YETI (YET to Intervene) Proactive Interventions by Multimodal AI Agents in Augmented Reality Tasks

Saptarashmi Bandyopadhyay
University of Maryland, College Park
College Park, MD, USA
saptab1@umd.edu

Vikas Bahirwani
Google
Mountain View, CA, USA
vrb@google.com

Lavisha Aggarwal
Google
Seattle, WA, USA
lavishaggarwal@google.com

Bhanu Guda
Google
Mountain View, CA, USA
bhanuguda@google.com

Lin Li
Google
Mountain View, CA, USA
linspeaking@google.com

Andrea Colaco
Google
Mountain View, CA, USA
andreacolaco@google.com

Abstract

Multimodal AI Agents are AI models that have the capability of interactively and cooperatively assisting human users to solve day-to-day tasks. Augmented Reality (AR) head worn devices can uniquely improve the user experience of solving procedural day-to-day tasks by providing egocentric multimodal (audio and video) observational capabilities to AI Agents. Such AR capabilities can help the AI Agents see and listen to actions that users take which can relate to multimodal capabilities of human users. Existing AI Agents, either Large Language Models (LLMs) or Multimodal Vision-Language Models (VLMs) are reactive in nature, which means that models cannot take an action without reading or listening to the human user's prompts. Proactivity of AI Agents on the other hand can help the human user detect and correct any mistakes in agent observed tasks, encourage users when they do tasks correctly or simply engage in conversation with the user - akin to a human teaching or assisting a user. Our proposed YET to Intervene (YETI) multimodal agent focuses on the research question of identifying circumstances that may require the agent to intervene proactively. This allows the agent to understand when it can intervene in a conversation with human users that can help the user correct mistakes on tasks, like cooking, using Augmented Reality. Our YETI Agent learns scene understanding signals based on interpretable notions of Structural Similarity (SSIM) on consecutive video frames. We also define the alignment signal which the AI Agent can learn to identify if the video frames corresponding to the user's actions on the task are consistent with expected actions. These signals are used by our AI Agent to determine when it should proactively intervene. We compare our results on the instances of proactive intervention in the HoloAssist multimodal benchmark for an expert agent guiding a user to complete procedural tasks.

1. Introduction

Recent advances in artificial intelligence have led to the widespread adoption of AI assistants across various platforms and modalities. While these systems, such as Siri for voice interaction and Gemini [18] for text-based communication, have demonstrated significant utility in task automation, they remain constrained by their single-modality architectures. This limitation presents a critical gap in human-AI interaction, particularly in scenarios requiring real-time, context-aware assistance.

Multimodal Vision-Language Models (VLMs) have emerged as a promising solution to bridge this modality gap, offering multimodal understanding that more closely aligns with human perception. However, current VLM-based assistive systems predominantly operate in a reactive paradigm, responding only to explicit user queries. This re-

Preprint.

active nature significantly limits their effectiveness in two critical scenarios: (1) novice learning environments, where users lack the domain knowledge to formulate appropriate queries, and (2) safety-critical operations, where immediate intervention may be necessary before user recognition of potential hazards.

To address these limitations, we propose **YET** to **I**ntervene (YETI), a novel framework (seen in Figure 1) for proactive AI intervention in augmented reality (AR) environments. Our approach leverages lightweight, real-time algorithmic signals to enable proactive assistance through AR interfaces such as smart glasses [19]. This system bridges the gap between cloud-based AI capabilities and real-world applications by enabling direct visual observation of user activities.

Our work builds upon recent developments in proactive AI assistance, particularly the HoloAssist dataset [21], which demonstrates the potential for real-time AI intervention in complex spatio-temporal tasks. While HoloAssist provides valuable insights into human-AI collaboration scenarios, such as computer assembly and coffee preparation, existing implementations face significant computational challenges.

Current state-of-the-art approaches for proactive intervention require extensive computational resources and multi-modal sensor data, including RGB streams, hand and head pose estimation, sensor readings like IMU (Inertial Measurement Unit), and depth information. The complexity of acquiring and processing this data in real-time presents a significant barrier to practical deployment. In contrast, YETI employs efficient algorithmic signals that can be computed on-the-fly, dramatically reducing the computational overhead while maintaining high intervention accuracy.

