POLITECNICO DI TORINO
Repository ISTITUZIONALE

Enhancing Smart Home Interaction through Multimodal Command Disambiguation

Original
Enhancing Smart Home Interaction through Multimodal Command Disambiguation / Calò, T., DE RUSSIS, L.. - In: PERSONAL AND UBIQUITOUS COMPUTING. - ISSN 1617-4909. - 28:6(2024), pp. 985-1000. [10.1007/s00779-024-01827-3]

Availability:
This version is available at: 11583/2990415 since: 2025-04-15T09:41:47Z

Publisher:
Springer

Published
DOI:10.1007/s00779-024-01827-3

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

28 November 2026

Personal and Ubiquitous Computing (2024) 28:985–1000
https://doi.org/10.1007/s00779-024-01827-3

ORIGINAL PAPER

Enhancing smart home interaction through multimodal command disambiguation

Tommaso Calò[1] · **Luigi De Russis**[1]

Received: 29 February 2024 / Accepted: 5 July 2024 / Published online: 22 July 2024
© The Author(s) 2024

Abstract
Smart speakers are entering our homes and enriching the connected ecosystem already present in them. Home inhabitants can use those to execute relatively simple commands, e.g., turning a lamp on. Their capabilities to interpret more complex and ambiguous commands (e.g., make this room warmer) are limited, if not absent. Large language models (LLMs) can offer creative and viable solutions to enable a practical and user-acceptable interpretation of such ambiguous commands. This paper introduces an interactive disambiguation approach that integrates visual and textual cues with natural language commands. After contextualizing the approach with a use case, we test it in an experiment where users are prompted to select the appropriate cue (an image or a textual description) to clarify ambiguous commands, thereby refining the accuracy of the system's interpretations. Outcomes from the study indicate that the disambiguation system produces responses well-aligned with user intentions, and that participants found the textual descriptions slightly more effective. Finally, interviews reveal heightened satisfaction with the smart-home system when engaging with the proposed disambiguation approach.

Keywords Smart home · Automation · Large language models · Concept disambiguation

1 Introduction

In the field of ubiquitous computing, *smart environments* refers to setups where a network of devices and sensors is embedded in various components, such as light bulbs, appliances, wearables, and the overall built environment. This integration enables these devices and appliances to detect and react to the needs of users, facilitating a responsive and interactive living or working space [1, 2]. The application of this form of intelligence in the built environment has diverse implications, spanning from building management [3] to infrastructure [4], healthcare [5], and home settings [6]. Our research specifically targets the residential aspect, with a focus on smart home systems. These systems are defined by their integration of Internet of Things (IoT) devices, which collectively enable the observation, detection, and control of different "things" within the home environment. Such systems have the potentiality to enhance aspects such as quality of life, comfort, and the efficiency of resource use [7–9]. The adoption of smart home technology is influenced by several critical factors, including customizability, automation, accessibility, reliability, and low latency [10]. These elements play a significant role in determining the extent to which people are willing to integrate and interact with smart home systems in their daily lives.

Automation systems for smart homes have relied heavily on explicit command structures and manual programming, exemplified by tools like IFTTT [11], or on pre-defined scenarios. This reliance often necessitated users to conform their communication to the system's capabilities, rather than the system adapting to the natural variability of human language and preferences [12–14]. A major challenge in these systems has been the direct interpretation of commands, especially when dealing with the inherent ambiguities in natural language. This challenge is notably evident in scenarios where users issue subjective requests, such as "prepare the living room for a relaxing evening." Traditional systems often represent commands as fixed dense vectors (embeddings) in a high-dimensional space, obtained through mapping text with a pretrained language model. These representations are static and cannot be easily refined or adapted based on subsequent

✉ Luigi De Russis
luigi.derussis@polito.it

Tommaso Calò
tommaso.calo@polito.it

[1] Politecnico di Torino, Corso Duca degli Abruzzi, 24, Torino 10129, Italy

interactions or user feedback. Due to the fixed nature of the dense vector representations, these systems find it challenging to accurately interpret subjective commands in terms of specific, actionable environmental adjustments that can be made to fulfill the user's request. [15–18]. Under-specified commands, while often clear to humans, pose challenges for systems, leading to user frustration due to existing systems' limitations in handling complex commands outside rigid structures [15–19].

Recent advancements in smart home technology have explored the use of large language models (LLMs) to enhance system responses to user commands. For instance, the Sasha system [20] employs LLMs for improved interpretation and execution of complex or vague commands. Sasha's approach includes a decision-making pipeline where key actions, such as device selection and routine checks, are managed by an LLM. Similarly, the SAGE system [21] utilizes LLMs to offer more nuanced smart home interactions, particularly for commands that require contextual understanding. These systems mainly utilize text-based inputs to process and respond to user commands, similarly to smart speakers. Such kind of textual descriptions can, however, not be particularly effective in fully capturing the user's intent [12, 22]. To address limitations in conveying intent via natural language, we implement an ambiguity detector for smart homes. As exemplified in Fig. 1, when a user command is received, if the detector identifies it as ambiguous, the system generates three textual disambiguation options or the corresponding image-based resolutions. The system then presents these to the user for validation, rather than immediately acting on its own interpretation. Users can confirm the option that more accurately matches their intent. By integrating this multimodal interaction with existing natural-language command systems, we aim to create a more intuitive interface aligning with human communication. We hypothesize that this verbal and visual approach works especially well for subjective requests like setting a room's ambience, where images provide sharper guidance. The main contributions of this paper are fourfold: (1) We propose a novel multimodal disambiguation approach for smart home systems that leverages large language models and visual cues to clarify ambiguous user commands. (2) We develop a prototype system that integrates ambiguity detection, image and text generation, and user interaction to demonstrate the feasibility of our approach. (3) We evaluate quantitatively the capability of the LLM on when and how to ask users to disambiguate with the right modality. (4) We conduct a 13-participant user study to evaluate the effectiveness of our system and gain insights into user preferences and perceptions regarding multimodal disambiguation in smart home contexts.

2 Related works

Recent studies have explored the use of machine learning to enhance the usability and adaptability of smart home systems. For example, Manu et al. [23] have delved into automating home functions based on activity recognition, using deep learning algorithms to interpret users' activities

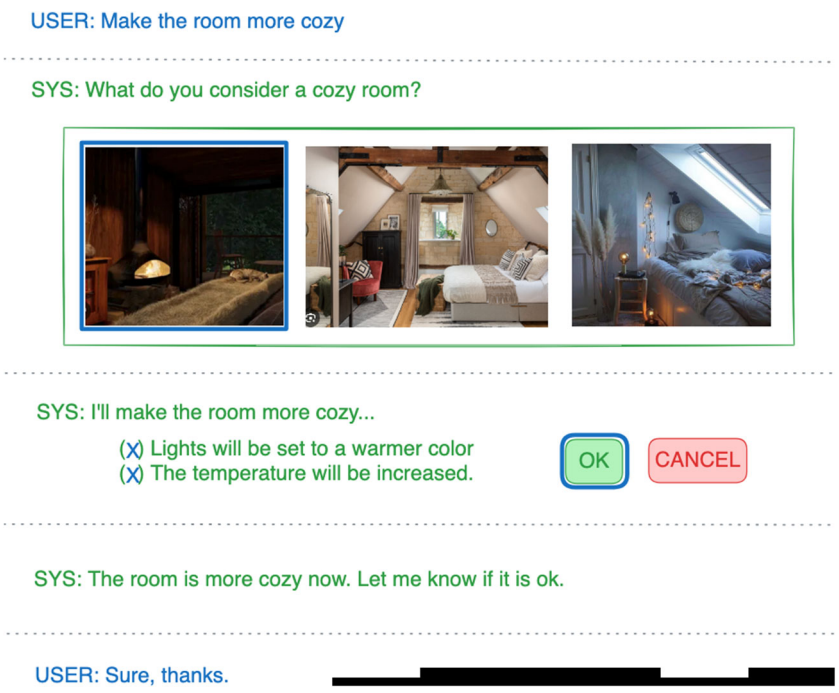

Fig. 1 The system captures a user's request to "Make the room cozier," to which the system responds by presenting, in this case, three visual options to help understand the user's interpretation of "cozy." After the user selects their preferred ambiance through an image, the system confirms the execution of actions like adjusting lights and temperature to achieve the desired coziness

