Enhancing Ambient Assisted Living with Multi-Modal Vision and Language Models: A Novel Approach for Real-Time Abnormal Behavior Detection and Emergency Response

by

Adil Zhiyenbayev

Submitted to the Department of Data Science
in partial fulfillment of the requirements for the degree of

Master of Science in Data Science

at the

NAZARBAYEV UNIVERSITY

June 2024

Author .....
Department of Data Science
05 April 2024

Certified by .....
Adnan Yazici
Department Chair, School of Engineering and Digital Sciences
Thesis Supervisor

Certified by .....
Atakan Varol
Department Chair, School of Engineering and Digital Sciences
Thesis Supervisor

Accepted by .....
Yelyzaveta Arkhangelsky
Acting Dean, School of Engineering and Digital Sciences

Enhancing Ambient Assisted Living with Multi-Modal Vision and Language Models: A Novel Approach for Real-Time Abnormal Behavior Detection and Emergency Response

by

Adil Zhiyenbayev

Submitted to the Department of Data Science
on 05 April 2024, in partial fulfillment of the
requirements for the degree of
Master of Science in Data Science

Abstract

The global demographic forecast predicts a surge to over 1.9 billion individuals by 2050, escalating the demand for efficient healthcare delivery, particularly for the elderly and disabled, who frequently require caregiving due to prevalent mental and physical health issues. This demographic trend underscores the critical need for robust long-term care services and continuous monitoring systems. However, the efficacy of these solutions is often compromised by caregiver overload, financial constraints, and logistical challenges in transportation, necessitating advanced technological interventions. In response, researchers have been refining ambient assisted living (AAL) environments through the integration of human activity recognition (HAR) utilizing advanced machine learning (ML) and deep learning (DL) techniques. These methods aim to reduce emergency incidents and enhance early detection and intervention. Traditional sensor-based HAR systems, despite their utility, suffer from significant limitations, including high data variability, environmental interference, and contextual inadequacies. To address these issues, vision language models (VLMs) enhance detection accuracy by interpreting scene contexts via caption generation, visual question answering (VQA), commonsense reasoning, and action recognition. However, VLMs face challenges in real-time application scenarios due to language ambiguity and occlusions, which can degrade the detection accuracy. Large language models (LLMs) combined with text-to-speech (TTS) and speech-to-text (STT) technologies can facilitate direct communication with the individual and enable real-time interactive assessments of a situation. Integrating real-time conversational capabilities via LLM, TTS, and STT into VLM framework significantly improves the detection of abnormal behavior by leveraging a comprehensive scene understanding and direct patient feedback, thus enhancing the system's reliability. A qualitative evaluation showed high system usability results in a subjective questionnaire during real-time experiments with participants. A quantitative evaluation of the developed system demonstrated high performance, achieving detection accuracy and recall rates of

93.44% and 95%, respectively, and a specificity rate of 88.88% in various emergency scenarios before interaction. After the interaction stage, the performance was boosted to 100% accuracy due to increased context from user's responses. Furthermore, the system not only effectively identifies emergencies but also provides contextual summaries and actionable recommendations to caregivers and patients. The research introduces a multimodal framework that combines VLMs, LLMs, TTS, and STT for real-time abnormal behavior detection and assistance. This study aims to develop a comprehensive framework that overcomes traditional HAR and AAL limitations by integrating instructions-driven VLM, LLM, human detection, TTS, and STT modules to enhance emergency response efficiency in home environments. This innovative approach promises substantial advancements in the field of AAL by providing timely and context-aware detection and response in emergencies.

Thesis Supervisor: Adnan Yazici
Title: Department Chair, School of Engineering and Digital Sciences

Thesis Supervisor: Atakan Varol
Title: Department Chair, School of Engineering and Digital Sciences

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.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The population is experiencing a significant change in demographics, characterized by a rapid increase in population exceeding 8.5 billion in 2030 and 9.7 billion in 2050 [66]. The demographic change poses several challenges, especially concerning the healthcare needs of people, both physical and mental. As people age, they will likely experience health issues and risks. Also, various health-related incidents like respiratory, cardiovascular, chronic issues, mechanical injuries, and other incidents can occur occasionally, increasing the need for continuous monitoring and health management.

The most vulnerable groups who require a system for continuous monitoring and emergency prevention are the aged and disabled people. For illustration, more than one-fifth of individuals 60 years of age and older encounter mental or neurological conditions, excluding those related to cerebral pains. These conditions account for 6.6% of the general disability rate in this age demographic statistic [59]. The number of disabled people is estimated at 1.3 billion, accounting for 16% of the global population [55]. Also, disabled people face more challenges with transportation to medical services compared to those without disabilities. Hence, there is a growing need for a long-term care service that can monitor and progress the well-being of the disabled and older adults.

Apart from vulnerable groups, there is a large number of healthy people in home incidents caused by physical or mental damage [1]. The most common accidents are

head injuries, open wounds, poisoning, and falls. According to a retrospective study among children and adolescents for home incidents and emergency cases, the leading cause of emergency accidents are due to falls (53.7%), showing a high prevalence of unintentional home incidents. [18] shows that common causes of death are poisoning (82600 cases in 2021) and falls (29100 cases in 2021) in a home environment, highlighting the fact that injuries or emergency accidents are occurring more often at home compared to emergency cases in public places. Therefore, there is a significant need for early emergency detection to cope with home incidents and prevent the lethal outcomes.

However, current assistive care services have a few drawbacks. First, it is an increased load on high-quality caregivers and conventional long-term services, which can pose financial challenges for individuals needing constant supervision and medical help [69]. Secondly, people under strict hospital supervision may experience psychological stress and isolation in an institutional setting, resulting in health issues and a decay in mental and emotional well-being [11]. Hence, there is a need for automated assistive technology that can improve a sense of independence and reduce feelings of isolation and home emergency incidents. Researchers propose various ambient assisted living (AAL) setups designed for the long-term monitoring and timely detection of emergency cases [11, 37, 9, 61]. AAL systems include patient monitoring via various sensors and human activity recognition (HAR) with machine learning (ML) or deep learning (DL) techniques.

Current approaches can detect abnormal behavior in people by using wearable or non-intrusive sensors [37, 9, 61, 3, 34, 7]. Mostly, the ML and DL approaches try to extract knowledge from the sensors to relate the data and the home context to detect abnormal or normal activity of the patient [3]. For example, if a person falls or holds the chest, then these actions can be tracked as abnormal action that requires calling the ambulance. Anton et al. collected the raw sequence of data from various sensors, tracked the patterns (common sequences of actions performed at specific timestamps) of people's activities and utilized the data from sensors as input for ML models to make them learn the normal and abnormal patterns of actions [3]. How-

ever, sensor-based methods have several drawbacks. Firstly, the person may perform daily activities in different sequences or with varying intensity, making it challenging for ML and DL models to establish consistent patterns for anomaly detection. In consequence, large fluctuations can appear in the order of activities. The other factor is that building robust ML and DL models requires a large volume of labelled training data, which is time-consuming to collect. Also, acquiring extensive datasets for anomaly detection in home environments, such as sensitive activities like falls or medical emergencies, can be difficult because of ethical considerations, privacy concerns, and the practical challenges of monitoring a diverse range of activities. Thirdly, sensor data alone may lack sufficient context to accurately distinguish between normal and abnormal activities. ML and DL models can struggle to interpret sensor readings in complex situations and differentiate between simple actions and complex emergency scenarios. Another drawback is that interference with the environment can cause noise in data due to lighting conditions, movements of other people and occlusions from the location of a camera. This can lead to false alarms and loss of scene details, reducing the accuracy of the models in the detection and classification processes. Lastly, ML and DL approaches, which use sensor data as input to infer knowledge and common patterns, require continuous monitoring and adaptation to the specific environment and person. However, implementing such conditions can be complex and resource-intensive.

In general, the fluctuating order of the activities, the small volume of training data, the limited context, interference with environmental factors, adaptation, and learning challenges are the main disadvantages of sensor-based systems with conventional ML and DL models [54], [4], [76]. Therefore, finding a solution capable of overcoming the deficiencies of traditional approaches is vital, especially in handling context-based limits.

To cover the limitations of sensor-based solutions for HAR and anomaly detection, vision language models (VLM) can be utilized because these models understand the context of the scene and perform many visual recognition tasks such as image classification, object detection, semantic segmentation, image captioning, and visual-

question answering (VQA). After pretraining with massive image-text pairs, VLM models can conduct autoregressive tasks like generating captions or descriptions for the images and answering questions related to the image [45]. To illustrate, VLM is used in medicine, where medical images (chest radiographs) and radiology reports were used to train the model to classify the disease on the image, achieving good results in a zero-shot classification (ability to classify the samples that were not used in the training process) for several datasets [77]. Another notable application of VLM is exemplified by MedFuseNet [70] and PathVQA [53], specifically used to tackle and provide insightful responses to visual inquiries in the medical domain. These sophisticated systems apply advanced image processing and linguistic comprehension, enhancing a deeper understanding of medical data. That is, VLMs can comprehend both the visual scene and its contextual nuances in the scope of medicine. VLMs introduce a new era in medical diagnostics and detection through their ability to recognize intricate details within the medical field, including the patients and their environment, answer queries, generate valuable context from the images and find the anomalies in the pictures.

Nevertheless, VLMs have not yet found their way into real-time applications for assisting people in emergency scenarios because of the difficulties in the ambiguity of language interpretation and, most importantly, uncertainty due to data quality [73]. As a result, the number of false alarms or erroneous responses can increase, causing frustration or confusion for a user. For instance, the image or video taken by the camera can be corrupted due to changes in temperature, lightness, atmosphere conditions, or the model is hallucinated due to limited context, poor data quality, and complexity of language understanding. As a result, the model cannot see and comprehend what is depicted in the video or image, and the model incorrectly detects abnormal or normal behavior, generating erroneous responses [73]. Thus, there is a need for additional models to communicate with a person in real-time to validate the results inferred from VLM and collect additional context, decreasing the rate of false positives or false negatives.

As a solution to the problems related to VLM, large language models (LLM)

can be used as a validation step after VLM that can understand, generate human language and interact. Also, LLMs can interpret the context, generating coherent and accurate responses. However, it is essential to consider that only LLMs with well-designed prompts can effectively validate the results from VLMs [73]. Good prompting is crucial for eliminating uncertainty inherent in VLM's outputs.

In this setup, a large language and visual assistant (LLaVA) is a VLM that integrates both visual assistant and LLM that was pretrained on vast amount of datasets. The LLaVA model shows excellent zero-shot (no additional training) capabilities in comprehending, describing and analyzing various visual aspects. This gives ability to recognize actions and visual cues that were not used in the training process. Thus, the LLaVA model is capable to interpret the sings of normal and abnormal behaviour of a person.

The visual assistant from LLaVA processes the visual data to align them with textual data, creating a multimodal representation. This representation is fed to the LLaVA's LLM to generate the response. For instance, LLM can answer questions related to abnormal cases based on the image context. Also, The LLM from LLaVA can be used to assess the correctness of the initial VLM response by conversing with a person where the explicit context of a situation can be retrieved. This interaction is achievable through additional text-to-speech (TTS) and speech-to-text (STT) models, enabling seamless human-computer interaction (HCI) [65], [20]. Through TTS and STT, a person can effectively communicate with LLM, describing their concerns and symptoms as a valuable context, which is then used as input for LLM to validate the results of the VLM and the person's responses. As a result, it can decrease the number of false positives and false negatives in classifying the action as emergency or non-emergency behaviour. Overall, a framework consisting of several models like VLM, LLM, STT, and TTS can give promising results on abnormal behaviour detection and assistance for people, achieving state-of-the-art (SOTA) accuracy in detecting anomalies in actions.

1.1 Motivation

The motivation of the work is to bridge the gap between technological advancements and the pressing healthcare needs of the global population via multi-modal system that leverages visual, textual and audio data modalities together. Also, the motivation is to promote a future where people can live independently and securely within their familiar environments and provide instant assistance in emergencies.

This thesis introduces an innovative framework combining VLM and LLM as a context generator, decision-maker, and validation tool. TTS and STT models stand out as the tools for real-time communication between the primary model and the person. The thesis's objectives are the following:

- Develop and implement a multimodal framework based on VLM, LLM, TTS, and STT technologies. This requires seamlessly integrating these components to enable effective interaction between people and the system, synchronising various processes and models, and providing instructions to direct the VLM and LLM for detecting anomalous patterns from visual data.

- Conduct real-time experiments involving interactions between the multimodal framework and the person. These experiments will assess the system's usability, effectiveness, and user experience in various scenarios, including emergencies and day-to-day activities.

- Evaluate the framework's capabilities in VQA, generating contextually relevant questions and suggestions in a zero-shot manner. This involves testing the system's ability to generate accurate, contextually relevant descriptions and responses based on visual inputs and instructions.

- Evaluate the framework's performance in detecting abnormal behavior of people. The assessment will involve qualitative and quantitative analysis of the system's ability to identify the emergency case and differentiate between normal and abnormal actions. Also, the provision of contextually relevant suggestions to the caregiver is conducted.

A main contribution of this work is the novelty in the approach for abnormal behavior detection and AAL, utilizing advanced language and image processing techniques enabled by VLMs. The system provides real-time validation and assistance through chatbot LLM, integrated with TTS and STT. The system ensures intuitive communication and users' practical expression of concerns. Real-time experiments evaluate usability, effectiveness, and user experience, contributing to the advancement of AAL technology by promoting independence, security, and immediate assistance in emergencies. This approach shows a novel contribution to AAL, HAR, and HCI, combining multiple modalities (VLM, TTS, STT, human detection model) and technologies in a unified pipeline for abnormal behavior detection.

LLaVA's extensive pretraining on diverse datasets allows it to comprehend various health-related concepts, human behaviors, expressions, and emergency protocols. This knowledge, derived from datasets with general human activities, gives the model to handle tasks involving assessing a person's physical and emotional states. By integrating prompt engineering directly into LLaVA's code, the model efficiently transfers its general capabilities to focus on emergency detection. Specifically, prompts guide LLaVA in detecting early signs of emergencies, such as protective postures or facial expressions indicative of distress or pain. Since the LLaVA model understands the protective postures, different emotions, various injuries, and hazardous objects, it can look specifically for these signs in the image to detect the emergency at the first stage. In the second stage, structured dialogue templates were given to the model to ensure that interactions with users remain focused on extracting crucial information for accurate emergency assessment. The model also analyzes historical interaction data and visual cues to evaluate the situation's severity and determine the necessity for medical intervention after interaction with a user. These prompts serve as a strategic guide that enhances LLaVA's ability to apply its broad knowledge to specific, emergency-related tasks. This ensures effective performance even in scenarios for which it was not explicitly trained. Finally, the multimodal fusion technique is used during the interaction and after interaction. During the interaction, the LLaVA model is instructed to generate new contextually relevant questions by

analyzing fused data (image embedding, textual responses of a user and generated previous questions). This fusion aims to generate contextually relevant questions to guide user-model interaction. The focus is maintaining context and ensuring new questions are unique and non-repetitive. After the interaction with a user, the image embedding and textual embedding from the history of the interaction are combined as input and valid context. The LLM takes this context (image embedding and outputs from the STT model) to evaluate the severity of a situation and make a decision on whether to send alerts to caregivers and call an ambulance. This fusion aims to combine visual and audio that were converted to textual information for decision-making. It integrates data from previous stages to assess the situation and make a final determination on emergency response.

Overall, the developed system for abnormal behavior detection is designed as a supportive solution for rapid emergency response and prevention. It automatically alerts emergency services and provides them with the situational context, making it an essential tool for the general population. The proposed system pipeline can be integrated into any home environment, ensuring safety for the elderly, individuals with disabilities, and healthy individuals.

The outline of the paper consists of several sections. Section 2 includes information about the literature review on non-invasive and intrusive sensors for HAR for people, vision-based sensors, combined sensors, VLMs, and LLMs for HAR. Following this, section 3 gives an overview of the proposed method, including detailed information about the modalities, the system's workflow, the datasets used in the development of the LLaVA model, and the hardware and software modules used in this work. Next, section 4 introduces the experiment procedure and evaluation of the work with qualitative and quantitative analysis. Finally, the last section 5 concludes the paper, refreshing the main findings, proposed solution, and future implications and limitations of the work.

Chapter 2

Related works

This section includes information about existing methods for general HAR and abnormal behaviour detection. The main methods, results, and limitations of each type of approach are also discussed in each subsection.

