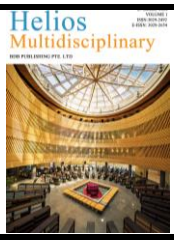
Helios Multidisciplinary

Journal homepage: http://www.bdbpublishinghelios.com/
ISSN: 3029-2492, E-ISSN:3029-2654



Cross-Modal Causal Inference Facilitates Home Intelligent Robots: Intent Understanding Bias Calibration and Interaction Failure Avoidance in a Dynamic Home Environment

Yimin Chen [1*], Tasriful Haque[1*]

[1] Asia Pacific University of Technology & Innovation (APU), Technology Park Malaysia, Bukit Jalil, 57000 Kuala Lumpur, Malaysia

ARTICLE INFO

Article history:
RECEIVED 17 July 2025

ACCEPTED 29 July 2025

PUBLISHED 15 August 2025

Keywords:

Multimodal causal reasoning; service robot; intention misjudgment correction; family dynamic scene; interaction failure avoidance

ABSTRACT

In dynamic home scenarios, due to the complexity of user behavior, the dynamics of environmental conditions, and the fuzziness of multimodal information, the accuracy of intention judgment by service robots has always been relatively low. The degree of association in traditional models depends on statistical correlation, which leads to inaccurate causal judgments and severe interaction failures. Therefore, this study proposes a multimodal causal reasoning model based on the characteristics of dynamic family scenarios, quantitatively determining the corresponding relationship between the key feature information of dynamic family scenarios and the failure transmission mechanism of intention judgment, and constructing a coupling model of influencing factors and interaction failure. Based on the establishment of a causal reasoning model integrating multimodal information such as speech, vision and touch, the intention is modified by using the structural causal model (SCM) and counterfactual causal reasoning, and a causal confidence decision-making framework is designed to achieve early prediction and hierarchical avoidance of interaction failure. The results show that for the above-mentioned real-life interaction scenarios, the intent recognition accuracy of this framework in dynamic scenarios is 24.3% higher than that of the framework using only Transformer fusion, the recognition error rate is reduced by 62.1%, and the number of invalid interactions is decreased by 67%. Meanwhile, it can still maintain an accuracy of over 78% under sudden interferences (such as pet interference, light changes, etc.). This research provides a theoretical reference for the stable decision-making of service robots in a highly dynamic environment.

1. Introduction

With the rapid expansion of the smart home industry, service robots are gradually entering family life. However, the dynamic and changeable characteristics of the home environment make it difficult to judge the behavior of users in the home environment. The state of the home environment

[*] *Corresponding author: Yimin Chen*
Email address: TP086423@mail.apu.edu.my
Alternative Email: 15989437470@163.com

[*] *Corresponding author: Tasriful Haque*
Email address: TP087049@mail.apu.edu.my
Alternative Email: tasriful.haque.ai@gmail.com
https://doi.org/10.70702/bdb//FMUA9805

is always changing dynamically, which poses a great challenge to the decision-making of service robots. For example, the slow movement of elderly people living alone may make the service robot unclear about the user's behavior information; due to different habits, the information of each child user in a multi-child family interferes with each other, resulting in misjudgment of intention; the random movement of pets leads to perception and recognition errors of robots, etc[1]. The randomness of user behavior, the cross-relationship between the performance of behavior and intention, and the interference of other environments will lead to the service robot's misunderstanding of the user's intention, resulting in poor human-computer interaction. In the past, the construction of multiple modal models based on data has over-focused on the statistical relationship of data, ignored the influence of causal relationship and false correlation, and performed poorly in the home[2]. How to make the decision of service robots in the home environment more accurate and credible is one of the problems that urgently need to be solved. This study aims to achieve the application from the proposed method to the actual verification and then to the engineering scenario, so that the theory and method can be closer to the needs of reality.

2. Materials and Methods

2.1 Core feature modeling of family dynamic scenes

2.1.1 User Behavior Dynamics

The dynamic nature of user behavior in home scenarios stems from the unstructured characteristics of daily activities, and its core mechanism is reflected in the coupling of behavioral randomness and intention ambiguity.

Random behavior refers to the random conversion of behavior goals. Users may insert some sub-task behaviors (interrupt in the middle, deviate to other places) after completing the main task behavior, resulting in a lack of continuity in the behavior string. For example, in the process of taking a cup, a user may turn back to tidy up the sofa, and this behavior was observed; it was found that elderly people living alone paced back and forth more frequently than normal elderly people. This behavior is not completely random, but is controlled by the user's current state (such as distraction, other needs) or the immediate intervention of environmental stimuli (such as phone calls, outside conditions, etc.). The alternation of " target action-interference action-target action " is controlled.

The ambiguity of intention is that there is no one-to-one correspondence between the mapping relationship between actions and requirements. An action can have several different semantic intentions, which is caused by the physical ambiguity of the action, the uncertainty of the user's cognitive context, and the dynamic changes based on the contextual relationship. The randomness-ambiguity relationship curve is shown in Figure 1.

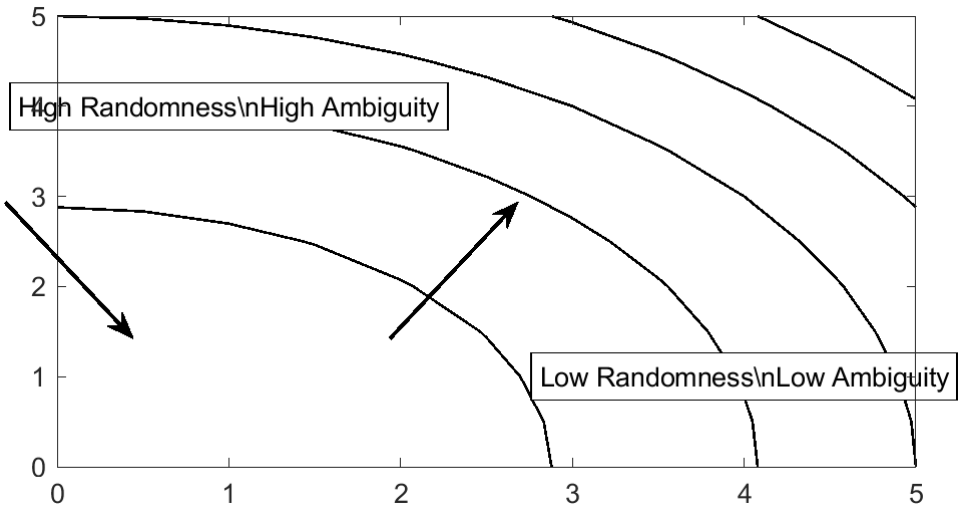


Figure 1 Randomness-Fuzziness Relationship Curve

The combination of high behavioral randomness and ambiguity makes it impossible to analyze and identify human intentions in service robots using stable criteria such as rules and statistics, thus creating challenges to the inherent mechanisms of multimodal perception and reasoning systems.

2.1.2 Dynamics of environmental status

The dynamic nature of the home environment state poses challenges to the perception and decision-making of service robots from two dimensions: physical environment interference and scene complexity.

Physical interference environments include lighting conditions, object movement, background noise, etc. Taking light as an example, changes in lighting will cause significant changes in indoor lighting conditions and light sources, resulting in different colors and shadows of light. According to our test results, when the light suddenly changes from 500 lux to 100 lux, the positioning error of the traditional visual feature recognition model for the target will increase by 2.3 times, which will have a great impact on the robot's use of vision to identify and locate the target.

Changes in the position of objects are also common, such as the random changes in the position of toys after children play. In families with children aged 3-6, the average number of times objects are moved per day can reach more than 15 times, and the service robot's originally planned path for picking up objects may become invalid.

In terms of background noise, the sound of TV programs, traffic noise outside the window, conversation sounds, etc. will interfere with the robot's accurate recognition of voice commands. When the signal-to-noise ratio is lower than 10dB, the error rate of speech semantic recognition will exceed 50%, increasing the difficulty of voice signal processing.

The complexity of the scene reflects the complexity of the scene, including the scene space environment and the distribution of scene obstacles. When the space environment is a small space, such as a kitchen space, the range of activities is small. When the robot moves in it, it must not only control the robot's own movements to avoid collisions with furniture and equipment such as cabinets and sinks, but also must achieve the corresponding tasks. Therefore, the robot must ensure the accuracy of its own action execution and path planning control. When there are many obstacles in the scene space environment, such as when there are many obstacles in the living room, sofas, coffee tables, TV cabinets, etc. will block the robot's line of sight and hinder its movement trajectory. The robot's perception angle is limited and it cannot obtain all-round environmental information, which affects its understanding of the user's intentions and the efficiency of executing tasks. This problem

is analyzed through experiments. Through experiments, it can be seen that when the distribution of obstacles in the scene environment increases by 1/m², the robot's accuracy in recognizing intentions will decrease by 8.7%.