from accelerometer data. Another area of focus is voice-based home assistants, which are designed to comprehend users' spoken commands and carry out the corresponding actions. A notable example by Rani et al. [24] is the development of a voice-controlled home automation system, utilizing natural language processing (NLP) and Internet of Things (IoT) technologies to manage basic household appliances. Commercially available intelligent assistants like Bixby, Google Assistant, and Alexa employ advanced NLP techniques to offer a user-friendly interface. These systems can handle various commands and queries, from shopping and setting reminders to controlling devices and automating various home functions. However, these contemporary home assistants often encounter challenges in accurately interpreting and responding to implicit or under-specified user commands [12, 22, 25]. Previous research has explored task-specific models to enhance the understanding of user commands. Noura et al. introduced the VISH grammar and dataset for modeling smart home utterances, aiming to improve natural language models' comprehension of user goals [22, 26]. This method enhances command structures by incorporating a goal-oriented grammar, allowing users to issue commands like "the blinds are too low" to adjust blinds. However, it lacks support for completely under-specified commands or persistent goals requiring reasoning about a home's devices and sensors. Palanca et al. proposed a multi-agent system to meet user goals, where agents coordinate based on an ontology describing their capabilities and interrelations [25]. The latest advancement in this field has been the introduction of LLMs, which have brought a new dimension to smart home systems by enhancing their ability to understand and interact with users more effectively. LLMs are comprehensive language models trained on a wide array of corpora, encompassing a significant portion of the textual and coded content available on the internet [27, 28]. These models are distinguished by their impressive performance across various tasks, achieving this without requiring significant model modifications for individual use cases [29, 30]. For instance, GPT-3, a notable LLM, has been applied in diverse areas, such as controlling robots through natural language [31, 32] and modeling social dynamics in small community settings [33]. The effectiveness and versatility of LLMs are largely attributed to the breadth and diversity of their training data. This data, covering multiple disciplines and contexts, encapsulates a wide range of semantic relationships that are not typically present in data used for task-specific models [27, 34].

In the context of smart home environments, LLMs show promise in narrowing the gap between the implicit goals of users and the specific, actionable responses required to achieve these objectives. Sasha [20] and SAGE (Smart home Agent with Grounded Execution) [21] are recently introduced systems leveraging the capabilities of LLMs in the context of IoT automation. Sasha is designed to creatively fulfill user goals while mitigating the issue of irrelevant or inaccurate responses commonly seen in earlier smart home assistants. This approach allows Sasha to target relevant devices and align with user preferences more effectively. SAGE functions as an LLM-based autonomous agent tailored for smart home applications, aiming to provide a natural, human-like interaction experience. SAGE integrates personal preferences, physical grounding (knowledge of home devices and their capabilities), and external grounding (awareness of external factors like weather or schedules). This comprehensive understanding enables SAGE to react to complex instructions and execute tasks more aligned with user expectations. Systems like Sasha [20] and SAGE [21] have marked an advancement in smart home technology by integrating Large Language Models (LLMs). These systems utilize the extensive knowledge of LLMs to enhance user interaction within the domain of home automation. However, their applications have primarily been centered around text-based interactions, which can be challenging to fully capture user's intent. In contrast, our research seeks to augment these LLM-based systems by introducing multimodal interaction capabilities, aiming to improve the efficiency and user experience in disambiguating commands. Specifically, we introduce a system that allows users to clarify and disambiguate concepts using visual cues in addition to text inputs. The use of various input and output modalities, tailored to user preferences and contextual needs, has been explored in different contexts [35–37]. This concept has been explored through different frameworks and architectures, demonstrating its feasibility in smart environments, including smart homes [36, 38]. For instance, systems have been developed to enable users with limited physical mobility to control smart home devices through a combination of modalities like eye blinking, speech, and touch [39]. Recent studies have emphasized the importance of context-based interactions [40]. Innovations in this area include user awareness through face identification for augmented reality-based smart home control [41] and personalization of content and automation based on user modeling and context data from smart city sensors [42]. Additionally, systems have been designed to adapt the graphical features of interfaces and the content presented to users, further enhancing the user experience [43]. Commercial assistants that offer multiple modalities often lack adaptability, typically presenting information in a uniform manner irrespective of user or context. Addressing this, adaptive interaction systems have been developed, capable of adjusting to the specific user(s) and context. Such an adaptability is achieved using information from embedded sensors in interaction devices or dedicated home sensors [38]. In the context of smart homes, adaptation to user preferences and environmental context has been instrumental in enhancing the interaction experience [40–44].

Our work relates to adaptive interaction principles through the ambiguity detector assessing commands and determining whether visual cues could provide clarification. This aligns with systems that dynamically adjust modality to best fit the user's context [36, 44]. Incorporating visual cues alongside textual descriptions for disambiguation represents a novel enhancement for smart home interaction, addressing limitations of systems failing to fully capture nuanced intent [12, 22]. Allowing multimodal expression via images or text can improve response accuracy and relevance [37]. By combining the language understanding capabilities of LLMs [27, 28] with the intuitive and contextually rich information provided by visual cues, we aim to enhance how users interact with these environments, making them more intuitive, responsive, and aligned with the complexities of human communication [35].

3 Use case for interactive disambiguation

To further contextualize and exemplify the benefits of our approach, we introduce a use case. The aim of this use case is to illustrate the differences in user experience and system performance between typical text-only methods and our proposed multimodal disambiguation approach. By presenting two scenarios based on the same user and context, we highlight the advantages of incorporating interactive disambiguation and visual cues in smart-home command interpretation and the limitations inherent in existing natural-language systems.

Paul, a graphic designer, returns home from an exceptionally hectic day at the office. His living room, usually a sanctuary of relaxation, feels stark and uninviting. Craving a serene atmosphere to unwind, Paul turns to his new smart home system, hoping to transform the space into an oasis of calm. He utters the command, "Set a relaxing mood in the living room." The system, designed to interpret such requests, recognizes the ambiguity in Paul's words. 'Relaxing' could mean different things to different people - some might find solace in dim lighting and soft music, while others might prefer the warmth of a simulated fireplace.

Using Multimodal (Image and Text) Interaction: *The system activates its Multimodal Concept Disambiguation process. It infers that the visual modality could be suitable to disambiguate the user request and swiftly generates a series of images, each depicting a unique interpretation of what a 'relaxing mood' could entail. One image shows the room bathed in soft, warm lighting with gentle music notes in the background, another presents a cozy setup with a virtual fireplace and ambient lighting, while a third image showcases a more natural setting with green hues and sounds of nature. Paul, viewing these options on his smart TV, feels immediately drawn to the image of the room with the virtual fireplace. It resonates with his idea of a peaceful evening - the warmth of the fire, the soft flicker of flames, creating a soothing visual and auditory experience. He selects this image, effectively communicating his preference without the need for complex descriptions. The system, upon receiving Paul's selection, springs into action. It adjusts the room's lighting to replicate the warm glow from the chosen image, activates the virtual fireplace on the large screen, and even subtly adjusts the room's temperature to enhance the feeling of warmth.*

Using Text-Only Interaction: *Without detecting ambiguity or engaging clarification, the system lacks context to interpret Paul's idea of "relaxing." It opts for a generic action – dimming the lights and playing soft instrumental music. While this response is within the realm of what could be considered relaxing, it doesn't quite align with Paul's personal preference for the evening. He then decides to refine the command in an attempt to better communicate their specific desires to the system. Paul issues another command, "Increase the warmth of the lighting and add a visual element like a fireplace." The system responds by slightly increasing the warmth of the lighting, but it struggles with the abstract concept of adding a 'visual element like a fireplace.' It interprets this as displaying images of fireplaces on the smart TV in the living room, rather than creating an immersive fireplace experience. Although the room now has warmer lighting and images of fireplaces, it still lacks the cozy, immersive ambiance Paul had envisioned. The system's limitations in understanding and translating Paul's nuanced request become evident. Paul tries once more, refining their command: "Make the lighting mimic a fireplace's glow and play fireplace sounds." This time, the system adjusts the lighting to a flickering, orange hue and plays sound effects of crackling fire. While this is closer to Paul's vision, the experience still feels somewhat artificial and lacks the seamless integration of visual elements that Paul desires.*

Through these iterative refinement cycles, it becomes clear that the text-only system, despite its advancements, faces significant challenges in accurately interpreting and executing more subjective, nuanced commands. Each refinement brings Paul closer to their desired outcome, yet the process is time-consuming and somewhat frustrating, highlighting the system's limitations in understanding and delivering on the full spectrum of human preferences without additional,

more specific input. This series of refinements highlights a key limitation of text-only commands: they often fail to convey the depth and complexity of visual information. While Paul knows what he wants, articulating it in words that the system can accurately interpret proves challenging. Each iteration, though closer to the desired outcome, requires effort and precise language that may not come naturally to all users.

