



Video Language Models for Generic Task Assistance

Arin Shinde, YSP Student, *Shrewsbury High School*

Nyasa Luharuka, YSP Student, *Lexington High School*

Kurosh Kavoori, Student, *Nashua High School South*

Bishoy Galoaa, College of Electrical and Computer Engineering, *Northeastern University*

Sarah Ostadabbas, Department, *Northeastern University*



Northeastern University
College of Engineering



Northeastern University
**Michael B. Silevitch and
Claire J. Duggan Center
for STEM Education**

Abstract

Artificial intelligence struggles to reason through long, multi-step human activities. Important information may be distributed across first-person video, third-person video, body-motion signals, and spoken instructions. We developed a synchronized multimodal data-collection and visualization method for studying task completion with mistakes and corrections. Participants completed tool organization, grocery-shopping, and food preparation tasks, and IKEA chair assembly and disassembly while being recorded with Aria Gen2 smart glasses, exocentric cameras, and Mudra Link wristbands. These inputs were synchronized, divided into temporal action segments, connected to transcripts, and organized in a Parquet format. Interactive interfaces were created to display the video, biosignal, transcript, and task-segment data on a shared timeline. This work creates the infrastructure needed to train future video-language models to provide step-by-step task assistance in a variety of fields from robotics to medicine.

Background

Retained surgical items are objects unintentionally left inside a patient following an operation. They are classified as serious, preventable safety events and may result in pain, infection, or death. The Agency for Healthcare Research and Quality reports an estimated incidence of approximately 1.3 retained surgical items per 10,000 procedures.

Current prevention strategies rely heavily on manual counts, communication between surgical team members, and technologies such as radiographic detection. However, it was reported that out of 190,000 operations, most retained-object events occurred. This demonstrates the need for multiple complementary safety systems.

An AI assistant capable of observing the procedure could potentially function as an additional safety layer. To do this effectively, the system must understand more than a single video frame. It must track objects, recognize the current procedural step, and identify whether an item has been removed or not.

With the use of multiple different perspectives from the Aria Gen2 glasses, exocentric cameras, and Mudra Link bands, this can provide additional aid to a model to visually understand. Our objective was therefore to build a reusable pipeline that synchronizes multiple perspectives and sensor types during tasks.

Experimental Methods

Stage 1: Multimodal Data Collection

- Human activities recorded using many different types of devices
 - Aria Gen2 glasses: egocentric, first-person video of participant's hands and the working area
 - Logitech Cameras: exocentric, wider scene and multiple angles to reduce occlusion
 - Mudra Link wristbands: biosignal data on intentionality and wrist movement

Stage 2: Expert-Observer Recording Framework

- Expert explained how to complete a task and observer performed actions
- Paired examples of spoken guidance, visible actions, and mistakes along with their corrections
 - Putting tools back in drawers
 - Making a peanut butter and jelly sandwich
 - Making a hotdog
 - Assembly and disassembly of an ikea chair

Stage 3: Synchronization and Temporal Segmentation

- Streams were not fully aligned
 - Used a clap performed at beginning of recording to create a synchronization event
- Temporal segmentation needed to split up a clip into separate steps
- Used Python and FFmpeg scripts to align streams and divide recordings into meaningful procedural segments

Stage 4: Data Processing and Organization

- Data structured into Parquet format
 - Different modalities could be accessed using a shared timeline
- Created an interactive visualization interface to display multiple camera views, transcript, and Mudra signals

Stage 5: Final Suturing Demonstration

- Used a synthetic suturing kit and simulated wound
 - Three gauze pads were included in procedure
- AI assistant provided procedural guidance while monitoring activity captured by cameras
 - When closure began with one pad remaining, assistant gave a warning

