Towards Building Condition-Based Cross-Modality Intention-Aware Human-AI Cooperation under VR Environment

Ziyao He
marsroscope@stu.xjtu.edu.cn
Xi'an Jiaotong University
MOE KLINNS Lab
Xi'an, Shaanxi, China

Shiyuan Li
lsy1170574738@stu.xjtu.edu.cn
Xi'an Jiaotong University
MOE KLINNS Lab
Xi'an, Shaanxi, China

Yunpeng Song
yunpengs@xjtu.edu.cn
Xi'an Jiaotong University
MOE KLINNS Lab
Xi'an, Shaanxi, China

Zhongmin Cai*
zmcai@sei.xjtu.edu.cn
Xi'an Jiaotong University
MOE KLINNS Lab
Xi'an, Shaanxi, China

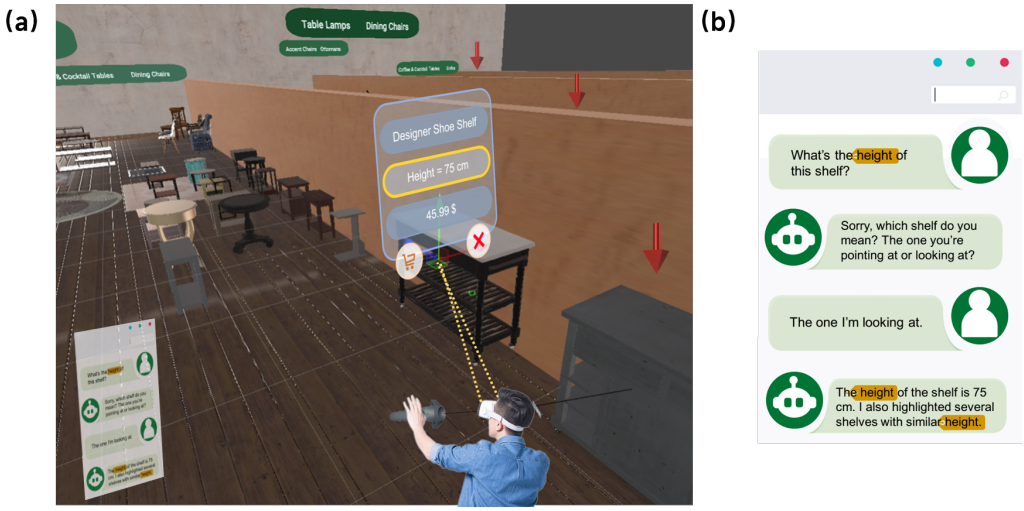

Figure 1: (a) The VR shopping system that follows our designed cross-modality intention-aware human-AI cooperation framework. The utilized interactions include dialogue interfaces (pop-up in the lower-left corner of the user's view), float labels (highlighting user interest), controller operations (enabling movement and selection), and red-arrow indicators (positioning and navigation). (b) The chatbot interface pops up when users talk with the intelligent agent within the system. Note that the highlighted "height" is for demonstration only, and is not displayed during user study.

ABSTRACT

To address critical challenges in effectively identifying user intent and forming relevant information presentations and recommendations in VR environments, we propose an innovative condition-based multi-modal human-AI cooperation framework. It highlights the intent tuples (intent, condition, intent prompt, action prompt) and 2-Large-Language-Models (2-LLMs) architecture. This design, utilizes "condition" as the core to describe tasks, dynamically match user interactions with intentions, and empower generations of various tailored multi-modal AI responses. The architecture of 2-LLMs separates the roles of intent detection and action generation, decreasing the prompt length and helping with generating appropriate responses. We implemented a VR-based intelligent furniture purchasing system based on the proposed framework and conducted a three-phase comparative user study. The results conclusively demonstrate the system's superiority in time efficiency and accuracy, intention conveyance improvements, effective product acquisitions, and user satisfaction and cooperation preference. Our framework provides a promising approach towards personalized and efficient user experiences in VR.

*Corresponding Author

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642360

CCS CONCEPTS

• **Human-centered computing → User interface programming**; *Ubiquitous and mobile computing theory, concepts and paradigms*; **Ubiquitous and mobile computing systems and tools**.

KEYWORDS

Virtual Reality, Human-AI Cooperation, Intention Detection, Action Generation

ACM Reference Format:

Ziyao He, Shiyuan Li, Yunpeng Song, and Zhongmin Cai. 2024. Towards Building Condition-Based Cross-Modality Intention-Aware Human-AI Cooperation under VR Environment. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3613904.3642360

1 INTRODUCTION

Nowadays, an increasing number of human-AI cooperation systems are emerging within the Virtual Reality (VR) environment to boost task completion efficiency. Effectively harnessing AI's multi-modal capabilities in interactions to facilitate productive human-AI cooperation has become a critical area of research. To enable efficient human-AI cooperation, two key tasks must be addressed in this process. Firstly, it is essential to accurately ascertain users' specific intentions from various input modalities, which can then be translated into actionable AI objectives. Secondly, based on the recognized intentions, the design of AI's information presentation must be tailored to assist decision-making. For instance, in a VR shopping scenario, these tasks encompass inferring the user's true purchasing needs or preferences based on interactions and subsequently offering relevant product recommendations or adjusting the presentation of information to aid browsing and decision-making. However, in today's information-saturated world, AI's filtering and presentation methods may not always align perfectly with user needs [13]. Overabundant results can burden users with secondary filtering, while overly specific recommendations limit comparison and potentially lead to unsatisfactory purchases. Therefore, AI needs to adapt its presentation methods based on user intent and behavior to efficiently provide essential and suitable information.

To effectively identify user intent and form relevant information presentations and recommendations, we identify three main challenges: user expression ambiguities, AI recognition and recommendation inaccuracies, and tailored response adjustments regarding users' varied expressions.

First, user expression can be ambiguous, relying on implicit knowledge or intuitive assumptions, such as "Items that fit in my car's trunk". More importantly, users often leverage information from multiple modalities in VR, such as "the item I see". This diversity makes it challenging for AI to comprehend user intents when lacking contexts, leading to user frustration and multiple revisions to clarify their intent. Secondly, from the perspective of AI recognition and recommendation, generating AI responses aligned with user intent poses another challenge. Taking the Large Language Model (LLM) as an example [5], its reasoning capabilities and response appropriateness are influenced by the prompt's quality such as length [12], which can result in response deviations and difficulty in control. For instance, when recommending products, overly lengthy prompts with comprehensive product information may yield irrelevant or off-intent recommendations. Lastly, dynamically adjusting tailored responses based on the changing user expressions during interactions is crucial. This adaptation should consider two aspects: First, users may adjust their intent based on AI feedback, such as altering search criteria or adding constraints. Second, users may modify ways of expression as interactions with AI progresses. Users would accordingly adjust their interaction strategies as their awareness of AI capabilities enhances, which is closely related to mental model updates [3, 18].

Regarding the aforementioned challenges, we introduce a condition-based cross-modality intention-aware human-AI cooperation framework. In this framework, we propose the construction of an intent library in the form of intent tuples *(intent, condition, intent prompt, action prompt)*, and the 2-LLMs (Intent LLM and Action LLM) architecture for intent recognition and action generation. Utilizing the intent library, Intent LLM identifies users' precise intent from multi-modal interactions in the form of matching intent tuples. Subsequently, according to the condition-based user intentions, the system could perform filtering and retrieve qualified results (such as products). Then the Action LLM is responsible for coordinating the dynamic generation of multi-modal AI responses tailored to the user's intent, including visual presentations, dialogues, and directional cues. This design takes a task-oriented approach closely aligned with human-mental-model [18], using "condition" as the core to describe tasks and generate various forms of cues, guiding users to quickly understand AI capabilities and express their needs clearly. Simultaneously, the architecture of 2-LLMs breaks down the mission into sub-tasks: intent detection and action generation, and performs filtering before action generation, hence decreasing the prompt length for each LLM to handle and helping with constraining LLM's output. Moreover, this framework is guided by changes in "conditions", referencing users' multi-modal context. Consequently, it adeptly aligns evolving user expressions or intent shifts with the coded intent tuples in a cohesive manner, facilitating better determinations of the user's current intent and timely adjustments.

To validate the effectiveness of condition-based human-AI cooperation, we implemented a furniture purchasing system in VR environment and composed a three-phase comparative user study. Objective results demonstrate that the system with condition-based human-AI cooperation mechanism outperforms traditional user interaction modes in time efficiency and accuracy, and helps with detecting user's multi-modal intentions and reducing the LLM's inappropriate responses to some extent. Analysis of user behaviors bolsters the evidence of the system's proficiency in assisting effective conveyance of user intentions and facilitating the acquisition of expected products. Moreover, subjective questionnaires attest to heightened satisfaction, reduced physical load, and improved perceived operational efficiency.