In contrast, a multimodal approach incorporating visual and textual cues for disambiguation allowed Paul to directly select an image that captures his idea of a relaxing environment. Visuals can convey nuances such as color, intensity, movement, and atmosphere more immediately and comprehensively than text. This would not only have saved time but also eliminated the guesswork and iterative refinement needed with text-only commands, leading to a more efficient and satisfying user experience.

4 Method

The architecture of the proposed multimodal disambiguation system, depicted in Fig. 2, uses a cohesive and dynamic approach to interpreting and responding to user commands. With its capability to learn and adapt over time, the system proposes a smart-home framework that understands and evolves with the user's unique preferences and behavior. In details, the proposed system integrates several interconnected components to interpret, clarify, and execute user commands, namely:

Context store The Context Store functions as a central repository within the smart home AI system, archiving data retrieved from user interactions and information of the environment. It is updated and managed through an LLM, which is programmed to process and refine the information using structured prompts.

Context advisor The Context Advisor leverages LLM's capabilities to propose concepts that align closely with the user's environment and their requests, retrieving and utilizing data from the Context Store. Serving as an essential intermediary, the Context Advisor transforms comprehensive contextual information into actionable insights. This process is crucial for fusing the context and the outcomes of the disambiguation process, ensuring that the system's responses and actions are both relevant and attuned to the individual user's needs and preferences. The LLM consults the context conditioned on the user's instructions, ensuring the AI's understanding aligns with the user's actual environment and preferences. This interaction between the Context Store and the Context Advisor is designed to be dynamic, allowing the system to adapt to the user's changing needs and evolve the context over time. Structured prompts are used to extract the information from the Context Store, leading to more contextualized knowledge; the engineered prompts are reported in Appendix A. Over time, as the user interacts with the system and their living environment shifts, the Context Store's data is continuously updated. This approach is essential for maintaining a smart home system that is attuned to the user, providing tailored responses and actions that resonate with the user's needs and current situation.

Concept store The Concept Store, distinct from the Context Store, focuses on archiving user-selected interpretations and their representations, thus forming a "memory" of user

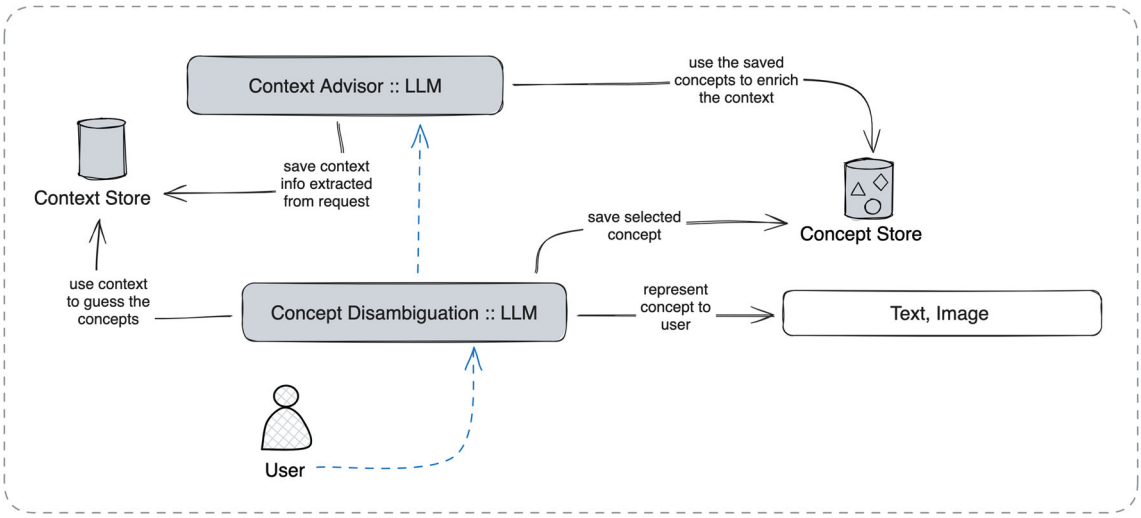

Fig. 2 The diagram shows the process where the AI interacts with a user to clarify ambiguous instructions in a smart home setting. The AI takes input from the user and the context store, consults the concept advisor to create a concept based on the environment, and then stores this concept. The Concept Disambiguation module uses LLMs to present the user with different modalities (text or image) to represent the concept, which the user selects for the AI to act upon, completing the feedback loop of understanding and action within the smart home ecosystem

preferences. While the Context Store is concerned with environmental data, the Concept Store is key to modeling user interactions and preferences, ensuring that future system responses align closely with the user's historical choices and expectations. This differentiation enhances the system's predictive capabilities and accuracy in responding to user commands.

Concept disambiguation The Concept Disambiguation subsystem of the smart-home AI employs an LLM to resolve ambiguous user commands. It is the central element of the proposed approach, the one closer to the user. To perform disambiguations, the subsystem involves a multi-step process: starting with a prompt engineered with a dataset of ambiguous commands tailored to smart homes, the system classifies these based on ambiguity levels. For commands where visual cues can aid in clarification, the system generates distinct, non-overlapping visual cue captions and corresponding AI-generated images. Users then select the image that aligns with their intent, allowing the Context Advisor to formulate a tailored policy for home automation. The detailed workings of this process will be further elaborated in the next subsection.

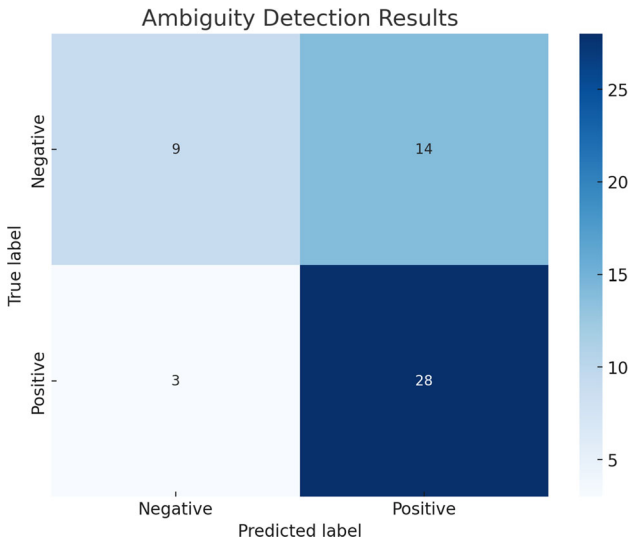

Fig. 3 Confusion matrix for ambiguity detection in smart home AI: the matrix depicts the system's proficiency in discerning ambiguous from clear user instructions. True positives and negatives correspond to accurate detections, while false positives and negatives highlight the instances of misclassification. The matrix reveals the system's inclination to prioritize ambiguity identification

4.1 Concept disambiguation

The subsystem for Concept Disambiguation lies at the core of our smart-home AI, leveraging LLMs to disambiguate human commands. By discerning user intent, it ensures that every instruction is comprehended in full and translated into actions that resonate with the user's true will. It includes five sub-components and mechanisms that confer such discernment to the system:

Multimodal choice for ambiguity resolution

We curated a dataset of 55 real-world smart home command examples, spanning ambiguous instructions like "set calming ambiance in the kitchen" and unambiguous commands like "turn on the kitchen lights." Each example is labeled with potential ambiguity triggers and the preferred modality for disambiguation (text, image). To evaluate zero-shot inference, we assessed a large language model's ability to categorize the commands as ambiguous or not without any training on this dataset. Despite no exposure, the model achieved 70% accuracy aligned with human judgments, demonstrating reliable ambiguity detection for even unseen instructions.

Figure 3 presents a confusion matrix that illustrates the performance of our smart home AI's ambiguity detection system. It demonstrates a strong tendency of the system to favor the detection of ambiguity (high false positives), which is preferable in smart home settings to prevent any misunderstanding of user commands.

