Enhancing Ambient Assisted Living: Multi-Modal Vision and Language Models for Real-Time Emergency Response

Adil Zhiyenbayev[1*], Rakhat Abdrakhmanov[1*], Adnan Yazici[1] and Huseyin Atakan Varol[2]

[1*]Department of Computer Science, School of Engineering and Digital Science, Nazarbayev University, Kabanbay Batyr 53, Astana, 010000, Astana, Kazakhstan.

[2]Institute of Smart Systems and Artificial Intelligence, Nazarbayev University, Kabanbay Batyr 53, Astana, 010000, Astana, Kazakhstan.

*Corresponding author(s). E-mail(s): adil.zhiyenbayev@nu.edu.kz; rakhat.abdrakhmanov@nu.edu.kz;
Contributing authors: adnan.yazici@nu.edu.kz; ahvarol@nu.edu.kz;

Abstract

Recent advancements in ambient assisted living (AAL) technologies have effectively harnessed sophisticated machine learning and deep learning techniques, especially in human activity recognition (HAR), to enhance emergency responses and preventive care. This research introduces a cutting-edge multi-modal system that incorporates an advanced vision-language model (VLM) to boost detection capabilities within AAL settings. Utilizing complex deep learning algorithms, the system adeptly interprets scene contexts to produce precise captions, enable visual question answering, and facilitate commonsense reasoning. This functionality is complemented by an interactive chatbot interface that employs state-of-the-art (SOTA) large language models (LLMs) equipped with text-to-speech and speech-to-text capabilities, allowing for real-time, interactive assessments that greatly improve the detection of abnormal behavior. Moreover, the system employs prompt engineering to refine anomaly detection at various stages efficiently, eliminating the need for additional extensive training or computational resources. It autonomously executes critical decisions, such as dispatching ambulances and generating timely descriptive alerts for caregivers, thus ensuring prompt action in urgent situations. An exhaustive evaluation of the system was performed to verify its effectiveness and usability. The qualitative analysis confirmed its high usability among caregivers, while quantitative assessments revealed a detection accuracy of 93.44%, a recall rate of 95%, and a specificity rate of 88.88%. Importantly, user interactions with the model provided extra contextual information, enhancing the detection accuracy to 100%. Overall, this multi-modal system marks substantial progress in AAL technology, not only enhancing emergency recognition and response capabilities but also providing caregivers with actionable insights and recommendations in real-time. The deployment of deep learning within AAL environments promises substantial reductions in emergency incidents and notable improvements in the quality of life for the elderly and disabled.

Keywords: Ambient assisted living, Human activity recognition, Vision-language models, Large language models, Speech models, Prompt engineering

1 Introduction

A forecast predicts a global population exceeding 9.7 billion individuals by 2050, highlighting growing demand for healthcare services, particularly for aging and disabled populations [1]. This trend underscores the importance of robust long-term care services and continuous monitoring systems. However, challenges such as caregiver overload, financial constraints, and logistical obstacles hinder the timely provision of assistance in emergency situations [2].

To address these challenges, researchers have proposed various autonomous Ambient Assisted Living (AAL) setups aimed at long-term monitoring and timely detection of emergency cases [3–6]. AAL systems typically involve patient monitoring using diverse sensors and Human Activity Recognition (HAR) techniques powered by machine learning algorithms to classify abnormal and normal actions.

Current methodologies rely on wearable or non-intrusive sensors to identify abnormal behavior in individuals by analyzing patterns in their daily activities [4–9]. These techniques aim to distinguish between normal and abnormal behaviors, such as falls or other emergency actions requiring medical attention [7]. However, traditional sensor-based approaches encounter several challenges. Variability in individuals' daily activity patterns can pose difficulties for machine learning models in consistently identifying anomalies. Moreover, developing robust models necessitates extensive collection of labeled training data, often hindered by ethical, privacy, and practical concerns in home environments. Additionally, sensor data may lack contextual information crucial for accurately discerning between normal and abnormal activities. Environmental factors such as poor lighting or obstructions can further degrade data quality, leading to false alarms and diminished model accuracy. Furthermore, these models require ongoing recalibration and adaptation to individual behavior and environmental changes, increasing complexity and resource requirements.

To address sensor-based system limitations, we explore Vision Language Models (VLMs). VLMs excel in various visual tasks, including image classification, object detection, and semantic segmentation, with applications in medicine [10]. For instance, VLMs trained on medical images and reports achieve impressive zero-shot (make predictions on new samples without any prior training) classification results [10]. Additionally, initiatives like MedFuseNet and PathVQA leverage VLMs to provide insightful responses to medical visual inquiries [11, 12]. These systems utilize advanced image processing and linguistic comprehension to deepen understanding of medical data.

VLMs hold promise for medical diagnostics and anomaly detection by comprehending visual scenes and contextual nuances [10]. However, their integration into real-time emergency applications faces challenges, including language ambiguity and data quality issues, potentially leading to false alarms [11, 12]. Environmental changes and limited contextual understanding may further impede accurate behavior detection.

To decrease the hallucination of VLM, Large Language Model (LLM) capabilities that include human-like text generation and chatbot functionalities can be used to interact with the person directly and collect the concerns and symptoms from a person as a valuable context. Also, this interaction can be facilitated through Text-To-Speech (TTS) and Speech-To-Text (STT) models, enabling seamless Human-Computer Interaction (HCI) [13, 14]. Furthermore, pre-trained LLMs are now adaptable through prompt engineering that can tailor them to specific tasks without additional training and computational costs [15]. The prompts with instructions allow LLMs to efficiently adapt and perform highly specialized tasks, like emergency detection, without extensive data collection and additional training.

The proposed system incorporates a Large Language and Visual Assistant (LLaVA) [16], which utilizes VLM integrated with LLM. The LLaVA model shows strong zero-shot capabilities in comprehending, describing, and analyzing various visual aspects. Also, the LLaVA model can be adapted for specific tasks via prompt engineering. Specific instructions and templates can guide the model to perform the visual tasks for Abnormal Behaviour Detection (ABD), and in-context learning (dynamically generated context and prompts that guide and adapt the model to the tasks that were not used during the training) [17]. Also, integrating TTS, STT, and LLM's chatbot capabilities from the LLaVA model can provide additional context to increase detection accuracy and decrease false alarms. Overall, a framework consisting of several models like VLM, LLM, STT, and TTS can give promising results on ABD and assistance for people.

This study introduces a multimodal system integrating visual, textual, and audio data to address healthcare needs, enhancing assessment accuracy and reliability by identifying abnormal behaviors and emergencies while engaging users. Key contributions include:

- The system seamlessly integrates VLM, LLM, TTS, and STT, enabling two-stage abnormal behavior detection and real-time response to complex scenarios. Multimodal data fusion improves decision-making by combining visual, textual, and auditory inputs.
- Extensive real-time experiments conduct qualitative and quantitative analyses across various scenarios, generating a new dataset with abnormal and normal scenarios.

- Utilizing the LLaVA model, the system employs advanced image and language processing techniques to accurately distinguish between normal and emergent behaviors without additional emergency dataset training. Prompt engineering enables context-based model adaptation.
- Integration of TTS and STT enables direct user communication, reducing false positives and negatives in ABD by gathering more contextual information.
- The system achieves a remarkable 93.44% accuracy using only images, reaching 100% with interaction history. Additionally, it consistently detects emergencies within an average of 154 seconds, demonstrating efficient response time.

The paper is organized into the following sections. Section 2 reviews related works on vision-based sensors, VLMs, and LLMs for HAR and ABD. Section 3 provides an overview of the proposed method, detailing modalities, system workflow, and hardware and software modules. Section 4 presents the experiment procedure, qualitative and quantitative results, and evaluation. Finally, Section 5 concludes the paper by summarizing the main findings, and the proposed solution, and discussing future implications and limitations.

2 Related Works

This section covers existing methods for general HAR and ABD, as well as VLMs and LLMs applied to HAR and the medical field. Additionally, it includes a comparison with the state-of-the-art (SOTA) approaches.

2.1 Vision-Based Sensors For HAR and ABD

Vision-based methods for HAR rely on cameras to capture activities, utilizing these visual inputs to extract 3D key joint data for deep learning (DL) models. Kaur et al. use multi-view cameras to develop contactless monitoring tools for neurological gait classification [18]. However, these efforts often face challenges such as small sample sizes and sensitivity to environmental conditions like lighting. [8] employs depth cameras to isolate human figures for feature extraction, preprocessing this data with techniques like linear discriminant analysis (LDA) for training on models like hidden Markov model (HMM). However, HMMs often struggle with complex data patterns and lack context due to ignoring background information. Further advancements include Batool and Javeed that uses RGB-D cameras with linear discriminant classifiers and recurrent neural networks (RNN) to achieve high accuracy in activity recognition [19]. [20] enhances model performance in activity classification with transformer architectures and synthetic data from generative adversarial networks (GANs).

Apart from HAR, other work with vision sensors focuses on detecting violence and dangerous behaviour [21]. This work proposes a two-stage complex action recognition framework for automatic violence detection at real-time surveillance by decomposing them into simpler components. For visual analysis, they employed a two-stream architecture utilizing the YOLO model for detecting spatial features and optical flow techniques for temporal analysis. This approach allowed for the real-time identification and classification of complex, violent activities within surveillance footage.

These studies demonstrate the potential of vision-based sensors in applications ranging from healthcare monitoring to surveillance and HCI. However, they still require considerable training on HAR and emergency datasets. Also, environmental interference can affect the quality of the images, reducing the robustness of vision-based approaches in HAR and ABD.

2.2 VLMs and LLMs for HAR and Medical Field

Recent advances in DL have introduced VLMs, which are pre-trained on extensive image-text pairs to learn vision-language correspondence, facilitating zero-shot and few-shot (recognize new classes with minimum training data) predictions on unseen data [22]. VLMs, utilizing deep convolutional networks (DCNN) and transformer architectures, have revolutionized traditional classification tasks by generating descriptive texts/reports/captions from images and answering questions about the image, enhancing dynamic image-text interactions. VLMs can recognize and adapt to a broad array of classes due to pre-training on large datasets and a self-supervised nature. CLIP VLM exemplifies this by learning a multimodal embedding space that effectively aligns visual and textual information, demonstrating high accuracy in zero-shot predictions on ImageNet [23]. Similarly, Prismer VLM leverages ensemble learning for efficient image captioning and Visual Question Answering (VQA) tasks, significantly reducing computational complexity while maintaining high performance [24]. BIKE VLM extends this approach to video understanding, integrating mechanisms that improve recognition by considering the temporal aspects of videos, although it demands considerable computational resources [25].

Apart from general HAR, VLMs can be used in medicine, with some works like MedFuseNet [11], MedViLL [26], RepsNet [27] and UniXGen [28] are used for addressing visual questions related to medical field or generating the medical reports. Also, LLMs are being adopted in medicine as chatbots for commonsense reasoning, assisting, and interaction tasks via the rapid generation of responses to questions based on the queried text prompts and images. For instance, Bosselut et al. propose the MEDITRON LLM that is pre-trained on clinical practice guidelines (CPGs) for guiding caregivers and patients in making evidence-based decisions about diagnosis, treatment, and management [29]. Clinical-BERT is pre-trained on medical reports from radiographs to generating reports [30]. [31] generates the reports in natural language from numeric and temporal personal health information. Other works transformed general LLMs for the healthcare domain via instructional or supervised fine-tuning [32–34]. However, fine-tuning and training require large amount or diverse high-quality datasets to perform and adapt well on medical tasks. On the contrary, other works use prompting to align the LLMs for the medical field, requiring no additional training and computational costs [32, 35–37].

Overall, these works address the challenge of dealing with a limited number of classes, generalization, scalability, adaptation, advanced multimodal analysis, and the provision of detailed contextual information. Since VLMs and LLMs are trained on large dataset corpus that include different objects, attributes, human-object interaction, and human-human interaction, there is potential for further research to explore their use as primary tools for health diagnostics, VQA, user interaction, anomaly detection, report generation and patient and caregiver guidance within the healthcare domain.

2.3 Comparison with the SOTA

In the discussion of various healthcare frameworks for ABD, HAR, and Activities of Daily Living (ADL), our proposed work stands out due to several distinctive properties outlined in Table 1. The comparison is based on ten attributes, including multi-modal capabilities, causal reasoning, adaptation to new classes without additional training, self-supervised abilities, user interactivity, personal and environmental contextual data incorporation, real-time implementation and detection, as well as simple and complex case detection.

Table 1: Comparison of attributes in ABD, HAR and ADL among SOTA approaches and proposed system. Each cell contains a symbol indicating whether a particular feature is present ("✓") or absent ("✗"). Partial value in the cell means the feature is not explicitly used or changed.