In summary, this paper's main contributions conclude as follows:

- We propose a condition-based cross-modality intention-aware human-AI cooperation framework in VR environment. It leverages the intent library of (intent, condition, intent

prompt, action prompt) tuples and the 2-LLMs architecture to accurately identify user intentions and generate tailored multi-modal AI responses, enhancing efficient human-AI cooperation.
- We design and implement an intelligent virtual furniture shopping system based on the proposed framework. This system harnesses the LLMs and leverages users' multi-modal behaviors to assist swift acquisition of desired products through float labels, indicators, and voice responses.
- We conducted a comparative user study, and the results highlight that our framework enhances intent-aware human-AI cooperation, outperforming traditional interaction modes. Statistical data on user behaviors, intentions, and subjective feedback support its effectiveness in improving users' intention conveyance, interaction tendencies, subjective satisfaction, etc.

2 RELATED WORKS

2.1 Multi-Modality-Based User Intention Detection in VR Environment

User behaviors from various modalities can be meticulously documented in VR, each encapsulating distinct dimensions of user intent. Therefore, much work has been done to combine multiple measurable modalities, including eye-tracking [7, 10, 11, 14], gestures [30, 34], speech [21, 27], and virtual keyboard-mouse interactions [41], to comprehensively assist in analyzing the user's current intent. This enables the system to respond aligning more closely with the user's behavior. For instance, Gebhardt et al. [14] integrated eye-tracking data and employed reinforcement learning techniques to learn when to display or conceal object labels as a cooperative strategy for virtual shopping. Yao et al. [40] employed several machine learning models in parallel to process user speech, spatial information, and device status to jointly predict user intent in a VR-simulated smart home. These attempts, amalgamate data from multiple modalities and employ machine learning algorithms for data prediction. Our work, however, leverages the powerful data-processing capabilities of the LLMs to incorporate smoothed user gaze, pointing, and verbal expression into the user log, enabling a comprehensive analysis of user intent.

2.2 User-Intention-Based VR Responses

Generating appropriate VR responses based on recognized user intent is another crucial task. These VR responses, also multi-modal, encompassing text, speech [42], visual label update [38], changes in system UI [24], variations in the virtual assistant's appearance [42], video playback [1], etc. Furthermore, such responses need to account for user states. For instance, they must ensure information necessity while reducing the visual burden caused by information overlap [32], such as presenting labels only when needed [19]. Additionally, as the presentation format of virtual information can influence user emotions [2, 16], a combination of multiple modalities in expression is generally more effective in stimulating user interest, for instance, combining speech and text to enhance comfort perception and increase purchase intent [33, 42]. However, previous works typically concentrated on optimizing response in single-modality or proposing criteria from the UI design perspective. Our work, while following these paradigms, focuses on coordinating and generating tailored multi-modal AI responses. Our goal is to enable users to more effectively utilize the tailored multi-modal information in VR to enhance user efficiency and satisfaction.

2.3 Chatbots in VR

Improved chatbot capabilities are fueling a rising interest in VR chatbot design. For instance, in VR, Lin et al. [22] developed a chat platform for students seeking psychological assistance and stress management counseling, Matsumoto et al. [25] designed a self-care chat system to alleviate physical and psychological fatigue among hospital workers, and Xie et al. [39] created chatbots facilitating school tasks including course navigation, homework submission, and answering questions for students and teachers. Our work, shares a similar pursuit as effective user intent capture, execution, and human-like responses delivered by chatbots to enhance task efficiency and user experience in VR. Notably, pioneering efforts have been made in the context of e-shopping within VR, including [33]'s establishment of VR shopping agents for investigating user perceptions, preferences, and usability assessments of virtual conversational agents. With recent development of LLMs [23], such designs will garner further discussion and research, underscoring the significance and demand for our research on human-AI cooperation in VR that harnesses the formidable capabilities of LLMs to propel the development of more robust AI assistants.

3 BUILDING CROSS-MODALITY INTENTION-AWARE HUMAN-AI COOPERATION UNDER VR ENVIRONMENT

3.1 The Designing of Intent Tuples

While devising human-AI cooperation in VR, we introduce a framework centered around constructing (*intent*, *condition*, *intent prompt*, *action prompt*) tuples as the intent library, thus enabling condition-based intent recognition. Within these intent tuples, "intent" signifies the current user intent category, and "condition" represents the task attributes that map to the corresponding intent. The "intent prompt" serves the purpose of classifying the intent and corresponding "condition", while "action prompt" consequently directs the system in generating a series of actions to execute based on the intent and condition.

For example, consider a VR shopping scenario with the user requesting, "Which one is cheaper? (pointing at: item A, item B)", the user wishes to make informed selections, such as choosing items within a limited budget. We designate this intent as "Compare" within our condition-based framework (Table 1), which has multiple pending "conditions" including different items for comparison and various product attributes to compare. Hence, the "condition" in this situation encompasses (product: item A, item B; price: minimum). In this case, to correctly identify the intent and corresponding conditions, the "intent prompt" for the LLM should include statements as " this intent implies the user's need to compare multiple products in specific attributes. Return names of the products, the attributes and the comparison requirement (e.g., maximum)." Subsequently, according to the condition-based intention, the system can query the database for the prices of item A and B.

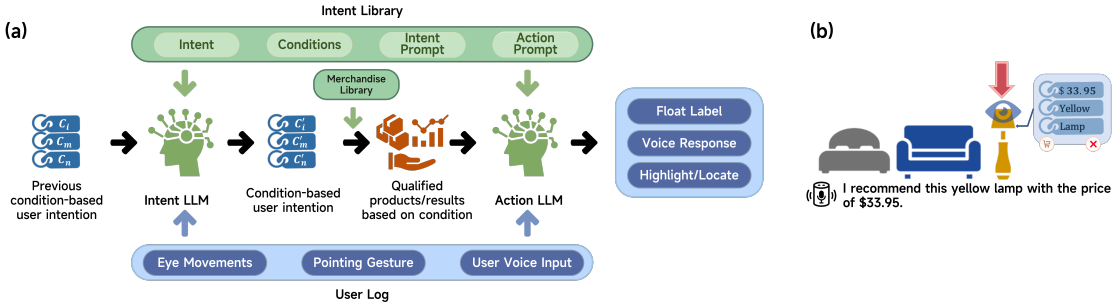

Figure 2: (a) How condition-based user intention coordinates through VR environment, here c_i, c_m, c_n represent specific conditions such as product name, material, price, etc. (b) An example of generated actions of AI in our system. For designing other systems, expansion or reduction of user input and AI responses is applicable without significantly affecting the framework's usability.

Then, based on the retrieved results, the "action prompt" includes the directions for generating a sequence of AI feedback, such as voice responses, informative pop-ups based on the user's focus on price, etc (Table 1). Here, the concept of "condition" leverages task attributes comprehensible and describable by both human and AI as the foundation to describe cooperative tasks. It aligns with the shared mental model of human-AI cooperation [3, 18], enabling better mutual understanding among human-AI teams.

The design of the intent tuples draws inspiration from the concept of "intent" and "slot value" from natural language processing [35]. However, merely designing a chatbot capable of querying the product database falls short of meeting complex user demands. To achieve intricate actions, it is imperative to synchronize the values of the "conditions" to coordinate multiple AI's actions (e.g., highlight products with "width" near 100cm) and incorporate them as an integral part of the "prompt" transmitted to the system as an execution directive.

3.2 The Utilization of Tuples with the 2-LLM Architecture

The intent tuples also provide a comprehensive framework for orchestrating AI execution across diverse modalities in VR. Fig. 2(a) illustrates how such condition-based intent tuples are utilized to form dynamic executions in our targeted shopping environment. In this process, we introduce two LLMs: one for intent and condition recognition, and the other for coordinating complex operations using "conditions". This design of 2-LLMs considers LLM's intent recognition ability without extensive training data and its flexibility in generating natural and programming language.

3.2.1 Intent LLM. The prompt feeding into the intent LLM for intent detection includes three parts: the previous condition-based user intentions, user logs, and the collections of intent and corresponding "intent prompt" from the intent library. Here, the previous condition-based user intention functions as context information from user's past interactions, which is initialized as empty. The user logs could include eye movements, pointing gestures, and voice input. Specifically in our system, we utilized user's current or nearest gazed-upon product or label (with a one-second threshold) as eye-tracking data, and user's current clicked or pointed product or label with the handle as pointing-modal data. This design is driven by the fact that users often use multi-modal expressions to convey intentions in VR, such as "this product (gazing at: item A)". In other VR applications, users' other eye movements or unique postures also provide valuable context that can be summarized and integrated into the user log. In this way, the intent LLM is capable of detecting user intents based on the user's multi-modal behaviors, in the form of matching intent tuples in the intent library.