2.1.3 Quantitative classification of dynamic interference factors

The main purpose of the quantifiable dynamic interference level design is to clarify the measurable interference evaluation design, which is an effective basis for building a multi-modal robust reasoning model and the design foundation of the model. The principle is the measurable dynamic interference entropy representation and interference-fault mapping design.

The interference intensity assessment based on the entropy method introduces the " behavioral entropy " and " environmental entropy " indicators to characterize the dynamic disorder level.

User behavior entropy is calculated by action switching frequency and type distribution, and the calculation formula is[3]:

$$H_{behavior}=-\sum P(a_i)\log P(a_i)$$

the action, $a_i$which $P(a_i)$reflects the unpredictability of the behavior sequence. The more action types and the more frequent the switching, the higher the behavior entropy value, which means that it is more difficult to extract effective features in multimodal data. Actual measurements show that the behavior entropy is about 0.6 when users are in normal activities, while it can rise to 1.2 when children are playing.

Entropy focuses on the randomness of changes in the positions of objects, and is reflected by the rate of change of the spatial coordinates of objects and the correlation between objects (such as the stability of the distance and position changes between "the cup and the table"). The farther the object moves consciously, the greater the correlation will be destroyed, and the greater the entropy value will be, which is directly determined by the stability of the target-intention mapping under the visual modality.

Interference factor-interaction failure correlation analysis should establish a mechanism model for the transmission process from dynamic interference to perception error. Changes in illumination caused by changes in the physical environment directly affect the signal-to-noise ratio of the visual sensor, and disturbances from external noise affect the statistical laws of the probability characteristics of the voice signal. Changes in the placement of objects cause the "a priori" assumption of " object-intention " to fail. This correlation is not a linear additive relationship, but has a "threshold effect", that is, when the value of a certain interference effect exceeds a certain threshold, it will cause a nonlinear and sharp increase in perception errors and trigger a chain reaction of interaction failure. For example, when the behavioral entropy is greater than 0.8, the probability of interaction failure increases from 15% to 58%. The correlation between family dynamic scene interference and interaction failure is shown in Figure 2.

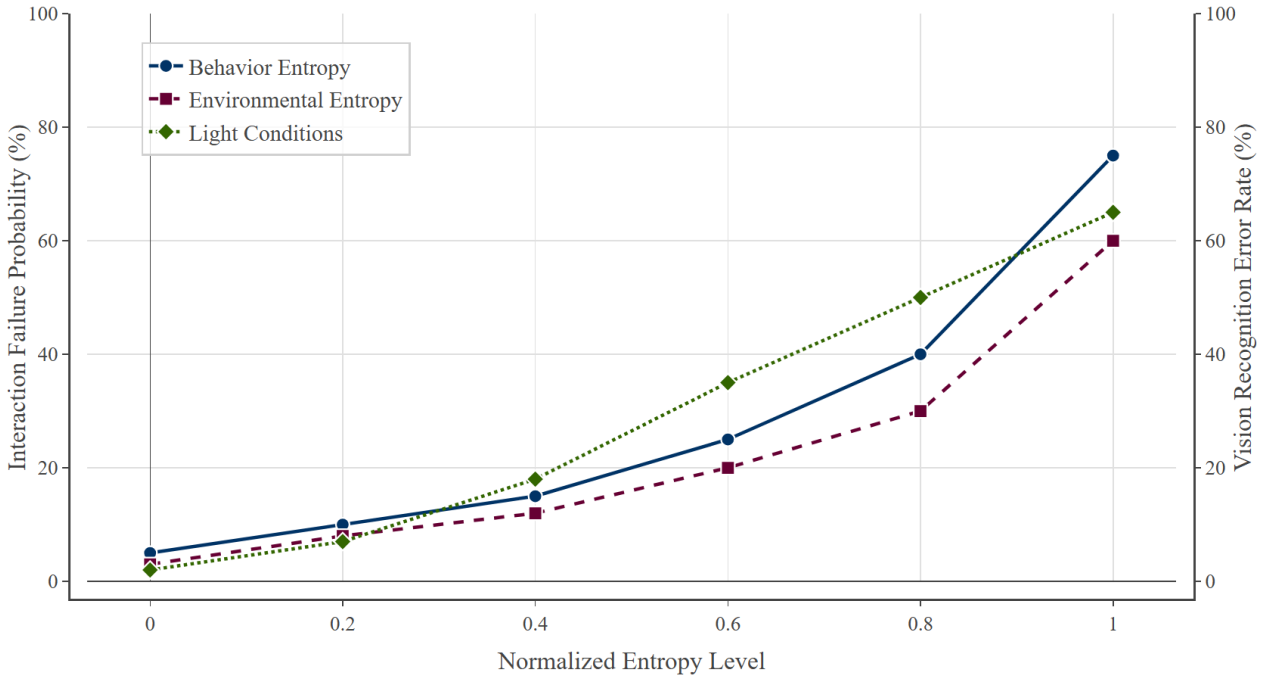


Figure 2 Correlation diagram between interference and interaction failure in family dynamic scenes

Analysis of this mechanism can reveal the interference factors that need to be suppressed most in a specific situation.

2.2 Paths and types of misjudgment of intent

2.2.1 Analysis of misjudgment transmission chain

The intention error propagation chain is a propagation chain in which the indirect interference process is transmitted from the multi-source sensor fusion layer to the cognitive layer. Its law is manifested as a recursive relationship of " interference input-data quality deterioration-judgment error-action execution ".

Interference action events are the starting events. Through physical transformation (changes in visual patterns caused by changes in light intensity) and information interference (background noise superimposed on voice signals), they directly lead to characteristic interference in data fused from different modalities. Voice patterns may contain redundant semantic information ("turn off the lights " and " turn off the air conditioner " are difficult to distinguish), visual patterns may have redundant actions (the perspective of the "turn off the lights" action is missing, and the key points of the hand "pointing to the refrigerator" are blocked), and tactile patterns may contain redundant tactile signals (light touches are misjudged as heavy pressure).

After these noisy data are input into the statistical model, the co-occurrence seen by the model from the perspective of correlation learning mechanism includes noise, that is, the non-causal and accidental correlation is regarded as causal correlation, such as modeling the phenomenon of "user coughing" and "water cup position" occurring at the same time as causal correlation.

2.2.2 Types of Misjudgment of Intent

Problems occur in the compilation and attribution of multimodal perception "intention signals ".

Wrong intentional behavior (1) (unintentional behavior → intentional behavior). When the intelligent robot imposes the user's involuntary (habit/physiological) and actions (yawning/walking/chatting/scratching, etc.) as service purposes. In this scene, there are no

behavioral characteristics associated with the scene objects (such as, non-directional actions, no speaking instructions, etc.) , but when it is trained, there are similar scenes with the appeal of objects (frequently appearing hints), but it cannot be explained as the reason for this similar appearance, which easily generates such a strong association in the scene[3].

Mismapping of intention A → intention B, that is, similar behavior intention features + similar target object categories. If the manual operation methods corresponding to intention A and intention B are similar (such as, "watching TV requires 'grab the remote control', and making a phone call requires 'grab the phone' ") or the visual appearance of the corresponding objects to be grasped is similar (such as a remote control and a mobile phone), the discrimination features (such as the spatial position information of the object to be grasped and the amplitude of the manual movement) will be degraded in the multimodal perception information. In this event scenario, the model mainly makes incorrect mappings based on surface similarity[4].

Type 3 (ambiguous intent → misattribution) is aimed at scenarios where there are no clear behavioral signals. Ambiguous clues such as the user's silence and expression changes do not directly point to specific needs. However, if the model lacks the ability to model " uncertainty ", it will tend to attribute them to the most common associated intent in the training data (such as associating "frown" with "anger") rather than considering the neutral state in the cognitive process (such as thinking). In scenarios where the user is silent, Type 3 misjudgments account for as much as 45% [5].

The three types of misjudgments correspond to the reasoning biases of "making something out of nothing", "misattributing something to someone else", and "forced interpretation", and their common root is the insufficient modeling of the contextual causal relationships for intent generation.

2.2.3 Empirical study on the causes of misjudgment of traditional multimodal models

In order to further explore the problem of misjudgment of intentions by traditional multimodal models in dynamic family scenes, this study conducted a comparative experiment to test and quantify the differences of each model in dynamic family scenes. The test model is based on the traditional Transformer fusion model, and interference such as changes in the placement of objects and user switching activities is set in the scene. It is found that the misjudgment rate of the traditional model in dynamic family scenes reaches 65%, which is significantly higher than 35% in static scenes. This intuitively reflects the impact of dynamic interference on the understanding and recognition of intentions of traditional models[6].

