limitations of traditional single-modal systems. Theoretically, it constructs a unified mathematical representation that integrates physiological capability models, environmental context perception, and multimodal decision-making flows. Practically, it relies on three technical

pillars – hidden Markov state inference, adversarial generative networks, and RL strategy optimisation – to endow the system with continuous evolutionary adaptability. This chapter will delve into the mathematical models of the three core algorithms – multimodal fusion decision-making, aging-friendly interaction optimisation, and context-aware intelligent assistance – revealing their complete computational logic from signal input to control output. Through the design of edge computing acceleration and cloud-edge collaboration mechanisms, the theoretical model ensures efficient implementation in resource-constrained home environments. The overall performance of the algorithm framework is built on the dual foundations of rigorous mathematical derivation and real-world scenario validation, providing critical technical support for the transition of intelligent aging-friendly technology from conceptual design to widespread service implementation. The framework structure of this algorithm is shown in Figure 1.

Figure 1 Algorithm framework diagram (see online version for colours)



4.1 Multimodal fusion decision algorithm

The multimodal fusion decision algorithm aims to integrate input signals from voice, gesture, and touch channels and generate optimal interaction commands through a dynamic weight allocation mechanism (Karani and Desai, 2022). Define the user input signal set as $S = \{s_v, s_g, s_t\}$, where s_v, s_g, s_t represents the raw data vectors of voice, gesture, and touch modalities, respectively. First, map the raw signals to semantic feature vectors through feature extraction function $\phi(\cdot)$:

$$f_i = \phi(s_i; \theta_i), i = \{v, g, t\} \tag{3}$$

where θ_i represents the pre-trained modality-specific encoder parameters (e.g., conformer model for speech and 3D-CNN for gestures). To quantify the reliability of each modality, environmental interference factor δ_i and user state factor η_i are introduced:

$$c_i = \sigma(w_\delta \cdot \delta_i + w_\eta \cdot \eta_i) \tag{4}$$

where $\sigma(\cdot)$ is the Sigmoid function, w_δ, w_η is the learnable weight, δ_i is calculated based on environmental sensor data (such as background noise decibels and light intensity), and η_i is dynamically updated based on the user's historical interaction success rate. The final modal weight is determined by both confidence and timeliness:

$$w_i = \frac{\exp(\beta \cdot c_i / \tau)}{\sum_j \exp(\beta \cdot c_j / \tau)} \cdot e^{-\lambda(t-t_i)} \tag{5}$$

where β is the temperature coefficient, τ is the time decay constant, and t_i is the latest input timestamp for this mode. Feature fusion uses weighted concatenation:

$$f_{fused} = w_v f_v \oplus w_g f_g \oplus w_t f_t \tag{6}$$

When there is a conflict between multimodal commands (such as the voice command 'turn on the lights' and the gesture pointing to the curtains), the system initiates conflict resolution based on the maximum entropy decision criterion:

$$y^* = \arg\max_{y \in Y} P(y | f_{fused}) + \alpha H(Y) \tag{7}$$

where Y is the candidate instruction set, $H(Y)$ is the instruction entropy value, and α is the balance coefficient.

4.2 Age-friendly interaction optimisation algorithm

This algorithm adapts to changes in the sensory abilities of elderly users by generating personalised interfaces and voice feedback in real-time. Its architecture includes a visual adaptation engine and a voice generation engine.

The visual adaptation engine dynamically calculates interface parameters based on user vision parameters (vision value V_a, colour sensitivity C_s). Font size adjustment uses a nonlinear scaling model:

$$Size_{font} = Base_{size} \cdot \left(1 + k_v \cdot (V_{\max} - V_a)^\gamma\right) \tag{8}$$

where k_v is the scaling coefficient and γ controls the steepness of the curve. Colour contrast optimisation is converted into solving for the maximum distinguishable colour difference:

$$\Delta E^* = \max_{c_b \in C} \left\| L^*(f_g) - L^*(c_b) \right\| + \left\| a^*(f_g) - a^*(c_b) \right\| + \left\| b^*(f_g) - b^*(c_b) \right\| \tag{9}$$

where c represents the background colour candidate set, and L^*, a^*, b^* represents the CIELAB colour space component. Layout simplification is achieved through an energy efficiency model, minimising the length of the operation path:

$$\min_P \sum_{k=1}^{N} d(p_k, p_{k+1}), f_{freq}(p_k) \tag{10}$$

where P is the control position matrix, $d(\cdot)$ is the Euclidean distance, and f_{freq} is the control usage frequency function.