3.2.2 Action LLM. Based on the detected intent and corresponding "conditions", the system could then perform filtering and retrieve the relevant results (such as the products from the merchandise library). Therefore, the prompt that feeds into the action LLM for action generation includes four parts: the user logs, the condition-based user intentions from Intent LLM, the "action prompt" from the intent library that corresponds with the detected intent, and the retrieved products or results. This process results in the generation of various action orders. Specifically, in our designed shopping environment (Fig. 2(b)), actions encompass adjusting float labels based on user-focused conditions, generating voice responses adhering to communication norms, and highlighting qualified products with red arrows. Users can then directly locate and navigate to suggested products by clicking on the indicators. Through decoupling intent detection and action generation, the prompt length for each LLM can be significantly shortened, as well as formatting the input-output in a simple form, thereby enhancing the appropriateness of their output.

3.3 Intent Tuples for Shopping Tasks

The condition-based human-AI cooperation framework we designed, conceptually adapts to a variety of multi-modal human-AI tasks, with various definitions of the task-specific intent tuples that necessitate thoughtful design. Our evaluation focuses on e-shopping, a scenario of current interest for academia and industry. Hence, we tailored an intent library, referring to works as [20, 26], as shown in Table 1, specifically for e-shopping as a reference for future design and user study. Correspondingly, examples of actual prompts are provided in Appendix A.

Table 1: The List of (intent, condition, intent prompt, action prompt) Tuples and Examples. User examples and explanations are for comprehension only, not part of the intent library.

User Example	Explanation	(Intent,	Condition,	Intent Prompt,	Action Prompt)
What's the height of this lamp?	View specific conditions of the products.	Get_Info	Product: Lamp X, Height	This intent implies the user's needs to view specific conditions of the products. Return the name of the products and the type of requested conditions.	1. Generate responses with the acquired condition values. 2. Prioritize the relevant conditions in the float label for user reference.
I want a green sofa.	View all products with specific conditions.	Get_Product	Category: Sofa, Color: Green	This intent implies the user's needs to see all items with specific conditions. Return the conditions and the value of each condition.	1. Generate suggestions with filtered products in response; 2. Highlight eligible products with arrows; 3. Prioritize the conditions in the float label.
Which is taller? These two sofas.	Compare differences between multiple products in certain conditions.	Compare	Product: Sofa 1, Sofa 2; Height: Maximum	This intent implies the user's need to compare multiple products in specific conditions. Return names of the products, the conditions and the comparison requirement (e.g., maximum).	1. Compare conditions for mentioned products, return qualified items; 2. Generate product suggestions in response; 3. Highlight eligible products; 4. Prioritize the relevant condition in the float label.
Not the white lamp, the black ones.	Correct the previously selected product or conditions.	Refine	Category: Lamp, Color: Black	This intent implies that the user needs to correct the product or condition previously selected. Return the name of the corrected product or condition as well as its value, retain the unchanged conditions and products.	1. Generate responses or product suggestions based on the updated product conditions or filter relevant items; 2. Highlight eligible products (if suggesting products); 3. Prioritize the corresponding condition in the float label.
I mean the one I'm looking at. (Lamp Y)	Clarify the selected products or conditions.	Disambiguate	Product: Lamp Y	This intent implies the user is clarifying the chosen product or condition, typically arising from a previous 'Unclear_Condition' intent. Return the clarified product or condition.	Update the required product or condition and execute the actions of Get_Info, Get_Product, or Compare based on user intent.
Add this lamp to the cart. (Lamp Y)	Add current products to the shopping cart.	Add_to_Cart	Product: Lamp Y	This intent implies the user's need to add the current product to the cart, return the product name.	Add the product to the cart.
Is there a lamp with sheep decoration?	Cases with non-existent conditions, unclear referred products (e.g., conflicting pointed and gazed items), or ambiguous descriptions (e.g., affordable, not too tall).	Unclear_Condition		This intent implies that one of the previous intents can be inferred, but no valid condition or product can be returned.	1. Introduce users to query options and assistant's capabilities; 2. Offer relevant recommendations, highlight products and update float labels when applicable.
Is a lamp 100 cm tall too big for a kid's study desk?	Non-task-related queries or require suggestions not supported by data in the merchandise library.	Chat	None	This intent includes all the non-shopping-task-related cases or requires responses not supported by the merchandise library.	Generate chat responses if possible.

4 SYSTEM DESIGN AND USER STUDY

To validate the effectiveness of the proposed condition-based cooperation framework in VR, we pose the following research questions:

RQ1: Does the condition-based multi-modal human-AI cooperation, compared to cooperation lacking condition-based mechanism, lead to improved efficiency in cooperation?

RQ2: What impact does the condition-based multi-modal human-AI cooperation have on various human interaction behaviors, including conversations, gaze-based interactions, and reviewing behaviors?

RQ3: Do users perceive effective assistance in the condition-based multi-modal human-AI cooperation system? What are their subjective attitudes?

To address these research questions, we devised a VR-based cooperative shopping system with a built-in virtual assistant and conducted comparative experiments to substantiate the effectiveness of our proposed framework.

4.1 System and Task Design

Virtual Environment. We designed a VR furniture purchasing scenario as in Fig. 3(a), where users are presented with four corridors, each flanked by several furniture items. These furniture models, comprising 145 real items across 12 categories such as sofas, kitchen islands, coffee tables, etc., were sourced from the "SIMMC" dataset[9, 20, 26]. The available "condition" of the products includes name, category, color, price, length, width, height, material, style, and intended rooms. Simulating real-world furniture stores' layout, these items were arranged according to the categories with identifying green float labels above (as in Fig. 1(a)), although similar categories might span multiple zones.

User Tasks. The users need to cooperate with the AI assistant within our system, and utilize various interactions introduced in section 3 to compare and select products based on preferences and our defined purchase intentions. During this process, users could converse with the AI through voice for suggestions. They could

Figure 3: (a) The overview of the designed virtual shopping environment, (b) The shopping cart design in VR environment in our system, (c) An example of the pop-up shopping page with detailed product information, (d) An interaction example from our control system.

either click on a product to access the shopping page with detailed information (Fig. 3(c)) through the controller, or directly see pop-up labels of the current gazed-upon product aligning with their current interest. Both float labels and shopping pages have the option of add_to_cart (Fig. 2(b) and 3(c)). Once users complete their selections, they would move to the "shopping cart" along the perimeter of the space, as in Fig. 3(b), where users could review product details and adjust their choices. Users could remove products by clicking on the ellipses behind each product, and may submit their shopping list multiple times by clicking the "submit" button. However, successful task completion is contingent upon users providing the correct product list, unless they exceed the maximum time limit that allows most users for completion. We then record users' total submission counts and time spent. This setup reflects real-world scenarios where users, after unsatisfying purchases, may spend extra time and effort on returns or exchanges to find products that meet their needs.

Specifically, we designed four task scenarios simulating real-world customer purchasing intents:

(1) Picking products based on the presented images of them.
(2) Picking products based on their partial information.
(3) Picking suitable combinations of products based on specific constraints, such as an overall design style, color tone, total budget, spatial dimensions, etc.
(4) Comparing and picking the most satisfying products based on an extended shopping list with product images and names and more detailed shopping requirements.

For each task scenario, we designed six sub-tasks to be completed, each requiring selecting specifically three products. To ensure assessability, each sub-task has a verifiable correct product or condition combination. Through pilot studies, we ensured the distances users needed to travel in VR environment were comparable within task scenarios, also avoiding excessive length or brevity, to prevent evident differences in difficulty. Specific task examples are provided in Appendix B.