2.1 Non-Intrusive Sensor-Based Solutions For HAR

Non-wearable sensors are gaining popularity in HAR since this technology does not require attaching sensors to the body and allows them to be installed anywhere in the room. This type of installation allows for the creation of a network of these devices that can record data in various formats, enabling communication among sensors and servers. Usually, radio-frequency fluctuations created by people are collected by different sensor types, allowing for the extraction of meaningful information and features (spatial and temporal features) to classify the actions.

The authors from [3] utilize various sensors to measure carbon dioxide levels, humidity, motion, and temperature to analyze the raw data from sensors and get valuable information. ML methods like the random forest (RF) algorithm are used to predict room occupancy in different time frames. So, if the person is not in the room or stays there for too long, the action is classified as abnormal or deviated from normal. The research demonstrates that these models can effectively identify regular behavioural patterns and establish efficient management and decision-making guide-

lines. However, some limitations include the relatively small sample size, dependence on specific context, and transferability issues. Similarly, Ghayvat et al. suggest using the same sensors, including force, electronic appliance usage, and push buttons, to extract behavioural patterns like timestamps and duration to classify daily activities [23]. Another approach [24] uses ready and available dataset ARUBA [21], which is split into 10 classes, which are eating, bathroom visits, relaxation, meal preparation, sleep, work, cleaning, dishwashing, entering the house, and leaving the house. The collected data from sensors is transformed into images to be used to train deep convolutional neural network (DCNN) models. The outcomes of this approach reveal an F1 score of 0.79 for 10 classes and 0.951 for 8 activities.

[57] demonstrates the framework for emergency case detection and response using activity tracking and flood and fire sensors. The system collects data on a person's daily routine and the time between the last activities to determine behavioral patterns and identify an emergency case. To gain more context on the situation, the authors include a disaster information module (flood and fire sensors) that tracks the occurrence of the disaster in the city or apartment. In an emergency, the disaster information model triggers home sensors to understand the person's condition, compute the risks of danger, and inform the ambulance about the emergency. [2] presents a pilot deployment of an AAL sensor type system in a real house setting to detect health issues at an early stage, using indoor sensors (motion, contact, and bed sensors) and outdoor sensors (beacon sensors placed in specific places in the city).

The non-intrusive sensors effectively detect abnormal changes in the behavior of the patients. However, there is still a limitation in contextual understanding of the scene since the sensors do not provide a direct picture of what is happening. Following this, non-intrusive sensors are limited in scalability and generalizability as the deployments are based on a limited number of participants and a specific geographical area (a room layout). Another problem lies in the data size and availability. The data from sensors are raw and need to be pre-processed before going to the model to learn normal and abnormal patterns in human actions. Also, the ML and DL models need to be trained on large amounts of data from unobtrusive sensors, which are scarce

and time-consuming to collect and gain the knowledge to detect anomalies. Finally, the system with non-intrusive sensors faces challenges related to high noise in the data, potential sensor malfunctions, and a relatively small sample size, which could impact its accuracy and reliability.

2.2 Intrusive Sensor-Based Solutions For HAR

Wearable sensors are popular approaches for HAR as they can monitor and understand human behaviors in various contexts. These small gadgets, which come with various sensors like heart rate monitors, accelerometers, gyroscopes, and GPS trackers, have become indispensable for monitoring and evaluating physical activity.

Smartphones are used in [13] as wearable sensors to track the movements of stroke patients and healthy persons. Smartphones are fastened to participants' waists, and accelerometer and gyroscope data are gathered to determine the tasks being performed. This study's findings suggest that cell phones can be used in HAR.

Other studies [52, 72, 60] solely rely on wearable sensors. [72] utilizes a wearable system for anomaly detection related to heart, electrocardiography (ECG), photoplethysmograph (PPG), and phonocardiogram (PCG). [60] proposes a system focused on heart anomalies like arrhythmia and atrial fibrillation, using a support vector machine (SVM) for classification. [52] employs inertial sensors for fall detection and heart-related anomalies. While the number of end-to-end systems for health-related anomaly detection is limited, studies like [72] and [60] demonstrate alternative approaches using wearable sensors like ECG, PPG, and PCG to detect heart, breathing, and apnea abnormalities.

Santiago et al. utilize a pendant designed for activity recognition, explicitly focusing on detecting falls [67]. The pendant is responsible for sending the data and notifications about the fall to the smartphone. Communication is done via Bluetooth, which makes life easier for older adults because they do not have to keep their phones with them all the time. [6] utilizes the available ADAPT dataset [10], which includes various classes of activities such as sitting, standing, walking, transitions, shuffling,

leaning, lying, ascending stairs, and descending stairs. SVM is employed with wearable sensors for activity recognition. The sensors are set on the chest, wrist, lower back, and thigh. The performance of their two-sensor system (thigh and lower back) achieves an F1 score of 87.2% and an accuracy of 88%. However, it is essential to note some limitations of this study, including an unbalanced dataset, a high number of false positives for the lying class, and a relatively small amount of training data. These limitations can impact the generalizability and accuracy of the model in real-world scenarios.

Overall, all the methods with intrusive sensors have common disadvantages. One of them is wearable sensors that are designed for specific use cases, and adapting them to new activities or scenarios can require significant training, customization, and time. Even though wearable devices provide good contextual information, interference from other devices, like electromagnetic interference or signal obstruction, can affect the sensitivity and accuracy of the sensors. Also, a person can forget or refuse to attach the sensors to the body, as the wearable sensors might be inconvenient.

2.3 Vision-Based Sensors For HAR

Vision-based methods for HAR are based on different types of cameras that take pictures or record the video of an activity to extract 3D key joints for DL models. [30] implements a multi-view camera to capture 3D joint keypoints to be used in DL techniques. In this study, video recordings of individuals with healthy gait (HOA), multiple sclerosis (MS), and Parkinson's disease (PD) are used to develop a contactless, low-cost, and highly accurate remote monitoring tool for neurological gait classification. However, this approach has certain limitations, including a relatively small sample size with gender disparities, susceptibility to noise and variations in lighting conditions, and constraint to a specific number of classes without the opportunity to infer new ones in real-time. Additionally, this approach primarily relies on patient joint key points as training data while overlooking background context information. [34] utilizes the depth data of 9 daily activities. The activities include

walking, eating-drinking, exercising, cooking, sitting down, standing up, falling back, watching TV, and lying down. They used a depth camera to extract human figures from other objects and the background, generating the body skeleton features. These features are then preprocessed by linear discriminant analysis (LDA) and trained with a hidden Markov model (HMM). Nevertheless, HMM models are not the best options for abnormal behaviour detection since they struggle with learning complex data patterns (changing patterns of activities). Also, the model does not consider the background information, resulting in a limited context of the situation. Similarly, [8] proposes an RGB-D camera-based approach to extract the skeleton data from images. They utilized linear discriminant classifier (LDC), long short-term memory networks (LSTM), and recurrent neural networks (RNN) for activity identification. The authors developed a model capable of identifying a wide range of human activities with a high accuracy of 92.83%, showing its potential utility in applications such as surveillance, healthcare monitoring, and HCI. [64] focuses on DL transformer with attention to activity recognition and classification, using human skeleton estimation to represent the pose of the human body. This paper highlights the effectiveness of synthetic data generated by generative adversarial networks (GANs) in enhancing the training dataset, particularly for underrepresented classes. As a result, they substantially improved model performance, achieving remarkable results (accuracy of 99.50%, and a precision of 88.96%).

2.4 Combined Sensor

These methods are based on integrating intrusive, non-intrusive, and vision sensors. These sensors create a dynamic synergy that surpasses the constraints of each technique separately by combining the benefits of conventional, contact-based sensing with the ease and adaptability of contactless and remote sensing.

Fan et al. utilize both intrusive and non-intrusive sensors [22]. A wide range of sensors, including motion detectors, door sensors (including door locks), humidity and temperature sensors, light sensors, water dispenser monitors, air conditioning sensors,

water leak detectors, and pressure sensors, are installed in the homes of the elderly volunteers for the experiments. The elderly also wear smartwatches, which record their heart rates, sleep habits, and number of steps taken. To enable prompt alerts and communication, the acquired data, which includes the patient's daily activities and health metrics, is routinely shared with their families and medical providers. In [58], a hybrid framework for abnormal behaviour detection is introduced, including activity and behaviour monitoring, personalization, and emergency services through methods like feedback-based human behaviour modelling, pattern mining, clustering for anomaly detection, and employing DL and ML for recognizing activities. The framework was validated with wearable sensors and camera systems across multiple activities, showing a high accuracy range of 82-92% in detecting simple (walking, sitting, sleeping, standing, and running) and complex behaviours (eating, exercising, and going to the toilet). A combination of visual and intrusive sensors detect abnormal behaviour through unsupervised learning with a specific threshold. Even though the framework shows high results, the sample size is limited. As a consequence, a low sample size can affect the system's generalizability and applicability to a broader population. Also, the system's adaptability to sudden changes in behaviour or lifestyle and its ability to learn from new patterns over time without manual intervention is not explicitly discussed. [17] combines non-intrusive and vision sensors together, creating a comprehensive system to detect a person's abnormal behaviour. These sensors collect visual data (3D coordinates) and data from various environmental sources. Also, this method integrates real-time 3D representation via a game engine to get the context of the environment and the person, and it enables vocal interactions between the user and the system via TTS and STT to gather additional information about the state of the person for better responsiveness of the system. This framework achieves improved context awareness through real-time 3D representation, continuous conversation with the user for additional context, and improved reasoning abilities. However, this system is still limited to a certain number of classes, resulting in a possible miss in coverage of unseen emergencies.

2.5 Vision-Language Models and Large Language Models for HAR

Recently, a new DL paradigm has appeared: VLM pretraining with zero-shot and few-shot prediction. VLMs are pretrained via image-text pairs from the internet or other datasets in large amounts, tracking and learning vision-language correspondence knowledge [82]. This enables few-shot (training with a few samples) and zero-shot (predictions without additional training) predictions on unseen data without additional fine-tuning to specific tasks. DCNN-based and transformer-based network architectures are the two primary types used by VLMs to extract image features. DCNN architectures use pretrained DL models to perform classification tasks for image feature learning. At the same time, transformer-based VLMs successfully generate textual descriptions for images and understand the relationships between textual and visual content [82]. VLMs enhance traditional classification tasks, as seen in previous sensor-based methods, by enabling the generation of textual descriptions (not just labels) from images, facilitating dynamic interactions between image and text, and identifying a broader range of classes or unseen ones.

One of the most promising VLMs is contrastive language-image pretraining (CLIP). CLIP learns a multimodal embedding space by simultaneously training text and image encoders to maximize the similarity of embeddings for real pairs and minimize the similarity of incorrect pairings. This VLM uses ResNet-50 for image encoding, while transformer architecture is used for text encoding, aligning visual and textual information effectively. CLIP demonstrates tremendous results for zero-shot prediction on the publicly available dataset ImageNet [19] with 14 million images of different objects and categories [62], achieving 76.2% accuracy comparable to the performance of the original ResNet-50 model. However, there are some limitations regarding systematic tasks like counting objects in an image and fine-grained classification tasks.

[45] uses Prismer VLM, which implements ensemble learning with domain experts to encode the image, enabling cost-efficiency and high performance. Prismer succeeds in image captioning and VQA tasks. The main benefit of Prismer VLM is data ef-

ficiency in terms of training since it already uses many pretrained models that help reduce the number of trainable parameters and computational complexity. Prismer achieves SOTA results in image captioning for a few-shot and zero-shot performance with fewer training data. Also, this model reached a high accuracy of almost 80% for the ImageNet dataset. Wu et al. introduce BIKE VLM, which achieves high accuracy results (86.7% and 94.7%) on the action recognition dataset Kinetics-400 [31] and ActivityNet [12], using CLIP [62] as the backbone model. The high results are reached with the integration of video-attribute-association mechanisms to extract auxiliary information from the videos and video concept spotting to identify important moments or concepts within a video, taking into account the temporal aspect of the video [78]. Both mechanisms in BIKE improve video recognition and representation. Nevertheless, BIKE requires lots of computational resources to be trained and finetuned to specific tasks.

Apart from general HAR, VLMs can be used in medicine, with some works like MedFuseNet [70] and PathVQA [53] for addressing visual questions within the medical field. Also, LLMs are being adopted in medicine as a chatbot for commonsense reasoning, understanding, assisting, and user interaction tasks via the rapid generation of responses to the questions based on the queried text prompts and images. This suggests that VLMs and LLMs can be used for health diagnostics, question answering, and user interaction within the healthcare domain. [48] proposes Video-ChatGPT that integrates a pretrained visual encoder and LLM, enabling active conversation between the user and chatbot based on the provided video or picture. Video-ChatGPT uses a CLIP visual encoder to get spatial and temporal features from the frames and LLM Vicuna [43] to initialize conversation and produce meaningful responses based on the question and image.

Most LLMs and VLMs were trained in a self-supervised manner (on vast datasets without explicit labels), resulting in increased generalization across diverse scenarios and adaptation to various environments. One of the examples is CLIP VLM, which was trained with image-text pairs and contrastive learning (maximizing the similarity between images and their corresponding captions while minimizing the similarity

between unrelated image-text pairs), allowing the model to understand and relate visual and textual information without explicit labels. Consequently, CLIP can generalize across different scenarios, making it adaptable to environmental changes and versatile in different tasks. This adaptability is especially valuable in applications like abnormal behaviour detection, where context and multimodal information are critical.

Moreover, working with large-scale datasets in a self-supervised way allows CLIP and similar VLM models to continuously learn and adapt to new cases or tasks, reducing the need for manual labelling and re-training. This efficiency enhances scalability and makes VLM and LLM models suitable for real-time applications, as they can process and understand data quickly.

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 are trained on large datasets 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 caption generation, image search, retrieval, visual commonsense reasoning, and VQA related to images in order to enhance context understanding and detect anomalies in the scene. Also, the integration of LLMs enables real-time communication (multimodal chat-bot interaction with the user) with the user to gain additional context and make a decision based on the provided context.

Chapter 3

Methodology

In this section, the leading solution is a system that combines VLM, LLM, human detection model, TTS, and STT models for abnormal behavior detection and HAR. The unique attributes and capabilities of the models, such as generalization capabilities, zero-shot capabilities, improved detection, the refinement of emergency detection, and reasoning through instruction tuning, are discussed.

3.1 Large Language-And-Vision Assistant Model

The LLaVA-1.5 model is used as the backbone model in HAR and emergency detection, encompassing VLM and LLM capabilities together. The model is combined with visual encoder CLIP-ViT-L-336px [62] and LLM Vicuna-1.5 [15], which are used in the work [43].

According to the figure 3-1, the primary model connects the visual features with the textual embedding space to align them and make the model learn this paired embedding in a self-supervised way. The model training is done in a two-stage process. In the first stage, a projection matrix (multi-linear layer) is meticulously crafted to bridge the gap between visual features extracted by CLIP and textual embedding from a pre-trained LLM Vicuna. Both the visual encoder and language model are frozen except for the projection matrix, in which weights are updated based on image-text pairs. This is done to make the projection layer learn how to align visual features

(vectors) with textual embeddings. After that, a projection layer is utilized to map image features to language tokens, which then serve as the input for the language model [43]. In the second stage, the LLM and the projection matrix are finetuned on a massive dataset of the instructional dataset with diverse instructions such as answering the questions, describing the objects, actions and reasoning from images in which the original descriptions/captions of the pictures are the answers or ground truth [36, 44, 43]. The model continues to learn the mapping between visual features and words while keeping the image encoder fixed. The model is exposed to different instructions related to visual content, the model becomes a powerful chatbot, learning to talk about pictures, follow user instructions and respond accordingly. Also, the model is finetuned for multimodal science questions, in which the ground truths are detailed explanations and lectures. The model is exposed to multiple-choice questions, which helps it understand, reason, and answer science-related questions.

The LLaVA model exists in various versions, each distinguished by the number of parameters and the extent to which they have been finetuned for specific tasks and domains [43]. Among these, the general LLaVA-1.5 version with 13 billion parameters stands out as the primary model with VLM and LLM capabilities in this setup for HAR and emergency detection. The model's extensive training on vast datasets comprising general images and text provides a robust foundation for comprehending visual concepts, object and human relationships, and human actions [43]. This knowledge is invaluable for scene understanding and identifying unusual postures, situations, environments, and emotions within an AAL system. For instance, it can infer a fall from visual cues like someone lying on the floor with a pain expression, even without visible injuries.