Features	Size (MB)	× SSIM	× CObj
Depth Estimation	137,408	6,543	6,870
Eye Gaze (E)	617	29	31
Hand Pose (H)	53,749	2,660	2,688
Head Pose	1,141	54	57
IMU (I)	1,132	54	57
SSIM (Ours)	**21**		
Alignment Cobj (Ours)	**20**		

Table 1. HoloAssist Feature sizes scaled with our Features

The YETI framework has comparable precision performance with different HoloAssist benchmark baseline models and shows better performance in some settings, specially higher recall and F-measure, all while using light-weight features that take 6500 times less memory, as seen in Table 1. This dramatic reduction in computational requirements enables real-time operation on resource-constrained AR devices, making proactive AI assistance practical for everyday use cases. Our framework thus represents a significant step toward deploying intelligent assistive systems in real-world applications, particularly in scenarios requiring immediate, context-aware intervention.

2. Related Works

2.1. Egocentric Interaction Datasets

Recent advances in egocentric vision have produced several datasets that capture human interactions and activities. HoloAssist [21] presents a large-scale egocentric dataset focusing on collaborative physical manipulation tasks between two people, providing detailed action and conversational annotations. This dataset offers valuable insights into how human assistants proactively and reactively intervene, correct mistakes, and ground their instructions in the environment.

Parse-Ego4D [1] introduces a benchmark for evaluating AI agents' capability to make unsolicited action suggestions based on user intent signals. We argue that as this benchmark evaluates the AI agents' response to user queries, it does not truly measure proactive behavior.

While Ego-Exo4D [11] provides a comprehensive multimodal, multiview dataset capturing both egocentric and exocentric perspectives in expert-learner scenarios, it primarily focuses on skilled single-person activities without addressing proactive communication. Similarly, existing datasets like Ego4D [10] and EPIC-Kitchens [7], while rich in activity and object annotations, lack direct mappings to actionable recommendations.

2.2. Proactive AI Agents and Communication

Proactive communication in AI agents encompasses several key aspects [8]: Intelligence (the ability to anticipate task developments), Adaptivity (dynamic adjustment of timing and interventions), and Civility (respect for user boundaries and ethical standards). Emerging research has demonstrated the value of proactive AI agents across various domains, including personal assistance, predictive maintenance, healthcare monitoring, and voice assistance [4, 5, 12].

2.3. Language Models for Proactive Assistance

Recent developments have shown promising results in using Multimodal Vision Language Models (VLMs) and LLMs for proactive assistance. ProAgent [23] introduces a framework that leverages LLMs to create agents capable of dynamically adapting their behavior and inferring teammate intentions. The effectiveness of these models has been further demonstrated through fine-tuning on ProactiveBench [13], which significantly enhances the proactive