Comparative Systems. To assess the effectiveness of condition-based cooperation, we designed two control systems referring to prevalent AI-assisted shopping systems[33]. We retained the multiple interaction and display modalities, but removed the condition-based propagation mechanism. The first control system no longer shows the dynamically updated float labels of the products, but directly displays the same shopping page when users either gaze at or click on a specific product (as Fig. 3(c)). Simultaneously, users engage in dialogues with a single LLM that utilizes complete product information and user's voice input as the prompt. The LLM offers product recommendations based on user requests, presenting them with images of recommended products and corresponding highlighting arrows, as in Fig. 3(d). Users can click on each image to expand product details as in Fig. 3(c), but are required to physically approach the corresponding product to add it to the shopping cart. In this configuration, the loss of the condition-based cooperation mechanism results in AI responses that no longer "intelligently" cater to user interests, and dynamic visual cues are absent. The second control system solely retains the product details when users

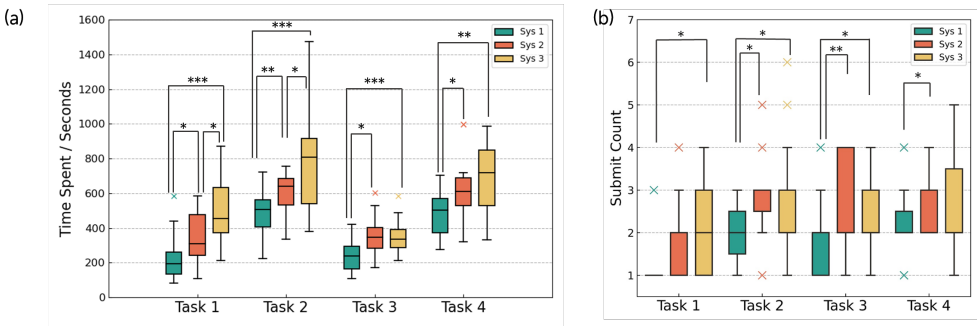

Figure 4: (a) Distribution of average task completion time for sub-tasks across 3 systems in 4 task scenarios, (b) Average submission counts to select the correct items per sub-task across 3 systems in 4 task scenarios. It partially reflects task completion accuracy. To enhance visibility, the y-axis starts at 0.5 in this plot, for users need to submit at least once. (The central black bar in each box plot represents the mean. * = p < .05, ** = p < .01, *** = p < .001, t-test)

gaze at or click on products, functioning as a relatively primitive blank control.

4.2 User Study Design

Hardware and Platform Settings. The comparative study based on the designed systems utilizes the HTC Vive Pro 2 headset as the VR testing platform, the GPT-3.5 Turbo API as the built-in LLM, and Microsoft Azure's speech service for speech recognition and voice responses.

Users and Tests Settings. We recruited 15 participants in total from a local university (7 female and 8 male), aged 21 to 29. These participants had some VR experience and a keen interest in human-AI cooperation in VR. Before formal tasks, the participants completed a sample task to familiarize the environment and operations. Then, they would select three time intervals, each separated by at least one day, and each to complete tasks in one of the three systems, to prevent potential carryover effects from prior experiences. As mentioned above, we designed 4 task scenarios, each with 6 sub-tasks, making a total of 24 sub-tasks. Hence, the participants were randomly divided into three groups of five. Within each user group, a non-repetitive random selection process was implemented, each time 2 out of the 6 sub-tasks per task scenario were chosen to be completed within one of the three systems. This methodology ensured that all participants, across the three systems, completed the full set of 24 sub-tasks without repetition. Subsequently, during data analysis, averaging procedures were employed to mitigate the potential impact of learning effects and subtle task difficulty variations. The participants have the choice to take breaks between sub-tasks to prevent discomfort or fatigue in VR environment. After finishing all sub-tasks, they were asked to complete a questionnaire and a brief interview. The questionnaire comprises seven-point scale questions based on evaluations as NASA-TLX [17] and UMUX [4], to assess physical and mental efforts, self-efficacy, and satisfaction. During interviews, participants could express opinions on encountered difficulties, preferences, and attitudes towards the systems. All participants were compensated at twice the local hourly wage for their actual participation time.

Table 2: The Analysis of Multi-modal Intent Detection and LLM Responses.

Percentage of	Sys1	Ratio	Sys2	Ratio
Multi-Modal Expressions (Users)	6.99%		9.87%	
⟶ Correctly Identified	4.52%	64.70%	1.23%	12.50%
Inappropriate Responses (LLMs)	8.33%		14.69%	
⟶ False Products	4.54%	54.55%	4.64%	31.58%
⟶ False Null Response	3.79%	45.45%	6.57%	44.74%
⟶ Non-existent Products	0.00%	0.00%	3.48%	23.68%

4.3 Results

For comparison, the condition-based main system is labeled as Sys1, the control system that keeps LLM interactions is Sys2, and the blank control is Sys3.

4.3.1 RQ1: Time and Efficiency. As shown in Fig. 4, Sys1 significantly enhances time efficiency and reduces the mistakes in submission compared to the control systems. Compared with Sys2, powered with LLM and location cues but lacks the condition-based cooperative mechanism, Sys1 substantially reduces time (30.15% on average) and submission counts (29.06% on average) across most tasks, enabling users to find expected products faster with fewer errors. Notably, in task 3 and 4, Sys2 and Sys3 show no significant differences in time efficiency or submission count. This suggests that, even with LLM support, the improvement of task efficiency remains limited with inappropriate responses and inefficient information presentation.

Regarding efficiency in interpreting users' multi-modal expressions (such as "this product (gaze at: item A)"), Sys1, equipped with condition-based multi-modal integration, achieved significantly higher precision than Sys2, which relied solely on speech dialogue context (Table 2). We also compare the efficiency in suggesting appropriate products in Table 2. Sys1 exhibited a lower inappropriate response rate from all listed perspectives. Filtering the merchandise library before action generation prevented the issue of recommending "Non-existent Products". Manual comparison shows that, in

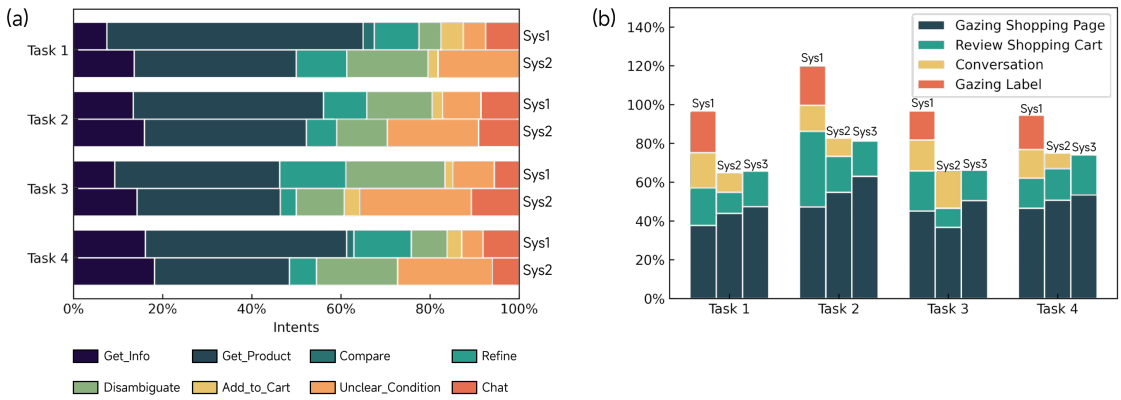

Figure 5: (a) Distribution of recognized user intent in Sys1 and manually marked user intent in Sys2 during interactions with the voice assistant (Sys3 is excluded as it lacks a voice assistant). Although GPT can correct most minor speech recognition errors, some way-off errors remain uncorrectable. We manually screened and excluded such errors to ensure proper intent distribution, as speech recognition calibration was not part of the system design. (b) Analysis of the proportion of time spent on various behaviors in the three systems (Since users may engage in gazing behaviors while conversing with the system, cumulative percentages may exceed 100%. Additionally, time spent on user mobility and aimless searching is challenging to quantify, hence not separately presented. This chart primarily aims to qualitatively compare differences in user operational behaviors across different systems.)

Sys1, the "False Null Responses" result from identifying wrong intent, non-existent conditions or redundant conditions, while "False Products" result from incorrectly detecting user's current focused conditions. Despite challenging to categorically explain, Sys2's inappropriate responses may stem from lengthy prompts and inadequate recognition of users' multi-modal intents.

4.3.2 RQ2: Behavioral Study. As shown in Fig. 5, users exhibit overall more proactive engagement with Sys1's assistant. On average, users engage in 5.0 conversational turns per sub-task in Sys1, compared to 3.375 turns in Sys2 ($p < 0.01$), with allocating approximately 10% more time to converse with Sys1 ($p < 0.05$). As in Fig. 5(a), Sys1 notably boosts effective query intents, decreases "unclear_condition" instances, and exclusively detects "compare" and "add_to_cart" intents in specific tasks. This indirectly indicates that users in Sys1 explore more interaction possibilities, aligning with observations of increased expression refinement, while in Sys2, users tend to abandon corrections after repeated disappointing responses and opt for manual screening.