In enhancing the smart-home system's ability to handle ambiguous commands, an important step involves analyzing each command in the dataset to assess whether visual cues can effectively resolve ambiguities. For each command marked as ambiguous in our dataset, a further evaluation is conducted to decide if visual cues are a suitable method for disambiguation. In fact, some ambiguities might be better resolved through additional textual information or other means. This approach is designed to use the most effective modality for each specific situation, thereby enhancing the system's ability to respond to user commands in the most effective way.

Creation of non-overlapping visual cue captions

In instances where a command from our dataset is marked as suitable for visual clarification, our system employs a process to create distinct, non-overlapping visual cue captions. This step is designed to provide a wide range of interpretations for the ambiguous concepts. We leverage a LLM to produce captions that offer a unique interpretation of the ambiguous concepts. To evaluate the LLM's effectiveness in generating a diverse array of concepts, we conduct comparisons between captions created through free generation and those crafted under constraints using metrics derived from the embedding space of a different pretrained language model [29]. This analysis helps us determine the level of semantic differentiation between the captions, providing a quantitative measure of the diversity in the concepts generated by the LLM and ensuring that the captions are capturing distinct facets of the ambiguity. We generated three captions for a set of 49 instructions for both a condition group and a control

group. With triplets sampled five times to enhance the statistical reliability of our findings. The condition group required the generation of captions that do not overlap in meaning, aiming to ensure that each caption provided a distinct interpretation. In contrast, the control group involved generating captions without restrictions on conceptual overlap, allowing for a broader range of interpretations. Statistical analysis revealed a significant difference in the means of the norms of variances between the condition and control groups (p-value: 0.003). Figure 4 presents a boxplot that compares the variances in embeddings between captions generated under conditions of free ambiguity and non-overlapping ambiguity. The boxplot visually illustrates this difference, showing a higher median variance in the condition group as compared to the control group. This indicates that the captions generated under the condition of non-overlapping ambiguity exhibit greater diversity, providing more distinct options for disambiguation.

This suggests that when the AI is tasked with generating non-overlapping interpretations, the resulting captions are more diverse, which is beneficial for providing complete options to users for disambiguation purposes.

User selection and policy generation

Presented with images or textual descriptions, the user engages in the final act of selection, pinpointing the depiction that best matches their intent. The Context Advisor then formulates a policy incorporating the user's choice, meshing the selected interpretation with the smart home's contextual data. This personalized policy is what steers the home automation system, ensuring that the user's initial ambiguous command materializes into an outcome that matches their expectations. In this way, our system not only resolves ambiguities but does so by engaging the user, learning from their choices, and continuously refining its understanding of their preferences.

4.2 Implementation

We implemented the multimodal disambiguation system using a Python server backend and React front-end interface.

We leverage the OpenAI API for access to GPT-4 [45], to categorize commands as ambiguous or not in a zero-shot setting. The components described in our architecture are orchestrated through the LangChain [46] framework. Non-overlapping captions are generated with GPT-4 and evaluated using embedding space metrics to validate diversity, as detailed in Sect. 4.1. Given the captions, we employ Dall-E 3 [47] to generate corresponding images representing possible visual disambiguation options. The experiment employs a React front-end [48] to allow the user to select the textual or visual option that best matches their intent. The overall system is designed for modularity. As better language or generative models are released, they can be readily integrated to enhance disambiguation quality. We plan to open-source key components to support further research.

5 User study

We conducted a user study to evaluate the effectiveness of the disambiguation approach and the suitability of the gen-

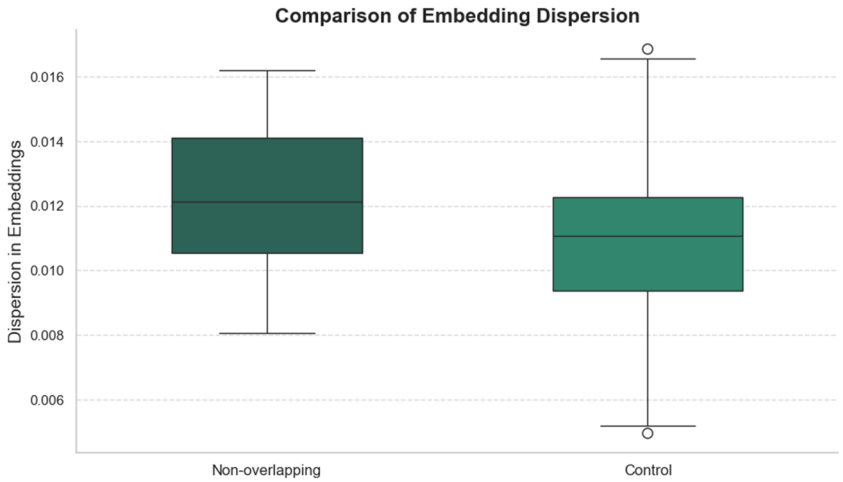

Fig. 4 Boxplot of variance in embeddings for caption generation: This chart compares the conceptual diversity of captions generated by the AI under "free ambiguity" versus "non-overlapping ambiguity" conditions. The "free ambiguity" condition refers to the AI's generation of captions when no constraints are placed on the ambiguity of the concepts, allowing for a wide range of possible interpretations. In contrast, the "non-overlapping ambiguity" condition imposes a requirement that the generated captions must not overlap in meaning, promoting distinctness and specificity in the representations of the ambiguous concept

erated visual and textual representation in smart homes. We engaged thirteen participants from varied backgrounds. The study employed both quantitative and qualitative measures. Quantitatively, participants rated the system's responses in terms of intent alignment and efficacy. Qualitatively, participants shared their experiences and opinions after interacting with the system. They discussed their preferences between text and visual-based outputs, ease of use, and overall satisfaction with the system's performance. This feedback offered deeper insights into the user experience, highlighting the practical implications of the system's performance and areas for improvement. It is important to note that the user study focuses on assessing the effectiveness and perception of the multimodal disambiguation process within a single interaction session. Thus, it is not meant to evaluate how the system learns and adapts to user preferences over time, leveraging the in-context learning capabilities of LLMs. We will leave this for further study on a longer time scale.

5.1 Procedure

Before the experiment, participants were assessed for their familiarity with smart home systems, including any previous experiences with voice or text-based assistants (e.g., Alexa or Google Assistant). They were asked about their expectations for ease of use, responsiveness, and accuracy in these systems, as well as their preference for how they communicate with smart devices.

Then, participants were asked to execute a fixed series of ambiguous commands. They includes, for example, "*Create a relaxing ambiance when I arrive home*," "*Turn on the lights if a child enter the house*," or "*Set romantic lights in the kitchen*." The complete list of commands is detailed in Appendix B. For each ambiguous command, participants were presented with either a textual or visual disambiguation option, which was randomly determined and balanced across participants to control for potential order effects. The participants were then asked to evaluate the effectiveness of the provided disambiguation option in clarifying the meaning of the ambiguous command.

In the post-experiment phase, participants reflected on the system's responses and discussed the effectiveness of both modalities. They compared their experiences with both textual and image-based responses, shared their views on ease of use, and suggested improvements for future interactions with smart home systems.

5.2 Participants

We recruited participants through convenience and snowball sampling through our social circles. We balanced our population by asking potential participants to complete a demographic survey to minimize self-selection bias. We selected a total of thirteen participants. Participants signed an informed consent form before participating in the study.

The age range of the participants varied from 23 to 34 years. Regarding gender distribution, there was a mix of both male and female participants, with slightly more males. The study was conducted in English and Spanish, the native languages of the participants.

In terms of familiarity with smart home systems, the participants exhibited a wide range of experience. Some had very little interaction, using smart home systems specifically rather than in their daily activities, while others were more knowledgeable and regularly used devices like Smart TVs, phone assistants, and specific features such as Alexa for music.

The information on participant profiles is summarized in Table 1, which also provides a quick overview of their demographics and experience with smart home systems.

5.3 Analysis of quantitative results

The experiment aimed to assess the alignment of the system's response with the participant's intent, as well as the efficacy of each modality.

Participants rated the system's responses on a scale from 1 to 5, with 1 being the least effective and 5 the most effective, in terms of *intent alignment* and *efficacy*. The concept of intent alignment refers to how well the system's disambiguation matched the participant's true intent, while efficacy pertains to the overall effectiveness and practicality of the response.