Furthermore, LLaVA-1.5 with 13 billion parameters version allows for adaptation to many more situations, ranging from simple to complex, compared to models specifically trained on simple activities of daily living (ADL) or finetuned on the limited data for emergency detection. Given that everyday activities may occasionally resemble emergencies solely from a visual perspective, LLaVA-1.5's adeptness at contextual understanding and reasoning derived from its extensive general pretraining

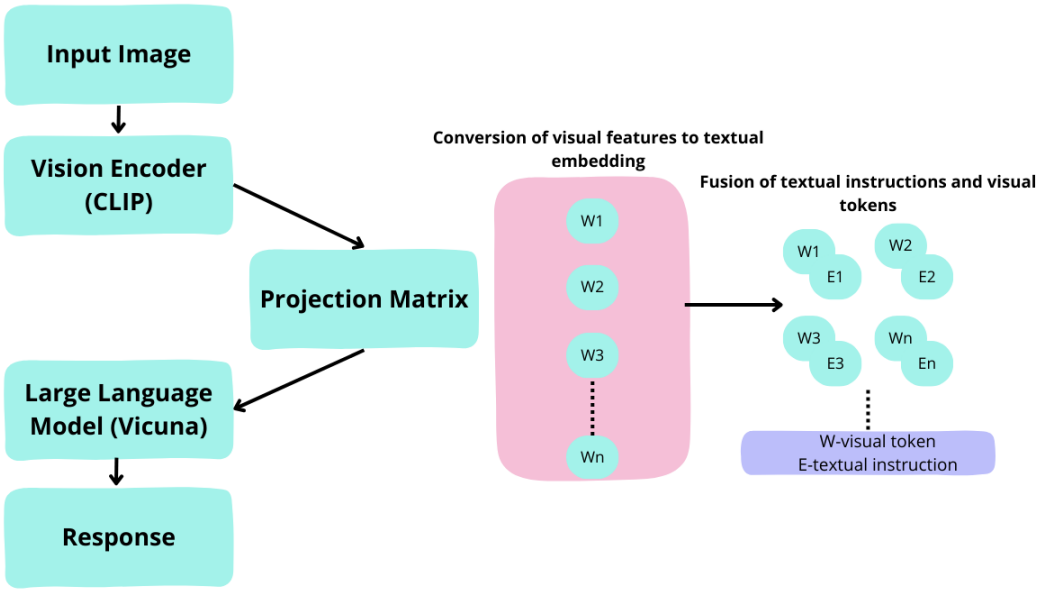

Figure 3-1: The figure depicts the architecture of LLaVA, including LLM Vicuna, vision transformer CLIP, conversion of visual features to textual tokens and fusion with textual instructions. All the fused information is given as input for the Vicuna LLM to generate responses based on both the visual context and user-provided textual instruction. Ultimately, the model's responses can be answers to questions (VQA), generation of context-related questions, or generation of captions and decision-making.

data proves its ability to discern between routine activities and emergencies. For example, the person holding a chest can be tracked as a possible heart attack scenario at first sight, but the person also smiles when looking at a person's facial expression. LLaVA can analyze the situation and identify it as non-emergency by looking at visual features and contextual information (gestures, facial expressions, and environmental settings). The LLaVA model then concludes that the person is holding their chest in a context of apparent happiness, which is not indicative of emergency.

Apart from the adaptability and generalization abilities of the model, the general version of LLaVA with 13 billion parameters demonstrated superior performance compared to SOTA VLMs such as BLIP-2, InstructBLIP, Qwen-VL-Chat on 11 out of 12 evaluated academic VQA benchmarks, using a smaller number of parameters but with better visual instruction tuning (a process that adapts LLMs to describe, reason and relate to visual information in addition to textual data). Benchmarks

include VQA-v2 [25], GQA [29], VisWiz [26], ScienceQA-IMG [47], TextVQA [79], POPE [41], MME [80], MMBench [46], MMBench-Chinese [46], SEED-Bench [39], LLaVA-Bench (In-the-Wild) [44], MM-Vet [81]. That is improved performance on tasks that require understanding, reasoning, conversation, detailed description, and following textual instructions related to visual data. Also, the smaller version of LLaVA with 7 billion parameters has shown less efficiency than the larger version on the datasets mentioned above [**?**]. Further, LLaVA was compared with GPT-4 across tasks in conversation, detailed description, and complex reasoning, attaining 85.1% of the relative performance score achieved by GPT-4, showing almost similar performance of SOTA model GPT-4 [44]. Furthermore, LLaVA outperformed GPT-4, achieving 90.92% accuracy for the Science QA dataset. Therefore, LLaVA, as an alternative to GPT-4, is utilized as the primary model for its strengths in understanding, reasoning, and conversational abilities. Also, its superior performance across multiple VQA benchmarks suggests an advanced capability to handle diverse, context-rich tasks, essential in our work involving deep scene and context analysis and reasoning from subtle visual cues to detect abnormalities in visual scenes.

3.2 Instruction Following Ability for Emergency Detection

VLMs sometimes produce unpredictable results due to their complex structure, which includes billions of parameters [73]. These models predict the subsequent word in a sequence by analyzing the statistical connections between words, leading to a degree of unpredictability in their responses. Also, achieving high performance on tasks for which the model has not been trained is challenging. Prompt engineering helps to reduce the randomness and inconsistencies in answers. LLaVA emerges as an exceptional choice due to its self-supervised nature, superior generalization abilities, enhanced instruction-following abilities (model's capacity to perform various visual tasks like recognizing activities, deducing the sequence of events from images and

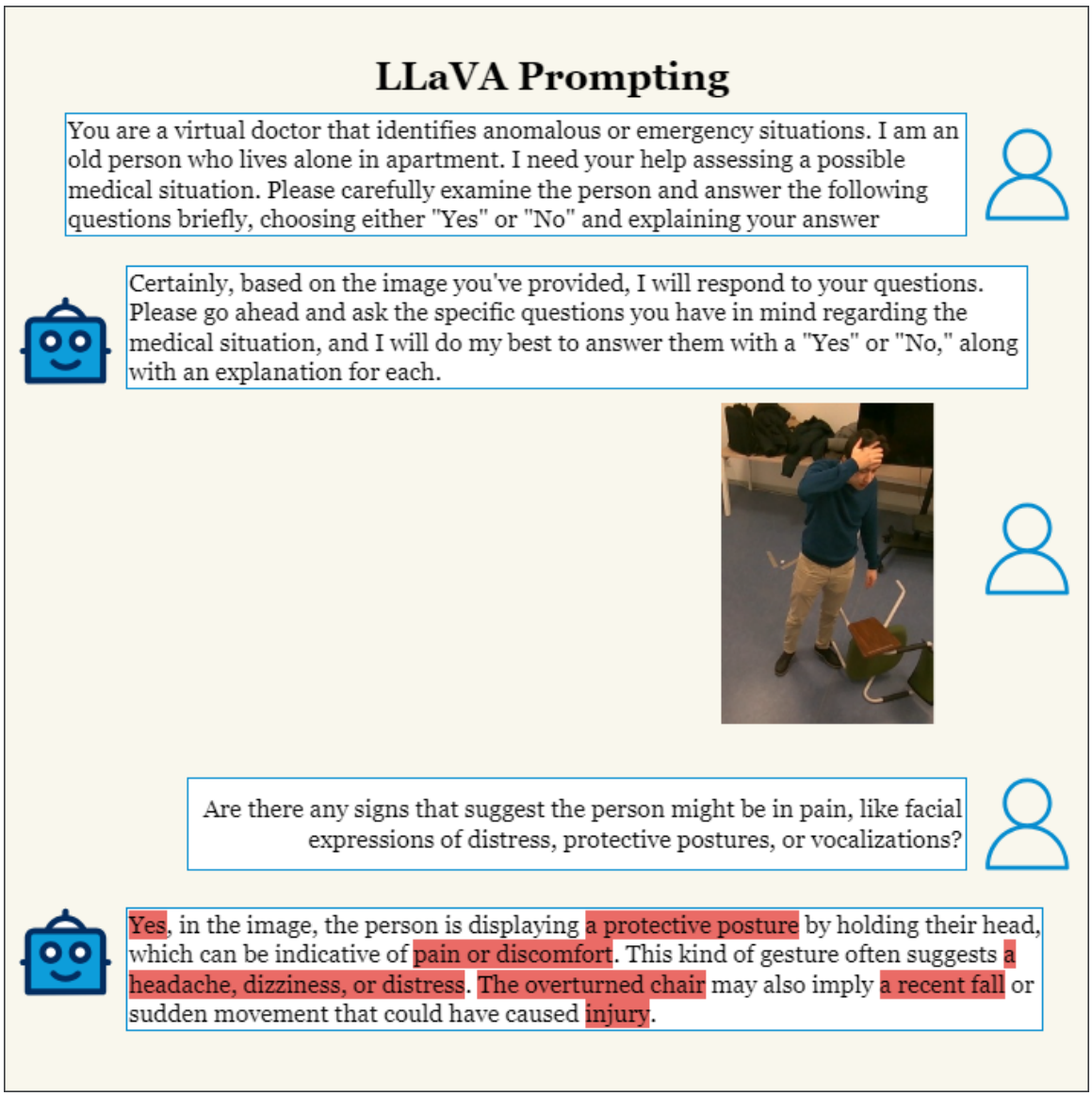

Figure 3-2: The figure depicts the prompt and instruction tuning for the LLaVA to determine abnormal behaviour and respond accordingly to user requests.

others based on the textual queries or instructions), and deep understanding of various contexts, concepts, image associations with textual descriptions, underpinned by its training on diverse datasets and visual instructional dataset. The empirical studies in [73] proved the ability of the LLaVA model to generalize to new classes and cases, excelling in analyzing and understanding medical images or tasks such as disease diagnosis without the model's fine-tuning or additional training. For instance, LLaVA

showed promising results in analyzing medical images like MRI and chest X-rays, even though they were not specifically trained on these medical datasets. LLaVA achieves great results in generalizing to unseen data due to its diverse set of multimodal image-instruction pairs from general sources, which cover a wide range of subjects and scenarios. Also, by incorporating the principles of transfer learning (knowledge in one field is applied to another domain), LLaVA can apply the visual and linguistic knowledge gained from training on general image-text pairs to the complex domain of medical imaging analysis. In total, its adeptness at generalizing allows it to excel in tasks beyond its training scope. At the same time, its receptiveness to specific prompt engineering reduces the effort needed to infer meaningful answers.

LLaVA's instruction-following feature is the main advantage in emergency detection. LLaVA can be tuned or instructed to look for specific visual and textual inputs that might indicate abnormality via prompt engineering that outlines specific behaviours or signs of emergencies or abnormal activities [73]. This way, using its existing knowledge, the model can look for specific patterns and concepts and understand what actions are anomalous and non-anomalous. For instance, since the LLaVA model understands many actions and human behaviour due to datasets like Kinetics, it can transfer its knowledge to find anomalous actions that deviate from normal ones. Prompts are used to focus the model's attention and guide its decision-making process. Prompts such as "Assess for signs of distress or unusual behaviour by looking at the gestures and facial expression" direct the model to look for deviations from the normal activities learned from the Kinetics dataset and others. When the model detects a person holding a chest with a painful expression, it is not just identifying the action based on visual cues; it is also interpreting the context provided by the prompt that such behaviour (a person holding a chest and grimacing might be an emergency case like a heart attack) may indicate an emergency.

Other example prompts like "Are there any signs that suggest the person might be in pain, like facial expressions of distress, protective postures?" or "Is there any evidence that the person might have fallen, like disarranged furniture or objects" allow LLaVA, with its advanced visual and textual comprehension and instruction-

following ability, to focus on critical aspects of the environment. LLaVA's architecture and training datasets allow the creation of highly specific and nuanced prompts with instructions that can guide the model to meticulously analyze visual scenes and textual descriptions for subtle cues of distress, unusual behaviours, or environmental indicators of emergencies.

The model needs detailed visual instructions and expected response output to help the model follow and perform the user's instructions. This way, LLaVA can be tasked to identify less apparent signs of emergencies, such as subtle indicators of discomfort or distress in a person's face or surroundings. This detailed level of prompting enables LLaVA to apply its advanced reasoning capabilities to detect signs that are not immediately evident [73]. To present the capabilities of LLaVA, an example in figure 3-2 shows the ability of LLaVA to take into account the visual instructions and reason on the potential cues of abnormality in the scene. At first, we determine the role of the model and commands to answer "Yes" or "No" on the questions with explanations related to anomaly behaviour detection as the expected answers. As a result, it helps the model decrease the hallucinations in answers and pay attention to anomalous behaviour and specific response format.

Since the LLaVA is not specifically trained on the emergency dataset, the LLaVA's chatbot answers the questions as instructed by the prompt and generates a good explanation of why this image depicts the emergency. The model points out protective posture as an indication of pain or discomfort, suggesting possible symptoms by looking at the gesture. Also, the model noticed the fallen chair in the image and exposed it as an anomaly sign, deducing the possible fall from a chair. It shows the model's reasoning in inferring causal relationships from the image.

LLaVA has proficiency in handling a sequence of interconnected prompts, allowing for a structured chain of inquiries that build upon each other [73]. Starting with identifying potential hazards that could indicate a fall and progressing to assessing signs of injury or distress, this approach prompts LLaVA to conduct a thorough analysis by synthesizing multiple observations. Additionally, customization for specific environments or individuals enables LLaVA to dynamically assess situations across various

contexts, significantly enhancing its emergency detection capabilities. The other critical component of the system is LLaVA's dynamic response formatting, which ensures the answer is exactly as the user requested in the instructions [73]. For example, if the user specifies that responses should be in the form of a "yes" or "no", the model will adhere to this instruction and format its answers accordingly. This feature shows that LLaVA can give unambiguous answers, which is significant for situations when consistent answers are needed to track and log incidents or short responses are needed to take fast action in emergency situations. Through this comprehensive approach, LLaVA's application in emergency detection shows the transformative potential of VLMs and LLMs in critical real-world scenarios, offering a nuanced, effective tool for early detection and response to potential emergencies.

3.3 Dataset and Generalization Ability

Since the primary model, LLaVA-1.5, is partially pretrained on general images from open-source datasets like common objects in context (COCO) [42], GQA [29], OCR-VQA [50], TextVQA [79], and VisualGenome [35] and trained on a subset (LCS-558K), consisting of image-text pairs from the large-scale open network (LAION) [68], conceptual captions (CC) [14] and SBU [83], captioned by BLIP VLM [40], LLaVA can understand general visual concepts and relationships between language and images and has the potential to understand and detect the emergency scenarios or abnormal behaviour in AAL systems [75]. Additionally, including BLIP synthetic captions expands the model's understanding of objects, scenes, people, animals, relationships and other knowledge.

Finally, LLaVA is trained on the 158K unique language-image instruction-following data, 58K conversations, 23K detailed description, and 77k complex reasoning samples assisted by language only GPT-4 and human annotations [43, 44] and finetuned on science question answering (ScienceQA) dataset [33], resulting in improved instruction-following ability (reasoning, describing and understanding from the visual context). This improves the model's ability to reason and understand complex queries related

to scientific knowledge. This multifaceted training and finetuning, which includes exposure to various instruction sets, answers and synthetic data, allows the LLaVA to generalize to new and unseen scenarios. Following this, LLaVA can understand and recognize many actions and cases that were not directly covered during its training phase. Also, LLaVA's knowledge and features from self-supervised learning are generalizable. Thus, the LLaVA model can use its learned representations to infer properties about unseen inputs based on similarities to learned feature sets. Therefore, LLaVA represents a powerful tool that, despite not being specifically trained or finetuned on particular actions or scenarios, holds the potential to recognize and perform different tasks on new data.

3.4 Text-To-Speech and Speech-To-Text

TTS and STT are the tools used in natural language processing (NLP) [51]. STT is an automatic speech recognition (ASR) model that can transform speech into text. These signals are digitized (dividing the acoustic features) and compared to phonemes (parts of human speech). After comparing each phoneme, the ASR model maps words and symbols with a specific phoneme. After that, the model examines the speech's context to determine the connections between the words that are said.