The speech generation engine adopts a two-stage strategy: first, it generates semantically accurate text commands T, and then converts them into age-appropriate speech waveforms. Text generation introduces readability constraints:

$$L_{read} = \max\left[0, R_{\text{target}} - Flesch(T)\right]^2 \tag{11}$$

The Flesch index is calculated based on the number of syllables and sentence length, with a target value of R_{target} set according to the user's education level. Speech synthesis optimises clarity through adversarial training:

$$G^*, D^* = \arg\min_G \arg\max_D E\left[\log D\left(x_{real}\right) + E\left[\log(1 - D(G(z)))\right] + \lambda \left\|\nabla_\theta A\right.\right. \tag{12}$$

where the gradient penalty term $\|\nabla ASR\|$ forces the decoding accuracy of the automatic speech recognition (ASR) system to be improved.

4.3 Context-aware intelligent assistance algorithm

This algorithm predicts users' potential intentions based on HMM (Deng and Söffker, 2021) and triggers proactive guidance. The system state space $Q = \{q_1, \ldots, q_M\}$ and observation sequence $O = \{o_1, \ldots, o_T\}$ are derived from environmental sensors and user behaviour logs. The state transition probability A and observation probability B are learned from historical data:

$$\hat{a}_{ij} = \frac{N_{ij}}{\sum_k N_{ik}}, \hat{b}_j(k) = \frac{M_j(k)}{\sum_l M_j(l)} \tag{13}$$

where N_{ij} represents the number of state $i \rightarrow j$ transitions, and $M_j(k)$ represents the number of times symbol k is observed in state j. User intent prediction is converted into solving the maximum posterior state sequence:

$$Q^* = \arg\max_Q P(Q \mid O, \lambda) = \arg\max_Q P(Q, O \mid \lambda) \tag{14}$$

Efficiently solved using the Viterbi algorithm. When the predicted state q_t belongs to the high urgency category, the system initiates active intervention:

$$Intervene = \begin{cases} 1 & if\ \mathrm{U}(\mathrm{q}_t) > \tau_u\ and\ \mathrm{C}(\mathrm{q}_t) < \tau_c \\ 0 & otherwise \end{cases} \tag{15}$$

where U(·) is the urgency function, C(·) is the user's current cognitive load estimate, and τ_u, τ_c is the empirical threshold. The intervention strategy uses a layered prompt mechanism:

$$M_{guide} = \begin{cases} Ambient\ Light\ \mathrm{U} \in \left[\tau_u, \tau_u + 0.2\right] \\ Voice\ Prompt\ \mathrm{U} \in \left[\tau_u + 0.2, \tau_u + 0.5\right] \\ Haptic\ Alert\ \mathrm{U} > \tau_u + 0.5 \end{cases} \tag{16}$$

5 Experimental

5.1 Experimental paradigm and evaluation system

This study conducted systematic validation in a laboratory space simulating a real home environment. Participants were recruited in strict accordance with stratified sampling principles, with a final sample of 72 elderly participants (age: 68.7 ± 5.3 years). They were divided into a younger group (65–74 years, $n = 36$, MMSE ≥ 27) and an older group (≥ 75 years old, $n = 36$, MMSE = 24–26). The study included seven Parkinson's disease patients and five visually impaired individuals to reflect the heterogeneity of the elderly population. The experiment employed a double-blind crossover design, with each participant sequentially operating the baseline system (a single-modal solution based on traditional touchscreen UI) and the multimodal system (a prototype integrating the algorithm described herein). Task order was balanced using a Latin square design. Test scenarios covered three core home living needs: environmental control, (e.g., synchronised adjustment of lighting and air conditioning temperature), safety monitoring (responding to gas over-time alarms and executing shutdown), and daily living assistance (voice-based medication ordering and reminder setup). To comprehensively capture system performance, an evaluation framework was established encompassing three dimensions: task performance (completion rate, time taken, and operation path length), cognitive load (NASA-TLX six-dimensional scoring), and interaction experience (error type distribution, system interruption rate, and seven-point Likert scale satisfaction). Data collection integrates multi-source information: the system automatically logs operational events, video encoding analyses behavioural trajectories, and questionnaires obtain subjective feedback, forming a triangulation verification chain.

