Towards Adaptive and Holistic
AR Task Guidance

DISSERTATION

Submitted in Partial Fulfillment of

the Requirements for

the Degree of

DOCTOR OF PHILOSOPHY (Computer Science)

at the

NEW YORK UNIVERSITY
TANDON SCHOOL OF ENGINEERING

by

Guande Wu

August 2025

Towards Adaptive and Holistic AR Task Guidance

DISSERTATION

Submitted in Partial Fulfillment of
the Requirements for
the Degree of

DOCTOR OF PHILOSOPHY (Computer Science)

at the

NEW YORK UNIVERSITY
TANDON SCHOOL OF ENGINEERING

by

Guande Wu

August 2025

Approved:

Department Chair Signature

June 17, 2025

Date

Approved by the Guidance Committee:

Major: PhD in Computer Science

Cláudio T. Silva

Cláudio T. Silva
Institute Professor
NYU Tandon School of Engineering

06-04-2025
Date

Huamin Qu

Huamin Qu
Chair Professor
The Hong Kong University of Science and Technology

06-05-2025
Date

Qi Sun

Qi Sun
Assistant Professor
NYU Tandon School of Engineering

6/16/2025
Date

R. K

Robert Krueger
Assistant Professor
NYU Tandon School of Engineering

15-06.25
Date

Microfilm or other copies of this dissertation are obtainable from

Vita

Guande Wu was born in Jingning She Autonomous County, Zhejiang, China. He attended Lishui Experimental School and Lishui High School. He received a B.Eng. degree in Software Engineering from Zhejiang University (2016). During his undergraduate studies, he visited the VIDi Lab at UC Davis, where he conducted research on visualization (2019), and worked in the ZJUIDG Lab on urban visualization (2018-2020). He also interned at Microsoft Research Asia in Beijing (2019-2020). Due to COVID, he spent half year in Tongji University iDVx lab working in the narrative visualization and NYU Shanghai for a Teaching Assistant job (2020). He started his Ph.D. in January 2021, working in various areas, including human-AI collaboration, video summarization and visual analytics. During his Ph.D., he interned at Adobe Research and Amazon AWS AI.

Acknowledgements

I am grateful to my mother and family members, who have always believed in me and were always rooting for my success.

I would like to thank my advisor, Cláudio Silva, for the support, guidance, and exchange of ideas throughout my Ph.D. studies. Thank you for allowing me to work on challenging projects and constantly pushing me to bring my research to a higher level. I would also like to thank the members of my Ph.D. committee, Huamin Qu, Qi Sun and Robert Krueger, for their expertise, ideas, and feedback.

I would like to thank my girlfriend, Danjing Chen, and our cat, Sisi, who have supported me throughout my PhD and beyond. I once believed love would be black and white—but with her, it is golden.

I would also like to thank the funding agencies that supported the work presented here: DARPA Perceptually-enabled Task Guidance (PTG), DARPA Automating Scientific Knowledge Extraction and Modeling (ASKEM) program, and C2SMART, a Tier 1 University Center awarded by U.S. Department of Transportation under contract 69A3351747124, and NYU SoE Fellowship. Any opinions, findings, and conclusions or recommendations expressed in this dissertation are those of the author and do not necessarily reflect the views of the funding agencies.

I would like to thank my first cousin once removed and my friend Yusheng Wu. I am deeply grateful to my very first lifelong friends, Jiale Xiang, Zining Wang, and Enqiao Xu—whom I've known since middle school. I also want to thank the friends who have become like family to me: Canwen Xu, Fenfei Guo, and Chen Zhang. I am especially thankful to my friend Chengbo Zheng, whose work in HCI and thoughtful discussions have inspired my research, and to his girlfriend, my friend Yujie Zheng, whose background in psychology and kind support have been greatly appreciated.

I also want to thank my collaborators during my PhD, including João Rulff, Sonia Castelo, Gromit Chan, Shunan Guo, Jing Qian, Shaoyu Chen, Erin McGowan, Chenyi Li, Ryan Rossi, Jane Holfswell, Eugney Kwon, Michael Middleton, Chen Zhao, Roque Lopez, Jianben He. I am also grateful to the friends I met during my PhD and past life—many of whom have been both labmates and companions—Chenjing Wu, Kefei Liang, Huiyu Pan, Yuxuan Qin, Qingyang Liu, Jun Yuan, Jorge Wagner, Thales Goncalves, Laura Melgar Garcia, Raphael Meyer, Parikshit

Solunke, Yurong Liu, and Chen Chen—for their support, friendship, and the many moments we shared along the way.

Guande Wu
August 2025

To my family and friends, and in loving memory of my grandmother.

ABSTRACT

Towards Adaptive and Holistic AR Task Guidance

by

Guande Wu

Advisor: Prof. Cláudio T. Silva, Ph.D.

Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy (Computer Science)

August 2025

Task guidance with augmented reality (AR) provides real-time, context-aware instructions, helping users complete complex tasks efficiently and accurately. However, existing AR task guidance often lacks adaptability and continuity, offering generic instructions without considering user needs or providing meaningful post-task insights. This research addresses these limitations by developing an adaptive and holistic AR guidance framework that supports users during and after task execution. To comprehensively enhance AR task guidance, this framework addresses three key goals: improving task efficiency, adapting to dynamic scenarios, and providing long-term insights for post-task analysis. First, to directly enhance task performance, the framework introduces an adaptive text simplification method tailored for AR scenarios, reducing cognitive load and optimizing text comprehension to improve in-task efficiency. Second, to broaden the system's adaptivity across

diverse contexts, the framework incorporates adaptive guidance using BDI-based user modeling and LLM agents, enabling context-aware and proactive guidance. Finally, to support the post-study review and complete a holistic framework, the framework integrates a visual documentation system with interactive video summarization, generating actionable insights from AR task recordings to support long-term performance review and continuous improvement.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Augmented reality (AR) task guidance delivers real-time, context-aware instructions by overlaying digital information onto the user's physical environment [292]. These systems have been increasingly adopted in domains such as aviation training (e.g., training pilot's flight procedure) [145], healthcare (e.g., assisting surgeons during operations) [216], and industrial assembly and maintenance (robot arm maintenance) [69, 188, 387], where they assist users in completing complex, multi-step tasks with improved efficiency and accuracy [212]. Despite these benefits, many existing AR task guidance systems offer static, one-size-fits-all instructions, with limited adaptability to user needs or situational changes. Furthermore, most systems focus solely on in-task assistance [147], providing little support for post-task review or learning.

In this dissertation, we aim to address these limitations by developing an adaptive and holistic AR task guidance framework. The proposed framework enhances task performance through adaptive in-task support while extending its impact beyond immediate task execution through post-task analysis. Our approach is grounded in three key contributions:

1. **Adaptive Text Simplification for AR Task Guidance.** We introduce a method for simplifying instructional text tailored to AR contexts. The approach reduces cognitive load by optimizing text presentation based on user context and task complexity. This work builds on the ARTiST system [325], demonstrating its effectiveness in improving real-time task performance.

2. **Adaptive User Modeling for Proactive Task Guidance.** We develop an adaptive user modeling framework based on the Belief-Desire-Intention (BDI) model. This enables proactive, context-aware assistance that aligns with the user's goals and situational needs. The framework extends the intent modeling and collaboration mechanisms explored in Your Co-Workers Matter [326] and Satori [161].

3. **Interactive Task Recording Summarization and Analysis Tool for Post-Task Review.** To support the user's continuous improvement and task documentation, we introduce a visual analytics approach that generates interactive summaries from AR task recordings, facilitating post-task review and performance analysis. This contribution comprises two complementary components: IntentVizor [323], which introduces a general-purpose, query-guided video summarization model and interactive interface; and InsightAR, which extends this approach to AR task guidance by incorporating a domain-specific summarization pipeline and visual analytics interface for identifying user errors and generating improvement suggestions.

1.1 Motivation

AR task guidance has gained increasing attention as an effective tool for delivering real-time, context-aware instructions in various domains [212, 292]. In industrial manufacturing, AR systems assist workers in assembling complex products by overlaying step-by-step instructions directly onto physical components [212, 283]. In healthcare, AR guidance supports medical professionals in performing procedures with improved precision [190]. Similar applications extend to equipment maintenance [387], education [109], and wayfinding [263, 295], where AR helps users carry out tasks more efficiently and accurately. As AR task guidance becomes more widespread, ensuring its usability across different contexts and user groups presents new challenges. Effective AR task guidance requires not only accurate instructions but also adaptability to dynamic task environments and user needs [212, 317]. Users may have varying levels of expertise, cognitive capacity, and situational awareness, which impacts their interaction with AR systems. Without adaptive support, static or overly complex instructions can increase cognitive load, hinder task performance, and reduce overall user experience.

Challenges

Despite its potential, AR task guidance faces several key challenges that limit its effectiveness in practical applications:

1. **High Cognitive Load.** AR task guidance often presents information directly in the user's field of view. If not designed carefully, this can lead to information overload, increasing the cognitive demands on the user [33]. Complex instructions, excessive visual elements, or poorly timed prompts can distract users and hinder task performance. Reducing cognitive load is critical to maintaining user focus and ensuring efficient task execution in AR environments.

2. **Adaptivity to Dynamic Task Contexts.** Tasks performed in AR environments are often dynamic, requiring the guidance system to respond to changing conditions, user actions, and individual differences [317, 322]. Many existing systems offer static, one-size-fits-all instructions [212, 387], which do not account for variations in user expertise, task complexity, or situational factors. Adaptive guidance, tailored to the user's current state and environment, is essential for providing relevant and effective support during task execution.

3. **Holistic Framework for Post-Task Analysis.** Most AR task guidance systems focus on in-task support [147] but offer limited support for structured post-task analysis or after-action review [125]. Without structured post-task analysis, it is difficult for users or organizations to identify areas for improvement, learn from past experiences, or support long-term skill development. A holistic framework that integrates task guidance with visual summarization and post-task analysis can provide actionable insights and documentation to enhance future performance and learning.

1.2 Contributions

In this dissertation, we describe three contributions that address the challenges described in Section 1.1. The main contributions can be summarized as follows.

1. **Adaptive Text Simplification for AR Task Guidance.** We develop an adaptive text simplification method designed specifically for AR task guidance scenarios. The method dynamically adjusts instruction complexity and presentation to reduce cognitive load and improve in-task efficiency. This approach is demonstrated in the ARTiST system [325], which optimizes textual guidance for AR environments.

2. **Adaptive User Modeling for Proactive Task Guidance.** We introduce an adaptive user modeling framework based on the Belief-Desire-Intention (BDI) model. This enables us to build a proactive AR task guidance system called Satorito proactively assist users by anticipating their needs and responding to dynamic task contexts. The framework also builds on the work in intent modeling and human-AI collaboration presented in *Your Co-Workers Matter* [326]

3. **Visual Documentation and Post-Task Analysis for AR Task Recordings.** We propose a holistic framework that integrates visual documentation with post-task analysis. Building on our interactive summarization work IntentVizor [323], which combines visual analytics with user-controlled summarization through graph convolutional networks, we now develop an AR-recording-specific tool InsightAR for documenting egocentric task recordings. This tool extends traditional video summarization by ensembling diverse information such as task errors and improvements while incorporating task knowledge into the analysis process. The tool can generate interactive summaries from AR task recordings, providing actionable insights for long-term performance improvement and learning.

1.3 Organization

The remainder of this dissertation is organized as follows. First, Chapter 2 reviews related work on AR task guidance, adaptive user modeling, and task recording analysis. Next, Chapter 3 introduces our adaptive text simplification approach for AR task guidance, which reduces cognitive load and improves in-task efficiency. Chapter 4 presents the adaptive user modeling framework based on the Belief-Desire-Intention (BDI) model, enabling proactive and context-aware guidance. Chapter 5 describes the work on the interactive summarization with the visual analytics support. Chapter 6 describes our visual documentation tool for post-task analysis, which generates interactive summaries from AR task recordings to support long-term performance improvement. Finally, Chapter 7 concludes the dissertation by summarizing the contributions and outlining directions for future research.

Chapter 2

Related Work

Augmented reality (AR) task guidance has gained increasing traction in various domains, such as industrial assembly, healthcare, and education. While existing systems have demonstrated the potential to improve task performance, several challenges remain in enhancing adaptivity, reducing cognitive load, and supporting post-task analysis. In this chapter, we review related work across three main areas relevant to this research: (1) AR task guidance and text presentation, (2) adaptive and proactive guidance through user modeling, and (3) visual documentation and post-task analysis. We conclude with an overview of the specific challenges of analyzing egocentric video data, which informs the design of our framework.

2.1 Foundations and Challenges in AR Task Guidance

2.1.1 Existing Systems and Applications

Augmented Reality (AR) task guidance systems have been increasingly used to support procedural tasks in domains such as cooking [43, 357], surgery [17], and industrial maintenance [212, 387]. By leveraging head-mounted displays (HMDs), these systems present task-relevant instructions directly in the user's physical workspace [12, 171, 291]. This approach allows users to receive step-by-step guidance without diverting attention away from the task environment. The concept of task guidance systems was introduced by Ockerman et al. [209], who envisioned them as structured references for supporting procedural tasks such as inspection and assembly. Early AR task guidance systems focused on digitizing traditional manuals and displaying them through AR interfaces. In industrial maintenance contexts, AR-based systems have been shown to reduce cognitive load and improve task performance compared to conventional paper manuals [71, 301]. Similar benefits have been observed in other domains, including military [99] and healthcare applications [17], where AR task guidance has contributed to reducing error rates and increasing user satisfaction [284, 375]. To enhance task understanding, AR task guidance systems often include visual elements such as text annotations and graphical highlights [128, 206, 255]. These features can help direct user attention to critical information. However, many existing systems offer limited customization for AR environments, frequently transferring content from paper-based manuals without considering the constraints of AR devices, such as restricted field-of-view (FoV) [71, 128]. This can lead to challenges in readability and comprehension, particularly in dynamic or complex tasks. In addition to traditional task guidance systems, some AR applications incorporate virtual agents to assist with procedural tasks. These have been developed for domains including assembly and manufacturing [24, 133, 169, 298], surgery [63, 231], maintenance [23, 73, 134], and cooking [50]. While these systems often demonstrate task-specific effectiveness, they are generally limited in their ability to generalize across different contexts. Command-based interactions are commonly employed to increase user control and system responsiveness [128]; however, they typically require explicit user input, which can interrupt workflow and limit usability in hands-busy scenarios. In this thesis, we define *AR task guidance* as systems that deliver instructional support through AR interfaces, including those that employ virtual agents or command-based interactions. Our work builds on this foundation by addressing the limitations of static and input-dependent systems, proposing adaptive AR task guidance methods that can proactively respond to user needs without requiring explicit commands.

2.1.2 Challenges in Task Guidance

Many existing AR task guidance systems rely on rule-based approaches, where actions or reminders are triggered by predefined conditions such as user inputs, task progress, or environmental events. These systems are often simple to design and implement [146], and they provide users with readily available instructions or support that can be accessed on demand or at predefined stages in the task sequence.

However, rule-based task guidance has several limitations. First, these systems typically require extensive manual configuration to define rules, triggers, and workflows for each task scenario. Second, they often depend on explicit user interactions to confirm task progress or request further instructions, which can interrupt task flow and reduce efficiency. For example, Sara et al. [250] demonstrated an AR maintenance system in which technicians were required to manually confirm the completion of each task step, using either touchpad controls or voice commands, before receiving subsequent instructions. Such manual confirmation can increase cognitive load, particularly in hands-busy or safety-critical environments.

Moreover, rule-based systems generally lack adaptivity. They are often limited to predefined instructions that do not account for variations in user expertise, changes in task conditions, or unforeseen events during task execution. As a result, users may receive instructions that are irrelevant, redundant, or poorly timed, leading to inefficiencies and potential errors. These limitations highlight the need for adaptive task guidance systems that can dynamically adjust to users' needs and task contexts without requiring explicit input.

2.1.3 Text Presentation in AR

Text presentation is a fundamental feature of AR task guidance systems. It enables the display of instructional content that aligns with and supplements the physical environment, supporting users as they perform procedural tasks [12, 229]. Effective text presentation in AR can improve user comprehension by providing contextual information directly within the task space.

However, presenting text in AR poses several challenges. One primary issue is the potential for occlusion, where overlaid text may block or interfere with the user's view of physical objects. Hardware limitations, such as the restricted field-of-view (FoV) of head-mounted displays, further constrain the available display area and require careful placement of text elements [38]. To mitigate occlusion, prior work has proposed techniques that incorporate depth information [104, 321], inter-frame motion analysis [156], and two-dimensional collision detection algorithms [29, 293]. Despite these efforts, occlusion and collision remain difficult to avoid, especially in complex and dynamic environments.

In addition to occlusion, the legibility of text in AR depends on factors such as font selection, text placement, and presentation style. Orlosky et al. [210] proposed an automated text placement algorithm that adapts to both physical and virtual backgrounds to improve readability. Rzayev et al. [248] conducted an empirical study on text display types and positions during sitting and walking tasks, concluding that top-right placement increases cognitive load and reduces comprehension. Matsuura et al. [191] further investigated the readability of six Japanese fonts displayed on HMDs while walking. They found that fonts with thin horizontal and vertical lines decreased legibility due to the effects of motion and vibration.

Furthermore, the coordinate system used for text display can impact user experience. Three main coordinate systems are commonly employed in AR: world-locked, head-locked, and body-locked displays [18]. Body-locked displays, which adjust text position relative to the user's body, have been shown to be effective in scenarios where users need to move while completing tasks.

While previous work has addressed font selection, layout design, and spatial positioning, existing AR task guidance systems have largely overlooked the optimization of text content itself. Instructions from traditional manuals are often transferred directly into AR environments without consideration for the cognitive load imposed by lengthy or complex text. Our contribution on the *ARTiST* project [325] specifically addresses this gap by introducing an adaptive text simplification approach for AR task guidance. This method dynamically simplifies instructional text, reducing sentence complexity and length while preserving essential information. By tailoring text presentation to the constraints of AR head-mounted displays and the needs of users in task contexts, our work on *ARTiST* aims to improve readability, reduce cognitive load, and enhance task performance in AR environments.

2.2 Adaptive Task Guidance in AR

2.2.1 Proactive Task Guidance

Proactive task guidance systems are designed to recognize contextual information and infer user intentions, even when these intentions are not explicitly communicated by the user [57, 232, 256]. Unlike reactive systems, which depend on direct user input or predefined triggers, proactive approaches anticipate user needs and provide assistance without requiring human intervention [140, 251, 354]. Proactive task guidance has been explored across various domains, including healthcare [234, 257, 338], navigation [223], and laboratory education [259].

These systems have shown potential in improving usability [258], fostering user trust [138], and enhancing task efficiency [351]. In AR task guidance, proactive systems can leverage environmental context to predict user goals and deliver timely, context-aware recommendations [113, 195, 197, 269]. For instance, gaze-moderated systems, such as iBall, demonstrate how gaze data can be integrated into visualizations to promote task engagement and attentiveness [45].

However, most existing proactive guidance systems rely on predefined rules based on location, time, or task events to trigger interventions [197]. For example, Ren et al. [245] proposed a proactive interaction design for smart product-service systems that use sensor data—such as physical location, light intensity, and environmental temperature—to infer user states like attention levels. While these methods advance proactive interaction, they often depend on explicit contextual signals that may not accurately reflect the user's actual needs. This reliance can result in assistance that is ineffective or obtrusive [141, 331].

One key challenge in achieving truly proactive task guidance is the difficulty of accurately interpreting user intentions. Understanding user goals typically requires more than explicit cues (e.g., verbal commands); it often depends on implicit, nonverbal signals, such as gaze direction, body posture, and physical interactions with objects [128]. Successfully analyzing these implicit cues is essential for reliable intention recognition and timely intervention.

Recent advances in vision-language models present new opportunities to improve multimodal understanding of user behavior in AR task guidance. Our contribution on the *Satori* project [161] leverages these advancements by introducing a multimodal input framework that integrates both voice and visual cues to better infer user intentions. By dynamically modeling user goals and the task context, our approach enables adaptive and proactive task guidance in AR environments, reducing the reliance on predefined rules and explicit user input.

2.2.2 User Modeling for Task Guidance

2.2.2.1 Understanding User Intention

Understanding user intention is critical to providing adaptive AR task guidance, as accurately inferring user intent ensures timely and relevant guidance. Early work by Broder [31] classified web search intentions into navigational, informational, and transactional types, laying a foundation for intention modeling research. Dearman et al. [56] further extended this by categorizing sharing intentions into nine distinct types, broadening the taxonomy's scope to collaborative interactions. Similarly, Church et al. [48] examined how users' intentions vary with context, such as location or activity (e.g., commuting), influencing the design of context-sensitive information retrieval systems like SocialSearchBrowser. Expanding on this, Li et al. [166] defined a comprehensive design space of follow-up actions, categorized into 17 types, to better support proactive guidance on mobile AR platforms. Collectively, these studies emphasize that accurately modeling user intention requires consideration of both explicit user inputs (e.g., verbal requests) and implicit context cues (e.g., body movements, environmental signals). However, prior approaches have primarily focused on explicit intention signals, providing limited insights into implicit cues essential for adaptive AR task guidance. Our work addresses this gap by proposing a multimodal approach to intention recognition, leveraging visual and semantic cues to dynamically infer user intent.

2.2.2.2 Theory of Mind and the BDI Framework

To effectively support adaptive guidance, systems must accurately model users' internal states, including their beliefs, desires, and intentions. The Theory of Mind (ToM), originating from psychology, describes an individual's ability to attribute mental states—such as beliefs, desires, and intentions—to oneself and others, facilitating behavior interpretation and prediction [37, 74, 227]. Computational implementations of ToM have recently attracted attention for enhancing the realism and adaptability of human-AI collaboration [328].

The Belief-Desire-Intention (BDI) framework [26, 49, 120, 158] operationalizes key ToM concepts, offering a structured model of human decision-making widely applied in cognitive modeling and multi-agent systems [130, 153, 221, 239]. Within the BDI framework, *beliefs* represent the user's knowledge about the environment; *desires* reflect high-level goals or preferences; and *intentions* denote immediate goals that direct the user's actions [26]. The BDI model has been extensively adopted in agent-oriented programming languages (e.g., AgentSpeak(L) [238], JADEX [27], GOAL [100]), demonstrating its utility in designing interpretable and responsive systems [22, 131, 240].

However, existing BDI-inspired systems typically focus on structured decision environments, whereas AR tasks involve complex, multimodal interactions, including egocentric visual inputs, gestures, and speech [20, 173]. Thus, directly applying traditional BDI approaches to AR settings is challenging. Our contribution in the *Satori* project [161] addresses this challenge by adapting BDI modeling principles specifically for AR environments. Leveraging recent advances in Large Language Models (LLMs) [22], our approach enables real-time inference of user states, bridging visual, semantic, and contextual information for adaptive task guidance.

2.2.2.3 Modeling User State in Human-AI Collaboration

Accurately modeling the user state is a long-standing challenge in Human-Computer Interaction (HCI), essential for adaptive computing systems across various domains [16, 194]. Past research has explored modeling user goals and intents [339], user expertise levels [302], and memory states in AR/MR contexts [95, 267]. However, these studies often narrowly target specific aspects, with limited attention given to comprehensive modeling of user states.

Particularly in AR task guidance, accurately capturing a user's state requires going beyond explicit task goals and incorporating implicit factors, such as user beliefs and background knowledge. Existing research seldom addresses user beliefs, despite their importance in accurately interpreting user actions and guiding decisions [80, 151, 297, 323]. Moreover, current approaches frequently fail to clearly differentiate between high-level goals (desires) and immediate action-oriented intents [135].

Our contribution explicitly addresses this gap by proposing a unified user modeling approach based on the BDI framework. By integrating beliefs, desires, and intentions into a cohesive representation of the user's state, our model supports adaptive AR task guidance systems capable of responding effectively to dynamic user contexts and needs.

2.2.3 Leveraging LLMs for Adaptive Guidance

2.2.3.1 LLMs for Task Guidance

Recent advancements in Large Language Models (LLMs) have demonstrated significant promise in supporting diverse task guidance applications. Most prior research on LLM agents has focused primarily on single-agent scenarios, such as web navigation and text-based games [75, 127, 330, 346, 380]. Techniques like chain-of-thought prompting [127, 319, 361], self-consistency decoding [312], task decomposition [377], and error reflection [347] have significantly improved the planning and reasoning capabilities of these models.

Recent work has extended LLM-based systems into multi-agent collaborative settings [88, 382], though most existing studies either involve homogeneous groups of LLM agents or only evaluate overall task outcomes without thoroughly examining collaborative interactions [105, 290, 329]. Our contribution on the *CoBlock* project [326] explicitly addresses this limitation by designing evaluations to directly assess and highlight collaborative interactions among LLM-based agents within diverse task scenarios.

2.2.3.2 Theory of Mind for Human-LLM Collaboration

Adaptive AR task guidance requires accurate interpretation and prediction of user intentions and actions. The Theory of Mind (ToM), initially defined by Premack and Woodruff [227], provides a psychological framework for understanding mental states such as beliefs, desires, and intentions, facilitating effective social interactions and collaboration [37, 74]. Computational ToM frameworks, such as Wu et al.'s cognitive knowledge graph COKE [328], formalize these mental states, demonstrating their potential for AI systems to interpret and predict complex human behaviors.

The Belief-Desire-Intention (BDI) model, derived from ToM principles [26, 49, 120, 158], provides a structured approach to modeling human decision-making. In the BDI model, *beliefs* represent users' understanding of their environment, *desires* correspond to their high-level goals, and *intentions* reflect immediate action plans guiding behavior [130, 153, 221, 239]. Previous studies demonstrate that the BDI framework enhances interpretability and responsiveness of agent behaviors in interactive systems [22, 27, 100, 238, 240].

Integrating the principles of ToM and the BDI framework with LLMs presents an opportunity to build adaptive AR systems capable of sophisticated human-AI collaboration. Recent research has explored the extent to which ToM-like reasoning emerges naturally within LLMs [114, 136, 181]. Building on these insights, our contribution on the *Satori* project [161] explicitly incorporates partner-state modeling approaches derived from ToM and BDI into LLM-driven prompting strategies. By embedding ToM principles directly within the prompting process, our method enables the AR guidance system to dynamically infer user mental states and intentions from implicit multimodal cues, including visual and gestural information. This approach allows the adaptive guidance system to proactively provide context-sensitive, personalized support to users, thus enhancing collaborative interaction in AR environments.

2.3 Task Guidance Process Analysis

Analyzing task guidance performance through recorded AR interactions provides essential insights for identifying task-related errors and improving long-term efficiency. Recorded task data, especially from egocentric videos, contains rich multimodal information capturing user interactions, environmental context, and procedural actions. Due to the complexity inherent in egocentric recordings, traditional video analysis techniques face challenges in accurately capturing context-sensitive details. This section discusses existing approaches to analyzing task-oriented videos, starting from general video summarization methods and multimodal analytical tools, followed by the unique challenges presented by egocentric recordings. Throughout, we highlight specific gaps addressed by our contributions on the *InsightAR* project.

2.3.1 Tools for Video Summarization

Video summarization aims to create concise representations by capturing key events, objects, actions, and salient moments from video content [93, 276, 323]. Summaries commonly take the form of visual keyframes or textual descriptions [205], facilitating efficient skimming and reducing cognitive load during video reviewing tasks [316].

However, single-modality approaches relying solely on visual content often fail to capture essential semantic relationships among actions. For instance, keyframe-based summaries cannot effectively convey action-based errors or procedural missteps in videos such as those showing coffee preparation. Multimodal summarization techniques address this limitation by combining visual content with corresponding textual descriptions, thus highlighting relationships between actions, objects, and events [96, 163, 384]. Although multimodal summaries provide improved semantic insights for various domains including surveillance and medical training [115], current tools offer limited interactivity and user control over the summarized content. Users have difficulty filtering information based on domain-specific criteria or easily transitioning between summary and original footage for detailed inspection. Our contribution in the *InsightAR* project directly addresses this issue by integrating multimodal video summarization with interactive capabilities specifically tailored for analyzing egocentric task recordings.

2.3.2 Tools for Multimodal Analysis of Task Performance

Multimodal analysis tools employ machine learning techniques to extract and interpret information from task-oriented videos. Such tools can identify procedural anomalies, including missed steps or incorrect sequences, task-specific errors like improper tool usage, and key events critical for assessing performance. Recent research has applied multimodal analysis to diverse tasks, ranging from everyday activities such as cooking to specialized domains like surgical training and industrial assembly [342, 343]. Advances in multimodal learning and computer vision have significantly increased the sophistication of analyzing procedural activities [176].

However, existing multimodal analysis methods typically focus on extracting low-level features, such as actions or object interactions, and often fail to generate actionable insights without integrating user domain knowledge [307]. To overcome this limitation, visual analytics systems like Performance Lens [371] and interactive task analysis frameworks [142] combine automated computational analyses with human expertise, enabling richer and more contextually meaningful interpretations of task performance data.

Despite these improvements, most current systems primarily analyze structured data, such as key performance indicators (KPIs) and textual records [252], rather than video data, even though video is the most intuitive medium for capturing task execution. This limitation is particularly prominent for egocentric video analysis, where subtle contextual details require expert interpretation and flexible analytical interactions. Our contribution in the *InsightAR* project fills this gap by combining advanced multimodal video analyses with interactive features, allowing domain experts to better explore, interpret, and derive insights from egocentric task videos.

2.3.3 Challenges of Egocentric Video Analysis

The increasing availability of wearable cameras and AR headsets, such as HoloLens, Meta Quest, and GoPro, has greatly facilitated the adoption of egocentric video recording. This has resulted in extensive multimodal datasets, including Ego4D [90], HowTo100M [199], and EPIC-KITCHENS-100 [52]. These datasets include rich sensor data such as integrated measurement unit (IMU) readings, gaze tracking data, and spatial mapping, enabling detailed insights into human-object interactions, environmental context, and task performance [42].

However, analyzing egocentric videos poses unique challenges. First, wearable cameras generate inherently unstable footage due to head movements, causing inconsistent lighting, changing viewpoints, and frequent occlusion, complicating traditional vision-based analyses [90, 309]. Second, critical objects and events often move in and out of the camera's limited field of view, requiring advanced techniques to maintain temporal continuity for accurate recognition and tracking [285]. Finally, egocentric perspectives yield visual patterns that significantly differ from traditional third-person viewpoints, demanding specialized analytical methods and models explicitly trained on egocentric datasets [87].

Current methods remain largely limited to basic action recognition and rarely provide sufficient contextual understanding for identifying procedural errors or anomalies. Furthermore, existing egocentric analysis systems often lack interactivity, preventing domain experts from incorporating their expertise during the analytical process. Our contribution in the *InsightAR* project directly addresses these limitations by offering interactive, multimodal analysis capabilities, specifically designed to enhance the usability and interpretability of egocentric video data in task guidance scenarios.

Chapter 3

ARTiST: Automated Text Simplification for Task Guidance in Augmented Reality

3.1 Introduction

AR has evolved into a transformative technology with far-reaching applications across multiple domains, including education [3, 110, 327], entertainment [179, 216, 225, 230], collaborative work [18, 19, 62], and professional training [12, 19, 288, 374]. Notably, AR superimposes digital content onto the physical world in real-time to facilitate more efficient task execution [228, 334]. As a result, AR applications have been increasingly employed for task guidance in manufacturing [212, 387], education [11, 109], and surgery [216]. AR devices have been widely adopted in the manufacturing industry, for example, to reduce reliance on guidance materials or other devices outside of the immediate work environment [147].

A head-mounted display (HMD) is a type of AR device that allows for multimodal interactions while the user maintains focus on the immediate work environment [70, 147]. Given their hands-free nature, HMDs are frequently used for text-based task guidance. However, compared to desktop displays, HMDs have a relatively small field of view (FoV) limiting the amount of text they can display; longer instructions may occlude a user's view resulting in lower-productivity and higher cognitive load [33], as well as safety risks [216]. As a result, AR text-based

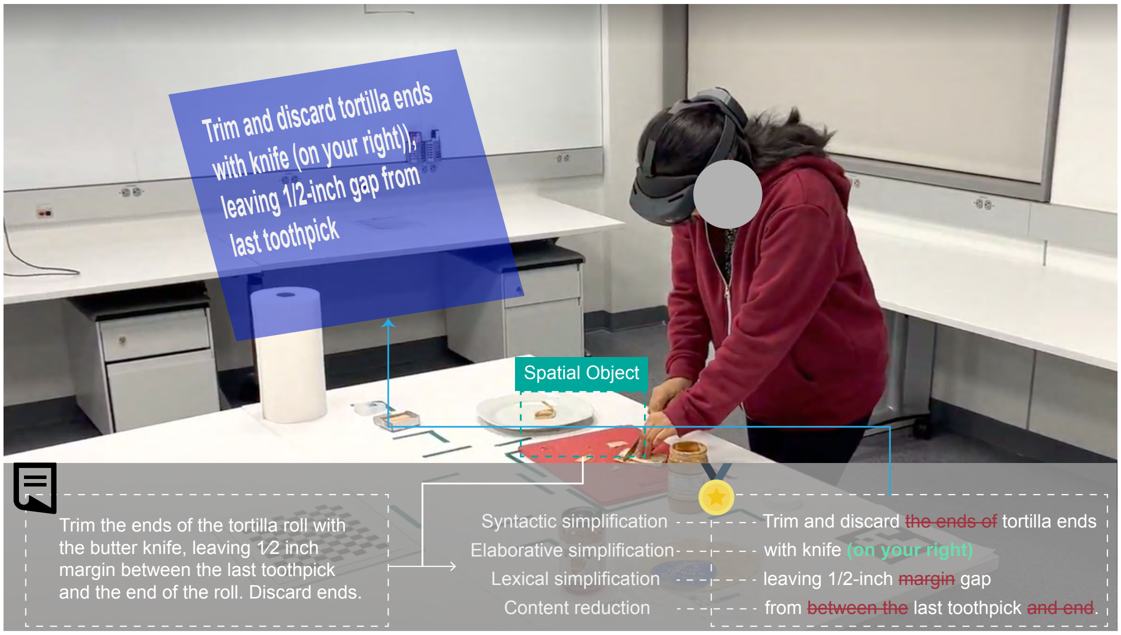

Trim and discard tortilla ends with knife (on your right), leaving 1/2-inch gap from last toothpick
Spatial Object
Trim the ends of the tortilla roll with the butter knife, leaving 1⁄2 inch margin between the last toothpick and the end of the roll. Discard ends.
Syntactic simplification
Elaborative simplification
Lexical simplification
Content reduction
Trim and discard ~~the ends of~~ tortilla ends
with knife (on your right)
leaving 1/2-inch ~~margin~~ gap
from ~~between the~~ last toothpick ~~and end~~.

Figure 3.1: ARTiST is a text simplification system that is designed for augmented reality (AR) head-mounted display (HMD) environments. Our system combines the findings from a formative study with a novel few-shot prompting to integrate four established text simplification techniques for AR-specific contexts. The example text shown in the bottom-right corner of the figure has been simplified using our approach. The red text indicates removals whereas the green highlights the addition of spatial information. The resulting simplified text is displayed directly in the HMD.

instructions require optimization for better utility.

Text simplification offers one potential solution. This process has historically been used to reduce the complexity or length of text for users [274] in non-AR settings, making it more readily understandable. To the best of our knowledge, however, there are presently no established methods for adapting and simplifying text for better utility in the specific context of AR. Furthermore, applying existing methods to AR raises several concerns. Firstly, traditional text simplification methods typically work to facilitate comprehension for individuals with limited reading skills [41, 265], an audience that may not overlap with the AR userbase. Secondly, these methods have not been designed or fine-tuned to accommodate AR-specific constraints, such as the small FoV, a restricted display area, or the necessity of users performing physical tasks concurrent with reading [216]. Finally, text simplification presents an opportunity to use spatial information AR content by describing a physical object's color, location, or direction. Textually indicating the location of a physical object can, for example, assist users in AR task execution [216, 387].

Accordingly, we aim to implement a text simplification system for AR by tailoring the existing methods to the AR context, with the goal of reducing cognitive load on users and improving their task performance. To do so, we build on insights from prior work [44, 265, 273, 274] and our own formative study to understand the specific challenges of AR text interfaces as well as their limitations and potentials. The formative study contains three parts: a literature survey, an open-ended exploration, and an expert interview. We found that both participants and experts addressed issues related to long-text-induced reading challenges (e.g., cognitive load) and comprehension. Interviews with participants and experts further elicited three design guidelines that helped build ARTiST, an automated text simplification system with few-shot prompting. This system leverages the multi-task capabilities of the large language model (LLM) GPT-3 in combination with our newly formed simplification techniques, eliminating the need for extensively annotated data [236]. We crafted prompts based on chain-of-thought principles, considering text simplification in AR [318]. Specifically, ARTiST introduces two novel simplification methods: the "plan-of-technique" prompt and "error-aware" model calibration, enhancing the effectiveness and reliability of text simplification.

A sample workflow for ARTiST is available in Figure 3.1

We tested ARTiST via two studies that entailed an empirical evaluation with 16 participants. The first study included task guidance to make pour-over coffee and set up a meeting room according to specific criteria. The second study asked participants to perform video editing on an iPad using AR instructions given through HMD. These studies explore how our proposed system can better benefit participants over the unmodified text and existing methods by assessing related performance metrics, cognitive load, and subjective ratings. The results indicate that ARTiST significantly improved task guidance performance by reducing the number of errors participants made, increasing the number of steps they correctly memorized, and reducing their cognitive load.

In summary, our work includes the following contributions:

1. The ARTiST, a novel system for text simplification in AR using few-shot prompts and customized GPT-3. This system incorporates chain-of-thought, plan-of-technique, and error-aware calibration to tailor text simplification for AR.

2. Results and design guidelines from a formative study that includes a literature review, an open-ended exploration with seven participants, and an expert interview with three field experts for text simplification in AR.

3. A 16-participant empirical evaluation of ARTiST against baseline and existing methods, which shows that ARTiST significantly reduces errors and overall cognitive load with similarly higher subjective ratings on text readability, memorability, guidance, and trust among users.

To further support the development of the field, we open-source our implementation[1].

3.2 Formative Study

This study involves three parts: a literature review, an open-ended exploration, and expert interviews to understand the needs around text simplification in AR. We wish to explore the following aspects of text simplification:

[1] Code is available at https://github.com/VIDA-NYU/artist

[**RQ1**] Which text simplification methods from the field of natural language processing (NLP) can be effectively applied in an AR context?

[**RQ2**] Would text simplification improve comprehension, per its benefit for low-literacy readers in traditional platforms?

[**RQ3**] Can text simplification lead to increased user satisfaction, and hence to a more positive AR experience overall?

3.2.1 Part I: Survey of Text Simplification

To address **RQ1**, we initiated a comprehensive review of existing literature on traditional text simplification; our goal was to identify NLP techniques that might be useful for AR applications. Our review commenced with an in-depth examination of three seminal survey papers [4, 265, 273]. We extended our scope by traversing both the references cited in these papers and consequent citations of them to gain an encompassing understanding of current methodological approaches. Subsequently, we identified four NLP techniques pertinent to our inquiry: content reduction (**A1**), syntactic simplification (**A2**), lexical simplification (**A3**), and elaborative simplification (**A4**). We describe the four techniques below.

3.2.1.1 A1: Content reduction

Content reduction in text simplification aims to achieve clarity and conciseness by eliminating or restructuring non-essential elements without altering the core message [207]. Strategies include removing non-essential information, shortening sentences, and eliminating repetition. This technique is particularly beneficial in constrained display environments like those of AR, where succinct, clear text enhances user interaction and comprehension [25].

3.2.1.2 A2: Syntactic simplification

Syntactic simplification involves rephrasing complex grammatical structures into simpler ones while still retaining the original meaning [254, 273]. Existing methodologies often target specific complex linguistic features such as coordination, subordination, relative clauses, passive constructions, and extended sentence lengths [4, 273].

3.2.1.3 A3: Lexical simplification

Lexical complexity often arises from the use of intricate words and phrases. To mitigate this, one widely employed strategy is to replace complex lexical items with simpler synonyms. This form of lexical simplification has seen extensive application in the context of second-language learning, primarily because it aids in comprehension and vocabulary acquisition for learners who may not be familiar with advanced or specialized terminology [211].

3.2.1.4 A4: Elaborative simplification

Elaborative simplification entails providing explanations of complex concepts. This technique is prevalent in professional textbooks, which frequently encompass specialized or technical subject matters [124]. In AR, spatial information becomes increasingly critical for user comprehension and task performance. Consequently, elaborative simplification can be especially beneficial for clarifying spatial metrics and locations. Spatial metrics refer to numerical measurements, such as distances or sizes denoted in units like inches or centimeters. These metrics often need to be elaborated to provide context or improve comprehension. Similarly, spatial locations, which may involve GPS coordinates or relational positioning (e.g., ”next to,” ”above,” ”beneath”), can be clarified through elaborative simplification to facilitate user orientation and task execution in AR environments [106].

3.2.1.5 Target application and users

Traditional text simplification techniques are normally targeted at non-native speakers or people with cognitive or literacy limitations, e.g., autism [54, 65, 341], aphasia [35, 40, 208], dyslexia [78, 101, 112, 242, 243, 244], hearing impairment [5, 6, 7, 8, 303] and language learners [177]. The associated benefits are largely attributed to simplified grammar structures and the use of common words, which can significantly reduce information processing time.

In a similar vein, AR users may encounter reduced reading capability due to the challenges associated with the AR setting. Studies have demonstrated that AR users experience reduced reading speed [241], lower comprehension [25, 76, 77], and increased cognitive load [61]. For instance, Rau et al. report that readers' response time in AR is longer than that associated with desktop reading [241]. Hardware limitations comprise a major contributing factor, impacting refresh rate, resolution, and FoV, and ultimately impeding text display due to registration errors [103, 182], extra latency, and visual artifacts. Moreover, users' rapid movements and the surrounding open environment can result in unstable text displays. Prolonged use of AR devices may also lead to eye strain and fatigue due to constant accommodation and vergence adjustments, making reading more challenging than on traditional displays. Finally, AR displays often overlay digital information on real-world information, affecting reading comprehension and focus.

Yet, the main reason for the reduced legibility of AR text is the user's elevated cognitive load in the immersive environment; evidence shows that AR users can experience high pressure and increased cognitive load [61].

Accordingly, drawing inspiration from the research elaborated above, we *aim to investigate whether text simplification techniques can benefit users in AR environments and mitigate the challenges described.*

3.2.2 Part II: Open-Ended Exploration

To explore the effectiveness of the four previously identified text simplification techniques (**A1-4**) in AR and further investigate **RQ2**, we conducted an open-ended exploration with seven participants. According to the literature, the simplification techniques (**A1-4**) can improve comprehension in paper-based reading. However, these techniques may need to be modified for AR-specific challenges and their benefits in AR require further investigation. Therefore, the open-ended exploration aimed to assess how text simplification techniques might be revised for this context.

3.2.2.1 Participants

Seven participants (four male and three female) were recruited from a school mailing list to experience text simplification in AR. Three out of the seven are native

English speakers. All participants have some prior experience with AR (2/7 are frequent users, 4/7 are occasional users, 1/7 is VR-only).

3.2.2.2 Tasks

The open-ended exploration involves two tasks: cooking and gem-hunting. These tasks were selected for being common and applicable to AR scenarios [143, 187, 357]. In the cooking task, participants used the AR system to make a pinwheel sandwich. The AR interface showed step-by-step instructions for ingredient preparation, assembly, and cooking [25]. These instructions are adapted from a wikiHow article on how to make a pinwheel. [2] For the gem-hunting task, participants followed clues displayed on the AR device to find a gem hidden in a room. Clues included puzzles, patterns, and spatial information. The task manual is derived from a party game website.

3.2.2.3 Method

Since each task contains multiple steps, the original and simplified text for each step were displayed side-by-side to participants. . Text simplification was manually performed based on the principles of the four existing techniques taken from the literature (i.e., *content reduction*, *syntactic simplification*, *lexical simplification*, and *elaborative simplification*), with each simplification technique being used an equal number of times. Since this is an exploratory study, quantitative data is not collected. Participants are asked to think aloud while performing their AR tasks. A semi-structured interview collects participants' thoughts on text simplification in AR, its potential, and its challenges.

3.2.2.4 Procedure

Initially, participants completed both tasks using the original text instructions. They shared any challenges they faced in understanding the AR interface. Next, simplified versions of the text were presented. Participants compared and evaluated readability and comprehension. On average, the exploration lasted about one hour. We coded our interview notes and think-aloud notes and summarized participant

[2] https://www.wikihow.com/Make-a-Pinwheel

feedback on the four text simplification techniques. The open-ended study was supervised by the university-approved IRB, and participants were compensated at an hourly rate of $20.

3.2.2.5 Results

During the study, we found text content and semantics affect the reading experience in AR.

Text length in AR. Text in an AR environment introduces unique challenges that are not present in traditional display mediums. Users can scroll or zoom to manage lengthy texts in conventional formats; these interactions are more challenging to execute in the AR setting [34, 203]. Occlusion and visual clutter are some of the issues pointed out by our participants (P2), who mentioned, "*The displayed text takes up too much space and occludes the table.*" Lengthy text segments also distract users' attention away from physical tasks. P3 found it challenging to focus on the task of *sliding floss under the tortilla, perpendicular to the length of the roll,* due to the distracting nature of the extended text. These distractions sometimes pose safety risks: P7 was at risk of cutting their finger while engrossed in reading. Furthermore, text length negatively impacts how well information is retained as processing time increases with longer text segments [83, 126]. This was evident in the gem-finding task, where P4 and P5 forgot a crucial step that they had been given earlier after reading a lengthy sentence. Text length thus requires careful design in AR.