Whisper STT is utilized in this work as it has shown SOTA results. This model is trained on 680000 hours (563000 hours for the English language and 117000 hours for the other 96 languages) of labelled data from the web. The backbone model for pretraining is a transformer [74] that consists of an encoder and decoder. The audio data is initially split into 30-second chunks and converted to a mel spectrogram (a visual representation of the audio's frequencies in terms of time). After going through several convolutional layers, activation function, and pre-activation residual blocks, the decoder learns the relationship between the word tokens and audio sequences, in which the most probable words are chosen as the decoder's output [63]. Whisper's decoder also utilizes special tokens alongside regular words during the generation process. These tokens instruct the model on tasks beyond basic word prediction,

like language identification or adding timestamps. Apart from transcription tasks, Whisper is capable of speech translation and speech activity. Whisper STT has outlined an almost human-like performance in understanding unseen speech patterns. In emergency situations, people might speak differently due to stress or unfamiliar situations. Whisper can still transcribe effectively without needing specific training on emergency scenarios. Also, Whisper performs well even in noisy environments, such as a crowded room or during a fire alarm, and can handle long audio recordings, transcribing or translating to specific language [63]. Whisper demonstrates impressive results on the word error rate (WER) metric, achieving a competitive 2.5% error rate on the widely used LibriSpeech dataset [56]. However, its true strength lies in its ability to generalize beyond this in-distribution data. While some supervised models trained specifically on LibriSpeech might boast slightly lower WER, their performance suffers when tested on real-world audio recordings that deviate from the training data [63]. Whisper excels at handling these unseen speech patterns, including pub noise, making it a more adaptable and robust solution for real-world scenarios.

In TTS, the text is transformed into linguistic features, which can include phonemes, words, or other linguistic units. Then, the linguistic features of the input and the "speaker" are embedded together to be encoded. After that, the decoder determines the energy, pitch, duration and spectral features (mel-spectrogram) to transform into acoustic features, which are then processed by vocoders to transform into wave forms [16].

Piper TTS is utilized in the proposed system, a neural network optimized to perform well on a Raspberry Pi 4 [5, 27]. Since the system's responsiveness in emergency cases is significant, the speed of the Piper fits the constraint of real-time scenarios. For instance, a medium-sized Piper model can synthesize speech as fast as it receives text, minimizing delay [27]. Also, the quality of the translation to audio is not compromised but preserved. Its lightweight design eliminates the need for a powerful computer, potentially reducing overall system complexity. In total, Piper's speed and efficiency make it a great choice for real-time applications on resource-constrained devices and emergency detection tasks.

In conclusion, TTS and STT models are used in this methodology since they enable a better understanding of contextual information through direct communication with the patient from the discussed paper [17]. The methodology leverages a medium-sized Piper TTS for real-time user-model interaction, while a small-sized Whisper STT offers a robust tool for understanding with high accuracy and robustness to noise, particularly valuable in situations where nuanced and precise context is needed. This combination, highlighting Piper's speed and Whisper's adaptability and robustness, ensures smooth user interaction and a deeper grasp of conversational context. In emergency cases, it is crucial for the chatbot to retrieve the user's answers, which might be obscured by background noise or the nuances of different accents, and voice its responses quickly and effectively.

3.5 Proposed method

The proposed approach combines continuous monitoring and user-model interaction blocks to detect, interact and confirm emergencies. Here's a breakdown of the key components of the system's hardware, models, and workflow.

3.5.1 Hardware

- Server Computers: Equipped with sufficient processing power (NVIDIA DGX A100) to run the LLaVA model, audio/video recording software, and data storage tools.

- Camera: One at the top corner of the experimentation room strategically placed the camera to capture a clear view (like a CCTV camera's view) of the living space and enable person detection. The camera is connected to the laptop.

- Microphones: "Rode Wireless GO" microphones are placed on the cloth of the participants to ensure clear audio capture of conversations between the user and the system.

- Audio Speaker: To deliver voice prompts and questions generated by the model using Piper TTS [27].

- Local laptop: To access the server computer through a remote connection, send and receive images and text files. Also, the local computer runs Piper TTS [27], Yolov8 (human detection) [71], and Whisper STT models [63].

3.5.2 Models

- YOLOv8: This DL model continuously analyzes camera footage to detect the presence of a person in the living space [71].

- LLaVA: This VLM plays a crucial role in emergent scenario identification and confirmation of an emergency and interaction with the user:

  - Question Generation: Based on the observed situation and context, LLaVA generates specific questions about the potential emergency [43].

  - Analysis and Decision: It analyses the user's responses to its questions and image embedding to determine whether the situation is an emergency or not [43].

- Piper TTS: This TTS engine converts LLaVA's generated questions into natural-sounding voice prompts delivered through the speakers [27].

- Whisper STT: This STT technology captures the user's answers to the model's questions with accurate transcription for LLaVA's analysis and generation of new contextually relevant questions [63].

- Emergency Alert System: Simulates emergency notifications for the experiments (sending alerts with information about the accident to family members or care providers).

The high-level view of the proposed method is depicted in figure 3-3. YOLOv8 recognizes whether there is a person in the photo and triggers the camera to take a

photo. Then, if there is a person, the photo is preprocessed (zooming in on the photo so that the model focuses attention on the person and objects standing nearby) and sent to the DGX server, where LLaVA performs initial emergency detection using VQA. If there is a potential emergency, the interaction between the person and the LLaVA model is initiated. If not, the system returns to the start and looks for the person. LLaVA is prompted to generate the questions, which are then transferred to the local system and voiced using Piper TTS. The person then answers the question, and the voice is recorded. Using Whisper ASR, it is translated into text, which is then transferred to the LLaVA model on the server part. After four questions, the LLaVA model is prompted to decide whether there is an emergency and whether we should call an ambulance based on the fused information from different modalities (image embedding and history of interaction between the user and model). After that, the LLaVA model is also asked to generate a suggestion for the user. For now, the last step is sending an email with the history of interaction with the user and suggestions. However, it is also possible to make a phone call to the ambulance with the case description. If an emergency is not detected, the system starts the process again. In the end, the LLaVA model is also instructed to classify the situation among 5 emergency and 3 non-emergency cases.

The detailed explanation of the system's workflow with state diagrams is presented in figure 3-7. The system consists of a continuous monitoring part (detailed information about the process of human detection, transfer of files between server and local part, and abnormality detection through VQA with a threshold) and user-model interaction block (detailed information of how LLaVA starts a real-time interaction with a user, how the LLaVA makes a final decision for emergency detection, how user transfers the answers to the model, how the model understands when the person is unresponsive and when the model decides to call the ambulance and send alerts).

3.5.3 Continuous Monitoring Part

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

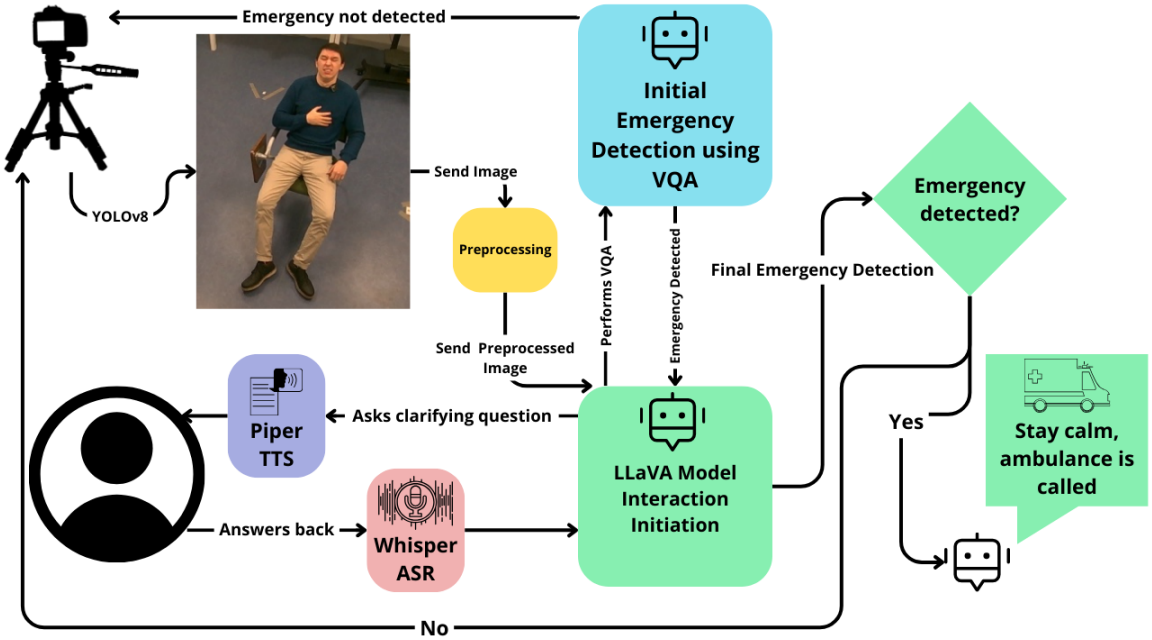

Figure 3-3: High-level representation of the system.

Image Capturing: The system employs a pretrained YOLOv8 model to identify people within the scene. The YOLOv8 is pretrained on massive classes, including animals, objects and people. The model's focus is directed towards triggering the camera to take a photo if the person is present in the scene. Thus, the passing animals or other objects will not trigger the camera. The camera is turned on at regular intervals, and when YOLOv8 detects the person, the camera captures one frame. The confidence threshold for detecting the person was set higher than 0.85 to avoid a blurred picture of a person. Any detected objects with a confidence score below 0.85 will be disregarded, ensuring that only detections with a high probability are considered to trigger the camera. After taking the photo, the captured frame is transmitted to a server-side visual assistant powered by a combined LLM, specifically, the LLaVA model. The photo taken by the camera is preprocessed to cut only the person and environment nearby to make the model focus on the actions of the person and objects. After that, the preprocessed image and "mode.txt" file with value 0 are sent to the server part to turn on the VQA process.

Abnormality detection through VQA: The system employs a predefined set of

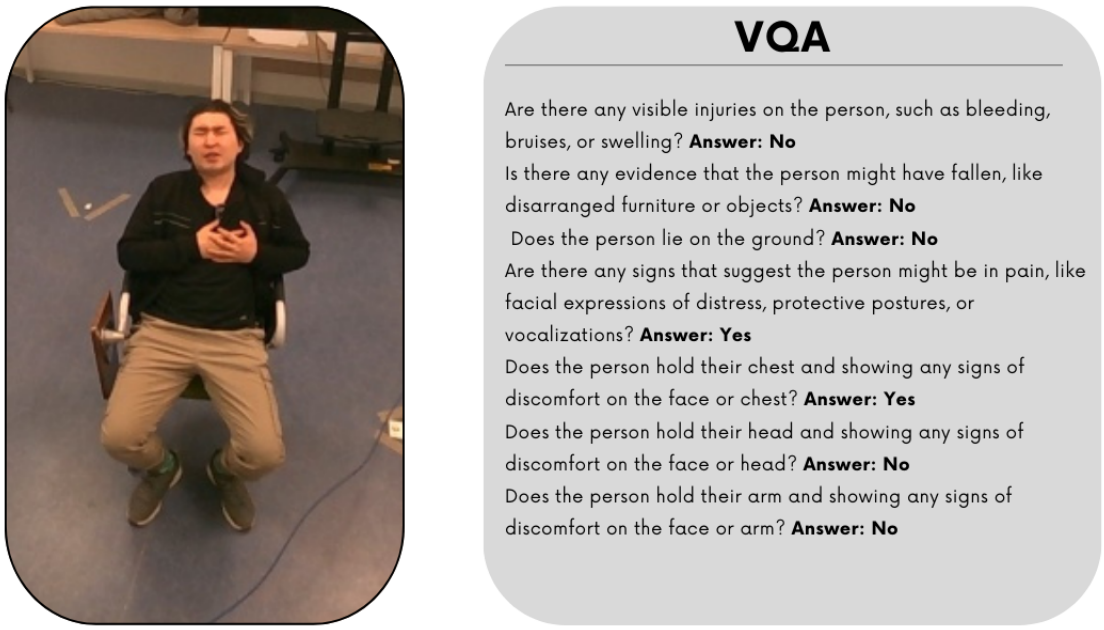

Figure 3-4: VQA of the abnormal behaviour case. This figure represents the example of the heart attack scenario and LLaVA's VQA based on the visual context. The model is prompted to answer in a short form by "Yes" or "No" to save time in an emergency. The bolded texts are the LLaVA's answers to the questions.

questions designed to detect early signs of abnormalities before interaction with a user. The VQA consists of seven questions that were carefully designed and experimented with to find the efficient instructions that are the most understandable for the LLaVA model. The model answers multiple choice questions, where a "Yes" answer signifies a potential abnormality or emergency. The model follows the instructions from the questions and makes a decision based on the context of the image and its reasoning. All the questions and answers are saved in the file and sent to the local part. The example of the emergency and non-emergency scenario with the model's answers on VQA to detect a potential abnormality is depicted in the figures 3-4, 3-5. Based on the heart attack scenario figure, the model determines the abnormality score by answering the questions with "Yes" on signs of pain, distress, discomfort, and holding the chest. In the figure with the person sitting on the computer, the model detects no abnormality and answers all questions with "No".

Thresholding: The collected answers are used to calculate an abnormality score. If

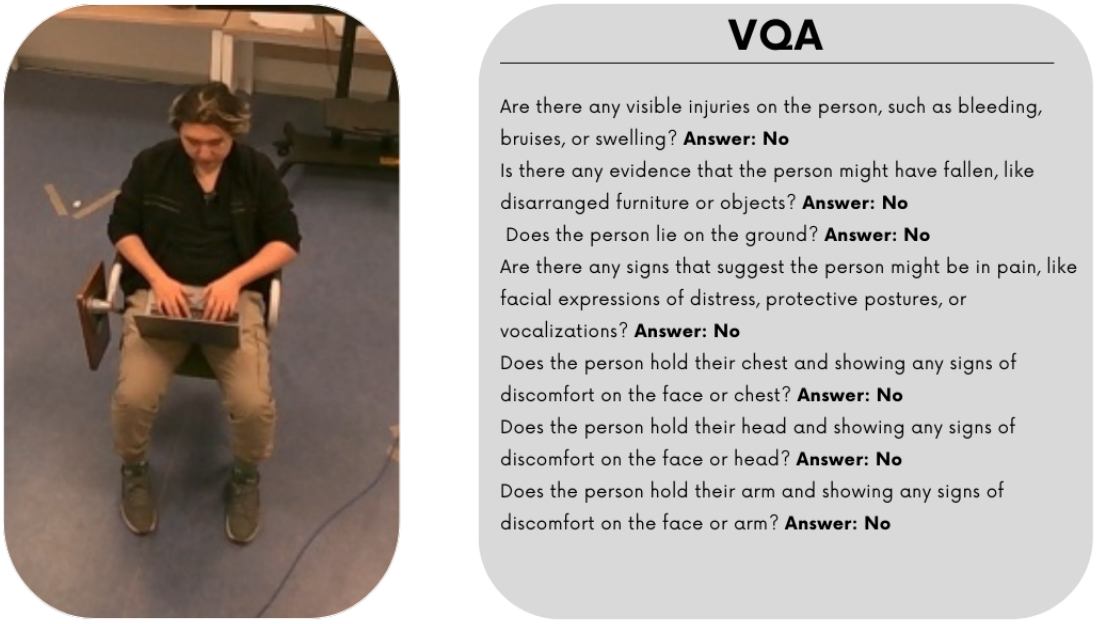

Figure 3-5: VQA of the non-abnormal behaviour case. This figure represents the scenario of a person working on the computer, and LLaVA's performs VQA based on the visual context. The model is prompted to answer in a short form by "Yes" or "No" to save time in an emergency. The bolded texts are the LLaVA's answers to the questions.

this score (number of "Yes" answers) exceeds a predefined threshold (non-anomalous action threshold), the real-time conversation with a user is triggered. A user-model interaction block grasps additional context from the model's dialogue with the user. However, before triggering the user-model interaction block, the system asks the user whether help is needed to confirm the further interaction. If the user answers "Yes", the "mode.txt" file is set to 1 to trigger the user-model interaction block. Also, if the user does not confirm help, the system automatically triggers the interaction block. This confirmation check is done for the user's convenience in case of a false alarm.

3.5.4 User-Model Interaction Block

In this block, the system interacts directly with the individual to verify and decide about the person's state. It asks four questions based on the visual context and saves the user's responses for analysis. After that, the system makes a final decision about

the person's state and creates a contextually relevant suggestion.