The analysis revealed that for the visual modality, the average score for intent alignment was $\mathbf{3.92} \pm 0.97$, while the average efficacy score was about $\mathbf{3.69} \pm 1.01$. Conversely, the textual modality yielded an average intent alignment score of $\mathbf{3.88} \pm 0.78$ and an average efficacy score of $\mathbf{4.04} \pm 1.10$. The results, reported in Fig. 5, indicate minor differences in both intention alignment and efficacy between the two modalities. While the visual modality exhibited slightly higher intention alignment, the textual modality showed marginally greater efficacy. However, these differences are not pronounced, suggesting that neither modality consistently outperforms the other in these aspects. The study observed that both modalities were generally effective, albeit with minor variations in specific scenarios. The textual modality was recognized for its directness and clarity, whereas the image modality provided a more intuitive and visual means of communication with the smart home system.

This finding suggests that the choice between text and image modalities may be more dependent on the specific context and user preference rather than a distinct advantage of one over the other. For instance, the text modality, with its slightly greater efficacy, might be more suitable for scenarios where precision and clear instructions are impor-

Table 1 User study's participants

#	Age	Occupation	Gender	Familiarity with smart home systems
P1	25	Master student in CS	Male	Some experience with smart home devices for automation
P2	29	Master student in CS	Male	Studied smart home systems and their energy efficiency
P3	24	Master student in CS	Female	Developed smart home applications as part of coursework
P4	31	PhD student in human-computer interaction	Male	Conducted studies on user experience with smart home interfaces
P5	26	Master student in CS	Male	Implemented a voice-controlled smart home prototype
P6	30	PhD student in cybersecurity	Male	Investigated security vulnerabilities in smart home networks
P7	24	Early childhood educator	Female	Very little, used specifically rather than in daily activities
P8	27	Software engineer	Male	Limited to Smart TV and phone assistant, not used for house control
P9	25	Photographer	Female	Aware of systems but no personal use or knowledge
P10	29	Management assistant	Female	Knowledgeable about multiple functionalities, used Alexa for music
P11	23	Student	Male	Limited familiarity, used Google Home for music
P12	34	Early childhood assistant	Female	Limited familiarity, used Google Home for music
P13	28	Software engineer	Male	Familiar with IoT aspects of smart home systems

tant. On the other hand, the visual modality, which showed slightly higher intent alignment, could be more effective in situations that require a more holistic or subjective interpretation, such as setting a mood or ambiance in a room. These findings support the idea that in scenarios where the desired outcome is more about creating a feel or an experience rather than executing a precise task, the visual cues offered by the visual modality can provide a more intuitive and comprehensive understanding of the user's intent. Furthermore, this modality is also fundamental in disambiguating concepts that may not be easily accessible or expressible through plain language. For instance, conveying the concept of "coziness" or "relaxation" can be more effectively achieved through visual representation rather than text, as these concepts can have

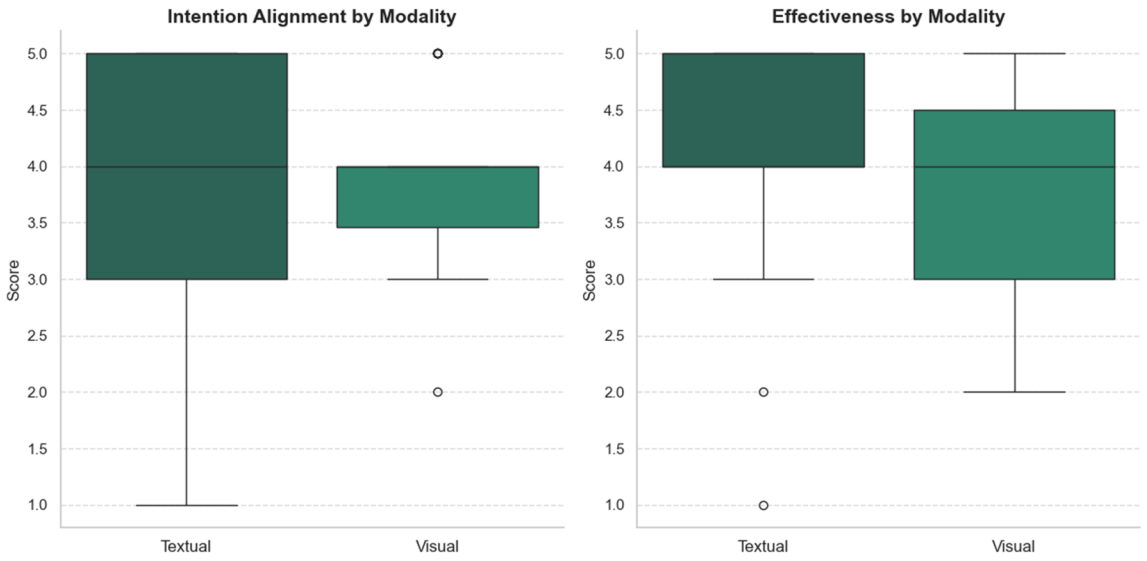

Fig. 5 Intention alignment and efficacy for both textual and image-based disambiguation results

varied interpretations that are better captured visually. This ability to bridge the gap in communication where language may fall short makes the visual modality a valuable tool in enhancing user interaction with smart home systems.

5.4 Analysis of qualitative results

We conducted a qualitative analysis of user responses following the guidelines proposed by Braun and Clarke [49]. Two researchers were involved in this, while the coding was performed by one of them. The analysis started by categorizing the material from the post-experiment phase at the sentence level through open codes, later grouped into four broader themes. The analysis revealed several key themes, reported in Table 2, that provide valuable insights into the effectiveness, usability, and user perceptions of the system. Each theme is described in the following paragraphs. In the following subsections, we will explore these themes in detail, discussing the alignment of the system with users' intentions, the effectiveness of visual and textual modalities, comparative perceptions of modalities, and suggestions for further improvements.

Congruence between user intentions and system disambiguation

To ensure user satisfaction with the disambiguation system, it is essential that the system's created choices closely match the users' actual intentions. Following our experiment, we conducted interviews to determine whether the system's disambiguation accurately reflected users' underlying intent. Participants generally expressed a high level of satisfaction with the system's response alignment to their commands. Participant 4 (P4) captured this sentiment, noting that the system's responses often matched their expectations: "*The image reflects what i imagined when i wanted a relaxing ambiance, since there is a soft light turned on by the bed, where i would probably relax when i arrive home*." Similarly, Participant 10 (P10) found the system's responses to be consistent with their preferences: "*It adjusted the devices to my preferences, most of the responses were consistent with what I wanted in each environment, it had some inaccuracies but it worked*" (1.1). However, some participants highlighted instances where the system's responses only partially aligned with their expectations, requiring further refinement. For example, Participant 1 (P1) mentioned, "*The system almost meets the expectations in every step, both the text-based and image, but in some cases the system responses weren't fully accurate*." Participant 9 (P9) also noted that the system's responses lacked specificity in some cases: "*I think the system's response was good on average but it wasn't specific enough with some of the commands*" (1.2). The need for adjustments in specific parameters or settings was also noted by several participants. Participant 6 (P6) expressed a preference for slight modifications in the system's response: "*I like this atmosphere, but I would prefer a little bit more of light*"(1.2). In some cases, participants found the system's responses to be misaligned or inaccurate. Participant 8 (P8)

Table 2 Final themes and primary subthemes for the qualitative data analysis, resulting from the thematic analysis

#	Description
1	**Congruence between user intentions and system disambiguation**
1.1	Close alignment with initial interpretation
1.2	Partial alignment requiring refinement
1.3	Misalignment or inaccurate responses
1.4	Need for additional clarification or contextual information
2	**System effectiveness in fulfilling user commands**
2.1	Fulfillment of user commands
2.2	Contextual awareness and practicality
3	**User preferences and perceptions of visual and textual modalities**
3.1	Preference for visual modality in specific scenarios
3.2	Preference for textual modality in specific scenarios
3.3	Situational preferences based on user background and experience
3.4	Complementary nature of visual and textual modalities
4	**Feedback and suggestions for system improvement**
4.1	Enhancing system performance and accuracy
4.2	Expanding system capabilities and supported commands
4.3	Providing flexibility in interaction modalities

pointed out a discrepancy between the system's response and their own interpretation: "*The color is not the same as the one shown in the image. The temperature of the room should be increased according to the system, however for me the kitchen can get pretty warm*". Participant 13 (P13) also experienced misalignment with their expectations: "*In a relaxing environment I don't listen to music, it was far from what I initially thought*" (1.3). The requirement for additional clarification or contextual information was also evident in some participants' responses. Participant 7 (P7) noted, "*It took into account the light and did not take into account the temperature, which I consider to be important when working in the kitchen*" (1.4). Overall, while the system generally aligned well with users' intentions, the feedback from participants emphasizes that the responses were in some cases underspecified. Although a fundamental issue, we argue that this could potentially be mitigated through multiple interaction turns and long-term adaptation to user preferences.