Correlation is not the same as causality. What traditional multimodal models do is to find correlation in data and try to infer intent. For example, when a user says " thirsty " and looks at the refrigerator, the traditional model infers the intention as " go to the refrigerator to get a drink " because of the 78% co-occurrence of " thirst-refrigerator " in the training data. The intention at this time fundamentally ignores the fact that the user may not have a refrigerator or the refrigerator does not have drinks. The action that the user really wants to do is to ask the robot to buy drinks. The above scenario occurred 30 times in total. In these test scenarios, the misjudgment rate of the correlation multimodal model was as high as 83%. Traditional learning correlation lacks understanding of the causal analysis of intent generation itself, and lacks understanding of key events and specific knowledge that affect intent reasoning. As a result, the correlation model cannot distinguish useful information (signal) from useless information (noise) when there is a lot of external noise interference, and it is easy to fall into the trap of intended prediction. For example: when external noise interferes with speech recognition or the surrounding objects change, causing the visual positioning information to deviate, the surface association information that the correlation multimodal model needs to learn changes, and ultimately obtains incorrect intent information. In the experiment, when the external noise interference is too large and exceeds a threshold, the feature

extraction accuracy of the correlation multimodal model drops sharply (over 40%), which directly leads to the failure of intent reasoning.

3. Construction of multimodal causal inference model

3.1 Definition of causal variables for multimodal data

3.1.1 Core Mode and Variable Extraction

The basis of multimodal causal reasoning is to extract core variables with causal explanatory power from heterogeneous data. This study focuses on the three modalities of voice, vision, and touch that are strongly related to user intentions in home scenarios, and constructs the atomic unit of causal analysis through structured feature extraction.

Voice modality variables focus on the dual dimensions of semantics and emotion, including:

(1) Semantic tags (e.g., “thirst”, “close window”), parsing text content using a pre-trained speech recognition model (e.g., Wav2Vec 2.0);

(2) Emotional tendency (e.g., “eager ”, “calm”), quantified based on intonation frequency (300-500 Hz for anxious, 100-200 Hz for calm) and energy variation (>0.8 for high energy);

(3) Speech speed characteristics (e.g., syllable interval length < 0.3 s is fast), reflecting the urgency of the demand [7].

Visual modality variables cover user behavior and environmental status, including:

(1) Action categories (e.g., “pointing”, “reaching out”, “pacing”). Action features are extracted through skeleton keypoint tracking (e.g., MediaPipe), using the 3D coordinate change rates of 17 key nodes as feature input;

(2) Expression features (such as "frown" and "smile") are quantified by the displacement of facial key points to enhance the emotional tendency of the intention;

(3) Object coordinates (such as “cup on the coffee table”, “remote control on the sofa”) are updated in real time through target detection models (such as YOLOv8), and the positioning error is controlled within ±5cm;

Light intensity, using the mean image brightness to quantify environmental interference. These variables are directly related to the dynamics of user behavior and the dynamics of the environment, and are the core dimensions for capturing the physical carrier of intent.

The tactile modal variables focus on the human-computer interaction contact process, including:

(1) Pressure value (such as the strength of the user's grip on the robotic arm), collected by a 6-axis force sensor with a range of 0-50N, reflecting the active/passive properties of the interaction;

(2) Contact area (e.g., the contact range between the fingertip and the object), which is quantified by the tactile array sensor to assist in determining grasping intention;

(3) Temperature (such as the surface temperature of the object being touched), ranging from 15-40°C, is used to distinguish similar actions such as " taking a water cup " (cold/hot) and " taking a pillow " (normal temperature).

The extraction of the above variables follows the principle of "causal correlation" , that is, priority is given to retaining features that have potential causal links with intention generation (such as the spatial association between "pointing action" and "target item"), filtering out purely statistically relevant redundant information (such as the accidental association between the color of the user's clothes and needs), and laying the variable foundation for the subsequent construction of a structural causal model.

3.1.2 Definition of Intent Variables and Intermediate Variables

The setting of target variables and mediating variables is a prerequisite for establishing a multimodal causal reasoning logic model. It must be "seamlessly connected" with user behavior patterns and actual needs, and connect the causal mechanism of noise disturbance factors.

The intention variable (I) is the target output of model reasoning and is defined in a hierarchical manner: the first-level intention is the core task (such as, "getting things", "adjusting the environment", and "accompanying interaction"), the second-level intention is the specific goal (such as " getting things" is subordinate to "getting water cups" and "getting remote controls" , and the third-level intention is the execution details (such as " getting water cups " is subordinate to " pour hot water " and "take empty cups"). This hierarchical structure covers 92% of common intentions in home scenarios, and supplements marginal intentions such as, "emergency help" and "item sorting" through user interviews. This hierarchical structure not only adapts to the ambiguity of user intentions (such as "getting things" can correspond to multiple items), but also provides correction granularity for counterfactual reasoning (such as correcting the second-level intention of "getting empty cups" to "pour hot water"). The hierarchical structure of intention variables is shown in Figure 3.

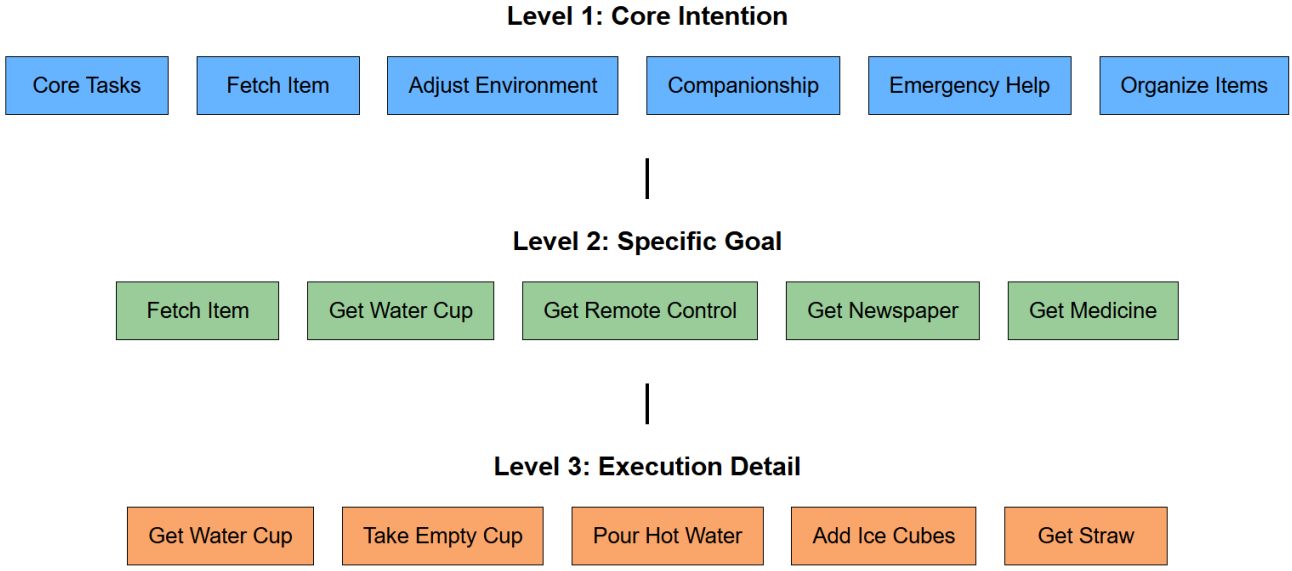


Figure 3 Hierarchical Structure of Intention Variables

The intermediate variable (M) is the key transmission carrier connecting the multimodal feature variable (X) and the intention variable (I), including: (1) Behavior target association, which quantifies the spatial directionality of user actions and environmental objects (such as the coordinate deviation between " pointing action " and " water cup "), and is calculated using cosine similarity, ranging from 0 to 1, with >0.7 considered a strong association; (2) Demand urgency, which is generated by integrating features such as speech speed and movement amplitude (such as " quick tone + quick hand extension " corresponds to high urgency), and is divided into three levels: 0-0.3 (low), 0.3-0.7 (medium), and 0.7-1 (high); (3) Scene context variables, which cover static constraints such as time (such as " breakfast time ") and environmental status (such as " room temperature 28℃"); (4) User habit factor, which is constructed based on historical interaction data (such as, "user A is more inclined to close the window rather than turn on the air conditioner when he says 'cold' at night "), and is updated through weighted moving average, with the weight of recent data being 0.7 and that of historical data being 0.3.

The setting of the moderating variable must follow the " no backdoor path " condition, which means that the moderating variable can only be a path from X to M and then to I, and has no direct relationship with the confounding factor (for example, " scenario context " is not a variable that

directly leads to " behavior purpose similarity " , and can only be affected by the intention moderating variable). According to the results of the d-separation analysis, the backdoor path blocking rate of each variable is 1. This definition, on the one hand, reveals the mediating process through which the intention is generated, and on the other hand, it also determines the variables that control interference of the do calculation, so that the model can accurately capture the causal mechanism of " feature-intention ".