Feedback on AR text simplification techniques All participants agreed that *content reduction* is beneficial in AR. For instance, they found the sentence, *Roll the tortilla into a log shape,* more effective than the original text: *Roll the tortilla from one end to the other into a log shape.* Most participants mentioned that adding a clause to further explain text may not be necessary (preferring *syntactic simplification*). When asked about replacing complex words with simpler ones (*lexical simplification*), most participants (6/7) did not indicate word complexity as an issue. For example, the word *perpendicular* was not found to be more opaque than *at the right angle too,* and most of the participants (5/7) preferred *perpendicular* because it was shorter (4/7) and more precise (3/7). In addition, most participants (6/7) expressed that added details (*elaborative simplification*) were unnecessary. P6 said that "*the 'which' clause is verbose and takes up too much space*" (referring to the instruction *use the keys to unlock the first drawer below the desk, which should be located to your right*). For spatial elaboration, most (6/7) found it helpful when the reference object was present in the scene. P6 remarked that indicating, "*'finger size' helps me make a quick estimate of the size.*" P3 commented that indicating a spatial location in the text is helpful, and a majority of participants (4/7) said that spatial information can complement AR spatial indicators such as bounding boxes or virtual arrows in the scene.

3.2.3 Part III: Expert Interview

To verify the initial insights gained from the literature review and the open-ended exploration, we further conducted a semi-structured interview with three experts from the industry. All interviewed experts possess extensive experience with AR task guidance systems. Our objective in these interviews was to address **RQ3** by eliciting their insights on text simplification for AR and exploring potential usage scenarios.

3.2.3.1 Expert background

Each of the three experts (E1-E3) interviewed has over three years of professional experience in AR interface development.

- E1 is an AR interface designer at a research and development (R&D) company that is currently working on a HoloLens application to support field surgery. E1's users are primarily skilled professionals such as teachers and emergency medical technicians (EMTs) who use AR devices to instruct them as they identify and treat injuries such as gunshot wounds.

- E2 is an interface developer at a document solution corporation. E2 collaborated with engine mechanics to develop a HoloLens-based instructional application for displaying engine maintenance documents.

- E3 is an AR/VR researcher with top-tier publications and extensive experience in HoloLens application development. E3 has developed AR applications for everyday tasks such as cooking for non-professional users.

3.2.3.2 Method

The interview addressed the experts' backgrounds and experiences, the challenges of AR text interface design, their assessment of the need for text simplification, and the potential benefits and drawbacks associated with it. Additionally, we presented the four commonly used text simplification methods and solicited their opinions on them.

3.2.3.3 Results

We describe the results in the following subsections.

Benefits of text simplification in AR All experts recognized the need to simplify text in AR. They believed this would reduce user impatience and the likelihood of mistakes. E2 said that mechanics, for instance, might be habituated to how they perform a specific task and so rush through it without noticing updates to the process. When the related instructional text is simplified, however, they are more likely to read the instructions. E2 explained: "*One of the things that happens is the procedure changes. Users can easily go on a routine and assume they know how to do it without actually reading the instructions.*" E1 and E3 also mention that the simplified text could reduce the cognitive load and mitigate user anxiety, another set of benefits. For example, E3 indicated that: "*Reading the long text may make the users anxious,*" but this may not be the case for shorter pieces of text.

All experts indicated that simplified text reduces the chance of visual occlusion. Object occlusion (virtual objects being blocked visually by physical objects [293]) is one such instance of this. This leads to users being unable or only partially able to read the AR text, causing frustration and diminished performance. E2 mentioned that: "*(Sometimes in engine maintenance) we're gonna have a wall full of the tools, (and sometimes) we are gonna have an engine in front of you, and (so) finding someplace in the visual display is gonna be a challenge.*" E3 also mentioned that shortening and simplifying text could reduce occlusion.

Both E2 and E3 mentioned that shorter text facilitates the AR reading experience since zooming or scaling long sections of text while reading on an HMD is challenging. E2 emphasized that: "*None of the users liked pinching and zooming,*" highlighting the need for methods that do not require additional interactions.

Finally, the experts mentioned the tremendous opportunity to use text simplification as a way to help automate the conversion of text from traditional digital media (e.g., PDF) to AR. All experts conveyed that the process of creating text instructions for AR is still sub-optimal and requires extra labor. E2 stated: "*All the documents we work with start as PDF or Word documents. We basically output them to AR (devices)*". In contrast, E1 and E3 mentioned the need to make modifications to the text displayed in AR. For example, E1 attempts to shrink text or split long sections of text into multiple steps to make them shorter, saying: "*We try to keep the words as quick, punchy, and actionable as possible.*" E3 also mentioned adjusting font sizes and colors to improve legibility in the AR environment. Although full-text automation involves fitting text to the AR scene with different formats, styles, or colors, E2 pointed out that automated text simplification would still be useful as the current manual approach requires expertise that novice workers may not possess. In addition, it is not feasible to manually revise all text when new sections are added regularly.

Challenges in text simplification in AR Current AR applications lack automated solutions and established practices for text simplification (E1-3). All experts concur that manual text revision is impractical due to the constant influx of new text and the absence of a standard framework for AR text readability. This drives home the need for automated methods to adapt existing documents for presentation in AR. However, text simplification for AR poses the following challenges, and current methods are not directly applicable (E1-3).

- All experts raised concern over avoiding accidental changes to meaning during text simplification. This concern is unique to AR because users perform physical actions *live* from textual instructions. E2 elaborated: "*When working with mechanical systems in real-world scenarios, failure to follow instructions accurately could lead to catastrophic consequences,*" highlighting the importance of retaining the integrity of the original text's meaning.

- Removing duplicated content is crucial in AR given that such redundancies could increase cognitive load for AR users who are already tasked with interpreting and acting upon visual overlays. All experts agreed that elaborative simplification should weigh toward eliminating redundancies rather than adding explanatory details, which is traditional in conventional text simplification.

- Traditional text simplification techniques must be re-adapted for AR (E1-3) as they are primarily geared toward enhancing readability for low-literacy individuals and do not address the attention constraints, high cognitive load, and FoV issues typical in AR. Therefore, the development of an AR-specific text simplification tool presents a challenging yet vital task, as it must harmonize these design goals to suit the unique demands of AR settings.

3.2.3.4 Feedback on existing text simplification techniques

ID	**Technique**	**E1**	**E2**	**E3**
A1	Content reduction		✓	
A2	Syntactic simplification	✓	✓	✓
A3	Lexical simplification	✓		
A4	Elaborative simplification	✓	✓	✓

Table 3.1: Expert (E1-3) feedback on simplification techniques (A1-4). A check indicates that the expert assesses that the given technique would be useful for AR.

Table 3.1 summarizes the traditional simplification techniques our experts use in their everyday work. All experts do manual *content reduction* when creating AR instructions. E3 employs lexical simplification with the aim of retaining the text's original meaning. All experts agree that simplifying syntax, length, and grammar is beneficial for AR interfaces. However, the use of elaborative simplification needs more scrutiny in AR settings as "*the subtle balance between the content and text length must be considered*"(E1, E2, E3). For example, E1 mentioned that engine maintenance manuals often include explanations of different engine parts that may be unfamiliar to users, and such explanations should not be removed. E2 brought up that both object and numeric elaboration can be beneficial when users need to quickly identify numerous targets in AR. Elaborating AR text to describe objects in the scene is one potential application. E2 explained that using a reference object that is similar in size to the dimensions given in the text (when the object is visible) would facilitate spatial awareness. For example, E2 said that a phrase like *Move* *the handle to seven inches to left* can be elaborated as *Move the gear to seven inches left, or the length of a screwdriver.* Again, as the experts mention, consideration needs to be given to balancing text length against the need for additional content in AR.

3.2.4 Design Guidelines and Updated Simplification Techniques

3.2.4.1 Design guidelines

Through summarizing the literature survey, the open-ended exploration, and insights shared by our AR experts, we derived design guidelines and updated the four selected simplification techniques for AR task guidance.

[DG1] **Meaning preservation is paramount in text simplification.** Preserving the original text meaning [15, 274]) is the main objective when applying text simplification techniques. This finding is in line with both our interview sessions and observations. Since almost all simplification techniques may compromise original meaning [207, 273], it is essential that any substituted words convey the same meaning as their original counterparts [58].

[DG2] **Text simplification must consider both AR-specific challenges, such as issues with FoV and cognitive load, while exploring AR-specific opportunities.** Traditional text simplification techniques (e.g., syntactic simplification, lexical simplification, etc.) do not address challenges associated with AR devices, such as reading the overlayed text while doing a physical task, the constraints of a small FoV, and users' increased cognitive load while completing a task. Minimizing the display space required to render text reduces the chance of visual occlusion while optimizing syntactic structures reduces cognitive load.

[DG3] **Text simplification in AR should give priority to text length over grammatical correctness.** Traditional text simplification techniques usually prioritize grammatical correctness [4, 273]. However, we find that priority should instead be given to text length and clarity in AR. This was gleaned from the open-ended exploration, where participants expressed the need to minimize occlusion caused by text length and indicated that less strict grammar did not notably affect their comprehension if meaning was preserved. Further expert interviews supported the assessment that AR users tend to skim lengthy texts, not paying strict attention to grammatical correctness.

3.2.4.2 Updated simplification techniques (A1-4)

Based on our findings, we update the four simplification techniques to fit users' needs in the AR context. We discuss the benefits and address discrepancies within the experts' feedback below.

A1: Content reduction We found that content reduction is beneficial in AR, as both the literature review and experts suggest. However, removed content may contain important task instructions, and its absence may alter the original meaning [**DG1**]. Furthermore, as suggested by [**DG3**] and observation of the open-ended exploration, prepositions and pronouns can be cut for more concise.

A2: Syntactic simplification The results from the formative study support syntactic simplification as beneficial in AR contexts, given that complex grammatical structures can consume user attention. However, as with content reduction, syntactic simplification may alter the original meaning [273], necessitating adherence to [**DG1**]. Furthermore, simplified grammatical structure can result in overall longer text, contradicting [**DG3**]. To mitigate this, syntactic simplification should be applied only when it does not increase the number of lines of the displayed text, as addressed by E3.

A3: Lexical simplification Lexically simplified phrases may deviate from original meanings and lengthen the text, conflicting with [**DG1**] and [**DG3**]. To address this, we propose two constraints for lexical simplification: Firstly, it should not alter task-related terms, and, secondly, it should not increase the number of lines of text.

A4: Elaborative simplification Elaborative simplification elicited nuanced opinions from the experts. In the NLP literature, elaborative simplification is described as benefiting second-language learners by elucidating abstract terms. However, as noted by E1-3, explaining terms may not benefit AR users and will likely lead to increased text length (contrary to [**DG3**]). Therefore, common-sense explanations and explanations of background knowledge should be excluded from elaborative simplification to support concision. However, E1-3's feedback indicates that elaboration of the spatial context and numerical measures offers greater utility within the AR context. For instance, the user can benefit from spatial positional information such as *the cup on your left* when multiple cups are present. Additionally, when conveying numeric measures (e.g., *seven inches*), experts advised elaborating by referencing the size of objects already present within the scene, such as the diameter of a plate. By incorporating spatial context and numerical measure, elaborative simplification can be adapted within AR to adhere to [**DG2**] and enhance task performance.

3.3 ARTiST System

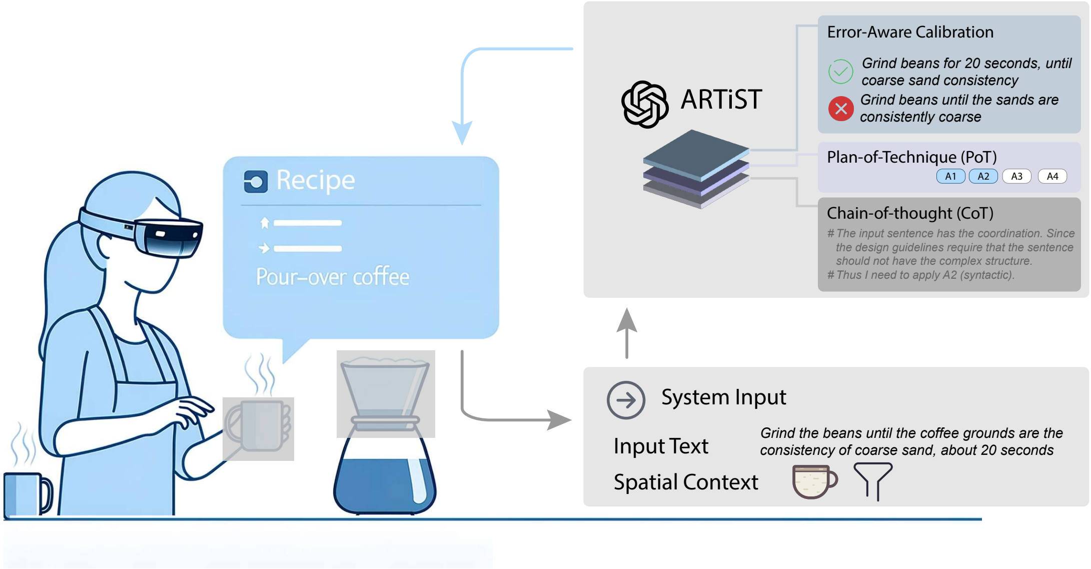

Recipe
Pour–over coffee
ARTiST
Error-Aware Calibration
Grind beans for 20 seconds, until coarse sand consistency
Grind beans until the sands are consistently coarse
Plan-of-Technique (PoT)
A1
A2
A3
A4
Chain-of-thought (CoT)
# The input sentence has the coordination. Since the design guidelines require that the sentence should not have the complex structure.
# Thus I need to apply A2 (syntactic).
System Input
Input Text
Grind the beans until the coffee grounds are the consistency of coarse sand, about 20 seconds
Spatial Context

Figure 3.2: Method Overview: ARTiST uses OpenAI's GPT-3 model, prompted with chain-of-thought and technique-as-plan methods, to generate simplified text candidates. The candidates are calibrated to reduce the likelihood of potential errors. The resulting simplified text is then displayed within a HoloLens 2 application. The spatial context is captured by detecting the objects in the scene to support the elaborative simplification.

In this section, we describe the design of ARTiST, which has been developed using the updated simplification techniques (Table 3.1) and design guidelines derived from the formative study.

ARTiST employs three novel methods, shown in Figure 3.2, to customize the GPT-3 model to stably output the desired simplification results. These include utilizing the chain-of-thought method to enhance GPT-3's reasoning capabilities and the plan-of-technique method for selecting the most appropriate techniques from **A1-4** ([**DG2**] and [**DG3**]). Additionally, ARTiST implements error-aware calibration to ensure the preservation of the original text's meaning([**DG1**]).

3.3.1 Plan-of-Technique Prompting

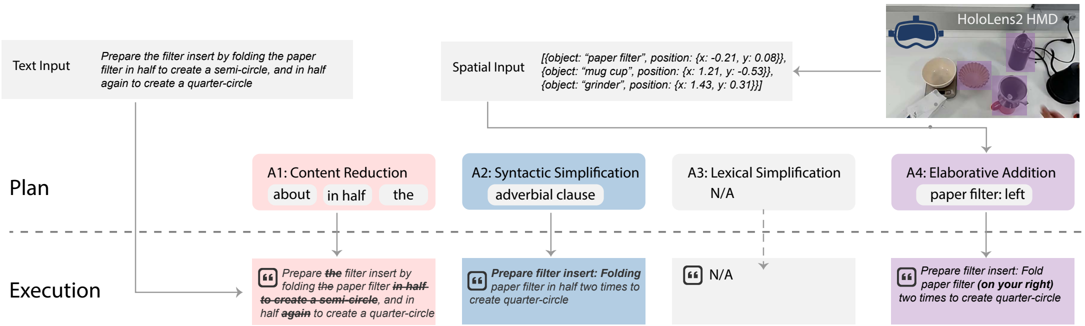

HoloLens2 HMD
Text Input
Prepare the filter insert by folding the paper filter in half to create a semi-circle, and in half again to create a quarter-circle
Spatial Input
[{object: "paper filter", position: {x: -0.21, y: 0.08}}, {object: "mug cup", position: {x: 1.21, y: -0.53}}, {object: "grinder", position: {x: 1.43, y: 0.31}}]
Plan
A1: Content Reduction
about
in half
the
A2: Syntactic Simplification
adverbial clause
A3: Lexical Simplification
N/A
A4: Elaborative Addition
paper filter: left
Execution
Prepare ~~the~~ filter insert by folding ~~the~~ paper filter ~~in half to create a semi-circle~~, and in half ~~again~~ to create a quarter-circle
Prepare filter insert: Folding paper filter in half two times to create quarter-circle
N/A
Prepare filter insert: Fold paper filter (on your right) two times to create quarter-circle

Figure 3.3: Plan-of-technique: The input text and the spatial context are fed into the LLM, which first generates a plan of the simplification techniques. The techniques will be sequentially applied to the input text to generate the final simplified text.

The plan-of-technique method is designed to structure the simplification process through a plan of different simplification techniques (**A1-4**). These techniques guide the GPT-3 model in executing the simplification as intended. This planning-and-execution model has been widely adopted in code generation [363], open-world agents [315], and robotics [275] for controllable and stable outputs. Figure 3.3 shows how input texts and the spatial context are fed into GPT-3 to generate the simplification plan.

Step-by-step execution ensures that all necessary simplification techniques can be applied. Our preliminary experiments reveal that GPT-3 sometimes forgets the techniques and design guidelines. One explanation for this is that LLMs like GPT-3 are typically trained on generic corpora without access to specialized design guidelines. Our plan-of-technique thus decomposes the simplification process into different simplification steps, mitigating forgetfulness. Moreover, such a structured pipeline can elicit the GPT-3's multi-hop reasoning capability shown in other NLP tasks [1, 270, 281].

In text simplification, multiple simplification techniques can sometimes conflict with each other and require multi-hop reasoning to resolve. For example, elaborative simplification (**A4**) may conflict with content reduction (**A1**). The plan-of-technique guides GPT-3 to consider the different techniques before executing the actual simplification actions, thereby reducing potential conflicts.

3.3.2 Chain-of-Thought Prompting

Chain-of-thought prompting is used to further enhance GPT-3's multi-hop reasoning capabilities and resolve potential technique conflicts. In few-shot prompting, a series of exemplars are created to instruct GPT-3 on how to generate the desired output based on the input text. Chain-of-thought augments the exemplars with intermediate reasoning steps, leading to the final output [318]. Drawing upon the proven efficacy of chain-of-thought's applications in diverse fields [129, 271, 306], we incorporate chain-of-thought into both the planning and execution phases of the plan-of-technique method. This decision aligns closely with [**DG2**] and [**DG3**], which stress the importance of adaptively applying traditional text simplification techniques (**A1-4**) to cater to AR-specific needs.

We use an example to show how the chain-of-thought method supports the plan generation in the plan-of-technique method. To simplify the sentence, *Grab a pair of 10 to 12 lb (4.5 to 5.4 kg) dumbbells and lie on your back with your arms behind you and your legs extended and raised to a 45-degree angle*, we prompt GPT-3 to generate the thoughts about the input text's applicability to AR context. GPT-3 identifies the sentence as overly lengthy, containing more than three phrases, and thus includes syntactic simplification in its simplification plan. The plan involves three steps of syntactic simplification: 1. splitting the sentence at the first *and* because the length of the two joined clauses is too long; 2. splitting the sentence at the second *and* for the same reason. 3. Adjusting the passive voice in *your legs extended and raised* for better readability. After generating the plan, we continue to prompt GPT-3 to apply the simplification techniques outlined in the plan, yielding the result: *Grab a pair of 10 to 12 lb (4.5 to 5.4 kg) dumbbells. Lie on your back with your arms behind you. Extend your legs and raise them to a 45-degree angle.*

3.3.3 Error-Aware Model Calibration

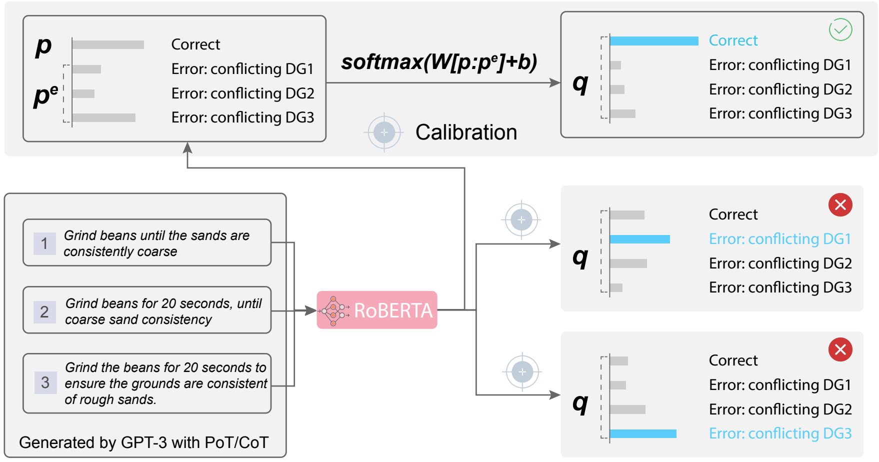

p
pe
Correct
Error: conflicting DG1
Error: conflicting DG2
Error: conflicting DG3
softmax(W[p:pe]+b)
q
Calibration
1 Grind beans until the sands are consistently coarse
2 Grind beans for 20 seconds, until coarse sand consistency
3 Grind the beans for 20 seconds to ensure the grounds are consistent of rough sands.
Generated by GPT-3 with PoT/CoT
RoBERTA

Figure 3.4: Error-aware model calibration: ARTiST prompts GPT-3 to generate a set of candidate results, which are subsequently analyzed by a RoBERTa-based error classification model (depicted in pink block) to detect any violations of design guidelines. The predicted scores of errors are calibrated with the affine matrix. Scores are adjusted using an affine matrix to ensure that the final selection is the output with the highest probability of correctness.

To align with [**DG1**] and prioritize meaning preservation in text simplification [15, 274], we propose an error-aware calibration method. Outputs from LLMs are often unstable and exhibit a bias toward certain answers due to the intrinsic bias of the LLMs and the influence of the prompt text, especially when applied to new tasks. Text simplification in AR has requirements that differ significantly from those of traditional NLP tasks, potentially exacerbating the impact of such intrinsic bias on LLM inference. For instance, LLM outputs may disproportionately reflect the influence of the last example in the prompt text, and within the context of our text simplification, the simplification techniques chosen may also be biased by the simplification techniques used in this last example [372]. To mitigate these issues, our error-aware calibration mechanism adjusts the output probabilities by applying an affine matrix [30], which is learned from a set of annotated datasets. This transformation does not directly rely on the prompt text and can alleviate LLM bias [272, 348, 349, 364]. Moreover, to strengthen LLM against common errors in AR, we enhance the annotated dataset with negative samples that violate [**DG1**] and risk altering the original meaning.

Shown in Figure 3.4, we use model calibration to stabilize language models in text generation [348, 349]. The affine transformation is defined as:

$$q = softmax(Wp + b), \tag{3.1}$$

where p refers to the probability of the generated simplified text, q is the calibrated probability, and W and b are learned parameters. We simplify this computation by treating W as a diagonal matrix following [91, 372]. The calibrated errors are identified from our open-ended exploration and expert interview. We then use the RoBERTA model to predict these errors by comparing the simplified text T^* and the original text T as, $p^e = p^e_1, p^e_2, ...p^e_m = f(T, T^*)$, where p^e_i is the probability of an error and m is the total number of errors. The errors include altering the meaning [**DG1**], producing text that is syntactically complex [**DG2**] and or too long [**DG3**]. Since access GPT-3's weights are not publicly accessible, we use RoBERTa instead to predict the error label p^e [175]. Therefore, we modify Equation 3.1 by incorporating p^e,

$$q = softmax(W[p;p^e] + b), \tag{3.2}$$

For each original text sample, we generate $n = 5$ simplified text samples and calibrate them. The final output is determined by the calibrated probability. The parameter values of W and b are learned from a small set of manually crafted data samples[349]. We first craft a set of gold-standard text simplification samples (64) $D = \{(T_1, Y_1, \hat{q}_1), (T_2, Y_2, \hat{q}_2),$
$\cdots, (T_k, Y_k, \hat{q}_k)\}$ where $(T, Y, \cap e)$ refers to the input text T, the simplified result Y, and whether the erroneous indicator $\hat{q}$. $\hat{q}$ indicates whether Y is correctly simplified from T and, if not, labels the error in Y. Since W and b have limited dimensions, we can learn the values of W and b through the gradient descent with the logistic

loss function $|q \cap e|^2$.

$$\begin{aligned} \mathcal{L} &= -\hat{q}log(q) - (1-\hat{q})log(1-q) \\ &= -\hat{q}log(softmax(Wp+b)) \\ &- (1-\hat{q})log(1-softmax(Wp+b)), \end{aligned} \tag{3.3}$$

where $\mathcal{L}$ is the loss function used to learn W and b.

3.3.4 Elaborative Simplification with Spatial Information

Following [**DG2**] and implementing elaborative simplification, we enrich the AR text by generating information on the spatial location of objects and object dimensions if they are presented in the original text. The object is detected and located with the Detic model, which runs on the backend server and provides the spatial information to LLM [381]. As shown in the expert interview, many objects may exist in the working environment, while only a subset of them can be useful. To align with [**DG2**], we require LLMs to select the objects that are mentioned for the first time in the text. Identified object locations are used to signify a spatial relationship to the user, adding a layer of contextual understanding that goes beyond identification. For example, the text *Then place the coffee mug with the dripper* can be elaborated with the coffee mug's detected position *on your right* to form the result *Then place the coffee mug on your right with the dripper.* The spatial location is determined before the user clicks *next step* and the elaborated content does not change during the execution of the step to avoid distracting the user. The Detic model may incur prediction errors and mismatches in object location due to user movement and latency. For instance, the user's movement can alter the object's location relative to the user, and the Detic model's results may continue to indicate the location of the object before the user's movement. We mitigate this issue by predicting only the spatial relationships between the object and the user (e.g., *the object is to the right of the user*). Therefore, the minor errors and latency in the Detic model do not significantly impact the final result. Furthermore, in our open-ended exploration, we observed that during the step transition, users typically do not engage in significant movement, thereby reducing the likelihood of potential mismatches. When displayed text includes a numerical measurement and an object with comparable dimensions is identified in the AR environment, the system automatically substitutes the numerical value with a description of the detected object.

3.3.5 System Implementation Details

We implement ARTiST's functionality using OpenAI's GPT-3 APIs and build a text simplification server with Flask. We run the Detic model on the server and incorporate its object detection result into the prompt text for GPT-3. The interface is developed using the PTGCTL architecture [304], with the HoloLens 2 component implemented in Unity.

3.4 Evaluation

Task	Step	Original	Simplified
Task 1.1	1	To create a coffee, first please carefully place the pour-over dripper over the coffee mug.	Place dripper (on your left) on coffee mug.
	7	Transfer the coffee grounds to the filter cone. Then place the coffee mug with the dripper on a digital scale and set it to zero.	Move grounds to filter cone. Set coffee mug with dripper on scale, zero it.
Task 1.2	2	Once the desk is clear, bring the power strip on the desk and connect the Charger to the power strip so the meeting attendants can use.	Put power strip on desk, connect phone charger to it.
	5	Next, place cups of water and papers on each chair. Each person should have one cup of water and paper;	Place water, paper onto desk in front of chairs.

Table 3.2: Four example system outputs in Study 1. The original, unmodified text (baseline) is in the third column; the last column shows the simplified condition with text output from ARTiST showing on the last column.

The formative study elicited that text-based AR guidance often creates a high cognitive load and that following AR guidance can be challenging due to the HMD's small display, low readability, and user error. Although our system attempted to address these limitations by integrating text simplification into AR, the actual effects on users' cognitive load, performance, and sense of usability required further exploration. To better understand these effects, we conducted two empirical studies. The first (Study 1) focuses on the overall cognitive cost of our system and its effect on performance over unmodified text, and the second (Study 2) focuses on a comparison against other AR text simplification methods and what can be learned from them. Tasks in both studies are everyday tasks that could benefit from AR task guidance [233]. Both studies comprise within-subject designs. Although subtasks in Study 1 have a between-subject component, our primary focus and point of investigation is the text condition. We investigate the following user study research questions:

[**UQ1**] In what ways does our proposed method impact cognitive load in AR?

[**UQ2**] In what ways does our proposed method affect task performance with text in AR?

[**UQ3**] How does our proposed method compare to other text simplification methods in AR?

While the first study focuses on **UQ1** and **UQ2**, the second explores **UQ3**. We pre-determined the study order so that half of the participants start with Study 1 and the other half with Study 2. Regardless of the order in which they engage the studies, participants are asked to review the study procedures and can only continue after giving their consent on the IRB-approved consent form.

3.4.1 Participants

Both studies involve 16 participants (average age 25, nine male and seven female). Half have previous experience using head-mounted AR and were recruited through electronic flyers and emails using snowball sampling.

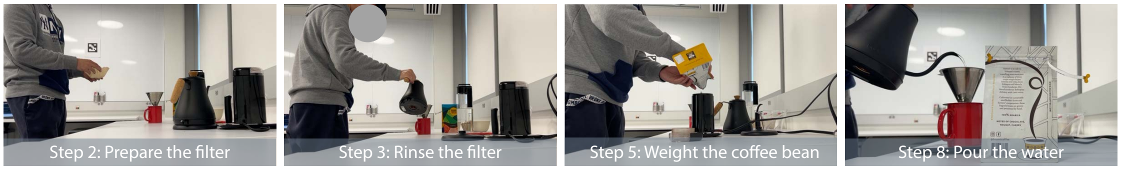


Figure 3.5: Task 1.1: Sample frames from user recording. The task requires participants to make pour-over coffee based on a nine-step online tutorial. The frames were sampled from steps 2, 3, 5, and 8.

3.4.2 Study 1: The Effect of Text Simplification on Guidance Tasks

We conducted an empirical study to evaluate the effect of textual simplification on users' cognitive load, performance, and other subjective ratings. We select two common physical tasks that benefit from AR guidance and collect data from real users. To avoid the learning effect while keeping task difficulty levels similar, both subtasks are physical activities that are performed in the same room (See Figures 3.5 and 3.6), have instructions of similar lengths, and do not require prior knowledge.

3.4.2.1 Experiment setup

We present the involved tasks and conditions in Study 1.

Task. The task contains two similar subtasks that have sequential instructions to guide users. In both subtasks, we display the AR text in a dark grey box to ensure visibility. We also adjust the font size to 9pt and have participants confirm that all text is legible. No single instruction is long enough to be cut off by the display. In terms of subtask assignment, we alternate the order of subtasks for each participant to balance the order effect.

- *Task 1.1: Pour-over Coffee.* This subtask contains nine step-by-step instructions that guide participants to make a pour-over coffee (Figure 3.5). The instructions are taken from an online tutorial on *how to make pour-over coffee.* [3] Participants need to read the text to complete the task.

[3] https://www.wikihow.com/Make-Pour-Over-Coffee

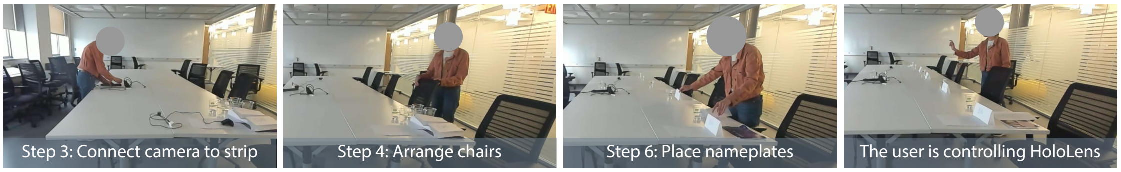

Step 3: Connect camera to strip
Step 4: Arrange chairs
Step 6: Place nameplates
The user is controlling HoloLens

Figure 3.6: Task 1.2: Sample frames from user recording. The task requires the participants to arrange objects in a meeting room based on a seven-step office menu. The frames were sampled from steps 3, 4, and 6.

- *Task 1.2: Meeting room preparation.* This subtask requires that participants follow AR instructions to arrange objects in a meeting room based on a seven-step office menu (Figure 3.6). The instructions are digitized from an online manual.

Conditions. This study has two conditions: a baseline condition that uses the original imported text and a simplified condition using ARTiST. Each participant will perform one subtask (either Task 1.1 or 1.2) with the baseline condition and the other subtask with the simplified condition. We use a pre-generated table to alternate the order of all trials so that each participant will perform tasks in different orders under both conditions. In total, all conditions and subtasks are evaluated an equal number of times. Samples from the simplified and baseline condition can be found in Table 3.2.

Apparatus. Participants wear a Microsoft HoloLens 2 and use hand gestures and voice commands to interact with the AR menu. These interactions are native to HoloLens 2, and the AR interactions comprise standard button tapping, translating, and spatial movement. Video and audio recording devices are set up to collect participants' feedback and qualitative data.

3.4.2.2 Procedure

The experimenters welcome the participants in a physical room; the physical tools necessary for task performance (e.g., coffee machine and ingredients) are present. To maintain ethical standards and comply with the IRB guidelines, each participant is given an informed consent form before the evaluation. Upon signing, each participant is paid an hourly rate of $20 and is fitted with the Microsoft HoloLens 2 headset. An ill-fitting HoloLens 2 can be detrimental to the AR experience, caus-

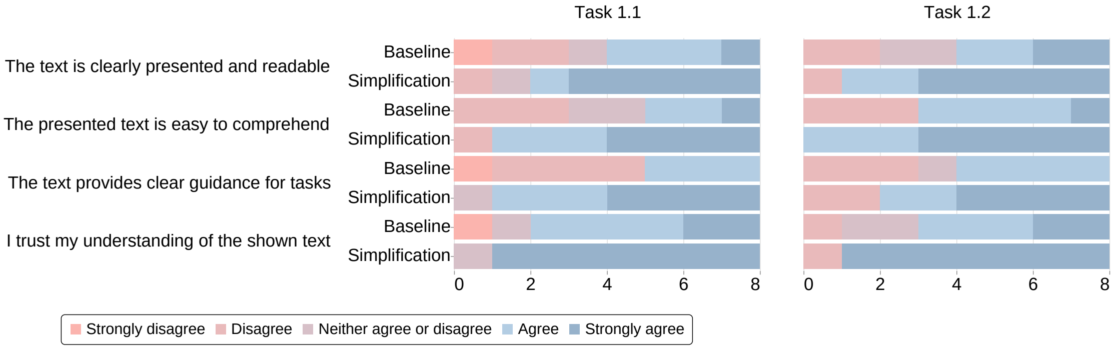


Figure 3.7: Study 1 results on a five-point Likert scale. Ratings are collected on a scale from “strongly disagree” to “strongly agree” in response to four questions assessing the readability, ease of comprehension, guidance, and trust in both simplified and baseline text versions. The horizontal bar graphs above visually represent the distribution of these ratings. The distribution reveals more positive responses for the simplified text across all questions and tasks.

ing blurry text. A series of initial calibrations are performed to ensure interface functionalities.

After all participants successfully interact with the AR interface, including its menus and buttons, using hand gestures, and indicate they can see the AR text clearly on the HMD, experimenters then explain the two subtasks and ask participants to practice thinking aloud. Meanwhile, video and audio recordings were set up before the trial began. Participants begin the study by air-tapping the AR button marked *Start* at the center of the HMD’s screen.

Once the task starts, step-by-step text instructions are automatically displayed in AR. Participants are not informed which condition they are using and are asked to think aloud while we observe and record the trials. Any anomalies or potential safety issues are continuously monitored by the experimenter. At the end of each subtask, we collect the participant’s subjective ratings on text readability, comprehensibility, guidance performance, trust, and cognitive load using a NASA TLX form. A semi-structured interview is conducted to better understand their experience.

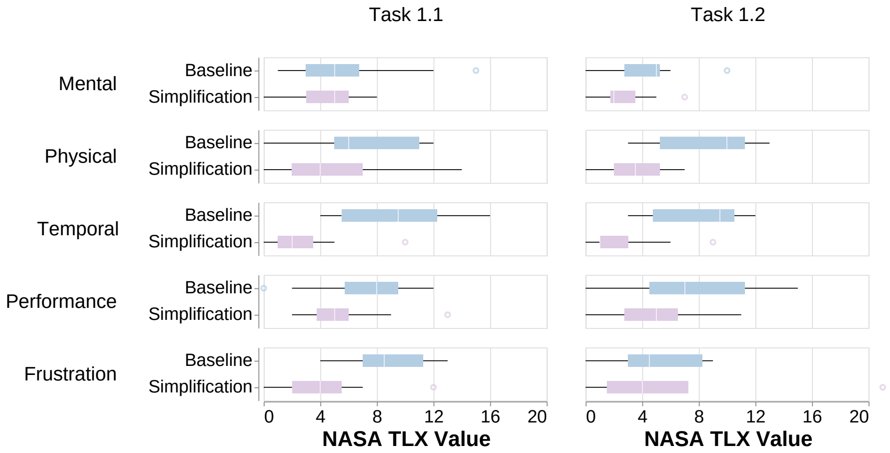


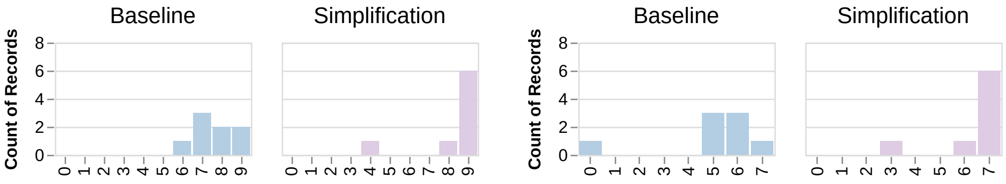
Figure 3.8: Study 1 results on NASA Task Load Index (TLX) values. The y-axis represents the different aspects of the NASA TLX, while the x-axis shows the TLX values. The simplified text significantly surpasses the baseline in reducing temporal demands and in enhancing performance and reducing frustration, demonstrating its advantages for overall task load.

3.4.2.3 Data collection

We collect quantitative data to measure performance. We specifically record the number of errors and the number of steps participants recall (i.e., memorability); in addition, we explore self-evaluated performance via subjective ratings. The experimenter counts the number of errors during participants' trials. Memorability is measured because one major challenge in AR guidance is that users only recall limited AR information during physical tasks; remembering steps reduces the need to split attention between AR and the task. Subjective ratings are inspired by the System Usability Scale (SUS), and we collect five-point Likert ratings on AR text readability, comprehensibility, guidance, and trust. We explain that trust reflects how confident the user is with their task performance.

Cognitive load is a primary user performance limitation in AR guidance tasks, and we use a NASA TLX 8(a)(b) form to measure it. Raw TLX scores are used and summative results are analyzed based on Hart's recommendations [94].

Experimenters also collect qualitative data via video and audio recordings of the study. Interview notes, think-aloud notes, and observations are also collected for later analysis. The sampled frames for the study can be found in Figures 3.5 and 3.6.



(a) Task 1.1 (b) Task 1.2

Figure 3.9: Number of recalled steps in the tasks. Study 1 results on the number of errors participants made while performing Tasks 1.1 and 1.2. The x-axis indicates the number of steps successfully recalled, while the y-axis shows the count of participants who recall that number of steps.

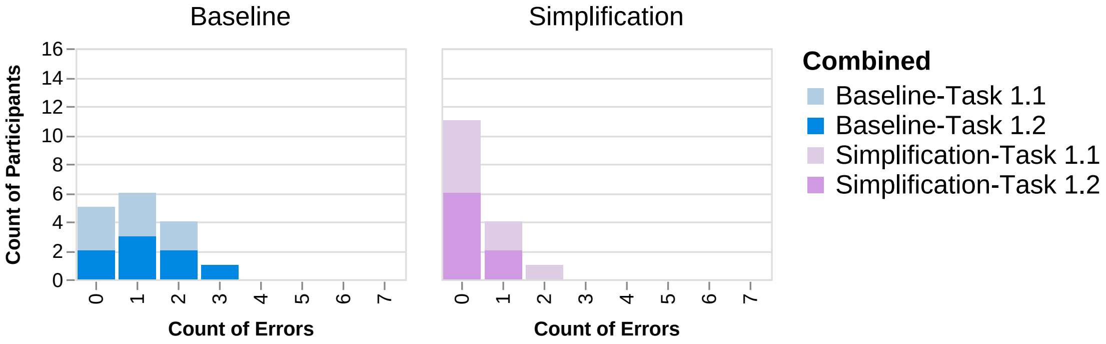


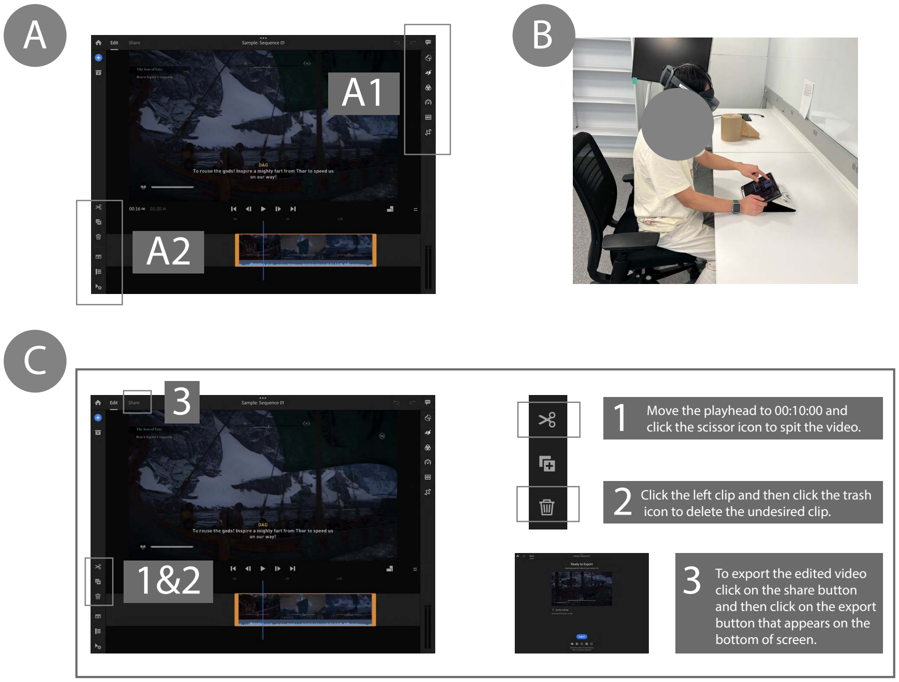
Figure 3.10: Study 1 results on the number of errors made in Tasks 1.1 and 1.2. The bar graph compares the error count between the baseline and simplified conditions, with the x-axis recording the number of errors and the y-axis depicting the number of participants who make those errors. The data illustrates that participants commit fewer errors when following the simplified text.

3.4.3 Study 1: Results and Discussion

3.4.3.1 Quantitative results

For Study 1, the subjective rating results are presented in Figure 3.7, and the NASA-TLX results are presented in Figure 3.8. The results for the number of recalled steps (memory) are shown in Figure 3.9, and the task error results are provided in Figure 3.10.

Using Mann-Whitney's U test, we assess differences among TLX scores, recall, and error data and use the Friedman test to assess differences among subjective ratings. These tests were chosen because the data are non-parametric. The TLX analysis shows that the simplified condition significantly reduces the overall cog-



1 Move the playhead to 00:10:00 and click the scissor icon to spit the video.

2 Click the left clip and then click the trash icon to delete the undesired clip.

3 To export the edited video click on the share button and then click on the export button that appears on the bottom of screen.

Figure 3.11: (A) Adobe Premiere Rush Interface: The interface showcases a video player positioned at the center of the screen, accompanied by a timeline below. The top-right section (A1) features buttons for graphics, effects, color, speed, audio, and transform functionalities. The bottom-left section (A2) contains buttons for editing tools and the project panel. (B) User Record: This section captures the user's interactions and activities during the Study 2 session. (C) Example Task Description: An illustration of a sample task description used in the study, providing users with instructions for completing a specific editing task.

nitive load for both subtasks ($U = 52,\ z = 2.84,\ p < 0.01$). Further evaluation of recall and error found no significant improvement in recall for the simplified condition over the baseline ($U = 85,\ z = 1.63,\ p = 0.10$), but showed the simplified condition significantly reduced the number of errors counted by users over the baseline ($U = 72.5,\ z = -2.09,\ p = 0.024$), see Figure 3.10. Test on subjective ratings indicated significant differences in all categories: readability ($\chi^2 = 10.71,\ p = 0.013$), comprehensibility ($\chi^2 = 15.00,\ p = 0.002$), guide ($\chi^2 = 10.71,\ p = 0.013$), and trust ($\chi^2 = 18.00,\ p = 0.001$).

3.4.3.2 Qualitative results

We coded the transcribed video and audio data along with notes from thinking aloud and observation. Codes sharing similarities were then grouped into themes to summarize analogous findings.