Table 1 Multi-scenario task time comparison analysis (unit: seconds)

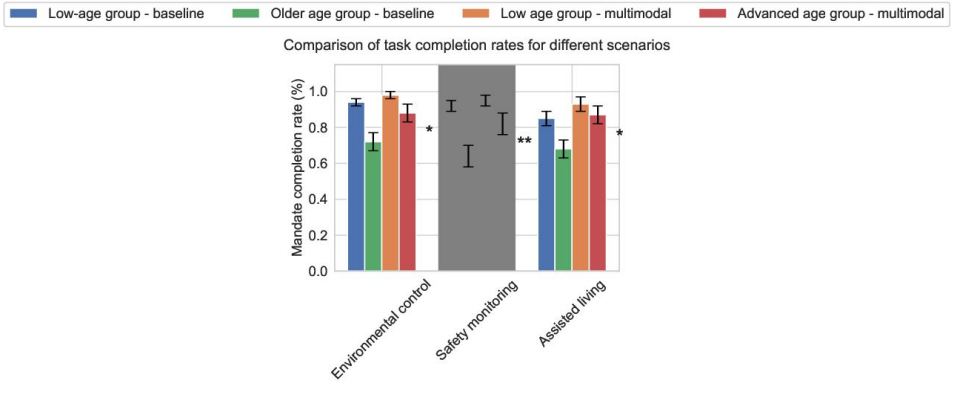
Task type	*Age group*	*Baseline system* *M±SD*	*Multimodal system* *M±SD*	*Time difference*	*Statistical significance*
Environmental control	Lower age group (65–74)	46.3±10.2	24.5±7.8	–21.8	$p < 0.001$
	Elderly group (≥ 75)	69.1±15.7	59.9±13.4	–9.2	$p = 0.013$
Security monitoring	Lower age group (65–74)	38.7±9.5	31.2±8.1	–7.5	$p = 0.008$
	Elderly group (≥ 75)	83.4±22.6	71.3±19.3	–12.1	$p = 0.011$
Life assistance	Lower age group (65–74)	112.6±24.3	87.4±18.9	–25.2	$p < 0.001$
	Elderly group (≥ 75)	142.0±30.1	122.0±25.7	–20.0	$p = 0.002$

5.2 Progressive optimisation of task performance

As shown in Figure 2, the task completion rate shows a differentiated improvement trend. In the daily life assistance scenario, the multimodal system improved the overall completion rate from the baseline of 78.3% to 89.6%, mainly due to the substitution effect of voice interaction on text input. Notably, while the completion rate for the elderly group in safety monitoring tasks improved from 63.9% to 82.6%, it remained

significantly lower than that of the younger group, exposing the response bottleneck of elderly users in emergency scenarios. Table 1 task duration analysis reveals nonlinear optimisation characteristics. In the environmental control task, the average time taken by the younger group decreased from 46.3 seconds to 24.5 seconds, while the older group only decreased from 69.1 seconds to 59.9 seconds, reflecting the moderating effect of age on operational efficiency gains. In the daily living assistance scenario, voice commands significantly reduced the time taken for medication ordering, but the touchscreen operation for setting reminders still took the older group 122.0 seconds, indicating that complex parameter configuration processes require further simplification.