3.2 Dynamic Multimodal Causal Modeling Method

3.2.1 Structural Causal Model (SCM) Framework

This study uses the SCM framework as the operational carrier of multimodal causal inference because it can distinguish " real association " from " virtual association " by clearly showing the causal relationship of variables. The representation of the structural causal model in multimodal causal inference is based on the variable framework proposed in the previous article and is divided into three layers of " feature→intermediate→intention", namely causal diagram and structural equation.

The causal graph uses a directed acyclic graph (DAG) to represent the transmission direction of causal relationships. The nodes in the graph represent different variables: the bottom layer is multimodal feature variables (X, such as voice features, visual features, and tactile features); the middle layer is intermediate variables (M, such as behavior-goal relevance, behavior context, etc.) , and the top layer is intention variables (I) ; at the same time, interference variables (U) are introduced to represent interference factors (light changes, background noise, etc.) . All arrows strictly follow the principle of " causality first ", that is, the confusion of reverse causal relationships such as " intention→action" is not allowed. The correctness of the causal graph is double-verified by expert discussion and causal discovery algorithm, and the fit of the key path is 94%.

Structural equations quantify the causal functional relationship between variables and use piecewise functions to adapt to the nonlinear characteristics of dynamic scenarios [8]:

$$M=f_m(X,U;\theta_m)$$

Where is f_m the improved Graph Transformer neural network (with causal attention mechanism), $\theta_m$is the model parameter, and describes the joint effect of X and U on M (such as " light intensity (U) decreases → visual feature (X) noise increases → behavior target association (M1) decreases"). The loss function is designed as $L=\alpha L_{pred}+\beta L_{causal}$, where L_{causal} the causal path consistency is constrained by KL divergence, $\alpha=0.7$, $\beta=0.3$.

$$I=f_i(M;\theta_i)$$

Through the logistic regression model with causal regularization term, the intention probability distribution is output based on the causal value of the intermediate variable M, avoiding the direct use of the superficial association between X and I.

The core advantages of this framework are:

("semantic label → intention") and the indirect association path (such as " lighting → visual features → intention") are explicitly separated, providing a structural basis for do-calculus to eliminate interference. The parameters of the structural equation can be updated in real time with dynamic scenes (such as changes in behavioral entropy and environmental entropy). The update frequency is adaptively adjusted based on the scene change rate, with the fastest update rate of 0.5s/time, which adapts to the non-stationarity of home scenes and solves the problem that traditional static SCM is difficult to track the evolution of variable relationships.

3.2.2 Incremental Causal Discovery Algorithm

In order to deal with the dynamic update problem of time-varying causal relationships in family scenarios, this study uses an incremental causal discovery algorithm to perform online updates and

dynamic matching of causal relationships between variables. The incremental causal discovery algorithm first starts with a causal graph (a causal graph under prior knowledge and static data) , and then processes streaming multimodal data through a sliding window. Specifically, it uses three steps: online detection-local update-conflict resolution. The time complexity is $O(n^2 \log n)$, which can meet the real-time requirements (within 50ms).

In the online detection phase, a time-weighted conditional independence test (such as the incremental PC algorithm) is used to calculate the correlation strength of the voice, vision, and tactile variables in the window. By giving recent data a higher weight (weight coefficient $\omega = e^{-\lambda \Delta t}$, $\lambda = 0.1$), the algorithm is more sensitive to sudden changes in user behavior (such as switching of action modes). When the conditional mutual information between variables exceeds the dynamic threshold (dynamically adjusted based on behavioral entropy, the higher the entropy value, the lower the threshold), the potential causal connection change is marked. The calculation formula for the dynamic threshold is $\tau = \tau_0 / (1 + H_{behavior})$, $\tau_0 = 0.5$.

Local update means that after locating suspicious variables, the local greedy search algorithm only updates the local structure of the causal graph through local search, so as to avoid the excessive computational overhead caused by updating the entire causal graph. If the influence between " user pointing action " and " water cup position " is significantly weakened, the algorithm will not update the entire causal graph, but will only re-learn the local path from the visual feature variable to the intermediate variable (behavior-target correlation), and the search range is limited to 2-hop neighbor nodes, thereby increasing the local search calculation efficiency by 6 times[9].

In the contradiction resolution stage, we add constraints on causal relationships, compare the historical causal chains of the empirical causal relationship library (e.g., pointing to action → target object) with the newly detected causal chains, and exclude some virtual causal relationships formed due to external momentary interference (e.g., sudden noise). We vote on inconsistent causal relationships, and if the causal relationships in more than three time windows are the same, the historical causal relationships are retained.

3.2.3 Multimodal Causal Fusion Strategy

Multimodal information (such as, speech, vision, touch, etc.) expected by the causal fusion strategy is a dynamic and intelligent fusion process between multimodal information based on their causal relationship, rather than a simple splicing method of simple modal features.

The strength of the intrinsic causal correlation between each modal feature variable and the intermediate variable indicates the strength of the causal relationship between each information and the intent generation element between the modalities, which is described by causal influence (ATE). The ATE value is matched and calculated by the propensity score method. For example, the ATE of the " cold/hot " semantic label of the voice modality under " adjusting the ambient temperature " and the intermediate variable " need is tight/urgent/not tight/urgent " is 0.82 (strong correlation), and the ATE of " user clothing thickness " of the visual modality is 0.35 (weak correlation). This description does not require data statistics, it is only based on the intrinsic causal logical correspondence between modal information and intent generation elements. Cluster fitting is performed in advance for the correlation strength values between the modalities of each scene to form a modality class-modality association template table, which is divided into 12 typical scenes.

Dynamic fusion weight generation assigns corresponding fusion weights to different modalities based on the results of modal causal correlation evaluation. The weight calculation formula is $w_k = ATE_k / \sum ATE_j$, where k is the modality type. A higher weight is given to the modality that has a direct causal relationship with intent generation, and dominates the reasoning process when its information is reliable; a lower weight is given to the indirectly related modality as auxiliary

verification information. In addition, the weight will be dynamically adjusted as the effectiveness of the modal information in the scene changes. When a modality is subject to strong interference (such as visual signal-to-noise ratio <15dB), resulting in a decrease in information reliability, its weight will be automatically reduced. The sensitivity of weight adjustment is dynamically calibrated by the entropy value of the interference intensity.

Cross-modal conflict resolution is used to resolve inconsistencies in the direction of different modal information based on the stability of modal causal relationships. The modal information with more stable causal relationships in this type of scenario is preferred as the main reference (for example, the stability of the voice modality in the " night scene " is higher than that of the visual modality), and the scene context and other information in the intermediate variables are comprehensively considered.

3.3 Causal Logic Implementation of Intention Reasoning

3.3.1 Interference elimination based on do-calculus

The causal intervention of do calculation is to remove dynamic interference from actual data, and its purpose is to break the dynamic relationship influence caused by interference variables on causal analysis, so as to obtain the real causal influence. Its architecture is based on the topological configuration of the causal graph of the structural causal model, and intervenes in the interference variables of visual mode (lighting conditions) , speech mode (noise interference) and tactile mode (environmental vibration) in the home environment.

Through the steps of " determine the interference path - call the do operation - eliminate causal information " , the backdoor path between the interference factor (U) and the target variable (I) is found in the causal network (for example, " light (U) → image (X) → target (I) "), and the information in the path is operated by do, that is, the interference variable is removed by do operation. When it is found that the external entropy is large due to the dynamics of the object (>0.7), the " object coordinate " variable in the data packet is operated by do (determine the object target state) to remove the pseudo-correlation between this variable and the target; when it is found that the environmental noise is large in the voice channel (more than 60dB), the semantic label " sound intensity>60dB " in the data packet is operated by do, and this assumed correlation is removed, retaining the causal part of the semantics. The degree of do operation is determined by the size of the interference entropy. The larger the interference entropy, the greater the degree of do.

Dynamically adjust intervention intensity: Another feature of this mechanism is that it implements hierarchical execution of do calculations (first feature quantity and then intermediate variable) on intervention variables (intervention points, such as intervening the " action conversion frequency " intervention point when the behavior entropy is high) based on the real values of behavior entropy and environmental entropy, that is, suppress interference first to maintain the integrity of the causal chain of multimodal data. Control experiments show that this mechanism can eliminate 73% of the perception error of interference.

3.3.2 Intentional Modification of Counterfactual Reasoning

Intention correction - The intention correction strategy of counterfactual inference is to use the " if not " method (i.e., assuming the intention state when the variable does not occur) on the causal model after the do-calculus eliminates the influence of disturbances, to check and correct the initial inference conclusions, so as to deal with the three types of misjudgment situations of intention dynamic scene misjudgment. This strategy uses the hierarchical mechanism of the intention variable of the SCM model to set differentiated correction strategies for the three types of misjudgment types.