Spatial information can assist users. Participants acknowledged the benefits of spatial information in reducing cognitive load. Elaboration on objects' spatial location eliminates the need to search for them, reducing user effort. P11 reported feeling nervous when presented with multiple objects and new mentions of objects in the AR interface. P11 mentioned, "*Sometimes it is overwhelming to face many objects, and the location word (on your left) helps (you) find the object.*" The reduced effort and pressure were also confirmed by P1, who stated, "*Even though it won't save much time, the elaboration on the object eases my (sense of) pressure.*"

Text length and structural complexity affect participants' performance. Most participants report that shorter text is beneficial. Some participants reported that shorter text takes less time to process (P2, P5, P7) and felt it was "*easier to understand*" each step during a subtask when the text was shorter (P6-7). This is reflected in the TLX scores, as simplified conditions yielded better cognitive load scores than the original texts. Participants further report that using shorter sentences leads to better comprehension and confidence (P2, P10-12, P14). Multiple participants pointed out that they naturally "*skim*" text in AR, and stated that complex sentence structures lead to skipping important information and misunderstandings. More than half of the participants further stated that the simplified text improved their trust. When asked to explain their reasons for skimming text, screen resolution, screen size, and the urgency of completing physical tasks (impatience) while wearing a headset were identified. These observations reflect what experts from the formative study indicate.

Simplified text improves task guidance. Participants respond positively to breaking longer sentences into shorter ones (syntactic simplification). ARTiST divides long sentences into shorter ones by adding verbs (elaborative simplification). P5 said "*Shorter sentences with clear actionable directions make it easy to know what to do,*" while P6 said, "*It is more convenient to follow smaller step instructions.*" The participants' positive feedback reflects the benefits introduced by the design guidelines proposed earlier.

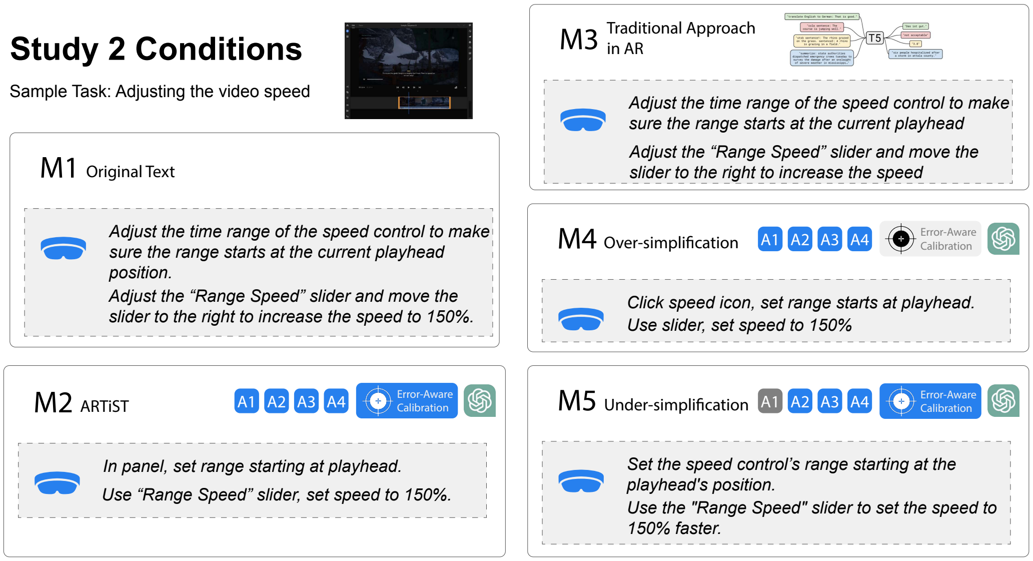
3.4.3.3 Discussion

Our results verify that ARTiST significantly improves cognitive load (**RQ2**) and reduces task performance errors, generating significantly higher subjective ratings (**RQ1**). Shortened sentences and syntactic simplification contributed to decreased cognitive load, as participants indicate that the simplified text is easier to read. Additionally, shorter sentences enable participants to quickly skim the text to grasp core concepts, which may also play a part in reducing cognitive load. As we mentioned earlier, reducing cognitive load could help to improve the usability of AR guidance and have a positive effect on users' safety.

ARTiST significantly improves subjective ratings on all four metrics. However, the system yields no significant change in memorability. Both the baseline and simplified conditions reached a fairly high recall count. A possible explanation for this is the fact that all tasks are physical tasks, and participants may rely on their performance more than the text for recall. Yet the simplified text resulted in fewer user errors than the unmodified text, suggesting that the system successfully retains critical information for tasks. Overall, participants felt better guided by the simplified instructions and more confident (i.e., trust), signaling a positive effect on their overall performances.

3.4.4 Study 2: Comparing Text Simplification Methods

In the previous study, we evaluated ARTiST against unmodified AR text. The goal of this study is to further understand how ARTiST's process compares with other methods for text simplification. However, almost all currently used methods are not tailored for AR. As such, we selectively integrated these methods into the AR context while keeping their traditional functionalities. This study is a within-subject study that includes five different methods with one task. The study recruited the same set of participants ($N = 16$). In addition to the HoloLens 2 used in Study 1, this study makes use of an iPad as an additional apparatus for task performance.



Adjust the time range of the speed control to make sure the range starts at the current playhead position.
Adjust the "Range Speed" slider and move the slider to the right to increase the speed to 150%.

In panel, set range starting at playhead.
Use "Range Speed" slider, set speed to 150%.

Adjust the time range of the speed control to make sure the range starts at the current playhead
Adjust the "Range Speed" slider and move the slider to the right to increase the speed

Click speed icon, set range starts at playhead.
Use slider, set speed to 150%

Set the speed control's range starting at the playhead's position.
Use the "Range Speed" slider to set the speed to 150% faster.

Figure 3.12: All five conditions for Study 2. M1 is the original text. M2 is the ARTiST condition. M3 is the state-of-the-art T-5 model applied to AR. M4 is ARTiST without engaging error-aware calibration (over-simplification). M5 is ARTiST without engaging content reduction(under-simplification). The text in the method's grey box represents the text after simplification. A1, A2, A3, A4, and error-aware calibration legends denote whether any of these components are used for the condition.

3.4.4.1 Experiment setup

We present the tasks and conditions of Study 2 in this section.

Task. The task is designed to have participants wear a HoloLens 2 while also using an iPad. AR instructions for video editing are displayed on the HoloLens 2. To minimize the learning effect and for repeated trials, we employ subtasks with similar interactions and difficulty levels but different content. We adopt a series of video editing jobs from Adobe Premiere Rush's official tutorial to AR[4] to test each method. They involve interaction primitives such as selection, pan, and translation. The task contains five subtasks, including video clipping (S1), speed control (S2), graphics overlay (S3), video filter application (S4), and aspect ratio adjustment (S5). These subtasks are chosen because they have similar interaction difficulty but require diverse types of interactions (e.g., tap, drag, and pinch). Each

[4] Adobe Premiere Rush. `https://helpx.adobe.com/premiere-rush/tutorials.html`

subtask has three steps and takes about three minutes to complete based on our preliminary testing. Adobe Premiere is installed on the iPad participants use to perform the video editing. An example can be found in Figure 3.11.

Conditions. To understand how our system differs from other simplification methods and investigate the implications for user performance, we explore five conditions, which are shown in Figure 3.12. Beyond making a comparison to the unmodified text (i.e., baseline), we also compare against the state-of-the-art T-5 text simplification model [237]. Further, because our formative study revealed that sentence complexity and grammar structure (text length) have a foremost effect on text reading in AR (which was also indicated by experts in the formative study), we also explore an over-simplification and an under-simplification condition in this study. These two simplification methods represent different levels of length modification relative to the original sentence. Over-simplification is achieved by removing the error-aware calibration step for maximum simplification at the cost of factual information. Under-simplification is achieved by removing the content reduction technique in ARTiST. We describe the five different conditions below:

- M1: Original text
- M2: ARTiST's approach;
- M3: Traditional state-of-the-art text simplification with T-5 [237] fine-tuned on WikiAuto dataset [117]
- M4: Over-simplification without error correction (i.e., does not force factuality, see 3.2.4 for details)
- M5: Under-simplification without content reduction (**A1**)

We use a pre-generated table to balance the learning and ordering effect for the five conditions across the five subtasks. We assign one condition to one of the subtasks to form pairs, and each pair includes three steps (i.e., three trials). For example, the pair M2-S1 stands for the M2 condition used in subtask 1. The pre-generated table ensures that each participant performs these pairs in a unique order. Each participant performs five condition-subtask pairs or 15 trials, for a total of 240 trials. Overall, all subtasks and conditions are evaluated an equal number of times.

3.4.4.2 Procedure

In the beginning of this study, participants are asked to wear the HoloLens while sitting and holding an iPad. Experimenters explain that their task entails editing videos that are displayed on the iPad. The videos are 30-second clips of stock footage, and participants are told they will use the onboard video editing tool on the iPad to seek, crop, change filter, and change the aspect ratio. After a short warm-up period to familiarize participants with iPad functionality and fit the HoloLens, we confirmed that participants can read the AR text clearly, similar to what we did in Study 1. During task performance, condition-specific AR text is displayed to the participants; they are asked to follow the text to perform the editing task. Experimenters count the number of errors made during the trials, and participants are asked to think aloud as they engage in their tasks. At the end of each condition, the experimenter collects recall data, subjective ratings, and TLX scores. A semi-structured interview is conducted at the end of the study to understand the participants' overall impression of each of the five conditions. We collect both quantitative and qualitative data in a similar way to Study 1.

3.4.5 Study 2: Results and Discussion

3.4.5.1 Quantitative results

For Study 2, the subjective rating results are presented in Figure 3.13, and the NASA-TLX results are presented in Figure 3.14. The results for the number of recalled steps (memory) and task errors are shown in Figure 3.15.

For non-parametric data, we used the independent-sample Kruskal-Wallis' test with repeated measures for performance metrics (error, memory recall, and subjective rating) and for cognitive load with Dunn's Test as post-hoc analysis with Bonferroni correction for multiple tests. For error analysis, we found an overall significant effect ($H(4) = 17.189, p = 0.001$), with post-hoc analysis showing that M2 ($p = 0.014$), M3 ($p = 0.014$), M4 ($p = 0.014$), and M5 ($p = 0.014$) reduced errors significantly compared to the baseline. We found that there is an overall difference across the conditions ($H(4) = 12.572, p = 0.014$) in terms of participants' ability to recall the instruction steps. Post-hoc analysis revealed a significant difference between the original text M1 and ARTiST ($p = 0.025$). No differences are found be-

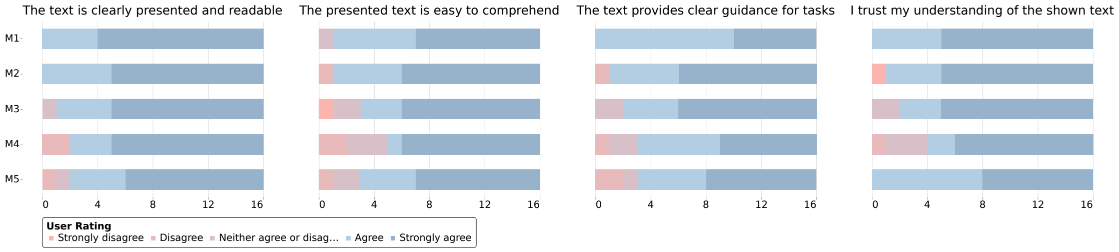


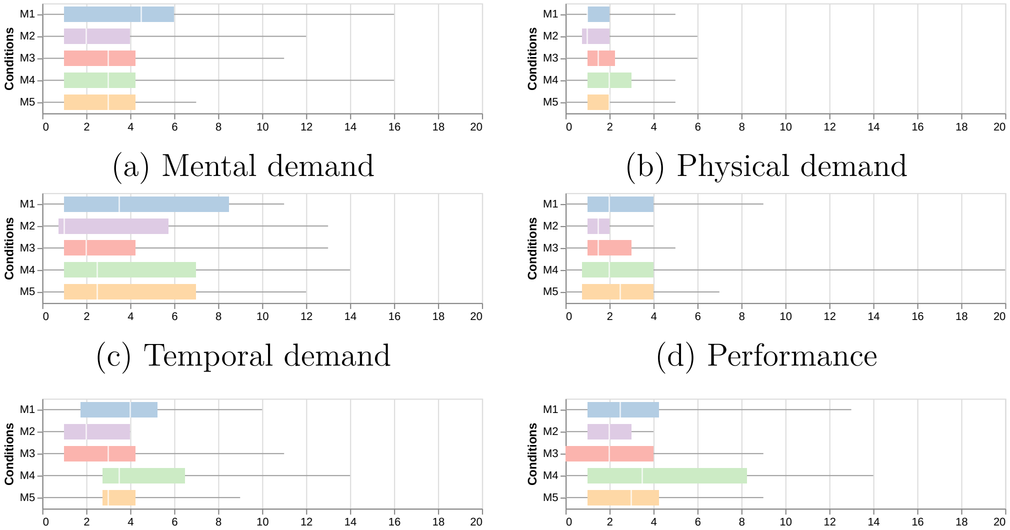
Figure 3.13: Study 2 results on the subjective Likert scale. Ratings were collected on a scale from "strongly disagree" to "strongly agree" in response to four questions assessing the readability, ease of comprehension, guidance, and trust in both simplified and baseline text versions. The horizontal bar graphs represent the distribution of these ratings, and the results for the four questions are laid out horizontally.

tween the original text M1 and M3 ($p = 0.179$), M4 ($p = 0.319$), and M5 ($p > 1.00$). No differences are found between the four simplified conditions (M1-4). As for TLX scores, we found that M2 significantly reduced overall cognitive load compared to M1 ($p = 0.043$): for a detailed breakdown refer to Figure 3.14. There are no significant differences among the five conditions in readability ($H(4) = 0.934, p = 0.934$), comprehensibility ($H(4) = 0.389, p = 0.983$), guidance ($H(4) = 2.444, p = 0.655$) and trust ($H(4) = 1.530, p = 0.821$); See Figure 3.13 for details.

3.4.5.2 Qualitative results

We identify a series of qualitative findings based on the quantitative metrics and the coded recordings.

The level of simplification has a mixed effect on error rates and subjective ratings. While participants reported that they could understand any of the simplified texts (M2-5) better than the unmodified text (M1), we noticed that there is no uniform effect on the level of simplification relative to task errors (P11-12, P3, P5). While several participants experienced increased errors due to over-simplified text omitting important information (P3), others made mistakes due to verbose text that was not simplified enough (P11-12, P5), causing them to overlook important information. This also aligns with the fact that the M1 condition yielded the most user errors. One of the behaviors observed is that the longer the sentences, the less patience a participant appears to have and the faster they skim.



(a) Mental demand

(b) Physical demand

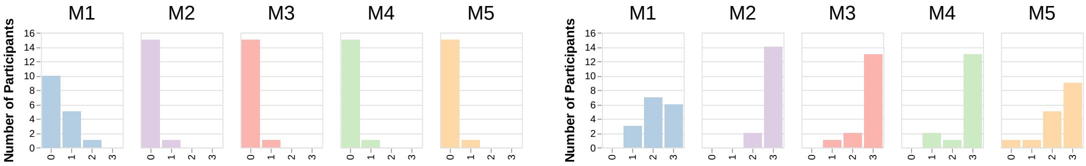
(c) Temporal demand

(d) Performance

(c) Effort

(d) Frustration

Figure 3.14: Study 2 results on NASA Task Load Index (TLX) values. The y-axis represents the different aspects of the NASA TLX, while the x-axis shows the TLX values. The results indicate that condition M2 significantly outperforms other conditions in terms of the user's effort.



(a) Error

(b) Memorized Steps

Figure 3.15: Study 2 results for error count and memorability evaluation. In panel (a), the x-axis represents the number of errors made by participants, while the y-axis shows the count of participants corresponding to each error count. The results indicate that methods M2, M3, M4, and M5 are effective at reducing the number of errors, underscoring the advantages of text simplification in enhancing task success. In panel (b), the x-axis displays the number of task steps correctly recalled after the task, and the y-axis shows the number of participants. This panel demonstrates the impact of the simplification process on the participants' abilities to recall information, with conditions M2, M3, and M4 showing an improved number of steps recalled over M1.

Often these behaviors lead to missing details while carrying out the task, such as when P11 and P12 adjusted the wrong button during task performance. Similarly, both over- and under-simplification methods affected participants' sense of readability and memorability in different ways. When asked about their experience with the M5 condition, P6 reported that "*the simplified one uses the more understandable words,*" but P10 mentioned that the texts are "*not simplified enough and can be thrown away.*" P7 also reported that M5 increased the number of previous steps they could recall, as M5 presents "*clear and memorable instructions.*"

The effect of different text simplification methods on cognitive load. Participants reported during the interview that simplified texts (M2-5) have a lower cognitive load. While reading unmodified text became tedious during task guidance (P2, P9-10, P14), the simplified text could be less so (P2). However, participants reported that over-simplified text (M4) increases cognitive load, as important information is often removed resulting in extra processing time needed (P1, P10). (P1, P10). Moreover, participants indicate that they believe they perform better with the ARTiST condition: As P7 mentioned, "*I am pretty sure I successfully completed the steps,*" while P11 said, "*I feel it increased my performance.*"

3.4.5.3 Discussion

In exploring **UQ3**, we found ARTiST impacted TLX ratings as it was the only condition that significantly reduced cognitive load for participants. Figure 3.14 shows that TLX variance is much lower for performance with ARTiST, which is in line with our observation that most participants show stable performance with the ARTiST condition.

All four simplified text conditions (M2-5) significantly reduce error rate, but do not necessarily increase recall. This finding reflects our [**DG3**], which addresses the importance of text length in AR. Regardless of the level of simplification, all four conditions shortened the text in some way. The results indicate that only ARTiST significantly improved recall while reducing error rate. This indicates that ARTiST helps users to improve performance in short-span tasks like video editing. In addition, the current state-of-the-art text simplification (M3) does not reduce high cognitive load nor improve memory for AR readers, while ARTiST improved on both. This suggests that the direct application of text simplification to AR might not be optimal and is in line with the results from the formative study.

ARTiST is also the only condition that significantly reduced the TLX scores (**UQ2**). Sentence length and structure may play an important part in reduced cognitive load as participants noted reduced processing time and more ready comprehension. This reduction addresses the concerns (high cognitive load) brought up by experts during our formative study in Sec. 3.2.3.

Finally, both over- and under-simplification conditions received mixed feedback from participants. This could be linked to their personal reading habits when wearing an HMD. We observed that participants who comment positively on the over-simplified condition are typically impatient readers when they have the HoloLens 2 on. Others, however, complained that the over-simplified condition does not provide enough detail or is missing critical information, creating obstacles to task completion. Our qualitative results showed that participants who took extra effort going after missing details scored higher in their TLX ratings. These findings reflect the results from the formative study that both text length and meaning preservation are important.

3.5 Final Considerations

In conclusion, this chapter presents ARTiST, an automated text simplification system tailored for head-mounted AR devices. We first identify the challenges in AR text presentation via a formative study that includes a survey of the literature, an open-ended exploration with seven participants, and interviews with three experts. The findings lead to design guidelines that help form the ARTiST system. The system leverages OpenAI's GPT-3 models through few-shot learning for automated text simplification. Using chain-of-thought prompting, we present two novel techniques tailored for AR text simplification: a plan-of-technique and error-aware calibration to ensure meaning preservation. We validate our system via a 16-participant empirical study, resulting in significant improvements in users' performance, reduced cognitive load, and better subjective ratings when compared to unmodified text, the state-of-the-art T-5 language model, and other methods. These findings underscore the efficacy of our system in enhancing text readability and mitigating cognitive load during task guidance in AR environments.

Chapter 4

Satori: Towards Proactive AR Assistant with Belief-Desire-Intention User Modeling

4.1 Introduction

Satori, a ghost-like deity from Japan, is fabled to read human minds and respond to thoughts before they unfold into action. While such supernatural power once belonged strictly to the realm of folklore, modern AI technologies are now beginning to emulate a similar ability to predict human intent and actions and even provide proactive assistance during task interactions [137]. Such *proactive* virtual or digital assistance, which determines optimal *content* and *timing* without explicit user commands, is gaining traction for its ability to enhance productivity and streamline workflow by anticipating user needs from context and past interactions [200]. However, there is currently limited research on how to best design and implement such systems.

Most current assistance in AR remains *reactive*, responding to user commands or environmental triggers without the capacity for *active* engagement. These systems require that users initiate interactions, which is inefficient in AR where users typically have limited attention to spare. In response to this, some AR assistance

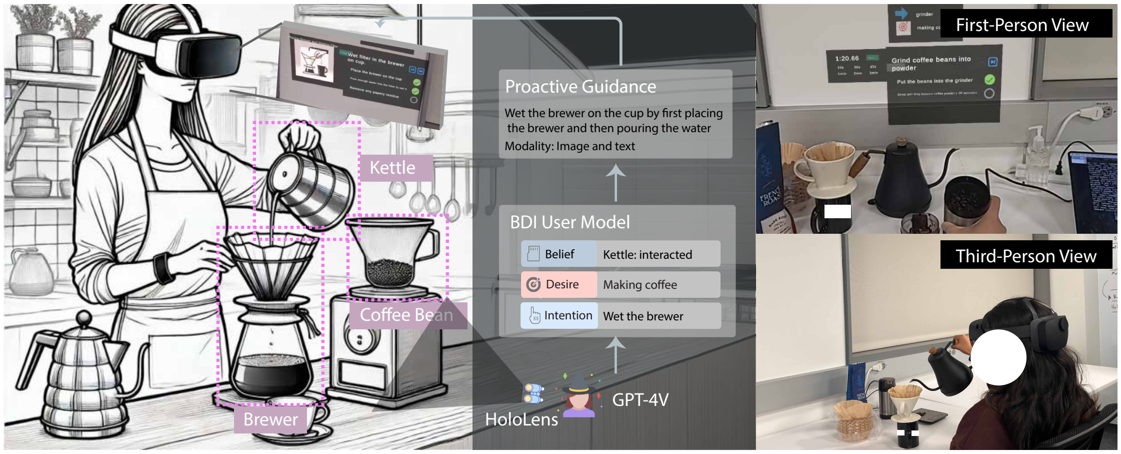


Figure 4.1: Satori is a mind-reading monkey-shaped creature in Japanese folklore. Our system extends this concept to highlight the importance of incorporating the user's state (i.e., knowledge and intentions) while building proactive AR assistants. The Satori system combines the tracked objects, the surrounding environment, task goals, and user actions with a large-language model (LLM) model to provide AR assistance to the user's immediate needs. This kind of **proactive AR assistance** is achieved by implementing the Belief-Desire-and-Intention (BDI) psychological model with advice from two formative studies with a total of 12 experts. The *belief* component reflects whether the users know where the task object is, and how to perform certain tasks (e.g., task goals, high-level knowledge); the *desire* component is the **actionable goal**; and the *intention* component is the **immediate next step** needed to complete the actionable goal.

incorporates proactive elements; for instance, they may provide maintenance guidance based on recognized objects or components [144, 198, 298]. Yet, these systems are often built on fixed rules and lack adaptability and reusability. They are limited in responding effectively to the user's surrounding environment or interpreting their actions over time. As a result, these systems struggle to guide users across multiple, consecutive steps and instead tend to function as discrete task-only assistance.

Designing proactive assistance for AR is particularly challenging due to the necessity of understanding the user's state, short-term goals, and surrounding environment. Further, timely assistance is crucial due to constraints on user attention. Providing assistance too early, too late, or simply too frequently can increase cognitive load and negatively impact the user's experience [14, 300].

This chapter addresses these gaps by first identifying the in-depth benefits, and challenges of designing a proactive AR assistance by conducting two formative studies and then exploring the design of a system through Satori. The first study with six professional AR designers revealed several design challenges such as: 1) limited generalizability and reusability of current non-proactive AR assistance, 2) difficulties in accurately detecting user intentions, and 3) the need to balance general advice with task-specific solutions. The professionals recognized that using proactive AR assistance could potentially improve scalability and efficiency, but also highlighted the technical challenges related to accurately tracking and understanding users' actions.

Building on the findings from the first study, the second formative study engaged six experts—three human-computer interaction (HCI) researchers and three psychology researchers—in dyadic interviews to explore design strategies for more proactive AR assistance. The design sessions found four key design considerations: 1) understanding human actions; 2) recognizing surrounding objects and tools; 3) assessing the current task; and 4) anticipating immediate next steps. Following experts' suggestions, these findings were later integrated with the well-established belief-desire-intention (BDI) model [26, 49, 84, 158], resulting in an AR-specific adaptation that guided the development of our system, Satori.

To adapt the BDI model for AR assistance, Satori needs to address the challenges brought up in the formative studies and account for the limitations of the

AR headset. Inspired by the theory underpinning the BDI model, we build Satori using an ensemble of egocentric vision models combined with a multimodal large language model (LLM) to determine timing, content, and user action in everyday AR assistance. The system is a multi-modal proactive assistance wherein the user's environment, nearby physical objects, action history, and task goals are input to predictively determine the content and timing of the assistance. Our approach ensures that the AR assistance delivers relevant information at appropriate moments, enabling a new and more seamless experience for AR users.An overview of the workflow is presented in Figure 4.1.

We evaluated Satori over four everyday AR tasks and compared it to a Wizard-of-Oz system (i.e., baseline) designed by six professional AR designers. We found that Satori's proactive guidance was as effective, useful, and comprehensible as the AR assistance created by the designers. User ratings also indicated that Satori's timing prediction performs similarly to the baseline. Additionally, Satori's guidance allowed participants to switch between tasks without the need for pre-training or scanning. Our findings suggest that our application of the BDI model not only successfully understood users' goals and actions but also captured the semantic context of given tasks, reducing the need to craft AR assistance for every specific scenario and improving its generalizability and reusability.

To summarize, the contributions include:

1. Identifying benefits, challenges, and design requirements for creating a proactive AR assistance, derived from two formative studies with twelve experts and applied using concepts from the BDI model in AR environments.

2. Design and implementation of Satori, a proactive AR assistance system applied with BDI's concepts that combine LLM with a series of vision models to infer users' current tasks and actions, providing appropriately timed step-by-step assistance with dynamically updated content.

3. A 16-user empirical study shows that Satori delivers performance comparable to designer-created AR assistance in terms of timing, comprehensibility, usefulness, and efficacy.

4.2 Formative Study 1: Design with Professional AR Designers

We first conduct a formative study to explore the problem space and potential benefits of proactive AR assistance. The study begins with a semi-structured interview on participants' background knowledge, followed by designing four different common AR interaction scenarios. A final apparatus combining participants' design feedback is created for later study.

4.2.1 Participants

Using email and snowball sampling, we recruited six professional AR designers (three female and three male, age: $\bar{x} = 30$). As we wanted to collect insights from experienced individuals, all participants selected were professionals with at least three years of experience working on developing AR applications. Participants were paid $30 per hour.

4.2.2 Tasks

The study was conducted in two sessions: a semi-structured interview and a design session for four different everyday AR scenarios with assistance. Each participant was asked to design two out of the four scenarios for a balanced scenario distribution. Each scenario was designed by three different AR designers.

In the first session, we collected participants' prior working experience using AR assistants, the challenges they faced in creating them, and their assessment of the assistants' potential benefits and applications. Additionally, we discussed the concept of proactive AR assistance with participants and collected their insights on potential benefits and use scenarios. In the second session, participants were asked to design AR assistants for two everyday scenarios out of the four. These two scenarios were assigned in a pre-determined order to balance the total number of designs. We use WikiHow [1] to obtain detailed, step-by-step instructions as the **task background information** for participants. These instructions ($averagesteps : \bar{x} = 7$) provide the framework to make guidance, and participants

[1] https://www.wikihow.com/

can elaborate (e.g., adding additional steps) at their will. Aside from the text instructions, we recorded videos in first-person view using the original instructions to provide a visual reference and interaction context for participants. Given instructions, images, and videos depicting the scenarios, participants were asked to design: 1) if a piece of guidance is needed for a particular step; 2) when the guidance should appear and for how long; 3) the modality of the guidance; 4) the content of the guidance. The above questions focus on the questions of "if", "when", "how", and "what" in AR assistance, which is a common architecture for guiding users in the literature and current practice [173].

4.2.3 Procedure

Since the AR designers reside in different time zones, the experiment was conducted remotely via Zoom after obtaining their informed consent. Participants were asked to introduce their background, describe their daily work, and discuss their projects related to AR assistance. We further inquired about their insights into the advantages and disadvantages of AR assistance, including challenges faced during development and challenges faced by end users. Finally, we presented the concept of proactive AR assistance and solicited their opinions on potential challenges and applications, as well as feasibility.

After the semi-structured interviews, participants received digital forms containing materials to design AR assistance for their assigned tasks, including textual descriptions and contextual images and videos. During this phase, participants were introduced to the interface and how to use its operations to, for example, create interaction prompts for a step/sub-step or select what information the user should be presented within what modality. The experimenters addressed any questions participants raised via Zoom.

On average, the study's first session lasted approximately 28 minutes ($\bar{x} = 28$), while the second session took around 60 minutes ($\bar{x} = 60$), totaling around 90 minutes. All participants successfully completed the design task. Since every scenario was designed twice by two participants, the final AR assistance design was merged together by the experimenters based on common modalities and a union of participant-generated instructions. Inconsistencies were resolved through discussion.

4.2.4 Results

4.2.4.1 Benefits in conventional AR assistance

AR assistance is beneficial in providing real-time, contextual information that improves user awareness.. Such guidance has ability to reveal forgotten or overlooked information. For instance, P1 emphasized that "*I find AR assistance most useful when it helps the user realize something they might not know... they might forget about an object, or not be aware that this object could be used in this situation... then (with AR assistance) they have this eureka moment.*"

AR assistance is also typically intuitive for users to follow, which reduces interaction cost and supports decision-making. . P2 and P3 highlighted that by overlaying visual cues such as arrows or animations directly onto the environment, AR could help users quickly comprehend otherwise difficult tasks such as examining electrical circuits. P3 stated that "*in tasks with spatially sensitive movements... AR is a proper medium because users intuitively understand what they need to do.*" P3 further explained that users who received spatially directed AR guidance for operating a machine (e.g., turning knobs or pressing buttons) found it more intuitive than 2D instruction books or manuals. Additionally, P4 brought up that being able to provide spatial guidance reduces interaction costs for tasks that require frequent operations, simplifying users' decision-making process.

4.2.4.2 Challenges in conventional AR assistance

Pre-designed AR assistance is hard to scale to diverse contexts.. AR designers often create designs based on their assumptions about the user's environment. However, users may interact with objects that fall outside these initial assumptions. As P1 noted, "*It's hard to cover all the edge cases of what a person might have... I assume they're in an indoor space, but that might not be the case,*" highlighting the complexity of accommodating varied environments.

AR assistance lacked an interaction standard.. P5 noted that there is not a standardized approach in the expansive interaction design space, especially when compared to traditional 2D interaction. P3 expressed that creating 3D visual assets from scratch was usually complicated.

Predicting action timing and user intention remains challenging. . Both P3 and P4 noted the difficulty in defining an accurate mapping between user actions and AR responses. P4 emphasized that misinterpreting user behavior can result in irrelevant or unhelpful guidance (e.g., recommending a taxi when the user intends to walk). P3 also emphasized the difficulty faced by task experts who do not have engineering expertise, stating, "*Suppose I am a designer and I know nothing about coding, but I still want to make AR assistance for users. How should I do that?*"

4.2.4.3 Benefits of Proactive AR Assistance

Proactive AR assistance is automatic without needing user input.. During the later part of the interview, participants envisioned the potential benefits of applying for proactive AR assistance on common tasks, from both the AR developers' and users' points of view. Three participants described automatic AR assistance as **proactive assistance** as P4 pointed out that such assistance anticipates the user's intentions and actively provides guidance based on the user's surrounding environment. **Reduces development time and increases efficiency.**. Half of the participants (P1, P2, and P6) agreed that proactive assistance could tremendously reduce development time on similar AR assets, animations, and programming logic (e.g., a panel shows up when a user touches an object). For instance, P2 remarked, "*We will definitely see a huge improvement in the efficiency of the content creation through this auto-generation process.*" P1 said that automatic assist can simplify the repetitive design process in "*adding labels, recognizing objects, and generating guidance*". She continued to offer an example of a cooking app where such automation would be particularly useful in identifying ingredients or suggesting cooking steps.

Improve scalability.. Both P1 and P3 highlighted how automatic AR assistance could generalize across different domains. According to P1, "*If we have a pipeline... using computer vision, it would save a lot of time... could have a universal pipeline to create guidance.*" Moreover, P3 pointed out that such assistance may be adapted as authoring tools like spatial programming and program-by-demonstration, increasing the accessibility for non-developer users.

Reducing information overload.. Participants (P3, P5) pointed out that proactive assistance could automatically detect the user's intention during AR interaction, presenting live-updated information in need, thus reducing information overload. It may also gain trust from users since the proactive assistance might make users believe that the system understands their intentions well.

4.2.4.4 Challenges of Proactive AR Assistance

Cross-domain scalability is difficult.. P1 raised concern over the feasibility of a universal system that could operate across different devices and domains. P3 further added that scalability remains a primary hurdle even for the most experienced AR designers because domain-specific knowledge is usually required to provide effective guidance. "*Scalability is the main issue... AR systems must lie in a specific domain, and it's hard to do this for every domain.*" P6 brought up the fact that proactive assistance must be able to adapt to even unforeseen circumstances, which requires a deep understanding of the task at hand. Even with the help of LLMs, further training and customization of the tasks have been necessary, as LLMs are generally not domain-specific.

Detecting user intention is a primary challenge, as errors lead to confusion.. Four participants (P2, P4, P5, and P6) emphasized the difficulty of accurately detecting users' intentions in AR. P5 brought up the limited field of view (FoV) in AR headsets and the low accuracy of detection algorithms as two main issues, although the former (limited FoV) might be among the causes of the latter (low accuracy). P5 commented that "*...sometimes, the system might trigger guidance when the user doesn't need it, which could lead to confusion...*" Similarly, P4 discussed how AR software in the industry has struggled to fully apprehend complex user environments and actions, causing confusion. This view is also shared by P2, who mentioned that proactive assistance might confuse users if it lacks self-explanatory features. P2 stated that "*if (the system is) fully automatic, you need the system to have some type of feedback. Automation without feedback may confuse the user.*"

Adapting the AR instruction to users' active duties is challenging.. P6 stressed that a proactive system should automatically adjust general advice to task- (and) environment-specific solutions. AR systems must remain relevant to the user's current goal, offering guidance that is actionable and appropriate.

4.2.4.5 Design results for four common scenarios

All participants used user-centered and object-centered strategies to determine when assistance should appear. Participants using the user-centered strategy focused on actions by, for example, *showing an instruction when the user got stuck on a step* or was about to get stuck. They also created instructions to indicate the user's completion of a step or unexpected situations. Participants who were conversely focused on object-centered strategies designed AR assistance that appeared in response to objects of interest. For example, one participant designed a reminder to *change the mop pad* when *the old pad is dirty*.

Participants' designs comprised multiple modalities, such as text, visuals, audio, and sometimes even tools (e.g., a timer). Notably, they tended to combine modality ("how") with specific contents ("what"), see Table 4.1. While most participants chose to use text-based assistance to provide an overview of step-by-step instructions, information about the object, or reminders, they also designed three types of visuals: overlays (e.g., arrows, progress indicator, checkpoint cue), images, and animations. In addition, audio was repeatedly used to sound a warning, pronounce guidance, or indicate completion.

4.2.4.6 Wizard-of-Oz system

Each participant created two AR assistance designs for two distinct tasks, resulting in a total of 12 designs for four tasks. These designs were later combined into a Wizard-of-Oz (WoZ) system. The system contains in-situ image, voice, and text-based AR assistance displays. We combined similar timing, modality, and content to form one AR assistance per task. Images were sourced from task instructions on WikiHow, and text and voice guidance were developed by combining participant designs and WikiHow instructions. We then implemented the four AR assistance architectures in Unity and employed WoZ to trigger the assistance on time and accurately via a wireless keyboard controlled by a human experimenter. To visualize instructions, we overlaid them directly on static images to indicate where the interaction should happen, how many materials should be used, etc. Animations

Modality	Detailed Assistance Type	Content
text	text	overview; instruction information; reminder
visuals	animations	instruction
	image	instruction
	arrows	location; interaction point
	checkpoint cue	step completion; warning
audio	sound cue	step completion; warning
	voice	instruction
tools	timer	count time

Table 4.1: Types of assistance provided across different modalities suggested by expert AR designers. The overlays are used to indicate locations or to indicate how to interact with apparatus in the scene; a progress indicator reflects how far the user is into the task. The image and animation are designed to illustrate actions and positions and show "how" to complete the current step. The checkpoint cues, according to participants, are used to indicate step completion. The timer counts time for time-sensitive steps, such as making pour-over coffee.

were achieved by looping multiple image sequences, similar to a GIF animation. The resulting system was video-recorded over Microsoft HoloLens and sent back to participants for recognition. All agreed with how each step was implemented after discrepancies were resolved either through clarification or modification of the apparatus.

4.3 Formative Study 2: Co-Design with Psychological and HCI Experts

Building on the previous formative study, the second formative study sought to gain insights into the design of a proactive AR system by consulting experts. We recruited six experts, three from computer science and three from psychology (E1-6). The study focused on **how to design** the system and **the probable methods** for executing said design by discussing critical factors, interaction flows, and system architectures via two dyadic interviews. We paired experts with complementary backgrounds to form three groups (Groups A, B, and C) as Table 4.2 shows. Their ideas and designs motivated later system implementations.

4.3.1 Dyadic Interviews

During the dyadic interviews, each pair of participants worked together to respond to open-ended questions and goals [202]. The first interview incorporated *participatory design* to explore potential solutions; the second interview focused on designing detailed interaction flows and system architecture. During the first interview, a set of goals and known challenges were presented to the groups to establish context; we included common AR assistance scenarios such as kitchen food preparation, classroom education tasks, and factory workflows.

4.3.2 Known Challenges

We presented participants with known challenges drawn from two sources, a literature survey and the results of the first formative study. The literature survey, which was furnished by searching *AR assistance, embodied assistant, and immersive assistant* on Google Scholar and ACM DL, is described in the following subsections. Two authors separately reviewed these papers, coded the challenges, and formed themes from the coding. In total, 25 common challenges were identified and grouped using thematic analysis [28].

4.3.2.1 C1: Triggering assistance at right time is challenging.

AR assistance must be triggered at the appropriate time during AR interaction. Poor timing strategy may confuse users and negatively impact user trust [139]. If a user is occupied or under stress, for example, frequently or inappropriately displaying AR assistance may be distracting or compound stress. Existing practice in AR assistance regulates the timing and display frequency using the user's intent and actions [251] or fixed intervals. However, these methods do not consider the user's goal and lead to sub-optimal performance.

4.3.2.2 C2: Reusability and scalability in AR assistance are a problem.

Most existing AR assistance systems are designed with ad hoc solutions, where the assistance (e.g., image, text, or voice) is individually developed [197, 223, 234] and later adapted for re-use because each interaction scenario is likely to be unique.

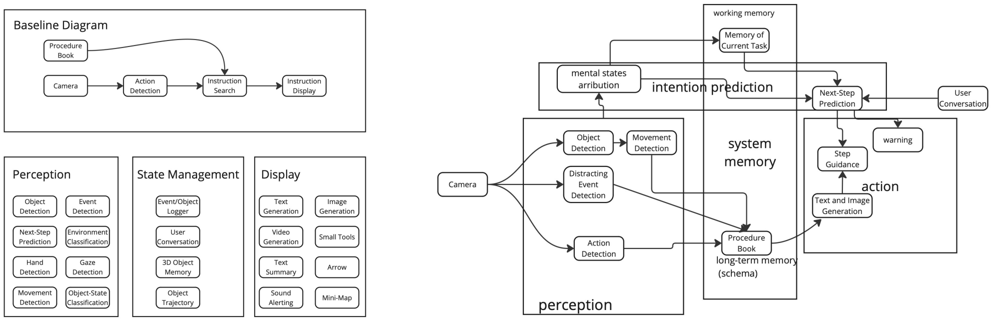
This creates repetitive labor, a concern raised by professional AR designers in our previous formative study.

4.3.2.3 C3: Task interruption and multi-task tracking pose challenges.

In everyday scenarios, users commonly handle multiple tasks at once and encounter interruptions. This creates challenges for AR assistance because oftentimes the system does not recognize that the user is goal switching and so responds incorrectly [21]. In these cases, efficacy will be affected, which can be detrimental to the user's trust in system [185, 385].

Expert	Background	Gender	Group
E1	HCI	M	A
E2	Psychology	F	A
E3	Computer Vision & Psychology	F	B
E4	Psychology	M	B
E5	HCI	M	C
E6	Psychology	M	C

Table 4.2: The table shows the experts' backgrounds in the co-design. We paired one computer science expert with one psychology expert per group. In total, three groups participated in the co-design.



(a) Initially presented diagram.

(b) Sample result.

Figure 4.2: During the first session (participatory design), experts need to collaborate on creating an ideal assistant framework based on the presented diagram and modules. At the bottom of Figure (a), the experts can find the system components for perception. Figure (b) is a result of the original diagram illustrated by one expert group (Group B).

4.3.3 Interview One: Participatory Design

To formalize **how to design a proactive system** capable of determining what to show users for task completion, we presented the known challenges and background knowledge to the experts as described in Section 4.3.2. During the presentation, we described the interaction context, explained the capabilities of current AR technology, and clarified any concerns the experts raised. Each group was then asked to discuss: 1) the information necessary for the AR system to act proactively; 2) any necessary system features, methods, or functions; 3) the perspective helpfulness of user modeling; and 4) ways to mitigate known challenges.

Each group was then moved into their own private discussion room. After a 50-minute open-ended discussion, we provided each group with a list of commonly used tracking, computer perception, contextual understanding, and display technologies and let them select which to use, see Figure 4.2 for reference. Experts were invited to add "imaginary" categories or functions to this list if they considered it theoretically useful. Their modified lists were illustrated using Miro [2].

4.3.4 Interview Two: Adaption of Design Models for AR

The second session involved reconvening the same groups of experts for dyadic interviews. Initially, we presented the outcomes of the first session alongside our synthesized framework, seeking confirmation that it accurately reflected their initial ideas. This was followed by an open discussion where the experts delved into the framework's details and made adjustments to further refine it. This session, which lasted approximately one hour for each of the three groups of experts, was essential for finalizing the design framework for the AR assistant.

4.3.5 Data Collection

Since the interviews were conducted over Zoom, we screen-recorded and transcribed the interviews using Zoom's auto-transcription feature. Two authors independently analyzed the video recordings and transcriptions, coding the findings into insights. The insights were then combined into the following findings based

[2] https://miro.com

on thematic analysis, and discrepancies were resolved through discussion.

4.3.6 Results

The BDI model may be a good candidate for supporting proactive guidance. During the interviews, all three psychology experts (E2, E4, and E6) mentioned that considering **What the user sees and understands in the surroundings** is important for predicting when guidance should appear (C1). For instance, E4 emphasized, "*... it is important to model the human's mental space, so we can adjust the AR (assistance's) timing.*" All the psychology experts introduced *belief-desire-intention*, describing it as well-established and straightforward, as well as a classic cognitive model for understanding human behavior, intention, and goals.

When describing ideas to implement the BDI model within the AR context, the expert groups outlined how *belief* supports the filtering of duplicated or unnecessary assistance and acts as a screening step to narrow the assistance's scope. They further outlined that *desire* refers to the goals of a given task. In AR, this means the system should model the user's actions and goals (Groups B and C). Finally, the expert groups indicated that *intention* comprises a small step toward the goal and affects the timing and content of the AR assistance (Group A and C). Together these adaptations of the BDI model help to construct a novel pipeline toward proactive AR assistance.

Determining the user's intention is essential to proactive guidance. . Group A and Group C first brought up the importance of understanding user intention, which they construed as the *immediate step being undertaken in the context of the guidance.* The group claimed that knowing the intention of the user is beneficial for effectively determining the content of the assistance and its timing. Additionally, when discussing how to design "next-step prediction" in practice, E6 suggested that computer vision models might be able to infer the user's intention. However, E5 thought otherwise and commented that the common method of inferring intention using egocentric short-term memory cannot predict intention reliably. All groups agreed that new methods are required to infer user intention.

Formative Study 1	**Results**
	Benefits Could be automatic. Reduces development time and increases efficiency. Improves scalability. Reduces information overload. ***Challenges*** Cross-domain scalability is difficult. Detecting user intention is a primary challenge, as errors lead to confusion. Adapting the AR instruction to users' active duties is challenging.
Formative Study 2	**Results**
	The BDI model may be a good candidate for supporting proactive guidance. Determining the user's intention is essential to proactive guidance. Understanding high-level goals improves transparency and efficiency. Using the potential of modern LLMs might offer a better understanding of context, environment, objects, and actions.

Table 4.3: **The table summarizes the main results from the two formative studies.**

Understanding high-level goals improves transparency and efficiency in task switching.. Groups B and C discussed transparency challenges in human-AI collaboration. Interaction can be improved if the information on tasks, objects, and goals is available to both the system and the user simultaneously. On the users' end, this is essential to support multi-tasking with task guidance as users are constantly aware of "how the system interprets the current situation" (Group A, B). On the system's end, knowing the user's high-level goals (e.g., task goals) can support multi-tasking effectively and automatically (E1 and E2). Additionally, providing the step-by-step reasoning that leads toward task completion is beneficial for users in that it allows them to maintain trust while collaborating with AI (Groups A and C).

Using the potential of modern LLMs for understanding context, environment, objects, and actions.. E5 has extensive experience in traditional computer vision models and expressed concern that current computer vision models may not be sufficient due to the inaccuracy of action and intent prediction. Even if users' intentions (i.e., immediate goals) can be detected, the predicted intention cannot be used to the fullest extent because these models often lack the ability to understand the user's environment or make accurate decisions based on intent predictions. E1, who has significant experience in LLM development, suggested that multimodal LLMs like GPT-4V could offer a solution because of their advanced reasoning capabilities. Exploring prompting techniques may help to detect context, environment, objects, and actions.