System effectiveness in fulfilling user commands

While the alignment with user intentions was the system's ability to match the user's true intention after the disambiguation process, the system's response effectiveness was assessed in terms of how well it fulfilled user commands. P2 expressed satisfaction with the system's ability to accurately execute commands, stating, "*The system effectively understood and carried out my intention.*" (2.1). This comment highlights the system's effectiveness in accurately executing desired actions. However, P4 noted limitations in the system's ability to handle complex or specific commands: "*I think the system can do a good job with basic tasks, but it would struggle to give more detailed instructions.*" (2.1). This suggests that the system may have limitations in executing commands that require a high level of specificity or complexity. In terms of contextual awareness and practicality, P6 appreciated the system's consideration of their environment and preferences: "*I liked how the system took into account different aspects that cannot be easily expressed in words.*" (2.2). This comment emphasizes the importance of contextual awareness in providing effective and personalized responses. On the other hand, P8 encountered situations where the system's suggestions were not entirely practical or realistic: "*Sometimes the system would recommend settings that didn't quite make sense for my situation.*" (2.2). This feedback indicates that the system's effectiveness could be improved by offering more practical and realistic solutions based on the user's context. These perspectives highlight the importance of accurate execution of commands, contextual awareness, and practicality in the effectiveness of smart home systems. While the system demonstrated strengths in these areas, it needs improvement in handling complex or specific commands and providing more practical suggestions tailored to the user's environment and preferences.

User preferences and perceptions of visual and textual modalities

The study also aimed to understand participants' comparative perceptions of the visual and textual modalities, exploring their preferences in specific scenarios and the potential complementary nature of the two modalities. Several participants expressed a preference for the visual modality in certain situations. P11 noted, "*Images, because it is easier to use. However you should put numbers on the images to represent some properties, like temperature*" (3.1), suggesting that the visual modality can be particularly useful for simple, straightforward interactions. P9 also highlighted the effectiveness of images in conveying the desired ambiance: "*I think images are more effective because for some people it's easier to understand something if shown visual examples than text that could be interpreted in different ways, based on personal experiences*". Similarly, P13 mentioned, "*Images, because an image conveys more information, sometimes it is more difficult to explain something in writing than visually*" (3.1), indicating that the visual modality can more naturally communicate the desired atmosphere or setting to users. The textual modality also received positive feedback regarding its efficacy. Participant 2 (P2) noted, "*text is easier to use because the human imagination does the rest, images instead force you to think about the case which is in the image*." Some participants preferred the textual modality for specific scenarios. P1 stated, "*I think the text one could be used when the scenario where more elements of the smart home needs to be actuated and thus a single image may not provide enough context*" (3.2) suggesting that the textual modality can be more suitable for complex scenarios involving multiple smart home elements. P7 also preferred the textual modality for its descriptive capabilities: "*I would say text, because it can be more descriptive, so it can express more details*". P4 expressed a similar preference, noting, "*I would like to have a text based system (maybe that i interact with and responds via voice) since it gives more detail on what i expected*" (3.2). These comments highlight the strengths of the textual modality in providing detailed and specific instructions. Interestingly, some participants' preferences were influenced by their background and prior experience with smart home systems. Participant 5 (P5) mentioned, "*I used images mainly in the context of mobile applications of smart devices providers (e.g. including maps, dashboards to control devices): the image approach used here was quite different, but still effective*" (3.3), therefore users' familiarity with existing smart home interfaces can shape their perceptions and preferences regarding modalities. Lastly, several participants recognized the complementary nature of the visual and textual modalities. P12 stated, "*Yes both modalities, because they are easy to understand and they make the answer more specific and according to my personal taste*." Similarly, P13 mentioned, "*Yes, I would like to use*

both, because it will give the system more information about what I specifically want, and that will give a more accurate response" (3.4). This supports the idea that the coexistence of visual and textual disambiguation together can provide a more comprehensive and personalized user experience.

Feedback and suggestions for system improvement

During the study, participants provided valuable feedback and suggestions for improving the disambiguation system. These suggestions primarily focused on enhancing system performance and accuracy, expanding system capabilities, and providing flexibility in interaction modalities. Regarding system performance and accuracy, P4 expressed confidence in the text-based modality, stating, "*I think the text based would be very effective, also because i imagine that if the system responds me in a way that made me think it didn't understand my needs, i could always ask again and precise the point it missed or where it was wrong*" (4.1), indicating that the system's ability to iterate and refine its understanding based on user feedback is crucial for accurate interpretation of user needs. P8 also highlighted the potential for personalization, mentioning, "*Yes, I would because I could improve and customize my smart home system with my preferences and the system could understand me better*." Additionally, P12 suggested incorporating contextual information for more personalized responses: "*For activities with children, for example if I will be watching with adults I would remove the light completely, but with children I should leave a warm light with 30% brightness. If I do chores, the light should be 100% white*" (4.1), underscoring the importance of leveraging user preferences and contextual information to enhance the system's performance and accuracy. Participants also provided suggestions for expanding the system's capabilities and supported commands. P1 mentioned, "*beneficial in everyday life, problematic just in case it doesn't execute the desired commands*," highlighting the need for the system to handle precisely a wider range of user instructions and scenarios (4.2). Flexibility in interaction modalities emerged as another area for improvement. Several participants expressed a desire for voice-based interaction to complement the visual and textual modalities. P2 mentioned, "*i would consider to do so if there were a vocal mechanism to receive commands*," and Participant 3 (P3) suggested, "*Maybe allowing the user to use their voice again could be useful in certain scenarios (for instance, when they are far from the device)*" (4.3). Additionally, some participants requested the ability to switch between visual and textual modes and select multiple options. P8 stated, "*I like the first and second option, even though I did not think about selecting the 3rd option, it is good, I'd like to have the possibility of selecting multiple options*", and P13 suggested, "*Expand the number of images, allowing you to select a maximum of two images*" (4.3). The desire for flexibility in interaction modalities highlights the suggestion of a more adaptable approach.

6 Limitations and future work

While our study provides valuable insights into the effectiveness and user experience of a multimodal disambiguation system for smart homes, we acknowledge the following limitations.

The study involved a sample size of thirteen participants, which may limit the generalizability of our findings to the broader population of smart home users.

Second, the study utilized a fictional smart home environment based on a predefined configuration and using a fixed set of ambiguous commands. While this approach allowed for a controlled experimentation, it may not fully capture the complexity and variability of real-world smart home settings.

Third, the study focused on a single interaction session between participants and the system, and it may not fully account for the potential long-term effects of using such a system on user behavior, preferences, and satisfaction.

To address these limitations, future studies could explore the integration of the multimodal disambiguation system with real smart home environments. This would also allow to investigate user interactions with the system over an extended period. In addition, future studies could explore the integration of more advanced visual disambiguation techniques, such as dynamic or interactive visualizations, to better convey complex changes in the smart home environment. By building upon the findings of this study and addressing its limitations, we can understand how to harness the power of LLMs in smart home environments in the most appropriate and effective manner.

7 Conclusions

To address natural-language difficulties interpreting ambiguous, nuanced, and subjective commands, we proposed a multimodal disambiguation approach consisting of ambiguity detection, distinct textual and image generation, and user-driven concept selection between visual or textual representations. Quantitative and qualitative analysis from a user study indicated general effectiveness and improved user engagement but also revealed areas for refinement. The major challenge is to ensure effective and aligned disambiguation under complex and underspecified instructions. Future work directions include enhancing the multimodal system, larger-scale studies with more diverse participant populations, exploring the integration of the system with real-world

smart home environments, and investigating the long-term effects of using the system through longitudinal studies.

Appendix A Prompts for smart-home system disambiguation

The appendix section of this paper provides detailed insights into the operational framework of the smart home system's AI, focusing on its capability to process and clarify ambiguous user instructions. This section is pivotal for understanding the AI's methodologies and its interaction mechanisms with users, which are fundamental to the system's efficiency and user satisfaction. The appendix elaborates on three specific prompts that illustrate the AI's decision-making and disambiguation processes.