Generation of the questions: Once the continuous monitoring block detects potential abnormal behaviour and triggers the user-model interaction, the LLaVA model is requested with instructions to create the set of questions by itself based on the visual context and previous responses of the user. The questions are saved in a "question.txt" file that is sent to the local part for speech models.

Activation of TTS: The questions generated by the LLaVA model are then vocalized using the Piper TTS. This ensures that the individual can hear and understand the questions clearly, which is critical for accurate assessment. The user is expected to answer these questions one by one.

Activation of STT: After vocalizing the question generated by the model, the Whisper STT model is activated to capture the user's responses. The model intelligently stops recording user responses if the person is silent for three seconds, a feature that can be customized. This advanced STT technology accurately transcribes the spoken words into text, despite potential challenges such as background noise or speech impediments. The transcribed answers are compiled into the "answers.txt" file.

Analysis of the interaction: The "answers.txt" file is then sent back to the server, where it is parsed and analyzed by the LLaVA model. At this stage, the system assesses the severity of the situation based on the context of the answers provided by the individual and previous image embedding. If the answers.txt file is empty, it would indicate that the individual has not responded to any question, and the system will deduce that the person is unresponsive. This case is treated as a critical emergency, and the protocol to call an ambulance is immediately initiated. However, if the file contains answers, the LLaVA model evaluates the content to decide if there is a need to call emergency services. It considers the context of the situation by looking through the history of the user's responses to the questions and the image context (the posture, emotion, environment objects, signs of wound, and others) that are combined in a late fusion. If the analysis suggests that the user is in distress or facing an emergency, the system will proceed to call an ambulance. Figure 3-6 shows

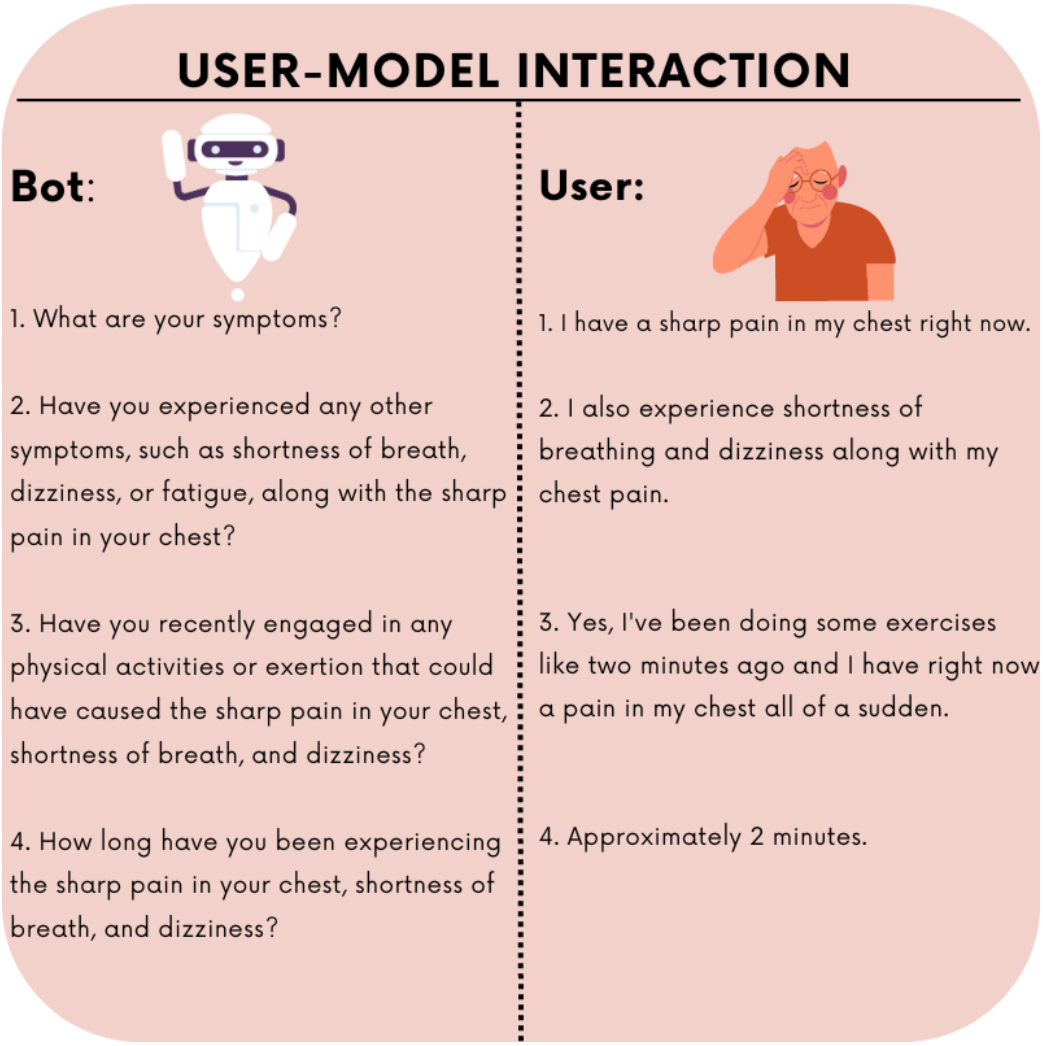

Figure 3-6: User-model interaction. The figure depicts an example of a real-time interaction during an emergency scenario like a heart attack. On the left are chatbot-generated questions created based on a user's previous responses and image embedding; on the right are the user's answers to questions.

an example of interaction between the user and the LLaVA model for a heart attack scenario.

Decision-making: The iterative process of generating questions and analyzing user responses continues until the LLaVA model decides whether to call an ambulance. Throughout this interaction, the system is designed to act autonomously, making informed decisions to ensure the user's safety and well-being in potential emergencies. If the system detects the emergency, the model will generate the overall context of

the situation, suggest what to do for the user, and send the information to the doctor or caregiver for further steps.

Overall, the detailed framework shown in figure 3-7 exemplifies a comprehensive approach to emergency detection and response. While the continuous monitoring stage ensures persistent monitoring for potential abnormalities, its interactive stage facilitates real-time assessment to decide whether to call the ambulance. The system produces a detailed context inferred from the image and the conversation in the end.

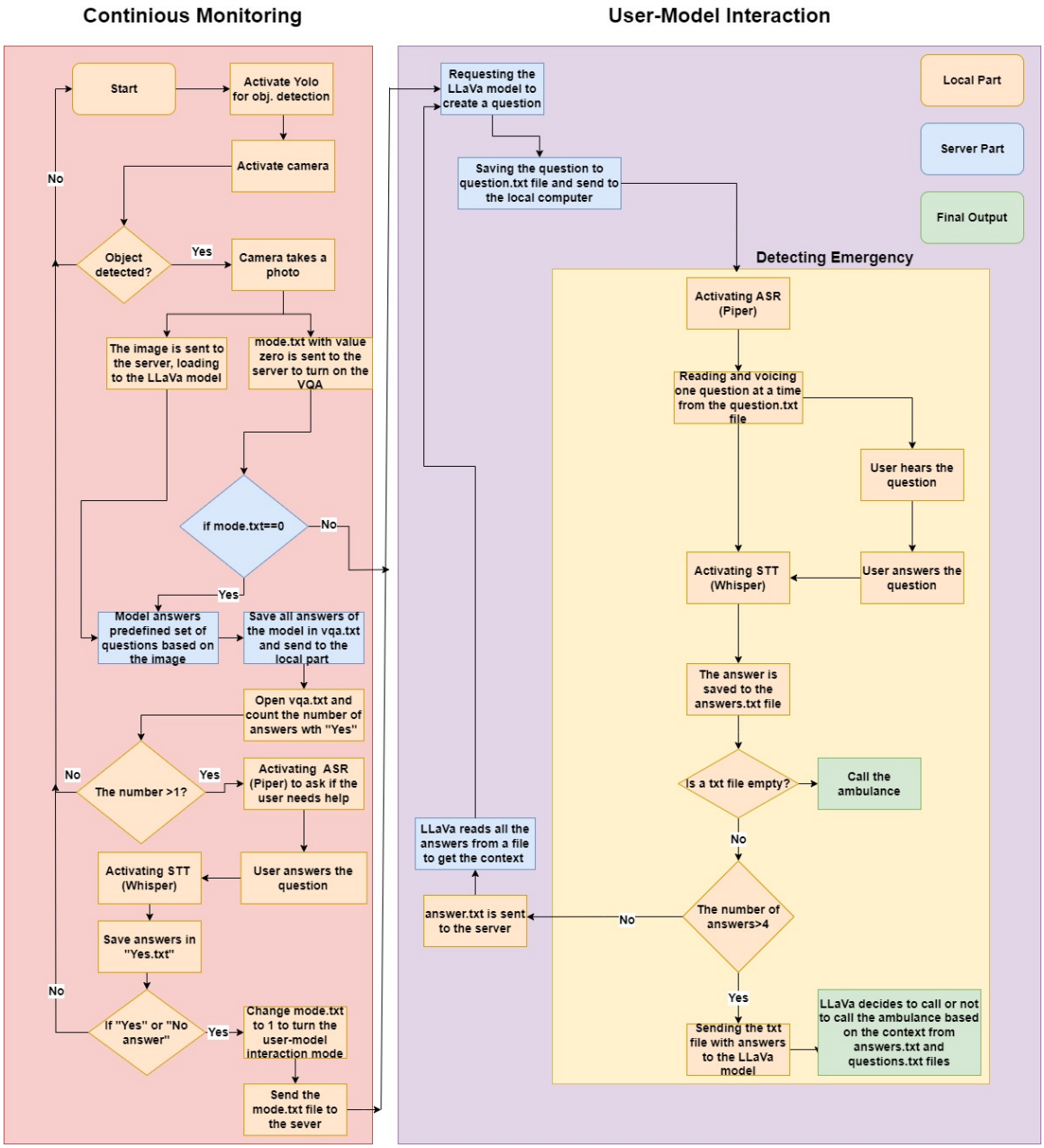

Figure 3-7: Proposed framework. On the left side is a continuous monitoring block involving human detection, VQA tasks, abnormality score calculation, thresholding, and triggering of the user-model interaction block. On the right side, the user-model interaction block involves the generation of questions, multimodal fusion, activating speech models, collecting the user's responses, and making a final decision about the severity of the incident.

Chapter 4

Results

This chapter presents the experiments conducted with voluntary participants, detailed information on the experimental procedure, qualitative and quantitative evaluations, and a comparison of the proposed approach with SOTA works on HAR, AAL and abnormal behaviour detection.

4.1 Experiments

This section provides information on how the experiments were conducted, including the ethics, participants, and experimental procedure.

4.1.1 Ethics

This research was certified by the Nazarbayev University Institutional Research Ethics Committee (NU IREC). The experiments were conducted using all confidentiality, safety, and ethics measures. There were no conflicts of interest between the participants and the researchers, which implies that the results were legitimate.

4.1.2 Participants

There were 24 participants in total: 9 women and 15 men. The spreadsheet for participating in experiments was shared via email among the students of Nazarbayev

University. The participants' ages ranged from 20 to 25, and most were bachelor's and master's degree students.

4.1.3 Experimental Procedure

Firstly, The participants were asked to come to the specified room for the experiments. They were then informed about the experimental procedure and asked to read and sign the consent form. The participants were asked to perform eight cases, including five emergencies and three normal scenarios. The five emergency scenarios include heart attack, heart attack with fainting, head or brain injury (or neurological disease), broken leg, and open wound. The three scenarios with normal activities include watching TV, reading a book, and sitting with a laptop. The reason for choosing these scenarios is to test the system's ability to generalize on common emergency cases. However, because LLaVA is pretrained on vast datasets and has the instruction-following capability, the model can be generalized from related concepts, recognizing new normal and abnormal cases. Some of the examples of emergency and non-emergency cases are presented in the figure 4-1. Here is a detailed description of all of the scenarios:

- *Heart attack*: The participant is sitting on a chair with their hand or hands on their chest and mimicking a painful expression on their face. The symptoms include chest pain, difficulty breathing or shortness of breath, fatigue, dizziness, and others.

- *Heart attack with fainting*: The participant is also sitting on a chair with their hands on their chest and a painful expression on their face. However, this time, they cannot speak and respond to the system, so the model can reason for a possible fainting or loss of consciousness.

- *Head/brain injury or neurological disease*: The participants act like they have lost their memory and do not know where they are. In addition to memory loss and disorientation, they mimic headaches in this scenario.

- *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 and stand up, losing feelings in their leg, swelling, skin color change on their leg, and others.

- *Open wound*: The person has also accidentally fallen and cut their hand with a sharp object, causing unstoppable bleeding. In this case, the participants wrap their hands or legs in a painted red bandage and pretend to be in pain by holding it. The symptoms include bleeding that the person cannot stop, severe pain, losing consciousness, dizziness, and others.

- *Watching TV*: The participant is sitting on a chair and watching TV without showing any discomfort.

- *Reading a book*: The participant is sitting on a chair and reading a book.

- *Sitting with a laptop*: The participant is sitting on a chair with a laptop and pretending to be working or playing a video game.

4.2 Evaluation

The qualitative results of the experiments are evaluated in two ways. Firstly, participants were given the subjective questionnaire with ten questions in total. The objective of the survey is to evaluate the experiment in terms of difficulty, complexity, speed, consistency, and other criteria. The questionnaire was constructed using NASA-Task Load Index (NASA-TLX) [28] and System Usability Scale (SUS) [38]. Secondly, participants were also asked to assess the performance of the LLaVA model by asking concise questions and providing the user with contextually relevant suggestions.

The quantitative results of the experiments are evaluated in several ways: 1. The answers from the VQA task are compared to ground truths to obtain accuracy and

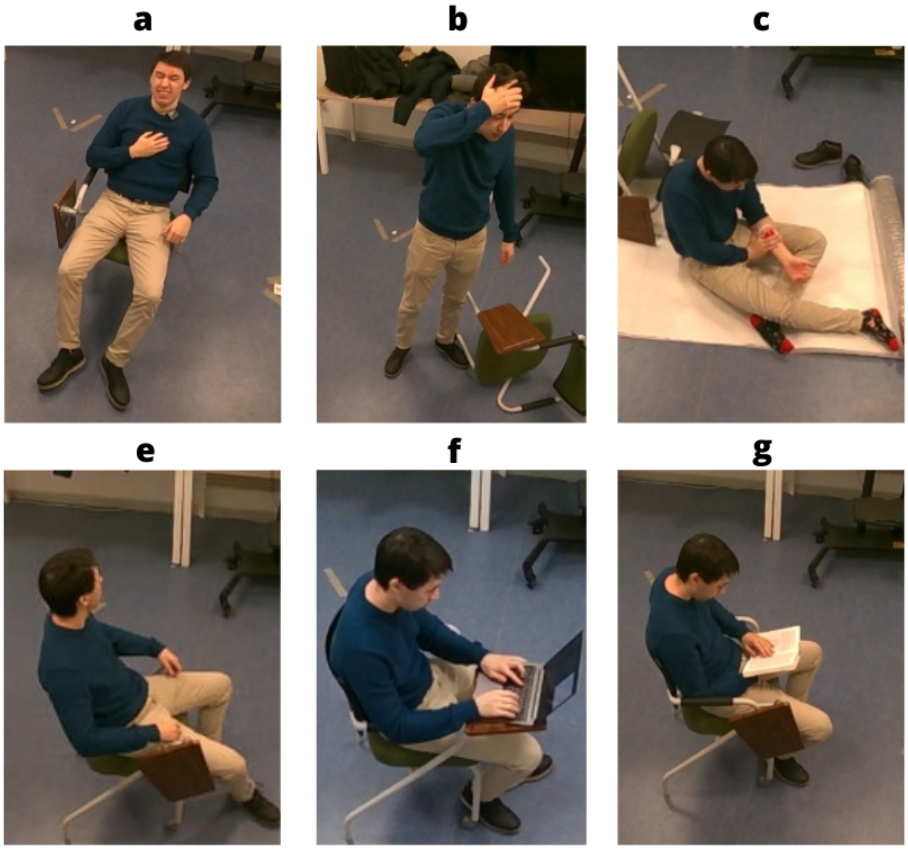

Figure 4-1: Emergency and non-emergency scenarios. The figure shows three examples of emergency cases: a) heart attack, b) head injury, c) open wound, and three non-emergency cases: e) watching TV, f) sitting with a computer, and g) reading a book.

understand how well the model answers the questions with instructions for each scenario. 2. The system's ability to turn on the interaction after the VQA task is evaluated by counting the correct activation of the interaction with the user when there is an emergency and vice versa. This binary activation was evaluated using accuracy, recall, precision, F1 score, and specificity metrics. 3. After the interaction process, the system's final decision (whether the analyzed case is emergent and needs an ambulance) is evaluated. The system is instructed to do a binary classification (emergency or non-emergency) after the interaction. 4. Since the system has multimodal contextual information from different stages, it was instructed to do a classification task on 8 classes (5 emergency activities and 3 normal activities). The

system's time performance is evaluated in terms of time and how fast the system's internal processes are in detection and assistance in emergency cases.