4.4 Design Requirements

Based on the findings of the two aforementioned formative studies, as summarized in Table 4.3, we propose the following design requirements for consideration in proactive AR assistance.

D1 Proactive AR assistance can be challenging to implement due to difficulty in timing its appearance, updating assistance to fit the user's environment, and understanding the user's goals and actions. The BDI model offers a new opportunity to provide real-time, in situ, updated AR content.

D2 AR assistance should convey appropriate content via an appropriate modality at the right time. It should also support users switching tasks or actively managing task life-cycle (i.e., beginning, pausing, and ending).

D3 Assistance should try to be transparent to gain users' trust, feed back the system's reasoning and detection, and provide easily accessible information about current and overall goals in the AR environment.

D4 LLMs could be used to improve scalability and re-usability. Using LLMs might offer a viable way to analyze complex environments, model user actions and goals, track progress, and update assistance content in situ. The result would be a more adaptive, scalable system for various common tasks.

4.5 Satori System Design

Guided by the design requirements, we present the implementation of our proposed Satori system. The goal of the implementation is automatic multimodal AR assistance (e.g., instructions, images, illustrations) with appropriate timing and content that is adaptive to the user's immediate surroundings. Through Satori, we aim to automatically update content to match the context and environment of the interaction, reducing the need for repetitive instructional information toward task completion.

We first use the BDI model as a blueprint to design a workflow to achieve proactive assistance. Next, we detail the implementations for timing prediction and assistance prediction. Finally, we describe our interface and interaction design while ensuring transparency and interpretability.

4.5.1 Implementing the BDI Model for AR Assistance

Architecture: We account for the unique characteristics of AR devices and technologies, such as small fields of view, the need for continual real-time environmental mapping, and the blend of physical and digital information. We describe how to apply the BDI model in terms of its components. This approach has been used when applying the BDI model to other fields to support intention and goal analysis [345, 367]. We follow a similar approach and implement the system architecture as in Figure 4.3. The results of the implementation are also summarized in Table 4.4.

4.5.1.1 BDI-guided chain-of-thought

On a high level, the BDI model aligns with the concept of chain-of-thought (CoT) [318] in LLM. CoT is a form of reasoning that allows the LLM to deliver assistance in a structured manner by sequentially following logical steps. By conceptualizing the BDI model as a series of thoughts, the model can systematically produce the appropriate assistance. Each thought in the process is marked with a hashtag, enabling the LLM to decompose complex tasks into manageable steps, thereby enabling reasoning functions (e.g., action prediction, task prediction, guidance, etc.)

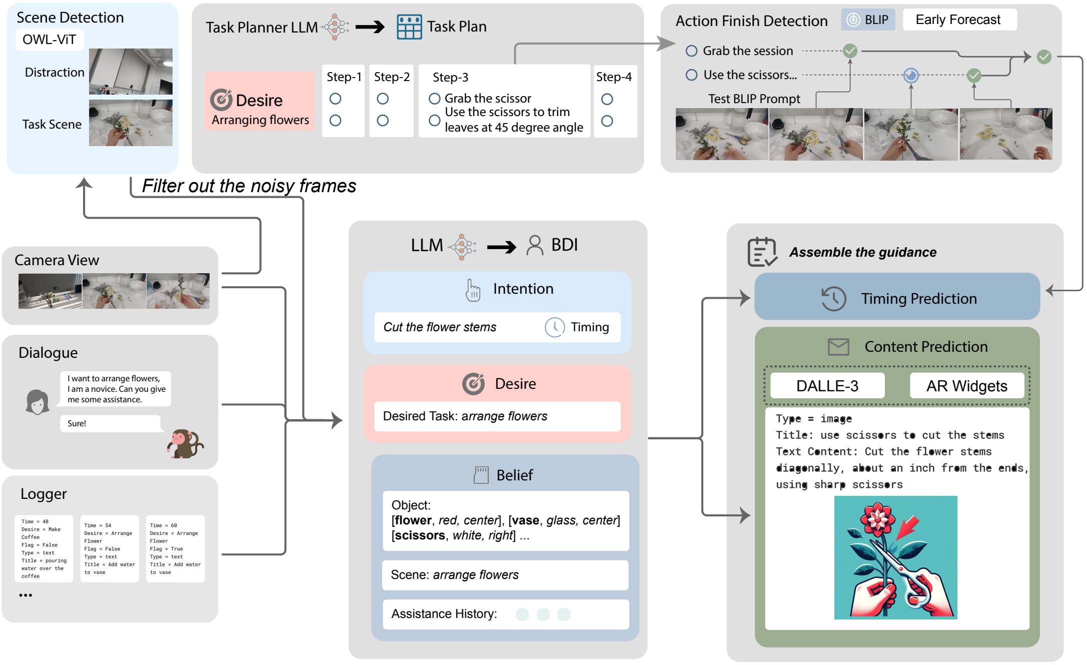

Scene Detection
OWL-ViT
Distraction
Task Scene
Task Planner LLM
Task Plan
Desire
Arranging flowers
Step-1
Step-2
Step-3
○ Grab the scissor
○ Use the scissors to trim leaves at 45 degree angle
Step-4
Action Finish Detection
BLIP
Early Forecast
○ Grab the session
○ Use the scissors...
Test BLIP Prompt
Filter out the noisy frames
Camera View
Dialogue
I want to arrange flowers, I am a novice. Can you give me some assistance.
Sure!
Logger
Time = 48 Desire = Make Coffee Flag = False Type = text Title = pouring water over the coffee
Time = 54 Desire = Arrange Flower Flag = False Type = text Title = Add water to vase
Time = 60 Desire = Arrange Flower Flag = True Type = text Title = Add water to vase
...
LLM
BDI
Intention
Cut the flower stems
Timing
Desire
Desired Task: arrange flowers
Belief
Object: [flower, red, center], [vase, glass, center] [scissors, white, right] ...
Scene: arrange flowers
Assistance History:
Assemble the guidance
Timing Prediction
Content Prediction
DALLE-3
AR Widgets
Type = image
Title: use scissors to cut the stems
Text Content: Cut the flower stems diagonally, about an inch from the ends, using sharp scissors

Figure 4.3: The figure is a system overview of the BDI user model. The system processes inputs from the camera's view, dialogue (voice communication between the user and the GPT model), and the historical logger (records of prior assistance). These inputs are sent to different BDI components for analysis and inference using a combination of local models and LLMs to generate proactive guidance and determine the appropriate modality and assistance timing. To ensure assistance appears and disappears at the right time, a task planner LLM generates a step-by-step task plan based on the inferred desire, with multiple checkpoints assigned to each step. These checkpoints are monitored by the action finish detection module, which determines task completion by verifying checkpoint progress. In addition, the system employs an early forecasting module to minimize latency.

in AR assistance. The following subsections describe how we conceptualize the BDI model.

4.5.1.2 Belief

Human *belief* is a complex psycho-neural function integrally connected with memory and cognition [178, 226]. Precise modeling of human belief within the constraints of AR technology is not feasible without access to human neural signals. To approximate the user's belief state within AR constraints, we propose a two-fold method via capturing scenes and objects from the AR's visual input and via user action history from task performance.

The **scene** provides information on the user's surrounding physical setting, the context of the ongoing task, and changes in their goals and actions. We represent the scene via the label predicted by the image classification model. The label prediction uses an OWL-ViT model [201], which is the zero-shot object detection model. The scene detection is implemented with the zero-shot image recognition model CLIP model [235].

Object information could be used to locate and filter task-relevant objects in the scene from others. To achieve this, we used two different models for object detection: Detr model to detect objects in the scene in zero-shot [36]; and LLaVA model to detect objects that are being held/touched/moved by human hands [174]. We did not use fixed-label set models because they cannot cover the entire case. We did not use the traditional object detection models in this case because these models are trained to predict a fixed set of labels, limiting generalizability.

Action and assistant history is used to ensure the guidance does not repeat. Due to the nature of linear task guidance, completed steps or instructions should not reappear. In our earlier testing, we noticed that the model prediction may give the same instructions that had appeared previously despite task progression. As a result, we implemented a history log to reduce such repetitions. This history contains user interaction logs, the AR assistance content, descriptions, progress, modalities, and images.

4.5.1.3 Desire

This component represents the user's high-level goals, or task goals for the AR system. From cleaning a room, to preparing food, to organizing a shelf, high-level goals are short-term tasks users aim to accomplish. Inspired by recent work that successfully used LLM to understand instructional tasks, we infer the user's goals using a GPT-4V, which takes the current camera frame as input to predict the high-level goals. Image frames are downsampled to 1 fps and sent to the LLM with a prompt specifying the need to understand "what the user is doing, at what place". The resulting label from the GPT-4V contains the task's general description (e.g., moving a table, arranging a desk, etc.).

However, the current LLM does not always predict the goals correctly. Our initial testing revealed 85% accuracy in predicting the correct task goals in a common household setting. As a result, instead of Satori immediately beginning to instruct the user after detecting their goals, it first asks users to **confirm** the predicted tasks or goals. This allows the users to begin AR guidance only if they accept Satori's suggestion of a given task, ensuring error-free task launching.

4.5.1.4 Intention

The results of the formative studies established that the concept of intention from the BDI model could affect the content, timing, and upcoming actions required to complete a task. To predict the user's upcoming actions, we rely on perceptual information (D1), including visual cues and user interactions with objects. We use a combination of localized models with LLM to balance the time cost for timing prediction. As for content prediction, we use customized prompts and CoT coupled with GPT-4V's semantic understanding to determine what type of assistance might be needed.

4.5.2 Timing Prediction

To determine when assistance should appear the system must first detect a user's action and then the corresponding assistance follows. We begin with a step-by-step pipeline to predict when an action will occur. The first naive implementation performs action forecasting after the previous action is completed. This is achieved by concatenating the last four frames and sending them to GPT-4V model via OpenAI's API at 1 fps. However, since the model prediction from LLM is not instant, the user must wait for the prediction to display after actions are finished, resulting in their interaction experience being interrupted. To correct this, we use a combination of **action forecasting** and **early forecasting** to reduce the interaction latency and provide a seamless experience. When the system is running, it continuously executes action forecasting using LLM; meanwhile, parallel early forecasting focuses on detecting action completion. Once detected, cached actions from the continuous action forecasting are immediately retrieved and the assistance is displayed. This way the user no longer has to wait (what was typically about 3 extra seconds) after their action was finished to move forward.

4.5.2.1 Action forecasting

We propose a multimodal LLM to forecast upcoming user actions. This is challenging due to the vast range of potential future actions, the ambiguous nature of user goals, and the misalignment with the label set. We start with constraining the forecasting process by incorporating the user's high-level goals, thus narrowing down the range of possible actions. We then prompt these actions to the LLM using a search-and-reflect framework consisting of three stages:

1. **Analysis Stage:** The LLM first analyzes the current task goals and corresponding task plan (see Section 4.5.5), breaking it down into actionable steps.

2. **Prediction Stage:** After analyzing the goals and plans, the LLM determines the upcoming actions. This involves using contextual cues (e.g., physical objects, scenes, and the user's action history) and the results from the task planner to converge on several probable actions.

3. **Reflection Stage:** The LLM further narrows to the single predicted action (or next step) by integrating the objects and tools in the scene. Actions that require missing or unavailable objects are eliminated, ensuring that only viable actions are suggested. This filtering helps refine the prediction further by aligning it with the actual scene context, reducing irrelevant or impossible options.

4.5.2.2 Early forecasting with finished action detection

Early forecasting focuses on minimizing response time and serves as a flag to retrieve cached action forecasting results. The action finish detection detects a series of checkpoints (see Section 4.5.5), or mini-goals within each step. If all checkpoints are reached, the action detection is complete. It is important to reduce the detection noise, such as the user not looking at the task or another person coming into view. Since there are no pre-trained models or large-scale datasets for detecting when an action is finished, we use the zero-shot learning capabilities of the vision-language model and propose an ensemble-based approach to balance latency and effectiveness. We ensemble the local image captioning model BLIP-2 [164] with the online GPT-4V model. BLIP-2 model has lower accuracy, and this pipeline double-checks its result with the GPT-4V model, which produces more reliable action prediction results based on our initial tests. BLIP-2 model also continuously outputs the prediction of where the user is looking, notifying the AR assistant if the user is distracted and filtering out noise.

4.5.3 Dynamic Content Generation

The content of AR assistance comes in different forms and via different modalities; inspired by the AR designers in the first formative study, we implemented text, image, sound, and tools for Satori. Each has different functions and use cases relative to scene context and user actions.

1. **Text:** We use white text on a black, transparent container to ensure readability. The text primarily contains general instructions (task names, titles, etc.), interface information, and step-by-step guidance. All text is dynamically generated from either the LLM's response or sub-steps from the task planner module, see Figure 4.3 for details.

2. **Image:** Images are generated in situ using DALL-E 3 to depict actions and objects, see examples in Figure 4.4. For more complex actions, we employ multiple images. See the appendix for the implementation details.

3. **Sound:** We use the headset's text-to-speech module for: 1) answering user's

spoken responses; 2) reading instructions aloud; and 3) confirming task completion.

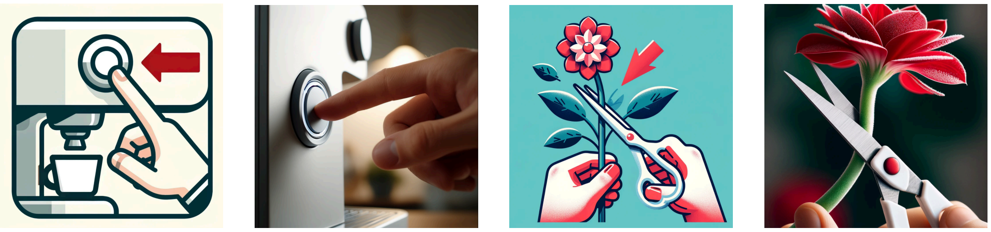
4. **Tools:** We implemented three example tools as a demonstration. Additional tools could be added to the pipeline if needed. *A voice-assistant* that is triggered by the keyword “Hello Tori” will listen and respond to voice input and can be used to command system actions with words such as “yes” or “cancel”. If the system thinks the task step requires time counting (e.g., boiling water, microwaving, grinding powders), a *timer* automatically appears. This is achieved by comparing the objects in the scene with objects needed for the current step in the task using the LLM’s reasoning ability. *Object indicator* locates the “objects of interest” in the current step. This is done through the object detection methods described in the earlier Section 4.5.1.2.

4.5.4 Inferring Modality

We use a GPT-4V to determine the modality using a set of rules in a prompt. The rules map a relationship between the intention and the current step to the corresponding modality. Based on the suggestions from the second formative study, we implemented four rules and their corresponding modality mapping: 1) for intention or steps involving a tool or interaction with materials the LLM returns an image; and 2) if the action is time relevant, the LLM gives a sequence of images; 3) if time counting is needed, the LLM shows the timer tool; and 4) if the step is challenging, the LLM asks for audio feedback. These rules are not mutually exclusive and could generate a combination if multiple conditions are met.

4.5.5 Task Planner for Checkpoints

This component first retrieves the most compatible task from a task database once the user’s goal is set (see Sections 4.5.1.2 and 4.4). It then provides detailed step-by-step instructions and layout **checkpoints or sub-steps** for the AR assistance. Each checkpoint is an actionable sub-step to reach the current step completion. The benefit is twofold: 1) It increases system transparency and builds trust for users as each checkpoint is explicitly listed on the AR interface, and 2) it decomposes the step prediction into smaller milestones for the system, increasing overall prediction validity.



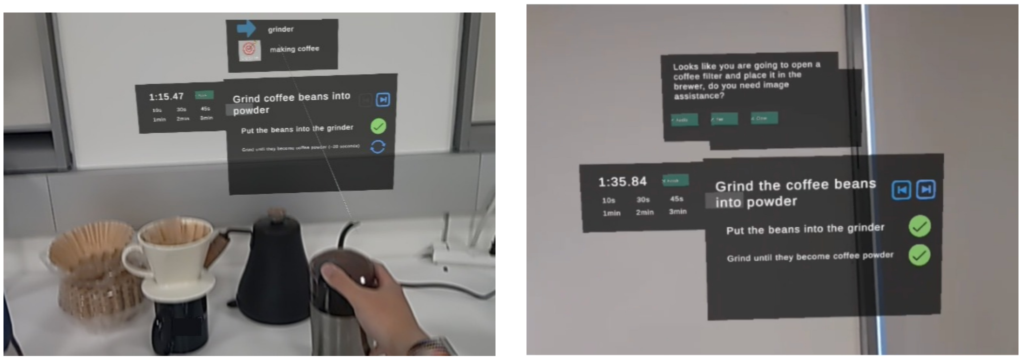
(a) Satori (b) Naive (c) Satori (d) Naive

Figure 4.4: Comparison of the naively generated images from the GPT model (i.e., Naive) with our proposed prompts (i.e., Satori). (a) "*One hand presses a white button on a white espresso machine. A large red arrow points to the button. No background, in the style of flat, instructional illustrations. Accurate, concise, comfortable color style.*" (b) "*One hand presses a white button on a white espresso machine.*" (c) "*Cut stem of a red flower up from bottom, with white scissors at 45 degrees. One big red arrow pointing to bottom of the flower stem. In the style of flat, instructional illustrations. No background. Accurate, concise, comfortable color style.*" (d) "*Cut stem of a red flower up from bottom with white scissors at 45 degrees.*"

4.5.6 Interface and Interaction Design

Figure 4.5 displays the interface with assistance, including active task (e.g., *Making Coffee* in the example), text instructions, images, and tools such as the object indicator (e.g., *coffee grinder*) or timer. Aligning with the design requirement to remain transparent D3, Satori's interface shows how the system tracks the user's task, their progress, and objects of interest. For example, the object indicator not only shows the object that the user needs to interact with but also points to the object's physical location relative to the user.

The voice feature is also supported to let users communicate with Satori hands-free. The Voice interface is activated if the user calls out the activation phrase or during any confirming stage such as goals or step confirmation. This allows the users to quickly express their intentions without interrupting the tasks on hand.


grinder
making coffee
1:15.47
10s 30s 45s
1min 2min 3min
Grind coffee beans into powder
Put the beans into the grinder
Grind until they become coffee powder (~20 seconds)
Looks like you are going to open a coffee filter and place it in the brewer, do you need image assistance?
1:35.84
10s 30s 45s
1min 2min 3min
Grind the coffee beans into powder
Put the beans into the grinder
Grind until they become coffee powder

(a) the user's desire (b) the confirmation page.

Figure 4.5: (a) In this example, the user is grinding the coffee beans. The interface shows the task goal as "Making Coffee" and the upcoming action or step as "Grind coffee beans into powder." The action checkpoints marked with green checks indicate the number of sub-steps that are completed. The action checkpoints marked with a blue circle indicate the number of sub-steps that are in progress. Once all sub-steps are checked, the current step is considered complete; and (b) A task assistance confirmation appears when the system detects step completion. The confirmation prompts the user, asking if they are about to use a coffee filter and whether they need assistance.

4.5.6.1 Human-AI interaction design

In the early testing, we found that users could become overwhelmed if the predicted action changed abruptly. This is because no existing systems can perfectly predict user action, and not every action is meaningful for the task (i.e., behavioral noise). Moreover, due to the nature of step-by-step guidance, prediction errors tend to accumulate across steps, and, without human correction, errors in earlier steps may propagate to later steps. Therefore, we opted to use a confirmation panel to determine whether the system's task or action prediction matched the user's intention, as shown in Figure 4.5. For example, in the coffee-making task, if AR assistance failed to detect that the coffee beans had been grounded, it might continuously prompt the user to grind the beans. With Satori, the system prompts a confirmation page, waiting for the user to confirm action completion. No additional information will appear to the user before they confirm the step with either the pinch button or voice. Similarly, when a new task or step is detected, the confirmation page displays, and users decide whether it matches their needs or the current step, as shown in Figure 4.5(b).

BDI Comp.	Definition	AR Guid. Comp.	Inference Method	Example Usage
Belief	Representation of the world	Scene understanding	OWL-ViT for zero-shot scene classification	Minimizing distractions caused by head movement
		Task-relevant object detection	DETR for object detection, verified by LLM	Locating objects to improve task efficiency
		User action history	Logged by an in-memory logger and inferred by LLM	Preventing repeated instructions for completed steps
Desire	Goals or objectives	High-level task goal	LLM-based scene analysis with user confirmation	Assisting task transitions with accurate goal identification
Intention	commitments that are actively pursued to achieve goals	Next intended action	GPT-based inference with CoT reasoning	Providing step-by-step guidance for upcoming actions
		Timing of next action	Checkpoint-based early forecasting	Reducing latency in delivering next guidance

Table 4.4: This table illustrates how three components— Belief, Desire, and Intention– in the BDI model are adapted for AR task guidance. BDI Comp. refers to the BDI components and AR Guid. Comp. refers to the AR system's task guidance components. **Belief** is represented through scene understanding, task-relevant object detection, and user action history to minimize distractions, locate objects, and avoid repeated instructions. **Desire** captures the user's high-level task goals, inferred through LLM-based scene analysis and confirmed by the user to ensure accuracy. **Intention** includes predicting the following intended action using GPT-based inference with chain-of-thought reasoning and determining the timing of next actions with checkpoint-based early forecasting.

4.6 Evaluation

We evaluated the Satori prototype through an open-ended exploratory study, focusing on the following research questions:

RQ1 Can Satori provide the correct assistant content at the right *timing*?

RQ2 Can Satori provide *comprehensible and effective* guidance?

RQ3 How does our system's guidance compare to that of professional AR experts?

4.6.1 Study Setup

4.6.1.1 Tasks

For our main tasks, we chose four everyday tasks that are comparable in difficulty but different in their goals and required skills, as shown in Figure 4.6. The four tasks were initially sampled from WikiHow [3] and were subsequently rewritten to ensure a consistent task load. Each task asked for specific sequencing and approach, minimizing users' ability to jump ahead of the instructions using prior knowledge. The task orders were pre-determined and counter-balanced for all 16 participants to avoid the ordering effect. The tasks were as follows:

1. *Arranging Flowers:* Participants arranged a variety of flowers in a vase, testing the system's ability to provide accurate and aesthetic guidance.

2. *Connecting Nintendo Switch:* This task involved setting up a Nintendo Switch with a monitor, evaluating the system's technical guidance, and troubleshooting support.

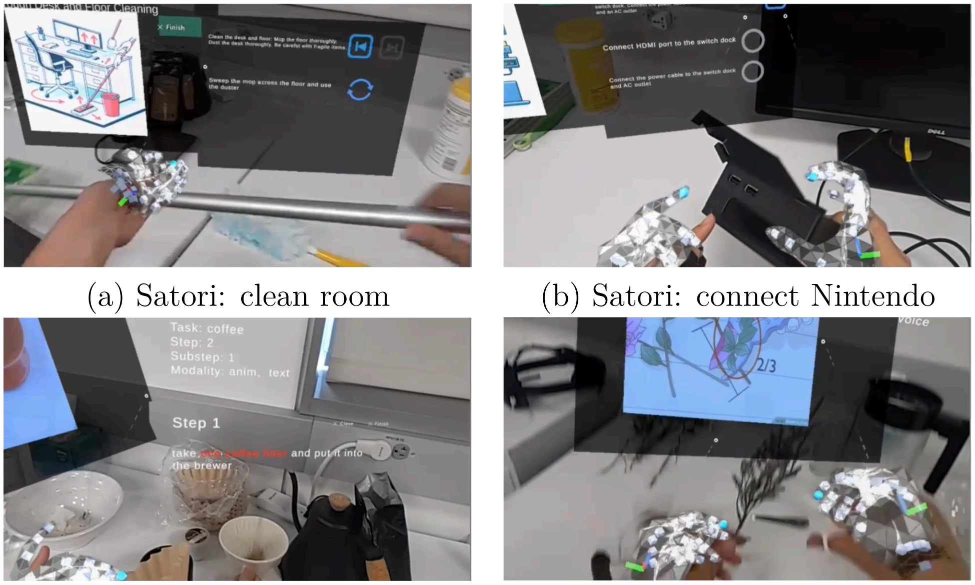
3. *Room Cleaning:* Participants assembled a mop and a duster, and cleaned a desk and the floor; the AR assistant suggested assembly instructions and a cleaning strategy.

4. *Making Coffee:* This task required making coffee using the pour-over method, with the AR assistant providing instructions on tool usage and pouring techniques.

[3] https://www.wikihow.com/

4.6.1.2 Conditions

Participants were presented with two conditions, Wizard-of-OZ (WoZ) and Satori. The tasks (indexed as 1, 2, 3, and 4) and conditions were presented in a counterbalanced order to mitigate the learning and other sequencing effects.


Desk and Floor Cleaning
× Finish
Clean the desk and floor: Mop the floor thoroughly. Dust the desk thoroughly. Be careful with fragile items.
Sweep the mop across the floor and use the duster
Connect HDMI port to the switch dock
Connect the power cable to the switch dock and AC outlet
Task: coffee
Step: 2
Substep: 1
Modality: anim, text
Step 1
× Close
× Finish
take one coffee filter and put it into the brewer
voice
2/3

(a) Satori: clean room

(b) Satori: connect Nintendo

(c) WoZ: make coffee

(d) WoZ: arrange flowers

Figure 4.6: Evaluation tasks using either Satori or a Wizard-of-Oz baseline. (a) The participant is assembling a mop during the room-cleaning task; and (b) The participant is connecting an HDMI cable to a Nintendo Switch dock during the connecting Nintendo Switch task; and (c) The participant is preparing a filter during the coffee-making task; and (d) The participant is trimming flower stems during the flower-arranging task.

4.6.1.3 Participants

A total of 16 participants (P01-P16, 11 male, 5 female) were recruited via a university email group and flyer. The average age was 23.8, with a maximum age of 27 and a minimum age of 21. 10 of the 16 participants had AR experience prior to the study. Each participant was compensated with a $30 gift card for their participation. Information on general wellness was collected from participants both before and after the study, and no motion sickness was observed following the study.

4.6.1.4 Apparatus

We used a Microsoft HoloLens 2 headset as the AR device for the study. Participants used the Satori system or WoZ system described earlier while performing the tasks. The headset connects to a server with an Nvidia 3090 graphics card to fetch real-time results.

4.6.1.5 Procedure

The study began with a brief tutorial introducing participants to the interface of the two AR systems. Afterward, participants were assigned four everyday tasks. They started with either the WoZ system or the Satori system before alternating to the other condition. After completion of each task, participants evaluated their experience using a usability scale and assessed their cognitive load using the NASA Task Load Index (NASA TLX). We also conducted a brief recorded interview, asking participants about the advantages, disadvantages, usefulness, and timeliness of the two systems. The experiments were supervised by the Institutional Review Board (IRB), and all task sessions were video-recorded. These recordings were securely stored on an internal server that is inaccessible from outside the university. Participants provided consent, and their personal identity was strictly protected. We collected data on participants' well-being both before and after the experiment and observed no significant adverse effects. The duration of the entire study was two hours on average. All participants completed the four tasks using both systems.

4.6.2 Data Collection

We used the following metrics to measure the users' perspective on how Satori's content and timing compared to the AR designer's version. Since content is automatically generated, we measured comprehensibility, helpfulness, and overall cognitive load to assess whether our system is capable of generating similar content utility without overwhelming the users.

4.6.2.1 User-rated scale

For RQ4.6 and RQ4.6, we opted to use a seven-point Likert scale (similar to Lewis et al., [159]), ranging from “strongly disagree” to “strongly agree” to measure the timeliness, ease of use, effectiveness, comprehensibility, and helpfulness of the AR assistance. Eleven questions were asked in total. For the complete set of 11 questions, see Table 4.11. We computed the mean and confidence intervals for each question using the bootstrapping method. Specifically, 1,000 bootstrap samples were generated from the original data set for computation with 95% confidence intervals for the estimation of the uncertainty around the mean.

4.6.2.2 NASA Task Load Index

We used the raw 100-point NASA TLX [94] form to measure the cognitive load with the six subcategories. Mean and confidence intervals were calculated for the sum of all ratings and each of the subcategories using the bootstrapping technique. 1,000 bootstrap samples were drawn from the original dataset with 95% confidence intervals to measure the uncertainty surrounding the mean.

4.6.2.3 One-Sided Wilcoxon Signed-Rank Test

A one-sided Wilcoxon signed-rank test was used to determine whether the user-rated scale and the TLX ratings were significant. The goal was to test whether Satori performed similarly to the AR assistance designed by professionals in AR; however, simply verifying that there is no significant difference between them does not ensure the two conditions are similar. Instead, we aimed to test whether Satori was no worse than the WoZ by a predefined margin Δ [85, 148, 157].

The test defines $D_i = X_{Ai} - X_{Bi}$ as the difference between the scores for each participant i under Conditions S (Satori) and W (WoZ), respectively. The adjusted difference accounting for the margin is given by $D'_i = D_i - \Delta = X_{Ai} - X_{Bi} - \Delta$. The hypotheses for this non-inferiority test are:

$$H_0 : \text{median}(D') > 0 \quad \text{(A is worse than B by more than } \Delta\text{)},$$

$$H_1 : \text{median}(D') \leq 0 \quad \text{(A is no worse than B by at most } \Delta\text{)}.$$

Similar to the vanilla Wilcoxon signed-rank test, this procedure involves ranking the absolute adjusted differences $|D'_i|$, calculating the sum of ranks for positive (W^+) and negative (W^-) differences, and using the test statistic $W = \min(W^+, W^-)$

to compute a one-sided p-value. This p-value indicates whether we can reject H_0 in favor of H_1. We chose the margin value $\Delta_{TLX} = 2.5$ for NASA TLX and Δ_{us} for the usability scale as they represent half of the rating interval.

4.6.3 System Evaluation Preparation

We used the GTEA [168], EgoTaskQA [116], study recordings, and our dataset to evaluate Satori. The GTEA dataset contains egocentric videos of participants performing daily life tasks, and the EgoTaskQA dataset contains questions about humans' beliefs in the world and the model's understanding of humans' beliefs. We used the GTEA dataset with 71 labels and leave-one-subject-out cross-validation. Since the EgoTaskQA dataset has a large amount of data in the test set, we sampled 200 data points for the evaluation. We use the indirect split, which has a more complicated relationship between the actions and the questions. The user study recordings consist of 14 participants who performed the four tasks described in this section. Two participants' recordings were lost due to data corruption. In addition, we added 4 more sets of the four tasks (totaling 16 videos) as our dataset for evaluation. GTEA, EgoTaskQA, and our dataset are used to evaluate the BDI model output, and user study recordings are used to evaluate modality and guidance timing.

4.7 Results

4.7.1 System Evaluation

We evaluated Satori's module-level performance on the GTEA dataset and the video dataset we recorded from the empirical study and testing. For *desire* task prediction, Satori achieved a balanced accuracy of 100% on GTEA dataset and our dataset (Table 4.5). Satori achieved 66.50% in *belief* inference, matching the state-of-the-art HCRN model [149] on EgoTaskQA dataset 69.53% (Table 4.6). The results on intention forecasting (timing and intention) revealed a 78.38% precision to predict user actions (Table 4.7). For modality prediction, Satori reached an average of 75.12% recall in deciding the modality that matches the WoZ designed by AR experts (Table 4.8). We discuss the implications of these results in the discussion session.

Dataset	GTEA	Our Dataset
Satori	100.00	100.00

Table 4.5: Desire inference includes understanding high-level task goals. We evaluated this module using the GTEA dataset and our dataset, which is annotated by three experimenters. Satori achieved a balanced accuracy of 100% on both datasets

Dataset	EgoTaskQA	Our Dataset
Task	Scene Understanding	Object Understanding
HCRN	69.50	N/A
Satori	66.50	57.90

Table 4.6: Belief inference includes scene understanding and task-relevant object understanding (object labels, locations) and their interaction history with the user. As for the evaluation, the goal is to understand the reasoning capability for scene understanding and object understanding. We evaluated this module using the EgoTaskQA dataset and our dataset to compare with the HCRN model. The EgoTaskQA dataset consists of questions about humans' understanding of the scene and the model's understanding of humans' beliefs. For our dataset, three experimenters annotated the highlighted object labels, locations, and interaction states separately. Satori reached a similar accuracy (66.50%) to that of the HCRN model (69.50%).

4.7.2 Usability Scale

We present the participants' raw scale data across all tasks in Figure 4.7 and processed statistics in Table 4.11. The questions are listed in Table 4.9. We found that there was no significant difference between most of the Satori and the WoZ conditions, suggesting that Satori's overall performance matched the wizard-of-oz designed by AR experts ($p_{non\_inferiority} < 0.05$). (e.g., Q1: $p = 0.099$, Q2: $p = 0.094$, Q3: $p = 0.090$, Q6: $p = 0.273$). However, non-inferiority tests demonstrated that Satori was not worse than the WoZ condition (e.g., Q1: $p = 0.001$, Q2: $p = 0.000$, Q6: $p = 0.001$) with a margin of $\delta = 0.5$.

L.A. Time	GTEA			Our Dataset		
	Recall	Prec.	F1	Recall	Prec.	F1
0s	63.04	78.38	69.88	**65.61**	62.52	58.89
1s	54.35	75.76	63.29	55.00	48.40	46.06
3s	39.95	65.73	49.43	52.31	44.44	45.24

Table 4.7: This table shows the module-level evaluation of intention (action) forecast. L.A. Time refers to Look-Ahead Time, Prec. refers to Precision score. We evaluated our methods on the GTEA dataset and our dataset. Three experimenters annotated user action labels in our dataset. Aside from the settings Satori uses (Look-Ahead Time = 0s), we also present results for two other hypothetical conditions if we predict the action 1s or 3s earlier. For our settings, our methods reached 78.38% on the GTEA dataset and 62.52% on our dataset.

4.7.2.1 Content.

Satori's adaptive AR content provides similar comprehensibility ($p = 0.099$, non-inferiority $p = 0.001$) and helpfulness ($p = 0.094$ and $p_{non\_inferiority} = 0.001$) to complete a guidance task compared to the baseline. Dynamic assistance almost matches with pre-designed assistance ($p = 0.357$, non-inferiority $p = 0.001$). This is in line with later interview results, where a majority (12/16) believed that Satori was able to provide assistance that appropriately matched the context of their tasks. Satori's image content is well-received, for example, P1 said that "*the picture [of the second one] is very nice and it looks good.*" Images in the WoZ are also useful, as P8 remarked that "*Guidance as a whole (text, images, and animations) was very helpful. Whereas, text alone as shown in the image lacks information.*"

4.7.2.2 Timing.

Satori provides timely guidance to users (Q3: $p = 0.090$ and $p_{non\_inferiority} = 0.001$) with appropriate frequency (Q10: $p = 0.156$ and $p_{non\_inferiority} = 0.002$). In fact, participants describe the experience as impressive (P16) and can display assistance in need (P3). Although occasional network latency has been reported (P4, P6), they comment that the overall experience was "not bad"(P6) and "...sometimes delayed, but I think it's like, it's okay.." (P4).

Task	Guidance Timing	Modality
Arranging Flowers	94.34	94.34
Connecting NS	79.49	74.15
Room Cleaning	80.49	73.17
Making Coffee	75.00	63.75
Average	81.69	75.12

Table 4.8: The table shows the modality prediction results using the user study videos for the four tasks: arranging flowers, connecting Nintendo Switch (NS), cleaning a room, and making coffee. Three experimenters labeled the assistance appearances and compared them with the WoZ. Our methods reached an average of 75.12% when referring to the same assistance type as the designers'. The guidance timing columns show the holistic evaluation on whether Satori generated the proper assistance at the proper time without modality.

4.7.2.3 Effectiveness.

We found that Satori performs better than the baseline in inferring intention (Q4: $p < 0.05$) and at appearing locations (Q5: $p < 0.05$). Most participants rated between "agree" to "strongly agree" that AR assistance appears at proper locations in space in both Satori ($\bar{x} = 6.48$) and the WoZ($\bar{x} = 5.95$). In general, participants felt positive regarding Satori's assistance effectiveness. P3 stated, "*I liked that it combines the various modalities of text, audio, and image to generate guidance, I believe that was helpful on multiple occasions where I might have been uncertain with only a single modality.*" P14 commented, "*The guidance helps me a lot, especially in coffee making. It provides me with very detailed instructions including time, and amount of coffee beans I need. I would have to google it if I don't have the guidance.*" P8 noted that "*For task like arranging the flower vase, the intricate details like trim the leaves, cutting the stem at 45 degrees, etc. are very necessary details that I might not have performed on my own.* "

In terms of the system's learnability questions (Q7: $p = 0.179$ and $p_{non\_inferiority} = 0.001$) and engagement (Q8: $p = 0.145$ and $p_{non\_inferiority} = 0.002$), Satori scored similarly to that of the baseline. P3 remarked that "*not a singular component by itself, but all components together do make me more engaged.*" P10 expressed a sense of active involvement in the task, stating that "*Yes. It may automatically detect my progress to make me more engaged in the task.*"

4.7.2.4 Satori as a proactive AR assistant in everyday life.

Most participants agreed that Satori has the potential to be generalized to everyday scenarios (Q11: $p = 0.277$ and $p_{non\_inferiority} = 0.005$). P9 said that "*maybe when we need to assemble furniture, instead of going through the manual back and forth all the time, we can just have this system to guide us.*" Furthermore, most participants acknowledged that they would not need additional training to use the system (Q7: $p = 0.179$ and $p_{non\_inferiority} = 0.001$), suggesting possible applications for more general purposes. With some training, as P10 mentioned, "*(The system can be used for) learning to complete a difficult task.*"

Table 4.9: Survey Questions

ID	Question Content
Q1	I can easily comprehend content via text/audio/image guidance.
Q2	I can easily understand how to perform my tasks with the guidance.
Q6	I am able to complete my work quickly using this system.
Q7	It was easy to learn to use this system.
Q8	How engaged I am using the system?
Q9	The system's guidance matches the context.
Q10	Overall, the system's guidance frequency and timing are appropriate.
Q11	Overall, I think the system helps my work.

Table 4.10: List of survey questions assessing user perceptions of system guidance, evaluating factors such as ease of comprehension, effectiveness, timing, engagement, contextual appropriateness, and overall satisfaction. The statistical analysis result is presented in Table 4.11.

4.7.3 NASA TLX Result on cognitive load

We found no significant difference between Satori and WoZ on all TLX measures. Detailed analysis within the six sub-categories of NASA TLX revealed no significant difference among the six subcategories of NASA TLX between the two conditions, see Table 4.12 and Figure 4.8 for details.

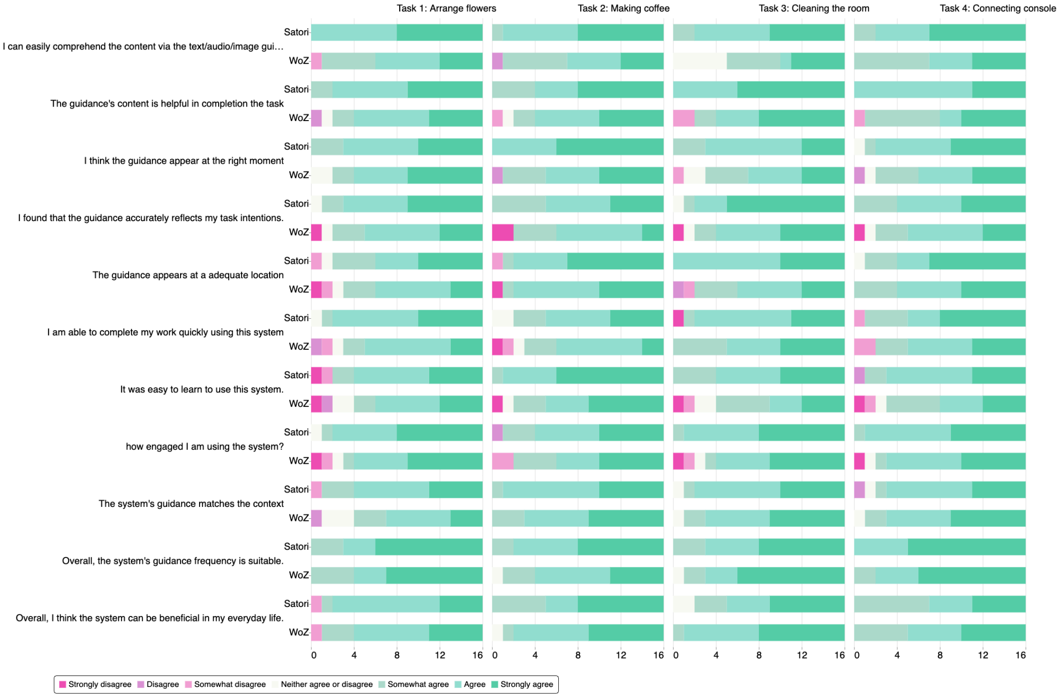


Figure 4.7: Color-coded seven-point Likert scale ratings are shown in the figure for the twelve-participant study. The figure compares the responses for Satori and WoZ systems across four tasks: Arranging Flowers, Making Coffee, Cleaning the Room, and Connecting a Console. Each bar represents the distribution of responses for a specific usability question, highlighting differences in user satisfaction, comprehensibility, and task support provided by both systems.

Question	Condition	Mean	95% CI	Vanilla		Non-Inferiority	
				W	*p*-value	W	*p*-value
[Q1]	Satori	6.25	[6.00, 6.75]	26.500	0.099	**89.500**	**0.001**
	WoZ	5.94	[5.25, 6.50]				
[Q2]	Satori	6.22	[5.75, 6.75]	26.000	0.094	**131.000**	**0.000**
	WoZ	5.80	[5.50, 6.50]				
[Q3]	Satori	5.97	[6.00, 6.50]	17.500	0.090	**125.000**	**0.001**
	WoZ	5.53	[5.00, 6.25]				
[Q4]	Satori	6.48	[6.00, 7.00]	**11.500**	**0.016**	**134.500**	**0.000**
	WoZ	5.95	[5.62, 6.50]				
[Q5]	Satori	6.23	[5.88, 6.75]	**15.000**	**0.032**	**131.500**	**0.000**
	WoZ	5.66	[5.25, 6.25]				
[Q6]	Satori	6.08	[5.50, 6.62]	30.000	0.273	**108.500**	**0.003**
	WoZ	5.75	[5.25, 6.50]				
[Q7]	Satori	6.48	[6.00, 7.00]	22.000	0.179	**103.500**	**0.001**
	WoZ	6.06	[5.75, 7.00]				
[Q8]	Satori	6.16	[5.88, 6.50]	20.500	0.145	**109.500**	**0.002**
	WoZ	5.75	[5.38, 6.50]				
[Q9]	Satori	6.27	[6.00, 6.75]	32.500	0.357	**91.000**	**0.001**
	WoZ	6.05	[5.62, 7.00]				
[Q10]	Satori	6.30	[5.75, 6.75]	30.000	0.156	**97.500**	**0.002**
	WoZ	5.92	[5.75, 6.50]				
[Q11]	Satori	5.94	[5.50, 6.50]	30.000	0.277	**105.500**	**0.005**
	WoZ	5.58	[5.25, 6.25]				

Table 4.11: The table summarizes the mean scores and 95% confidence intervals (CI) for each system (our Satori system and WoZ designed by the AR designer) across usability scale questions using non-inferiority tests. The "Vanilla" columns provide the Wilcoxon signed-rank test results (W statistic and p-values) for significant differences between systems. The "Non-Inferiority" columns show W statistics and p-values testing if Satori's performance is non-inferior to WoZ within a set margin. The highlighted cells indicate established non-inferiority, suggesting that Satori performs comparably or better than WoZ over system performance and usability. The question content is available in Table 4.9

Metric	Condition	Mean	95% CI	Vanilla		Non-Inferiority	
				W	*p*-value	W	*p*-value
M. D.	Satori	34.06	[17.50, 43.81]	60.500	0.744	78.000	0.316
	WoZ	33.12	[16.25, 50.00]				
P. D.	Satori	32.34	[11.25, 50.00]	55.000	0.776	80.000	0.281
	WoZ	30.47	[10.00, 43.75]				
T. D.	Satori	28.52	[21.25, 37.50]	46.000	0.274	65.000	0.388
	WoZ	26.41	[15.00, 32.50]				
P.	Satori	16.17	[7.50, 21.25]	44.000	0.593	**95.500**	**0.022**
	WoZ	17.27	[7.50, 20.00]				
E.	Satori	28.20	[17.50, 37.50]	52.500	0.464	58.500	0.353
	WoZ	26.02	[15.00, 36.25]				
F.	Satori	19.84	[10.00, 28.75]	41.500	0.175	**126.000**	**0.001**
	WoZ	26.95	[11.25, 34.38]				

Table 4.12: This table shows the results for NASA TLX questions and non-inferiority tests using the mean scores and 95% confidence intervals (CI) for Satori and WoZ systems across six dimensions: Mental Demand (M. D.), Physical Demand (P. D.), Temporal Demand (T. D.), Performance (P.), Effort (E.), and Frustration (F.). The Vanilla Wilcoxon signed-rank test results and non-inferiority tests (highlighted) indicate whether the Satori system performs comparably or better than the WoZ system in terms of cognitive load.