A.1 Ambiguity detection prompt

This prompt details the AI's initial step in processing user commands—identifying and categorizing ambiguities within these instructions. It outlines the criteria for determining the ambiguity level of a concept and how the AI should respond to these findings. The prompt underscores the AI's analytical capabilities, highlighting its role in extracting ambiguous concepts from instructions and determining their potential interpretations based on the home environment context.

Listing 1 Ambiguity Detection Prompt

```
zero_shot = ‘‘’’
You are a support AI inside a smart home. Your goal
is to extract ambiguous concepts from the instruct–
ions, consider a concept ambiguous if it can be
interpreted differently by two people using the
provided home environment, ignore device ambiguity.
Assign them an ambiguity level using the following
criteria:
  – ‘‘high’’, indicating that it is completely
unclear the concept means in the environment;
  – ‘‘medium’’, indicating that the concept is
subjective and may have several interpretations;
  – ‘‘none’’, meaning that the concept can be
infered from the environment.

If ambiguous concepts were detected, respond with
the following JSON format, otherwise respond with
an explanation of why ambiguities were not found.
[
    {{
        ‘‘concept’’: <concept>,
        ‘‘explanation’’: <why it is ambiguous>,
        ‘‘ambiguity’’: <ambiguity level>
    }}
]

Environment: {environment}

User Instruction: {instruction}
```

A.2 Image option prompt

This section explains how the AI generates visual disambiguation options for ambiguous concepts identified in user instructions. It describes the process of creating search queries related to ambiguous words, which are then used to generate images. These images serve as visual cues for users to clarify their intent, showcasing the AI's ability to employ multimodal inputs for enhancing command interpretation accuracy.

Listing 2 Image Option Prompt

```
zero_shot = ‘‘’’
You are a AI Assistant for creating prompts to
feed an image generator.

You want to clarify the ‘‘Ambiguos word’’ give you
in the ‘‘User Instruction’’ for knowing how you
should act with the devices that you have in the
given ‘‘Environment’’. For that, you need to
create a search query related with the ‘‘Ambiguos
word’’ and ‘‘User instruction’’that could be used
in a search engine.

Respond just with the following JSON format:
{{
    ‘‘concept’’: <concept>,
    ‘‘place’’: <place>,
    ‘‘search_query’’: <search_query>
}}

Ambiguos concept: {ambiguos_word}
User Instruction: {user_instruction}
Environment: {context}
```

A.3 Text option prompt

This prompt illustrates the alternative method of text-based disambiguation, where the AI provides users with explicit textual options related to the ambiguous concept. It focuses on how these options are crafted based on the devices and their properties within the given environment, offering a direct and straightforward approach for users to specify their intent.

Listing 3 Text Option Prompt

```
zero_shot = ‘‘’’
```

```
You are a support AI inside a smart home.
You want to clarify the ''Ambiguos concept'' that
is given you, related with the ''User Instruction''
for knowing how you should act with the devices
that you have in the given ''Environment''.
For that, you need to give to the user 5 text
explicit options about what is the meaning just of
the ''Ambiguos concept'' for the ''User
Instruction''.
The options have to be related exclusively with the
devices of the place and their properties according
to the given ''Environment''.

Respond just with the following JSON format:
{{
    ''concept'': <concept>,
    ''place'': <place>,
    ''devices'': [<devices>],
    ''options'': [<options>]
}}

Ambiguos concept: {ambiguos_word}
User Instruction: {user_instruction}
Environment: {context}
```

Appendix B User study materials

B.1 Commands

- Set romantic lights in the kitchen
- Make my room cozy
- Please let me know when a child leave the house
- I want to turn on the light in the room only when there is an elderly in it
- If it is dark outside close the garage please
- Set my room light like a dawn
- Create a relaxing ambiance when I arrive home
- Turn off the lights in the kitchen after dinner
- Let me know when many people get in to the house
- Set warm lights in the kitchen when it's cold outside
- Adjust the kitchen ambiance to feel festive
- Make the bedroom atmosphere lively
- Set the kitchen lights for a party mood
- Adjust the bedroom lighting for a movie night
- Create an energetic vibe in the living room
- Set a calming ambiance in the kitchen
- Adjust the bedroom lights for a cozy reading experience
- Dim the lights in the bedroom for a romantic evening
- Adjust the kitchen lights for a calm and soothing feel
- Set the bedroom lights for a serene atmosphere
- Adjust the kitchen ambiance for a productive cooking session

B.2 Environment

```
environment:
  devices:
    - name: KitchenLamp
      place: kitchen
      type: lamp
      properties:
        - name: state
          type: bool
          value: true
        - name: color
          type: string
          value: FFFFF
        - name: brightness
          type: int
          value: 100
    - name: MyTV
      place: bedroom
      type: tv
      properties:
        - name: state
          type: bool
          value: true
    - name: Bedroom Speaker
      place: bedroom
      type: speaker
      properties:
        - name: connected
          type: bool
          value: true
        - name: volume
          type: int
          value: 50
    - name: House Lock
      place: entrance
      type: lock
      properties:
        - name: locked
          type: bool
          value: false
    - name: Bedroom Thermostat
      place: bedroom
      type: thermostat
      properties:
        - name: temperature
          type: float
          value: 21
```

```
  - name: Kitchen Thermostat
    place: kitchen
    type: thermostat
    properties:
      - name: temperature
        type: float
        value: 19
sensors:
  - name: Entrance Movement Sensor
    place: entrance
    type: motion
    properties:
      - name: motion
        type: timestamp
        value: 1696499498
  - name: Entrance Camera
    place: entrance
    type: camera
    properties:
      - name: connected
        type: bool
        value: true
```

Acknowledgements The authors want to thank all the participants to the study for their availability and Iván Contreras Perez, who contributed to the work as part of his M.S. thesis.

Funding Open access funding provided by Politecnico di Torino within the CRUI-CARE Agreement.

Data Availability In addition, the dataset generated during the current study is available from the corresponding author on reasonable request.

Declarations

Conflict of Interest The authors declare no Conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Dunne R, Morris T, Harper S (2021) A survey of ambient intelligence. ACM Comput Surv (CSUR) 54(4):1–27
2. Weiser M (1999) The computer for the 21st century. ACM SIGMOBILE Mobile Computing and Communications Review 3(3):3–11
3. Kim D, Yoon Y, Lee J, Mago PJ, Lee K, Cho H (2022) Design and implementation of smart buildings: a review of current research trend. Energies 15(12):4278
4. Branny A, Møller MS, Korpilo S, McPhearson T, Gulsrud N, Olafsson AS, Raymond CM, Andersson E (2022) Smarter greener cities through a social-ecological-technological systems approach. Curr Opin Environ Sustain 55:101168
5. Acampora G, Cook DJ, Rashidi P, Vasilakos AV (2013) A survey on ambient intelligence in healthcare. Proc IEEE 101(12):2470–2494
6. Alaa M, Zaidan AA, Zaidan BB, Talal M, Mat Kiah ML (2017) A review of smart home applications based on internet of things. J Netw Comput Appl 97:48–65
7. Lutolf R (1992) Smart home concept and the integration of energy meters into a home based system. In: Seventh international conference on metering apparatus and tariffs for electricity supply, pp 277–278
8. Ki C-WC, Cho E, Lee J-E (2020) Can an intelligent personal assistant (IPA) be your friend? Para-friendship development mechanism between IPAs and their users. Comput Hum Behav 111:106412. https://doi.org/10.1016/j.chb.2020.106412
9. Wilson C, Hargreaves T, Hauxwell-Baldwin R (2015) Smart homes and their users: a systematic analysis and key challenges. Pers Ubiquit Comput 19:463–476
10. Reisinger MR, Prost S, Schrammel J, Fröhlich P (2022) User requirements for the design of smart homes: dimensions and goals. J Ambient Intell Humaniz Comput, 1–20
11. IFTTT (2023) IFTTT. https://ifttt.com/. Retrieved May 1, 2023
12. Clark M, Newman MW, Dutta P (2017) Devices and data and agents, oh my: how smart home abstractions prime end-user mental models. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 1(3):1–26
13. Yu H, Hua J, Julien C (2021) Dataset: analysis of IFTTT recipes to study how humans use Internet-of-Things (IoT) devices. In: Proceedings of the 19th ACM conference on embedded networked sensor systems, pp 537–541
14. Upadhyay P, Heung S, Azenkot S, Brewer RN (2023) Studying exploration & long-term use of voice assistants by older adults. In: Proceedings of the 2023 CHI conference on human factors in computing systems. CHI '23. Association for Computing Machinery, New York, USA. https://doi.org/10.1145/3544548.3580925
15. Pradhan A, Lazar A, Findlater L (2020) Use of intelligent voice assistants by older adults with low technology use. ACM Transactions on Computer-Human Interaction (TOCHI) 27(4):1–27
16. Luger E, Sellen A (2016) "Like Having a Really Bad PA": the gulf between user expectation and experience of conversational agents. In: Proceedings of the 2016 CHI conference on human factors in computing systems, pp 5286–5297
17. Kim S, Choudhury A (2021) Exploring older adults' perception and use of smart speaker-based voice assistants: a longitudinal study. Comput Hum Behav 124:106914
18. Cowan BR, Pantidi N, Coyle D, Morrissey K, Clarke P, Al-Shehri S, Earley D, Bandeira N (2017) "What Can I Help You With?": infrequent users' experiences of intelligent personal assistants. In: Proceedings of the 19th international conference on human-computer interaction with mobile devices and services, pp 1–12
19. Upadhyay P, Heung S, Azenkot S, Brewer RN (2023) Studying exploration & long-term use of voice assistants by older adults. In: Proceedings of the 2023 CHI conference on human factors in computing systems, pp 1–11
20. King E, Yu H, Lee S, Julien C (2024) Sasha: creative goal-oriented reasoning in smart homes with large language models. Proc ACM Interact Mob Wearable Ubiquitous Technol 8(1). https://doi.org/10.1145/3643505
21. Rivkin D, Hogan F, Feriani A, Konar A, Sigal A, Liu S, Dudek G (2023) SAGE: smart home agent with grounded execution. arXiv preprint arXiv:2311.00772