Study	Multi-Modal	Causal Reasoning	Adapt new classes	Self-supervised	Interactive	Person Context	Env. Context	Real-time Detection	Simple Case Detection	Complex Case Detection
[38]	✓	✗	✗	✗	✗	✓	✗	✗	✓	✗
[39]	✓	✓	partial	✗	partial	✓	✓	✓	✓	✓
[40]	✗	✗	✗	✗	✗	✗	✓	✗	✓	✗
[41]	✓	✗	✗	✗	✗	✓	✗	✓	✓	✗
[42]	✓	✗	✗	✗	✗	✗	✓	✓	✓	✓
[8]	✗	✗	✗	✗	✗	✓	✗	✗	✓	✓
[43]	✗	✗	✗	✗	✗	✓	✗	✓	✓	✓
[44]	✗	✓	✗	✗	✗	✗	✓	✓	✓	✓
[45]	✗	✗	✗	✗	✗	✗	✓	✗	✓	✓
[19]	✗	✗	✗	✗	✗	✓	✗	✓	✓	✓
[18]	✗	✗	✗	✗	✗	✓	✗	✓	✓	✓
[20]	✗	partial	✗	partial	✗	✓	✗	✗	✓	✓
[46]	✗	✗	✗	✗	✗	✓	✗	✓	✓	✓
[21]	✗	✗	✗	✗	✗	✓	✓	✓	✓	✓
[47]	✓	✓	partial	✗	✗	✓	✓	✓	✓	✓
Proposed	✓	✓	partial	✓	✓	✓	✓	✓	✓	✓

According to Table 1, the proposed system, along with a few others, adopts a multimodal approach, integrating various data modalities [38, 41, 42], unlike other studies which rely solely on visual features or 3-axial accelerometer and gyroscope sensors or activations from environmental sensors for action classification [8, 18–21, 43–46].

The following attribute is causal reasoning, which shows the system's ability to build causal links between objects and events and explain its reasoning and decision in ABD. Causal reasoning is marked in the proposed system since the LLaVA model was trained on the instructional dataset (answering the questions that required reasoning) and can infer causality from observed data. Also, this feature is present in the works capable of capturing temporal dependencies and sequential patterns [39, 44, 47].

While many works lack the ability to adapt to new classes without extra training, the proposed system stands out. The LLaVA model's self-supervised training enables it to learn inherent features from vast amounts of unlabeled data, enhancing generalizability across domains and tasks. When encountering new classes, the model compares image-text pairs with its knowledge base, leveraging contextual clues to identify similar outputs and adapt accordingly. However, the model's adaptability to entirely new classes may be limited. While transfer learning can aid in this adaptation by transferring knowledge from related tasks, its effectiveness depends on the similarity between the new and original classes. If the new classes are too dissimilar, transfer learning alone may not suffice. Therefore, the model requires supervision to ensure comprehension of new classes, and additional fine-tuning could further improve its ability to understand new concepts.

Interactive features, such as STT and TTS models for communication, are unique to the proposed system. At the same time. some works [39, 47] include partial interaction through alerting systems and messaging with patients and caregivers.

Contextual understanding, where systems process information based on personal and environmental factors, is evident in works using wearable sensors and some with visual sensors and DL models. Most works utilize skeletal features, motion data, and personal medical parameters, with some incorporating environmental sensors. The proposed system also considers contextual and environmental factors, addressing the individual's state and environmental anomalies.

Real-time detection, seen in studies [18, 19, 21, 39, 41–44, 46, 47] and the proposed system, allows for quick data processing and response. However, many studies do not explicitly compare their system's time performance for abnormality detection to our proposed system.

Simple case detection, identifying basic activities like sitting or walking, is common in most studies, indicating effective identification of straightforward patterns. Complex case detection, recognizing intricate patterns, is also achievable in most works, although it often requires extensive training and data preprocessing. In contrast, the proposed solution utilizes zero-shot performance, in-context learning, and prompt engineering for ABD without extensive training.

Overall, this comparison highlights the proposed system's comprehensive support for vital attributes, suggesting a robust and adaptable approach to ABD and HAR.

3 Methods

This section delves into our system, which integrates VLM, LLM, human detection, TTS, and STT models for ABD and HAR. We delve into the system's distinctive features, such as its generalization and zero-shot capabilities, improved detection accuracy, emergency refinement, and reasoning facilitated by prompt engineering and in-context learning.

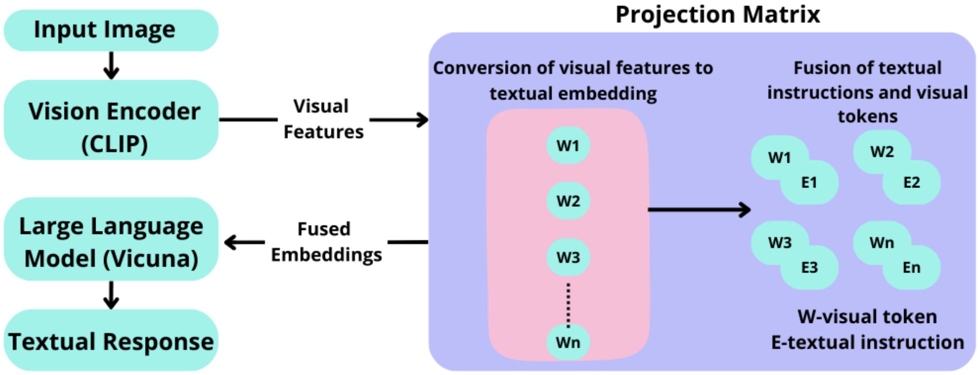


Fig. 1: The figure illustrates the LLaVA architecture, featuring LLM Vicuna, vision transformer CLIP, and the conversion of visual features to textual tokens, fused with textual instructions. This fused information serves as input for Vicuna LLM, enabling response generation based on both visual context and user instructions.

3.1 Large Language-And-Vision Assistant Model

The LLaVA-1.5 model serves as the backbone for HAR and emergency detection, integrating VLM and LLM capabilities. It combines the visual encoder CLIP-ViT-L-336px [23] and LLM Vicuna-1.5 [48], as detailed in [16].

As illustrated in Figure 1, the LLaVA model undergoes a two-stage training process, aligning visual features with textual embeddings in a self-supervised manner. Initially, a projection matrix connects CLIP-extracted visual features with textual embeddings from the pre-trained LLM Vicuna. Subsequently, this projection layer maps image features to language tokens for the LLM input. The second stage involves fine-tuning both the LLM and the projection matrix on a diverse instructional dataset, enhancing the model's ability to interpret and respond to visual and textual queries.

Furthermore, the model undergoes fine-tuning for multimodal science questions (ScienceQA), where it learns from detailed explanations and lectures [49]. This exposure to multiple-choice questions improves its understanding, reasoning, and ability to answer science-related queries. The model's versatility extends to becoming a powerful chatbot, capable of discussing images, following instructions, and responding to multimodal science queries with detailed explanations [16, 50, 51]. This exposure to various instruction types and science-related questions enhances its reasoning and response capabilities related to images.

The LLaVA model, available in various versions, is characterized by parameter count and task-specific fine-tuning [16]. Notably, the general LLaVA-1.5, with 13 billion parameters, serves as the primary model with VLM and LLM capabilities in our system. Its extensive training on large datasets of general images and text forms a robust foundation for comprehending visual concepts, object relations, and human actions [16]. This knowledge is crucial for scene understanding and identifying unusual postures, situations, environments, and emotions within an AAL system. Unlike models trained on basic activities or narrowly fine-tuned datasets, LLaVA-1.5 can distinguish complex emergencies from routine activities using advanced contextual understanding. For example, it can discern non-emergency scenarios by analyzing visual cues combined with contextual data, such as facial expressions and environment, enhancing reliability in real-world applications.

Furthermore, LLaVA with 13 billion parameters outperformed SOTA VLMs like BLIP-2, InstructBLIP, and Qwen-VL-Chat on 11 out of 12 academic VQA benchmarks, showcasing superior visual instruction tuning. These benchmarks include VQA-v2 [52], GQA [53], and others [51]. This translates to improved performance in tasks requiring understanding, reasoning, conversation, detailed description, and following textual instructions related to visual data. Additionally, the smaller 7-billion-parameter version of LLaVA exhibited lower efficiency compared to the larger version on the mentioned datasets [16].

LLaVA was compared with ChatGPT-4 across conversation, detailed description, and complex reasoning tasks, achieving 85.1% of ChatGPT-4's relative performance score [51]. Moreover, LLaVA surpassed ChatGPT-4 with 90.92% accuracy on the Science QA dataset, establishing itself as a formidable alternative. Consequently, LLaVA is chosen as the primary model for its proficiency in understanding, reasoning, and conversational abilities.

These findings position LLaVA as a robust alternative to ChatGPT-4, particularly adept at tasks demanding deep contextual analysis and reasoning. Its superior performance across multiple VQA benchmarks indicates an advanced ability to handle diverse, context-rich tasks, essential for our work involving deep scene and context analysis, and reasoning to detect abnormalities effectively.

3.2 Instruction Following Ability for Emergency Detection

VLMs, with their intricate structures containing billions of parameters [54], may yield unpredictable outcomes as they predict subsequent words in sequences based on statistical word connections. However, prompt engineering mitigates randomness and inconsistencies in their responses. LLaVA emerges as the primary model, boasting self-supervised and contrastive learning, superior generalization, and enhanced instruction-following capabilities [16]. [55] underscores LLaVA's reasoning prowess in emotion recognition, accurately inferring emotions by understanding scene context and complex emotional states.

Through transfer learning, LLaVA demonstrates promising outcomes in medical imaging analysis, such as MRI and chest X-rays, without additional training [54]. This adaptability to novel data and contexts, coupled with responsiveness to prompt engineering, renders LLaVA highly effective, particularly in ABD. Consequently, LLaVA can be tailored or instructed through prompts to discern specific visual and textual inputs indicative of emergencies or abnormal activities. This approach empowers the model to recognize specific patterns and concepts, distinguishing between anomalous and non-anomalous actions.

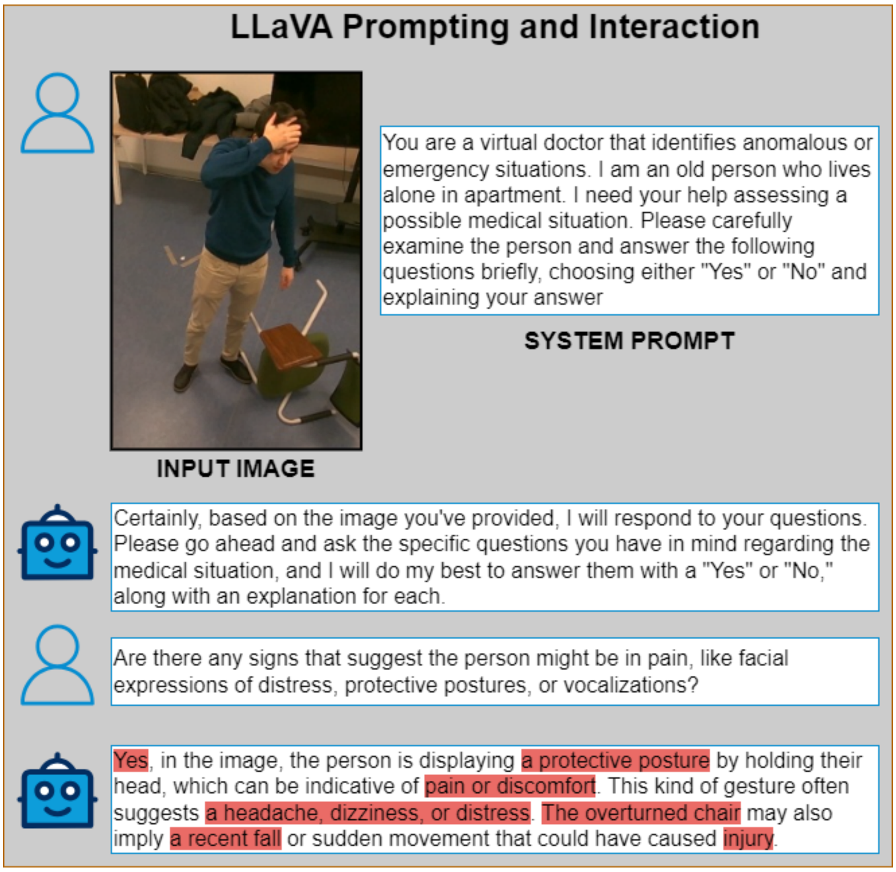


Fig. 2: The figure depicts the prompt and instruction tuning for the LLaVA to determine abnormal behaviour and respond accordingly to user requests. The response is person-specific, referring to the person with guesses regarding the anomalies like protective postures, gestures and environmental objects (marked with red).