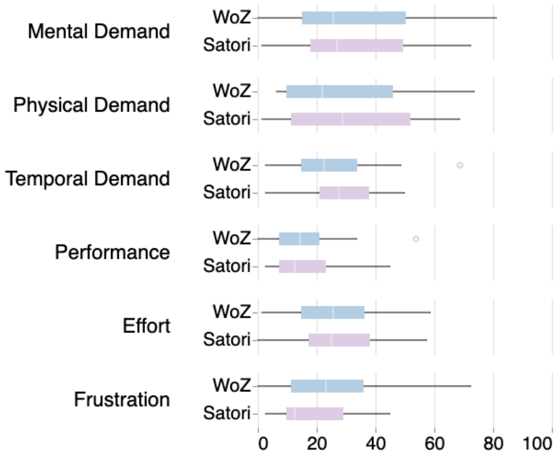


Figure 4.8: The box plot of NASA-TLX results illustrates the distribution of cognitive load ratings across six dimensions: Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration. Each box represents the interquartile range (IQR) with the median marked by a horizontal line, showing the variability and central tendency of participants' workload ratings for both systems. The comparison highlights differences in perceived workload between the WoZ and Satori conditions, providing insights into the effectiveness and usability of each approach.

4.8 Final Considerations

This chapter presents *Satori*, a proactive and adaptive AR task guidance system integrating the Belief-Desire-Intention (BDI) framework with Large Language Models (LLMs). Through multimodal input analysis, including implicit cues from gestures and gaze patterns, Satori dynamically generates context-aware instructions tailored to users' intentions. Our formative studies with 12 domain experts highlighted essential design requirements, emphasizing precise user modeling and adaptive guidance to maintain interaction fluency and reduce cognitive load.

An empirical evaluation involving 16 participants indicated that Satori successfully provided clear, efficient, and timely instructions, comparable in effectiveness to manually designed AR task guidance but requiring significantly less manual effort. Although interaction latency and limited field-of-view remain challenges inherent to current AR hardware, users perceived the multimodal instructions positively, suggesting high practicality. Moreover, the integration of interactive confirmation mitigated potential errors without significantly disrupting user flow. Overall, Satori demonstrates the potential of proactively modeling user intention to improve the usability, adaptability, and scalability of AR guidance systems.

Chapter 5

Intentvizor: Towards Generic Query Guided Interactive Video Summarization

5.1 Introduction

This chapter presents IntentVizor, an interactive video summarization method that leverages graph convolutional networks and a visual analytics interface. This work lays the foundation for our goal of providing interactive multimodal analysis of AR task recordings and identifying task-related insights. The proposed summarization model serves as the basis for our egocentric AR task recording summarization method. Furthermore, this chapter introduces the concept of interactive query-driven summarization, which involves users in the summarization process by allowing them to control the summarization objectives (intentions). This approach directly informs our subsequent work on interactive documentation and summarization of task recordings in Chapter 6.

Traditional video summarization methods usually generate concise/representative summary that contains the entities and events with high priority from the video and with low repetition and redundancy using unsupervised [47, 119, 121, 183, 214, 340, 365], supervised [66, 215, 217, 362, 369, 370, 383] and reinforcement learning ways[9, 378]. However, such a summary cannot satisfy the needs of users and be of low practical value. As the elongated video, especially when captured in

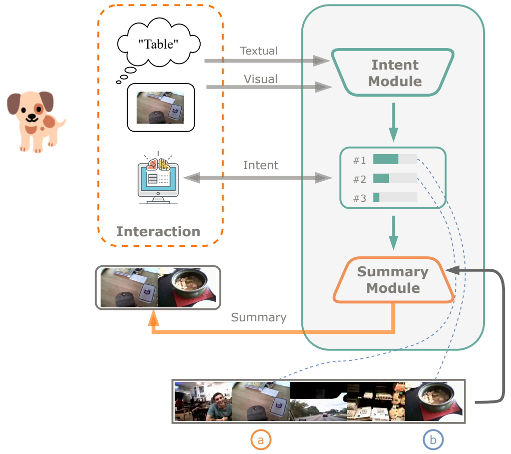


Figure 5.1: Illustration of our IntentVizor framework. We take query "Table" as an example. Generic queries, including text/video snippets related to "Table" are inputs of the model. The intent module transforms these queries into a probability distribution over the basis intents, followed by the summary module, which generates a video summary by combing the basis intents and their probability values. As the user can find the underlying visual semantic meaning of each basis intents (e.g., in the figure, basis intent #1: the dining table; #2: the working table), they can adjust the distribution of these basis intents through our proposed interface (Fig. 5.4) to satisfy their needs, and the final generated summaries can be updated accordingly/iteratively.

the realistic scenario, may cover a wide range of topics, only fractional content of specific topics will meet the user's needs. Based on this observation, the user query-driven summarization model, which considers the user's preference, has gradually attracted researchers' attention.

The basic idea for query-driven summarization is to use the text query to guide the generation of video summaries. A popular dataset for this query-driven summarization was the textual query dataset, proposed by Sharghi et al. [266]. The summarization model proposed in the chapter was trained to predict a subset of the video shots (5 seconds per shot) closely related to the textual query. For the follow-up works, the attention mechanism [196, 332, 333] and generative adversarial networks[366] based summarization models are also introduced to achieve better summarization performance. However, the performance of these models was still not satisfying as the textual query is not enough to represent the users' preferences. To be more specific, first, the user cannot express their detailed needs with few fixed input textual queries at the very beginning of summarization. They may have multiple needs and want to adjust the priority of different needs. Second, the textual query can be ambiguous. People can have different understandings of a word in communication, let alone the model trained on a fixed word dictionary. Therefore, the model should be interactive to loop users into the summarization, and other query formats (e.g., visual query) should be considered to better represent the user preference with lower ambiguity.

To propose a generic model for queries from different modalities and allow users to interact during the summarization process, in this chapter, we propose a novel framework named as IntentVizor. We borrow the concept ***Intent*** from the Information Retrieval (IR) community to define the users' need, independent of the query modalities[31, 132, 360]. However, our intent differs from the traditional definition in IR with different representation and extraction: (1) We represent the intent by an adjustable distribution over the basis intents rather than the pre-defined categories[31] , taxonomies[32, 350] or in a distributed representation space[92, 353]; The basis intents are defined as the learned and basic components of the user's needs. Compared with the traditional definitions[31, 32, 92, 350, 353], our method enables interactive manipulation, satisfying the user's diverse and subtle needs. (2) We extract a unified intent from the queries of different modalities instead of only the textual query to avoid the ambiguity problem as mentioned before.

The intentVizor framework consists of two modules, i.e., the intent module for extracting the intent from the query and the summary module for summarizing the video with the intent. To effectively correlate the video features with the generic query/intent in the two modules, we design a flexible network structure named Granularity-Scalable Ego-Graph Convolutional Network (GSE-GCN). This GSE-GCN will work as a shared backbone for both the summary module and the intent module. Besides this backbone, the two modules each have an intent head and a summary head separately.

To sum up, we structure our contributions as follows:

- To the best of our knowledge, our IntentVizor framework is the first attempt to introduce generic queries to better satisfy the user's diverse needs. We also propose a novel dataset for the visual-query-guided video summarization based on UTE videos.

- We formulate the video summarization as an interactive process, where the user can fine-tune its intent iteratively with our proposed novel interface. This idea is further explored in our proposed work in Chapter 6
- We propose a novel GSE-GCN structure to effectively correlate the generic queries of multi-modalities with the input video.

5.2 IntentVizor Framework

The IntentVizor framework targets at (1) interactive control over the video summarization process; (2) support of the generic multi-modality query. This section first shows that the two requirements can be satisfied by modelling the multi-modality queries as a unified and interactive user intent. Then, we will describe GSE-GCN, which is designed to better deal with multi-modality queries.

5.2.1 Unified and Interactive User Intent

5.2.1.1 Problem Setting

We introduce a novel problem setting with our proposed unified and interactive intent. The canonical setting for query-focused video summarization is to output a representative and concise subset of video shots based on the inputting video $\boldsymbol{v}$ of T shots and text query q_t. We re-define the task by generalizing the text query q_t into the generic query q. Then, we propose to predict not only a final video summary, but also a unified and interactive user intent ζ for the multi-modality queries. ζ can be learned implicitly like a latent variable. We assume that there are a set of basis intents as $Z = \{\zeta_1, \zeta_2, ..., \zeta_k\}$ and the user intent ζ is chosen from the basis intents according to a categorical distribution conditioned on the query q as $\zeta \sim p(\zeta|q, v)$. Given the user query q, the distribution $p(\zeta|q, v)$ is parameterized by the probability vector of basis intents, $\boldsymbol{p(\zeta)} = [p(\zeta_1|q, \boldsymbol{v}), p(\zeta_2|q, \boldsymbol{v}), ..., p(\zeta_k|q, \boldsymbol{v})]^T$.

In practice, the query can be either textual, visual, or other formats. In this work, we only implement the models for textual and visual queries. Following the previous works[266], we represent the text query by two text concepts as $q_t =$

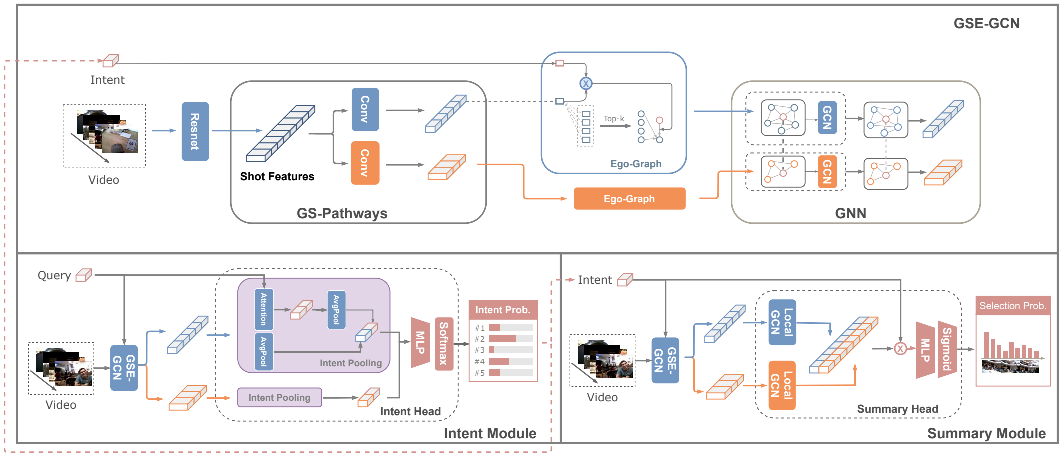


Figure 5.2: **GSE-GCN** exploits two notions i.e., GS-Pathway and Ego-Graph. The input video will be processed by two convolutional networks to produce two segment-level feature sequence of coarse and fine granularity. Then, each sequence will be processed to generate a Ego-Graph, where the intent/query vertex is an ego-vertex with all the video segments are connected. After feeding the graph into GCN, the two pathways will be produce the corresponding segment-level features. **Intent Head** pools the segment features into a distributed representation, which will be processed by a MLP with softmax to produce the intent probability. **Summary Head** exploits the local-GCN module to produce the shot-level features, which will be used to predict the shot selection probability.

.

$\{c_1, c_2\}$, where c_1, c_2 are two concepts. By comparison, we represent the visual query by a set of representative shots in the original video as $q_v = \{u_1, u_2, ..., u_P\}$ where P is a constant number.

Then, for each shot s, we denote $\eta_s \in \{True, False\}$ as whether s should be selected in the summarization. We assume that η_s is sampled from a Bernoulli distribution conditioned on the intent as

$$p(\eta_s) = p(\eta_s|\zeta, \boldsymbol{v}). \tag{5.1}$$

Finally, we can condition the shot selection probability $p(\eta_s)$ on the user query as

$$p(\eta_s|q) = \Sigma_{i=1}^{i \leq k} p(\zeta_i|q, \boldsymbol{v}) * p(\eta_s|\zeta_i, \boldsymbol{v}). \tag{5.2}$$

Instead of the deterministic intent ζ, we characterize the user's needs by the distribution $p(\zeta|q,\boldsymbol{v})$, which weights different basis intents as Equation 5.2 shows. Such a notion follows the perspective of Bayesianism as the latent variable (intent) is a random variable instead of a deterministic value. The user can iteratively adjust the probability vector $\boldsymbol{p(\zeta)}$ to fine-tune its intent.

Since the shot selection probability is often viewed as a summarization score when $\eta = True$, we use the shot score and selection probability interchangeably in this chapter. To implement Equation 5.1, and 5.2, we design two modules $\boldsymbol{p}(\zeta|q) = g(q,\boldsymbol{v}:\theta_g)$ (intent module) and $p(\eta_s|\zeta) = h(\zeta,\boldsymbol{v}:\theta_h)$ (summary module), where θ_g and θ_h are the parameters of g and h

$$p(\eta_s|q,\boldsymbol{v}) = \Sigma_{i=1}^{i\leq k} g_i(q,\boldsymbol{v}:\theta_g) * h(\zeta_i,\boldsymbol{v}:\theta_h). \tag{5.3}$$

Given the ground truth labels, we can optimize the parameters θ_g, θ_h of our modules by the BCE Loss as

$$\mathscr{L}_{BCE}(\theta_g,\theta_h) = \Sigma_{t=1}^{t\leq T} log(p(y_t|q,\boldsymbol{v})), \tag{5.4}$$

where y_t is the ground truth label for the t^{th} shot.

5.2.1.2 Non-Linear Activation

The Equation 5.3 strictly follows the selection probability's theoretical definition in Equation 5.2. However, it restricts the capacity of the intent module because the resulting probability is simply the linear combination of $h(\zeta_i,\boldsymbol{v}) * g_i(q,\boldsymbol{v})$. To address the issue, we trade off the strictness for better performance by adding a non-linearity layer on every basis intent score. Specifically, we employ shifted ReLU[2] as the non-linearity activation.

$$p(\eta_s|q) = \Sigma_{i=1}^{i\leq k} ReLU(g_i(q,\boldsymbol{v}) * h(\zeta_i,\boldsymbol{v}) - \delta), \tag{5.5}$$

where δ refers to the threshold value for the shifted ReLU.

5.2.2 GSE-GCN: Granularity-Scalable Ego-Graph Convolutional Networks

As the shared backbone of intent module g and summary module h, the GSE-GCN exploits two newly proposed components, i.e., Granularity-Scalable Pathways (GS-Pathways) and Ego-Graph Convolutional Network (E-GCN), to better deal with the temporal multi-granularity and sparsity of correlation respectively.

5.2.2.1 Granularity-Scalable Pathways (GS-Pathways)

(a) Coarse Pathway

Layer	Kernel	Stride	Channel	Output Size
Conv1	5	8	1024	$[L//8, 1024]$
MaxPool1	2	1	1024	$[L//8, 1024]$
Conv2	5	1	1024	$[L//8, 1024]$
MaxPool2	3	2	1024	$[L//16, 1024]$

(b) Fine Pathway

Layer	Kernel	Stride	Channel	Output Size
Conv1	5	1	256	$[L//2, 256]$
MaxPool1	2	2	256	$[L//2, 256]$
Conv2	5	1	256	$[L//2, 256]$
MaxPool2	2	2	256	$[L//4, 256]$

Table 5.1: Hyperparameter settings of the granularity-scalable pathways. L denotes the length of the original video.

Models with the constant temporal granularity may fall short in aligning the video events/actions of multi-granularity with the user query/intent. We have shown in Figure 5.3 that the actions of different temporal lengths and movement speeds should be processed with the features of different temporal granularity. The issue raises the necessity of a granularity-scalable model. To realize it, we propose a flexible structure with two pathways of different granularity. The idea is similar with [67] technically while being motivated by different concerns. For each pathway, we aggregate shot-level features into segment-level features (a segment spans 4 and 16 shots with the fine and coarse pathways, respectively) by a convolutional network. We list the hyper-parameters in Table 5.1. The produced segment-level

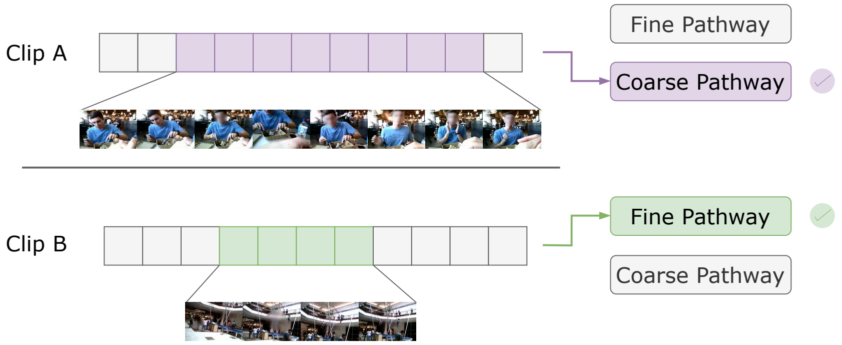


Figure 5.3: The eating action (Clip A) with a longer length should be processed with the coarser-grained features by the coarse pathway. By comparison, the jumping event (Clip B) with far faster movement should be processed with the finer-grained features.

features are fed into our E-GCN described below to align with the query/intent.

5.2.2.2 Ego-Graph Convolutional Networks

The correlations between the different video segments and query/intent can be relatively sparse given a long video. For example, if the user queries *"walking"*, there can be only a fraction of video content correlated with walking people. Besides, the query-related video content can also have a sparse relationship with other video segments, especially those having a long temporal distance. Thus, correlating all the video segments (e.g., transformer-based models) can be time-inefficient and space-inefficient. We propose to exploit the notion of dynamic edge convolution [314] and construct a graph $\mathscr{G} = (\mathscr{V}, \mathscr{E})$ by connecting the video segments and query/intent dynamically. The graph's vertices $\mathscr{V}$ comprises of video segments extracted from the above GS-Pathways and the query/intent. To dynamically model the correlations between the video segments and the user intent, we connect them with the edge set $\mathscr{E}$ consisting of three types of edges, i.e., intent edge $\mathscr{E}_i$, semantic edge $\mathscr{E}_s$, and temporal edge $\mathscr{E}_t$.

Intent Edge connects the segment vertices with the centric intent vertex, which is why we call the graph as Ego-Graph. We map the intent embedding and segment feature into a mutual latent space dynamically by two MLPs. Then we can derive the intent-segment edge set $\mathscr{E}_z$ as,

$$\mathscr{E}_z = \{(w', w_t), w_t \in \mathscr{V}_T\}, \tag{5.6}$$

where $\mathscr{V}_T$ refers to the vertex set of the video segments and w' refers to the mapped query/intent vertex.

Semantic Edge connects the video segments with the correlated semantics. Motivated the sparsity of correlation, we follow [337] and connect the top-k related vertices for each video segment vertex in $\mathscr{V}_T$.

$$\mathscr{E}_s = \{(w_t, w_{n_t(k)})|t = 1,2,...,T; k = 1,2,...,K\}, \tag{5.7}$$

where $w_{n_t(k)}$ is the k^{th} nearest neighbor of the vertex w_t in the feature space and K is a constant number.

Temporal Edge connects the edges temporally adjacent. Each vertex has a forward edge to the next vertex and a backward edge to the last vertex except the two ends of the segment sequence. We represent the two sets of edges as:

$$\mathscr{E}_t^f = \{(w_t, w_{t+1}|t = 1,2,...,T-1\}, \tag{5.8}$$

$$\mathscr{E}_t^b = \{(w_t, w_{t-1}|t = 2,3,...,T\}, \tag{5.9}$$

where $\mathscr{E}_t^f$ includes the forward temporal edges, $\mathscr{E}_t^b$ includes the backward temporal edges and $\mathscr{E}_t = \mathscr{E}_t^b \cup \mathscr{E}_t^b$.

Edge Convolution After obtaining the graph, We apply edge convolution as our graph convolution operation[314]. Following Xu et al.[337], We employ the convolution operation to perform the efficient edge convolution on the obtained graph.

5.2.2.3 Local Graph for Shot-Level Features

The output features of edge convolution are at segment-level. To reconstruct the shot feature sequence from the segment features, We build the local Ego-Graph for each segment. The graph consists of one segment feature vertex connected with the all spanned shot vertices. We also add the semantic and temporal edges to the graph. After applying edge convolution on the constructed graph, we can obtain a shot-level feature sequence.

5.2.2.4 Implementation of the Modules

Both intent and summary modules are implemented based on GSE-GCN with different inputs and outputs. The summary module performs element-wise multiplication on the intent embedding and the Local-GCN-processed shot features to get a similarity vector. Then it exploits an MLP with Sigmoid activation to generate the selection probability of shots. By comparison, the intent module exploits an MLP head with Softmax to generate the intent distribution. Since the intent module is designed for the queries of different modalities, there is a slight difference between the visual-query and textual-query. The intent module for textual query strictly follows the GSE-GCN structure, while the intent module for the visual query models the query shots as individual vertices instead of one merged vertex.

5.3 Experiments

Method	Video-1			Video-2			Video-3			Video-4			Avg.		
	Pre.	Rec.	F-1	Pre.	Rec.	F-1	Pre.	Rec.	F-1	Pre.	Rec.	F-1	Pre.	Rec.	F-1
QC-DPP [266]	49.86	53.38	48.68	33.71	62.09	41.66	55.16	29.24	36.51	21.39	63.12	29.96	40.03	60.25	44.19
CHAN [333]	54.73	46.57	49.14	45.92	50.26	46.53	59.75	64.53	58.65	25.23	51.16	33.42	46.40	53.13	46.94
HVN [118]	52.55	52.91	51.45	38.66	62.70	47.49	60.28	62.58	61.08	26.27	54.21	35.47	44.57	**58.10**	48.87
QSAN [332]	48.41	52.34	48.52	46.51	51.36	46.64	56.78	61.14	56.93	30.54	46.90	34.25	45.56	52.94	46.59
Nalla et al. [204]	54.58	52.51	50.96	48.12	52.15	48.28	58.48	61.66	58.41	37.40	43.90	39.18	<u>49.64</u>	52.55	<u>49.20</u>
Ours	62.19	45.23	51.27	50.43	57.81	53.48	73.45	53.56	61.58	28.24	56.47	37.25	**53.58**	<u>53.27</u>	**50.90**

Table 5.2: Textual Query Dataset: Comparison with the previous state-of-the-art approaches.

Method	Video-1			Video-2			Video-3			Video-4			Avg.		
	Pre.	Rec.	F-1	Pre.	Rec.	F-1	Pre.	Rec.	F-1	Pre.	Rec.	F-1	Pre.	Rec.	F-1
Linear Baseline	59.24	45.33	49.75	21.49	26.71	23.62	56.09	44.42	49.22	14.44	33.1	19.77	37.82	37.39	35.59
Attention Baseline	45.01	33.96	37.71	38.86	48.01	41.09	57.7	48.75	50.66	18.00	41.5	24.75	39.89	43.06	38.55
Ours	58.17	44.91	49.43	42.52	52.69	46.64	65.45	51.92	57.49	21.15	49.23	29.19	**46.82**	**49.69**	**45.69**

Table 5.3: Visual Query Dataset: Comparison with the baselines.

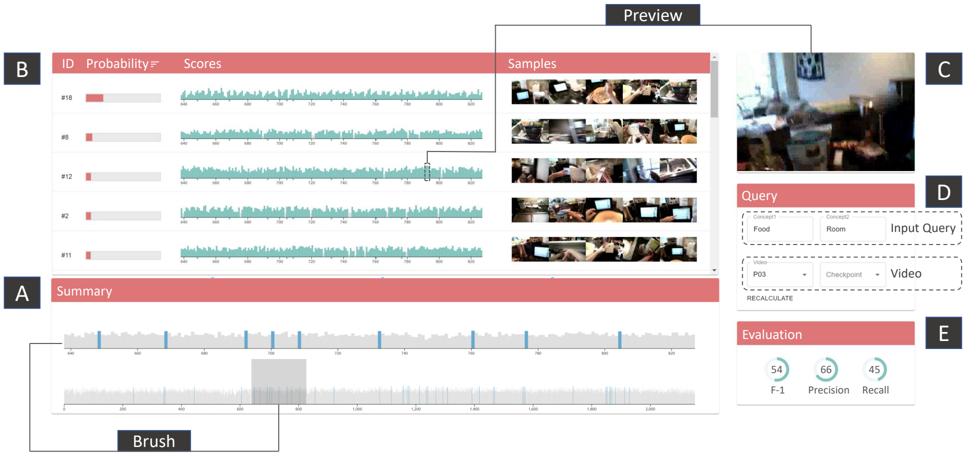


Figure 5.4: Prototype Overview. **A**: **Summary View** presents two temporal bar charts, which shows the overall scores and the summarized shots. The bottom bar chart shows the overview of all the shots, while the top bar chart zooms into the detail decided by the brush in the bottom chart. **B**: **Intent View** list all the basis intents with their probability, shot scores, and representative samples. The samples are selected with the highest score. **C**: **Preview View** plays a GIF of the user-hovering shot. In this case, the user hovers on the highlighted shot in intent #12, which includes a room scenario. **D**: **Query View** allows the user to change the query and makes the model rerun. **E Evaluation View** shows the quantitative result of the summary.

5.3.1 Implementation Detail

We exploit PyTorch [219] to implement our model on an NVIDIA RTX 8000. We have 20 basis intents, each represented by a 128-D embedding vector. For the summary module, we employ a 3-layer GCN and a 3-layer MLP after the GCN layers. We set the number of GCN layers and MLP layers as 2 and 3 for the intent module. Models are trained by an Adam optimizer with a base learning rate of 1e-4. We employ a warm-up strategy [89] to linearly increase the learning rate from 0 to the base learning rate in 10 epochs. After that, we reduce the learning rate to one-tenth of the previous value every twenty epochs.

5.3.2 Experiment Setting

5.3.2.1 Text Query Dataset

We conduct our text-query experiments on the query-driven video summarization dataset [266]. The dataset includes the four videos in UT Egocentric(UTE) dataset[154]. Each of the videos (3-5 hours) is captured in daily life scenarios. Each query in the dataset is represented by two concepts among the total 48 concepts.

5.3.2.2 Visual Query Dataset and Dataset Baselines

We build our visual-query dataset based on the text-query dataset. For each annotated summarization, we employ the eigenvector centrality as the criteria to pick the most representative shots as the query shots. Detailed examples and descriptions can be found in the supplementary materials. As the visual query dataset is newly introduced and no previous work exists, we compare our approach with two baselines, i.e., linear prediction and attentional query model, which can be founded in the supplementary.

5.3.2.3 Evaluation Protocol

To compare with the previous approaches fairly, we employ the semantic evaluation protocol proposed by Sharghi et al.[266]. This protocol is based on the semantic similarity between the machine-generated and the ground-truth video shots. The similarity is generated through finding the maximum weight matching of the bipartite graph computed based on the semantic interception-over-union (IOU). The summed weights of the matched shot pairs are used to compute precision, recall, and F-1 measure. Note that for the visual query dataset, we mask out the query shots in the evaluation stage. To evaluate the interactive intent, which involves the human subjectivity, we develop a prototype and propose a case study in qualitative analysis.

5.3.3 Comparative Analysis

The comparison among our method and SOTA methods for the Textual Query Task can be found in Table 5.2. We observe that our method achieves the highest

F-1 value of 50.90%. The result proves that our method can better identify the correlation between the query and summary. We also present the experimental result on the visual query task in Table 5.3. We find our method significantly outperforms the baselines by 7%, although the general performance is inferior to the performance on the text query task.

5.3.4 Ablation Analysis

We evaluate the effects of the proposed methods and find the optimum model setting by an ablation study.

5.3.4.1 Ego-Graph Convolutional Networks

I.M.	S.M.	Pre.	Rec.	F-1
Transformer	Transformer	44.82	44.52	42.68
Transformer	Ego-GCN	49.00	47.89	46.15
Ego-GCN	Transformer	47.09	47.26	44.75
Ego-GCN	Ego-GCN	**53.58**	**53.27**	**50.90**

Table 5.4: Ablation study the proposed Ego-GCN. I.M. refers to the intent module when S.M. refers to the summary module.

Our proposed Ego-GCN approach aligns the video segment features with the query/intent. To validate the effectiveness, we replace Ego-GCN by the transformer-based approach[279, 280, 379] in the intent and summary modules iteratively. The experiment results can be found in Table 5.4. Our Ego-GCN can advance the model performance when added to either intent or summary module.

5.3.4.2 Local GCN

Pathway	Pre.	Rec.	F-1
Upsampling	38.04	37.48	35.88
Transpose Conv	47.53	47.41	45.18
Local GCN	**53.58**	**53.27**	**50.90**

Table 5.5: Ablation study for Local GCN.

We employ Local GCN to recover the shot-level features from the segment-level features. As shown in Figure 5.5, Local GCN's performance is superior to bi-cubic upsampling and transpose convolutional layer, which is used in [333].

5.3.4.3 GS-Pathway

Pathway	Pre.	Rec.	F-1
Shot-Level Feature	47.45	45.38	44.40
Coarse-Granularity	47.40	47.66	45.15
Fine-Granularity	50.18	50.23	47.81
Full-Model	**53.58**	**53.27**	**50.90**

Table 5.6: Ablation study for the GS-Pathway. The pathway *Shot-Level Feature* refers to the model that directly applies our Ego-GCN on the shot-level video features.

To validate the effects of our GS-Pathway, we compare our model with three variants with only one fixed pathway. We present the experiment results in Table 5.6. Our result shows our model surpasses the three variants, verifying the necessity of attending the segment features of multi-granularity.

5.3.4.4 When to fusion the pathways?

Stage	Pre.	Rec.	F-1
Early	**53.58**	**53.27**	**50.90**
Middle	49.53	48.66	46.66
Late	47.69	47.98	45.47

Table 5.7: Experiment result on the feature fusion stage.

The features of different pathways can fuse at different stages. To find the optimum of the model, we compare the variants with different fusion strategies, i.e, early, middle, late fusions. The early fusion strategy fusions the features before the dot product shown in Fig. 5.2. The middle fusion and late fusion happen before and after the MLP, respectively. As can be found in Table 5.7, fusion at the early stage is the best choice.

5.3.4.5 Do we need video as input of the intent modules?

Intent Module	Pre.	Rec.	F-1
Video Agnostic	50.06	48.78	47.15
Query Attention	49.26	47.85	46.27
Full Model	**53.58**	**53.27**	**50.90**

Table 5.8: Experiments on the video inputs in the intent module.

Our intent module use both the query and video as input to predict the user intent. However, the user intent can also solely rely on the user query, as some users might not have time to browse the original videos. Thus, it is necessary to learn if we can remove the video inputs from the intent module and let it infer only based on the user query. To do so, we compare the full model with two variants using a simpler intent module and a video-agnostic intent module. We present the experiment result in Table 5.8. Though there is a marginal performance decrease, the model with a video-agnostic intent module still outperforms most of the previous state-of-the-art approaches. The result shows it is reasonable to remove the video input for the intent module to promote the model's generalizability.

5.3.4.6 Can we transfer the summary module between different datasets?

Training	Pre.	Rec.	F-1
Canonical	46.82	49.69	45.69
Transferring	**47.15**	**51.08**	**46.40**

Table 5.9: Experiment on visual query task with transfer learning.

To validate the generality of the summary module, we experiment on the visual-query task in the transferring setting. We first train the summary module on the text-query dataset. Then, we reuse this pre-trained summary module and only train the intent module for the visual-query task. The results can be found in Table 5.9. The experiment result shows that the transferred model surpasses the model trained in the canonical setting, showing that the summary module is interchangeable for the queries of multi-modality.

5.3.5 Prototype and Qualitative Analysis

We demonstrate the interactivity of our framework with a prototype as shown in Fig. 5.4. The prototype can also work as qualitative analysis, to prove that our approach can generate the query-related summary with better interpretability. In the figure, we show an example case. The snapshot is taken when the user queries "Food" and "Room" for video-3. Note here we set that the user input is always two queries in our design of the prototype, although the proposed framework can deal with other numbers of queries. The user first brushes on the result view and focuses on the clip where more shots are captured in the summary. Based on the Intent View (B), we can identify intent #18, #8, #12, #2, and #11 in descending order. From the samples of each intent, we find the #intent #18 and #8 are closely related to the food cooking scenarios when #11 contains some food storage scenarios. The #12 and #2 are more likely to focus on the room scenarios. We also observe that there are some computer frames in #12 and #2. Previewing the related shot, we notice that the computer is the foreground object of the room, as Fig. 5.4. C shows. Thus, the snapshot shows that our model successfully captures the food and room scenarios. We can find that there are two types of food scenarios identified, i.e., food cooking and food storage. This finding also shows that our approach can provide finer-grained user intent representation.

5.4 Final Considerations

In this chapter, we propose IntentVizor, an interactive video summarization framework guided by the generic query. First, our framework introduces a novel concept "intent", which originally comes from the Information Retrieval (IR) community, to represent the multi-modality queries. Second, we develop a prototype to make the proposed framework interactive with the user. The user can control the intent to generate summaries satisfying their needs. Third, for the model part, two novel intent/summary modules are designed to better understand the generic queries and generate summaries accordingly/adaptively. Both quantitative and qualitative experiment results verify the superiority of our proposed approach. Four ablation studies also verify more potential extensibility of the proposed framework. For future work, we will solve the limitation above, and introduce more query modalities to better satisfy users' video summarization needs. IntentVizor serves as the foundation for the proposed work in Chapter 6, where the interactive summarization process is further employed and expanded.

Chapter 6

InsightAR: A Tool for Multimodal Summarization and Interactive Analysis of AR-based Egocentric Task Videos

6.1 Introduction

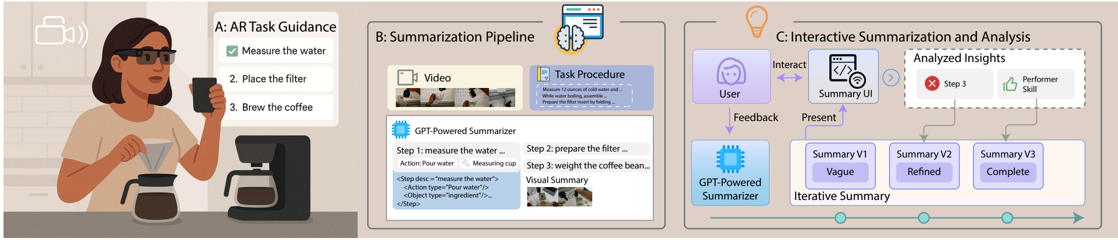

A: AR Task Guidance
☑ Measure the water
2. Place the filter
3. Brew the coffee
B: Summarization Pipeline
Video
Task Procedure
Measure 12 ounces of cold water and ...
While water boiling, assemble ...
Prepare the filter insert by folding ...
GPT-Powered Summarizer
Step 1: measure the water ...
Action: Pour water
Measuring cup
<Step desc = "measure the water">
<Action type="Pour water"/>
<Object type="ingredient"/>...
</Step>
Step 2: prepare the filter ...
Step 3: weight the coffee bean...
Visual Summary
C: Interactive Summarization and Analysis
User
Interact
Summary UI
Analyzed Insights
Step 3
Performer Skill
Feedback
Present
GPT-Powered Summarizer
Summary V1
Vague
Summary V2
Refined
Summary V3
Complete
Iterative Summary

Figure 6.1: InsightAR, a visual analytics tool for AR-based egocentric task video analysis. (A) The system captures real-world task guidance in AR, where users follow step-by-step instructions during hands-on activities. (B) The summarization pipeline integrates video understanding with language models to extract structured procedures and generate visual and textual summaries of task steps. (C) The interactive summarization and analysis interface enables users to iteratively refine summaries, identify key insights (e.g., errors or skills), and compare performance over multiple iterations.

The growing adoption of head-mounted devices such as wearable cameras [282], smart glasses [359], and augmented reality (AR) headsets [218, 305], has made egocentric video recording an increasingly popular method for capturing task performance. These recordings are used to document daily tasks, such as cooking [193] and driving [102, 186], as well as domain-specific tasks, including surgical operations [150, 192, 253] and industrial assembly [261]. By offering a first-person perspective of users' interactions with their environment and tools—such as surgical instruments, or manufacturing equipment—egocentric videos can help users identify their performance errors and potential areas for improvement. For instance, in auto repair, inspectors can use AR glasses to record machine inspections, which are later analyzed to detect execution errors [64, 262].

Despite the surge in egocentric recordings across various domains, analyzing them remains a labor-intensive and error-prone manual process due to the lack of tools tailored for task-oriented egocentric videos. Most existing systems are designed for third-person footage and fail to address challenges unique to egocentric videos—such as constant camera motion, abrupt viewpoint shifts, and frequent occlusions of manipulated objects. General-purpose visual analytics tools [260] are also limited when adapting to vision calibration or shifting perspectives within egocentric videos. While recent advances in computer vision have improved low-level video understanding (such as saliency, optical flow, and action recognition) of egocentric videos [152, 262], they remain insufficient for deeper contextual analysis of complex tasks, such as deviations from standard procedures. Egocentric videos uniquely capture fine-grained human-object interactions critical for training and assessment, but current AI-based tools—often built on non-egocentric data—do not capture the nuanced interaction richness of task execution [309]. Furthermore, these tools focus on broad analysis like user emotions or overall performance [97, 355], rather than assessing procedural adherence and errors critical for domain-specific insights. Moreover, existing video summarization and multimoda analysis tools either lack interactivity or fail to preserve the semantic structure of tasks [96, 115]. To our knowledge, there are no interactive visual analytics tools to support summarization and multimodal analysis of egocentric videos in task-oriented contexts, pointing to a critical gap in current research.

We bridge this gap through InsightAR, a visual tool for analyzing egocentric task videos using multimodal summarization and interactive analysis shown in Figure 6.1. To understand the design requirements for such a tool, we first conducted a formative study with 4 domain experts, including cooking instructors, surgical training supervisors, and industrial process analysts. Through semi-structured interviews, we explored their workflows and identified their needs for analyzing task-performance videos. Our findings revealed that users need tools that can: (a) automatically extract and summarize key events and highlights from task recordings, (b) align video content with task-specific domain knowledge, and (c) support interactive analysis of multimodal summaries to generate actionable insights.

Based on the findings from the formative study, we developed InsightAR, which has two main components - *multimodal summarization* and *interactive analysis*. To generate multimodal task video summaries, users record their tasks using AR or smart glasses and upload the footage to InsightAR. The system processes the data through three analytical pipelines, each operating at a different level of granularity. First, an **overview-level summary** provides a visual synopsis with key shots and a textual narrative, generated using a contrastive learning-based model (trained on the SUMME and Epic-Kitchen datasets [98, 123]), enabling users to understand the task flow without watching the entire video. Second, a **step-level summary** is created using the same model at a higher frame rate to capture key moments in greater detail, using a large language model (LLM) to generate captions for each keyframe and highlight mistakes or provide suggestions for improvement. Third, a **timeline summary** is generated that identifies key events, actions, errors, and anomalies using an ensemble of vision models (including BLIP [165]), to assist users to detect workflow inefficiencies and deviations from optimal performance. These summaries are presented in a web-based interface with an interactive analytical tool. Users can use this tool to study their performance through three interconnected views: the **summary view** that enables quick navigation through key segments by aligning visual highlights with textual descriptions and metadata; the **task knowledge graph** that supports comparison of task execution against standard procedures, automatically detecting discrepancies, errors, or missed steps; and the **annotation view** lets users to document insights, create time-linked notes, and compile observations for training or performance improvement.

To evaluate InsightAR's effectiveness, we conducted two case studies in cooking and medical training domains, along with a user study involving 16 participants. These evaluations demonstrate how users can identify procedural improvements and performance gaps through our system. We also performed a technical evaluation to assess the accuracy of our generated summaries. An overview of our approach and visual analytics components is illustrated in Figure 6.1.

In summary, our main contributions in this chapter are:

1. InsightAR, a tool for summarizing and analyzing egocentric task recordings. InsightAR generates multimodal video summaries by combining computer vision and multimodal LLM, and has an interactive interface for users to study and refine these summaries with varied granularity of insights.

2. A formative study with domain experts. We conducted semi-structured interviews with experts in cooking, surgical training, and industrial workflows to inform the design of InsightAR and identify key requirements for analyzing egocentric task videos.

3. Two case studies in medical and aviation domains showing the generalizability and application of our approach across distinct task-driven environments.

4. A user study with 16 participants. We evaluate the usability and effectiveness of InsightAR in supporting task analysis, procedural review, and insight generation

6.2 Formative Study

To guide the design of our egocentric task-video analysis tool, we conducted a formative study involving interviews with four domain experts (**E1**–**E4**). The study aimed to:

G1: Understand the current manual and automated workflows used by experts to analyze task recordings.

G2: Explore how multimodal summaries can support task analysis, and identify what information and strategies are needed to incorporate domain knowledge.

G3: Identify key features and capabilities that would make a task analysis system more effective for domain experts.

Expert Participants. E1 has six years of experience in AR-based pilot training and aerospace software development. **E2** is an AR task guidance developer focused on medical training, with expertise in task recording analysis. **E3** and **E4** are AR/VR researchers with top-tier publications and extensive experience in AR user behavior, task guidance, and recording analysis.

Study Protocol. Each interview lasted 75 minutes and was divided into three phases: (1) a 15-minute familiarization phase to discuss their workflows and frequency of task recording analysis, (2) a 45-minute interview to identify domain-specific analysis needs, including preferred summaries and autogenerated insights, and (3) a 10-minute brainstorming session where experts sketched their ideal interface for video task analysis.

6.2.1 Analysis:

We synthesized expert feedback into the following key insights that shaped the design requirements of our system:

6.2.1.1 Analyzing egocentric task videos is essential for designing task guidance systems, yet existing workflows are often labor-intensive:

All experts, with prior experience in AR-based task guidance, recognized the value of AR task recordings for post-hoc performance analysis and skill improvement. **E2** noted that *"military personnel [can] practice medical skills [...] with HoloLens and can watch their own recordings to improve the skills."* and further emphasized that *"looking into the made mistakes and unconscious actions, there are a lot of lessons can be learned [by the users]."* This aligns with prior research showing task recordings help extract lessons and enhance performance [150, 186]. However, the process remains labor-intensive, requiring manual review, segmentation, and annotation, as **E3** noted, *"[currently], we have to manually watch the videos and check every step [in the video]."*

6.2.1.2 Auto-summarization is useful for analysis and should provide multi-modal and multi-granular analysis:

Experts emphasized the inefficiency of manually reviewing lengthy task recordings, underscoring the need for automated, structured summarization to support analysis. **E2** noted that *"trainers may spend hours reviewing trainees' recordings, which is not sustainable (when the task scales up)."* When we suggested video summarization techniques as an option, experts confirmed that traditional summarization methods, which extract keyframes based solely on representativeness, are insufficient. Instead, they stressed the importance of a multimodal approach that *integrates key visual frames* (i.e., key video frames showing critical actions) with *textual descriptions* (explaining what is happening in the frames) to clarify both *what* happened and *why* it matters. **E4** stated that *"seeing key moments alongside explanations of what should be happening at each step helps quickly identify where trainees deviated from protocol."* Experts also noted that a single summary for the entire video is inadequate for structured task recordings divided into sequential steps. **E3** explained that *"the [task] recording is naturally segmented [in] different steps and the video-level summary may not be the best solution"*, pointing to the need for both overview and step-level summaries. Overview summaries should cover the entire process, highlighting key or erroneous steps, while step-level summaries should include detailed analysis of actions, objects, and hand movements [323]. **E4** reinforced this need for detailed step-level information, noting that *"there are a lot of features we can use for the step-level summary, such as the actions and events."* Overall, expert feedback underscored that egocentric task recordings required a specialized summarization approach that integrated visual and textual content at multiple levels of granularity, a capability lacking in generic video summarization systems.

6.2.1.3 Analyzing time performance and task adherence for skill evaluation and error identification:

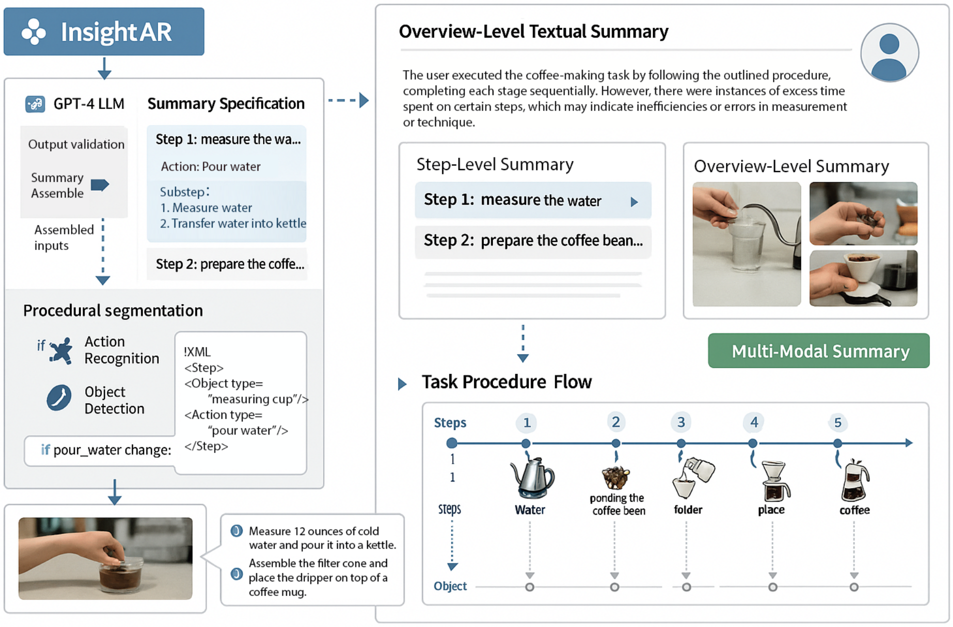
Our interviews revealed that time performance evaluation is crucial for assessing skill levels in egocentric task recordings, with experts regularly analyzing time spent on actions, tool use, and procedural steps. They stressed the need for effective visualization of time metrics, such as a dashboard, which **E4** described as *"a common practice [...] to evaluate the person's skill with the particular action"*. Task adherence is also key to evaluating task completion, with deviations signaling potential mistakes. **E1** emphasized that failure to comply with aviation checklists *"may cause catastrophic consequences."* While basic adherence analysis can identify obvious errors, experts noted it is inadequate for complex tasks with flexible ordering, where *"the steps for a task may be arranged in different orders"* (**E2**). This flexibility makes linear task representations inadequate for proper error detection. In **E2**'s example of emergency medical training, proper tourniquet preparation is a prerequisite for its application. This highlights the need for visualizations that clearly represent prerequisite relationships between task steps to ensure accurate task execution.