22. Noura M, Heil S, Gaedke M (2020) VISH: does your smart home dialogue system also need training data? In: Bielikova M, Mikkonen T, Pautasso C (eds) Web engineering. Springer, Cham, pp 171–187
23. Manu RD, Kumar S, Snehashish S, Rekha K (2019) Smart home automation using IoT and deep learning. Int Res J Eng Technol 6(4):1–4
24. Rani PJ, Bakthakumar J, Kumaar BP, Kumaar UP, Kumar S (2017) Voice controlled home automation system using Natural Language Processing (NLP) and Internet of Things (IoT). In: 2017 Third international conference on science technology engineering & management (ICONSTEM), pp 368–373
25. Palanca J, Val E, Garcia-Fornes A, Billhardt H, Corchado JM, Julián V (2018) Designing a goal-oriented smart-home environment. Inf Syst Front 20:125–142
26. Noura M, Heil S, Gaedke M (2020) Natural language goal understanding for smart home environments. In: Proceedings of the 10th international conference on the internet of things, pp 1–8
27. Gao L, Biderman S, Black S, Golding L, Hoppe T, Foster C, Phang J, He H, Thite A, Nabeshima N et al (2020) The Pile: an 800GB dataset of diverse text for language modeling. arXiv preprint arXiv:2101.00027
28. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I et al (2019) Language models are unsupervised multitask learners. OpenAI Blog 1(8):9
29. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A et al (2020) Language models are few-shot learners. Adv Neural Inf Process Syst 33:1877–1901
30. Wei J, Bosma M, Zhao V, Guu K, Yu AW, Lester B, Du N, Dai AM, Le QV (2022) Finetuned language models are zero-shot learners. In: International conference on learning representations. https://openreview.net/forum?id=gEZrGCozdqR
31. Liang J, Huang W, Xia F, Xu P, Hausman K, Ichter B, Florence P, Zeng A (2023) Code as policies: language model programs for embodied control. In: 2023 IEEE International conference on robotics and automation (ICRA), pp 9493–9500. https://doi.org/10.1109/ICRA48891.2023.10160591
32. Wu J, Antonova R, Kan A, Lepert M, Zeng A, Song S, Bohg J, Rusinkiewicz S, Funkhouser T (2023) TidyBot: personalized robot assistance with large language models. In: 2023 IEEE/RSJ International conference on intelligent robots and systems (IROS), pp 3546–3553. https://doi.org/10.1109/IROS55552.2023.10341577
33. Park JS, O'Brien J, Cai CJ, Morris MR, Liang P, Bernstein MS (2023) Generative agents: interactive simulacra of human behavior. In: Proceedings of the 36th annual acm symposium on user interface software and technology. UIST '23. Association for Computing Machinery, New York, USA. https://doi.org/10.1145/3586183.3606763
34. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M-A, Lacroix T, Rozière B, Goyal N, Hambro E, Azhar F et al (2023) Llama: open and efficient foundation language models. arXiv preprint arXiv:2302.13971
35. Këpuska V, Bohouta G (2018) Next-generation of virtual personal assistants (Microsoft Cortana, Apple Siri, Amazon Alexa and Google Home). In: Proceedings of the 2018 IEEE 8th annual computing and communication workshop and conference (CCWC). IEEE, Las Vegas, NV, USA, pp 99–103
36. Almeida N, Silva S, Teixeira A, Ketsmur M, Guimarães D, Fonseca E (2018) Multimodal interaction for accessible smart homes. In: Proceedings of the 8th international conference on software development and technologies for enhancing accessibility and fighting info-exclusion. ACM, New York, USA, pp 63–70
37. Liu C, Xie W, Zhang P, Zhan J, Xiao Z (2018) Considerations on multimodal human-computer interaction. In: Proceedings of the 2018 5th IEEE international conference on cloud computing and intelligence systems (CCIS). IEEE, Nanjing, China, pp 331–335
38. Almeida N, Teixeira A, Silva S, Ketsmur M (2019) The AM4I architecture and framework for multimodal interaction and its application to smart environments. Sensors 19:2587. https://doi.org/10.3390/s19112587
39. Contreras-Castañeda MA, Holgado-Terriza JA, Pomboza-Junez G, Paderewski-Rodríguez P, Gutiérrez-Vela FL (2019) Smart home: multimodal interaction for control of home devices. In: Proceedings of the XX international conference on human computer interaction. ACM, New York, NY, USA
40. Neßelrath R, Lu C, Schulz CH, Frey J, Alexandersson J (2011) A gesture based system for context-sensitive interaction with smart homes. Springer, Berlin/Heidelberg, Germany, pp 209–219. https://doi.org/10.1007/978-3-642-18167-2_15
41. Marques B, Dias P, Alves J, Santos BS (2020) In: Ahram T, Karwowski W, Pickl S, Taiar R (eds) Adaptive augmented reality user interfaces using face recognition for smart home control. Springer, Cham, Switzerland, pp 15–19. https://doi.org/10.1007/978-3-030-27928-8_3
42. Vlachostergiou A, Stratogiannis G, Caridakis G, Siolas G, Mylonas P (2016) User adaptive and context-aware smart home using pervasive and semantic technologies. J Electr Comput Eng 2016:4789803. https://doi.org/10.1155/2016/4789803
43. Gullá F, Ceccacci S, Menghi R, Cavalieri L, Germani M (2017) In: Cavallo F, Marletta V, Monteriù A, Siciliano P (eds) Adaptive interface for smart home: a new design approach. Springer, Cham, Switzerland, pp 107–115. https://doi.org/10.1007/978-3-319-54283-6_8
44. Chahuara P, Portet F, Vacher M (2017) Context-aware decision making under uncertainty for voice-based control of smart home. Expert Syst Appl 75:63–79. https://doi.org/10.1016/j.eswa.2017.01.014
45. OpenAI (2024) OpenAI API. https://openai.com/api/. Accessed 27 Mar 2024
46. LangChain (2024) A library for building applications with language models. https://github.com/LangChain/langchain. Accessed 27 Mar 2024
47. OpenAI (2024) DALL·E: a neural network-based image generation model. https://openai.com/dall-e. Accessed 27 Mar 2024
48. React (2024) A JavaScript library for building user interfaces. https://reactjs.org/. Accessed 27 Mar 2024
49. Braun V, Clarke V (2006) Using thematic analysis in psychology. Qual Res Psychol 3(2):77–101. https://doi.org/10.1191/1478088706qp063oa

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.