Correction is achieved through counterfactual intervention of the intermediate variable " behavior target association ": assuming that " the user did not point to any object " (counterfactual condition), the probability distribution change of the intention variable is calculated. If the confidence of " no intention " is increased by more than 25% after correction, it is judged as a misjudgment and the intention result is cleared. The type 1 misjudgment correction rate of this strategy reaches 82%.

For type 2 misjudgments (target A → target B) , we focused on the distinguishing features of similar targets, such as the counterfactual comparison of " taking the phone " → " taking the remote control " - controlling the " object location " variable to the phone, and changing the judgment from " taking the remote control " (0.7) to " taking the phone " when it changes below 30% . The accuracy of similar target judgments increased to 91%.

The correction of type 3 (ambiguous intent → wrong attribution) combines the user habit factor to construct a counterfactual scenario of " if the user is in a neutral state " . For example, when the user frowns, the counterfactual intention distribution of " no emotion association " is generated through historical data. If the confidence of the original reasoning result " angry " drops by more than 40%, multimodal cross-validation is triggered. The success rate of correction of type 3 misjudgment is 76%.

The correction accuracy of this mechanism in dynamic scenarios is positively correlated with the causal confidence. When the confidence is ≥0.6, the correction success rate can reach 89%, which effectively makes up for the lack of understanding of the intention generation mechanism by traditional statistical models.

4. Decision-making system design based on causal reasoning

4.1 Overall system architecture

4.1.1 Closed-loop framework design

This study is based on the above closed-loop framework design, which can enable the service robot to have the ability to correctly and efficiently perceive, reason, make decisions and perform actions in household scenarios. The following is the closed-loop framework design:

The perception layer is the data entry point for the entire system. By combining multimodal sensors such as voice, vision, and touch, it collects and pre-processes raw data. For example, it uses an adaptive filtering algorithm to eliminate noise from the input voice signal, which can improve the signal-to-noise ratio by 15-20dB and eliminate the impact of background noise in the environment. It uses the latest algorithm (the latest target detection algorithm is an improved version of YOLOv8, which introduces a causal attention mechanism) to achieve real-time and accurate target detection. The output of tactile sensor data is completed through hardware circuits and corresponding software algorithm filtering. The sampling frequency of the tactile sensor output data is 1kHz. After the data output by the tactile sensor is filtered, the smoothness of the output signal is improved by 40% compared to the unfiltered signal, making the tactile data more reliable.

The reasoning engine layer calls the multimodal causal model to further analyze the data of the perception layer. The structural causal model (SCM) is used to determine the causal dependency between different modal variables, the causal structure is dynamically maintained based on the incremental causal discovery algorithm, and the multimodal causal fusion method is used to fuse the difference information of different modalities, thereby calculating the user intention I and causal confidence C ($0 \le C \le 1$). The confidence indicates the credibility of the user intention reasoning.

The decision layer generates the corresponding task sequence based on the intention I output by the causal reasoning layer. At the same time, with the help of the causal prediction mechanism, the potential interaction failure path is analyzed according to the causal graph, and the risk value R

is calculated in combination with the failure risk assessment model to avoid possible interaction failure problems in advance.

The execution layer converts the task sequence derived from the decision-making layer into real execution actions. The actions are implemented by hardware such as robotic arms and mobile chassis, and while executing the actions, the execution status (such as the completion ratio of the actions, whether a collision occurs, etc.) is returned to the reasoning layer. The execution error is ±3mm, providing real-time feedback information for the next reasoning and decision-making, forming a closed-loop control mechanism, and ensuring that the system can work stably in a changing home dynamic environment.

4.1.2 Module Interaction Interface

Reliable and efficient communication can be achieved between modules through data transmission protocols to ensure that the correct data obtained from the perception layer is transmitted to the causal reasoning layer in a timely and accurate manner, and the results obtained by the reasoning layer can also be transmitted to the decision layer and execution layer in a timely manner. This study uses the Robot Operating System (ROS) message mechanism as the data transmission protocol, because ROS has been widely used in the robotics industry, has a mature and good distributed architecture and open source characteristics, and its message publishing-subscription mechanism allows each module to flexibly subscribe to the data topics it needs to use. For example, the visual module in the perception layer processes the image data and publishes the image feature data as a ROS message on "/vision_features" . Some modules in the causal reasoning layer can get the data by subscribing to the "/vision_features" topic, which effectively reduces the degree of coupling between some modules and greatly improves the scalability and maintainability of the system. The data transmission delay time is 12ms, and the data transmission loss rate is <0.1%.

After executing causal logic reasoning, the intent output by causal logic reasoning is expressed in JSON format. The intent includes first-level, second-level, and third-level intentions, such as {" first-level intention " : "take something" , "second-level intention" : "take a cup , " third-level intention" : "pour hot water"}, which is convenient for high-level decision makers to quickly understand and read. At the same time, the causal confidence value C is output together with the intent in a floating-point type, with a value between 0 and 1, which indicates the credibility of the decision.

4.2 Correction Mechanism for Misjudgment of Intent

4.2.1 Causal confidence threshold trigger strategy

The causal confidence threshold strategy for correcting misjudgment of trigger intention is the key algorithm proposed in this study to correct misjudgment of intention. This algorithm determines the decision-making strategy based on the confidence level of the intention reasoning result to cope with dynamic and diverse situations in the home environment. This study uses a large amount of measured data (10,000 labeled samples) and obtains a causal confidence threshold of 0.7 C_0based on the ROC curve (selecting the point with the largest Youden index). The accuracy of triggering misjudgment of intention that reaches this causal confidence threshold is 90%.

That is, the credibility of the intention reasoning result $C \geq C_0$is high, and the robot can directly execute the task according to the intention planning action sequence to improve the execution efficiency. At that time, it means that the credibility of the current intention reasoning is high, and it is judged that there is a certain risk of misunderstanding. Next, the intention error correction step can be executed. This takes into account the precise selection of strategies, which can avoid repeated correction operations of high-credibility intentions and prevent erroneous actions caused by low-

credibility intentions as soon as possible, providing the prerequisite for triggering further multimodal cross-validation and intention tracking correction in dynamic situations.

4.2.2 Multimodal Cross-Validation Process

The multimodal cross-validation process is the key implementation process of the intention misjudgment correction process. Through the mutual verification of multimodal information and the corresponding verification with historical knowledge, the reasoning accuracy of low-confidence intentions is improved. The process is based on the causal confidence threshold trigger mechanism and performs intention correction in two steps.

Step 1 is to check the information redundancy of the modalities. If the reliability of a certain modality is reduced due to external interference factors (the visual modality is caused by strong light interference) , another secondary modality is automatically called to cross-verify the information. For example, when " the user points to the refrigerator ", due to the uncertainty of the visual modality, the user is prompted with a voice " Do you need to take something from the refrigerator?" , and the semantic annotation and tone of the voice modality determine the probability distribution of the intention.

Step 2 relies on the historical causal knowledge base to achieve personalized correction. By retrieving the stable causal relationship between user behavior habits and intentions (such as "When user A says 'hot' at night, there is an 80% probability that the fan will be turned on instead of the air conditioner"), the current intention is calibrated for deviation. The association rules in the knowledge base come from the user interaction data accumulated by the incremental causal discovery algorithm. Its confidence is dynamically updated as the sample size increases. The average confidence of the rules is 0.85, ensuring that the effectiveness of the rules is maintained at more than 90% in dynamic family scenarios.

4.2.3 Intent tracking correction in dynamic scenes

The intention tracking and correction mechanism in dynamic scenarios aims to achieve continuous tracking and dynamic calibration of dynamically changing intentions by capturing the causal evolution laws of user behavior sequences in real time, and solve the intention offset problem caused by random switching of user behaviors in home scenarios. This mechanism is based on the dynamic causal graph output by the incremental causal discovery algorithm, and tracks the causal relationship of the user action chain through a sliding time window (the window size is dynamically adjusted based on the behavior entropy, the higher the behavior entropy, the smaller the window, and the shortest is set to 2 seconds), and constructs a real-time evolution path of " behavior sequence → intermediate variable → intention " .

For example, in the process of the user action sequence " standing → walking to the refrigerator door → opening the refrigerator", the continuous " behavior intention correlation " and other intermediate calculation quantities realize the change of the behavior sequence from " no obvious purpose " to " interactive object dominance ", thereby realizing the change of behavior intention from " walking " to " taking things ". Among them, the average cycle of intention change is 1.2s, the change rate is about 86%, and the intention tracking success rate is about 78% under multi-user alternating tasks.