Prompts are instrumental in focusing LLaVA's attention and guiding its decision-making process. As depicted in Figure 2, LLaVA was tasked with understanding and recognizing abnormal behavior, adopting the role of a virtual doctor, and providing "Yes" or "No" responses with explanations related to ABD. This strategy reduced hallucinations in responses and directed attention to anomalous behavior and specific response formats. Furthermore, tailored visual instructions, such as identifying signs of abnormality like protective postures or facial expressions of discomfort, augment LLaVA's anomaly detection capabilities [17].

As a result, the model adhered to instructions, focusing on protective posture (holding a head) and environmental abnormalities like a fallen chair. Additionally, LLaVA demonstrated adept reasoning regarding signs of abnormal behavior, deducing a potential fall from a chair resulting in a head injury. The detailed visual instructions and expected response output guide the model in following and executing user instructions effectively. Consequently, LLaVA can be tasked with identifying subtle indicators of emergencies, such as facial expressions or environmental cues of distress. For example, LLaVA exhibited consistent and robust performance in discerning subtle emotional clues from images and videos across multiple datasets like The EMOTIC, CAER-S, and BoLD, annotated with emotional labels [55]. This nuanced prompting enhances LLaVA's advanced reasoning capabilities to detect less obvious signs [54].

Overall, LLaVA excels in handling sequences of interconnected prompts, facilitating structured analysis that progresses from identifying potential hazards to assessing injury or distress comprehensively. Customization for specific environments or individuals enhances its ability to assess varied contexts dynamically, bolstering its emergency detection capabilities. Moreover, LLaVA's dynamic response formatting ensures clear and concise answers that meet user specifications. This is crucial for tracking incidents and taking swift action in emergencies. This feature underscores LLaVA's potential for use in HAR and ABD. Similar to the approach presented in Figure 2, the model can be prompted with visual instructions to perform VQA based on emergency signs in images.

3.3 Text-To-Speech and Speech-To-Text

TTS and STT are fundamental components of natural language processing (NLP) [56]. STT, an automatic speech recognition (ASR) model, converts speech to text, while TTS transforms text into human-like speech. In this study, Whisper STT is chosen for its cutting-edge performance. Trained on an extensive dataset comprising 680,000 hours of labeled audio data from the web, including 563,000 hours in English and 117,000 hours across 96 other languages [57], Whisper employs a transformer architecture. Initially, it processes audio into 30-second chunks and converts them into Mel spectrograms. The decoder then correlates word tokens with audio sequences, selecting the most probable words for output through a series of convolutional layers, activation functions, and residual blocks. Additionally, Whisper's decoder utilizes special tokens to enhance language identification and timestamping tasks.

Whisper goes beyond mere transcription, excelling in speech translation and activity detection, displaying near-human performance even in scenarios altered by stress or emergency situations, without specific training for such contexts. Its effectiveness extends to noisy environments, rendering it robust for real-world applications. Despite its impressive word error rate (WER) of 2.5% on the LibriSpeech dataset [58], Whisper's true strength lies in its ability to adapt to audio variations not seen during training, such as recordings with background noise, making it a versatile and dependable tool for a wide range of audio challenges [57].

In summary, leveraging the conversational capabilities of the LLaVA model, it can serve as a chatbot to engage with users, gather additional context, analyze past interactions, and make informed decisions during emergencies. Integrating TTS and STT into the system facilitates direct interaction between users and the model. Employing a medium-sized Piper TTS [59] ensures real-time user-model interaction, while a small-sized Whisper STT provides accurate understanding, even in noisy environments. This combination balances Piper's speed with Whisper's adaptability and noise robustness, enhancing user engagement and enabling the LLaVA model to grasp conversational context effectively.

3.4 Key Components of a System

The proposed approach integrates continuous monitoring and user-model interaction blocks to detect, interact with, and confirm emergencies. Below, we outline the essential components of the system's hardware, models, and workflow.

3.4.1 Hardware

The experimental setup comprises an NVIDIA DGX V100 server for running the LLaVA model and data storage. A strategically positioned Intel RealSense d455 camera, akin to a security camera setup, offers extensive coverage of the living space for person detection and is connected to a laptop. Participants wear "Rode Wireless GO" microphones for precise audio capture during interactions. Furthermore, a local laptop equipped with an audio speaker enables remote server access, handles image and text file transfers, and executes the Piper TTS, human detection model, and Whisper STT models.

3.4.2 Models

The setup integrates several technologies for monitoring and interaction during emergency scenarios. YOLOv8, a DL model for human detection [60], continuously monitors camera footage to identify individuals in the living space. The LLaVA VLM plays a central role in identifying emergencies, generating context-specific questions based on the situation, and analyzing user responses and image data to confirm the emergency [16]. Piper TTS converts LLaVA's questions into natural voice prompts [59], which are then relayed through speakers. User responses are transcribed accurately by Whisper STT [57], facilitating LLaVA's analysis and the generation of follow-up questions. Additionally, an emergency alert system simulates and dispatches notifications to relevant contacts.

3.5 System's Workflow

The proposed method is shown in Figure 3 and in Appendix A1, providing both high-level and detailed perspectives. The system comprises a continuous monitoring component responsible for human detection, file transfers between the server and local devices, and anomaly detection using VQA with predefined thresholds. Additionally, a user-model interaction block delineates how LLaVA initiates real-time interactions, handles user responses, confirms emergencies, identifies when a person is unresponsive and triggers actions such as calling an ambulance and sending alerts.

3.5.1 Continuous Monitoring Part

This system segment is tasked with constantly surveilling individuals to detect potential emergencies.

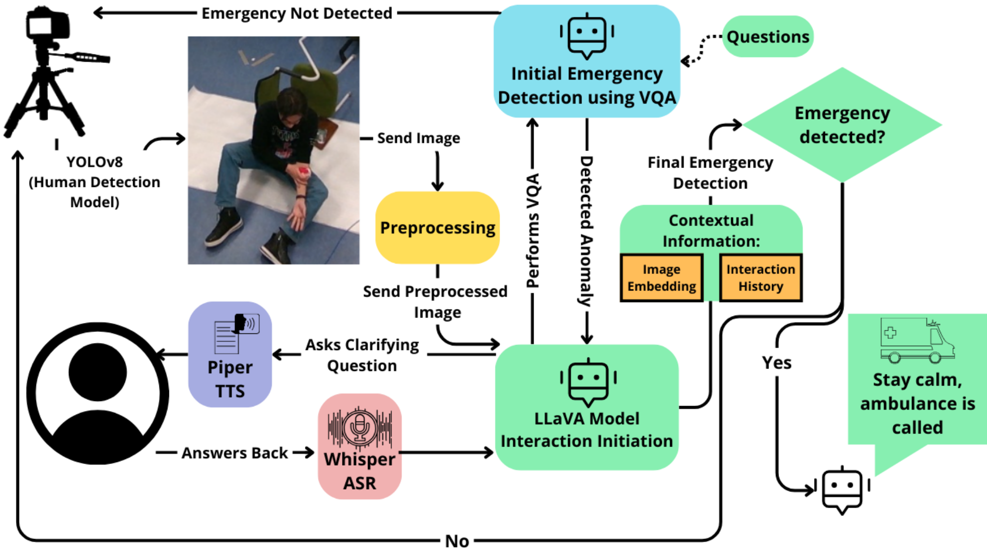


Fig. 3: High-level diagram of our emergency detection system: The camera detects a person in an emergency, capturing an image, which is preprocessed for analysis. The LLaVA model evaluates the image through a VQA session, answering predefined questions based on visual cues, generating an anomaly score. If the score surpasses a threshold, the model engages in user-model interaction mode, gathering detailed responses. Ultimately, based on contextual information, the LLaVA model decides whether to call an ambulance.

Image Capturing: The system utilizes a pre-trained YOLOv8 model to detect individuals within the scene, focusing on human presence amidst various other classes. When a person is detected, the camera captures a frame at set intervals. To ensure image clarity, YOLOv8's confidence threshold for person detection is empirically set to 0.85. The captured frame undergoes preprocessing to isolate the individual and their immediate surroundings, enhancing focus on their actions. Subsequently, the preprocessed image and a "mode.txt" file, signaling VQA initiation (with a value of 0), are transmitted to the server for further processing by the LLaVA model.

Abnormality detection through VQA: The system utilizes a predefined set of questions, meticulously crafted to identify early signs of abnormalities before user interaction. Comprising seven questions, VQA protocol was carefully curated and tested for optimal clarity and effectiveness with the LLaVA model. Each question prompts the model to provide a multiple-choice response, with a "Yes" indicating a potential abnormality. The model's decision-making process considers both the image context and its reasoning capabilities. Results from the VQA session, including questions and answers, are stored and transmitted to the local system. An illustrative example of the VQA process, focusing on detecting signs of a heart attack, is depicted in Figure 4. Additionally, Figure A2 in Appendix A shows the initiation of the VQA task for abnormality detection through a prompt directed at the LLaVA model.

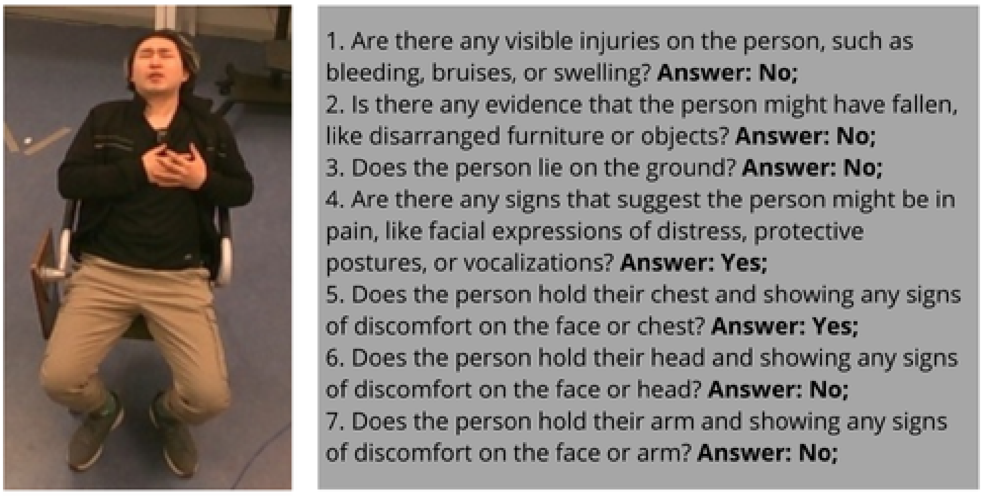


Fig. 4: VQA of the abnormal behaviour case. This figure represents the example of the heart attack scenario and LLaVA's VQA based on the visual context. The model is prompted to answer in a short form by "Yes" or "No" to save time in an emergency. The bolded texts are the LLaVA's answers to the questions.

Thresholding: Collected user responses are tallied to generate an abnormality score, representing the count of "Yes" answers. Should this score surpass a predetermined threshold (the non-anomalous action threshold), the system initiates real-time dialogue with the user. Before activating the user-model interaction block, the system prompts the user to confirm the need for further interaction. If the user affirms assistance is required ("Yes" response), the system updates the "mode.txt" file to 1, signaling the activation of the interaction block. Alternatively, if the user remains unresponsive, the system automatically triggers the interaction block. This confirmation step is implemented for user convenience, mitigating the possibility of false alarms from the VQA process.

3.5.2 User-Model Interaction Block

In this stage, the system engages directly with the individual, posing questions tailored to the visual context and preserving the interaction history. Leveraging this history alongside previous image embeddings, the model formulates contextually relevant follow-up queries. After analyzing this contextual information, the system determines the individual's state and provides a suggestion aligned with the context. Notably, the system adjusts its behavior dynamically, leveraging in-context learning principles, without requiring additional training on new data. This exemplifies the utilization of in-context learning within the interaction component [17].