4.3 Qualitative Evaluation of the Results

The qualitative evaluation of the results of the experiments was done through the subjective questionnaire with 10 questions. The questionnaire was filled out by 24 participants, 9 of whom were women and 15 were men. The possible answers for each of the questions were scaled from 1 to 5, and 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)

Table 4.1: Mean results for each question in the subjective questionnaire 4.3

Question	1	2	3	4	5	6	7	8	9	10
Female	1.44	1.89	1.33	2	1.78	4.67	1	4.89	1.44	4.78
Male	1.27	1.93	1.47	2.13	1.47	4.6	1.8	4.53	1.27	4.73
Overall	1.33	1.92	1.42	2.08	1.57	4.625	1.375	4.75	1.33	4.75

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 4.1 represents the mean results (mean answers from 1 to 5) for each question for male and female groups and overall means. First, it can be seen that the results are more directed toward the positive scales, and there might also be some statistical differences between male and female groups.

To analyze the results in more depth, statistical tests must be employed to verify the statistical differences between the groups.

4.3.1 Statistical Analysis

First, it is necessary to identify which statistical test can be used for the results. Although the data does not seem normally distributed, the normality of the data should be tested first. One of the most powerful normality tests is the Shapiro-Wilk test for normality [32]. Shapiro-Wilk test considers 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:

$$W = \frac{(\sum_{i=1}^{n} a_i x_(i))^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{4.1}$$

Here, n is 24, which accounts for the number of samples, a_i is a critical value from the table of the Shapiro-Wilk test, in which the constant values depend on the number of samples, $x_(i)$ is the answer of the participant from 1 to 5 scale and $\overline{x}$ is a sample mean for each question from table 4.1. The rejection of the null hypothesis is done

if the value of W is less than a critical value from the table or the value of W is not close to 1. After calculating the test statistic for one question, the resultant value was $W = 0.02$, while the perfect match for W is 1. Also, the value of W was compared with a critical value of 0.916 at 0.05 level (commonly used significance level) from the table of Shapiro-Wilk test with p values. Since the obtained statistical value of W is less than a critical value, then the null hypothesis H_0 is rejected.

As the data collected is not normally distributed, it can now be said that non-parametric statistical tests should be used here to identify whether there is a difference between the answers of male and female participants. Given two data groups (male and female), one of the most appropriate tests for the case is the Mann-Whitney U test [49]. Mann-Whitney U test considers 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:

$$U_1 = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1 \tag{4.2}$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - R_2 \tag{4.3}$$

In equations 4.2 and 4.3, R_1 and R_2 are the sum of the ranks for male and female answers. To obtain the ranks, two groups are pooled in one set, and the rank is set from the smallest to the largest (from 1 to 24) for the values from the smallest to the largest. n_1 and n_2 represent the number of males (15) and females (9). After the calculation of U_1 and U_2, the smallest value is chosen to be compared with the crucial value from the reference table of Mann-Whitney U test. If the observed value of U statistic is less than or equal to the critical value at the confidence level of 0.05, then H_0 hypothesis is rejected. Table 4.2 represents the calculated U statistic for

Table 4.2: 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**

each question and U critical for both one-tailed and two-tailed tests. In this case, we are trying to determine whether the male answers differ from the female, while the answers can differ in both ways (can be larger or smaller). This implies that the U-critical for the two-tailed test should be considered, meaning the differences in both directions need to be used. It can be seen that the null hypothesis H_0 can be rejected for questions 1, 3, 6, 7, 9, and 10 in table 4.2 since the $Ustatistical$ is less than or equal to the critical value ($Ucritical$ two-tailed). This implies a statistical difference between male and female answers to the earlier questions.

4.3.2 Discussion of Statistical Analysis and Questionnaire

To understand the difference between male and female answers, we need to consider the questions where there is a difference, one by one. The first question considers the difficulty of the experiments. We can see in table 4.1 that the mean answer for women is considerably larger than for men, meaning that it was more difficult for women to complete the experiment. The third question considers how much physical activity was required, and the mean answer for men, here, is larger. This implies that the experiment was more demanding for men, probably because of the scenarios that required them to lie on the ground. Next is the sixth question, which considers the intuitiveness of the system. Although the mean values for the male and female answers differ only slightly, as can be seen in table 4.1, it can be stated that, statistically, the interaction with the system was more intuitive for women. Women also tended to

find more inconsistency in the system's workflow and were more confident interacting with it. Notably, in question 7, all of the female participants chose answer 1, meaning that all of them found the system easy to use, while it was not that easy for some male participants.

4.3.3 Performance of the LLaVA Model in Generating Suggestions and Questions

In addition to the subjective questionnaire, the participants were also asked to evaluate the model performance based on the Likert Scale. The participants were asked to evaluate the questions based on the following Likert 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.

After gathering the answers, the average result was **3.75**. This means that in most cases, the model generated relevant, concise questions that guided the model in understanding the issue. Notably, the model sometimes asks questions that do not regard the person but the circumstances in which the emergency occurred and the person's surroundings. This might be the reason for the result not being perfect. Another reason can be that the participant provides lots of context during the first question, making it difficult for the model to develop a good question. Therefore, the model goes into details that are not important for the situation.

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.

The average value between all answers was also **3.75**. This means that the model generated fairly detailed suggestions that regarded the person's state identified the probable diagnosis, and gave corresponding advice. However, sometimes, the model did not have enough contextual information and was careful in diagnosing the user, resulting in suggestions that included a wide range of potential diagnoses. Also, some participants were short in their answers, resulting in the model's further guessing of possible diagnoses. This might be the cause for the resultant average value not being perfect.

4.4 Quantitative Evaluation of the Results

The effectiveness of the emergency detection system is quantitatively assessed through several key metrics that focus on its ability to correctly identify emergencies, initiate appropriate user interactions in emergency cases and making a final decision whether the user needs an ambulance (based on the severity of the situation). Also, the classification for 8 classes was conducted after the interaction block.

4.4.1 Evaluation of VQA Accuracy in Emergency Detection

The accuracy of the VQA mechanism plays a pivotal role in evaluating the monitoring system. This is the central part of the continuous monitoring block, in which the

abnormality score is calculated. In this unit, the decision to engage the interaction block is based on the computed abnormality score. The VQA's effectiveness was measured by its ability to correctly identify signs of various emergencies and answer the predefined set of questions correctly. The accuracy is used as the primary metric for evaluating VQA.

VQA Accuracy: The VQA accuracy is computed by evaluating the model's responses to a predefined set of questions to uncover signs of abnormality in various scenarios. The accuracy is calculated using the following formula 4.4:

$$\text{VQAaccuracy (AccVQA)} = \frac{\text{Number of Correct Answers}}{\text{Total Number of Questions}} \tag{4.4}$$

Each scenario presented to the system, ranging from heart attack to sitting on a computer, was accompanied by a series of questions designed to determine abnormality scores. The system's performance was gauged by comparing ground truth answers with observed answers ("Yes" and "No" answers on the questions from the figure 3-4). The table 4.3 below delineates the frequency of correct responses for each VQA question within the emergency context. Overall, we have 7 questions for one scenario, 5 emergency scenarios, 3 non-emergency scenarios, and 24 participants. The model answers a set of questions (7 predefined questions) across all emergency cases. That is 120 answers on the one question for all scenarios (24 participants x 5 emergency cases), 24 answers on the one question for one emergency case, and 840 answers (5 emergency cases x 24 participants x 7 questions) for all the questions for each emergency scenario.

The table 4.4 shows 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. Thus, there are 288 answers to one question for all non-emergency cases (3 normal cases x 4 repetitions x 24 participants), 96 answers to one question for one non-emergency

Table 4.3: VQA answers each question across all the emergency scenarios

Question	1	2	3	4	5	6	7	STD	AVG
Heart Attack	24	24	22	8	21	11	14	6.601	17.714
Fainting	24	24	23	10	23	8	12	7.319	17.714
Head Injury	24	18	17	19	6	22	0	8.802	15.143
Broken Leg	24	13	23	11	10	14	13	5.682	15.429
Open Wound	13	12	21	7	17	18	15	4.572	14.714
STD	4.919	5.762	2.490	4.743	7.231	5.550	6.140	-	-
AVG	21.8	18.2	21.2	11	15.4	14.6	10.8	-	-

Table 4.4: VQA answers each question across all the non-emergency scenarios

Question	1	2	3	4	5	6	7	STD	AVG
Watching TV	96	96	96	95	93	92	92	1.9	94.286
Reading a Book	96	96	96	96	96	96	96	0	96
Sit at a Computer	96	96	96	96	96	94	96	0.756	95.714
STD	0	0	0	0.577	1.732	2.0	2.309	-	-
AVG	96.0	96.0	96.0	95.667	95.0	94.0	94.667	-	-

case (24 participants x 4 repetitions), and 2016 answers to all the questions for each non-emergency scenario (96 answers to one question for one non-emergency x 3 non-emergency cases x 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 4.5 represents 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 tables 4.4 and 4.3.

Table 4.5: VQA Accuracy for Different Scenarios

Type	**Scenario**	**VQA ACC.**
Emergency	Heart Attack	73.81%
	Fainting	73.80%
	Head Injury	63.10%
	Broken Leg	64.29%
	Open Wound	62.00%
Non-emergency	Watching a TV	98.21%
	Reading a Book	100.00%
	Sitting on a Computer	99.70%

4.4.2 Discussion of the VQA

In emergency scenarios, the system's VQA is challenged with discerning health-related severe events such as heart attacks, fainting, and physical injuries. Our data demonstrates that the VQA system achieved varying accuracy across these scenarios. Specifically, heart attack and fainting scenarios showed relatively high accuracy of 73.81%, suggesting that the VQA system is adept at identifying critical, life-threatening situations. However, the accuracy for head injuries, broken legs, and open wounds was slightly lower, hovering around the 62-64% range from table 4.5.

The distribution of answers on the VQA questions, depicted in table 4.3, shows a varied distribution of correct answers in heart attack, fainting, and injuries, suggesting a differentiated sensitivity of the system to specific symptoms or visual cues.

For heart attack and fainting scenarios, the VQA responses are relatively consistent, indicating a solid model performance in recognizing classic signs of this particular emergency. In the case of head injuries and open wounds, there is a noticeable decrease in correct responses to a specific question. This may be attributed to the varied visual presentation of these emergencies, potentially leading to a lower VQA response rate. For instance, some participants covered the visible part of the face and showed very differently in emergency cases. 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 arms. Another reason for not answering correctly in the VQA task can be the camera's distance from the user, resulting in a poor understanding of the facial emotion and some protective postures. Also, the instruction for the 7th question can be vague because the LLaVA may perceive holding a hand on the head as a situation where a person is holding a hand with the other hand due to discomfort. Thus, more precise instructions are needed, specifically for the 7th question in the head injury case.

The systematic variation in correct answers across these questions is a key observation, indicating the model's differential response to varied indicators of emergencies. For instance, questions related to more overt signs of an emergency (e.g., lying on the ground, falling, bleeding, bruises, swelling, disarranged furniture) may yield higher correct response rates compared to subtler indicators (e.g., facial expressions of distress, discomfort). This provides valuable insight into potential areas for model refinement, guiding future research and development efforts.

The lower VQA response rate for subtler signs of emergencies (question 4 across most scenarios) suggests that this may be less visually apparent or that the model may benefit from further training on more varied samples and these specific indicators like pain in facial expression, distress, and protective postures.

Conversely, the model demonstrates high accuracy in non-emergency scenarios such as watching TV or reading a book (table 4.5). The VQA system reliably identifies these situations as non-critical, indicated by the uniformly high number of correct responses across all questions (table 4.4). This highlights the system's precision in

differentiating everyday activities from potential emergencies, enhancing user convenience by reducing unnecessary interventions.

4.4.3 Evaluation of Pre-Interaction VQA Binary Activation

The binary activation for emergency cases means the number of correct activations of the interaction with the user when the emergency case persists. The binary activation for non-emergency cases means the number of correct non-activations after VQA when a normal case persists. In the development and refinement of the monitoring system, selecting an optimal threshold for emergency detection was crucial. The chosen threshold was meticulously calibrated manually through experimentation to ensure that the system activates and initiates user interaction when an emergency is likely to occur, despite the possibility of occasional incorrect responses to specific questions in VQA. When the system activates the interaction module after comparing the abnormality score with the threshold, it considers the observed action anomalous before the interaction stage.

The design philosophy behind setting this threshold acknowledges the inherent complexity of real-world scenarios where clear answers may not always be available. By choosing an optimal threshold, we strike a delicate balance, minimizing the likelihood of the system overlooking an actual emergency due to a few incorrect question responses while reducing the incidence of false positives that could lead to unnecessary interactions with users.

The selected threshold works effectively as a filter. It accentuates the significance of the pattern of answers rather than individual inaccuracies. It allows for some margin of error in question responses while still maintaining a high sensitivity to potential emergencies. This approach recognizes the critical nature of emergency detection systems, where the cost of missing an actual emergency is far greater than the inconvenience of an occasional false alarm.

The outcomes of this activation are categorized into four types of classifications before the interaction stage:

Table 4.6: Confusion Matrix for Binary Activation Evaluation

		True Class		
		Emergency	Non-emergency	**Total=192**
Predicted Class	Emergency	114	8	122
	Non-emergency	6	64	70
	Total=192	120	72	

- True Positives (TP): The model correctly identifies a situation as an emergency when it is indeed an emergency.
- False Positives (FP): The model incorrectly identifies a situation as an emergency when it is not an emergency (a false alarm).
- True Negatives (TN): The model correctly identifies a situation as a non-emergency when it is not an emergency.
- False Negatives (FN): The model incorrectly identifies a situation as a non-emergency when it is actually an emergency (a missed detection).

The table 4.6 shows the distribution of TP, FP, FN, and TN. There is 1 activation for 1 participant and scenario. That is, 192 cases in total for 24 participants and 8 scenarios (the total number of participants is multiplied by the total number of cases). Table 4.7 presents the accuracy, precision, recall, F1 score, and specificity

Table 4.7: Performance Metrics of the VQA System

Metric	**Value (%)**
Accuracy	93.44
Precision	92.71
Recall	95.00
F1 Score	93.84
Specificity	88.88

for the system's binary activation based on the TP, FP, FN, and TN depicted in table 4.6. Since there is a binary classification system in activating the interaction, the outcomes are classified as either abnormal or non-abnormal in pre-interaction stage. Each metric offers insight into different aspects of the system's performance.

Accuracy provides a general measure of correctness, precision focuses on the exactness of the emergency predictions, recall emphasizes the system's ability to detect all actual emergencies, the F1 score conveys the balance between precision and recall, and specificity measures the system's ability to identify non-emergency situations correctly.

4.4.4 Discussion of Pre-Interaction VQA Binary Activation

The binary activation evaluation through the confusion matrix (Table 4.6) and the subsequent performance metrics (Table 4.7) provide insightful data on the operation and reliability of the VQA system with the threshold within the human care monitoring framework.

The confusion matrix's results are foundational in assessing the VQA's decision-making process in activating user interaction following a potential emergency. With 114 TP and 64 TN, the system has demonstrated a robust capability to identify emergencies and non-emergencies correctly. Such a high TP rate indicates the system's responsive nature, while the TN rate affirms its discernment, avoiding unnecessary activation when no emergency is present. Additionally, the FP rate is 8, which is relatively low, indicating that the system seldom misclassifies a non-emergency as an emergency, thus preventing undue distress to the users and caretakers. Equally important is the FN rate, which is 6. Thus, the system rarely fails to activate during an actual emergency.

The performance metrics and binary activation evaluation of our monitoring system's VQA module, detailed in Tables 4.6 and 4.7, exhibit a high degree of accuracy (93.44%) and precision (92.71%), with impressive recall (95.00%) and specificity (88.88%). The high accuracy and recall suggest that the system detects emergency cases in most cases. The specificity is also high, resulting in the system's ability to identify non-emergency situations accurately, and it is an essential factor in user convenience and system trust. The low rate of FP suggests that the system's threshold for initiating user interaction after VQA is well-calibrated. Despite the system's high reliability, the presence of FN rate points to the critical need for continuous refine-

ment, particularly in enhancing the system's capability to differentiate some signs of emergency.