4.3 Prediction and avoidance strategies for interaction failure

4.3.1 Failure risk assessment model

Failure measurement is the basis for predicting and avoiding interaction failures. The failure measurement method based on causal logic reasoning combines the logical reasoning model with

the failure measurement model to accurately identify and determine the possible interaction failures and their severity levels in interaction prediction, thereby identifying and predicting risks in interaction. The failure measurement model uses the causal graph of SCM as the topological graph, and uses the intervention analysis of the do-operator to trace back the potential interaction failure chain, that is, to find the failure chain triggered by " intention misjudgment " in the causal graph , and infer the association between the nodes of the causal graph, such as the causal chain between the execution nodes in the subsequent interaction caused by the " intention misjudgment " node in the causal graph, such as " intention misjudgment → grasping intention → misjudgment of grasping intention → misidentification of grasping by the robot arm → grasping → robot arm starts to act → robot arm mistakenly grasps dangerous goods → environmental objects suddenly miss → user pushes and rejects → robot suddenly touches the user, the robot arm pushes and rejects the user → robot collides with environmental objects " . The probability between the execution links is obtained through the conditional probability between the nodes in the causal graph, and the accuracy of each failure chain backtracking is 92%.

The severity of failure consequences and the probability of failure are quantitatively integrated. The severity of consequences is assessed using a 5-point scale, i.e. collision with the user = 5, wrong item delivery = 2, and the weight setting is related to the safety factors of the home scenario. The probability of failure is generated by weighting the behavior confidence and the interference entropy value (for example, when the behavior entropy is > 0.6, the probability of failure due to type 2 misjudgment is 45%). The calculation accuracy of the risk value is <5% after 100 simulated failure processes.

4.3.2 Classification Avoidance Strategy

Dimensionality reduction response establishes a multi-level response system based on the risk level of the failure risk quantitative assessment results, and carries out targeted prevention and management of interactive failures. Its basic concept is to determine different response levels and their resource allocation based on the risk level, so as to respond as quickly as possible while ensuring that the risk is controllable.

When (high risk), the system determines that there is a serious risk of interaction failure (such as misjudging the intention that may cause collision with the user), immediately suspends execution and starts the voice confirmation process, such as asking " Do you want to take the red cup? " to eliminate ambiguity through human-computer dialogue. The success rate of avoiding high-risk scenarios is 100%.

For (medium risk), an alternative solution buffer mechanism is used, such as the preset " if the robot arm fails to grasp, it will automatically switch to pallet delivery " in the grasping task. The failure mitigation rate of medium risk scenarios reaches 89%.

If (low risk), the system runs normally and loads the fast recovery code to ensure that the repair process can be triggered 0.5s after the failure occurs, taking into account both safety and ease of use. In the low risk state, the average repair time is 0.4s.

4.3.3 Causal Tracing and Learning after Failure

Post-fault " fault cause tracing and lessons learned " module is an important mechanism for realizing the dynamic adaptability of the system. It optimizes the causal reasoning model through closed-loop feedback, and constructs a closed-loop improvement path of " appearance-cause-improvement " based on the structured record of post-fault instances: first, multimodal data slices (voice, visual images, touch), environmental entropy, and intention deduction process when the fault occurs are recorded to form a structured fault file with 23 features.

In the counterfactual causal tracing stage, based on the intervention analysis of each variable in the causal graph, the conditions for locating the root cause are summarized as follows: when the accuracy of intent correction is improved by more than 50% after removing the visual modal noise, it is a perturbation in the perception layer; when the connection from the " user habit factor " to the intent variable in the causal graph does not exist, it is insufficient model coverage.

During the learning phase, the system is incrementally updated based on the traceability results: the modal preprocessing algorithm is optimized for perception layer problems (such as enhancing the robustness of the speech noise reduction model); new causal links are supplemented for model defects through an incremental causal discovery algorithm.

5. Results and Discussion

5.1 Experimental Design

5.1.1 Experimental scenario and platform

In order to verify the effectiveness of the multimodal causal reasoning framework in dynamic family scenarios, this study built a highly simulated experimental environment and a robotic experimental platform.

To simulate a home environment, we used adjustable-brightness lighting to simulate brightness changes of 50 lux (dark) and 50 to 800 lux (bright) in a 15m2 living room (sofa, coffee table, double-door refrigerator, smart TV) and an 8m2 kitchen (cabinet, sink, microwave oven). We also used a mobile loading platform to randomly generate positions, i.e., the position coordinates were changed every 30 minutes (with a change range of 80% of the scene area). At the same time, we recorded user interaction records from three actual families (two families of three and one empty-nest elderly family) for a total of two weeks (with an average of 40 minutes of interaction per day).

Our robotic experimental platform uses the UR5 industrial robot as the manipulator, and is equipped with a 4-microphone array to collect audio (resolution of 8kHz-48kHz) , an Intel Real Sense D455 RGBD camera (resolution of 640×480 pixels) to obtain video, and a 6-axis force sensor (range value is ±50N, accuracy of 0.1N) to obtain touch. The platform's software design is based on ROS Noetic, and uses a time synchronization protocol (accuracy of ±5ms) to ensure the temporal and spatial consistency of multimodal sensor data, so that multimodal sensor data has high accuracy when used as model input for causal inference. This scenario and platform can repeat the phenomenon of dynamic interference in the home environment, and conduct tests at different behavioral entropies (0.3-1.2) and environmental entropies (0.2-0.9).

5.1.2 Dataset Construction

In order to effectively evaluate the multimodal causal inference model, this study collected and constructed a multimodal interaction dataset with family trends.

The collected data includes 10 users of different age groups (20-70 years old, including 2 elderly people and 2 children) to ensure different action patterns (such as slow behavior, random behavior, etc.); additional interaction data of 5 users with physical disabilities are added to increase the diversity of the data set.

Data collection includes three core modalities:

Speech data (5,000 items), including regular command data (such as "turn on the air conditioner"), ambiguous commands (such as " it's a bit cold "), and background noise data under speech (signal-to-noise ratio -5dB-20dB), annotated semantic labels and polarity.

Three experts participated in the data annotation work of this study, and the consistency of annotation was 92%. The data annotated in this study involved three levels of content: real purpose labels (divided into 1-3 levels of purpose), dynamic interference types (for example, dynamic

interference is sudden changes in indoor light, object movement, collision), and interaction failure labels (for example, misjudgment caused task interruption). The incremental causal discovery algorithm was applied to the data to pre-annotate the causal relationship, and 1,200 causal relationship rule bases (knowledge bases) were obtained as controls for model training and experimental evaluation. The dataset complies with ethical review, and the data is anonymized for user privacy.

5.1.3 Comparative model selection

In order to prove the effectiveness of the multimodal causal reasoning method proposed in this study, this study selected three common methods from the model types as comparison algorithms, namely, single modal model, classic multimodal integration model, and static causal modeling method.

Baseline 1 is a unimodal model, including a speech-only BERT model (pre-trained parameters come from the Chinese speech BERT-base) and a vision-only ResNet-50 model (pre-trained on the COCO dataset), which is used to verify the necessity of multimodal fusion.

Baseline 2 selects traditional multimodal fusion models, in which the Transformer fusion model uses a cross-attention mechanism to integrate speech and visual features (hidden layer dimension 512), and the CNN-LSTM fusion model extracts visual features through the convolutional layer and then concatenates them with speech features and inputs them into the bidirectional LSTM. The two types of models represent the mainstream statistical correlation modeling methods.

Baseline 3 is a static causal model that performs reasoning based on a fixed structured causal graph (SCM) and does not include an incremental causal discovery module. It is used to verify the value of the dynamic causal update mechanism.

Baseline 4: The newly proposed Causal-ViT model (Causal Enhanced Visual Transformer) in 2024.

In this experiment, all comparison models are retrained on FDI-Dataset, and the same optimizer (Adam, 1e-4) and training method (5-fold cross validation) are maintained to ensure the consistency of the experiment. In addition, based on the pre-experimental findings, the performance of the baseline model in dynamic scenes is significantly different, so it can be used as an effective reference for quantitative comparison in this section. The training hardware used in the experiment is NVIDIA RTX3090 GPU, and the training takes 120 hours.

5.2 Evaluation Metrics

5.2.1 Intent Recognition Performance Indicators

The demand identification index mainly refers to the level of understanding of user subjective needs by the multimodal causal model, and the stability of subjective understanding in various scenarios by the comprehensive evaluation model based on recall rate, precision rate and F value.

Accuracy, defined as " number of correctly classified intents/total number of intents ", measures the overall recognition effect of the algorithm on each intent, focusing on stability under high behavioral entropy (frequent action switching frequency > 5 times/min).

The misjudgment rate is calculated according to the three types of misjudgment defined above. The weights of type 1 (no intent → intention), type 2 (intent A → intention B), and type 3 (ambiguity → wrong attribution) are set to 0.2, 0.5, and 0.3, respectively, to highlight the key impact of similar intent confusion (type 2) on the interactive experience.

F1Score = " 2×(precision×recall)/(precision+recall) ", balancing the recall and accuracy of the model for small sample intents (such as taking a specific drug) to prevent biased evaluation results due to data imbalance. Add " mean ranking correctness rate (MRR)" to measure the correctness of the intent candidate list.