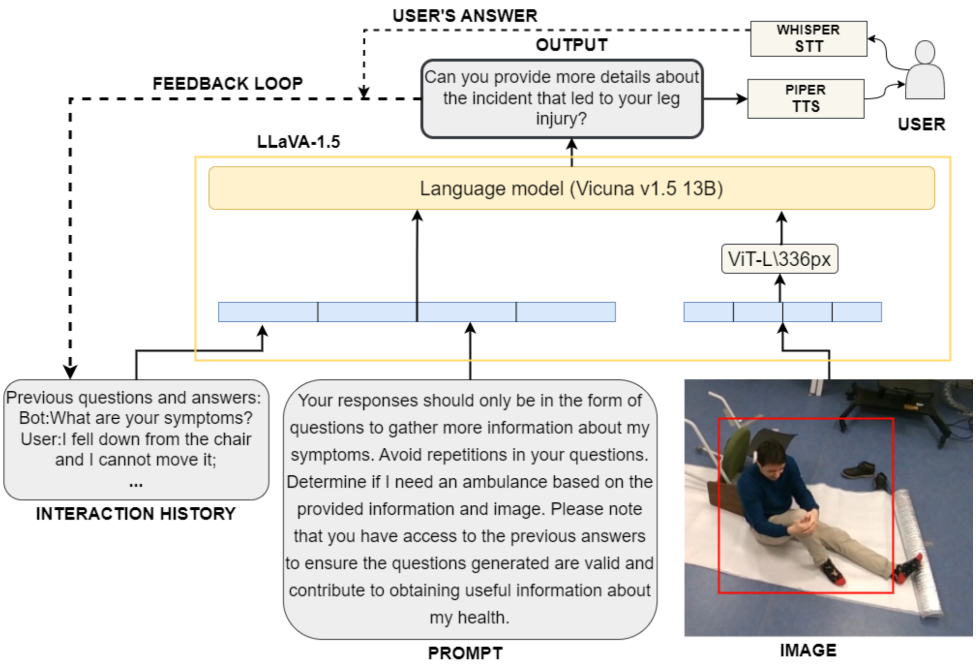


Fig. 5: An illustration of LLaVA's dynamic question generation process involves forwarding the prompt, image embedding, and interaction history to the Vicuna LLM. This prompts LLaVA to generate questions tailored to the context and the user's previous responses, ensuring coherence and relevance in follow-up questioning.

Generation of the questions: Once abnormal behavior is detected and confirmed by the user, the LLaVA model dynamically generates a set of questions tailored to the user's visual context and prior responses. Figure 5 illustrates this process, where the LLaVA model is tasked with creating contextually relevant and non-repetitive questions based on the prompt and visual cues. The prompt and context are provided as input to LLaVA's LLM (Vicuna v1.5 13B), which generates the follow-up questions accordingly. These questions are then stored in a "question.txt" file and sent to the local part for processing by speech models. Subsequently, the user's responses are converted to text and stored, forming part of the interaction history. LLaVA utilizes this history to generate more detailed and contextually relevant follow-up questions in subsequent interactions.

Activation of TTS and STT: Questions generated by the LLaVA model are audibly presented using Piper TTS, ensuring precise communication for effective user assessment. Subsequently, the Whisper STT model is activated to capture and transcribe the user's responses into the "answers.txt" file on the local side. Whisper STT intelligently halts recording after three seconds of silence, a customizable feature designed to manage conversational pauses.

Analysis of the interaction: The "answers.txt" file is then transmitted back to the server for analysis by the LLaVA model. Based on the user's responses and previous

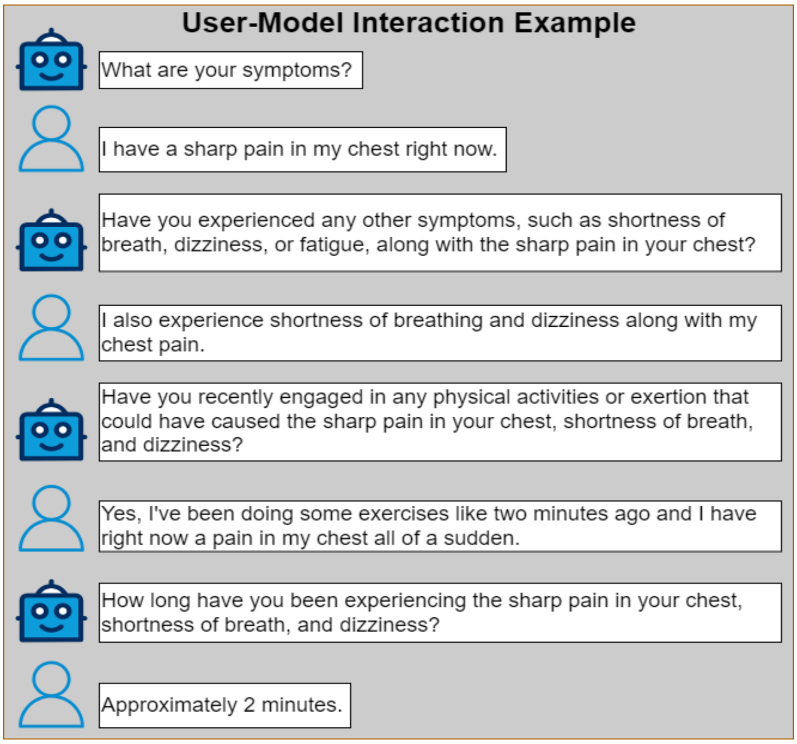


Fig. 6: User-model interaction during an emergency scenario, such as a heart attack, is illustrated in the figure. On the left, chatbot-generated questions, informed by the user's previous responses and image embedding, are displayed. On the right, the user's answers are shown.

image embeddings, the system gauges the severity of the situation. If the "answers.txt" file is empty, indicating no response from the individual, the system interprets it as unresponsiveness, triggering an immediate ambulance call. Conversely, if responses are present, LLaVA scrutinizes them alongside historical data and image analysis to ascertain the necessity for emergency services. If deemed necessary, the system initiates an ambulance call. Figure 6 illustrates a sample interaction between the user and the LLaVA model during a simulated heart attack scenario.

Decision-making: The iterative process of generating questions and analyzing user responses continues until the LLaVA model determines whether to call an ambulance. Operating autonomously, the system makes informed decisions to prioritize user safety and well-being during potential emergencies. Upon detecting an emergency, the model summarizes the situation, provides recommendations for the user, and forwards relevant information to medical professionals or caregivers for further assistance.

Overall, the detailed framework outlined in Appendix A1 demonstrates a comprehensive approach to emergency detection and response. The continuous monitoring stage ensures ongoing vigilance for potential abnormalities, while the interactive stage enables real-time assessment to determine the necessity of calling an ambulance.

By synthesizing information from images and conversations, the system generates a detailed context for informed decision-making.

4 Results

This section elaborates on the experiments conducted with voluntary participants, outlining the procedure, qualitative and quantitative assessments, and comparisons with SOTA works in HAR, AAL, and ABD.

4.1 Ethics and Participants

The experiments received approval from the Nazarbayev University Institutional Research Ethics Committee (NU IREC), guaranteeing adherence to confidentiality, safety, and ethical standards without conflicts of interest, thereby validating the results. Recruited via email among Nazarbayev University students, the study involved 24 participants (9 women and 15 men), predominantly bachelor's and master's degree students aged 20 to 25.

4.2 Experimental Procedure and Dataset

Participants were briefed on the experimental procedure and requested to read and sign a consent form. They were then tasked with performing eight scenarios, comprising five emergencies (heart attack, heart attack with fainting, head or brain injury, broken leg, open wound) and three normal activities (watching TV, reading a book, sitting with a laptop). These scenarios were selected to evaluate the system's ability to handle common emergency cases. Leveraging LLaVA's pre-training and instruction-following capabilities, the model can generalize from related concepts, identifying new normal and abnormal cases. Examples of these scenarios are shown in Figure 7. A novel dataset for ABD was compiled, consisting of 192 videos and images capturing emergency (e.g., heart attack, fainting) and non-emergency behaviors (e.g., watching TV). Videos were recorded using a mobile phone, while images were captured by an RGB camera positioned in the room's top corner. We collected 24 videos and frames for each behavior class, resulting in a total of 192 videos and images. Here is a detailed description of all the scenarios used to test the proposed system and dataset:

- *Heart attack*: The participant sits with hands on the chest, mimicking symptoms like chest pain, difficulty breathing, fatigue, and dizziness.
- *Heart attack with fainting*: Similar to the first, but the participant appears unable to speak or respond, simulating fainting or loss of consciousness.
- *Head/brain injury*: Participants act as if they've lost memory and orientation, mimicking headaches and disorientation.
- *Broken leg*: The person has accidentally fallen from a chair and cannot move their leg and stand up. In this scenario, the participants sit on the floor, holding their legs near a fallen chair. The symptoms include severe leg pain, inability to move, losing feelings in their leg, swelling, and others.

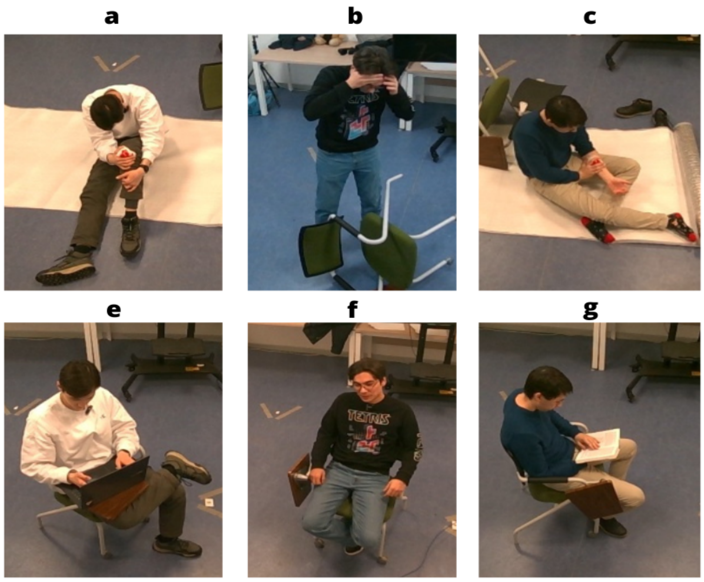


Fig. 7: Emergency and non-emergency scenarios. The figure shows three examples of emergency cases: a) broken leg, b) head injury, c) open wound, and three non-emergency cases: e) playing on laptop, f) watching TV, and g) reading a book.

- *Open wound*: Following a simulated fall from a chair, the participant pretends to have an unstoppable bleeding cut, wrapping the wound in a painted red bandage, displaying symptoms of severe pain and potential loss of consciousness.
- *Watching TV*: The participant sits comfortably, watching TV without any signs of discomfort.
- *Read a book*: The participant is engaged in reading a book while sitting on a chair.
- *Playing with laptop*: The participant is sitting on a chair with a laptop and pretending to be working or playing a video game.

4.3 Evaluation

The qualitative assessment of the experiments involved two main approaches. First, participants completed a subjective questionnaire comprising ten questions, aiming to evaluate various aspects such as difficulty, complexity, speed, and consistency, utilizing constructs from NASA-Task Load Index (NASA-TLX) and System Usability Scale (SUS) [61, 62]. Second, participants provided feedback on the performance of the LLaVA model through concise questions, assessing its ability to offer contextually relevant suggestions.

The quantitative evaluation of the experiments encompasses various assessments. Firstly, the VQA Task Accuracy measures how effectively the model handles scenario-specific instructions by comparing its responses to ground truths. Interaction Activation evaluates the system's ability to initiate interactions with users during

Table 2: Mean results for each question in the subjective questionnaire 4.4. Q - question, F - females, M - males, O - average of F and M.

Q	1	2	3	4	5	6	7	8	9	10
F	1.44	1.89	1.33	2	1.78	4.67	1	4.89	1.44	4.78
M	1.27	1.93	1.47	2.13	1.47	4.6	1.8	4.53	1.27	4.73
O	1.33	1.92	1.42	2.08	1.57	4.62	1.37	4.75	1.33	4.75

emergencies, considering metrics like accuracy, recall, precision, F1 score, and specificity. Final Decision Accuracy assesses the system's post-interaction capability to classify cases as emergencies needing an ambulance or non-emergencies. Additionally, Multimodal Classification utilizes multimodal contextual information to categorize responses into eight classes, distinguishing between five emergency and three normal activities. Finally, System Performance Time gauges the speed and efficiency of the system's internal processes in detecting and assisting in emergency situations.

4.4 Qualitative Evaluation of the Results

The qualitative evaluation of the experiment results involved a subjective questionnaire comprising 10 questions. 24 participants, including 9 women and 15 men, completed the questionnaire. Each question had response options ranging from 1 to 5. The questions are as follows:

1. How difficult did you find the experiment to complete? (1 is easy, 3 is medium and 5 is difficult)
2. How much mental and perceptual activity was required? (1 is not much mental activity, 3 is medium and 5 is too much mental activity)
3. How much physical activity was required? Was the task easy or demanding, slack or strenuous? (1 is easy, 3 is demanding and 5 is strenuous)
4. Did you find the system too slow? (1 is normal, 3 is slightly slow and 5 is too slow)
5. Did you find the system too fast? (1 is normal, 3 is slightly faster than expected and 5 is too fast)
6. How intuitive was the interaction with the system? (1 is not intuitive, 3 is average level of intuitiveness and 5 is very intuitive)
7. Did you find the system unnecessary complex? (1 is not complex at all, 3 is average level of complexity and 5 is very complex)
8. Did you find the system easy to use? (1 is not easy at all, 3 is average level of easiness and 5 is very easy)
9. Did you find too much inconsistency in the system's workflow? (1 is no inconsistencies at all, 3 is some inconsistencies and 5 is lots of inconsistencies)
10. Did you feel confident interacting with the system? (1 is not confident at all, 3 is average level of confidence and 5 is very confident)

Table 2 displays the mean responses (ranging from 1 to 5) for each question, segmented by male and female groups, along with overall means. Generally, responses tend toward positive scales, with potential statistical disparities between genders. Further analysis necessitates employing statistical tests to confirm differences between groups.