6.2.1.4 Seamless toggle between video content and procedural information required for efficient analysis:

Experts struggled to assess task step adherence due to the disconnect between video content and procedural documentation, often relying on manual cross-referencing—a process described as *"frequently switching between the procedural book and the video player"* (**E3**). To address this, they requested a system displaying video and task procedures side-by-side with clear visual links between segments, steps, actions, and objects. They emphasized interactive navigation—selecting procedural elements to jump to relevant video parts, and vice versa. **E2** highlighted the need to examine common error points by clicking on procedural steps: *"there are some places where users easily made mistakes and we need to quickly locate them [from the task manual]."* Likewise, **E3** emphasized verifying preconditions via backtracking: *"check if pre-conditions are fulfilled by clicking on the previous step of the one currently being analyzed."* These insights point to the need for an interactive system that tightly integrates procedural knowledge and video for efficient error detection and task validation.

6.2.1.5 Notes-taking and comparing recordings:

When asked about additional features for analysis, **E4** suggested a notebook or annotation panel for recording insights, while **E2** proposed a comparative view to



```
!XML
<Step>
<Object type=
    "measuring cup"/>
<Action type=
    "pour water"/>
</Step>
```

The user executed the coffee-making task by following the outlined procedure, completing each stage sequentially. However, there were instances of excess time spent on certain steps, which may indicate inefficiencies or errors in measurement or technique.

1 Measure 12 ounces of cold water and pour it into a kettle.

2 Assemble the filter cone and place the dripper on top of a coffee mug.

Figure 6.2: Workflow of InsightAR, illustrating the integration of backend processing and the user interface. The backend pipeline (left) performs procedural segmentation of egocentric AR task videos using action recognition and object detection, which are then structured by a summarizer developed on GPT4 into step-wise descriptions with substeps and XML outputs. The processed data is visualized in the user interface (right), which includes: an overview-level summary combining textual and visual highlights; step-level summaries for detailed task breakdown; a multi-modal summary that integrates visual and procedural information; and a task procedure flow that traces step dependencies and object transitions across the workflow. These linked views support efficient, interpretable analysis of complex AR task performances.

analyze multiple recordings, such as comparing a user's performance with experienced practitioners. **E1** emphasized the importance of task recordings for optimizing procedures, noting that task steps may need reordering based on action dependencies, like adding hot water before a tea bag. This feedback emphasizes the need for tools that visualize task dependencies and enable inter-video comparisons for procedural optimization.

6.2.1.6 Human-AI collaborative summarization for accurate insights:

Machine-generated video summaries often focus on low-level information and can be inaccurate, especially for egocentric task recordings that require domain expertise for higher-level insights. Our interviews revealed a need for a process that combines automated analysis with human input. **E3** emphasized the importance of allowing *"users have control on the summarized result"*, and both **E2** and **E3** emphasized integrating user-generated annotations with system-generated summaries for a more comprehensive analysis. Experts agreed that a collaborative approach between automated systems and human experts could produce more accurate and actionable insights than either could provide independently.

6.2.2 Design Requirements

Based on the above insights, we inferred that our tool should:

DR1 *Provide multi-modal and multi-granularity summaries.* To allow users to efficiently analyze lengthy task recordings, our system should generate visual summaries (keyframes) and textual descriptions at different levels: an overview-level summary of the entire workflow and a step-level summary that breaks down the video into specific actions and objects. This would let users quickly identify key segments, spot procedural errors, and check detailed steps if needed.

DR2 *Synchronize video content with task procedures through bidirectional navigation.* To allow users to efficiently identify procedural errors, verify step completion, and understand action sequences without manually correlating video content with the task document, the system should synchronize video content with task procedures through bidirectional navigation between video segments to their corresponding steps. This feature could map video segments to their corresponding steps allowing users to click it to highlight the related procedural step, and vice versa.

DR3 *Support iterative summarization through the human-LLM collaboration.* To allow users to refine task analysis, the system should dynamically incorporate user annotations into existing video summaries. The system should enable users to provide high-level insights, such as detecting errors and potential task improvements, which are used to re-summarize the video content. If the initial summary contains inaccuracies, user annotations on specific actions, steps, and objects should update the summary. This interactive approach could enhance the accuracy of the summaries by combining automated processing with user-driven insights.

These design requirements guided the development of our tool, InsightAR, which we detail in the following section.

6.3 InsightAR: A Visual Analytics Tool

We present InsightAR, a visual analytics tool for egocentric task recordings embedded with two core functionalities - multi-modal summarization and interactive analysis. This section outlines InsightAR's core components, workflow, and key UI features.

6.3.1 System Overview:

6.3.1.1 Multi-modal Summarization of Task Videos:

To generate multi-modal task video summaries, users record their tasks using AR or smart glasses and upload the footage to InsightAR, which processes it through three analytical pipelines, each offering summaries at different levels of granularity. (a) An overview-level summary is generated as a visual synopsis with key shots and a textual narrative, allowing users to grasp the task flow without watching the full video. The workflow and UI overview are presented in Figure 6.2. For this summary, we use a contrastive learning-based video summarization model trained on the SUMME and Epic-Kitchen datasets [98, 123]. (b) A step-level summary generated using the same model at a higher frame rate to extract key moments with greater detail. An LLM then generates captions for each keyframe and presents a breakdown of the task steps to highlight steps with mistakes and identify areas for improvement. (c) A timeline summary is generated with details about key events, actions, errors, or unforced anomalies throughout the video to help users identify workflow inefficiencies and exact moments of deviations from optimal performance. This summary is generated using an ensemble of vision models, including BLIP [165]. These multi-modal summaries are then presented in a web-based portal along with the analysis module for users to review their performance further (DR1)

6.3.1.2 Interactive Analysis of Summaries:

For further analysis and enabling the requirements from Section DR3, users can use InsightAR's web-based interactive module using three interconnected views. Using the summary view, users can quickly identify important segments, understand their context within the task, and efficiently navigate through specific moments without watching the entire recording. The system facilitates this exploration by visually aligning key moments with their textual descriptions and relevant metadata. Next, using a task knowledge graph, users can compare their task execution against the order of steps and actions of standard procedures to identify discrepancies, errors, or missed steps (DR2). For example, when analyzing a coffee-making task, users can immediately spot if they attempted to pour water without first placing a cup, as the system highlights discrepancies between the expected and observed task flow. Finally, using an annotation view, users can document their observations and generate insights. As they identify areas for improvement, users create and edit notes that the system automatically synchronizes with the video timeline. Using this view, users can record observations to review later or share with others.

6.3.2 UI Walkthrough and Features:

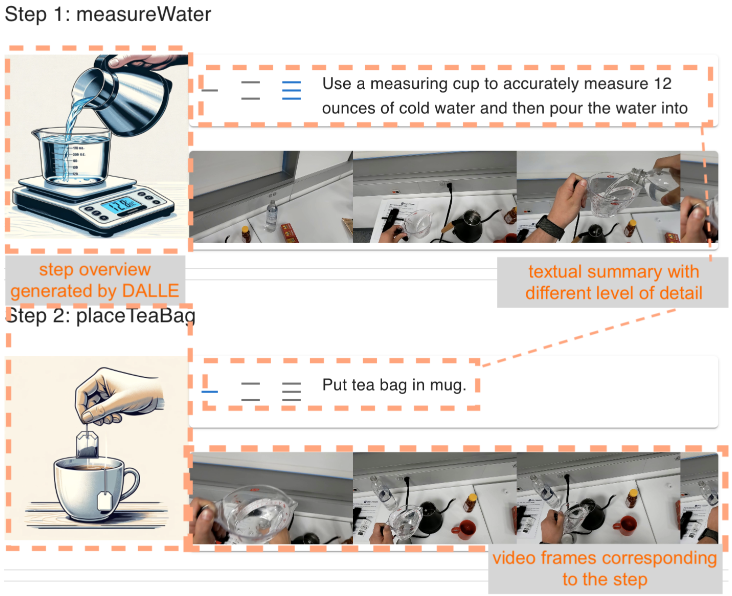
To demonstrate how InsightAR supports egocentric video analysis, we present a typical workflow using a pour-over coffee-making recording as our example.

6.3.2.1 Loading and Processing a Recording:

To start, users can upload their egocentric video recordings to InsightAR's user interface. For our coffee-making example, the user uploads a ∼10-minute recording captured via HoloLens 2 showing the complete pour-over coffee preparation process. InsightAR automatically processes the video through our multi-modal summarization pipeline, extracting actions, objects, and segmenting the recording into distinct steps.

6.3.2.2 Exploring the Multi-Modal Summary:

After the processing is completed, InsightAR presents the user with the multi-modal summary view shown in Figure 6.3. For our coffee example, this view organizes the task into seven key steps, from measuring water to the final pour-over coffee grounds. Each step provides an AI-generated illustration of the expected action and the representative keyframes extracted from that step. The interface also shows the object and action tags (such as “measuring cup” and “pour water”) along with a textual description of the step that can be adjusted for detail level. The user can toggle between overview and step-level summaries using the toggle at the top of the interface, allowing them to either grasp the entire coffee-making process quickly or examine each step in detail.



Step 1: measureWater

Use a measuring cup to accurately measure 12 ounces of cold water and then pour the water into

step overview generated by DALLE

textual summary with different level of detail

Step 2: placeTeaBag

Put tea bag in mug.

video frames corresponding to the step

Figure 6.3: Step-level summary shows the step-based video summaries of both textual and visual modalities. The user can switch between this and the overview-level summary.

6.3.2.3 Analyzing Task Flow:

To further understand step dependencies with the task, the user can switch to the Task Flow View shown in Figure 6.4. This visualization reveals how objects like the kettle, filter, and coffee grounds update states through different steps throughout the task. The user can observe that our system updates the kettle's state from "unprepared" to "prepared" when filled with water, and finally to "used" during the pouring action. The timeline view shows precise timestamps of the steps, for example, revealing that heating water consumed the most time (45 seconds) in the process. When the user clicks on a specific step node (e.g., "fold paper filter"), the system highlights related components and automatically displays the video player for playing the corresponding segment, allowing the user to further verify the proper technique of that step.

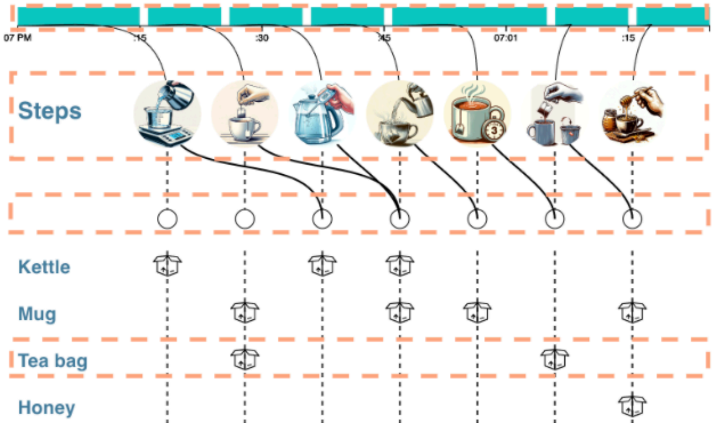


Figure 6.4: **Task Flow View** illustrates the dependencies between procedural steps as defined by the states of objects (unprepared, prepared, and used). Steps are represented by illustrations generated by DALLE-2. Lines connecting the nodes indicate dependencies between the steps. The diagram reflects the state of each object at the corresponding step. The duration of each step is displayed on the timeline above.

6.3.2.4 Reviewing Timeline Details:

For a chronological analysis of the steps, the user can explore the Video Timeline View shown in Figure 6.5. This view breaks down the coffee-making process into consecutive three-second segments with aligned events, actions, and objects. If the user makes an error, for example, pour water before checking the water temperature, then they are shown an error marker in the segment of that erroneous step. By clicking this segment, they're taken directly to that point in the video to observe their mistake. In this way, they can easily identify missteps, errors, and opportunities for improvement.

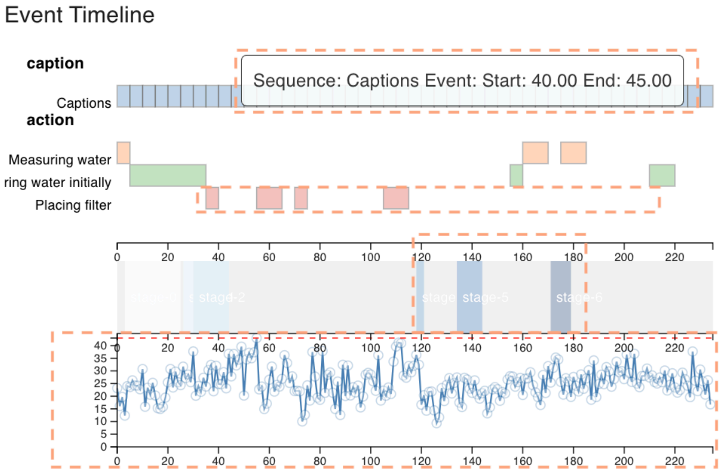
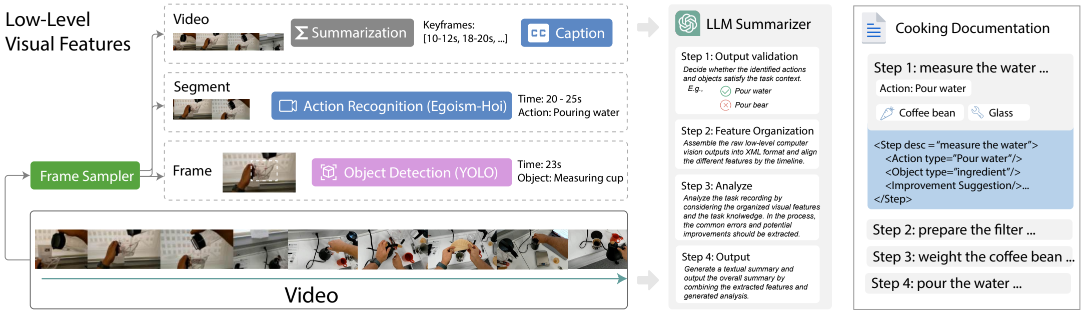


Figure 6.5: Video event timelines for chronological analysis show the captions of each video segment of ∼3 seconds and the associated events, objects, and errors.

6.3.2.5 Interactive Features Across Views

During analysis, the user can benefit from synchronized interactions across all views. When they identify a potential issue in the timeline (for example, spending too much time folding the filter), they can click on the corresponding segment to view the video evidence, reference the knowledge graph to verify if this step follows the correct sequence, compare their performance against reference recordings from expert baristas, and add annotations to document insights for future improvement. This integrated approach allows the user to efficiently identify areas for improvement in their coffee-making technique without watching the entire recording multiple times.



Step 1: Output validation — Decide whether the identified actions and objects satisfy the task context.

Step 2: Feature Organization — Assemble the raw low-level computer vision outputs into XML format and align the different features by the timeline.

Step 3: Analyze — Analyze the task recording by considering the organized visual features and the task knolwedge. In the process, the common errors and potential improvements should be extracted.

Step 4: Output — Generate a textual summary and output the overall summary by combining the extracted features and generated analysis.

```
<Step desc = “measure the water”>
   <Action type=”Pour water”/>
   <Object type=”ingredient”/>
   <Improvement Suggestion/>...
</Step>
```

Figure 6.6: Backend pipeline of InsightAR for multi-modal summarization. The system begins with a frame sampler that extracts representative frames from the input egocentric video. These are analyzed at multiple levels: full video clips for summarization and captioning, segments for action recognition (e.g., Egoism-HOI), and individual frames for object detection (e.g., YOLO). The extracted low-level visual features are passed to a LLM-based summarizer, which performs: (1) output validation to ensure context-appropriate recognition, (2) feature organization into a timeline-aligned XML structure, (3) analytical reasoning to identify key steps, errors, and improvements, and (4) generation of structured task documentation combining textual, visual, and semantic insights.

6.4 InsightAR's Backend Processing Pipeline

This section describes our machine learning-based pipeline for generating summaries from egocentric videos: extracting low-level features (actions, objects, and segments), using contrastive learning for visual summarization, and integrating these with LLM-generated text into a unified, interactive multi-modal summary.

6.4.1 Extracting Low-level Visual Features

6.4.1.1 Event, Action and Object Detection:

Figure 6.6 gives the overview of our pipeline to support **G1**, where we extract the action, object, event, and dense captions using computer vision. As the resulting outputs can be noisy due to imperfect CV models, we use a GPT-4 to analyze and correct predictions that contradict common sense.

6.4.1.2 Action Recognition:

Egocentric tasks involve unique actions not well-covered by datasets like Kinetics-400 [39, 122] and Activity-Net [358]. Instead, we leverage few-shot or zero-shot action recognition with custom labels via vision-language models [79, 247]. We use EgoVLP as our base model for action recognition, as it is pre-trained on egocentric datasets and learns joint text-video representations via contrastive learning [172]. Our system are tuned on a GPT4 model to generate candidate actions based on the task procedure book [294], cooking and assembling(Wikihow[1]).

6.4.1.3 Human Object Interaction:

The interaction between the task performer and object is important for the performance evaluation [286]. To identify the object that the performer interacts with, we incorporate Egoism-HOI, a model designed to discern and categorize objects that users interact with [155][2]. Egoism-HOI is trained to detect the objects in the egocentric videos.

6.4.1.4 Procedural Segmentation:

Procedural segmentation involves partitioning videos into distinct segments corresponding to specific sub-tasks, to ensure that the summary reflects the logical sequence of actions. We achieve this by using both rule-based step prediction and pre-trained vision-language models (VLMs). While the rule-based method offers robustness, it lacks generalizability. In contrast, VLMs, such as EILEV [352], provide better generalizability but are less robust. We combine rule-based approaches with a state-of-the-art VLM EILEV [352] for better segmentation results. EILEV takes input from the combinations of video snippets and textual queries to understand specific procedural steps. Using a sliding window approach, this model processes queries like *"Does the step [STEP_DESCRIPTION] finish in the video?"*, updating the placeholder with each task description. EILEV can accurately identify the boundaries between different steps of the procedure, ensureing alignment

[1] https://www.wikihow.com/Main-Page

[2] Egoism-HOI is the name of the proposed data in [155]. Since there is no explicit name for the model proposed, we refer Egoism-HOI to the model.

of each video segment task steps and improving the clarity and usability of the video summary.

6.4.1.5 Textual Summary:

Taking the outputs from action recognition, human-object interaction detection, and procedural segmentation, we then use the GPT-4 model to compile a structured document of the recording. As these outputs may contain noise or inaccuracies, especially when the recording data deviates from the models' training domains, we first prompt GPT-4 to validate the results against common sense. For example, if the action recognition model misidentifies *pour water* as *pour beer* in a coffee-making task, GPT-4 corrects the error using its extensive knowledge base. After validation, GPT-4 organizes the refined results into an XML-formatted document, structuring the video into distinct procedural steps while cataloging relevant actions and objects for each phase of the task (Figure 6.6).

6.4.2 Generating Multi-Granular Video Summaries

To support G2, users need video summaries to quickly skim recordings. Traditional video summarization selects representative frames from the full video using a summarizer ϕ,

$$s := \{f \in v\} = \phi(v), v = \{f_1, f_2, \cdots, f_T\} \tag{6.1}$$

However, to meet DR1, we require segment-specific summaries that highlight contrastive aspects of a given step compared to the rest of the video. For a query segment $q = \{f_{t_a}, \cdots f_{t_b}\}$ defined by the segment range $\{t_a, t_b\}$, summarization is reformulated as:

$$s = \{f_t | f_t \in v, t \geq a, t \leq b, \} = \phi(v, t_a, t_b), \tag{6.2}$$

Conventional methods lack this contextual focus, failing to emphasize what makes a segment unique.

To ensure both the *representativeness* and *contrastiveness* of the generated video summary, we introduce **a novel self-supervised video summarization**

framework that utilizes adversarial reconstruction and contrastive learning. This framework employs an adversarial approach to train a summarizer that can generate a summary capturing the information essential to reconstruct the original video [184, 324]. Thus, the generated summary is ensured to be representative. Contrastive learning aims to learn the representation by contrasting positive and negative pairs of samples. Positive pairs are constructed using in-query frames and the generated summary s, while negative pairs are formed by contrasting the generated summaries with frames sampled from the remainder of the video.

6.4.2.1 Adversarial Learning:

We adopt an adversarial training approach for the video summarizer, following prior work [184, 324]. Each video frame f_t is encoded via ResNet-50 to obtain embeddings r_t. A query segment q is encoded using an LSTM encoder $\psi_{encoder}$ to get e_q. The summarizer ϕ generates a summary s, which is then passed through an LSTM-based reconstructor ψ_{recon} to produce a reconstructed embedding e_r. A discriminator $\psi_{discriminator}$, also an LSTM, distinguishes between real (encoded) and reconstructed embeddings: $y_r = \psi_{discriminator}(e_r), y_q = \psi_{discriminator}(u_q)$, and the GAN loss is:

$$\mathcal{L}_{GAN} = log y_q - log(1 - y_r). \tag{6.3}$$

We train the discriminator to distinguish encoded from reconstructed features by minimizing $-\mathcal{L}_{GAN}$, and simultaneously train the encoder, reconstructor, and summarizer to minimize $\mathcal{L}_{GAN}$, encouraging the summarizer to generate representative summaries whose reconstructions are indistinguishable from real features.

6.4.2.2 Contrastive Learning:

Our contrastive samples are generated by sampling the query segment q and the rest of the video, which is not in q. We re-use the feature encoder $\psi_{encoder}$ to generate the features of sampled frames. We use a uniform random sample function $h(.)$ to select the frames randomly from a video segment.

- $e_{\text{pos}} = \psi_{encoder}(h(q))$ for an embedding of a positive sample (a sub-segment within the query segment q).

- $e_{\text{neg}} = \psi_{encoder}(h(v \setminus q))$ for an embedding of a negative sample (a sub-segment from outside Q).
- $e_{\text{anchor}} = \psi_{encoder}(s)$ for an embedding of the anchor point, which is the embedding of the generated summary in our problem.

The similarity measure between two embeddings e_i and e_j is defined as the cosine similarity:

$$\text{sim}(e_i, e_j) = \frac{e_i \cdot e_j}{\|e_i\|\|e_j\|} \tag{6.4}$$

The model aims to make the anchor closer to its positive samples than to any negative sample, defining a contrastive learning scenario.

$$\mathcal{L}_{sim} = -\log \frac{\exp(\text{sim}(e_{\text{anchor}}, e_{\text{pos}})/\tau)}{\exp(\text{sim}(e_{\text{anchor}}, e_{\text{pos}})/\tau) + \sum_{n=1}^{N} \exp(\text{sim}(e_{\text{anchor}}, e_{\text{neg}}^{(n)})/\tau)} \tag{6.5}$$

where τ represents a temperature parameter that scales the similarities, and N is the number of negative samples.

We can optimize the summarizer by combining $\mathcal{L}_{\text{sim}}$ and $\mathcal{L}_{\text{GAN}}$. We implement our models using PyTorch. The model is trained using the Adam optimizer with a learning rate of 1×10^{-3} and a weight decay of 5×10^{-4}. We set $\tau = 0.3$ and $N = 8$ in the contrastive learning setting.

6.4.3 Combining LLM-Based Summarization with User Feedback

InsightAR allows users to edit identified events, actions, and descriptions via a visual interface. These modifications are sent back to the LLM, which regenerates the summary incorporating user feedback, enabling domain experts to refine automated analyses. To improve reasoning capabilities, InsightAR integrates chain-of-thought prompting techniques with the LLM [319] and implements tool-usage functionality that allows GPT to call InsightAR's backend APIs for extracting actions from specific video ranges [167]. For example, if the summarizer detects a potential discrepancy—such as a missing object in a required step—it can query the object detection model to verify the observation. By combining the semantic understanding of LLMs with the perceptual strengths of vision models, InsightAR ensures increasingly accurate multi-modal summaries through iterative human-AI collaboration.

6.4.4 Technical Implementation

InsightAR uses a containerized architecture with Docker for consistent deployment across environments. We built the backend with Flask to handle video processing requests and API endpoints. MongoDB stores processed video data, annotations, and user insights. We implemented this backend pipeline with PyTorch and Hugging Face Transformers to process human-object interactions by downsampling videos to 2 frames per second (fps). For procedural segmentation and action recognition, we used a sliding window approach with 4-second segments and a 2-second stride to detect procedure completions. The system runs on dedicated hardware with two NVIDIA RTX 8000 GPUs (48GB VRAM each), thus enabling parallel video analysis while maintaining real-time responsiveness. We used CUDA acceleration for computer vision tasks, and OpenAI's commercial API to access multi-modal language models. The microservices-based architecture separates video processing, knowledge extraction, and UI components, ensuring scalability and maintainability. We enforced secure data access through RESTful APIs with authentication. For the frontend, we used React.js for component-based UI rendering and D3.js for interactive visualizations, including timelines and knowledge graphs.

Summary: In summary, InsightAR has a machine learning-based backend that integrates action recognition, human-object interaction detection, and procedural segmentation to generate multi-modal video summaries. Our technical contributions include enhancing procedural segmentation by combining rule-based prediction with vision-language models like EILEV for robustness and generalizability. For action recognition, we leverage EgoVLP and fine-tune LLAMA-13B with task-specific labels. The system supports human-AI collaboration by allowing experts to refine summaries through annotations and resummarization.

6.5 Case Study Evaluation with Domain Experts

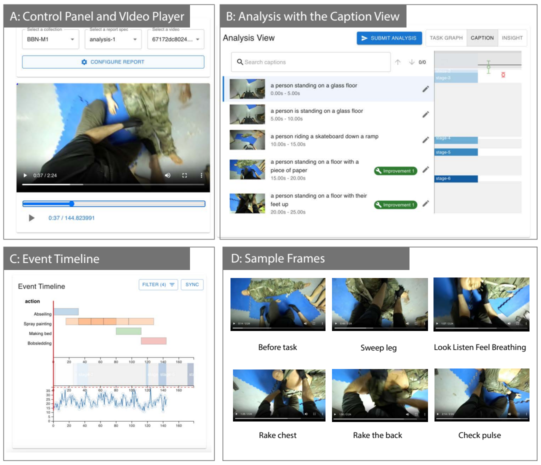
We now demonstrate how InsightAR generalizes to specialized domains through two case study applications: medical training and pilot training. We developed these applications in collaboration with domain experts from our formative study.

6.5.1 Case I: Assessing Task Performance for Battlefield Medical Training

Context: We collaborated with battlefield medical training experts to examine how InsightAR could support the analysis of emergency medical procedures. For these procedures, trainees typically practice skills such as tourniquet application and emergency rescue while wearing head-mounted devices like HoloLens 2, which provide task guidance and record the procedures for later review.

Data Collection: Our medical training expert provided a dataset containing recordings of **ten different medical tasks,** including nasopharyngeal airway (NPA) insertion, trauma assessment, tourniquet application, pressure dressing, wound packing, X-Stat application, bag valve mask ventilation, chest seal application, and needle chest decompression. Each task collection contained 10-20 videos along with corresponding procedural documentation. We processed these recordings using InsightAR and organized them into collections based on task type.

Analysis using InsightAR: In the following evaluation process, we use Bob to refer to the user. During evaluation, Bob analyzed tourniquet application and trauma assessment tasks using InsightAR. For the tourniquet task, he confirmed step completeness in the overview summary and identified a missed action, “hands twist windlass,” through the step breakdown. Using annotations, he documented the error, prompting the system to re-summarize the video with prioritized insights. In the trauma assessment, Bob used the recording collection view to compare performance metrics, identifying anomalies in step durations. He verified an incorrectly executed “Rake chest” step by reviewing linked footage and annotating the error. **This case study demonstrates how InsightAR efficient error detection, targeted feedback, and comparative analysis through multi-level summarization for a training task in a specialized domain like battlefield medical training.** The case study interface is shown in Figure 6.7.



a person standing on a glass floor
0.00s - 5.00s

a person is standing on a glass floor
5.00s - 10.00s

a person riding a skateboard down a ramp
10.00s - 15.00s

a person standing on a floor with a piece of paper
15.00s - 20.00s
Improvement 1

a person standing on a floor with their feet up
20.00s - 25.00s
Improvement 1

Figure 6.7: Case study on InsightAR for the medical task recording analysis. This case study demonstrates how InsightAR assists in evaluating medical procedures by analyzing egocentric video recordings from AR devices used during medical training. **(A)** The user can use Control Panel to switch between the videos and preview the video by Video Player. **(B)** The user can view the captions of the medical procedure and identify the potential insights from the Analysis View. **(C)**: The user can use the Timeline to view the actions in the video. **(D)**: The sample frames from the medical training video are presented.

6.5.2 Case II: Optimizing Procedures for Pilot Training

Context: We conducted a second case study focusing on helicopter pilot training, analyzing pre-flight and in-flight equipment check procedures in a high-fidelity simulator.

Data Collection: We used a collection of helicopter pilot training recordings in the simulator. The videos focus on the pre-flight phase.

We showcase how the pilot training expert user can use InsightAR to analyze recordings from pilots with varying experience levels (we use Amy to refer the user). She begins by selecting the pre-flight phase collection and reviewing a novice trainee's overview-level summary to confirm all required procedural steps were done, but a timeline visualization revealed an unusually long duration for one step. Upon navigating to the step-level summary, Amy discovered the idle time during which the trainee appeared inactive. She then confirmed via the synchronized video player that the trainee hesitated due to unfamiliarity with the procedure, prompting her to annotate specific suggestions for improvement. For comparative analysis, she then reviewed an experienced pilot's recording, using the Task Flow View to highlight and verify critical steps—confirming that common errors were absent and that shorter durations indicated genuine efficiencies. Finally, after documenting her assessment in the notes panel, InsightAR regenerated the summary with her annotations. **This case study shows how InsightAR supports comprehensive analysis of task recordings by enabling efficient navigation between different levels of summary, facilitating comparison across skill levels, and integrating expert comments into the analytical process.**

6.6 User Study Evaluation

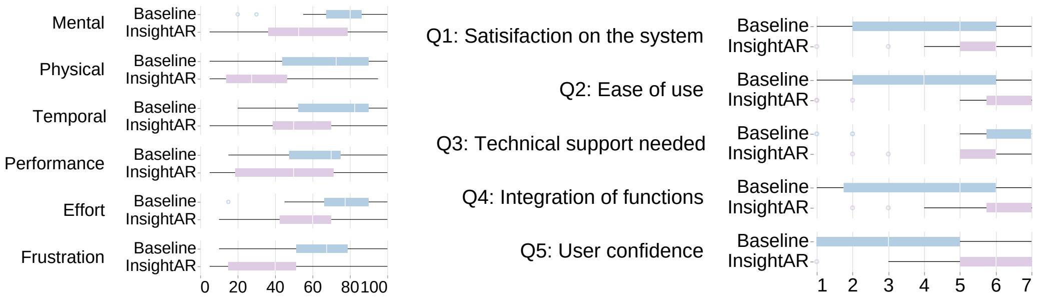
To evaluate our system's performance and user experience with the objective measures, we conducted a user study to evaluate the effectiveness by analyzing egocentric task recordings. This evaluation explores the benefit and efficacy of our system in determining the errors and potentials of insight summation. We ask the following research questions:

1. RQ1: Can our system facilitate users in identifying procedural errors and support more efficient insight summarization?

2. RQ2: What are the user experiences, benefits, and limitations of using an AI-assisted analysis system for multi-view, egocentric videos?

6.6.1 Experiment Design

The study is *within-subjects* with two conditions: (1) **baseline:** manual analysis of egocentric videos using standard video players aided by a browser showing the vision models's output data and the summary generated by a state-of-the-art video summarization model; and (2) **InsightAR**: using LLM-based automated error identification and video insights summation. A Latin square is used to mitigate the order/learning effects. For task analysis, we deliberately selected two egocentric coffee-making videos that had task performance errors. Each video contained 5 predefined errors of varying types: procedural errors (e.g., incorrect step



(a) NASA TLX Results

(b) SUS experiment results

Figure 6.8: Comparison between baseline and InsightAR on (a) NASA TLX metrics—mental demand, physical demand, temporal demand, effort, performance, and frustration, and (b) System usability scales (SUS).

sequence), technique errors (e.g., improper tool handling), and time inefficiencies (e.g., excessive dwelling between steps). Both videos were comparable in length (approximately 3 minutes) and complexity (7-8 steps).

6.6.2 Participants

We recruited 16 participants (9 male, 7 female, 0 non-binary; ages 23-34) via our institution's mailing lists and professional networks. Among them, 13 had experience in data visualization research and 16 in data analysis.

6.6.3 Tasks

Participants were asked to analyze two coffee-making videos using our InsightAR. Specifically, they were instructed to: (1) identify any errors or inefficiencies in the task performance, (2) document these issues with timestamps and write down insights, and (3) suggest improvements based on their analysis. Each participant completed two analysis sessions with a 5-minute break between conditions to prevent fatigue.

6.6.4 Procedure

We conducted the study remotely. Each session lasted approximately 60 minutes using the following procedure: First, participants completed an online consent form and a pre-study questionnaire about the demographic information and prior experience with video analysis. Next, they familiarized themselves with the coffee-making task and the reference instructions. They then viewed a visual tutorial demonstrating InsightAR's features to become acquainted with the interface. After each condition, participants completed questionnaires measuring system usability and task workload. The system usability was measured using the System Usability Scale (SUS) with a 10-item questionnaire assessing perceived usability. The task workload was measured using the NASA Task Load Index (NASA-TLX),assessing the workload across six dimensions (mental demand, physical demand, temporal demand, performance, effort, and frustration). Then, we also ask the questions on their feedback on the system usage and experience.

6.6.5 Analysis and Results

We analyzed our quantitative data using Mann-Whitney test with a significance level of $\alpha == 0.05$, reporting effect sizes using Cohen's d [86]. We analyzed the numbers of errors and insights using Wilcoxon signed-rank tests, as count-based data typically violates normality assumptions. The Likert-scale data from SUS and NASA-TLX questionnaires were analyzed with non-parametric Wilcoxon signed-rank tests. For error analysis, we use the procedure's step guidance as ground truth; actions or results of actions deviant from the guidance are considered errors. Each error will only count once in a video. We ask the participant to note any factual events, personal traits, and behavioral observations for insights. The experimenter examined the insights to ensure they were relevant to the video.

6.6.5.1 Quantiative Results

NASA TLX results (Figure 6.8 (a)) reflect that InsightAR significantlyreduces users' temporal load and performance load compared to the baseline approach ($p < 0.05$, $U = 184.50$, $z = 0.72$). This reduction in cognitive load suggests that the multi-modal summarization and interactive analysis components of InsightAR effectively streamline the task video analysis process, allowing users to identify important information more efficiently.

The SUS evaluation presented in Figure 6.8 (b) revealed that InsightAR achieved a better on the question on the user satisisfaction. This indicates that participants found the system to be more effective. Notably, participants rated the higher score on the system easy-to-use, suggesting that the interface design and interaction methods of InsightAR were intuitive and accessible even for first-time users with minimal training required.

Evaluation on error and insights identification showed that InsightAR identified significantly more insights ($p < 0.05$, $U = 76.00$ and $z = 0.36$) than that of the baseline. On average, InsightARhelped participants to identify ($N_{insights} = 2.19$,$STD_{insights} = 1.42$) per video, which is about 17% more than that of the baseline ($N_{errors} = 1.31$, $STD_{errors} = 0.77$). For error analysis, we did not find a significant effect on InsightAR($p = 0.093$, $U = 85.00$, $z = 0.33$) than that of the baseline. On average, InsightARfound ($N_{errors} = 2.12$, $STD_{errors} = 0.55$) errors, and the baseline ($N_{errors} = 1.00$, $STD_{errors} = 0.58$).

6.6.5.2 Qualitative Results

The analysis of participant's data revealed three key findings regarding improvement on video summarization analysis, ability to retrieve insights, and multi-user collaboration.

InsightAR enhanced multimodal video summarization and analysis Participants reported that InsightAR enabled them to analyze egocentric recordings more efficiently compared to manual methods. Rather than watching entire videos sequentially, participants could use the multi-level summary views to quickly identify areas of interest. As P14 noted, "*The step-level summary allows me to immediately focus on interesting frames without watching the entire recording.*" The multimodal summaries provided complementary information through both visual and textual formats, supporting different analytical approaches. Participants (P14, P3, P10, P5) found these summary formats more engaging and useful for skimming the video. P10 mentioned that "*I can quickly understand the video by skimming the textual and visual summaries together.*" Additionally, participants highlighted that the visual interface enables interaction with video summaries and connects them with the original video content, transforming static summaries into engaging, interactive presentation. For example, P10 added that "*the dynamic presentation (interaction on the summary view in the context) of the summaries is helpful*"

InsightAR supported deeper analytical insights through knowledge integration The integration of task domain knowledge with video content proved particularly valuable for generating in-depth analytical insights. Participants appreciated how the *Task Flow View* visualized dependencies between steps and object states, helping them understand what errors occurred and why they occurred. Comparing their performance with recordings from more experienced practitioners, participants were able to identify specific technique differences that would have been difficult to notice through traditional video review. For example, in the coffee-making case study, P11 identified that their excessive time on step 3 (folding paper filter) was due to an improper handling technique that became apparent when compared with expert recordings. P11 added that "*such timeline view clearly shows the time patterns.*" The system's ability to align the recorded actions with standard procedures through the knowledge graph allowed participants to detect procedural deviations that might otherwise be overlooked in manual review. As P6 remarked, "*That (Task Graph View) is an interesting way to inspire my memory of the task workflow.*"

InsightAR supported human-AI collaborative analysis The human-AI collaborative aspect of the analysis was useful as participants often used the system's automated findings as starting points, and then applied their domain knowledge to interpret and expand upon these observations. The LLM summarizer would re-summarize the video content based on the user's annotated findings and insights. This resulted in improved video summaries that were more helpful for enhancing task analysis results. For example, P6 mentioned that "*The summary can be refreshed (re-summarization) after I add the insights, which is cool.*"

6.7 Discussion

We found that InsightAR significantly enhanced task analysis quality by enabling users to identify more insights and errors (RQ1). The results showed that our system helped uncover 21.20% more insights compared to manual analysis. This improvement is driven by features such as multi-level summarization, the alignment of visual content with task knowledge graphs, and an interactive timeline that facilitates efficient navigation between related steps, boosting confidence in analyzing complex tasks. Our user study feedback analysis confirms that InsightAR successfully addresses the design requirements identified in our formative study, providing an effective tool for comprehensive task analysis from egocentric recordings. In the following paragraphs, we discuss the user experience, benefits, limitations, and potential of interacting with InsightAR (RQ2):

Expanding LLM's capabilities. The effectiveness of InsightAR lies in its hybrid architecture that combines low-level vision models with LLM-based summarization. This structure oercomes the limitations of LLM-only approaches in identifying visual events and actions [313]. This design, similar to recent works that pair LLMs with caption-based summaries [10], leverages specialized models (e.g., for action recognition [39] and object detection [386]) to enrich video understanding. Such paradigm can be expanded to other methods where video processing is required, which is not an advantage of existing LLMs. This paradigm extends to applications like video-QA [356]—where LLMs require action and object data to answer questions accurately. Similarly, in video-retrieval tasks [213], LLMs cannot directly parse a large set of video data efficiently, and vision models can produce the necessary embeddings and feature representations for effective video retrieval and comparison.

Automatically creating new task procedures. While InsightAR currently works with predefined procedures, our backend pipeline could be extended to automatically generate procedures from exemplar recordings. This capability would be valuable in domains where formal documentation lags behind practiced expertise. For example, in industrial settings, where expert technicians develop optimized workflows before formalizing them [220], extracting procedural steps to draft documentation could capture tacit knowledge. This reverse workflow of deriving procedures from practice, could accelerate knowledge transfer in rapidly evolving fields, such as manufacturing, and preserve expertise that might otherwise be lost [46].

Human-in-the-Loop. InsightAR integrates human input into the analysis process, enabling domain experts to augment system-generated summaries with critical contextual insights. In our cooking task case study, a participant noted that excessive time on step 3 was due to unfamiliarity with equipment rather than procedural complexity—an insight that automated analysis alone missed. The insight is also reflected in [373], where the authors noted that the LLM-based decision making may be not optimal and can be improved by the human participation. This human-AI collaboration combines automated efficiency with expert's contextual understanding. The system's annotation features allow users to document these insights for future training and improvements. This approach overcomes the limitations of fully automated systems that can fail to capture essential context.

Expanding to VR and stereo task guidance. For future, our approach can be extended to other formats such as Virtual Reality (VR) [170, 278] and stereo videos. For example, surgical training often employs stereo recordings to capture depth information crucial for procedural assessment, and while adapting InsightAR to these contexts requires specialized processing for 3D spatial data, the core principles of procedural alignment and multi-modal summarization remain applicable. As our medical training expert noted, "Having similar summary capabilities for our VR training simulations would significantly enhance debriefing sessions," enabling analysis of spatial interactions that that are difficult to capture in traditional 2D recordings, potentially revealing insights about ergonomics and efficiency in physical tasks that current approaches miss.

Potential for real-time analysis While InsightAR is currently designed for post-task analysis, it can be expanded for real-time analysis to provide immediate feedback to users during task execution [264]. This brings new potentials to training as instructors becomes more capable to intervene at critical moments and enabling self-correction for trainees. Implementing real-time analysis presents additional challenges in latency management, incomplete procedural information handling, and balancing computational requirements under the limitations of AR devices. By exploring edge computing solutions to analyze video streams locally while offloading complex processing to cloud infrastructure, a hybrid system structure can be applied for both immediate feedback and comprehensive post-task analysis.

Integrating reflection prompts to deepen understanding of performance. Reflection is the process of critically thinking about one's actions and decisions to deepen understanding (for example, prompting users to consider not just what happened, but why) and improve future performance [296, 299, 368]. Our system lays the groundwork for embedding reflection prompts tied to performance patterns (e.g., unusual time spent on particular steps) and allows users to document insights alongside objective data. These reflections can be added to video summaries, enriching them with both behavioral and cognitive perspectives. This could be valuable in training applications where mental processes matter as much as physical task execution.

6.8 Limitations and Future Work

While we evaluated the effectiveness of InsightAR, several limitations exist. InsightAR currently assumes that task procedures are linear, limiting its ability to analyze workflows that involve branching paths or emergency deviations, as highlighted by **E1** in the context of helicopter operations. Implementing graph-structured task schemas could better represent complex procedural relationships and improve flexibility in analysis. Another limitation is multimodal LLMs lack domain-specific knowledge, and action recognition models struggle with specialized tasks that deviate from their training data. Future advancements in domain adaptation could improve the accuracy of automated insights for specific use cases. Finally, scaling InsightAR to handle large datasets, such as hundreds of YouTube cooking videos suggested by **E3**, requires efficient video database management and multi-source summarization techniques to extract meaningful insights from vast video collections.

6.9 Final Considerations

This chapter presented InsightAR, a visual analytics tool for analyzing egocentric task recordings, particularly from head-mounted AR devices. InsightAR supports multi-modal summarization by generating visual keyframes and textual descriptions at both overview and step levels, allowing users to review task performance without watching full videos. Its interactive interface aligns task knowledge with video content, helping identify procedural errors and performance gaps. Our case studies in medical and pilot training, along with a user study of 16 participants, validated InsightAR's generalizability and effectiveness. By turning raw first-person videos into structured, explorable insights—minimizing manual review effort, InsightAR thereby reduced the need for manual review processes currently followed.

Chapter 7

Conclusions and Future Work

This dissertation presents a framework composed of three major components for implementing adaptive and holistic AR task guidance. The first component, ARTiST, introduces an adaptive text simplification method for AR task guidance that accounts for spatial context and AR-specific challenges. A formative study was conducted to understand the unique requirements of text simplification in AR, which differ from those in traditional NLP. We incorporate these findings into LLMs by considering the user's spatial context. A user study shows that our method significantly reduces both cognitive load and task errors through simplified text instructions. The second component focuses on user modeling in AR task guidance. Inspired by the belief-desire-intention (BDI) model from cognitive psychology, we develop a framework that enables LLMs to deliver adaptive and proactive support. In our system, beliefs, desires, and intentions are mapped to the user's context, goals, and next actions. The integration of this model into an LLM-based guidance system enables context-aware suggestions. Our evaluation shows that the BDI-based system achieves comparable effectiveness and user experience to a Wizard-of-Oz system operated by expert designers. The third component addresses post-guidance analysis, completing the holistic pipeline. This component builds on two works: IntentVizor and InsightAR. IntentVizor proposes a novel query-guided video summarization model that supports generic video summarization based on user queries and integrates user interaction via a visual analytics interface. Experiments on benchmark datasets demonstrate state-of-the-art performance, and a case study highlights the system's utility. InsightAR extends this by offering AR-task-specific summarization and analysis, enabling deeper insight into user performance and generating improvement suggestions using LLMs. Our user study indicates that InsightAR helps users identify more task errors and improves overall experience compared to the baseline.