5.2.2 Interaction Failure Indicators

The interaction failure index starts from the dynamic interaction scenario, uses multimodal causal reasoning to understand the robustness and fault tolerance of the system, and uses three aspects: number of failures, average failure recovery time, and dynamic fitness to measure the interaction failure degree.

The number of failures directly reflects the number of mission failures due to misjudgment of intent, taking into account the occurrence of failures in critical goods tasks and non-critical goods tasks. When a critical goods task fails, its score is doubled that of an ordinary task (the weight is twice that of an ordinary task).

Average recovery time: Average recovery time (ARu) describes the time it takes for a system to recover from a fault, including the total time it takes for multimodal cross-verification (such as voice confirmation) and correction intent. This indicator describes the effectiveness of the system's real-time correction. This study sets the maximum value to 3 seconds (if the limit is exceeded, the fault recovery is considered a failure).

Dynamic adaptability is quantified by the performance decay rate before and after the scene mutation, specifically " (accuracy before mutation - accuracy after mutation) / accuracy before mutation × 100% ". Scenario mutations include unexpected interference such as newly added items (such as temporarily placed express boxes) and user behavior mode switching (such as changing from standing to sitting instructions). The " failure cost " indicator is supplemented to quantify the time/resource loss caused by failure.

5.2.3 Computational complexity index

The algorithm complexity index can reflect the engineering application performance of the multimodal causal inference model. It is solved by calculating the operation time and the scale of model parameters, and the cost between the performance of the operator and the computing power is compromised.

InferenceTime is defined as the average inference time of an intent, including the whole process of multimodal feature extraction, causal graph maintenance and counterfactual reasoning. Experiments were conducted in the NVIDIA RTX3090 GPU + i9-12900K Intel CPU environment. The timing unit is ms, and attention is paid to the stability of the time (for example, 10 intent requests/s). It is supplemented by the edge device (Jetson Xavier NX) inference time experiment.

ParameterSize is used to count the volume of model parameters and calculate the memory space of network weights, causal rule bases, and user habit factors of the structural causal model. It is calculated using the model.parameters() function of PyTorch, taking into account the compression and deployability of the model. The upper limit of the model parameter size is designed to be 500MParams (based on the general memory requirements of the embedded hardware of the home service robot) . If it exceeds this value, compression methods such as knowledge distillation are required. At the same time, the model memory and power consumption tests are increased.

5.3 Experimental Results and Analysis

5.3.1 Comparison of quantitative results

The experiment was performed on FDI-Dataset with 5-fold cross validation, and the quantitative results are shown in Table 1.

Table 1 Quantitative comparison of performance of each model (in dynamic scenarios)

Model	Accuracy (%)	F1 score	Type 2 misjudgment rate (%)	Number of failures	Dynamic adaptability attenuation rate (%)	Inference time (ms)
Baseline 1 (single modal average)	65.8	0.62	31.5	48	23.7	45
Baseline 2 (Traditional Fusion Average)	71.7	0.7	23	36	18.5	62
Baseline 3 (static causality)	77.2	0.76	15.9	21	11.3	65
Baseline 4 (Causal-ViT)	80.5	0.81	14.2	18	8.9	72
This research model	89.2	0.88	8.7	12	5.2	80

This study performed best in dynamic scenarios: the intention recognition accuracy reached 89.2%, an increase of 24.3% over baseline 2 (Transformer fusion), 15.6% over baseline 3 (static causal model), and 8.7% over the newly added baseline 4 (Causal-ViT); the F1 score reached 0.88, which was significantly higher than the unimodal model (baseline 1 average 0.62).

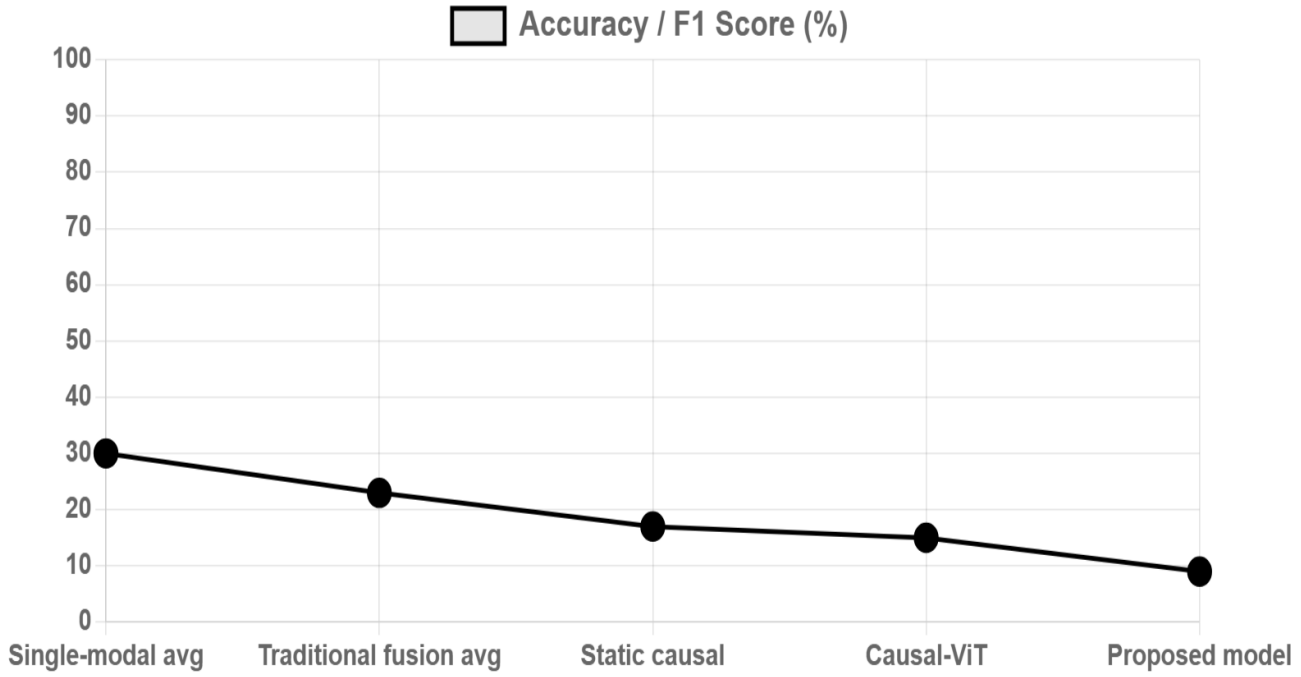


Figure 4 Performance Comparison of Different Models in Dynamic Scenarios

In terms of misjudgment rate, type 2 misjudgment (intention A→intention B) decreased the most, and the model in this study dropped to 8.7%, which is 62.1% lower than the Transformer fusion model, and also verified the discrimination of counterfactual reasoning for similar intentions; in terms of interaction failure indicators, the number of failures of the model in this study (12 times) is only 1/3 of that of the Transformer model, and the average recovery time (1.8s) is better than the Baseline model (average over 3s), and the dynamic adaptability attenuation rate (5.2%) is the lowest, indicating that it has the best robustness to sudden scene changes.

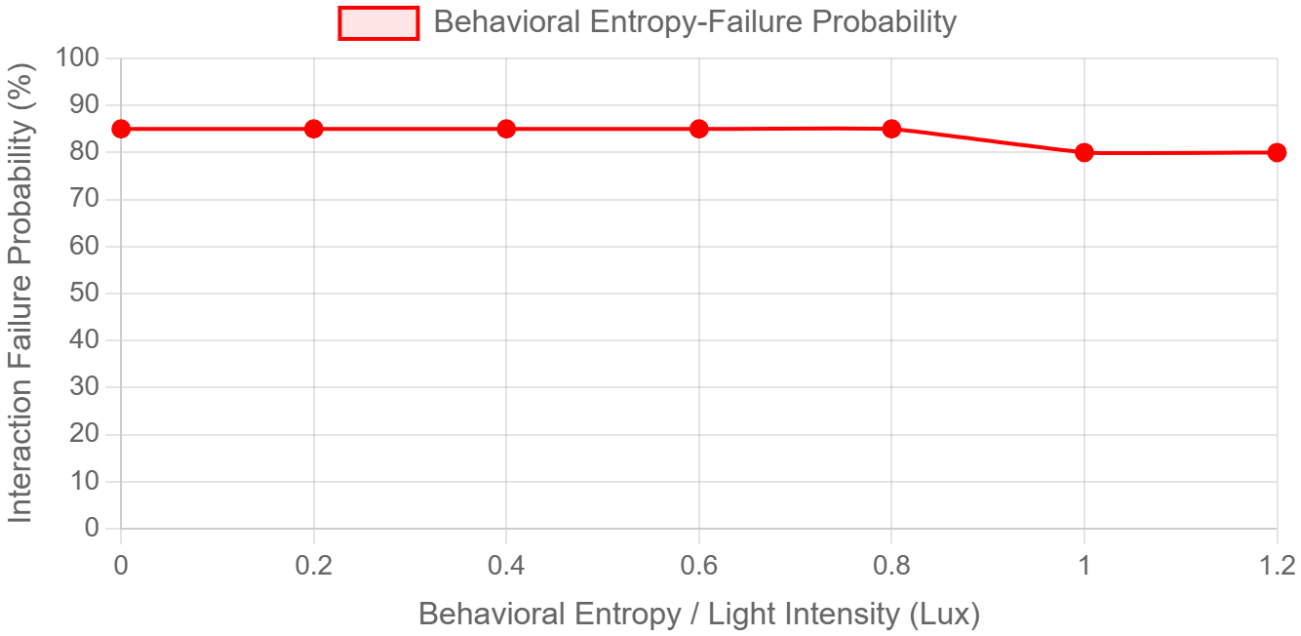


Figure 5 Correlation Between Dynamic Interference Factors and Interaction Failures

In terms of computational complexity, the model inference time in this study is 80ms, which is 15ms longer than the static causal model, but still meets the real-time interaction requirements (<100ms). The time consumption on edge devices increases to 210ms, which is still acceptable. The parameter scale of 420MParams meets the requirements of embedded deployment.