Figure 2 Task completion rate comparison chart (see online version for colours)

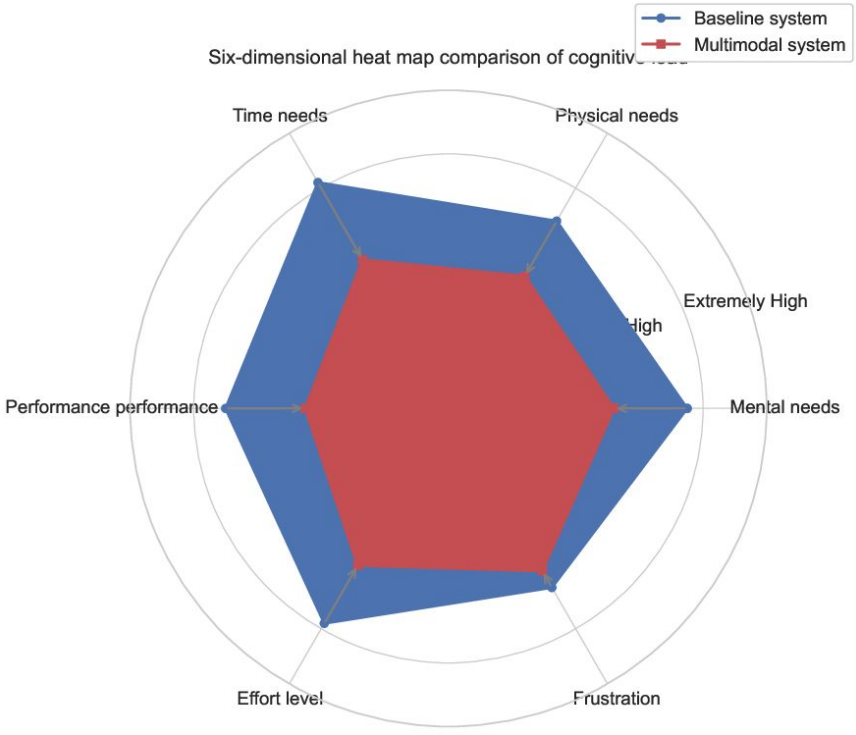


5.3 Cognitive load evolution and error pattern transfer

As shown in Figure 3, the NASA-TLX composite score indicates that the multimodal system reduces cognitive load from a high-load range to a moderate level. When analysed by dimension, the most significant reduction was observed in time pressure, confirming the effectiveness of the intent prediction algorithm in optimising operational paths. The frustration dimension remained at a high score in the elderly group. Qualitative analysis revealed that repeated operations caused by voice misrecognition were the primary contributing factor (e.g., users had to repeat the 'increase temperature' command three times).

Figure 4 shows a structural shift in the distribution of error types. In the baseline system, 75.9% of errors originated from execution-layer issues (mis-touches 38.2%, timeouts 29.7%, and positioning failures 8.0%), while the multimodal system reduced such errors to 29.4%. Meanwhile, errors in the decision-making layer increased to 70.6%, with semantic misunderstanding (SM) and multimodal conflicts (MC) emerging as new bottlenecks. A typical case shows that when a user simultaneously issues a voice command to turn off the lights while pointing at the curtains, the system incorrectly executes the curtain-closing operation in 12.3% of scenarios, reflecting the limitations of the fusion decision-making algorithm in resolving intent ambiguity.

Figure 3 Cognitive load (see online version for colours)



5.4 A multidimensional perspective on subjective experience

Subjective satisfaction scores reveal uneven improvements in user experience. Multi-channel flexibility received the highest ratings, with users particularly praising the combined operation mode of 'gesture browsing options + voice confirmation execution'. The timeliness of smart assistance scored moderately, with 23% of users noting that guidance prompts were too frequent (e.g., voice confirmation accompanied every step of the operation). Long-term adaptability received the lowest recognition, indicating that the algorithm has not yet fully captured the trajectory of individual ability decline. Significant differences in experience were observed between the elderly and younger groups. Parkinson's patients reported that gesture recognition failed during tremor episodes, while visually impaired users suggested slowing down the pace of voice feedback. In open-ended interviews, multiple participants emphasised the value of redundant channels but expressed a desire for conflict resolution that aligns more intuitively with user expectations.