As in Fig. 5(b), users proactively utilize labels in Sys1, dedicating an average of 19.85% of completion time to label observation, which leads to a 9.20% (Sys2, $p < 0.05$) and 20.14% (Sys3, $p < 0.01$) reduction in the time proportion spent examining shopping pages. Intriguingly, in Sys1 users spend more time reviewing final choices in specific task scenarios, highlighting its potential to support informed decision-making and attentive choices, considering the significantly lower submission count.

In conclusion, our design improves LLM intent recognition, user expression, and users' interaction tendencies with AI, enabling users to efficiently utilize information from multiple modalities.

4.3.3 RQ3: Subjective Feedback. As shown in Fig. 6, compared to Sys3, Sys1 received higher ratings for efficiency improvement ($M = 5.2$) than Sys2 ($M = 4.4$). Users also showed higher overall satisfaction with Sys1 ($M = 5.67$) compared to Sys2 ($M = 4.67$) and Sys3 ($M = 3.8$). Regarding workload, users found Sys1 ($M = 4.8$) more effective at reducing physical load than Sys2 ($M = 3.87$) and Sys3 ($M = 2.87$). However, the mental task load was not significantly different, possibly due to users perceiving reduced cognitive load in Sys2's direct product recommendations, despite an increase in false selections.

During interviews, some users appreciated the visual cues for efficient multi-modal navigation, finding them "helpful in quick-filtering relevant items"($P8, 13$). Others showed interest in "the consistency between the chatbot responses and visual updates"($P3$), regarding it more effective than "just conversing with AI". Some other users noted that the control systems "seem to have inaccurate recommendations"($P5$), reported feeling "cumbersome through long-time filtering"($P11$), and "discomfort from repetitive searches in VR"($P4$). These findings, aligning with objective results, altogether suggest condition-based cooperation's effectiveness in boosting efficient human-AI cooperation.

5 IMPLICATIONS, LIMITATIONS AND FUTURE WORK

5.1 Design Implications

5.1.1 Decreasing the Prompt Length. Qualitatively speaking, our design utilizes an intent library, task-decoupling for LLMs, and filters the merchandise library, resulting in a decreased prompt length. However, when evaluating quantitatively, challenges arise since LLMs utilize users' entire interaction history for responses, causing prompt length to drastically accumulate with interaction rounds, and directly comparing actual prompts may yield unrigorous results. Hence, we only compare the relatively fixed portion of the

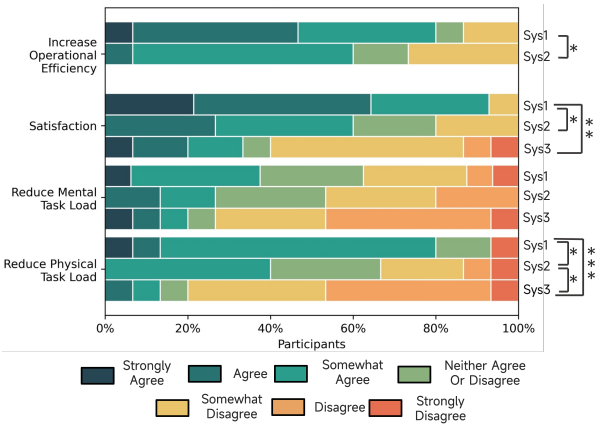

Figure 6: The scale distribution for the subjective questionnaire. The left-side y-axis represents user agreement levels with the following statements "Compared to System 3, Systems 1/2 improved my operational efficiency", "I am highly satisfied with System 1/2/3 regarding completing the tasks", "Systems 1/2/3 effectively reduced my cognitive workload", and "Systems 1/2/3 effectively reduced my physical workload".(* = p < .05, ** = p < .01, *** = p < .001, t-test)

prompt, excluding varied user logs. Sys1's Intent and Action LLMs significantly reduce prompt lengths by 82.54% and 92.87% compared to Sys2 (Action Prompt length averaged considering the frequency of corresponding intents). While enriching prompts with intent libraries might impact LLM performance, filtering the merchandise library and decoupling tasks has a stronger "de-prompting" effect.

5.1.2 Designing Intent Library. Our handcrafted intent library categorizes user behavior at a high level, leveraging reference literature and prior experience, which remains consistent with conventional chatbot intent design practices. Abstract shopping intent categories typically range between 3-20 depending on task complexity [8, 20, 31, 37]. In our case, 8 intents use 563 "tokens" in prompts [28]. Assuming a similar prompt increase per intent, even with 20 categories, the total is around 1407.7 tokens. Although it is hard to quantitatively correlate increased tokens with LLM response quality, Sys2's accuracy exceeding 85 % to perform selections from 145 items with 5635 tokens, suggests the probability of extending the "ceiling" of intent library size beyond 20 (GPT3.5-Turbo allows 16385 tokens). This provides the possibility of applying our approach to other shopping scenarios or even more. Moreover, to design functional intent libraries for various tasks, effective intent detection also necessitates careful design of distinguishing various intents without overlaps and the intent prompt utilized, other than the token-length consideration. Hence, we could gather insights from developer experiences and harness the LLM's summarization capabilities to partly automate such design, presenting a promising direction for future research.

5.1.3 Enhancing Controllability of LLMs in Practice. Ensuring LLM output controllability is a challenge [36], especially for developers utilizing off-the-shelf LLMs, like GPT, in applications. Adjusting model parameters or training data for quality control seems impractical. Nevertheless, our approach offers a different perspective in practice. Our 2-LLMs approach restricts simple condition-based input/output forms and decreases the prompt length, thereby improving the LLM output quality and accuracy thus enhancing their controllability. Comparatively, Sys2, feeding comprehensive product information and user dialogue, although occasionally swiftly providing user-desired products, experiences a significant increase in output irrationality due to lengthy prompts (Table 2). Consequently, some users, repeatedly encountering irrational responses, opt to conduct manual searches, paradoxically resulting in reduced efficiency. However, in Sys1, users generally receive more appropriate replies and are observed to more proactively cooperate with AI. This demonstrates our framework's effectiveness in promoting AI controllability in generating understandable and usable AI responses, which is crucial for successful human-AI cooperation.

5.1.4 The Applicability of Condition-Based Cooperation Framework. In this work, we validated the effectiveness of our condition-based human-AI cooperation framework in an e-commerce scenario. However, its applicability extends beyond shopping, encompassing diverse human-multi-AI interactions for various tasks. The intent tuples are comprehensible to both human and AI when describing tasks [18], and assist in swiftly discerning current task demands, orchestrating system queries, and enhancing multi-AI task assignment efficiency. Hence, our framework can also be adapted to task-attribute-based multi-AI control and coordination. For example, in cooperative scenarios like space station operations [15], multiple robots with distinct responsibilities require efficient management. Task attributes, such as velocity and temperature, serve as "conditions", allowing LLMs to capture operational requirements based on the most critical task attributes through conversations, thus identifying most responsible AIs for quick responses.

5.2 Limitation and Future Work

In our comparative experiments, we focused on fundamental VR interaction modalities—gaze fixation, controller manipulation, and vocal input—common to mainstream shopping systems. We utilized users' gaze fixation and controller manipulation data to enhance the interpretation of their linguistic expressions in VR, enhancing the identification of user intentions. However, we admit that analyzing intricate patterns in users' various multi-modal behaviors could also uncover implicit intentions beyond verbal expressions. Our current designed system may lack comprehensive analysis of such intentions, which might require additional methods such as electrooculography (EOG) [6] or electroencephalography (EEG) [29] for insightful decoding. However, our framework prioritizes intent classification and AI action generation based on intent tuples and conditions rather than prescribing specific modalities, thereby enriching the applicability of our work. We believe that, extracting and conveying these implicit intentions to LLMs through alternative modules could also advance our system's analytical capabilities, and enable more personalized responses and recommendations aligned with user behavior in future works.

Moreover, the VR environment also offers other diverse interaction possibilities, like physically moving furniture models for comparison. Additionally, the conversational assistant could be

tailored as a virtual persona for varied interactions [33] for further development. Our future research will delve deeper into VR interaction design, utilizing diverse user-centric interaction forms to carry out the condition-based framework. We also plan on adapting our framework into augmented reality (AR) devices for richer human-AI tasks in realistic settings.

Also, our intent tuples incorporated a limited amount of "intents" and intrinsic product attributes as "conditions". In future work, we plan to include additional information like user ratings, sales volume, return rates, etc. to enrich the "condition" for broader context of e-shopping. Also, exploring optimal methods for intent classification and "intent prompt" design could be an intriguing future research direction.