4.4.1 Statistical Analysis

First, it's essential to determine the appropriate statistical test for the results. While the data appears non-normally distributed, confirming this assumption requires a normality test. The Shapiro-Wilk test for normality [63] is among the most robust. It evaluates two hypotheses:

H_0: The data follows a normal distribution.
H_A: The data does not follow a normal distribution.

The Shapiro-Wilk test for normality uses the following test statistic:

$$W = \frac{(\sum_{i=1}^{n} a_i x_(i))^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{1}$$

In this scenario, n represents the sample size, which is 24 participants. a_i denotes a critical value obtained from the Shapiro-Wilk test table, varying based on the sample size. $x_{(i)}$ signifies the participant's response on a scale from 1 to 5, and $\overline{x}$ denotes the sample mean for each question from Table 2. The null hypothesis is rejected if the calculated W value is significantly less than 1 or falls below a critical threshold from the Shapiro-Wilk test table. For instance, for one question, the resulting W value was calculated as 0.02, considerably deviating from the ideal value of 1. By comparing W to a critical value of 0.916 at a significance level of 0.05, derived from the Shapiro-Wilk test table with p values, the null hypothesis (H_0) is rejected due to W being below the critical threshold.

Since the data collected does not exhibit normal distribution, non-parametric statistical tests are warranted to determine whether there is a distinction between the responses of male and female participants. With two distinct data groups (male and female), the Mann-Whitney U test [64] emerges as one of the most suitable options for this scenario. The Mann-Whitney U test entails two hypotheses:

H_0: Given two selected groups of data, X (male answers) and Y (female answers), there is no statistical difference between the two groups
H_A: Given two selected groups of data, X (male answers) and Y (female answers), there is a statistical difference between the two groups.

Mann-Whitney U test uses the U statistic calculated by taking the smallest of U_1 and U_2. U_1 and U_2, in turn, are defined as follows:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 \tag{2}$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 \tag{3}$$

In equations 2 and 3, R_1 and R_2 signify the sum of ranks for male and female responses, respectively. To assign ranks, both groups are combined into one dataset, and ranks are assigned from the smallest to the largest values, ranging from 1 to 24. Here, n_1 and n_2 denote the number of male (15) and female (9) participants, respectively. After computing U_1 and U_2, the smaller value is selected for comparison with the critical value from the Mann-Whitney U test reference table. If the observed U statistic is less than or equal to the critical value at a confidence level of 0.05, the null hypothesis (H_0) is rejected.

Table 3: U statistic and U critical for both one-tailed and two-tailed tests

Question	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
U statistical	***38***	44	**27**	44	49	**27**	**0**	41	**27**	**23**
U critical (one-tailed)	34	34	**34**	34	34	**34**	**34**	34	**34**	**34**
U critical (two-tailed)	**39**	39	**39**	39	39	**39**	**39**	39	**39**	**39**

Table 3 showcases the calculated U statistic for each question alongside the U critical values for both one-tailed and two-tailed tests. In this context, we aim to discern whether male responses differ from female responses, allowing for differences in both directions (larger or smaller responses). Thus, the U-critical for the two-tailed test is considered, encompassing variations in both directions. Observing Table 3, we note that the null hypothesis H_0 can be rejected for questions 1, 3, 6, 7, 9, and 10. This determination arises because the $U_{\text{statistical}}$ values are less than or equal to the critical value (U_{critical} two-tailed), indicating a statistical distinction between male and female responses to the aforementioned questions.

4.4.2 Discussion of Statistical Analysis and Questionnaire

To delve deeper into the differences between male and female responses to the experimental questions, a nuanced analysis is imperative. For instance, examination of the first question indicates that women perceived the experiments as more challenging, evident from their higher average scores compared to men, as detailed in Table 2. In contrast, the third question suggests that men reported higher physical demands, likely attributed to scenarios necessitating physical actions such as lying on the ground. Notably, the slight variance in means for the sixth question implies that women found the system more intuitive. Furthermore, it is noteworthy that all female participants found the system easy to use according to question 7, unlike some male participants. To conduct a more comprehensive analysis, statistical tests should be applied to ascertain the statistical disparities between the groups.

4.4.3 Performance of the LLaVA Model in Suggestion and Question Generation

Alongside the subjective questionnaire, participants were also tasked with assessing the model's performance using the Likert Scale [65]. They were prompted to rate the questions on the following scale:

1. The model generated questions without any relevance to the case scenario and cohesion;
2. The model generated questions that are somewhat related to the case scenario;
3. The model generated questions that are relevant to the case scenario and concise;
4. The model generated questions that are relevant to the case scenario, concise, and clearly directed towards figuring out the state of the user.

Furthermore, the participants were also asked to evaluate the suggestions generated by the model at the end of the interaction using the following Likert Scale:

1. The model generated a suggestion without any relevance to the case scenario and cohesion;
2. The model generated a suggestion that is somewhat related to the case scenario, however, is not concise enough and cohesive;
3. The model generated a suggestion which is relevant to the case scenario and concise;
4. The model generated a detailed suggestion that carefully regards the state of the user and provides concise and cohesive advice accordingly.

After analyzing the responses, the average score for the model's question generation was *3.75*, indicating generally relevant and concise inquiries that effectively guided its understanding of the situation. However, occasional questions unrelated to the person's condition but focusing on circumstances or surroundings during the emergency may have affected the score. Additionally, detailed participant responses during the initial question might have challenged the model's ability to formulate appropriate follow-up inquiries.

Similarly, the average score for the model's suggestions was also *3.75*, reflecting generally detailed advice and probable diagnoses. However, occasional lack of context in participant responses may have led to broader suggestions. This limitation might have caused the model to speculate on potential diagnoses, impacting the overall score.

4.5 Quantitative Evaluation of the Results

The quantitative assessment of the emergency detection system encompasses key metrics to evaluate its performance in identifying emergencies accurately, initiating timely user interactions during emergency situations, and determining the need for an ambulance based on the severity of the situation.

4.5.1 Evaluation of VQA Accuracy in Emergency Detection

The accuracy of the VQA mechanism is fundamental for assessing the monitoring system's performance. This component, integral to the continuous monitoring block, determines the abnormality score, which guides the decision to activate the interaction block. The VQA's effectiveness is gauged by its ability to accurately identify signs of emergencies and provide correct answers to a predefined set of questions. Accuracy serves as the primary metric for evaluating VQA.

VQA Accuracy: The accuracy of VQA is determined by assessing the model's responses to a predefined set of questions aimed at identifying signs of abnormality across various scenarios. It is calculated using the following formula:

$$\text{AccVQA} = \frac{\text{Number of Correct Responses}}{\text{Total Number of Responses}} \tag{4}$$

Each scenario presented to the system, ranging from a heart attack to using a laptop, was accompanied by a series of questions aimed at determining abnormality scores. The system's performance was evaluated by comparing the ground truth answers with the observed answers ("Yes" and "No") to the questions from Figure 4. Additionally, Table A1 from Appendix A illustrates the comparison between predicted and true values of the VQA task. Table 4 below presents the frequency of correct responses for each VQA question within the emergency context. Overall, there are 7 questions for each scenario, comprising 5 emergency scenarios and 3 non-emergency scenarios, with 24 participants. The model provides answers to a set of questions (7 predefined questions) across all emergency cases. This entails 120 answers to one question for all scenarios (24 participants × 5 emergency cases), 24 answers to one question for each emergency case, and 840 answers (5 emergency cases × 24 participants × 7 questions) for all questions for each emergency scenario.

Table 4: Correct VQA answers for all the emergency and non-emergency scenarios. AVG means average of the results and std means standard deviation.

Question	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**AVG (std)**
Emergency Cases								
Heart Attack	24	24	22	8	21	11	14	17.71 (6.60)
Fainting	24	24	23	10	23	8	12	17.71 (7.32)
Head Injury	24	18	17	19	6	22	12	16.86 (5.67)
Broken Leg	24	13	23	11	10	14	13	15.43 (5.68)
Open Wound	13	12	21	7	17	18	15	14.71 (4.57)
AVG (std)	21.8 (4.92)	18.2 (5.76)	21.2 (2.49)	11 (4.74)	15.4 (7.23)	14.6 (5.55)	13.2 (1.17)	-
Non-Emergency Cases								
Watching TV	96	96	96	95	93	92	92	94.28 (1.9)
Read a Book	96	96	96	96	96	96	96	96 (0)
Playing on Laptop	96	96	96	96	96	94	96	95.71 (0.76)
AVG (std)	96.0 (0)	96.0 (0)	96.0 (0)	95.67 (0.58)	95.0 (1.73)	94.0 (2)	94.66 (2.31)	-

Table 4 illustrates the frequency of correct responses for each VQA question within the non-emergency cases. For each question, there are 3 non-emergency scenarios, which were tested four times. That is, questions (1-7) were answered by the model for a specific non-emergency case 4 times to validate its effectiveness in detecting normal activity without enabling the system's interaction with a person. Consequently, there are 288 answers to one question for all non-emergency cases (3 normal cases $\times$ 4 repetitions $\times$ 24 participants), 96 answers to one question for one non-emergency case (24 participants $\times$ 4 repetitions), and 2016 answers to all the questions for each non-emergency scenario (96 answers to one question for one non-emergency $\times$ 3 non-emergency cases $\times$ 7 questions). Each column represents a specific VQA question that assesses the presence of emergency indicators. A higher count indicates more instances that were correctly identified by the VQA system. Table 5 displays the results of VQA accuracy for emergency and non-emergency cases (correct answers on the predefined set of the questions by the LLaVA model) based on the distribution of answers in Table 4.

Table 5: VQA Accuracy for Different Scenarios

Type	Scenario	VQA ACC.
Emergency	Heart Attack	73.81%
	Fainting	73.80%
	Head Injury	70.24%
	Broken Leg	64.29%
	Open Wound	62.00%
Non-emergency	Watching TV	98.21%
	Reading a Book	100.00%
	Playing on Laptop	99.70%

4.5.2 Discussion of the VQA

In emergency scenarios, the system's VQA performance varies, with heart attack, fainting, and head injury scenarios achieving relatively high accuracy results of 70-74%, indicating the system's proficiency in identifying critical situations. However, the accuracy drops to around 62-64% for broken leg and open wound cases, as detailed in Table 5. This variation suggests different sensitivities to specific symptoms or visual cues.

The data shows that the VQA system consistently recognizes classic emergency signs in heart attack and fainting scenarios. In contrast, the accuracy decreases for broken leg and open wound cases, which may come from the varied visual presentations and difficulties in capturing clear images of the affected areas or subtle signs like facial expressions due to the camera's distance or angle. Also, some participants hid their faces and arms from the camera, making it challenging for the model to see the emotion on their faces or the protective postures. Additionally, the participants exhibited similar emotions and protective postures in emergency cases differently.

Conversely, the results highlight the system's robust performance in non-emergency cases like watching TV or reading a book, where it reliably identifies these as non-critical with high accuracy, as shown in Table 4. This demonstrates the system's effectiveness in distinguishing between everyday activities and potential emergencies, reducing false alarms and enhancing user convenience.

4.5.3 Evaluation of Pre-Interaction VQA Binary Activation

The binary activation for emergency cases refers to the accurate initiation of user interaction when an emergency scenario is detected, while for non-emergency cases, it indicates the correct absence of interaction when the situation is deemed normal following VQA assessment.

In developing the monitoring system, determining the optimal threshold for emergency detection was crucial. Through manual testing, this threshold was meticulously fine-tuned to ensure the system triggers user interaction appropriately when an emergency is likely, despite potential inaccuracies in some VQA responses. Activation occurs when the system's abnormality score reaches or surpasses this threshold, prompting immediate response to the perceived emergency.

The rationale behind this threshold calibration lies in effectively managing the complexity of real-world scenarios, striking a balance between the need to avoid overlooking genuine emergencies and minimizing unnecessary interactions. This threshold accommodates a margin of error in question responses, prioritizing the overall pattern of answers to uphold high sensitivity to emergencies.

Table 6: Confusion Matrix for Binary Activation Evaluation

		True Class	
		Positive	**Negative**
Predicted Class	**Positive**	114	8
	Negative	6	64

Table 6 presents the breakdown of true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN). Each entry corresponds to one activation for one participant and one scenario. Therefore, the total number of cases considered is 192, resulting from 24 participants across 8 scenarios.