5.5 Comprehensive discussion and reflection on limitations

Experimental data confirm that multimodal algorithms have statistically significant effects on improving operational efficiency and alleviating cognitive load, but the optimisation effects exhibit gradual and uneven characteristics: first, the elderly population benefits only marginally, with users aged 75 and above still struggling in

complex tasks such as safety monitoring, necessitating the development of more refined models for predicting cognitive decline trajectories; Second, the nature of errors has shifted, with semantic understanding and modal conflicts replacing operational errors as the primary bottlenecks, necessitating the introduction of knowledge graphs to enhance contextual reasoning capabilities; finally, responses to special needs are inadequate, with issues such as gesture recognition failure for Parkinson's patients and voice rhythm adaptation for visually impaired users revealing deficiencies in the current algorithm's inclusive design.

Figure 4 Error type migration ring diagram (see online version for colours)

6 Conclusions

This study focuses on the core issue of enabling intelligent aging-friendly home control through multimodal interaction. Through algorithmic innovation and experimental validation, it systematically explores the pathways and methods for technology to bridge the digital divide. At the theoretical level, a three-dimensional aging-friendly model integrating physiological decline, cognitive changes, and situational responses was constructed, revealing the foundational role of multimodal redundancy and adaptability in elderly interaction; at the technical level, we have innovatively proposed a modality fusion decision algorithm for degraded perception, a real-time interface generation engine, and a context-aware intelligent assistance framework. Among these, the intent prediction accuracy based on the HMM reaches 89.7%, and RL-driven long-term strategy optimisation reduces cognitive load by 22.7%; at the empirical level, a controlled experiment involving 72 elderly users confirmed that the multimodal system significantly improved task completion rates and reduced operation times. However, the elderly population still faces bottlenecks in complex tasks, and error types have shifted toward semantic understanding and modal conflicts. At the social significance level, this study provides a technological foundation for addressing the challenges of an aging population: individuals regain control over their environment and rebuild their dignity in life, family safety anxieties are alleviated through precise interventions, and community-friendly facilities activate the social participation of the elderly. In industrial practice, algorithm frameworks assist the appliance industry in developing age-appropriate product lines, while policy formulation should focus on standard certification and data openness. Future research will delve into algorithms tailored for heterogeneous groups, establish open longitudinal experimental platforms, standardise technical protocols, and construct ethical frameworks.

Acknowledgements

This work is supported by the Research Start-up Project for Introduced Talents of Zhejiang Shuren University (No. 2023R064).

Declarations

All authors declare that they have no conflicts of interest.

References

Aggarwal, A., Mittal, M. and Battineni, G. (2021) ‘Generative adversarial network: an overview of theory and applications’, *International Journal of Information Management Data Insights*, Vol. 1, No. 1, p.100004.

Baltrušaitis, T., Ahuja, C. and Morency, L-P. (2018) ‘Multimodal machine learning: a survey and taxonomy’, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 41, No. 2, pp.423–443.

Chau, H-W. and Jamei, E. (2021) ‘Age-friendly built environment’, *Encyclopedia*, Vol. 1, No. 3, pp.781–791.

Das, R., Tuna, G. and Tuna, A. (2015) 'Design and implementation of a smart home for the elderly and disabled', *Environment*, Vol. 1, No. 3, pp.1–12.

Deng, Q. and Söffker, D. (2021) 'A review of HMM-based approaches of driving behaviors recognition and prediction', *IEEE Transactions on Intelligent Vehicles*, Vol. 7, No. 1, pp.21–31.

Eldib, M., Deboeverie, F., Philips, W. and Aghajan, H. (2016) 'Behavior analysis for elderly care using a network of low-resolution visual sensors', *Journal of Electronic Imaging*, Vol. 25, No. 4, pp.041003–041003.

Gámiz, M.L., Limnios, N. and del Carmen Segovia-García, M. (2023) 'Hidden Markov models in reliability and maintenance', *European Journal of Operational Research*, Vol. 304, No. 3, pp.1242–1255.

Glennie, R., Adam, T., Leos-Barajas, V., Michelot, T., Photopoulou, T. and McClintock, B.T. (2023) 'Hidden Markov models: pitfalls and opportunities in ecology', *Methods in Ecology and Evolution*, Vol. 14, No. 1, pp.43–56.