4.4.5 Evaluation of Post-Interaction Classification

After interacting with the user, the LLaVA model was instructed to do a binary classification task (emergency or non-emergency; call the ambulance or not to call the ambulance). After that, the model was instructed to make suggestions based on the collected information (history of interaction and image embedding). Also, a multi-classification task among 8 cases (emergency: heart attack, fainting, broken leg, open wound, head injury; non-emergency: watching a TV, playing a computer, reading a book) was performed by the model based on the collected context (history of the interaction, image embedding and suggestion). The LLaVA model was instructed to assess the severity of the situation and choose whether to call the ambulance or not. If the LLaVA model chooses to call the ambulance, this means defining the observed activity as an emergency and vice versa. So, the model chooses one option based on the collected context to make a binary classification. According to the figure 4-2, the LLaVA model was also instructed to do a multi-classification task based on the cumulative context like image embedding, history of interaction and final suggestion after the binary classification task and final suggestion of the LLaVA model.

Table 4.8: Classification Results

Classes	**Correct cases**	**Accuracy**
8 classes	192/192	100%
2 classes	192/192	100%

According to the table 4.8, the accuracy for 8 classes and 2 classes achieved 100%, showing excellent classification results, meaning that the model has enough contextually relevant information to make a final decision, determine the correct class and call the ambulance. 192 cases (24 participants x 8 scenarios) were evaluated for post-interaction classification. The same number of cases (192) were used for binary classification.

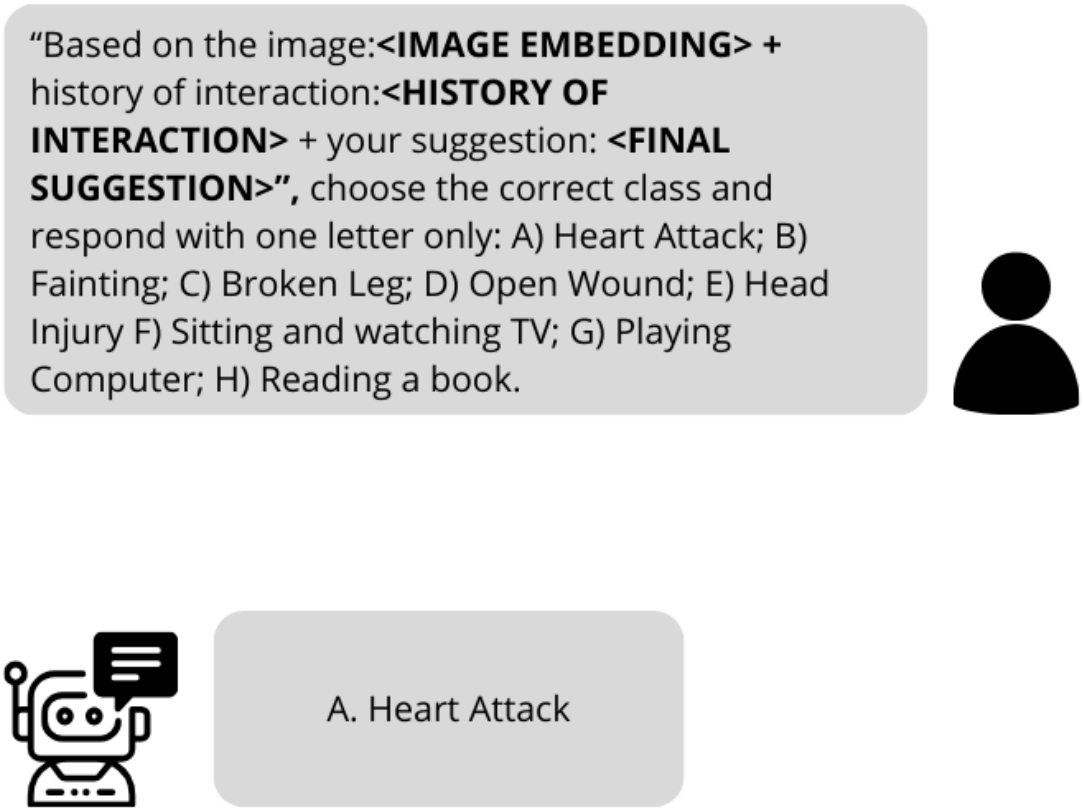

Figure 4-2: Instruction of the LLaVA model for classification task. The model was instructed to classify confirmed abnormal behaviour among 8 classes (5 emergency and 3 non-emergency) based on the cumulative contextual information like image embedding, interaction history and final suggestion.

4.4.6 Discussion of Post-Interaction Classifications of the Model

The remarkable efficiency of our system is further accentuated when the model transitions into the interaction mode with the user. In this stage, the system boasts a perfect record of accurately discerning the nature of an incident, whether it suggests calling an ambulance for an emergency or refraining from action in non-critical situations. The autonomy of the model in generating context-relevant questions and its sequential collection of user responses are central to this success. The LLaVA model does not rely on static scripts. Instead, it dynamically creates questions tailored to

each scenario's visual and contextual nuances. 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. However, the system was tested in ideal conditions without any background noise, showing great results in transcribing the user's responses to text correctly. Although Whisper STT performs well in pub noise, this ASR needs to be tested in a noisy environment to understand how well it recognizes and transcribes speech in background noise. Also, the Whisper model sometimes produces hallucinations when there is little noise or mumbling from a person, resulting in predicting words that the user did not speak.

The LLaVA model was instructed to make a decision on whether to call the ambulance or not after the interaction. Calling the ambulance means considering the observed scenario as an emergency, and not calling the ambulance means a non-emergency case. As a result, the proposed system achieved 100% accuracy in binary classification (emergency and non-emergency) and multi-classification (5 emergency classes and 3 normal classes) from table 4.8. If we compare the results in the pre-interaction stage with the post-interaction stage, it can be said that LLaVA's performance is increased due to the rich context gained during the user-model interaction combined with the final suggestion and image embedding. Overall, integrating the speech models in the system's pipeline for direct communication with a user allows the LLaVA model to gain perfect results in classification tasks for abnormal activities and ADL.

4.4.7 Comparison with SOTA Works

According to table 4.9, the comparisons with our proposed work and SOTA work on HAR, abnormal behaviour detection, gait dysfunction, falls, and ADL were conducted in terms of accuracy results. Although our proposed system (pre-interaction) does not outperform some SOTA works in HAR, our system still has high accuracy in detecting emergency cases. Also, most methods do not explicitly show abnormal behaviour detection accuracy but high accuracy for HAR or ADL. Moreover, our

proposed system is not limited to classifying a limited number of classes; our approach can classify even new actions and scenarios. The proposed work (post-interaction) for 8 classes and 2 classes achieves 100% accuracy, outperforming them in recognizing emergency cases and ADL.

Table 4.9: Accuracy comparison of the proposed system with the existing literature in HAR, AAL and abnormal behaviour detection. Proposed (pre-interaction) means the performance of our approach without the interaction part and responses from the user, while proposed (post-interaction) means the performance with the context of interaction and history of responses.

Citation	**Year**	**Method**	**Classes**	**Accuracy**
[13]	2016	Non-intrusive sensors +DT	6 ADL	94%
[24]	2018	Non-intrusive sensors +DCNN	10 ADL	98.54%
[6]	2019	Intrusive sensors +SVM	4 ADL	88%
[23]	2019	Non-intrusive sensors +Advanced Belief Model	14 ADL	99.74%
[34]	2019	Vision sensor +HMM	9 ADL	84.33%
[30]	2022	Vision sensors + MSResNet	2 Gait Dysfunctions +1 Normal Gait	93%
[64]	2023	Vision sensors +BERT	5 Falls +7 ADL	99.50%
Proposed (pre-interaction)	2024	Vision sensors +VLM + TTS +STT	1 emergency + 1 non-emergency	93.44%
Proposed (post-interaction)	2024	Vision sensors +VLM + TTS +STT	1 emergency + 1 non-emergency	**100%**
Proposed (post-interaction)	2024	Vision sensors +VLM + TTS +STT	5 emergency + 3 non-emergency	**100%**

According to table 4.10, the concrete attributes are compared within each study. The first attribute is the multimodal nature of the system, which stays for the use of different modalities and data. The proposed approach, [23], [13] and [6] present the multimodal approach, indicating that other studies primarily rely on single-modal

Table 4.10: Comparison of attributes in abnormal behaviour detection, 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
[13]	✓	✗	✗	✗	✗	✓	✗	✗	✓	✗
[24]	✗	✗	✗	✗	✗	✗	✓	✗	✓	✗
[6]	✓	✗	✗	✗	✗	✓	✗	✓	✓	✗
[23]	✓	✗	✗	✗	✗	✗	✓	✓	✓	✓
[34]	✗	✗	✗	✗	✗	✓	✗	✗	✓	✓
[30]	✗	✗	✗	✗	✗	✓	✗	✓	✓	✓
[64]	✗	**partial**	✗	**partial**	✗	✓	✗	✗	✓	✓
Proposed	✓	✓	**partial**	✓	✓	✓	✓	✓	✓	✓

approaches. For instance, the works [34], [30], and [64] use only the visual features to classify the actions. 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 abnormal behaviour detection. Causal reasoning is marked only 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. [64] has a partially self-supervised feature because of the usage of LLM BERT which was pretrained self-supervised, where it learnt general language representations without needing labelled data for specific tasks. However, BERT is not directly applying its pretrained language model capabilities. Instead, it is adapted and fine-tuned for activity recognition using the skeleton features, specific labels and supervised learning.

Adapting to new classes without additional training and a self-supervised nature is lacking in most works, except the proposed system. The LLaVA model was trained in a self-supervised manner, allowing the model to learn inherent features and structures from the massive amounts of unlabeled data, increasing generalizability across many domains and tasks. When the model sees new classes, it compares image-text pairs with its knowledge base and identifies similar outputs based on contextual clues, contributing to its ability to adapt to new classes. However, the LLaVA model can still

be limited in adapting to new classes. While transfer learning can help adapt LLaVA to new classes by transferring knowledge from related tasks, the effectiveness of this approach depends on the similarity between the new classes and the classes on which the model was originally trained. If the new classes are too distinct, transfer learning might not be sufficient. Thus, the model needs supervision to see if it understands new classes with instructions, and additional finetuning would benefit the model in understanding new concepts. Interactive features of the system are only present in the proposed system as STT and TTS models are used to facilitate direct communication and gain additional context of the situation.

The next attribute is contextual understanding, which refers to a system's ability to recognize and process information based on its context (personal and environmental factors). Contextual understanding of a person is present in the works [13], [6], [34], [30], [64]. Most of the works uses the skeleton features of a person from an image or motion data using wearable sensors. Other works [24], [23] use some environmental sensors to collect the context of the environment. Contextual understanding is also available in the proposed system, which considers not only the person's state but also the environment for abnormal signs. For example, the proposed system looked not only for facial expressions or protective postures but also the environment, like a fallen chair near the person, indicating a possible fall from the chair (see figure 3-2).

Real-time detection feature appears in these studies ([6], [23], [30] and proposed system), indicating that these systems can process data and act quickly. However, most of the works did not explicitly show the time performance of their system for abnormality detection. The next feature is simple case detection, which refers to identifying straightforward patterns or basic activities like sitting, standing, walking or other. Most studies commonly show this feature, indicating that these systems can effectively identify simple activities or cases. Complex case detection involves identifying more intricate or nuanced patterns that require advanced analysis or multimodal contextual understanding. Only a few studies ([23], [30], [64], [34] and the proposed study) demonstrate this capability, suggesting that they can handle more complicated scenarios. Overall, this table presents a comprehensive comparison of

existing studies based on multiple vital attributes. The proposed system stands out with support for most features, suggesting a more robust and adaptable approach to abnormal behaviour detection and activity recognition.

4.4.8 Time Evaluation of The System

In this section, the time performance of the system is presented in table 4.11, which shows the average time of the system to detect the anomalous scenario, interact with the person, and make a final decision with a suggestion. Also, the table demonstrates the system's time performance in detecting the emergency scenario; the result obtained is the average time for all emergency scenarios among 24 participants. The LLaVA model was run on two NVIDIA V100 graphics processing units (GPU). The human detection and speech models were run on a single GeForce GTX-1060 GPU.

The system shows fast detection and taking a photo of a person that requires less than a second, while the time to send the image or the files is relatively high compared to other system processes. Network latency is the main reason for the long time needed to transfer files and images between the server and the local part. As a solution, optimizing data transfer or a faster internet connection can enhance the time results. The processing time of the model to answer VQA is quite efficient because the model can answer the 7 questions in less than a second.

In the interactions stage, which starts after sending the trigger file ("mode.txt") with value 1 to turn on the interaction stage and ends with final suggestion generation, the model LLaVA generates 4 questions overall one by one, Piper TTS voices the question 4 times, and Whisper STT listens to the user 4 times. 4 time repetitions were chosen through experimentation. This number of repetitions is optimal for the model to create detailed and contextually relevant questions and understand the situation in detail. In the end, the model makes one decision and one suggestion based on the analyzed contextual information.

The time required for interaction (overall is 109 seconds) is considerably high compared to the processes in the continuous monitoring stage (overall is 29 seconds). The reason for that is a significant amount of time spent on voicing questions and

listening to the user's answers. That is, the model sometimes generates very long questions or details, resulting in a longer time for voicing the question. The model listens 4 times to the user if the person speaks, and if the person is silent for 3 seconds, the system stops further recording the user's response. This is the main reason for the long time spent on STT, as the user may answer the question briefly or for a long time. The final decision-making and suggestion-generation stages are fast, which is excellent as these are critical steps that could impact the outcome of an emergency.

Overall, the system performance in time is 2 minutes, 34 seconds to complete all the processes in the table 4.11. For a real-life emergency detection system, it's crucial to assess whether this duration is within an acceptable range for the scenarios conducted in the experiments. According to the qualitative results depicted in table 4.4, it can be seen that participants answered questions 4 and 5 (which are designed to evaluate the speed of the system) with the result of no more than 3, indicating that the system is relatively fast from a subjective questionnaire. Referring to the real-time results, the average time for the system to detect the emergency, interact, and make a final decision about calling the ambulance coincides with subjective survey results, as it can be seen from table 4.11 with actual results. Nevertheless, it is important to assess the system's time performance in a real environment (within hospitals, home environments, and other locations) with the elderly, disabled, and healthy people to assess the system's speed more accurately.

Table 4.11: Time performance of the proposed system. The average number was calculated based on 24 patients with 5 emergency care scenarios for each patient. The LLaVA model was run on two NVIDIA V100 GPUs on the server side. YOLOv8, Whisper, and Piper were run on a single GeForce GTX-1060 GPU on the local side. Italicized entries represent the time for processing one instance and are not accounted in the final time of the whole system

Process in the system	**AVG. time in sec.**
Continious Montoring Block	
Time for the model to detect the person (YOLOv8):	0.694
Time for taking the photo (Depth Camera)	0.405
Time to send one image to the server:	5.718
Time for sending the "mode.txt" file:	5.925
Time for LLaVA to answer on the VQA (7 questions):	4.324
Time for LLaVA to answer on one question:	*0.618*
Time to send "vqa.txt" to local:	1.730
Time for voicing the question (Piper TTS) to assure help:	6.867
Time for listening the user (Whisper STT):	3.681
User-Model Interaction Block	
Time for sending the "mode.txt" with value 1 to server:	6.497
Time for generation the four questions (LLaVA):	6.501
Avg. time for generation one question:	*1.625*
Time for sending the question to local part four times:	7.330
Time for sending one question to local part:	*1.833*
Time for voicing the question (Piper TTS) four times:	20.648
Time for voicing the one question (Piper TTS):	*5.162*
Time for listening the user's answers (Whisper STT) four times:	14.818
Time for listening the one user's answer (Whisper STT):	*3.705*
Time for sending "answer.txt" file to server four times:	23.403
Time for sending one "answer.txt" file to server:	*5.851*
Time for the model (LLaVA) to make a final decision:	2.0458
Time for generation of the suggestion (LLaVA):	6.916
Time for voicing the suggestion (Piper TTS):	21.0236
Overall time of the whole system:	**2 min 34 sec.± 12 sec.**

Chapter 5

Conclusion

In conclusion, the proposed work delves into the healthcare challenges posed by a growing population and increased healthcare needs, resulting in a shortage of timely assistance for people in emergencies. The suggested framework aims to provide a real-time system for continuous monitoring and user interaction by combining VLM, LLM, and TTS/STT as a novel approach to efficiently assist people in home accidents by detecting the emergency and providing contextually relevant information and suggestions to the caregivers. Although the LLaVA model sometimes answers incorrectly on the VQA, the chosen threshold for the questions allowed the system to achieve high accuracy, precision, specificity, recall and F1 score results before the interaction. The system also achieved perfect results in the final classification tasks for 2 and 8 classes, demonstrating robustness in understanding the situation using a multimodal context.