The system's binary activation performance metrics were calculated as follows: accuracy (93.44%), precision (92.71%), recall (95%), F1 score (93.84%), and specificity (88.88%). These metrics were derived from the TP, FP, FN, and TN values provided in Table 6. With a binary classification system for activating interactions, outcomes are categorized as either abnormal or non-abnormal during the pre-interaction stage. Additional examples of VQA with decisions that trigger the interaction block are presented in Table A1 and Table A2 from Appendix A.

4.5.4 Discussion of Pre-Interaction VQA Binary Activation

The confusion matrix (see Table 6) and resulting performance metrics provide crucial insights into the functionality of the VQA system within the human care monitoring framework. With 114 TP and 64 TN, the system adeptly identifies both emergencies

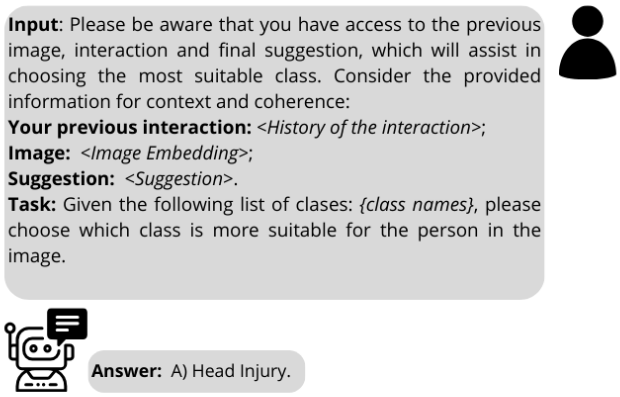


Fig. 8: LLaVA model's instructions for classification task: To classify abnormal and normal behavior based on cumulative context such as image embedding and interaction history. If ambulance call chosen, confirms abnormal behavior.

and non-emergencies, showcasing responsiveness and discernment. A low FP rate of 8 minimizes unnecessary distress for users and caregivers, while the presence of only 6 FNs indicates rare failures in emergency detection.

These performance metrics underscore the system's overall effectiveness, demonstrating robust emergency detection and accurate non-emergency identification. This suggests a well-optimized threshold for initiating user interaction. However, the presence of some FNs highlights the ongoing need for system refinement to further enhance emergency differentiation.

4.5.5 Evaluation of Post-Interaction Classification

After engaging with the user, the LLaVA model undertook a binary classification task (determining whether the situation warranted calling an ambulance) as depicted in Figure 8. If the LLaVA model opted to summon emergency services, it categorized the observed activity as an emergency; otherwise, it classified it as non-emergency. Subsequently, the model provided recommendations based on gathered information (interaction history and image embedding). Additionally, the LLaVA model was directed to perform a multi-class classification task (identifying among 5 emergency and 3 non-emergency classes) using cumulative context such as image embedding, interaction history, and final suggestions post-binary classification. Figure A3 from Appendix A illustrates how the LLaVA model was guided to execute the multi-class classification task.

The accuracy for both 8-class and 2-class classifications reached 100%, indicating outstanding classification outcomes, suggesting the model possesses ample contextually relevant information to make precise determinations, assign the correct class, and initiate ambulance calls. 192 cases (24 participants $\times$ 8 scenarios) were assessed for post-interaction binary classification and multi-class classification.

4.5.6 Discussion of Post-Interaction Classifications of the Model

The efficiency of our system is enhanced when the LLaVA model enters the interaction mode with the user, demonstrating perfect accuracy in deciding whether to call an ambulance in emergencies or to refrain in non-critical situations. By processing the user's responses in sequence, the model collects the context via real-time interaction with the user and the image context. This affords it a comprehensive understanding of the situation, generating relevant suggestion at the end. Integrating speech models for direct communication has evidently optimized the LLaVA model's ability to classify abnormal activities and daily living activities, achieving 100% accuracy for both binary classification (emergency or non-emergency class) and multi-class classification (5 emergency and 3 non-emergency classes). Some additional examples are presented in the Table A3 and Table A4 from Appendix A.

Table 7: Baseline performance metrics from various studies, including DT (Decision Tree), RF (Random Forest), SVM (Support Vector Machines), WKMC (Weighted K-Mean Clustering), ANN (Artificial Neural Network), and MSResNet (Multiscale Residual Network).

Study	**Year**	**Method**	**Classes**	**Accuracy (%)**
[38]	2016	Intrusive sensors + DT	6 ADL	94
[39]	2017	Intrusive and Non-intrusive sensors + DT + HMM	11 ADL	93
[40]	2018	Non-intrusive sensors + DCNN	10 ADL	98.54
[41]	2019	Intrusive sensors + SVM	4 ADL	88
[42]	2019	Non-intrusive + Adv. Belief Model	14 ADL	99.74
[8]	2019	Vision sensor + HMM	9 ADL	84.33
[43]	2020	Wearable sensors + Ensemble Classifier	1 Fall + 6 ADL	93.90
[44]	2020	Non-intrusive sensors + HHMM	11 ADL	65.20
[45]	2021	Wearable sensors + DCNN + LSTM	5 ADL	85.90
[19]	2022	Vision sensors + LSTM-RNN	6 ADL	92.83
[18]	2022	Vision sensors + MSResNet	2 Gait Dysfunctions + 1 Normal Gait	93
[20]	2023	Vision sensors + BERT	5 Falls + 7 ADL	99.50
[46]	2023	Wearable sensors + RF	6 ADL	99.30
[21]	2023	Vision sensors + YOLO + Linear Regression	11 violent actions + 1 Non-violent	80.50
[47]	2023	Intrusive and Non-intrusive sensors + WKMC + ANN	1 malicious + 1 non-malicious action	93.15
Proposed (pre)	2024	Vision sensors + VLM + TTS + STT	1 emergency + 1 non-emergency	93.44
Proposed (post)	2024	Vision sensors + VLM + TTS + STT	1 emergency + 1 non-emergency	**100**
Proposed (post)	2024	Vision sensors + VLM + TTS + STT	5 emergency + 3 non-emergency	**100**

While the system demonstrated outstanding performance in noise-free conditions during testing, its effectiveness in noisy environments remains untested. Although the Whisper STT model performed adequately in pub noise, additional testing in noisy settings is necessary to assess its ability to accurately transcribe speech amidst background noise. Additionally, the Whisper model occasionally generates hallucinations in high-noise or mumbled speech situations, resulting in the prediction of words not spoken by the user [66].

4.5.7 Time Evaluation of The System

The time performance of our system, detailed in Table 8, outlines the average duration for emergency detection, user interaction, and decision-making. Utilizing two NVIDIA V100 GPUs for the LLaVA model and a GeForce GTX-1060 GPU for human detection and speech models, the system swiftly captures photos in less than a second. However, network latency notably impacts image and file transfer times between server and local components.

During the interaction phase, triggered by transmitting a "mode.txt" file and concluding with final suggestion generation, the LLaVA model generates and processes four questions, with Piper TTS vocalizing and Whisper STT listening four times each. This sequence, found optimal during experiments, facilitates detailed and contextually relevant question generation. Interaction time averages 109 seconds, notably longer than the 29 seconds for continuous monitoring, primarily due to voicing and listening times varying based on user responses.

Overall, the system completes all processes in 2 minutes and 34 seconds. Participant feedback from a subjective questionnaire (see Table 2) indicates generally satisfactory speed. Future assessments should consider testing the system's time performance in real-world environments like hospitals or homes, especially with different user groups such as the elderly or disabled, to gauge effectiveness more accurately. Referring to real-time results, the average time for the system to detect emergencies, interact, and make final decisions about calling an ambulance aligns with subjective survey results, as shown in Table 8. Nevertheless, it's crucial to evaluate the system's time performance in real environments (within hospitals, homes, and other locations) with diverse user demographics to assess speed more accurately.

5 Conclusion, Future Work, Limitations

In conclusion, the proposed framework tackles healthcare challenges by offering real-time assistance in emergencies, leveraging VLM, LLM, and TTS/STT technologies for continuous monitoring and user interaction. Despite occasional inaccuracies in VQA responses, the system achieves high accuracy, precision, and recall, ensuring reliable emergency detection. Additionally, it demonstrates robustness in multi-class classification tasks.

However, the system faces limitations. Fine-tuning VQA responses for subtle indicators like facial expressions is needed, along with addressing image comprehension issues due to camera quality and distance. Enhanced language models and instructions can mitigate occasional hallucinations in generated questions. System latency could be reduced through edge computing, enhancing processing speed and reducing reliance on remote servers. Moreover, improving Whisper ASR's ability to filter noise

Table 8: Time performance of the proposed system. The averages were calculated across 24 patients, each experiencing 5 emergency care scenarios. The LLaVA model ran on two NVIDIA V100 GPUs on the server side, while YOLOv8, Whisper, and Piper operated on a single GeForce GTX-1060 GPU on the local side. Italicized entries denote processing time per instance and are not included in the overall system time.

Process in the system	**AVG. time (s)**
Continuous Monitoring Block	
Time for the model to detect the person (YOLOv8):	0.69
Time for taking the photo (Depth Camera)	0.40
Time to send one image to the server:	5.71
Time for sending the "mode.txt" file:	5.92
Time for LLaVA to answer on the VQA (7 questions):	4.32
Time for LLaVA to answer on one question:	*0.61*
Time to send "vqa.txt" to local:	1.73
Time for voicing the question (Piper TTS) to assure help:	6.86
Time for listening to the user (Whisper STT):	3.68
User-Model Interaction Block	
Time for sending the "mode.txt" with value 1 to server:	6.49
Time for generation of the four questions (LLaVA):	6.50
Avg. time for generation one question:	*1.62*
Time for sending the question to local part four times:	7.33
Time for sending one question to local part:	*1.83*
Time for voicing the question (Piper TTS) four times:	20.64
Time for voicing one question (Piper TTS):	*5.16*
Time for listening to the user's answers (Whisper STT) four times:	14.81
Time for listening to one user's answer (Whisper STT):	*3.70*
Time for sending "answer.txt" file to server four times:	23.40
Time for sending one "answer.txt" file to server:	*5.85*
Time for the model (LLaVA) to make a final decision:	2.04
Time for generation of the suggestion (LLaVA):	6.91
Time for voicing the suggestion (Piper TTS):	21.02
Overall time of the whole system:	**154$\pm$ 12**

and process speech accurately is crucial. Finally, diversifying the dataset to include various environments and scenarios is essential for comprehensive testing.

Future work will focus on refining VQA performance, integrating a database for interaction history, improving camera quality and positioning, enhancing language models, reducing system latency, enhancing Whisper ASR's noise handling, and diversifying the dataset for robustness.

References

[1] Sadigov, R., *et al.*: Rapid growth of the world population and its socioeconomic results. The Scientific World Journal **2022**, 1–8 (2022)

[2] Petrova, N., Khvostikova, D.: Prevalence, structure, and risk factors for mental disorders in older people. Advances in Gerontology **11**, 409–415 (2021)

[3] Buzzelli, M., Albé, A., Ciocca, G.: A vision-based system for monitoring elderly people at home. Applied Sciences **10**(1), 374 (2020)

[4] Lentzas, A., Vrakas, D.: Non-intrusive human activity recognition and abnormal behavior detection on elderly people: A review. Artificial Intelligence Review **53**(3), 1975–2021 (2020)

[5] Berger, A., Horst, F., Müller, S., Steinberg, F., Doppelmayr, M.: Current state and future prospects of eeg and fnirs in robot-assisted gait rehabilitation: a brief review. Frontiers in Human Neuroscience **13**, 172 (2019)

[6] Poongodi, M., Sharma, A., Hamdi, M., Maode, M., Chilamkurti, N.: Smart healthcare in smart cities: wireless patient monitoring system using IoT. The Journal of Supercomputing, 1–26 (2021)

[7] Antón, M.Á., Ordieres-Meré, J., Saralegui, U., Sun, S.: Non-invasive ambient intelligence in real life: Dealing with noisy patterns to help older people. Sensors **19**(14), 3113 (2019)

[8] Kim, K., Jalal, A., Mahmood, M.: Vision-based human activity recognition system using depth silhouettes: A smart home system for monitoring the residents. Journal of Electrical Engineering & Technology **14**(6), 2567–2573 (2019)

[9] Banos, O., Garcia, R., Holgado-Terriza, J.A., Damas, M., Pomares, H., Rojas, I., Saez, A., Villalonga, C.: mhealthdroid: a novel framework for agile development of mobile health applications. In: Proc. of the International Workshop on Ambient Assisted Living, pp. 91–98 (2014). Springer

[10] Windsor, R., Jamaludin, A., Kadir, T., Zisserman, A.: Vision-language modelling for radiological imaging and reports in the low data regime. arXiv preprint arXiv:2303.17644 (2023)