Although AR task guidance has been studied since the early 1990s, its practical deployment has long been limited by hardware and model capabilities. Recent advances in multimodal LLMs and AR/MR devices have revitalized interest in this area. While progress has been made in object recognition and guidance delivery, understanding user behavior and context remains a significant challenge.

In recent years, we have observed a resurgence of AR task guidance, enabled by the emergence of multimodal large language models (LLMs) and the growing availability of advanced AR/MR devices such as HoloLens 2 and Apple Vision Pro. These developments have made it possible to build more adaptive, perceptually grounded, and context-aware guidance systems than were previously feasible.

Multimodal LLMs enable new capabilities in AR task guidance by offering better transferability and perceptual grounding than traditional domain-specific models. These advantages allow the creation of strong task guidance systems through prompting strategies, as demonstrated in ARTiST and Satori. With the increasing availability of egocentric multimodal models [13, 72, 108, 268, 344], task guidance datasets [51, 53, 162, 222, 224, 311], and relevant surveys [246, 287], the foundation for LLM-driven AR guidance continues to grow. Through the development of our framework, we identify several open challenges associated with LLM integration in AR:

Understanding Temporal Dependency with LLMs Understanding temporal dependencies in task workflows is essential for generating appropriate next-step suggestions and identifying user errors in AR task guidance. However, current multimodal LLMs struggle with temporal reasoning across video sequences [59, 111], which impacts their ability to predict procedural errors or next actions in AR tasks. Although techniques like plan-of-techniques [325] and chain-of-thought prompting [318] partially address this, limitations remain. These issues may stem from sensitivity to input order and hallucinations in temporal reasoning. Future improvements in LLM training—especially with more temporally structured data—could help. Additionally, incorporating symbolic reasoning [82, 335] or human feedback [180, 310] may help mitigate temporal errors.

Hallucination LLMs remain susceptible to hallucination, particularly in unfamiliar or domain-specific tasks such as aviation [107, 189, 277, 289]. While research continues on mitigating hallucinations, effective strategies include incorporating domain knowledge through prompt engineering [249] or retrieval-augmented generation [213].

Latency and Model Serving Edge devices (e.g., mobile phones) can typically support models up to 7B parameters, which often lack the capacity for accurate perception and recommendation [336, 376]. Consequently, large models (e.g., 70B) are often hosted on remote servers, introducing latency issues. Because real-time responsiveness is critical in AR, we propose a hybrid solution demonstrated in Satori, where local models predict guidance timing and remote models handle reasoning. This approach reduces latency while maintaining reasoning quality.

The challenges discussed above suggest several promising areas for future research. In the following, we outline directions that aim to address limitations in adaptivity, planning, and domain transfer in AR task guidance.

Improved Adaptivity While our current system adapts to user context and task goals, further adaptivity based on user memory, long-term behavior, and collaboration remains to be explored. For example, in cooking tasks, knowing a user's taste preferences could guide ingredient recommendations. Emerging research has begun addressing long-term adaptivity [344], but practical systems remain underdeveloped. Our BDI user model provides a promising foundation, and future work could enhance it with personalized preferences and collaborative contexts.

Flexible Task Planning Many AR task systems follow predefined sequences, yet tasks such as industrial assembly require adaptive planning based on conditions like tool availability or operator skill [55]. Effective planning involves optimizing task sequences and enabling parallel execution where feasible. Proactive system-user communication is essential for co-planning such flexible procedures.

Domain Applications of AR Task Guidance Though not the primary focus of this dissertation, our framework has been applied in aviation and medical domains. To broaden its applicability, we open-sourced our code. However, domain adaptation requires addressing two challenges: the incorporation of domain knowledge—which may be missing from pretrained models—and the specific skill requirements of each domain. For example, general tasks (e.g., pouring water) differ substantially from surgical tasks, which demand fine-grained motion control and precision [60, 68, 320]. Integrating domain-specific knowledge via retrieval-based or structured prompts [81, 160, 249, 308] is necessary to meet these requirements.

Bibliography

[1] L. Adolphs, K. Shuster, J. Urbanek, A. Szlam, and J. Weston. Reason first, then respond: Modular generation for knowledge-infused dialogue. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7112–7132, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics.

[2] A. F. Agarap. Deep learning using rectified linear units (relu). *CoRR*, abs/1803.08375, 2018.

[3] M. Akçayır and G. Akçayır. Advantages and challenges associated with augmented reality for education: A systematic review of the literature. *Educational research review*, 20:1–11, 2017.

[4] S. S. Al-Thanyyan and A. M. Azmi. Automated text simplification: a survey. *ACM Computing Surveys (CSUR)*, 54(2):1–36, 2021.

[5] O. Alonzo. The use of automatic text simplification to provide reading assistance to deaf and hard-of-hearing individuals in computing fields. *ACM SIGACCESS Accessibility and Computing*, (132), mar 2022.

[6] O. Alonzo, S. Lee, M. Maddela, W. Xu, and M. Huenerfauth. A dataset of word-complexity judgements from deaf and hard-of-hearing adults for text simplification. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 119–124, Abu Dhabi, United Arab Emirates (Virtual), 2022. Association for Computational Linguistics.

[7] O. Alonzo, M. Seita, A. Glasser, and M. Huenerfauth. Automatic text simplification tools for deaf and hard of hearing adults: Benefits of lexical simplification and providing users with autonomy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, New York, NY, USA, 2020. Association for Computing Machinery.

[8] O. Alonzo, J. Trussell, M. Watkins, S. Lee, and M. Huenerfauth. Methods for evaluating the fluency of automatically simplified texts with deaf and hard-of-hearing adults at various literacy levels. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–10, New York, NY, USA, 2022. Association for Computing Machinery.

[9] E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris, and I. Patras. Ac-sum-gan: Connecting actor-critic and generative adversarial networks for unsupervised video summarization. *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

[10] D. M. Argaw, S. Yoon, F. C. Heilbron, H. Deilamsalehy, T. Bui, Z. Wang, F. Dernoncourt, and J. S. Chung. Scaling up video summarization pretraining with large language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 8332–8341. IEEE, 2024.

[11] K. Augestad, H. Han, J. Paige, T. Ponsky, C. Schlachta, B. Dunkin, and J. Mellinger. Educational implications for surgical telementoring: a current review with recommendations for future practice, policy, and research. *Surgical endoscopy*, 31:3836–3846, 2017.

[12] R. T. Azuma. A survey of augmented reality. *Presence: teleoperators & virtual environments*, 6(4):355–385, 1997.

[13] S. Bandyopadhyay, V. Bahirwani, L. Aggarwal, B. Guda, L. Li, and A. Colaco. YETI (YET to intervene) proactive interventions by multimodal AI agents in augmented reality tasks. *CoRR*, abs/2501.09355, 2025.

[14] J. Baumeister, S. Y. Ssin, N. A. ElSayed, J. Dorrian, D. P. Webb, J. A. Walsh, T. M. Simon, A. Irlitti, R. T. Smith, M. Kohler, et al. Cognitive cost of using augmented reality displays. *IEEE transactions on visualization and computer graphics*, 23(11):2378–2388, 2017.

[15] B. Beigman Klebanov, K. Knight, and D. Marcu. Text simplification for information-seeking applications. In *On the Move to Meaningful Internet Systems 2004: CoopIS, DOA, and ODBASE*, pages 735–747, Berlin, Heidelberg, 2004. Springer, Springer Berlin Heidelberg.

[16] D. Benyon and D. Murray. Adaptive systems: from intelligent tutoring to autonomous agents. *Knowl. Based Syst.*, 6(4):179–219, 1993.

[17] C. Bichlmeier, S. M. Heining, M. Rustaee, and N. Navab. Laparoscopic virtual mirror for understanding vessel structure evaluation study by twelve surgeons. In *2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*, pages 125–128. IEEE, IEEE, 2007.

[18] M. Billinghurst, J. Bowskill, N. Dyer, and J. Morphett. An evaluation of wearable information spaces. In *Proceedings. IEEE 1998 Virtual Reality Annual International Symposium (Cat. No. 98CB36180)*, pages 20–27. IEEE, IEEE, 1998.

[19] M. Billinghurst and A. Duenser. Augmented reality in the classroom. *Computer*, 45(7):56–63, 2012.

[20] D. Bohus, S. Andrist, N. Saw, A. Paradiso, I. Chakraborty, and M. Rad. SIGMA: an open-source interactive system for mixed-reality task assistance research - extended abstract. In *IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops, VR Workshops 2024, Orlando, FL, USA, March 16-21, 2024*, pages 889–890. IEEE, 2024.

[21] L. Bonanni, C. Lee, and T. Selker. A framework for designing intelligent task-oriented augmented reality user interfaces. In R. S. Amant, J. Riedl, and A. Jameson, editors, *Proceedings of the 10th International Conference on Intelligent User Interfaces, IUI 2005, San Diego, California, USA, January 10-13, 2005*, pages 317–319. ACM, 2005.

[22] R. H. Bordini, A. El Fallah Seghrouchni, K. Hindriks, B. Logan, and A. Ricci. Agent programming in the cognitive era. *Autonomous Agents and Multi-Agent Systems*, 34:1–31, 2020.

[23] D. Borro, Á. Suescun, A. Brazález, J. M. González, E. Ortega, and E. González. Warm: Wearable ar and tablet-based assistant systems for bus maintenance. *Applied Sciences*, 11(4):1443, 2021.

[24] C. Botto, A. Cannavò, D. Cappuccio, G. Morat, A. N. Sarvestani, P. Ricci, V. Demarchi, and A. Saturnino. Augmented reality for the manufacturing industry: The case of an assembly assistant. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops, VR Workshops, Atlanta, GA, USA, March 22-26, 2020*, pages 299–304. IEEE, 2020.

[25] M. Bower, C. Howe, N. McCredie, A. Robinson, and D. Grover. Augmented reality in education–cases, places and potentials. *Educational Media International*, 51(1):1–15, 2014.

[26] M. Bratman. Intention, plans, and practical reason. 1987.

[27] L. Braubach, A. Pokahr, and W. Lamersdorf. Jadex: A bdi-agent system combining middleware and reasoning. In *Software agent-based applications, platforms and development kits*, pages 143–168. Springer, 2005.

[28] V. Braun and V. Clarke. *Thematic analysis.* American Psychological Association, 2012.

[29] D. E. Breen, R. T. Whitaker, E. Rose, and M. Tuceryan. Interactive occlusion and automatic object placement for augmented reality. *Computer Graphics Forum*, 15(3):11–22, 1996.

[30] G. W. Brier et al. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

[31] A. Broder. A taxonomy of web search. In *ACM SIGIR Forum*, volume 36, pages 3–10, 2002.

[32] A. Z. Broder, M. Fontoura, E. Gabrilovich, A. Joshi, V. Josifovski, and T. Zhang. Robust classification of rare queries using web knowledge. In *Proceedings of the 30th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 231–238, 2007.

[33] J. Buchner, K. Buntins, and M. Kerres. The impact of augmented reality on cognitive load and performance: A systematic review. *J. Comput. Assist. Learn.*, 38(1):285–303, 2022.

[34] W. Büschel, A. Mitschick, T. Meyer, and R. Dachselt. Investigating smartphone-based pan and zoom in 3d data spaces in augmented reality. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services*, pages 1–13, New York, NY, USA, 2019. Association for Computing Machinery.

[35] Y. Canning, J. Tait, J. Archibald, and R. Crawley. Cohesive generation of syntactically simplified newspaper text. In *Text, Speech and Dialogue*, pages 145–150, Berlin, Heidelberg, 2000. Springer, Springer Berlin Heidelberg.

[36] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[37] S. M. Carlson, M. A. Koenig, and M. B. Harms. Theory of mind. *Wiley Interdisciplinary Reviews: Cognitive Science*, 4(4):391–402, 2013.

[38] J. Carmigniani and B. Furht. *Augmented reality: an overview*, pages 3–46. Springer, New York, NY, 2011.

[39] J. Carreira and A. Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 4724–4733. IEEE Computer Society, 2017.

[40] J. Carroll, G. Minnen, Y. Canning, S. Devlin, and J. Tait. Practical simplification of english newspaper text to assist aphasic readers. In *Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, pages 7–10. Association for the Advancement of Artificial Intelligence, 1998.

[41] J. Carroll, G. Minnen, D. Pearce, Y. Canning, S. Devlin, and J. Tait. Simplifying text for language-impaired readers. In H. S. Thompson and A. Lascarides, editors, *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 269–270, Bergen, Norway, June 1999. Association for Computational Linguistics.

[42] S. Castelo, J. Rulff, E. McGowan, B. Steers, G. Wu, S. Chen, I. Roman, R. Lopez, E. Brewer, C. Zhao, et al. Argus: Visualization of ai-assisted task guidance in ar. *IEEE Transactions on Visualization and Computer Graphics*, 2023.

[43] S. Castelo, J. Rulff, E. McGowan, B. Steers, G. Wu, S. Chen, I. Roman, R. Lopez, E. Brewer, C. Zhao, J. Qian, K. Cho, H. He, Q. Sun, H. Vo, J. Bello, M. Krone, and C. Silva. Argus: Visualization of ai-assisted task guidance in ar. *IEEE Transactions on Visualization and Computer Graphics*, 30(1):1313–1323, 2024.

[44] R. Chandrasekar, C. Doran, and S. Bangalore. Motivations and methods for text simplification. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*, 1996.

[45] Z. Chen, Q. Yang, J. Shan, T. Lin, J. Beyer, H. Xia, and H. Pfister. iball: Augmenting basketball videos with gaze-moderated embedded visualizations. In A. Schmidt, K. Väänänen, T. Goyal, P. O. Kristensson, A. Peters, S. Mueller, J. R. Williamson, and M. L. Wilson, editors, *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI 2023, Hamburg, Germany, April 23-28, 2023*, pages 841:1–841:18. ACM, 2023.

[46] P. Chi, N. Frey, K. Panovich, and I. Essa. Automatic instructional video creation from a markdown-formatted tutorial. In J. Nichols, R. Kumar, and M. Nebeling, editors, *UIST '21: The 34th Annual ACM Symposium on User Interface Software and Technology, Virtual Event, USA, October 10-14, 2021*, pages 677–690. ACM, 2021.

[47] W.-S. Chu, Y. Song, and A. Jaimes. Video co-summarization: Video summarization by visual co-occurrence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3584–3592, 2015.

[48] K. Church and B. Smyth. Understanding the intent behind mobile information needs. In *Proceedings of the 14th international conference on Intelligent user interfaces*, pages 247–256, 2009.

[49] P. R. Cohen and H. J. Levesque. Intention is choice with commitment. *Artificial intelligence*, 42(2-3):213–261, 1990.

[50] J. D'Agostini, L. Bonetti, A. Salee, L. Passerini, G. Fiacco, P. Lavanda, E. Motti, M. Stocco, K. T. Gashay, E. G. Abebe, S. M. Alemu, R. Haghani, A. Voltolini, C. Strobbe, N. Covre, G. Santolini, M. Armellini, T. Sacchi, D. Ronchese, C. Furlan, F. Facchinato, L. Maule, P. Tomasin, A. Fornaser, and M. D. Cecco. An augmented reality virtual assistant to help mild cognitive impaired users in cooking a system able to recognize the user status and personalize the support. In *2018 Workshop on Metrology for Industry 4.0 and IoT, Brescia, Italy, April 16-18, 2018*, pages 12–17. IEEE, 2018.

[51] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, et al. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):4125–4141, 2020.

[52] D. Damen, H. Doughty, G. M. Farinella, A. Furnari, E. Kazakos, J. Ma, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray. Rescaling egocentric vision. *CoRR*, abs/2006.13256, 2020.

[53] D. Damen, H. Doughty, G. M. Farinella, A. Furnari, E. Kazakos, J. Ma, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray. Rescaling egocentric vision: Collection, pipeline and challenges for EPIC-KITCHENS-100. *Int. J. Comput. Vis.*, 130(1):33–55, 2022.

[54] A. Dattolo and F. L. Luccio. Accessible and usable websites and mobile applications for people with autism spectrum disorders: a comparative study. *EAI Endorsed Transactions on Ambient Systems*, 4(13), 2017.

[55] A. de Giorgio, A. Maffei, M. Onori, and L. Wang. Towards online reinforced learning of assembly sequence planning with interactive guidance systems for industry 4.0 adaptive manufacturing. *Journal of manufacturing systems*, 60:22–34, 2021.

[56] D. Dearman, M. Kellar, and K. N. Truong. An examination of daily information needs and sharing opportunities. In B. Begole and D. W. McDonald, editors, *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work, CSCW 2008, San Diego, CA, USA, November 8-12, 2008*, pages 679–688. ACM, 2008.

[57] Y. Deng, W. Lei, M. Huang, and T.-S. Chua. Rethinking conversational agents in the era of llms: Proactivity, non-collaborativity, and beyond. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, SIGIR-AP '23, page 298–301, New York, NY, USA, 2023. Association for Computing Machinery.

[58] S. Devlin. The use of a psycholinguistic database in the simplification of text for aphasic readers. *Linguistic databases*, 1998.

[59] X. Ding and L. Wang. Do language models understand time? In *Companion Proceedings of the ACM on Web Conference 2025*, pages 1855–1868, 2025.

[60] B. J. Dixon, H. Chan, M. J. Daly, A. D. Vescan, I. J. Witterick, and J. C. Irish. The effect of augmented real-time image guidance on task workload during endoscopic sinus surgery. In *International Forum of Allergy & Rhinology*, volume 2, pages 405–410. Wiley Online Library, 2012.

[61] M. Dunleavy and C. Dede. Augmented reality teaching and learning. *Handbook of research on educational communications and technology*, pages 735–745, 2014.

[62] B. Ens, J. Lanir, A. Tang, S. Bateman, G. Lee, T. Piumsomboon, and M. Billinghurst. Revisiting collaboration through mixed reality: The evolution of groupware. *International Journal of Human-Computer Studies*, 131:81–98, 2019.

[63] D. Escobar-Castillejos, J. Noguez, F. Bello, L. Neri, A. J. Magana, and B. Benes. A review of training and guidance systems in medical surgery. *Applied Sciences*, 10(17):5752, 2020.

[64] M. Eswaran and M. R. Bahubalendruni. Challenges and opportunities on ar/vr technologies for manufacturing systems in the context of industry 4.0: A state of the art review. *Journal of Manufacturing Systems*, 65:260–278, 2022.

[65] R. Evans, C. Orăsan, and I. Dornescu. An evaluation of syntactic simplification rules for people with autism. In S. Williams, A. Siddharthan, and A. Nenkova, editors, *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pages 131–140, Gothenburg, Sweden, Apr. 2014. Association for Computational Linguistics.

[66] J. Fajtl, H. S. Sokeh, V. Argyriou, D. Monekosso, and P. Remagnino. Summarizing videos with attention. In *Proceedings of the Asian Conference on Computer Vision*, pages 39–54, 2018.

[67] C. Feichtenhofer, H. Fan, J. Malik, and K. He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6202–6211, 2019.

[68] E. A. Felinska, T. E. Fuchs, A. Kogkas, Z.-W. Chen, B. Otto, K.-F. Kowalewski, J. Petersen, B. P. Müller-Stich, G. Mylonas, and F. Nickel. Telestration with augmented reality improves surgical performance through gaze guidance. *Surgical Endoscopy*, 37(5):3557–3566, 2023.

[69] I. Fernández del Amo, J. A. Erkoyuncu, R. Roy, and S. Wilding. Augmented Reality in Maintenance: An information-centred design framework. *Procedia Manufacturing*, 19:148–155, 2018.

[70] C. G. Fidalgo, Y. Yan, H. Cho, M. Sousa, D. Lindlbauer, and J. Jorge. A survey on remote assistance and training in mixed reality environments. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2291–2303, 2023.

[71] M. Fiorentino, A. E. Uva, M. Gattullo, S. Debernardis, and G. Monno. Augmented reality on large screen for interactive maintenance instructions. *Comput. Ind.*, 65(2):270–278, 2014.

[72] A. Flaborea, G. M. D. di Melendugno, L. Plini, L. Scofano, E. D. Matteis, A. Furnari, G. M. Farinella, and F. Galasso. PREGO: online mistake detection in procedural egocentric videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 18483–18492. IEEE, 2024.

[73] J. Frandsen, J. Tenny, W. Frandsen Jr, and Y. Hovanski. An augmented reality maintenance assistant with real-time quality inspection on handheld mobile devices. *The International Journal of Advanced Manufacturing Technology*, 125(9):4253–4270, 2023.

[74] C. Frith and U. Frith. Theory of mind. *Current biology*, 15(17):R644–R645, 2005.

[75] H. Furuta, O. Nachum, K. Lee, Y. Matsuo, S. S. Gu, and I. Gur. Multimodal web navigation with instruction-finetuned foundation models. *CoRR*, abs/2305.11854, 2023.

[76] J. L. Gabbard, J. E. Swan, and D. Hix. The effects of text drawing styles, background textures, and natural lighting on text legibility in outdoor augmented reality. *Presence*, 15(1):16–32, 2006.

[77] J. L. Gabbard, J. E. Swan, D. Hix, S.-J. Kim, and G. Fitch. Active text drawing styles for outdoor augmented reality: A user-based study and design implications. In *2007 IEEE Virtual Reality Conference*, pages 35–42. IEEE, IEEE, 2007.

[78] N. Gala and J. Ziegler. Reducing lexical complexity as a tool to increase text accessibility for children with dyslexia. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 59–66, Osaka, Japan, 2016. The COLING 2016 Organizing Committee.

[79] Z. Gan, L. Li, C. Li, L. Wang, Z. Liu, and J. Gao. Vision-language pretraining: Basics, recent advances, and future trends. *Found. Trends Comput. Graph. Vis.*, 14(3-4):163–352, 2022.

[80] Q. Gao, W. Xu, M. Shen, and Z. Gao. Agent teaming situation awareness (ATSA): A situation awareness framework for human-ai teaming. *CoRR*, abs/2308.16785, 2023.

[81] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, Q. Guo, M. Wang, and H. Wang. Retrieval-augmented generation for large language models: A survey. *CoRR*, abs/2312.10997, 2023.

[82] Y. Ge, S. Romeo, J. Cai, R. Shu, M. Sunkara, Y. Benajiba, and Y. Zhang. Tremu: Towards neuro-symbolic temporal reasoning for llm-agents with memory in multi-session dialogues. *CoRR*, abs/2502.01630, 2025.

[83] D. Genzel and E. Charniak. Entropy rate constancy in text. In P. Isabelle, E. Charniak, and D. Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 199–206, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[84] M. Georgeff, B. Pell, M. Pollack, M. Tambe, and M. Wooldridge. The belief-desire-intention model of agency. In *Intelligent Agents V: Agents Theories, Architectures, and Languages: 5th International Workshop, ATAL'98 Paris, France, July 4–7, 1998 Proceedings 5*, pages 1–10. Springer, 1999.

[85] G. Z. Georgiev. *Statistical methods in online A/B testing.* Self-Published, 2019.

[86] G. E. Gignac and E. T. Szodorai. Effect size guidelines for individual differences researchers. *Personality and individual differences*, 102:74–78, 2016.

[87] G. Gkioxari, R. B. Girshick, P. Dollár, and K. He. Detecting and recognizing human-object interactions. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 8359–8367. Computer Vision Foundation / IEEE Computer Society, 2018.

[88] R. Gong, Q. Huang, X. Ma, H. Vo, Z. Durante, Y. Noda, Z. Zheng, S. Zhu, D. Terzopoulos, L. Fei-Fei, and J. Gao. Mindagent: Emergent gaming interaction. *CoRR*, abs/2309.09971, 2023.

[89] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *CoRR*, abs/1706.02677, 2017.

[90] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, M. Martin, T. Nagarajan, I. Radosavovic, S. K. Ramakrishnan, F. Ryan, J. Sharma, M. Wray, M. Xu, E. Z. Xu, C. Zhao, S. Bansal, D. Batra, V. Cartillier, S. Crane, T. Do, M. Doulaty, A. Erapalli, C. Feichtenhofer, A. Fragomeni, Q. Fu, A. Gebreselasie, C. Gonzalez, J. Hillis, X. Huang, Y. Huang, W. Jia, W. Khoo, J. Kolar, S. Kottur, A. Kumar, F. Landini, C. Li, Y. Li, Z. Li, K. Mangalam, R. Modhugu, J. Munro, T. Murrell, T. Nishiyasu, W. Price, P. R. Puentes, M. Ramazanova, L. Sari, K. Somasundaram, A. Southerland, Y. Sugano, R. Tao, M. Vo, Y. Wang, X. Wu, T. Yagi, Z. Zhao, Y. Zhu, P. Arbelaez, D. Crandall, D. Damen, G. M. Farinella, C. Fuegen, B. Ghanem, V. K. Ithapu, C. V. Jawahar, H. Joo, K. Kitani, H. Li, R. Newcombe, A. Oliva, H. S. Park, J. M. Rehg, Y. Sato, J. Shi, M. Z. Shou, A. Torralba, L. Torresani, M. Yan, and J. Malik. Ego4d: Around the world in 3,000 hours of egocentric video, 2022.

[91] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 2017.

[92] J. Guo, Y. Fan, Q. Ai, and W. B. Croft. A deep relevance matching model for ad-hoc retrieval. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, pages 55–64, 2016.

[93] M. Gygli, H. Grabner, H. Riemenschneider, and L. V. Gool. Creating summaries from user videos. In *Computer Vision - ECCV 2014 - 13th European Conference*, volume 8695, pages 505–520. Springer, 2014.

[94] S. G. Hart. Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 50, pages 904–908. Sage publications Sage CA: Los Angeles, CA, 2006.

[95] M. Harvey, M. Langheinrich, and G. Ward. Remembering through lifelogging: A survey of human memory augmentation. *Pervasive Mob. Comput.*, 27:14–26, 2016.

[96] B. He, J. Wang, J. Qiu, T. Bui, A. Shrivastava, and Z. Wang. Align and attend: Multimodal summarization with dual contrastive losses. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 14867–14878. IEEE, 2023.

[97] J. He, X. Wang, K. K. Wong, X. Huang, C. Chen, Z. Chen, F. Wang, M. Zhu, and H. Qu. Videopro: A visual analytics approach for interactive video programming. *IEEE Transactions on Visualization and Computer Graphics*, 30(1):87–97, 2023.

[98] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016.

[99] S. J. Henderson and S. Feiner. Evaluating the benefits of augmented reality for task localization in maintenance of an armored personnel carrier turret. In *2009 8th IEEE International Symposium on Mixed and Augmented Reality*, pages 135–144. IEEE, IEEE, 2009.

[100] K. V. Hindriks. Programming rational agents in goal. In *Multi-agent programming: Languages, tools and applications*, pages 119–157. Springer, 2009.

[101] F. Hmida, M. B. Billami, T. François, and N. Gala. Assisted lexical simplification for French native children with reading difficulties. In A. Jönsson, E. Rennes, H. Saggion, S. Stajner, and V. Yaneva, editors, *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, pages 21–28, Tilburg, the Netherlands, Nov. 2018. Association for Computational Linguistics.

[102] A. Holder, C. Elsey, M. Kolanoski, and M. Mair. Investigating the use of force in contemporary conflict: Researching military operations with audio, video and transcript data. 2022.

[103] R. L. Holloway. Registration error analysis for augmented reality. *Presence: Teleoperators & Virtual Environments*, 6(4):413–432, 1997.

[104] A. Holynski and J. Kopf. Fast depth densification for occlusion-aware augmented reality. *ACM Transactions on Graphics (ToG)*, 37(6):1–11, 2018.

[105] S. Hong, M. Zhuge, J. Chen, X. Zheng, Y. Cheng, C. Zhang, J. Wang, Z. Wang, S. K. S. Yau, Z. Lin, L. Zhou, C. Ran, L. Xiao, C. Wu, and J. Schmidhuber. Metagpt: Meta programming for a multi-agent collaborative framework, 2023.

[106] M. Hornacek, H. Küffner-McCauley, M. Trimmel, P. Rupprecht, and S. Schlund. A spatial ar system for wide-area axis-aligned metric augmentation of planar scenes. *CIRP Journal of Manufacturing Science and Technology*, 37:219–226, 2022.

[107] B. Hou, Y. Zhang, J. Andreas, and S. Chang. A probabilistic framework for LLM hallucination detection via belief tree propagation. In L. Chiruzzo, A. Ritter, and L. Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 3076–3099. Association for Computational Linguistics, 2025.

[108] Y. Huang, J. Xu, B. Pei, Y. He, G. Chen, M. Zhang, L. Yang, Z. Nie, J. Liu, G. Fan, D. Lin, F. Fang, K. Li, C. Yuan, X. Chen, Y. Wang, Y. Wang, Y. Qiao, and L. Wang. An egocentric vision-language model based portable real-time smart assistant. *CoRR*, abs/2503.04250, 2025.

[109] Y.-R. Huang, J. Zhang, T.-C. Liu, and K.-E. Chang. Designing an ar-based guidance and feedback system for learning assistance. In *2024 5th International Conference on Information Technology and Education Technology (ITET)*, pages 51–55. IEEE, 2024.

[110] M.-B. Ibáñez and C. Delgado-Kloos. Augmented reality for stem learning: A systematic review. *Computers & Education*, 123:109–123, 2018.

[111] M. F. M. Imam, C. Lyu, and A. F. Aji. Can multimodal llms do visual temporal understanding and reasoning? the answer is no! *CoRR*, abs/2501.10674, 2025.

[112] T. I. Ivanova. Ontology-based text simplification for dyslexics. *Science and Technology*, 3(10):34–47, 2017.

[113] P. Jain, R. Farzan, and A. J. Lee. Co-designing with users the explanations for a proactive auto-response messaging agent. *Proceedings of the ACM on Human-Computer Interaction*, 7(MHCI):1–23, 2023.

[114] M. Jamali, Z. M. Williams, and J. Cai. Unveiling theory of mind in large language models: A parallel to single neurons in the human brain. *CoRR*, abs/2309.01660, 2023.

[115] A. Jangra, S. Mukherjee, A. Jatowt, S. Saha, and M. Hasanuzzaman. A survey on multi-modal summarization. *ACM Comput. Surv.*, 55(13s):296:1–296:36, 2023.

[116] B. Jia, T. Lei, S.-C. Zhu, and S. Huang. Egotaskqa: Understanding human tasks in egocentric videos. *Advances in Neural Information Processing Systems*, 35:3343–3360, 2022.

[117] C. Jiang, M. Maddela, W. Lan, Y. Zhong, and W. Xu. Neural CRF model for sentence alignment in text simplification. In D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7943–7960, Online, 2020. Association for Computational Linguistics.

[118] P. Jiang and Y. Han. Hierarchical variational network for user-diversified & query-focused video summarization. In *Proceedings of the 2019 International Conference on Multimedia Retrieval*, pages 202–206, 2019.

[119] Y. Jung, D. Cho, D. Kim, S. Woo, and I. S. Kweon. Discriminative feature learning for unsupervised video summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8537–8544, 2019.

[120] D. Kahneman and A. Tversky. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*, pages 99–127. World Scientific, 2013.

[121] H. Kanafani, J. A. Ghauri, S. Hakimov, and R. Ewerth. Unsupervised video summarization via multi-source features. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*, page 466–470, 2021.

[122] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017.

[123] E. Kazakos, A. Nagrani, A. Zisserman, and D. Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition, 2019.

[124] J. Keil, A. Korte, A. Ratmer, D. Edler, and F. Dickmann. Augmented reality (ar) and spatial cognition: effects of holographic grids on distance estimation and location memory in a 3d indoor scenario. *PFG–Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, 88(2):165–172, 2020.

[125] N. L. Keiser and W. Arthur Jr. A meta-analysis of the effectiveness of the after-action review (or debrief) and factors that influence its effectiveness. *Journal of Applied Psychology*, 106(7):1007, 2021.

[126] F. Keller. The entropy rate principle as a predictor of processing effort: An evaluation against eye-tracking data. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 317–324, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[127] G. Kim, P. Baldi, and S. McAleer. Language models can solve computer tasks. *CoRR*, abs/2303.17491, 2023.

[128] K. Kim, L. Boelling, S. Haesler, J. Bailenson, G. Bruder, and G. F. Welch. Does a digital assistant need a body? the influence of visual embodiment and social behavior on the perception of intelligent virtual agents in ar. In *2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 105–114. IEEE, IEEE, 2018.

[129] S. Kim, S. J. Joo, Y. Jang, H. Chae, and J. Yeo. Cotever: Chain of thought prompting annotation toolkit for explanation verification. In D. Croce and L. Soldaini, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. EACL 2023 - System Demonstrations, Dubrovnik, Croatia, May 2-4, 2023*, pages 195–208. Association for Computational Linguistics, 2023.

[130] S. Kim, H. Xi, S. Mungle, and Y.-J. Son. Modeling human interactions with learning under the extended belief-desire-intention framework. In *IIE Annual Conference. Proceedings*, page 1. Institute of Industrial and Systems Engineers (IISE), 2012.

[131] D. Kinny, M. P. Georgeff, and A. S. Rao. A methodology and modelling technique for systems of BDI agents. In W. V. de Velde and J. W. Perram, editors, *Agents Breaking Away, 7th European Workshop on Modelling Autonomous Agents in a Multi-Agent World, Eindhoven, The Netherlands, January 22-25, 1996, Proceedings*, volume 1038 of *Lecture Notes in Computer Science*, pages 56–71. Springer, 1996.

[132] C. Kofler, M. Larson, and A. Hanjalic. User intent in multimedia search: a survey of the state of the art and future challenges. *ACM Computing Surveys (CSUR)*, 49(2):1–37, 2016.

[133] M. König, M. Stadlmaier, T. Rusch, R. Sochor, L. Merkel, S. Braunreuther, and J. Schilp. Ma 2 ra-manual assembly augmented reality assistant. In *2019 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, pages 501–505. IEEE, 2019.

[134] F. K. Konstantinidis, I. Kansizoglou, N. Santavas, S. G. Mouroutsos, and A. Gasteratos. Marma: A mobile augmented reality maintenance assistant for fast-track repair procedures in the context of industry 4.0. *Machines*, 8(4):88, 2020.

[135] C. Koo, Y. Joun, H. Han, and N. Chung. A structural model for destination travel intention as a media exposure: Belief-desire-intention model perspective. *International Journal of Contemporary Hospitality Management*, 28(7):1338–1360, 2016.

[136] M. Kosinski. Theory of mind might have spontaneously emerged in large language models, 2023.

[137] M. Kraus, M. R. G. Schiller, G. Behnke, P. Bercher, M. Dorna, M. Dambier, B. Glimm, S. Biundo, and W. Minker. ”was that successful?” on integrating proactive meta-dialogue in a diy-assistant using multimodal cues. In K. P. Truong, D. Heylen, M. Czerwinski, N. Berthouze, M. Chetouani, and M. Nakano, editors, *ICMI ’20: International Conference on Multimodal Interaction, Virtual Event, The Netherlands, October 25-29, 2020*, pages 585–594. ACM, 2020.

[138] M. Kraus, N. Wagner, Z. Callejas, and W. Minker. The role of trust in proactive conversational assistants. *IEEE Access*, 9:112821–112836, 2021.

[139] M. Kraus, N. Wagner, and W. Minker. Effects of proactive dialogue strategies on human-computer trust. In T. Kuflik, I. Torre, R. Burke, and C. Gena, editors, *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization, UMAP 2020, Genoa, Italy, July 12-18, 2020*, pages 107–116. ACM, 2020.

[140] M. Kraus, N. Wagner, and W. Minker. Modelling and predicting trust for developing proactive dialogue strategies in mixed-initiative interaction. In Z. Hammal, C. Busso, C. Pelachaud, S. L. Oviatt, A. A. Salah, and G. Zhao, editors, *ICMI '21: International Conference on Multimodal Interaction, Montréal, QC, Canada, October 18-22, 2021*, pages 131–140. ACM, 2021.

[141] M. Kraus, N. Wagner, and W. Minker. Prodial - an annotated proactive dialogue act corpus for conversational assistants using crowdsourcing. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 3164–3173. European Language Resources Association, 2022.

[142] K. Kucher, E. Zohrevandi, and C. A. Westin. Towards visual analytics for explainable ai in industrial applications. *Analytics*, 4(1):7, 2025.

[143] R. Kumaran, Y. Kim, A. E. Milner, T. Bullock, B. Giesbrecht, and T. Höllerer. The impact of navigation aids on search performance and object recall in wide-area augmented reality. In A. Schmidt, K. Väänänen, T. Goyal, P. O. Kristensson, A. Peters, S. Mueller, J. R. Williamson, and M. L. Wilson, editors, *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI 2023, Hamburg, Germany, April 23-28, 2023*, pages 710:1–710:17. ACM, 2023.

[144] Z.-H. Lai, W. Tao, M. C. Leu, and Z. Yin. Smart augmented reality instructional system for mechanical assembly towards worker-centered intelligent manufacturing. *Journal of Manufacturing Systems*, 55:69–81, 2020.

[145] G. Lallai, G. Loi Zedda, C. Martinie, P. Palanque, M. Pisano, and L. D. Spano. Engineering task-based augmented reality guidance: application to the training of aircraft flight procedures. *Interacting with Computers*, 33(1):17–39, 2021.

[146] J.-F. Lapointe, M. S. Allili, L. Belliveau, L. Hebbache, D. Amirkhani, and H. Sekkati. Ai-ar for&nbsp;bridge inspection by&nbsp;drone. In *Virtual, Augmented and Mixed Reality: Applications in Education, Aviation and Industry: 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 II*, page 302–313, Berlin, Heidelberg, 2022. Springer-Verlag.

[147] J.-F. Lapointe, H. Molyneaux, and M. S. Allili. A literature review of ar-based remote guidance tasks with user studies. In *Virtual, Augmented and Mixed Reality. Industrial and Everyday Life Applications*, pages 111–120, Cham, 2020. Springer International Publishing.

[148] L. L. Laster and M. F. Johnson. Non-inferiority trials: the ‘at least as good as’ criterion. *Statistics in Medicine*, 22(2):187–200, 2003.

[149] T. M. Le, V. Le, S. Venkatesh, and T. Tran. Hierarchical conditional relation networks for multimodal video question answering. *Int. J. Comput. Vis.*, 129(11):3027–3050, 2021.

[150] R. Lear, S. Ellis, T. Ollivierre-Harris, S. Long, and E. K. Mayer. Video recording patients for direct care purposes: Systematic review and narrative synthesis of international empirical studies and uk professional guidance. *Journal of Medical Internet Research*, 25:e46478, 2023.

[151] M. Lee, P. Liang, and Q. Yang. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In S. D. J. Barbosa, C. Lampe, C. Appert, D. A. Shamma, S. M. Drucker, J. R. Williamson, and K. Yatani, editors, *CHI '22: CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 29 April 2022 - 5 May 2022*, pages 388:1–388:19. ACM, 2022.

[152] S. Lee, Z. Lu, Z. Zhang, M. Hoai, and E. Elhamifar. Error detection in egocentric procedural task videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 18655–18666. IEEE, 2024.

[153] S. H. Lee. *Integrated human decision behavior modeling under an extended belief-desire-intention framework*. The University of Arizona, 2009.

[154] Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1346–1353, 2012.

[155] R. Leonardi, F. Ragusa, A. Furnari, and G. M. Farinella. Exploiting multimodal synthetic data for egocentric human-object interaction detection in an industrial scenario. *CoRR*, abs/2306.12152, 2023.

[156] V. Lepetit and M.-O. Berger. A semi-automatic method for resolving occlusion in augmented reality. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*, volume 2, pages 225–230. IEEE, IEEE, 2000.

[157] E. Lesaffre. Superiority, equivalence, and non-inferiority trials. *Bulletin of the NYU hospital for joint diseases*, 66(2), 2008.

[158] A. M. Leslie, T. P. German, and P. Polizzi. Belief-desire reasoning as a process of selection. *Cognitive psychology*, 50(1):45–85, 2005.

[159] J. R. Lewis. Ibm computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction*, 7(1):57–78, 1995.

[160] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[161] C. Li, G. Wu, G. Y. Chan, D. G. Turakhia, S. C. Quispe, D. Li, L. Welch, C. Silva, and J. Qian. Satori: Towards proactive AR assistant with belief-desire-intention user modeling. *CoRR*, abs/2410.16668, 2024.

[162] G. Li, K. Zhao, S. Zhang, X. Lyu, M. Dusmanu, Y. Zhang, M. Pollefeys, and S. Tang. Egogen: An egocentric synthetic data generator. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 14497–14509. IEEE, 2024.

[163] H. Li, J. Zhu, C. Ma, J. Zhang, and C. Zong. Multi-modal summarization for asynchronous collection of text, image, audio and video. In M. Palmer, R. Hwa, and S. Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 1092–1102. Association for Computational Linguistics, 2017.

[164] J. Li, D. Li, S. Savarese, and S. C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR, 2023.

[165] J. Li, D. Li, C. Xiong, and S. C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, and S. Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR, 2022.

[166] J. N. Li, Y. Xu, T. Grossman, S. Santosa, and M. Li. Omniactions: Predicting digital actions in response to real-world multimodal sensory inputs with llms. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–22, 2024.

[167] X. Li. A review of prominent paradigms for llm-based agents: Tool use, planning (including rag), and feedback learning. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, and S. Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 9760–9779. Association for Computational Linguistics, 2025.

[168] Y. Li, Z. Ye, and J. M. Rehg. Delving into egocentric actions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 287–295, 2015.

[169] H. Lie, K. Studer, Z. Zhao, B. Thomson, D. G. Turakhia, and J. Liu. Training for open-ended drilling through a virtual reality simulation. In *2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 366–375, 2023.

[170] H. Lie, K. Studer, Z. Zhao, B. Thomson, D. G. Turakhia, and J. Liu. Training for open-ended drilling through a virtual reality simulation. In *2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 366–375. IEEE, 2023.

[171] G. Lin, T. Panigrahi, J. Womack, D. J. Ponda, P. Kotipalli, and T. Starner. Comparing order picking guidance with microsoft hololens, magic leap, google glass xe and paper. In *Proceedings of the 22nd international workshop on mobile computing systems and applications*, pages 133–139, 2021.

[172] K. Q. Lin, J. Wang, M. Soldan, M. Wray, R. Yan, E. Z. Xu, D. Gao, R. Tu, W. Zhao, W. Kong, C. Cai, H. Wang, D. Damen, B. Ghanem, W. Liu, and M. Z. Shou. Egocentric video-language pretraining. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

[173] D. Lindlbauer, A. M. Feit, and O. Hilliges. Context-aware online adaptation of mixed reality interfaces. In *Proceedings of the 32nd annual ACM symposium on user interface software and technology*, pages 147–160, 2019.

[174] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[175] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[176] Y. Liu, D. Yang, Y. Wang, J. Liu, J. Liu, A. Boukerche, P. Sun, and L. Song. Generalized video anomaly event detection: Systematic taxonomy and comparison of deep models. *ACM Comput. Surv.*, 56(7):189:1–189:38, 2024.

[177] H. Lotherington-Woloszyn. Do simplified texts simplify language comprehension for esl learners?. *English for Specific Purposes*, 68:31–46, 1993.

[178] F. H. Lund. The psychology of belief. *The Journal of Abnormal and Social Psychology*, 20(1):63, 1925.

[179] M. R. Lyu, I. King, T. Wong, E. Yau, and P. Chan. Arcade: Augmented reality computing arena for digital entertainment. In *2005 IEEE Aerospace Conference*, pages 1–9. IEEE, 2005.

[180] S. Ma, Y. Lei, X. Wang, C. Zheng, C. Shi, M. Yin, and X. Ma. Who should I trust: AI or myself? leveraging human and AI correctness likelihood to promote appropriate trust in ai-assisted decision-making. In A. Schmidt, K. Väänänen, T. Goyal, P. O. Kristensson, A. Peters, S. Mueller, J. R. Williamson, and M. L. Wilson, editors, *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI 2023, Hamburg, Germany, April 23-28, 2023*, pages 759:1–759:19. ACM, 2023.

[181] Z. Ma, J. Sansom, R. Peng, and J. Chai. Towards A holistic landscape of situated theory of mind in large language models. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 1011–1031. Association for Computational Linguistics, 2023.

[182] B. MacIntyre, E. M. Coelho, and S. J. Julier. Estimating and adapting to registration errors in augmented reality systems. In *Proceedings IEEE Virtual Reality 2002*, pages 73–80. IEEE, IEEE, 2002.

[183] B. Mahasseni, M. Lam, and S. Todorovic. Unsupervised video summarization with adversarial lstm networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 202–211, 2017.

[184] B. Mahasseni, M. Lam, and S. Todorovic. Unsupervised video summarization with adversarial LSTM networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2982–2991. IEEE Computer Society, 2017.

[185] A. Mahmood, J. W. Fung, I. Won, and C. Huang. Owning mistakes sincerely: Strategies for mitigating AI errors. In S. D. J. Barbosa, C. Lampe, C. Appert, D. A. Shamma, S. M. Drucker, J. R. Williamson, and K. Yatani, editors, *CHI '22: CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 29 April 2022 - 5 May 2022*, pages 578:1–578:11. ACM, 2022.

[186] M. Mair, C. Elsey, P. V. Smith, and P. G. Watson. War on video: Combat footage, vernacular video analysis and military culture from within. *Ethnographic Studies*, (15):83–105, 2018.