By leveraging these technologies, the framework enhances abnormal behaviour detection in real time, addressing limitations in traditional methods. Integrating YOLOv8 for image capture, vision models, and LLM (LLaVa) for context generation and real-time conversations ensures informed decision-making. The framework introduces a systematic emergency response, including thresholding based on collected responses from VQA, a generation of questions based on the context, a collection of user's answers, a final decision of the model based on the collected contextual data, and initiating ambulance alerts and suggestions when needed. Even in cases

of unresponsiveness, the system automatically calls for assistance as the person is unresponsive most of the time.

5.0.1 Implications

Implementing the proposed system offers practical benefits for the elderly, people with disabilities, people in danger, caregivers, and healthcare professionals. For the people, there is enhanced safety through real-time emergency response and improved well-being via interactive communication, while caregivers and professionals have less burden with automated insights and timely alerts, allowing them to focus on cases that require direct human intervention. This research has the potential to redefine people's care standards, contributing to real-world improvements in the medical field. Although the system is dedicated to the healthcare sector, this research has a broader application in the fields that require real-time monitoring and decision-making, such as security, smart-home systems and information retrieval systems.

5.0.2 Limitations

Despite the promising results of the system, several limitations need to be mentioned. Firstly, the model's ability to answer the questions from VQA needs refinement by finetuning the model with specific scenarios that showed an accuracy of less than 70% (head injury, open wound and broken leg). Also, the model needs to be finetuned to answer the questions related to subtler indicators (e.g., facial expressions of distress and discomfort). Secondly, the system had challenges in comprehending the images due to the low quality and distance of the camera. Thus, a camera that provides good quality even at a distance will improve the system's performance in answering questions from VQA. Thirdly, the model sometimes produces hallucinations when generating questions for the user, which can be improved via a more capable language model and with more accurate and enhanced visual and textual instructions for the model. Thirdly, the system's time performance is still limited due to the system's architecture (local and server part connection) and network latency, which

increases the system's response time. This can be resolved by edge computing, in which the models are set in the edge device, and data processing can be done locally. This enables a reduction in reliance on distant servers and minimization of delays. Also, the files and images can be compressed via compression techniques, reducing the size and bandwidth to send from one point to another and effectively speeding up the data transfer process. Fourthly, the Whisper ASR faces challenges in recognizing multiple speakers and producing erroneous texts. Although the small pretrained version of Whisper is fast, this ASR has reduced accuracy and efficiency in filtering out background noise and focusing on the speech content compared to larger versions of Whisper. Whisper ASR sometimes produced random words in the low-noise environment during the experiments. The reason for random word generation can be that Whisper may misinterpret background noise as speech components, especially if it has been trained on noisy data, leading to associations between specific noise patterns and words. Additionally, the processing of distorted audio features can confuse the model, causing the decoder to pick any recognizable speech patterns and output them as words.

5.0.3 Future Work

In future work, the system performance can be improved by incorporating video instead of images to understand the context and detect emergency cases. Secondly, the LLaVA model can be finetuned to answer the questions related to subtler indicators (facial expressions of distress and discomfort) to improve its ability to perform VQA tasks and detect initial signs of abnormalities. Also, finetuning the model with the emergency images taken far away from the camera can make the system more robust and accurate in detecting the person's facial expressions, movements, and parts of the body. Thirdly, using skeleton joints as the input for the LLaVA model can be beneficial as valuable insights into users' movements and postures can be obtained, resulting in a better understanding of the actions and postures of a person. Fourthly, integrating the model with emotion recognition through the voice tone can enhance emergency detection by inferring their emotional state and detecting indicators of

distress, pain, or calmness. This information can provide additional context for interpreting the severity of the situation and tailoring the response accordingly. Fourthly, the performance of Whisper ASR can be increased, especially in noisy environments, if better decoder techniques are integrated; this could address persistent errors, such as the model getting stuck in repeat loops or failing to transcribe the beginnings and ends of audio segments accurately. Also, fine-tuning the model on varied datasets with high-quality and poor-quality data can make the model more robust in noise handling. Furthermore, the hyperparameter tuning, such as the temperature parameter (this parameter controls the randomness of predictions by scaling the logits before applying the softmax function) in the decoding process can manage random word generation during silent or noisy segments. Implementing a dynamic temperature adjustment that varies with the model's confidence level or environmental noise can optimize output accuracy. Fifthly, integrating a quantized version of the LLaVA model and speech models on an edge device can significantly increase a system's time performance. Finally, the experiments with the proposed system need to be conducted in a natural home or healthcare setting to validate its effectiveness, usability, and practicality in real-world scenarios.

Bibliography

[1] Alya Al Rumhi, Huda Al Awisi, Mahmood Al Buwaiqi, and Salim Al Rabaani. Home accidents among children: a retrospective study at a tertiary care center in oman. *Oman medical journal*, 35(1):e85, 2020.

[2] Hamdi Aloulou, Mounir Mokhtari, and Bessam Abdulrazak. Pilot site deployment of an iot solution for older adults' early behavior change detection. *Sensors*, 20(7):1888, 2020.

[3] Miguel Ángel Antón, Joaquín Ordieres-Meré, Unai Saralegui, and Shengjing Sun. Non-invasive ambient intelligence in real life: Dealing with noisy patterns to help older people. *Sensors*, 19(14):3113, 2019.

[4] Damla Arifoglu and Abdelhamid Bouchachia. Activity recognition and abnormal behaviour detection with recurrent neural networks. *Procedia Computer Science*, 110:86–93, 2017.

[5] Batyr Arystanbekov, Askat Kuzdeuov, Shakhizat Nurgaliyev, and Huseyin Atakan Varol. Image captioning for the visually impaired and blind: A recipe for low-resource languages. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1–4. IEEE, 2023.

[6] Muhammad Awais, Lorenzo Chiari, Espen Alexander F Ihlen, Jorunn L Helbostad, and Luca Palmerini. Physical activity classification for elderly people in free-living conditions. *IEEE journal of biomedical and health informatics*, 23(1):197–207, 2018.

[7] Oresti Banos, Rafael Garcia, Juan A Holgado-Terriza, Miguel Damas, Hector Pomares, Ignacio Rojas, Alejandro Saez, and Claudia Villalonga. mhealthdroid: a novel framework for agile development of mobile health applications. In *International workshop on ambient assisted living*, pages 91–98. Springer, 2014.

[8] Mouazma Batool and Madiha Javeed. Fundamental recognition of adl assessments using machine learning engineering. In *2022 19th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*, pages 195–200. IEEE, 2022.

[9] Alisa Berger, Fabian Horst, Sophia Müller, Fabian Steinberg, and Michael Doppelmayr. Current state and future prospects of eeg and fnirs in robot-assisted gait rehabilitation: a brief review. *Frontiers in human neuroscience*, 13:172, 2019.

[10] Alan Kevin Bourke, Espen Alexander F Ihlen, Ronny Bergquist, Per Bendik Wik, Beatrix Vereijken, and Jorunn L Helbostad. A physical activity reference data-set recorded from older adults using body-worn inertial sensors and video technology—the adapt study data-set. *Sensors*, 17(3):559, 2017.

[11] Marco Buzzelli, Alessio Albé, and Gianluigi Ciocca. A vision-based system for monitoring elderly people at home. *Applied Sciences*, 10(1):374, 2020.

[12] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015.

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

[14] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3558–3568, 2021.

[15] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023.

[16] Md Jalal Uddin Chowdhury and Ashab Hussan. A review-based study on different text-to-speech technologies. *arXiv preprint arXiv:2312.11563*, 2023.

[17] Alessandra Corneli, Leonardo Binni, Berardo Naticchia, and Massimo Vaccarini. Digital twin models supporting cognitive buildings for ambient assisted living. In *International Conference on Technological Imagination in the Green and Digital Transition*, pages 167–178. Springer, 2022.

[18] The National Safety Council. Deaths in the Home: Introduction - Data Details - Injury Facts — injuryfacts.nsc.org. `https://injuryfacts.nsc.org/home-and-community/deaths-in-the-home/introduction/data-details/`, 2021. [Accessed 09-04-2024].

[19] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[20] Johanes Effendi, Sakriani Sakti, and Satoshi Nakamura. Weakly-supervised speech-to-text mapping with visually connected non-parallel speech-text data using cyclic partially-aligned transformer. In *Interspeech*, pages 2257–2261, 2021.

[21] Labiba Gillani Fahad, Syed Fahad Tahir, and Muttukrishnan Rajarajan. Activity recognition in smart homes using clustering based classification. In *2014 22nd International conference on pattern recognition*, pages 1348–1353. IEEE, 2014.

[22] Xiaohu Fan, Qubo Xie, Xuebin Li, Hao Huang, Jian Wang, Si Chen, Changsheng Xie, and Jiajing Chen. Activity recognition as a service for smart home: ambient assisted living application via sensing home. In *2017 IEEE International Conference on AI & Mobile Services (AIMS)*, pages 54–61. IEEE, 2017.

[23] Hemant Ghayvat, Muhammad Awais, Sharnil Pandya, Hao Ren, Saeed Akbarzadeh, Subhas Chandra Mukhopadhyay, Chen Chen, Prosanta Gope, Arpita Chouhan, and Wei Chen. Smart aging system: uncovering the hidden wellness parameter for well-being monitoring and anomaly detection. *Sensors*, 19(4):766, 2019.

[24] Munkhjargal Gochoo, Tan-Hsu Tan, Shing-Hong Liu, Fu-Rong Jean, Fady S Alnajjar, and Shih-Chia Huang. 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.

[25] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 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*, pages 6904–6913, 2017.

[26] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018.

[27] Michael Hansen. Rhasspy/piper: A fast, local neural text to speech system.

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

[29] Drew A Hudson and Christopher D Manning. 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*, pages 6700–6709, 2019.

[30] Rachneet Kaur, Robert W Motl, Richard Sowers, and Manuel E Hernandez. 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.

[31] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

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

[33] Tushar Khot, Ashish Sabharwal, and Peter Clark. Scitail: A textual entailment dataset from science question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[34] Kibum Kim, Ahmad Jalal, and Maria Mahmood. 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.

[35] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

[36] Kwok-Yan Lam, Victor CW Cheng, and Zee Kin Yeong. Applying large language models for enhancing contract drafting. 2023.

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

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

[39] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.

[40] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[41] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.

[42] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[43] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.

[44] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[45] Shikun Liu, Linxi Fan, Edward Johns, Zhiding Yu, Chaowei Xiao, and Anima Anandkumar. Prismer: A vision-language model with an ensemble of experts. *arXiv preprint arXiv:2303.02506*, 2023.

[46] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023.

[47] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

[48] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023.

[49] Thomas W MacFarland, Jan M Yates, Thomas W MacFarland, and Jan M Yates. Mann–whitney u test. *Introduction to nonparametric statistics for the biological sciences using R*, pages 103–132, 2016.

[50] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE, 2019.

[51] Aurore Patricia Mroz. 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*, page 147, 2022.

[52] Adnan Nadeem, Amir Mehmood, and Kashif Rizwan. A dataset build using wearable inertial measurement and ecg sensors for activity recognition, fall detection and basic heart anomaly detection system. *Data in brief*, 27:104717, 2019.

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

[54] Ehsan Nazerfard, Zahra Atashgahy, and Alireza Nadali. Abnormal activity detection for the elderly people using convlstm autoencoder. 2021.

[55] World Health Organization. Disability — who.int. https://www.who.int/news-room/fact-sheets/detail/disability-and-health#:~:text=Key%20facts,earlier%20than%20those%20without%20disabilities., 2023. [Accessed 10-04-2024].

[56] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE, 2015.

[57] Prateek Pandey and Ratnesh Litoriya. Elderly care through unusual behavior detection: a disaster management approach using iot and intelligence. *IBM Journal of Research and Development*, 64(1/2):15–1, 2019.

[58] Ashish Patel and Jigarkumar Shah. Real-time human behaviour monitoring using hybrid ambient assisted living framework. *Journal of Reliable Intelligent Environments*, 6(2):95–106, 2020.

[59] NN Petrova and DA Khvostikova. Prevalence, structure, and risk factors for mental disorders in older people. *Advances in Gerontology*, 11:409–415, 2021.

[60] Paola Pierleoni, Alberto Belli, Andrea Gentili, Lorenzo Incipini, Lorenzo Palma, Sara Raggiunto, Agnese Sbrollini, and Laura Burattini. Real-time smart monitoring system for atrial fibrillation pathology. *Journal of Ambient Intelligence and Humanized Computing*, 12:4461–4469, 2021.

[61] Manoharan Poongodi, Ashutosh Sharma, Mounir Hamdi, Ma Maode, and Naveen Chilamkurti. Smart healthcare in smart cities: wireless patient monitoring system using iot. *The Journal of Supercomputing*, pages 1–26, 2021.

[62] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[63] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023.

[64] Heilym Ramirez, Sergio A Velastin, Sara Cuellar, Ernesto Fabregas, and Gonzalo Farias. Bert for activity recognition using sequences of skeleton features and data augmentation with gan. *Sensors*, 23(3):1400, 2023.

[65] V Madhusudhana Reddy, T Vaishnavi, and K Pavan Kumar. Speech-to-text and text-to-speech recognition using deep learning. In *2023 2nd International Conference on Edge Computing and Applications (ICECAA)*, pages 657–666. IEEE, 2023.

[66] Rahim Sadigov et al. Rapid growth of the world population and its socioeconomic results. *The Scientific World Journal*, 2022, 2022.

[67] Joseph Santiago, Eric Cotto, Luis G Jaimes, and Idalides Vergara-Laurens. Fall detection system for the elderly. In *2017 IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 1–4. IEEE, 2017.

[68] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[69] Suraj S Senjam and Hasheem Mannan. Assistive technology: The current perspective in india. *Indian Journal of Ophthalmology*, 71(5):1804–1809, 2023.

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

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

[72] Arijit Ukil and Soma Bandyopadhyay. Automated cardiac health screening using smartphone and wearable sensors through anomaly analytics. *Mobile Solutions and Their Usefulness in Everyday Life*, pages 145–172, 2019.

[73] Minh-Hao Van, Prateek Verma, and Xintao Wu. On large visual language models for medical imaging analysis: An empirical study. *arXiv preprint arXiv:2402.14162*, 2024.

[74] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[75] Tomás F Yago Vicente, Le Hou, Chen-Ping Yu, Minh Hoai, and Dimitris Samaras. Large-scale training of shadow detectors with noisily-annotated shadow examples. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, pages 816–832. Springer, 2016.

[76] Yan Wang, Xin Wang, Damla Arifoglu, Chenggang Lu, Abdelhamid Bouchachia, Yingrui Geng, and Ge Zheng. A survey on ambient sensor-based abnormal behaviour detection for elderly people in healthcare. *Electronics*, 12(7):1539, 2023.

[77] Rhydian Windsor, Amir Jamaludin, Timor Kadir, and Andrew Zisserman. Vision-language modelling for radiological imaging and reports in the low data regime. *arXiv preprint arXiv:2303.17644*, 2023.

[78] Wenhao Wu, Xiaohan Wang, Haipeng Luo, Jingdong Wang, Yi Yang, and Wanli Ouyang. 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*, pages 6620–6630, 2023.

[79] Zhengyuan Yang, Yijuan Lu, Jianfeng Wang, Xi Yin, Dinei Florencio, Lijuan Wang, Cha Zhang, Lei Zhang, and Jiebo Luo. Tap: Text-aware pre-training for text-vqa and text-caption. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8751–8761, 2021.

[80] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023.

[81] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

[82] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. 2023.

[83] Lei Zhu, Zijun Deng, Xiaowei Hu, Chi-Wing Fu, Xuemiao Xu, Jing Qin, and Pheng-Ann Heng. Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 121–136, 2018.