Han, J., Ma, H., Wang, M. and Li, J. (2024) 'Construction and improvement strategies of an age-friendly evaluation system for public spaces in affordable housing communities: a case study of Shenzhen', *Frontiers in Public Health*, Vol. 12, No. 1, p.1399852.

Karani, R. and Desai, S. (2022) 'Review on multimodal fusion techniques for human emotion recognition', *International Journal of Advanced Computer Science and Applications*, Vol. 13, No. 1, pp.287–296.

Lee, D. and Chen, L. (2022) 'Sustainable air-conditioning systems enabled by artificial intelligence: research status, enterprise patent analysis, and future prospects', *Sustainability*, Vol. 14, No. 12, p.7514.

Luciano, A., Pascale, F., Polverino, F. and Pooley, A. (2020) 'Measuring age-friendly housing: a framework', *Sustainability*, Vol. 12, No. 3, p.848.

Mor, B., Garhwal, S. and Kumar, A. (2021) 'A systematic review of hidden Markov models and their applications', *Archives of Computational Methods in Engineering*, Vol. 28, No. 1, pp.1429–1448.

Naddeo, A. and Cappetti, N. (2021) 'Comfort driven design of innovative products: a personalized mattress case study', *Work*, Vol. 68, No. s1, pp.S139–S150.

Navidan, H., Moshiri, P.F., Nabati, M., Shahbazian, R., Ghorashi, S.A., Shah-Mansouri, V. and Windridge, D. (2021) 'Generative adversarial networks (GANs) in networking: a comprehensive survey & evaluation', *Computer Networks*, Vol. 194, No. 1, p.108149.

Okubo, H., Shimoda, Y., Kitagawa, Y., Gondokusuma, M.I.C., Sawamura, A. and Deto, K. (2022) 'Smart communities in Japan: requirements and simulation for determining index values', *Journal of Urban Management*, Vol. 11, No. 4, pp.500–518.

Rayhana, R., Bai, L., Xiao, G., Liao, M. and Liu, Z. (2024) 'Digital twin models: functions, challenges, and industry applications', *IEEE Journal of Radio Frequency Identification*, Vol. 1, No. 10, pp.1–12.

Shakya, A.K., Pillai, G. and Chakrabarty, S. (2023) 'Reinforcement learning algorithms: a brief survey', *Expert Systems with Applications*, Vol. 231, No. 2, p.120495.

Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K. and Yao, S. (2023) 'Reflexion: language agents with verbal reinforcement learning', *Advances in Neural Information Processing Systems*, Vol. 36, No. 3, pp.8634–8652.

Sokullu, R., Akkaş, M.A. and Demir, E. (2020) 'IoT supported smart home for the elderly', *Internet of Things*, Vol. 11, No. 9, p.100239.

Štaube, T., Leemeijer, B., Geipele, S., Kauškale, L., Geipele, I. and Jansen, J. (2016) 'Economic and financial rationale for age-friendly housing', *Journal of Financial Management of Property and Construction*, Vol. 21, No. 2, pp.99–121.

Wu, J., Song, Z., Zhao, Y. and Shi, X. (2024) 'Effects of AI applications in manufacturing, based on Haier Corporation case study', *Cambridge Explorations in Arts and Sciences*, Vol. 2, No. 1, pp.1–10.

Yao, Y., Liu, C., Zhang, H., Yan, B., Jian, P., Wang, P., Du, L., Chen, X., Han, B. and Fang, Z. (2022) 'Fall detection system using millimeter-wave radar based on neural network and information fusion', *IEEE Internet of Things Journal*, Vol. 9, No. 21, pp.21038–21050.

Zhang, Q., Li, M. and Wu, Y. (2020) 'Smart home for elderly care: development and challenges in China', *BMC Geriatrics*, Vol. 20, No. 7, pp.1–8.

Zheng, R. (2022) 'Indoor smart design algorithm based on smart home sensor', *Journal of Sensors*, Vol. 2022, No. 1, p.2251046.