6 CONCLUSION

Our research presents a condition-based cross-modality intention-aware human-AI cooperation framework for VR environments. This framework, utilizing an intent library consisting of intent tuples (intent, condition, intent prompt, action prompt) and the 2-LLMs architecture, accurately identifies user intentions and generates tailored multi-modal AI responses. We implement a VR furniture shopping system within this framework and validate its effectiveness through a comparative user study. Results demonstrate improved task efficiency, intention conveyance, effective product acquisition, user satisfaction and cooperation preference, compared to traditional interaction modes. This study contributes to enhancing human-AI cooperation in VR, offering a promising approach for personalized, efficient user experiences.

ACKNOWLEDGMENTS

We are grateful to Professor Xiaohong Guan for his kind support of this work and anonymous reviewers for their insightful comments. This work is supported by the National Natural Science Foundation of China (No. 62102308) and the National Key R&D Program of China (No. 2018AAA0101501).

REFERENCES

[1] Renan Adachi, Emily M Cramer, and Hayeon Song. 2022. Using virtual reality for tourism marketing: A mediating role of self-presence. *The Social Science Journal* 59, 4 (2022), 657–670.

[2] Deeksha Adiani, Aaron Itzkovitz, Dayi Bian, Harrison Katz, Michael Breen, Spencer Hunt, Amy Swanson, Timothy J. Vogus, Joshua W. Wade, and Nilanjan Sarkar. 2022. Career Interview Readiness in Virtual Reality (CIRVR): A Platform for Simulated Interview Training for Autistic Individuals and Their Employers. *ACM Trans. Access. Comput.* 15, 1 (2022), 2:1–2:28. https://doi.org/10.1145/3505560

[3] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S. Lasecki, Daniel S. Weld, and Eric Horvitz. 2019. Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. In *Proceedings of the Seventh AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2019, Stevenson, WA, USA, October 28-30, 2019*, Edith Law and Jennifer Wortman Vaughan (Eds.), Vol. 7. AAAI Press, Palo Alto, California USA, 2–11. https://doi.org/10.1609/hcomp.v7i1.5285

[4] Simone Borsci, Stefano Federici, Silvia Bacci, Michela Gnaldi, and Francesco Bartolucci. 2015. Assessing user satisfaction in the era of user experience: Comparison of the SUS, UMUX, and UMUX-LITE as a function of product experience. *International journal of human-computer interaction* 31, 8 (2015), 484–495.

[5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). MIT Press, Cambridge, MA , USA, 1877–1901. https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

[6] Samantha Chan, Haimo Zhang, and Suranga Nanayakkara. 2023. Eye Movement Analysis of Human Visual Recognition Processes with Camera Eye Tracker: Higher Mean and Variance of Fixation Duration for Familiar Images. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 132, 8 pages. https://doi.org/10.1145/3544549.3585782

[7] Xiao-Lin Chen and Wen-Jun Hou. 2022. Gaze-Based Interaction Intention Recognition in Virtual Reality. *Electronics* 11, 10 (2022), 1647.

[8] Ann-Kristin Cordes, Benjamin Barann, Michael Rosemann, and Jörg Becker. 2020. Semantic Shopping: A Literature Study. In *53rd Hawaii International Conference on System Sciences, HICSS 2020, Maui, Hawaii, USA, January 7-10, 2020*. ScholarSpace, Maui, Hawaii, USA, 1–10. https://hdl.handle.net/10125/63885

[9] Paul A Crook, Shivani Poddar, Ankita De, Semir Shafi, David Whitney, Alborz Geramifard, and Rajen Subba. 2019. SIMMC: Situated Interactive Multi-Modal Conversational Data Collection And Evaluation Platform. *arXiv preprint arXiv:1911.02690* abs/1911.02690 (2019).

[10] Carlos Gomez Cubero and Matthias Rehm. 2021. Intention Recognition in Human Robot Interaction Based on Eye Tracking. In *Human-Computer Interaction - INTERACT 2021 - 18th IFIP TC 13 International Conference, Bari, Italy, August 30 - September 3, 2021, Proceedings, Part III (Lecture Notes in Computer Science, Vol. 12934)*, Carmelo Ardito, Rosa Lanzilotti, Alessio Malizia, Helen Petrie, Antonio Piccinno, Giuseppe Desolda, and Kori Inkpen (Eds.). Springer, Bari, Italy, 428–437. https://doi.org/10.1007/978-3-030-85613-7_29

[11] Brendan David-John, Candace Peacock, Ting Zhang, T. Scott Murdison, Hrvoje Benko, and Tanya R. Jonker. 2021. Towards gaze-based prediction of the intent to interact in virtual reality. In *2021 Symposium on Eye Tracking Research and Applications, ETRA 2020, Virtual Event, Germany, May 25-27, 2021, Short Papers*, Andreas Bulling, Anke Huckauf, Hans Gellersen, Daniel Weiskopf, Mihai Bace, Teresa Hirzle, Florian Alt, Thies Pfeiffer, Roman Bednarik, Krzysztof Krejtz, Tanja Blascheck, Michael Burch, Peter Kiefer, Michael D. Dodd, and Bonita Sharif (Eds.). ACM, New York, NY, USA, 2:1–2:7. https://doi.org/10.1145/3448018.3458008

[12] Andrew Gao. 2023. Prompt Engineering for Large Language Models. *Available at SSRN 4504303* 0 (2023), 1 8.

[13] Jing Gao. 2023. Design and Development of E-commerce Recommendation System Based on Big Data Technology. In *2023 4th International Conference on E-Commerce and Internet Technology (ECIT 2023)*, Vol. 1. Atlantis Press, Atlantis Press, Nanchang, CN, 36–40.

[14] Christoph Gebhardt, Brian Hecox, Bas van Opheusden, Daniel Wigdor, James Hillis, Otmar Hilliges, and Hrvoje Benko. 2019. Learning Cooperative Personalized Policies from Gaze Data. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology, UIST 2019, New Orleans, LA, USA, October 20-23, 2019*, François Guimbretière, Michael S. Bernstein, and Katharina Reinecke (Eds.). ACM, New York, NY, USA, 197–208. https://doi.org/10.1145/3332165.3347933

[15] Felix Gervits, Dean Thurston, Ravenna Thielstrom, Terry Fong, Quinn Pham, and Matthias Scheutz. 2020. Toward Genuine Robot Teammates: Improving Human-Robot Team Performance Using Robot Shared Mental Models. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems, AAMAS '20, Auckland, New Zealand, May 9-13, 2020*, Amal El Fallah Seghrouchni, Gita Sukthankar, Bo An, and Neil Yorke-Smith (Eds.). International Foundation for Autonomous Agents and Multiagent Systems, Auckland, New Zealand, 429–437. https://doi.org/10.5555/3398761.3398815

[16] Jerônimo G. Grandi, Zekun Cao, Mark Ogren, and Regis Kopper. 2021. Design and Simulation of Next-Generation Augmented Reality User Interfaces in Virtual Reality. In *IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops, VR Workshops 2021, Lisbon, Portugal, March 27 - April 1, 2021*. IEEE, Lisbon, Portugal, 23–29. https://doi.org/10.1109/VRW52623.2021.00011

[17] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, Amsterdam, The Kingdom of the Netherlandsl, 139–183.

[18] Ziyao He, Yunpeng Song, Shurui Zhou, and Zhongmin Cai. 2023. Interaction of Thoughts: Towards Mediating Task Assignment in Human-AI Cooperation with a Capability-Aware Shared Mental Model. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI 2023, Hamburg, Germany, April 23-28, 2023*, Albrecht Schmidt, Kaisa Väänänen, Tesh Goyal, Per Ola Kristensson, Anicia Peters, Stefanie Mueller, Julie R. Williamson, and Max L. Wilson (Eds.). ACM, Hamburg, Germany, 353:1–353:18. https://doi.org/10.1145/3544548.3580983

[19] Simon Julier, Yohan Baillot, Dennis G. Brown, and Marco Lanzagorta. 2002. Information Filtering for Mobile Augmented Reality. *IEEE Computer Graphics and Applications* 22, 5 (2002), 12–15. https://doi.org/10.1109/MCG.2002.1028721

[20] Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. 2021. SIMMC 2.0: A Task-oriented Dialog Dataset for Immersive Multimodal Conversations. In *Proceedings of the 2021 Conference on Empirical Methods in*

Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Cana, Dominican Republic, 4903–4912. https://doi.org/10.18653/v1/2021.emnlp-main.401

[21] Koki Kusano. 2023. Towards Immersive Inclusivity for C2C: How Immersive Multimodal Interactions Can Make Online Customer-to-Customer Shopping More Inclusive. In *Special Proceedings of Asian CHI Symposium 2023*. ACM, Virtual, 1–8.