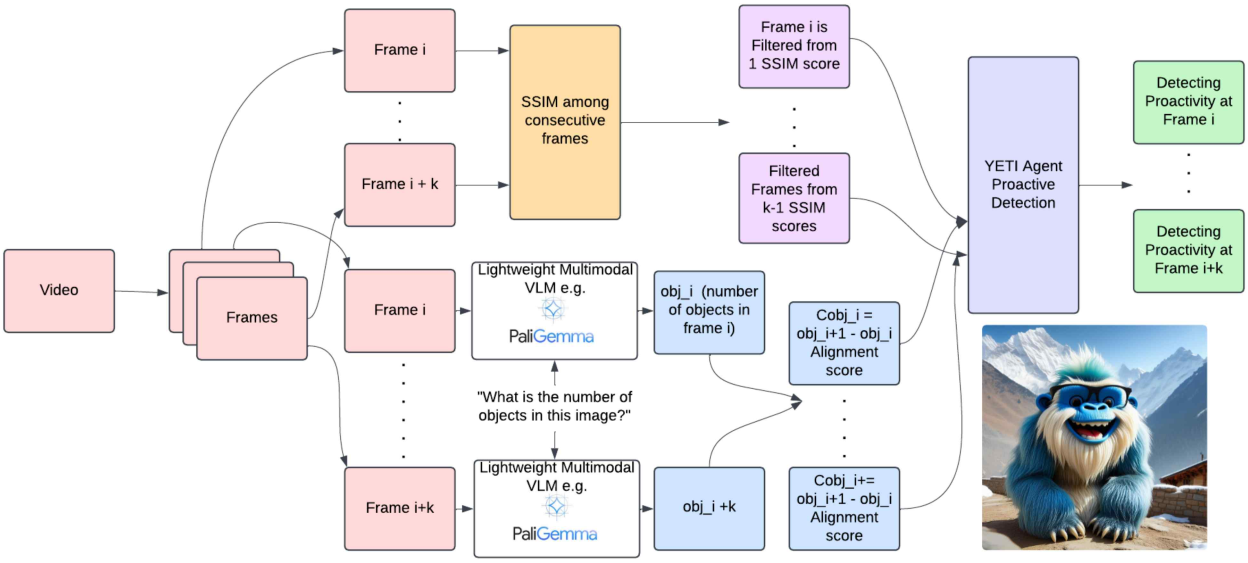
Video
Frames
Frame i
Frame i + k
SSIM among consecutive frames
Frame i is Filtered from 1 SSIM score
Filtered Frames from k-1 SSIM scores
YETI Agent Proactive Detection
Detecting Proactivity at Frame i
Detecting Proactivity at Frame i+k
Frame i
Lightweight Multimodal VLM e.g.
PaliGemma
obj_i (number of objects in frame i)
Cobj_i = obj_i+1 - obj_i Alignment score
"What is the number of objects in this image?"
Frame i+k
Lightweight Multimodal VLM e.g.
PaliGemma
obj_i +k
Cobj_i+= obj_i+1 - obj_i Alignment score

Figure 1. Overview of the YETI framework detecting the frames of proactive interaction or intervention by a Multimodal AI Agent. Our YETI Agent system generates lightweight features on-the-fly, enabling rapid decision-making for timely user assistance.

capabilities of LLM agents. In the context of assistive technology, Smart Help [6] demonstrates how proactive and adaptive support can be provided to users with diverse disabilities and dynamic goals across various tasks and environments.

Open-source VLMs are very popular as a starting point for AI assistants, especially Google's PaliGemma [3] Open-source VLM. Open-source VLMs do not have proactive interaction capabilities which is what we want to support in our research. PaliGemma generates a quick and accurate estimate of the number of objects in a given scene. PaliGemma was trained on a wide variety of datasets, including the TallyQA dataset [2], which is useful for taking a response to a question that asks for the number of objects in a given image.

3. Methodology

3.1. Proactive Augmented Reality Interaction Data of Cooperative Agents

The HoloAssist dataset [21] provides multimodal egocentric vision-language benchmarks focusing on Augmented Reality (AR)-based human-AI collaboration. AR devices are used to capture the Expert-User collaborative dynamics, recording the visual observations of an User Agent (human) collaborating with an Expert Agent (Instructor), which can be an AI Agent, on physical reasoning tasks, while documenting the dialogue between the two agents. The dataset comprises 482 unique Expert-User interaction sequences with videos and dialogues of the agents', spanning 20 diverse task domains, including but not limited to:

- Cooking procedures like making coffee
- Fixing items like motorcycles
- Assembling/Disassembling furniture
- Assembling Devices like Computers, Scanners, GPUs
- Maintaining Electrical systems like circuit breakers
- Configuring Devices like printers, cameras, switches

The User Agents wear the AR devices to record first-person perspective videos while executing procedural tasks. The AR devices simultaneously capture the Expert Agent's observations and guidance to the User Agents. The dataset's annotation schema encompasses a variety of interaction types. Some examples of interactions done by the Expert Agent include:

1. Proactive Interactions
   - High-level instructional guidance
   - Follow-up instructions without any user query
   - Interventional feedback
   - Error correction mechanisms
2. Reactive interactions:
   - Expert clarifications to user queries
   - User-initiated dialogues

The corpus encompasses 45.5 hours of video recordings generated by 350 distinct expert-user pairs, providing a rich foundation to study the spatio-temporal dynamics of when and how Proactive Multimodal AI Agents should engage in AR-assisted collaborative scenarios. This temporal aspect, when an AI Agent is yet to intervene, but should intervene, to improve collaborative task execution or correct mistakes