[11] Sharma, D., Purushotham, S., Reddy, C.K.: Medfusenet: An attention-based multimodal deep learning model for visual question answering in the medical domain. Scientific Reports **11**(1), 19826 (2021)

[12] Naseem, U., Khushi, M., Kim, J.: Vision-language transformer for interpretable pathology visual question answering. IEEE Journal of Biomedical and Health Informatics **27**(4), 1681–1690 (2022)

[13] Reddy, V.M., Vaishnavi, T., Kumar, K.P.: Speech-to-text and text-to-speech recognition using deep learning. In: Proc. of the International Conference on Edge Computing and Applications (ICECAA), pp. 657–666 (2023). IEEE

[14] Effendi, J., Sakti, S., Nakamura, S.: Weakly-supervised speech-to-text mapping with visually connected non-parallel speech-text data using cyclic partially-aligned transformer. In: Proc. of the Interspeech, pp. 2257–2261 (2021)

[15] Peng, C., Yang, X., Smith, K.E., Yu, Z., Chen, A., Bian, J., Wu, Y.: Model tuning or prompt tuning? a study of large language models for clinical concept and relation extraction. arXiv preprint arXiv:2310.06239 (2023)

[16] Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744 (2023)

[17] Hartsock, I., Rasool, G.: Vision-language models for medical report generation and visual question answering: A review. arXiv preprint arXiv:2403.02469 (2024)

[18] Kaur, R., Motl, R.W., Sowers, R., Hernandez, M.E.: A vision-based framework for predicting multiple sclerosis and parkinson's disease gait dysfunctions—a deep learning approach. IEEE Journal of Biomedical and Health Informatics **27**(1), 190–201 (2022)

[19] Batool, M., Javeed, M.: Fundamental recognition of adl assessments using machine learning engineering. In: Proc. of the International Bhurban Conference on Applied Sciences and Technology (IBCAST), pp. 195–200 (2022). IEEE

[20] Ramirez, H., Velastin, S.A., Cuellar, S., Fabregas, E., Farias, G.: Bert for activity recognition using sequences of skeleton features and data augmentation with gan. Sensors **23**(3), 1400 (2023)

[21] Lopez, D.J.D., Lien, C.-C.: Two-stage complex action recognition framework for real-time surveillance automatic violence detection. Journal of Ambient Intelligence and Humanized Computing **14**(12), 15983–15996 (2023)

[22] Zhang, J., Huang, J., Jin, S., Lu, S.: Vision-language models for vision tasks: A survey (2023)

[23] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., *et al.*: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763 (2021). PMLR

[24] Liu, S., Fan, L., Johns, E., Yu, Z., Xiao, C., Anandkumar, A.: Prismer: A vision-language model with an ensemble of experts. arXiv preprint arXiv:2303.02506 (2023)

[25] Wu, W., Wang, X., Luo, H., Wang, J., Yang, Y., Ouyang, W.: Bidirectional cross-modal knowledge exploration for video recognition with pre-trained vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6620–6630 (2023)

[26] Moon, J.H., Lee, H., Shin, W., Kim, Y.-H., Choi, E.: Multi-modal understanding and generation for medical images and text via vision-language pre-training. IEEE Journal of Biomedical and Health Informatics **26**(12), 6070–6080 (2022)

[27] Tanwani, A.K., Barral, J., Freedman, D.: Repsnet: Combining vision with language for automated medical reports. In: Proc. of the International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 714–724 (2022). Springer

[28] Lee, H., Lee, D.Y., Kim, W., Kim, J.-H., Kim, T., Kim, J., Sunwoo, L., Choi, E.: Unixgen: A unified vision-language model for multi-view chest x-ray generation and report generation. arXiv preprint arXiv:2302.12172 (2023)

[29] Bosselut, A., Chen, Z., Romanou, A., Bonnet, A., Hernández-Cano, A., Alkhamissi, B., Matoba, K., Salvi, F., Pagliardini, M., Fan, S., et al.: Meditron: Open medical foundation models adapted for clinical practice (2024)

[30] Wang, G., Liu, X., Ying, Z., Yang, G., Chen, Z., Liu, Z., Zhang, M., Yan, H., Lu, Y., Gao, Y., et al.: Optimized glycemic control of type 2 diabetes with reinforcement learning: a proof-of-concept trial. Nature Medicine, 1–10 (2023)

[31] Harris, J., Zaki, M.J.: Neural models for generating natural language summaries from temporal personal health data. Journal of Healthcare Informatics Research, 1–30 (2024)

[32] Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., et al.: Towards expert-level medical question answering with large language models. arXiv preprint arXiv:2305.09617 (2023)

[33] Han, T., Adams, L.C., Papaioannou, J.-M., Grundmann, P., Oberhauser, T., Löser, A., Truhn, D., Bressem, K.K.: Medalpaca–an open-source collection of medical conversational ai models and training data. arXiv preprint arXiv:2304.08247 (2023)

[34] Li, Y., Li, Z., Zhang, K., Dan, R., Jiang, S., Zhang, Y.: Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. Cureus **15**(6) (2023)

[35] Chowdhery, A., et al.: Scaling language modeling with pathways. PaLM (2022)

[36] Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., *et al.*: Large language models encode

clinical knowledge. Nature **620**(7972), 172–180 (2023)

[37] He, J., Li, F., Li, J., Hu, X., Nian, Y., Xiang, Y., Wang, J., Wei, Q., Li, Y., Xu, H., et al.: Prompt tuning in biomedical relation extraction. Journal of Healthcare Informatics Research, 1–19 (2024)

[38] Capela, N., Lemaire, E., Baddour, N., Rudolf, M., Goljar, N., Burger, H.: Evaluation of a smartphone human activity recognition application with able-bodied and stroke participants. Journal of Neuroengineering and Rehabilitation **13**(1), 1–10 (2016)

[39] Fan, X., Xie, Q., Li, X., Huang, H., Wang, J., Chen, S., Xie, C., Chen, J.: Activity recognition as a service for smart home: ambient assisted living application via sensing home. In: Proc. of the IEEE International Conference on AI & Mobile Services (AIMS), pp. 54–61 (2017). IEEE

[40] Gochoo, M., Tan, T.-H., Liu, S.-H., Jean, F.-R., Alnajjar, F.S., Huang, S.-C.: Unobtrusive activity recognition of elderly people living alone using anonymous binary sensors and DCNN. IEEE Journal of Biomedical and Health Informatics **23**(2), 693–702 (2018)

[41] Awais, M., Chiari, L., Ihlen, E.A.F., Helbostad, J.L., Palmerini, L.: Physical activity classification for elderly people in free-living conditions. IEEE Journal of Biomedical and Health Informatics **23**(1), 197–207 (2018)

[42] Ghayvat, H., Awais, M., Pandya, S., Ren, H., Akbarzadeh, S., Chandra Mukhopadhyay, S., Chen, C., Gope, P., Chouhan, A., Chen, W.: Smart aging system: uncovering the hidden wellness parameter for well-being monitoring and anomaly detection. Sensors **19**(4), 766 (2019)

[43] Kerdjidj, O., Ramzan, N., Ghanem, K., Amira, A., Chouireb, F.: Fall detection and human activity classification using wearable sensors and compressed sensing. Journal of Ambient Intelligence and Humanized Computing **11**, 349–361 (2020)

[44] Asghari, P., Soleimani, E., Nazerfard, E.: Online human activity recognition employing hierarchical hidden markov models. Journal of Ambient Intelligence and Humanized Computing **11**, 1141–1152 (2020)

[45] Chan, M.H., Noor, M.H.M.: A unified generative model using generative adversarial network for activity recognition. Journal of Ambient Intelligence and Humanized Computing **12**(7), 8119–8128 (2021)

[46] Thakur, D., Biswas, S.: Guided regularized random forest feature selection for smartphone based human activity recognition. Journal of Ambient Intelligence and Humanized Computing **14**(7), 9767–9779 (2023)

[47] Manocha, A., Kumar, G., Bhatia, M., Sharma, A.: Iot-inspired machine learning-assisted sedentary behavior analysis in smart healthcare industry. Journal of Ambient Intelligence and Humanized Computing **14**(5), 5179–5192 (2023)

[48] Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., *et al.*:

[49] Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.-W., Zhu, S.-C., Tafjord, O., Clark, P., Kalyan, A.: Learn to explain: Multimodal reasoning via thought chains for science question answering. In: The 36th Conference on Neural Information Processing Systems (NeurIPS) (2022)

[50] Lam, K.-Y., Cheng, V.C., Yeong, Z.K.: Applying large language models for enhancing contract drafting (2023)

[51] Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36** (2024)

[52] Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6904–6913 (2017)

[53] Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6700–6709 (2019)

[54] Van, M.-H., Verma, P., Wu, X.: On large visual language models for medical imaging analysis: An empirical study. arXiv preprint arXiv:2402.14162 (2024)

[55] Xenos, A., Foteinopoulou, N.M., Ntinou, I., Patras, I., Tzimiropoulos, G.: Vllms provide better context for emotion understanding through common sense reasoning. arXiv preprint arXiv:2404.07078 (2024)

[56] Mroz, A.P.: Integrating mobile-based text-to-speech (tts) and speech-to-text (stt) to advance proficiency and intelligibility in french. Technological Resources for Second Language Pronunciation Learning and Teaching: Research-based Approaches, 147 (2022)

[57] Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: International Conference on Machine Learning, pp. 28492–28518 (2023). PMLR

[58] Panayotov, V., Chen, G., Povey, D., Khudanpur, S.: Librispeech: an asr corpus based on public domain audio books. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5206–5210 (2015). IEEE

[59] Hansen, M.: Rhasspy/piper: A fast, local neural text to speech system (2022). https://github.com/rhasspy/piper

[60] Talaat, F.M., ZainEldin, H.: An improved fire detection approach based on yolo-v8 for smart cities. Neural Computing and Applications **35**(28), 20939–20954 (2023)

[61] Lewis, J.R.: The system usability scale: past, present, and future. International Journal of Human–Computer Interaction **34**(7), 577–590 (2018)

[62] Hart, S.G.: Nasa-task load index (nasa-tlx); 20 years later. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 50, pp. 904–908 (2006). Sage publications Sage CA: Los Angeles, CA

[63] Khatun, N., *et al.*: Applications of normality test in statistical analysis. Open Journal of Statistics **11**(01), 113 (2021)

[64] MacFarland, T.W., Yates, J.M., MacFarland, T.W., Yates, J.M.: Mann–whitney u test. Introduction to nonparametric statistics for the biological sciences using R, 103–132 (2016)

[65] Joshi, A., Kale, S., Chandel, S., Pal, D.K.: Likert scale: Explored and explained. British journal of applied science & technology **7**(4), 396–403 (2015)

[66] Koenecke, A., Choi, A.S.G., Mei, K., Schellmann, H., Sloane, M.: Careless whisper: Speech-to-text hallucination harms. arXiv preprint arXiv:2402.08021 (2024)

Appendix A Detailed Examples of Proposed System, Prompts, VQA, Binary and Multi-class Classification

This section shows a more detailed state diagram of the proposed system (see Figure A1), additional prompts (see Figures A2, A3) created to adapt the LLaVA model to perform VQA and multi-class classification for ABD. Tables for VQA (see Tables A1, A2), binary classification (see Table A3), and multi-class classification (see Table A4) are presented to demonstrate examples of the interactions between the user and the model and the final suggestions and decisions.