[22] Aislyn PC Lin, Charles V Trappey, Chi-Cheng Luan, Amy JC Trappey, and Kevin LK Tu. 2021. A Test Platform for Managing School Stress Using a Virtual Reality Group Chatbot Counseling System. *Applied Sciences* 11, 19 (2021), 9071.

[23] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Comput. Surv.* 55, 9, Article 195 (jan 2023), 35 pages. https://doi.org/10.1145/3560815

[24] Feiyu Lu, Shakiba Davari, Lee Lisle, Yuan Li, and Doug A. Bowman. 2020. Glanceable AR: Evaluating Information Access Methods for Head-Worn Augmented Reality. In *IEEE Conference on Virtual Reality and 3D User Interfaces, VR 2010, Atlanta, GA, USA, March 22-26, 2020*. IEEE, Atlanta, GA, USA, 930–939. https://doi.org/10.1109/VR46266.2020.1581100361198

[25] Atsuko Matsumoto, Takeshi Kamita, Yukari Tawaratsumida, Ayako Nakamura, Harumi Fukuchimoto, Yuko Mitamura, Hiroko Suzuki, Tsunetsugu Munakata, and Tomoo Inoue. 2021. Combined Use of Virtual Reality and a Chatbot Reduces Emotional Stress More Than Using Them Separately. *J. Univers. Comput. Sci.* 27, 12 (2021), 1371–1389. https://doi.org/10.3897/jucs.77237

[26] Seungwhan Moon, Satwik Kottur, Paul A. Crook, Ankita De, Shivani Poddar, Theodore Levin, David Whitney, Daniel Difranco, Ahmad Beirami, Eunjoon Cho, Rajen Subba, and Alborz Geramifard. 2020. Situated and Interactive Multimodal Conversations. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, Donia Scott, Núria Bel, and Chengqing Zong (Eds.). International Committee on Computational Linguistics, Barcelona, Spain, 1103–1121. https://doi.org/10.18653/v1/2020.coling-main.96

[27] Han Wei Ng, Aiden Koh, Anthea Foong, Jeremy Ong, Jun Hao Tan, Eng Tat Khoo, and Gabriel Liu. 2022. Real-Time Spoken Language Understanding for Orthopedic Clinical Training in Virtual Reality. In *Artificial Intelligence in Education - 23rd International Conference, AIED 2022, Durham, UK, July 27-31, 2022, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 13355)*, Maria Mercedes T. Rodrigo, Noburu Matsuda, Alexandra I. Cristea, and Vania Dimitrova (Eds.). Springer, Durham, UK, 640–646. https://doi.org/10.1007/978-3-031-11644-5_61

[28] OpenAI. 2023. Tokenizer. https://platform.openai.com/tokenizer. Accessed 12/12/2023.

[29] Mansi Sharma, Shuang Chen, Philipp Müller, Maurice Rekrut, and Antonio Krüger. 2023. Implicit Search Intent Recognition using EEG and Eye Tracking: Novel Dataset and Cross-User Prediction. In *Proceedings of the 25th International Conference on Multimodal Interaction* (Paris, France) *(ICMI '23)*. Association for Computing Machinery, New York, NY, USA, 345–354. https://doi.org/10.1145/3577190.3614166

[30] Sean Simmons, Kevin Clark, Alireza Tavakkoli, and Donald Loffredo. 2018. Sensory Fusion and Intent Recognition for Accurate Gesture Recognition in Virtual Environments. In *Advances in Visual Computing - 13th International Symposium, ISVC 2018, Las Vegas, NV, USA, November 19-21, 2018, Proceedings (Lecture Notes in Computer Science, Vol. 11241)*, George Bebis, Richard Boyle, Bahram Parvin, Darko Koracin, Matt Turek, Srikumar Ramalingam, Kai Xu, Stephen Lin, Bilal Alsallakh, Jing Yang, Eduardo Cuervo, and Jonathan Ventura (Eds.). Springer, Las Vegas, NV, USA, 237–248. https://doi.org/10.1007/978-3-030-03801-4_22

[31] Ning Su, Jiyin He, Yiqun Liu, Min Zhang, and Shaoping Ma. 2018. User Intent, Behaviour, and Perceived Satisfaction in Product Search. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018, Marina Del Rey, CA, USA, February 5-9, 2018*, Yi Chang, Chengxiang Zhai, Yan Liu, and Yoelle Maarek (Eds.). ACM, Marina Del Rey, CA, USA, 547–555. https://doi.org/10.1145/3159652.3159714

[32] Markus Tatzgern, Valeria Orso, Denis Kalkofen, Giulio Jacucci, Luciano Gamberini, and Dieter Schmalstieg. 2016. Adaptive information density for augmented reality displays. In *2016 IEEE Virtual Reality, VR 2016, Greenville, SC, USA, March 19-23, 2016*, Tobias Höllerer, Victoria Interrante, Anatole Lécuyer, and Evan A. Suma (Eds.). IEEE Computer Society, Greenville, SC, USA, 83–92. https://doi.org/10.1109/VR.2016.7504691

[33] Pedro Valente, Tiago Fornelos, Rafael Ferreira, Diogo Silva, Diogo Tavares, Nuno Correia, João Magalhães, and Rui Nóbrega. 2023. Beyond Browser Online Shopping: Experience Attitude Towards Online 3D Shopping with Conversational Agents. In *Human-Computer Interaction - INTERACT 2023 - 19th IFIP TC13 International Conference, York, UK, August 28 - September 1, 2023, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 14143)*, José L. Abdelnour-Nocera, Marta Kristín Lárusdóttir, Helen Petrie, Antonio Piccinno, and Marco Winckler (Eds.). Springer, York, UK, 257–276. https://doi.org/10.1007/978-3-031-42283-6_15

[34] Hongyue Wang, Zhiquan Feng, Xiaohui Yang, Liran Zhou, Jinglan Tian, and Qingbei Guo. 2023. MRLab: Virtual-Reality Fusion Smart Laboratory Based on Multimodal Fusion. *International Journal of Human–Computer Interaction* 0 (2023), 1–14.

[35] Henry Weld, Xiaoqi Huang, Siqu Long, Josiah Poon, and Soyeon Caren Han. 2023. A Survey of Joint Intent Detection and Slot Filling Models in Natural Language Understanding. *ACM Comput. Surv.* 55, 8 (2023), 156:1–156:38. https://doi.org/10.1145/3547138

[36] David Wingate, Mohammad Shoeybi, and Taylor Sorensen. 2022. Prompt Compression and Contrastive Conditioning for Controllability and Toxicity Reduction in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 5621–5634. https://doi.org/10.18653/v1/2022.findings-emnlp.412

[37] Chao-Yuan Wu, Amr Ahmed, Gowtham Ramani Kumar, and Ritendra Datta. 2017. Predicting Latent Structured Intents from Shopping Queries. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*, Rick Barrett, Rick Cummings, Eugene Agichtein, and Evgeniy Gabrilovich (Eds.). ACM, Perth, Australia, 1133–1141. https://doi.org/10.1145/3038912.3052704

[38] Nannan Xi and Juho Hamari. 2021. Shopping in virtual reality: A literature review and future agenda. *Journal of Business Research* 134 (2021), 37–58.

[39] Qitao Xie, Wenxi Lu, Qingquan Zhang, Lingxiu Zhang, Ting Zhu, and Jianwu Wang. 2023. Chatbot Integration for Metaverse - A University Platform Prototype. In *IEEE International Conference on Omni-layer Intelligent Systems, COINS 2023, Berlin, Germany, July 23-25, 2023*. IEEE, Berlin, Germany, 1–6. https://doi.org/10.1109/COINS57856.2023.10189232

[40] Powen Yao, Yu Hou, Yuan He, Da Cheng, Huanpu Hu, and Michael Zyda. 2022. Using Multi-modal Machine Learning for User Behavior Prediction in Simulated Smart Home for Extended Reality. In *Virtual, Augmented and Mixed Reality: Design and Development - 14th International Conference, VAMR 2022, Held as Part of the 24th HCI International Conference, HCII 2022, Virtual Event, June 26 - July 1, 2022, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 13317)*, Jessie Y. C. Chen and Gino Fragomeni (Eds.). Springer, Berlin, German, 94–112. https://doi.org/10.1007/978-3-031-05939-1_7

[41] Guanhua Zhang, Susanne Hindennach, Jan Leusmann, Felix Bühler, Benedict Steuerlein, Sven Mayer, Mihai Bâce, and Andreas Bulling. 2022. Predicting Next Actions and Latent Intents during Text Formatting. In *Proceedings of the CHI Workshop Computational Approaches for Understanding, Generating, and Adapting User Interfaces (2022-01-01)*. ACM, Louisiana,USA, 1–6.