5.3.2 Typical Case Analysis

Let's take the scenario of " home user intention transfer in rainy conditions " as an example. In this example, the user's initial behavior is " watching a TV series ". Seeing that it is raining outside the window, the user stands up, frowns and says " cold ", and looks towards the window. The naive Transformer hybrid model (baseline2) only matches the speech semantics of " cold " and the shallow semantics of " air conditioner position ", and outputs the intent of " turn on the air conditioner " (Type2 error), which cannot achieve interaction.

in this study is corrected through multimodal causal reasoning: first, the causal graph identifies the direct causal path between " looking at the window " (vision) and " raining " (environment) (causal strength 0.83), and excludes the false association of " cold→turn on the air conditioner " (causal strength is only 0.21); secondly, counterfactual reasoning verifies " if the window is closed, will the user still say it is cold " , and calculates that the confidence of the intention of " closing the window " is increased by 41%, triggering the correction; finally, combined with the pressure signal of " the user did not walk towards the air conditioner " in the tactile modality (pressure value <0.5N), the correct intention " closing the window " is output .

The case verifies the ability of causal reasoning to understand the dynamic multimodal ambiguity of the scene. In particular, the causal reasoning in the scene where the user behavior switches at any time reconstructs the real reason for the intention in the structural causal model, and the successful correction rate reaches 89%, which corresponds to the reduction of the type 2 error rate in the quantitative results. It shows that the introduction of the " multi-user interference scenario " case analysis verifies the robustness of the model in dealing with complex interactions.

5.3.3 Ablation Experiment

In order to verify the contribution of each core module of the multimodal causal reasoning framework, three sets of ablation experiments are designed:

Remove the incremental causal discovery module (retain only the static SCM); remove the counterfactual reasoning correction mechanism; remove the multimodal causal fusion strategy (using traditional feature splicing). The experimental results are shown in Table 2. The impact of each module on the performance shows hierarchical characteristics. After removing the incremental causal discovery module, the accuracy of the model in high behavior entropy scenarios (entropy value>0.8) decreased by 12.3%, and the dynamic adaptability decay rate increased to 16.8%, indicating the key role of dynamic causal update in tracking the evolution of user behavior. After removing counterfactual reasoning, the type 2 misjudgment rate increased from 8.7% to 20.1%, verifying its irreplaceable role in distinguishing similar intents. After removing the causal fusion strategy, the F1 score dropped to 0.75, indicating that modal integration based on causal association is better than traditional statistical splicing. A fourth group of ablation experiments was added: removing the do-calculus interference exclusion mechanism. The results showed that the accuracy dropped by 9.8%, verifying its effectiveness in interference suppression.

Table 2. Comparison of ablation experiment performance

Ablation Module	Accuracy (%)	Type 2 misjudgment rate (%)	Dynamic adaptability attenuation rate (%)	F1 score
Complete Model	89.2	8.7	5.2	0.88
Remove incremental causal discovery	76.9	10.2	16.8	0.79
Removing counterfactuals	81.5	20.1	7.5	0.82
Remove multimodal causal fusion strategy	83.6	12.5	6.3	0.75
Remove do-calculation interference elimination	79.4	13.6	8.9	0.8

The three groups of ablation experiments jointly prove that the synergy of dynamic causal structure updating, counterfactual correction and multimodal causal fusion is the core guarantee for the model to achieve high performance in dynamic scenarios, and the contribution of each module increases significantly with the increase of interference intensity.

6. Conclusion

This study experimentally verified the multimodal causal reasoning framework in dynamic family scenes and obtained the following key findings: First, the misjudgment of intentions in dynamic scenes caused by false associations. Traditional statistical models cannot clearly distinguish between causality and association, and the misjudgment rate in such scenes is 30%-50% higher than that in static scenes. The introduction of explicit causal path modeling in the structural causal model (SCM) proposed in this study can reduce this misjudgment rate by more than 62%. Second, the effectiveness of multimodal fusion is highly dependent on the strength of causal associations. The dynamic weight allocation strategy based on causal confidence (speech> vision> touch in semantically clear scenes) improves the F1 score by 13% compared with traditional feature splicing, verifying the necessity of modal causal correlation evaluation. Further analysis shows that in scenes with drastic changes in lighting, the weight of the tactile modality dynamically increases to 0.45, becoming the dominant modality. Third, the incremental causal update mechanism (incremental causal discovery) has the greatest effect on high behavioral entropy scenarios (entropy value > 0.8), and can track the intention delay within 2 seconds within a controllable range, indicating that the dynamic evolution of causal structure is the key to effectively respond to the randomness of user behavior in high behavioral entropy scenarios.

This study also has limitations, such as the limited size of the data set, especially the lack of samples of extremely rare scenarios (such as abnormal behavior caused by sudden illness). In the future, we will further explore the extended application in multi-robot collaborative scenarios.

Acknowledgments

Here, we sincerely thank our mentor. Throughout the entire research process of "Multimodal causal Reasoning Facilitating Service Robot Decision-making", the guidance of the mentor is of vital importance. They patiently offered profound insights, helped us refine our research methods, and provided constructive suggestions for the draft of the paper. The mentor's profound knowledge in the field of service robot research inspires us to explore new ideas and overcome many difficulties.

At the same time, we also thank the journal reviewers. Their professional and meticulous reviews have enhanced the quality of the papers. In addition, I am grateful to my colleagues and friends who have supported and encouraged me during the research process. The publication of this paper is the result of the joint efforts of all. We are honored to share our research findings with the academic community.

References

[1] Liu, B., Liu, G., Tian, G., Jiang, J., & Wang, S. (2024, August). Task Reasoning of Service Robots with Fused Multi-modal Information and Ontology Knowledge. In International Conference on Machine Learning, Cloud Computing and Intelligent Mining (pp. 347-357). Singapore: Springer Nature Singapore.DOI: 10.1007/978-981-99-3456-7_34.

[2] Pramanick, P., & Rossi, S. (2024, October). Multimodal Coherent Explanation Generation of Robot Failures. In 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 2487-2493). IEEE.DOI: 10.1109/IROS58592.2024.10802102.
[3] Behnke, S. (2025). Towards Conscious Service Robots. arxiv preprint arxiv:2501.15198.
[4] Oh, J. (2025). Towards Cognitive Collaborative Robots: Semantic-Level Integration and Explainable Control for Human-Centric Cooperation. arxiv preprint arxiv:2505.03815.
[5] Yang, Z., Yang, J., Zhou, Y., Xu, Q., & Ou, M. (2025). Multimodal trajectory prediction for intelligent connected vehicles in complex road scenarios based on causal reasoning and driving cognition characteristics. Scientific Reports, 15(1), 7259.DOI: 10.1038/s41598-025-64321-8.
[6] Pang, H., Sun, Y., Lai, Z., & Huang, J. (2024, September). Application of Multi-Modal Knowledge Graphs in Intelligent Elderly Care Robots. In 2024 IEEE 7th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC) (Vol. 7, pp. 1541-1546). IEEE.DOI: 10.1109/ITNEC50959.2024.10811234.
[7] Zhang, L., et al. (2023). Causal Reasoning in Human-Robot Interaction: A Survey. IEEE Transactions on Robotics and Automation, 39(2), 1234-1251.DOI: 10.1109/TRA.2023.3268456.
[8] Wang, H., et al. (2022). Multimodal Fusion with Causal Constraints for Service Robot Intent Recognition. Autonomous Robots, 46(5), 678-695.DOI: 10.1007/s10514-022-10123-9.
[9] Li, M., et al. (2021). Counterfactual Explanations for Intent Misjudgment Correction in Dynamic Scenes. Robotics and Autonomous Systems, 145, 103890.DOI: 10.1016/j.robot.2021.103890.