[187] I. Majil, M. Yang, and S. Yang. Augmented reality based interactive cooking guide. *Sensors*, 22(21):8290, 2022.

[188] E. Marino, L. Barbieri, F. Bruno, and M. Muzzupappa. Assessing user performance in augmented reality assembly guidance for industry 4.0 operators. *Computers in Industry*, 157:104085, 2024.

[189] A. Martino, M. Iannelli, and C. Truong. Knowledge injection to counter large language model (llm) hallucination. In *European Semantic Web Conference*, pages 182–185. Springer, 2023.

[190] C. Mather, T. Barnett, V. Broucek, A. Saunders, D. Grattidge, and W. Huang. Helping hands: using augmented reality to provide remote guidance to health professionals. In *Context Sensitive Health Informatics: Redesigning Healthcare Work*, pages 57–62. IOS Press, 2017.

[191] Y. Matsuura, T. Terada, T. Aoki, S. Sonoda, N. Isoyama, and M. Tsukamoto. Readability and legibility of fonts considering shakiness of head mounted displays. In *Proceedings of the 2019 ACM International Symposium on Wearable Computers*, pages 150–159, New York, NY, USA, 2019. Association for Computing Machinery.

[192] L. Mazer, O. Varban, J. R. Montgomery, M. M. Awad, and A. Schulman. Video is better: why aren't we using it? a mixed-methods study of the barriers to routine procedural video recording and case review. *Surgical endoscopy*, pages 1–8, 2022.

[193] M. McGraw-Hunter, G. Faw, and P. Davis. The use of video self-modelling and feedback to teach cooking skills to individuals with traumatic brain injury: A pilot study. *Brain Injury*, 20(10):1061–1068, 2006.

[194] M. F. McTear. User modelling for adaptive computer systems: a survey of recent developments. *Artif. Intell. Rev.*, 7(3-4):157–184, 1993.

[195] A. Meck, C. Draxler, and T. Vogt. How may I interrupt? linguistic-driven design guidelines for proactive in-car voice assistants. *Int. J. Hum. Comput. Interact.*, 40(22):7517–7531, 2024.

[196] S. Messaoud, I. Lourentzou, A. Boughoula, M. Zehni, Z. Zhao, C. Zhai, and A. G. Schwing. DeepQAMVS: Query-aware hierarchical pointer networks for multi-video summarization. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 1389–1399, 2021.

[197] C. Meurisch, M. Ionescu, B. Schmidt, and M. Mühlhäuser. Reference model of next-generation digital personal assistant: integrating proactive behavior. In S. C. Lee, L. Takayama, and K. N. Truong, editors, *Adjunct Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers, UbiComp/ISWC 2017, Maui, HI, USA, September 11-15, 2017*, pages 149–152. ACM, 2017.

[198] C. Meurisch, C. A. Mihale-Wilson, A. Hawlitschek, F. Giger, F. Müller, O. Hinz, and M. Mühlhäuser. Exploring user expectations of proactive AI systems. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 4(4):146:1–146:22, 2020.

[199] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips, 2019.

[200] O. Miksik, I. Munasinghe, J. Asensio-Cubero, S. R. Bethi, S. Huang, S. Zylfo, X. Liu, T. Nica, A. Mitrocsak, S. Mezza, R. Beard, R. Shi, R. W. M. Ng, P. A. M. Mediano, Z. Fountas, S. Lee, J. Medvesek, H. Zhuang, Y. Rogers, and P. Swietojanski. Building proactive voice assistants: When and how (not) to interact. *CoRR*, abs/2005.01322, 2020.

[201] M. Minderer, A. Gritsenko, A. Stone, M. Neumann, D. Weissenborn, A. Dosovitskiy, A. Mahendran, A. Arnab, M. Dehghani, Z. Shen, et al. Simple open-vocabulary object detection. In *European Conference on Computer Vision*, pages 728–755. Springer, 2022.

[202] D. L. Morgan, J. Ataie, P. Carder, and K. Hoffman. Introducing dyadic interviews as a method for collecting qualitative data. *Qualitative health research*, 23(9):1276–1284, 2013.

[203] A. Mulloni, A. Dünser, and D. Schmalstieg. Zooming interfaces for augmented reality browsers. In *Proceedings of the 12th International Conference on Human Computer Interaction with Mobile Devices and Services*, pages 161–170, New York, NY, USA, 2010. Association for Computing Machinery.

[204] S. Nalla, M. Agrawal, V. Kaushal, G. Ramakrishnan, and R. Iyer. Watch hours in minutes: Summarizing video with user intent. In *Proceedings of the European Conference on Computer Vision*, pages 714–730, 2020.

[205] M. Narasimhan, A. Rohrbach, and T. Darrell. Clip-it! language-guided video summarization. In M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 13988–14000, 2021.

[206] A. Nijholt. Towards social companions in augmented reality: Vision and challenges. In *Distributed, Ambient and Pervasive Interactions. Smart Living, Learning, Well-Being and Health, Art and Creativity: 10th International Conference, DAPI 2022, Held as Part of the 24th HCI International Conference, HCII 2022, Virtual Event, June 26 – July 1, 2022, Proceedings, Part II*, pages 304–319, Berlin, Heidelberg, 2022. Springer, Springer-Verlag.

[207] S. Nisioi, S. Štajner, S. P. Ponzetto, and L. P. Dinu. Exploring neural text simplification models. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 85–91, Vancouver, Canada, 2017. Association for Computational Linguistics.

[208] M. G. Obiorah, A. M. M. Piper, and M. Horn. Designing aacs for people with aphasia dining in restaurants. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[209] J. Ockerman and A. Pritchett. A review and reappraisal of task guidance: Aiding workers in procedure following. *International Journal of Cognitive Ergonomics*, 4(3):191–212, 2000.

[210] J. Orlosky, K. Kiyokawa, and H. Takemura. Managing mobile text in head mounted displays: studies on visual preference and text placement. *ACM SIGMOBILE Mobile Computing and Communications Review*, 18(2):20–31, 2014.

[211] G. H. Paetzold and L. Specia. A survey on lexical simplification. *Journal of Artificial Intelligence Research*, 60:549–593, 2017.

[212] R. Palmarini, J. A. Erkoyuncu, R. Roy, and H. Torabmostaedi. A systematic review of augmented reality applications in maintenance. *Robotics and Computer-Integrated Manufacturing*, 49:215–228, 2018.

[213] J. Pan, Z. Lin, Y. Ge, X. Zhu, R. Zhang, Y. Wang, Y. Qiao, and H. Li. Retrieving-to-answer: Zero-shot video question answering with frozen large language models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023 - Workshops, Paris, France, October 2-6, 2023*, pages 272–283. IEEE, 2023.

[214] R. Panda, N. C. Mithun, and A. K. Roy-Chowdhury. Diversity-aware multi-video summarization. *IEEE Transactions on Image Processing*, 26(10):4712–4724, 2017.

[215] P. Papalampidi, F. Keller, and M. Lapata. Movie summarization via sparse graph construction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13631–13639, 2020.

[216] P. Parekh, S. Patel, N. Patel, and M. Shah. Systematic review and meta-analysis of augmented reality in medicine, retail, and games. *Visual computing for industry, biomedicine, and art*, 3:1–20, 2020.

[217] J. Park, J. Lee, I.-J. Kim, and K. Sohn. Sumgraph: Video summarization via recursive graph modeling. In *Proceedings of the European Conference on Computer Vision*, pages 647–663, 2020.

[218] S. Park, S. Bokijonov, and Y. Choi. Review of microsoft hololens applications over the past five years. *Applied sciences*, 11(16):7259, 2021.

[219] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32:8026–8037, 2019.

[220] A. Patel, D. Turakhia, O. Brunner, A. Mertens, and J. Liu. Benchmarking open-endedness of vr training systems in manufacturing. In *2024 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pages 64–70. IEEE, 2024.

[221] A. Pauchet, N. Chaignaud, and A. El Fallah Seghrouchni. A computational model of human interaction and planning for heterogeneous multi-agent systems. In *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*, pages 1–3, 2007.

[222] R. Peddi, S. Arya, B. Challa, L. Pallapothula, A. Vyas, B. Gouripeddi, Q. Zhang, J. Wang, V. Komaragiri, E. D. Ragan, N. Ruozzi, Y. Xiang, and V. Gogate. Captaincook4d: A dataset for understanding errors in procedural activities. In A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

[223] V. Pejovic and M. Musolesi. Anticipatory mobile computing: A survey of the state of the art and research challenges. *ACM Computing Surveys (CSUR)*, 47(3):1–29, 2015.

[224] T. Perrett, A. Darkhalil, S. Sinha, O. Emara, S. Pollard, K. Parida, K. Liu, P. Gatti, S. Bansal, K. Flanagan, J. Chalk, Z. Zhu, R. Guerrier, F. Abdelazim, B. Zhu, D. Moltisanti, M. Wray, H. Doughty, and D. Damen. HD-EPIC: A highly-detailed egocentric video dataset. *CoRR*, abs/2502.04144, 2025.

[225] W. Piekarski and B. Thomas. Arquake: the outdoor augmented reality gaming system. *Communications of the ACM*, 45(1):36–38, 2002.

[226] N. Porot and E. Mandelbaum. The science of belief: A progress report. *Wiley Interdisciplinary Reviews: Cognitive Science*, 12(2):e1539, 2021.

[227] D. Premack and G. Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526, 1978.

[228] J. Qian, D. A. Shamma, D. Avrahami, and J. Biehl. Modality and depth in touchless smartphone augmented reality interactions. In *Proceedings of the 2020 ACM International Conference on Interactive Media Experiences*, IMX '20, page 74–81, New York, NY, USA, 2020. Association for Computing Machinery.

[229] J. Qian, Q. Sun, C. Wigington, H. L. Han, T. Sun, J. Healey, J. Tompkin, and J. Huang. Dually noted: Layout-aware annotations with smartphone augmented reality. In S. D. J. Barbosa, C. Lampe, C. Appert, D. A. Shamma, S. M. Drucker, J. R. Williamson, and K. Yatani, editors, *CHI '22: CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 29 April 2022 - 5 May 2022*, pages 552:1–552:15. ACM, 2022.

[230] J. Qian, T. Zhou, M. Young-Ng, J. Ma, A. Cheung, X. Li, I. Gonsher, and J. Huang. Portalware: Exploring free-hand ar drawing with a dual-display smartphone-wearable paradigm. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference*, DIS '21, page 205–219, New York, NY, USA, 2021. Association for Computing Machinery.

[231] L. Qian, A. Deguet, and P. Kazanzides. Arssist: augmented reality on a head-mounted display for the first assistant in robotic surgery. *Healthcare technology letters*, 5(5):194–200, 2018.

[232] R. C. Quesada and Y. Demiris. Proactive robot assistance: Affordance-aware augmented reality user interfaces. *IEEE Robotics & Automation Magazine*, 29(1):22–34, 2022.

[233] J. C. Quiroz, E. Geangu, and M. H. Yong. Emotion recognition using smart watch sensor data: Mixed-design study. *JMIR mental health*, 5(3):e10153, 2018.

[234] M. Rabbi, M. H. Aung, M. Zhang, and T. Choudhury. Mybehavior: automatic personalized health feedback from user behaviors and preferences using smartphones. In K. Mase, M. Langheinrich, D. Gatica-Perez, H. Gellersen, T. Choudhury, and K. Yatani, editors, *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp 2015, Osaka, Japan, September 7-11, 2015*, pages 707–718. ACM, 2015.

[235] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021.

[236] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[237] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2020.

[238] A. S. Rao. Agentspeak(l): BDI agents speak out in a logical computable language. In W. V. de Velde and J. W. Perram, editors, *Agents Breaking Away, 7th European Workshop on Modelling Autonomous Agents in a Multi-Agent World, Eindhoven, The Netherlands, January 22-25, 1996, Proceedings*, volume 1038 of *Lecture Notes in Computer Science*, pages 42–55. Springer, 1996.

[239] A. S. Rao and M. P. Georgeff. Modeling rational agents within a bdi-architecture. *Readings in agents*, pages 317–328, 1997.

[240] A. S. Rao and M. P. Georgeff. Decision procedures for BDI logics. *J. Log. Comput.*, 8(3):293–342, 1998.

[241] P.-L. P. Rau, J. Zheng, Z. Guo, and J. Li. Speed reading on virtual reality and augmented reality. *Computers & Education*, 125:240–245, 2018.

[242] L. Rello and R. Baeza-Yates. How to present more readable text for people with dyslexia. *Universal Access in the Information Society*, 16:29–49, 2017.

[243] L. Rello, R. Baeza-Yates, S. Bott, and H. Saggion. Simplify or help? text simplification strategies for people with dyslexia. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*, pages 1–10, New York, NY, USA, 2013. Association for Computing Machinery.

[244] L. Rello, R. Baeza-Yates, L. Dempere-Marco, and H. Saggion. Frequent words improve readability and short words improve understandability for people with dyslexia. In *Human-Computer Interaction – INTERACT 2013*, pages 203–219, Berlin, Heidelberg, 2013. Springer, Springer Berlin Heidelberg.

[245] M. Ren, L. Dong, Z. Xia, J. Cong, and P. Zheng. A proactive interaction design method for personalized user context prediction in smart-product service system. *Procedia CIRP*, 119:963–968, 2023. The 33rd CIRP Design Conference.

[246] I. Rodin, A. Furnari, D. Mavroeidis, and G. M. Farinella. Predicting the future from first person (egocentric) vision: A survey. *Comput. Vis. Image Underst.*, 211:103252, 2021.

[247] L. Ruan and Q. Jin. Survey: Transformer based video-language pre-training. *AI Open*, 3:1–13, 2022.

[248] R. Rzayev, P. W. Woźniak, T. Dingler, and N. Henze. Reading on smart glasses: The effect of text position, presentation type and walking. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–9, 2018.

[249] P. Sahoo, A. K. Singh, S. Saha, V. Jain, S. Mondal, and A. Chadha. A systematic survey of prompt engineering in large language models: Techniques and applications. *CoRR*, abs/2402.07927, 2024.

[250] G. Sara, G. Todde, and M. Caria. Assessment of video see-through smart glasses for augmented reality to support technicians during milking machine maintenance. *Scientific Reports*, 12(1):15729, 2022.

[251] R. Sarikaya. The technology behind personal digital assistants: An overview of the system architecture and key components. *IEEE Signal Process. Mag.*, 34(1):67–81, 2017.

[252] C. R. Sauer and P. Burggräf. Hybrid intelligence–systematic approach and framework to determine the level of human-ai collaboration for production management use cases. *Production Engineering*, pages 1–17, 2024.

[253] T. J. Saun, K. J. Zuo, and T. P. Grantcharov. Video technologies for recording open surgery: a systematic review. *Surgical innovation*, 26(5):599–612, 2019.

[254] C. Scarton, A. P. Aprosio, S. Tonelli, T. Martín-Wanton, and L. Specia. MUSST: A multilingual syntactic simplification tool. In S. Park and T. Supnithi, editors, *Proceedings of the IJCNLP 2017, System Demonstrations*, pages 25–28, Tapei, Taiwan, Nov. 2017. Association for Computational Linguistics.

[255] A. Schmeil and W. Broll. Mara-a mobile augmented reality-based virtual assistant. In *2007 IEEE Virtual Reality Conference*, pages 267–270. IEEE, IEEE, 2007.

[256] A. J. Schmid, O. Weede, and H. Worn. Proactive robot task selection given a human intention estimate. In *RO-MAN 2007 - The 16th IEEE International Symposium on Robot and Human Interactive Communication*, pages 726–731, 2007.

[257] B. Schmidt, S. Benchea, R. Eichin, and C. Meurisch. Fitness tracker or digital personal coach: how to personalize training. In K. Mase, M. Langheinrich, D. Gatica-Perez, H. Gellersen, T. Choudhury, and K. Yatani, editors, *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers, UbiComp/ISWC Adjunct 2015, Osaka, Japan, September 7-11, 2015*, pages 1063–1067. ACM, 2015.

[258] M. Schmidt, W. Minker, and S. Werner. How users react to proactive voice assistant behavior while driving. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 485–490. European Language Resources Association, 2020.

[259] P. M. Scholl, M. Wille, and K. V. Laerhoven. Wearables in the wet lab: a laboratory system for capturing and guiding experiments. In K. Mase, M. Langheinrich, D. Gatica-Perez, H. Gellersen, T. Choudhury, and K. Yatani, editors, *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp 2015, Osaka, Japan, September 7-11, 2015*, pages 589–599. ACM, 2015.

[260] J. Schöning and G. Heidemann. Visual video analytics for interactive video content analysis. In *Advances in Information and Communication Networks: Proceedings of the 2018 Future of Information and Communication Conference (FICC), Vol. 1*, pages 346–360. Springer, 2019.

[261] T. J. Schoonbeek, T. Houben, H. Onvlee, P. H. N. de With, and F. van der Sommen. Industreal: A dataset for procedure step recognition handling execution errors in egocentric videos in an industrial-like setting, 2023.

[262] T. J. Schoonbeek, T. Houben, H. Onvlee, P. H. N. de With, and F. van der Sommen. Industreal: A dataset for procedure step recognition handling execution errors in egocentric videos in an industrial-like setting. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2024, Waikoloa, HI, USA, January 3-8, 2024*, pages 4353–4362. IEEE, 2024.

[263] A. Seeliger, R. P. Weibel, and S. Feuerriegel. Context-adaptive visual cues for safe navigation in augmented reality using machine learning. *International Journal of Human–Computer Interaction*, 40(3):761–781, 2024.

[264] F. Sepulveda, Y. Fan, C. Gabbianelli, K. Studer, D. Turakhia, and J. Liu. Using act-r architecture in the design of intelligent tutoring systems for vr training of manufacturing skills. In *2024 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pages 87–92. IEEE, 2024.

[265] M. Shardlow. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*, 4(1):58–70, 2014.

[266] A. Sharghi, J. S. Laurel, and B. Gong. Query-focused video summarization: dataset, evaluation, and a memory network based approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2127–2136, 2017.

[267] J. Shen, J. J. Dudley, and P. O. Kristensson. Encode-store-retrieve: Enhancing memory augmentation through language-encoded egocentric perception. *CoRR*, abs/2308.05822, 2023.

[268] J. Shen, J. J. Dudley, and P. O. Kristensson. Encode-store-retrieve: Augmenting human memory through language-encoded egocentric perception. In U. Eck, M. Sra, J. K. Stefanucci, M. Sugimoto, M. Tatzgern, and I. Williams, editors, *IEEE International Symposium on Mixed and Augmented Reality, ISMAR 2024, Bellevue, WA, USA, October 21-25, 2024*, pages 923–931. IEEE, 2024.

[269] N.-J. Shih, H.-X. Chen, T.-Y. Chen, and Y.-T. Qiu. Digital preservation and reconstruction of old cultural elements in augmented reality (ar). *Sustainability*, 12(21):9262, 2020.

[270] K. Shuster, M. Komeili, L. Adolphs, S. Roller, A. Szlam, and J. Weston. Language models that seek for knowledge: Modular search & generation for dialogue and prompt completion. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 373–393. Association for Computational Linguistics, 2022.

[271] C. Si, W. Shi, C. Zhao, L. Zettlemoyer, and J. Boyd-Graber. Getting MoRE out of mixture of language model reasoning experts. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8234–8249, Singapore, Dec. 2023. Association for Computational Linguistics.

[272] C. Si, C. Zhao, S. Min, and J. Boyd-Graber. Re-examining calibration: The case of question answering. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2814–2829, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics.

[273] A. Siddharthan. Syntactic simplification and text cohesion. *Research on Language and Computation*, 4:77–109, 2006.

[274] A. Siddharthan. A survey of research on text simplification. *ITL-International Journal of Applied Linguistics*, 165(2):259–298, 2014.

[275] I. Singh, V. Blukis, A. Mousavian, A. Goyal, D. Xu, J. Tremblay, D. Fox, J. Thomason, and A. Garg. Progprompt: Generating situated robot task plans using large language models. *CoRR*, abs/2209.11302, 2022.

[276] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes. Tvsum: Summarizing web videos using titles. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 5179–5187. IEEE Computer Society, 2015.

[277] G. Sriramanan, S. Bharti, V. S. Sadasivan, S. Saha, P. Kattakinda, and S. Feizi. Llm-check: Investigating detection of hallucinations in large language models. In A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

[278] K. Studer, H. Lie, Z. Zhao, B. Thomson, D. G. Turakhia, and J. Liu. An open-ended system in virtual reality for training machining skills. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–5, 2024.

[279] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*, 2020.

[280] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7464–7473, 2019.

[281] W. Sun, P. Ren, and Z. Ren. Generative knowledge selection for knowledge-grounded dialogues. In A. Vlachos and I. Augenstein, editors, *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2077–2088, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[282] X. Suo, W. Tang, and Z. Li. Motion capture technology in sports scenarios: A survey. *Sensors*, 24(9):2947, 2024.

[283] A. Syberfeldt, O. Danielsson, M. Holm, and L. Wang. Visual assembling guidance using augmented reality. *Procedia Manufacturing*, 1:98–109, 2015.

[284] A. Tang, C. Owen, F. Biocca, and W. Mou. Comparative effectiveness of augmented reality in object assembly. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 73–80, New York, NY, USA, 2003. Association for Computing Machinery.

[285] H. Tang, K. J. Liang, K. Grauman, M. Feiszli, and W. Wang. Egotracks: A long-term egocentric visual object tracking dataset. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[286] S. Tang, D. Roberts, and M. Golparvar-Fard. Human-object interaction recognition for automatic construction site safety inspection. *Automation in Construction*, 120:103356, 2020.

[287] Y. Tang, J. Situ, A. Y. Cui, M. Wu, and Y. Huang. LLM integration in extended reality: A comprehensive review of current trends, challenges, and future perspectives. In N. Yamashita, V. Evers, K. Yatani, S. X. Ding, B. Lee, M. Chetty, and P. O. T. Dugas, editors, *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, CHI 2025, YokohamaJapan, 26 April 2025- 1 May 2025*, pages 1054:1–1054:24. ACM, 2025.

[288] L. Tatwany and H. C. Ouertani. A review on using augmented reality in text translation. In *2017 6th International Conference on Information and Communication Technology and Accessibility (ICTA)*, pages 1–6. IEEE, 2017.

[289] P. Taveekitworachai, F. Abdullah, and R. Thawonmas. Null-shot prompting: Rethinking prompting large language models with hallucination. In Y. Al-Onaizan, M. Bansal, and Y. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 13321–13361. Association for Computational Linguistics, 2024.

[290] X. Team. Xagent: An autonomous agent for complex task solving, 2023.

[291] B. H. Thomas, G. F. Welch, P. Dragicevic, N. Elmqvist, P. Irani, Y. Jansen, D. Schmalstieg, A. Tabard, N. A. ElSayed, R. T. Smith, et al. Situated analytics. *Immersive analytics*, 11190:185–220, 2018.

[292] P. C. Thomas and W. David. Augmented reality: An application of heads-up display technology to manual manufacturing processes. In *Hawaii international conference on system sciences*, volume 2. ACM SIGCHI Bulletin, 1992.

[293] Y. Tian, Y. Ma, S. Quan, and Y. Xu. Occlusion and collision aware smartphone ar using time-of-flight camera. In *Advances in Visual Computing*, pages 141–153, Cham, 2019. Springer, Springer International Publishing.

[294] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. Canton-Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288, 2023.

[295] C. Truong-Allié, A. Paljic, A. Roux, and M. Herbeth. User behavior adaptive ar guidance for wayfinding and tasks completion. *Multimodal Technologies and Interaction*, 5(11):65, 2021.

[296] D. Turakhia, P. Jiang, B. Liu, M. Leake, and S. Müller. The reflective maker: Using reflection to support skill-learning in makerspaces. In M. Agrawala, J. O. Wobbrock, E. Adar, and V. Setlur, editors, *The Adjunct Publication of the 35th Annual ACM Symposium on User Interface Software and Technology, UIST 2022, Bend, OR, USA, 29 October 2022- 2 November 2022*, pages 28:1–28:4. ACM, 2022.

[297] D. Turakhia, Z. Mroue, P. Jiang, and S. Mueller. Generating Reflection Prompts in Self-Directed Learning Activities with Generative AI. *An MIT Exploration of Generative AI*, sep 10 2024. https://mit-genai.pubpub.org/pub/kju0447a.

[298] D. G. Turakhia, P. Jiang, and S. Mueller. The reflective make-ar in-action: Using augmented reality for reflection-based learning of makerskills. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI EA '23, New York, NY, USA, 2023. Association for Computing Machinery.

[299] D. G. Turakhia, P. Jiang, and S. Müller. The reflective make-ar in-action: Using augmented reality for reflection-based learning of makerskills. In A. Schmidt, K. Väänänen, T. Goyal, P. O. Kristensson, and A. Peters, editors, *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems, CHI EA 2023, Hamburg, Germany, April 23-28, 2023*, pages 276:1–276:6. ACM, 2023.

[300] D. G. Turakhia, Y. Qi, L.-G. Blumberg, A. Wong, and S. Mueller. Can physical tools that adapt their shape based on a learner's performance help in motor skill training? In *Proceedings of the Fifteenth International Conference on Tangible, Embedded, and Embodied Interaction*, TEI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[301] A. E. Uva, M. Gattullo, V. M. Manghisi, D. Spagnulo, G. L. Cascella, and M. Fiorentino. Evaluating the effectiveness of spatial augmented reality in smart manufacturing: a solution for manual working stations. *The International Journal of Advanced Manufacturing Technology*, 94:509–521, 2018.

[302] K. P. Vaubel and C. F. Gettys. Inferring user expertise for adaptive interfaces. *Hum. Comput. Interact.*, 5(1):95–117, 1990.

[303] C. Vettori and O. Mich. Supporting deaf children's reading skills: the many challenges of text simplification. In *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility*, ASSETS '11, page 283–284, New York, NY, USA, 2011. Association for Computing Machinery.

[304] VIDA-NYU. ptgctl: A Python Library and Command Line Tool for the PTG API. `https://github.com/VIDA-NYU/ptgctl`, 2024. Available online: `https://github.com/VIDA-NYU/ptgctl`.

[305] S. Vlahovic, M. Suznjevic, and L. Skorin-Kapov. A survey of challenges and methods for quality of experience assessment of interactive VR applications. *J. Multimodal User Interfaces*, 16(3):257–291, 2022.

[306] B. Wang, X. Deng, and H. Sun. Iteratively prompt pre-trained language models for chain of thought. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 2714–2730. Association for Computational Linguistics, 2022.

[307] D. Wang, E. F. Churchill, P. Maes, X. Fan, B. Shneiderman, Y. Shi, and Q. Wang. From human-human collaboration to human-ai collaboration: Designing AI systems that can work together with people. In R. Bernhaupt, F. F. Mueller, D. Verweij, J. Andres, J. McGrenere, A. Cockburn, I. Avellino, A. Goguey, P. Bjøn, S. Zhao, B. P. Samson, and R. Kocielnik, editors, *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems, CHI 2020, Honolulu, HI, USA, April 25-30, 2020*, pages 1–6. ACM, 2020.

[308] Q. Wang, R. Ji, T. Peng, W. Wu, Z. Li, and J. Liu. Soft knowledge prompt: Help external knowledge become a better teacher to instruct LLM in knowledge-based VQA. In L. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 6132–6143. Association for Computational Linguistics, 2024.

[309] T. Wang, T. Yang, M. Danelljan, F. S. Khan, X. Zhang, and J. Sun. Learning human-object interaction detection using interaction points. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 4115–4124. Computer Vision Foundation / IEEE, 2020.

[310] X. Wang, H. Kim, S. Rahman, K. Mitra, and Z. Miao. Human-llm collaborative annotation through effective verification of LLM labels. In F. F. Mueller, P. Kyburz, J. R. Williamson, C. Sas, M. L. Wilson, P. O. T. Dugas, and I. Shklovski, editors, *Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI 2024, Honolulu, HI, USA, May 11-16, 2024*, pages 303:1–303:21. ACM, 2024.

[311] X. Wang, T. Kwon, M. Rad, B. Pan, I. Chakraborty, S. Andrist, D. Bohus, A. Feniello, B. Tekin, F. V. Frujeri, N. Joshi, and M. Pollefeys. Holoassist: an egocentric human interaction dataset for interactive AI assistants in the real world. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 20213–20224. IEEE, 2023.

[312] X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. H. Chi, S. Narang, A. Chowdhery, and D. Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[313] X. Wang, Y. Zhang, O. Zohar, and S. Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, editors, *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXX*, volume 15138 of *Lecture Notes in Computer Science*, pages 58–76. Springer, 2024.

[314] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics*, 38(5):1–12, 2019.

[315] Z. Wang, S. Cai, A. Liu, X. Ma, and Y. Liang. Describe, explain, plan and select: interactive planning with llms enables open-world multi-task agents. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[316] M. Wasim, I. Ahmed, J. Ahmad, M. Nawaz, E. Alabdulkreem, and Y. Ghadi. A video summarization framework based on activity attention modeling using deep features for smart campus surveillance system. *PeerJ Comput. Sci.*, 8:e911, 2022.

[317] M. Weerasinghe, A. Quigley, K. Č. Pucihar, A. Toniolo, A. Miguel, and M. Kljun. Arigatō: Effects of adaptive guidance on engagement and performance in augmented reality learning environments. *IEEE Transactions on Visualization and Computer Graphics*, 28(11):3737–3747, 2022.

[318] J. Wei, X. Wang, D. Schuurmans, M. Bosma, b. ichter, F. Xia, E. Chi, Q. V. Le, and D. Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc., 2022.

[319] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022.

[320] R. Wen, L. Yang, C.-K. Chui, K.-B. Lim, and S. Chang. *Intraoperative visual guidance and control interface for augmented reality robotic surgery.* IEEE, 2010.

[321] M. M. Wloka and B. G. Anderson. Resolving occlusion in augmented reality. In *Proceedings of the 1995 Symposium on Interactive 3D Graphics*, pages 5–12, New York, NY, USA, 1995. Association for Computing Machinery.

[322] J. Wolf. *Towards advanced user guidance and context awareness in augmented reality-guided procedures.* PhD thesis, ETH Zurich, 2022.

[323] G. Wu, J. Lin, and C. T. Silva. Intentvizor: Towards generic query guided interactive video summarization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10493–10502. IEEE, 2022.

[324] G. Wu, J. P. Lin, and C. T. Silva. ERA: entity-relationship aware video summarization with wasserstein GAN. In *32nd British Machine Vision Conference 2021, BMVC 2021, Online, November 22-25, 2021*, page 321. BMVA Press, 2021.

[325] G. Wu, J. Qian, S. C. Quispe, S. Chen, J. Rulff, and C. T. Silva. Artist: Automated text simplification for task guidance in augmented reality. In F. F. Mueller, P. Kyburz, J. R. Williamson, C. Sas, M. L. Wilson, P. O. T. Dugas, and I. Shklovski, editors, *Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI 2024, Honolulu, HI, USA, May 11-16, 2024*, pages 939:1–939:24. ACM, 2024.

[326] G. Wu, C. Zhao, C. T. Silva, and H. He. Your co-workers matter: Evaluating collaborative capabilities of language models in blocks world. In L. Ku, A. Martins, and V. Srikumar, editors, *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 4941–4957. Association for Computational Linguistics, 2024.

[327] H.-K. Wu, S. W.-Y. Lee, H.-Y. Chang, and J.-C. Liang. Current status, opportunities and challenges of augmented reality in education. *Computers & education*, 62:41–49, 2013.

[328] J. Wu, Z. Chen, J. Deng, S. Sabour, H. Meng, and M. Huang. COKE: A cognitive knowledge graph for machine theory of mind. In L. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 15984–16007. Association for Computational Linguistics, 2024.

[329] Q. Wu, G. Bansal, J. Zhang, Y. Wu, S. Zhang, E. Zhu, B. Li, L. Jiang, X. Zhang, and C. Wang. Autogen: Enabling next-gen LLM applications via multi-agent conversation framework. *CoRR*, abs/2308.08155, 2023.

[330] Z. Xi, W. Chen, X. Guo, W. He, Y. Ding, B. Hong, M. Zhang, J. Wang, S. Jin, E. Zhou, R. Zheng, X. Fan, X. Wang, L. Xiong, Y. Zhou, W. Wang, C. Jiang, Y. Zou, X. Liu, Z. Yin, S. Dou, R. Weng, W. Cheng, Q. Zhang, W. Qin, Y. Zheng, X. Qiu, X. Huan, and T. Gui. The rise and potential of large language model based agents: A survey. *CoRR*, abs/2309.07864, 2023.

[331] J. Xiao, R. Catrambone, and J. T. Stasko. Be quiet? evaluating proactive and reactive user interface assistants. In M. Rauterberg, M. Menozzi, and J. Wesson, editors, *Human-Computer Interaction INTERACT '03: IFIP TC13 International Conference on Human-Computer Interaction, 1st-5th September 2003, Zurich, Switzerland.* IOS Press, 2003.

[332] S. Xiao, Z. Zhao, Z. Zhang, Z. Guan, and D. Cai. Query-biased self-attentive network for query-focused video summarization. *IEEE Transactions on Image Processing*, 29:5889–5899, 2020.

[333] S. Xiao, Z. Zhao, Z. Zhang, X. Yan, and M. Yang. Convolutional hierarchical attention network for query-focused video summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12426–12433, 2020.

[334] J. Xiong, E.-L. Hsiang, Z. He, T. Zhan, and S.-T. Wu. Augmented reality and virtual reality displays: emerging technologies and future perspectives. *Light: Science & Applications*, 10(1):216, 2021.

[335] S. Xiong, A. Payani, R. Kompella, and F. Fekri. Large language models can learn temporal reasoning. In L. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 10452–10470. Association for Computational Linguistics, 2024.

[336] J. Xu, Z. Li, W. Chen, Q. Wang, X. Gao, Q. Cai, and Z. Ling. On-device language models: A comprehensive review. *CoRR*, abs/2409.00088, 2024.

[337] M. Xu, C. Zhao, D. S. Rojas, A. Thabet, and B. Ghanem. G-tad: Sub-graph localization for temporal action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10156–10165, 2020.

[338] S. Xu, C. Chen, Z. Liu, X. Jin, L. Yuan, Y. Yan, and H. Qu. Memory reviver: Supporting photo-collection reminiscence for people with visual impairment via a proactive chatbot. In L. Yao, M. Goel, A. Ion, and P. Lopes, editors, *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology, UIST 2024, Pittsburgh, PA, USA, October 13-16, 2024*, pages 88:1–88:17. ACM, 2024.

[339] S. B. Yadav. A conceptual model for user-centered quality information retrieval on the world wide web. *J. Intell. Inf. Syst.*, 35(1):91–121, 2010.

[340] G. Yalınız and N. Ikizler-Cinbis. Unsupervised video summarization with independently recurrent neural networks. In *27th Signal Processing and Communications Applications Conference (SIU)*, pages 1–4, 2019.

[341] V. Yaneva. Easy-read documents as a gold standard for evaluation of text simplification output. In *Proceedings of the Student Research Workshop*, pages 30–36, Hissar, Bulgaria, Sept. 2015. INCOMA Ltd. Shoumen, BULGARIA.

[342] D. Yang, S. Huang, Z. Xu, Z. Li, S. Wang, M. Li, Y. Wang, Y. Liu, K. Yang, Z. Chen, Y. Wang, J. Liu, P. Zhang, P. Zhai, and L. Zhang. AIDE: A vision-driven multi-view, multi-modal, multi-tasking dataset for assistive driving perception. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 20402–20413. IEEE, 2023.

[343] D. Yang, K. Yang, M. Li, S. Wang, S. Wang, and L. Zhang. Robust emotion recognition in context debiasing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 12447–12457. IEEE, 2024.

[344] J. Yang, S. Liu, H. Guo, Y. Dong, X. Zhang, S. Zhang, P. Wang, Z. Zhou, B. Xie, Z. Wang, B. Ouyang, Z. Lin, M. Cominelli, Z. Cai, Y. Zhang, P. Zhang, F. Hong, J. Widmer, F. Gringoli, L. Yang, B. Li, and Z. Liu. Egolife: Towards egocentric life assistant. *CoRR*, abs/2503.03803, 2025.

[345] Y. Yang, Y. Li, and X. Quan. UBAR: towards fully end-to-end task-oriented dialog system with GPT-2. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14230–14238. AAAI Press, 2021.

[346] S. Yao, H. Chen, J. Yang, and K. Narasimhan. Webshop: Towards scalable real-world web interaction with grounded language agents. In *NeurIPS*, 2022.

[347] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. R. Narasimhan, and Y. Cao. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[348] X. Ye and G. Durrett. Can explanations be useful for calibrating black box models? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 6199–6212. Association for Computational Linguistics, 2022.

[349] X. Ye and G. Durrett. The unreliability of explanations in few-shot prompting for textual reasoning. In *Advances in Neural Information Processing Systems*, volume 35, pages 30378–30392. Curran Associates, Inc., 2022.

[350] X. Yin and S. Shah. Building taxonomy of web search intents for name entity queries. In *Proceedings of the 19th International Conference on World Wide Web*, pages 1001–1010, 2010.

[351] N. Yorke-Smith, S. Saadati, K. L. Myers, and D. N. Morley. The design of a proactive personal agent for task management. *Int. J. Artif. Intell. Tools*, 21(1), 2012.

[352] K. P. Yu, Z. Zhang, F. Hu, and J. Chai. Efficient in-context learning in vision-language models for egocentric videos. *CoRR*, abs/2311.17041, 2023.

[353] H. Zamani and W. B. Croft. Estimating embedding vectors for queries. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*, pages 123–132, 2016.

[354] N. Zargham, L. Reicherts, M. Bonfert, S. T. Voelkel, J. Schöning, R. Malaka, and Y. Rogers. Understanding circumstances for desirable proactive behaviour of voice assistants: The proactivity dilemma. In M. Halvey, M. E. Foster, J. Dalton, C. Munteanu, and J. Trippas, editors, *CUI 2022: 4th Conference on Conversational User Interfaces, Glasgow, United Kingdom, July 26 - 28, 2022*, pages 3:1–3:14. ACM, 2022.

[355] H. Zeng, X. Wang, A. Wu, Y. Wang, Q. Li, A. Endert, and H. Qu. Emoco: Visual analysis of emotion coherence in presentation videos. *IEEE Trans. Vis. Comput. Graph.*, 26(1):927–937, 2020.

[356] K.-H. Zeng, T.-H. Chen, C.-Y. Chuang, Y.-H. Liao, J. C. Niebles, and M. Sun. Leveraging video descriptions to learn video question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

[357] K. Zhai, Y. Cao, W. Hou, and X. Li. Interactive mixed reality cooking assistant for unskilled operating scenario. In J. Y. C. Chen and G. Fragomeni, editors, *Virtual, Augmented and Mixed Reality. Industrial and Everyday Life Applications - 12th International Conference, VAMR 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19-24, 2020, Proceedings, Part II*, volume 12191 of *Lecture Notes in Computer Science*, pages 178–195. Springer, 2020.

[358] D. Zhang, X. Dai, and Y. Wang. Dynamic temporal pyramid network: A closer look at multi-scale modeling for activity detection. In C. V. Jawahar, H. Li, G. Mori, and K. Schindler, editors, *Computer Vision - ACCV 2018 - 14th Asian Conference on Computer Vision, Perth, Australia, December 2-6, 2018, Revised Selected Papers, Part IV*, volume 11364 of *Lecture Notes in Computer Science*, pages 712–728. Springer, 2018.

[359] D. Zhang, Y. Li, Z. He, and X. Li. Empowering smart glasses with large language models: Towards ubiquitous agi. In *Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 631–633, 2024.

[360] H. Zhang, X. Song, C. Xiong, C. Rosset, P. N. Bennett, N. Craswell, and S. Tiwary. Generic intent representation in web search. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 65–74, 2019.

[361] J. Zhang, X. Xu, and S. Deng. Exploring collaboration mechanisms for LLM agents: A social psychology view. *CoRR*, abs/2310.02124, 2023.

[362] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman. Video summarization with long short-term memory. In *Proceedings of the European Conference on Computer Vision*, pages 766–782, 2016.

[363] S. Zhang, Z. Chen, Y. Shen, M. Ding, J. B. Tenenbaum, and C. Gan. Planning with large language models for code generation. In *The Eleventh International Conference on Learning Representations*. OpenReview.net, 2023.

[364] S. Zhang, C. Gong, and E. Choi. Knowing more about questions can help: Improving calibration in question answering. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 1958–1970, Online, 2021. Association for Computational Linguistics.

[365] S. Zhang, Y. Zhu, and A. K. Roy-Chowdhury. Context-aware surveillance video summarization. *IEEE Transactions on Image Processing*, 25(11):5469–5478, 2016.

[366] Y. Zhang, M. Kampffmeyer, X. Liang, M. Tan, and E. P. Xing. Query-conditioned three-player adversarial network for video summarization. In *29th British Machine Vision Conference*, 2018.

[367] Y. Zhang, Z. Ou, and Z. Yu. Task-oriented dialog systems that consider multiple appropriate responses under the same context. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9604–9611. AAAI Press, 2020.

[368] Z. Zhang. Design for supporting reflection in design-based learning. 2023.

[369] B. Zhao, X. Li, and X. Lu. Hsa-rnn: Hierarchical structure-adaptive rnn for video summarization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7405–7414, 2018.

[370] B. Zhao, X. Li, and X. Lu. Property-constrained dual learning for video summarization. *IEEE Transactions on Neural Networks and Learning Systems*, 31(10):3989–4000, 2019.

[371] J. Zhao, M. Karimzadeh, L. S. Snyder, C. Surakitbanharn, Z. C. Qian, and D. S. Ebert. Metricsvis: A visual analytics system for evaluating employee performance in public safety agencies. *IEEE Trans. Vis. Comput. Graph.*, 26(1):1193–1203, 2020.

[372] Z. Zhao, E. Wallace, S. Feng, D. Klein, and S. Singh. Calibrate before use: Improving few-shot performance of language models. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR, 2021.

[373] C. Zheng, Y. Zhang, Z. Huang, C. Shi, M. Xu, and X. Ma. Disciplink: Unfolding interdisciplinary information seeking process via human-ai co-exploration. In L. Yao, M. Goel, A. Ion, and P. Lopes, editors, *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology, UIST 2024, Pittsburgh, PA, USA, October 13-16, 2024*, pages 91:1–91:20. ACM, 2024.

[374] T. Zheng, M. Ardolino, A. Bacchetti, and M. Perona. The applications of industry 4.0 technologies in manufacturing context: a systematic literature review. *International Journal of Production Research*, 59(6):1922–1954, 2021.

[375] X. S. Zheng, C. Foucault, P. Matos da Silva, S. Dasari, T. Yang, and S. Goose. Eye-wearable technology for machine maintenance: Effects of display position and hands-free operation. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 2125–2134, New York, NY, USA, 2015. Association for Computing Machinery.

[376] Y. Zheng, Y. Chen, B. Qian, X. Shi, Y. Shu, and J. Chen. A review on edge large language models: Design, execution, and applications. *ACM Computing Surveys*, 57(8):1–35, 2025.

[377] D. Zhou, N. Schärli, L. Hou, J. Wei, N. Scales, X. Wang, D. Schuurmans, C. Cui, O. Bousquet, Q. V. Le, and E. H. Chi. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[378] K. Zhou, Y. Qiao, and T. Xiang. Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pages 7582–7589, 2018.

[379] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. Corso, and J. Gao. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13041–13049, 2020.

[380] S. Zhou, F. F. Xu, H. Zhu, X. Zhou, R. Lo, A. Sridhar, X. Cheng, Y. Bisk, D. Fried, U. Alon, and G. Neubig. Webarena: A realistic web environment for building autonomous agents. *CoRR*, abs/2307.13854, 2023.

[381] X. Zhou, R. Girdhar, A. Joulin, P. Krähenbühl, and I. Misra. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, pages 350–368, Cham, 2022. Springer, Springer Nature Switzerland.

[382] X. Zhou, H. Zhu, L. Mathur, R. Zhang, H. Yu, Z. Qi, L. Morency, Y. Bisk, D. Fried, G. Neubig, and M. Sap. SOTOPIA: interactive evaluation for social intelligence in language agents. *CoRR*, abs/2310.11667, 2023.

[383] W. Zhu, J. Lu, J. Li, and J. Zhou. Dsnet: A flexible detect-to-summarize network for video summarization. *IEEE Transactions on Image Processing*, 30:948–962, 2020.

[384] H. Zhuang, W. E. Zhang, L. Xie, W. Chen, J. Yang, and Q. Sheng. Automatic, meta and human evaluation for multimodal summarization with multimodal output. In K. Duh, H. Gómez-Adorno, and S. Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 7768–7790. Association for Computational Linguistics, 2024.

[385] N. Zierau, C. Engel, M. Söllner, and J. M. Leimeister. Trust in smart personal assistants: A systematic literature review and development of a research agenda. In N. Gronau, M. Heine, H. Krasnova, and K. Poustcchi, editors, *Entwicklungen, Chancen und Herausforderungen der Digitalisierung: Proceedings der 15. Internationalen Tagung Wirtschaftsinformatik, WI 2020, Potsdam, Germany, March 9-11, 2020. Zentrale Tracks*, pages 99–114. GITO Verlag, 2020.

[386] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye. Object detection in 20 years: A survey. *Proceedings of the IEEE*, 111(3):257–276, 2023.

[387] J. Zubizarreta, I. Aguinaga, and A. Amundarain. A framework for augmented reality guidance in industry. *The International Journal of Advanced Manufacturing Technology*, 102:4095–4108, 2019.