[42] Shangshang Zhu, Wei Hu, Wenjie Li, and Yenan Dong. 2023. Virtual Agents in Immersive Virtual Reality Environments: Impact of Humanoid Avatars and Output Modalities on Shopping Experience. *International Journal of Human–Computer Interaction* 0 (2023), 1–23.

A PROMPT EXAMPLES

A.1 Intent LLM

Fixed Prompt

There are 8 user intents in total, namely:

GET_INFO: This intent implies the user's need to view specific conditions of the products. Return the name of the products and the type of requested condition.

GET_PRODUCT: This intent implies the user's need to see all items with specific conditions. Return the conditions and the value of each condition.

COMPARE: This intent implies the user's need to compare multiple products in specific conditions. Return names of the products, the conditions, and the comparison requirement (e.g., maximum).

REFINE: This intent implies that the user needs to correct the product or condition previously selected. Return the name of the corrected product or condition as well as its value, retain the unchanged conditions and products.

DISAMBIGUATE: This intent implies the user is clarifying the chosen product or condition, typically arising from a previous 'Unclear_Condition' intent. Return the clarified product or condition;

ADD_TO_CART: This intent implies the user is clarifying the chosen product or condition, typically arising from a previous 'Unclear_Condition' intent. Return the clarified product or condition.

UNCLEAR_CONDITION: This intent implies that one of the previous intents can be inferred, but no valid condition or product can be returned.

CHAT: This intent includes all the non-shopping-task-related cases or requires responses not supported by the merchandise library.

All label attributes include: "class_name, sale_price, x_dim, y_dim, z_dim, color, material, decor_style, intended_room." Among them, class_name represents the product category, including Accent Chairs, Area Rugs, Bookcases, Coffee & Cocktail Tables, Dining Chairs, End Tables, Kitchen Islands, Ottomans, Sofas, Table Lamps, Teen Bookcases. Please match the input to these categories. Sale_price represents the price, using an interval for representation, such as 100-200. x_dim represents length, y_dim represents width, z_dim represents height, all using intervals for representation, such as 100-200. Color represents color, material represents the material, matching as closely as possible. If no corresponding attribute is mentioned, set it as NULL. Classify the user's last request's intent, reply in the format: " Intent: XXX
product: XXX
condition: XXX
class_name: XXX
sale_price: XXX
x_dim: XXX
y_dim: XXX
z_dim: XXX
color: XXX
material: XXX
decor_style: XXX
intended_room: XXX "
(No need to include the words "product name" and "attributes"), and provide a new line for each response.

User Log:

User: I want to buy white sofas. (gazing at: NULL, pointing at: NULL);

System: I recommend several white sofas for you, highlighted. (Intent: GET_PRODUCT, class_name: Sofas, color: White).

User: Please help me find the cheapest one of these white sofas. (gazing at: Avery Sofa Bed, pointing at: Avery Sofa Bed);

System: Clarence Loveseat is cheaper, only $369.9 (Intent: COMPARE, product: Avery Sofa Bed; Clarence Loveseat, Price: Minimum).

User: I want a sofa that is the same color as this one. (gazing at: Keanu Loveseat, pointing at: NULL);

Note: The fixed prompt includes the "intent" and "intent prompt" in our intent library. Considering LLMs utilizes entire interaction history for output generation, providing this fixed portion once is sufficient. The user log, however, accumulates with each round. To clarify its components, we present the third-round log from a test case. Multi-modal information (excluding voice) follows the user's speech input in parentheses. The textual replies from the "System" in the first two rounds are generated by the Action LLM, and the condition-based user intention detected by the Intent LLM before response generation is included in the parentheses behind. This corresponds to the "previous condition-based user intention" illustrated in Fig. 2.

Figure 7: User test task samples for task 1.

A.2 Action LLM

Note: Only actions corresponding with the detected intent would be fed into action LLM, hence we only present the corresponding Action Prompt with Appendix A.1.

User Log: User: I want to buy white sofas. (gazing at: NULL, pointing at: NULL) [Intent: GET_PRODUCT, class_name: Sofas, color: White];

System: I recommend several white sofas for you, highlighted.

User: Please help me find the cheapest one of these white sofas (gazing at: Avery Sofa Bed, pointing at: Avery Sofa Bed)[Intent: COMPARE, product: Avery Sofa Bed; Clarence Loveseat, Price: Minimum];

System: Clarence Loveseat is cheaper, only $369.9.

User: I want a sofa that is the same color as this one. (gazing at: Keanu Loveseat, pointing at: NULL)[Intent: REFINE, class_name: Sofas, color: Gray];

Filtered products: Merrick Sofa, Sabine Sleeper Loveseat, Hattiesburg Sterling Sofa, Keanu Loveseat, Martinique Tufted Sofa ,Rosina Convertible Sleeper Sofa.

Action_REFINE:

(1) Generate a text response based on the updated product conditions or filtered relevant items. Tell the user the other actions you've deployed.
(2) Highlight eligible products (if suggesting products). (Modify Arrow.cs, make changes to the products[] parameter, and execute the file.)
(3) Prioritize the corresponding condition in the float label. (Modify GazeLabel.cs, make changes to the ExtractedRules[] parameter, and execute the file.)

"Generate suggestions with filtered products in response" implies creating a Product array.

"Highlight eligible products with arrows" means that, after generating the Product array, you need to call the highlight function.

"Prioritize the conditions in the float label" implies generating an Extracted Rule array.

After "COMPARE", you need to compare the most suitable products first and generate Product and Extracted Rule arrays accordingly.

Product array like $[product1, product2, \cdots]$, Extracted Rule array like $[color, price, \cdots]$

When responding, provide the statements you need to reply with, provide the actions that you believe need to be taken, and provide the Product array and Extracted Rule array.

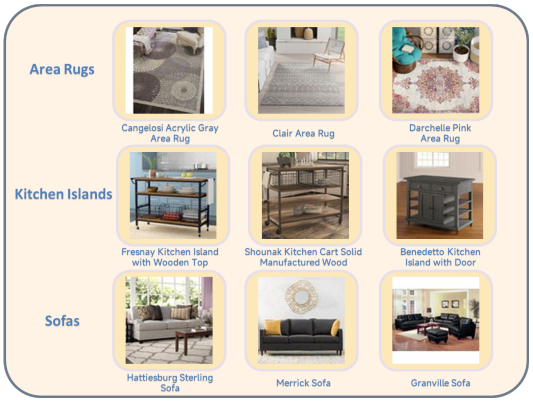

Figure 8: User test task samples for task 4.

A.3 Sys2 LLM

Now, you are a shopping assistant, and you need to record all the product information in the following order: product_name, class_name, sale_price, x_dim, y_dim, z_dim, color, material, decor_style, intended_room.

Dorset Barrel Chair, Accent Chairs, 202.99, 29.09, 29.54, 29.89, Blue birch wood, Traditional/Modern, Any room;

⋮

Chrysanthos Etagere Bookcase, Teen Bookcases, 199.99, 35.1, 71, 12, White solid and manufactured wood/wood veneers, Unknown, Living Room/Office;

Based on this information, filter out the product names that meet the users' requirements. If there are products that match the criteria, output in the following format: Omitting greetings, simply provide each product name separated by commas.

User: I want to buy white sofas.

Note: With around 145 items in the Merchandise Library, listing each one here would be extensive. Hence, we present two examples, while the actual Prompt comprises 145 lines of detailed merchandise information.

B USER TEST QUESTION SAMPLES

B.1 Task 1

Task Description: Your goal is to purchase the exact three pieces of furniture as shown in Fig.7.

B.2 Task 2

Task Description: Your goal is to buy three pieces of furniture that match the given description. Please consider the following:

(1) Ensure that the accent chair is not black, white, or gray, and its price should not exceed $200.
(2) Look for blue nylon material area rugs.
(3) Find a brown bookcase with a height not exceeding 150.

B.3 Task 3

Task Description: You want to create a furniture combination that satisfies the following requirements:

Requirements: Your envisioned home decor style is modern with a predominant use of wood. However, your space allows only an 80 cm width. Your challenge is to arrange three favorite furniture pieces, each from a different category, altogether within this limited space.

B.4 Task 4

Task Description: Based on the options presented in Fig. 8, your task is to identify one piece of furniture from each category that meets the specified requirements.

Requirements: The space has a modern theme, so select modern-style sofas and carpets. Avoid gray, as the homeowner dislikes it. The kitchen island should not exceed 100 cm in height. While material preferences are flexible, the homeowner leans towards products with metallic materials.

Received 14 September 2023; revised 12 December 2023; accepted 19 January 2024