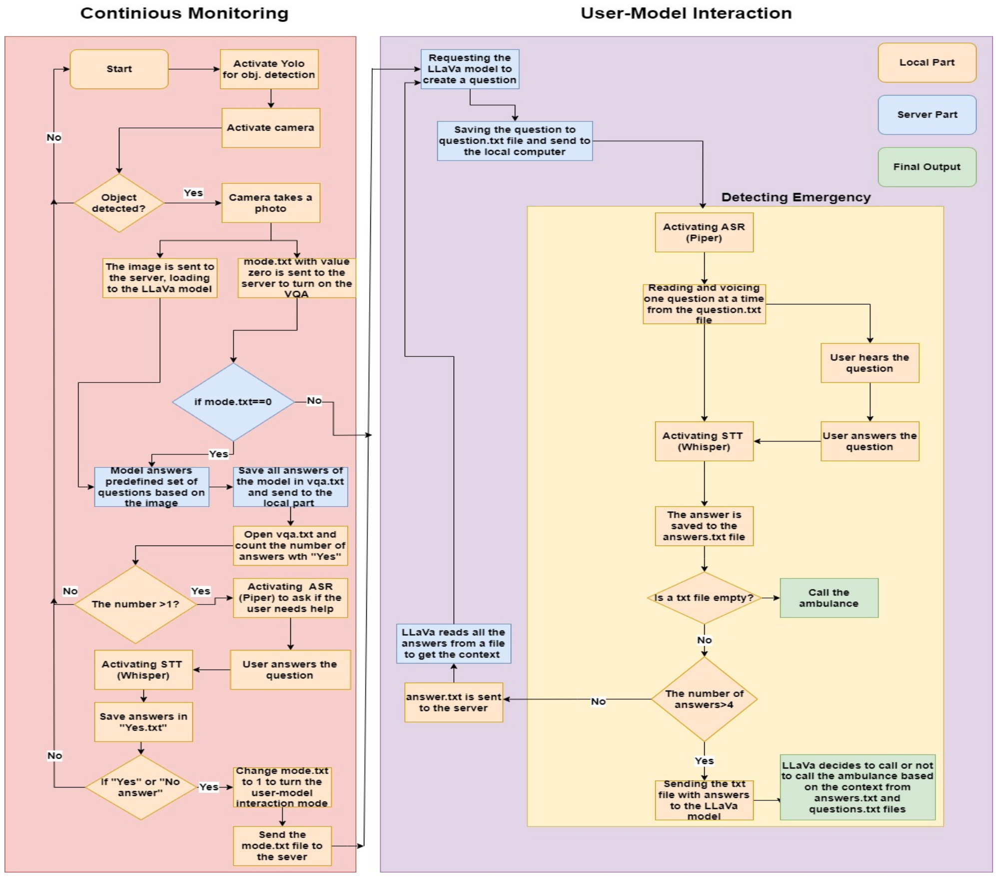


Fig. A1: State diagram of the proposed framework. On the left side is a continuous monitoring block involving human detection, VQA tasks, abnormality score calculation, thresholding, and triggering of the user-model interaction block. On the right side, the user-model interaction block involves the generation of questions, multimodal fusion, activating speech models, collecting the user's responses, and making a final decision about the severity of the incident.

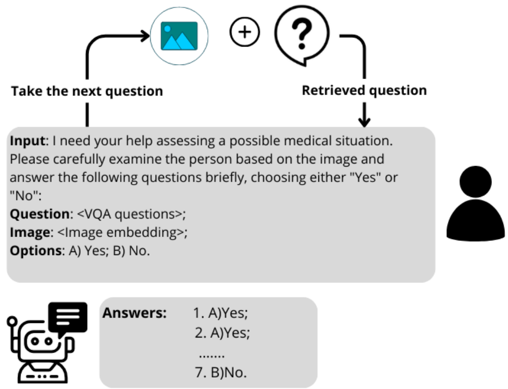


Fig. A2: Prompt instructions for the LLaVA model for the VQA task. The model was instructed to answer a predefined set of questions (1-7 questions) that included specific instructions and guidelines, adapting the LLaVA model for anomaly detection based on subtle visual clues (facial expression, gestures, protective postures, environmental anomalies).

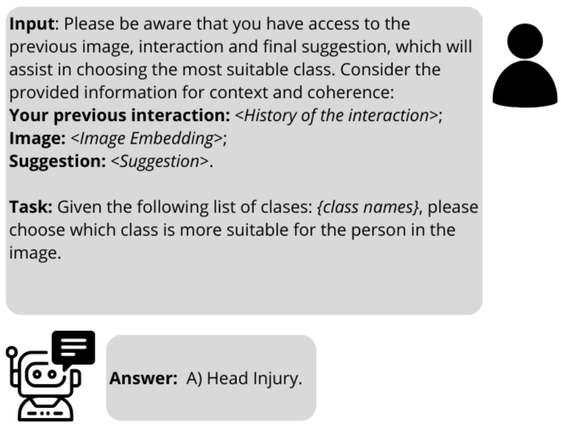


Fig. A3: Instruction of the LLaVA model for classification task. The model was instructed to classify confirmed abnormal behaviour among 8 classes (5 emergency and 3 non-emergency) based on the cumulative contextual information like image embedding, interaction history and final suggestion.

Table A1: VQA examples and initial abnormality detection

Image	Questions	Pred. VQA	True VQA	Decision
	**1.Question:** Are there any visible injuries on the person, such as bleeding, bruises, or swelling? **2.Question:** Is there any evidence that the person might have fallen, like disarranged furniture or objects? **3.Question:** Does the person lie on the ground? **4.Question:** Are there any signs that suggest the person might be in pain, like facial expressions of distress, protective postures, or vocalizations? **5.Question:** Does the person hold their chest and showing any signs of discomfort on the face or chest? **6.Question:** Does the person hold their head and showing any signs of discomfort on the face or head? **7.Question:** Does the person hold their arm and showing any signs of discomfort on the face or arm?	1. No; 2. No; 3. No; 4. Yes; 5. Yes; 6. Yes; 7. No.	1. No; 2. No; 3. No; 4. Yes; 5. Yes; 6. No; 7. No.	Abnormal
	**1.Question:** Are there any visible injuries on the person, such as bleeding, bruises, or swelling? **2.Question:** Is there any evidence that the person might have fallen, like disarranged furniture or objects? **3.Question:** Does the person lie on the ground? **4.Question:** Are there any signs that suggest the person might be in pain, like facial expressions of distress, protective postures, or vocalizations? **5.Question:** Does the person hold their chest and showing any signs of discomfort on the face or chest? **6.Question:** Does the person hold their head and showing any signs of discomfort on the face or head? **7.Question:** Does the person hold their arm and showing any signs of discomfort on the face or arm?	1. No; 2. Yes; 3. Yes; 4. Yes; 5. No; 6. No; 7. Yes.	1. No; 2. Yes; 3. Yes; 4. Yes; 5. No; 6. No; 7. Yes.	Abnormal

Table A2: Continued from previous page

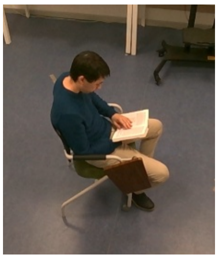
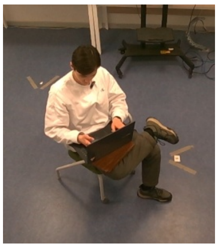
Image	Questions	Pred. VQA	True VQA	Decision
	**1.Question:** Are there any visible injuries on the person, such as bleeding, bruises, or swelling? **2.Question:** Is there any evidence that the person might have fallen, like disarranged furniture or objects? **3.Question:** Does the person lie on the ground? **4.Question:** Are there any signs that suggest the person might be in pain, like facial expressions of distress, protective postures, or vocalizations? **5.Question:** Does the person hold their chest and showing any signs of discomfort on the face or chest? **6.Question:** Does the person hold their head and showing any signs of discomfort on the face or head? **7.Question:** Does the person hold their arm and showing any signs of discomfort on the face or arm?	1. No; 2. No; 3. No; 4. No; 5. No; 6. No; 7. No.	1. No; 2. No; 3. No; 4. No; 5. No; 6. No; 7. No.	Non-abnormal
	**1.Question:** Are there any visible injuries on the person, such as bleeding, bruises, or swelling? **2.Question:** Is there any evidence that the person might have fallen, like disarranged furniture or objects? **3.Question:** Does the person lie on the ground? **4.Question:** Are there any signs that suggest the person might be in pain, like facial expressions of distress, protective postures, or vocalizations? **5.Question:** Does the person hold their chest and showing any signs of discomfort on the face or chest? **6.Question:** Does the person hold their head and showing any signs of discomfort on the face or head? **7.Question:** Does the person hold their arm and showing any signs of discomfort on the face or arm?	1. No; 2. No; 3. No; 4. No; 5. No; 6. No; 7. No.	1. No; 2. No; 3. No; 4. No; 5. No; 6. No; 7. No.	Non-abnormal

Table A3: Binary classification examples

Image	**History of Interaction**	**Ground Truth**	**Predicted**
	ASSISTANT: What are your symptoms, and how long have you been experiencing them? USER: My hand bleeds and it's for a few minutes. The bleeding is not stopping. I can feel my fingers. I'm sweating and I'm dizzy. ASSISTANT: Have you experienced any recent injuries or accidents that could be causing the bleeding in your hand? USER: Yes, yes, I scratched my hand with a nail on the wall. ASSISTANT: How severe is the pain in your hand, and have you tried any treatments to alleviate it? USER: Very severe. I don't have any medication. ASSISTANT: Have you noticed any other symptoms or changes in your hand or arm since the injury, such as swelling, discoloration, or difficulty moving? USER: Yes, I think I don't feel my fingers, I can't move them.	Emergency	Emergency
	ASSISTANT: What are your symptoms?¡/s¿ USER:I have pain in my chest, in heart region. I have shortness of breath and dizziness. ASSISTANT: Have you recently experienced any physical exertion or stress that could have contributed to your symptoms? USER: Yes, I had some physical activities today. ASSISTANT: How long have you been experiencing these symptoms, and have they worsened or improved over time? USER: I have such pain for about several months. ASSISTANT: Have you consulted a healthcare professional about your symptoms, and if so, what was their advice or diagnosis? USER: Not yet, I haven't consulted the professional.	Emergency	Emergency
	ASSISTANT: What are your symptoms? USER: My leg hurts. I can't move it. I think I'm sweating. ASSISTANT: How long have you been experiencing pain in your leg, and have you noticed any changes in your mobility or ability to perform daily activities? USER: No, everything was fine. I fell down from the chair a few minutes ago and now I'm electric. ASSISTANT: Can you describe the sensation of being "electric" in your leg? Is it a tingling, burning, or stinging sensation? USER: It is burning very hard. I can't move it. If I move it, it is really painful. ASSISTANT: Have you tried any pain management techniques or taken any medication to alleviate the burning pain in your leg? USER: No, I can't stand up.	Emergency	Emergency

Table A4: Multi-class classification examples

Image	**History of Int.**	**Suggestion**	**Ground Truth**	**Predicted**
	ASSISTANT: What are your symptoms, and how long have you been experiencing them? USER: My hand bleeds and it's for a few minutes. The bleeding is not stopping. I can feel my fingers. I'm sweating and I'm dizzy...	Based on the user's symptoms, it is important to seek immediate medical attention. The bleeding in the hand, difficulty moving the fingers, and dizziness suggest that the injury might be severe or have caused nerve damage. The user should call an ambulance or visit the emergency room to receive proper medical care and assessment. It is crucial to address these symptoms promptly to prevent further complications and ensure the user's safety and well-being.	Open Wound	Open Wound
	ASSISTANT: What are your symptoms? USER: I have pain in my chest, in heart region. I have shortness of breath and dizziness...	Based on the user's symptoms of chest pain, shortness of breath, and dizziness, it is important for them to seek immediate medical attention. These symptoms could be indicative of a serious medical condition, such as a heart attack or pulmonary embolism. The user should call an ambulance or visit the nearest emergency room as soon as possible to receive proper evaluation and treatment. It is also advisable for the user to consult a healthcare professional for further guidance and management of their symptoms.	Heart Attack	Heart Attack

Supplementary information To prove the correct operation of the proposed system, a video abstract which explains the project and shows some trailers from the experiment is provided via a short demo video along the manuscript.

Author Contributions Concept, design, system's workflow (continuous monitoring and user-model interaction stages), deployment of the models, collecting and preparing materials, hardware, and abnormal and non-abnormal scenarios in the experimental room were conducted by A.Z. and R.A. Also, the collection of the dataset for ABD, experiment execution, and quantitative and qualitative analyses were done by A.Z. and R.A. While A.V. and A.Y. came up with the idea for the project, supervised the project and provided guidance throughout the research, contributing to the conceptual framework, methodology, experimental procedure, VLM and LLM theory. Also, A.Y. and A.V. edited and approved the final version of the manuscript for submission.

Acknowledgments I would like to express my deep gratitude to my supervisors, Dr Adnan Yazici and Dr Atakan Varol, for providing guidance and feedback throughout the project. Thanks also to my teammate Rakhat Abdrakhmanov for helping me develop this project's system.

Funding This work was supported by the Institute of Smart Systems and Artificial Intelligence (ISSAI) at Nazarbayev University.

Data Availability and Materials Data is available on request from the authors. The code for the current study can be found in ISSAI's Github: https://github.com/IS2AI/Enhancing-Ambient-Assisted-Living-with-Multi-Modal-Vision-and-Language-Models.git

Declarations

Ethical Approval The experiments were authorized by the Nazarbayev University Institutional Research Ethics Committee (NU IREC), which ensured that all procedures adhered to confidentiality, safety, and ethical standards with no conflicts of interest, thereby legitimizing the results.

Consent to Participate Each participant read and signed the consent forms for participation in our experiments for data collection and system evaluation. Some participants also signed a photo/video release form to gain permission from individuals to use their images or videos in our research.

Consent to Publish N/A.

Competing Interests The authors declare no competing interests.

Open Access