Toward Abnormal Activity Recognition of Developmentally Disabled Individuals Using Pose Estimation

Taihei Fujioka* [1], Christina Garcia [2], Sozo Inoue [3]
[128]Kyushu Institute of Technology,

Abstract

In this study, we propose to optimize temporal parameters with pose estimation data of simulated abnormal activities of developmentally disabled individuals by incorporating behavior context to Large Language Models (LLMs). Facilities for the developmentally disabled face the challenge of detecting abnormal behaviors because of limited staff and the difficulty of spotting subtle movements. Traditional methods often struggle to identify these behaviors because abnormal actions are irregular and unpredictable, leading to frequent misses or misclassifications. The main contributions of this work is the creation of a unique dataset with labeled abnormal behaviors and the proposed application of LLMs to this dataset comparing results of Zero-Shot and Few-Shot. Our method leverages the context of the collected abnormal activity data to prompt LLMs to suggest window size, overlap rate, and LSTM model's length sequence tailored to the specific characteristics of these activities. The dataset includes labeled video data collected for four days from five normal participants performing eight activities with four abnormal behaviors. The data was collected with normal participants to simulate activities, and no individuals with disabilities. For evaluation, we assessed all normal versus abnormal activities and per abnormal activity recognition comparing with the baseline without LLM. The results showed that Few-Shot prompting delivered the best performance, with F1-score improvements of 7.69% for throwing things, 7.31% for attacking, 4.68% for head banging, and 1.24% for nail biting as compared to the baseline. Zero-Shot prompting also demonstrated strong recognition capabilities, achieving F1 scores above 96% across all abnormal behaviors. By using LLM-driven suggestions with YOLOv7 pose data, we optimize temporal parameters, enhancing sensitivity to abnormal behaviors and generalization across activities. The model reliably identifies short, complex behaviors, making it ideal for real-world caregiving applications.

[1]fujioka.taihei250@mail.kyutech.jp
[2]alvarez7.christina@gmail.com
[3]sozo@brain.kyutech.ac.jp

Keywords: Large Language Models, abnormal behaviour recognition, pose estimation, temporal

1 Introduction

The shortage of staff in facilities for persons with disabilities is a critical issue that significantly impacts the quality of facility operations and service delivery [1, 2]. According to a survey conducted by the Welfare and Medical Service Agency (WAM), as of 2023, 52.6% of facilities reported experiencing a shortage of staff, an increase compared to the 2020 survey [3]. Individuals with intellectual disabilities or severe mental illnesses face challenges such as decreased quality of care and compromised safety when proper diagnosis and care are not provided. Issues such as staff diagnostic capabilities and the fairness of care are particularly prominent, requiring multidisciplinary collaboration and the adoption of new technologies to address these challenges [4].

Currently, many facilities for persons with disabilities rely on direct observation and manual recording processes, which increase staff workload and make it difficult to quickly and accurately identify abnormal behaviors. However, manual recording has limitations, making it challenging to continuously collect sufficient data. In addition, data collection itself is performed by a system, meaning that certain challenges arise independently of human involvement. For example, system interruptions, mislabeling, and improper timestamps can limit the amount of collected data [6, 7]. To overcome these data collection challenges and enhance the accuracy of abnormal behavior detection, video-based human activity recognition technology has been studied as a crucial approach [5, 12]. While this technology holds promise for the early detection of abnormal behaviors and distinguishing them from daily activities, challenges remain in recognizing complex and atypical behaviors demonstrated by users. The ambiguous definition of human-perceivable abnormality limits accurate video detection [23]. Addressing these challenges requires the development of new algorithms that consider the context of behaviors and datasets specifically designed for abnormal behavior recognition [12].

Although the methodology of this study is based on the facility environment in Japan, abnormal behavior recognition using pose estimation is considered applicable on a global scale. Previous studies have demonstrated that cultural backgrounds influence the behavioral characteristics and support strategies for individuals with disabilities [29]. For example, Mori et al. (2021) investigated the role of cultural factors in the behavioral characteristics of individuals with developmental disabilities and emphasized that different cultural settings influence both the understanding of behaviors and the support strategies provided for individuals with disabilities. By incorporating such contextual knowledge into pose-based abnormal behavior recognition, our method has the potential to be adapted for use in different countries and cultural settings.

Among the recent approaches to complex activity recognition is decomposing classes into temporal sequences or adjusting the timestamps for improve labeling

[23, 24]. While sequential timestamp extensions can improve labeling accuracy, addressing abrupt and unpredictable abnormal behaviors remains challenging.

Recently, large language models have increased in application including in activity recognition [11]. As complex activities tend to have limited data due to their nature, another solution by Dobhal et. al. utilizing LLM [25] is generating synthetic data to improve performance with pose data incorporating general context from the activities. However, there are differences in the characteristics especially of unprecedented abnormal behavior. This study proposes generating temporal parameters by prompting LLMs, leveraging context and example data from each abnormal activity.

We propose addressing the differences between brief abnormal behaviors and prolonged normal behaviors by using prompting techniques to generate temporal parameters, comparing Zero-Shot and Few-Shot approaches. The system utilizes models like LSTM with adjusted window sizes and sequence lengths to classify behaviors while considering contextual information. Furthermore, we apply LLM to the unique dataset to generate window size, overlap rate, and length sequence with the aim of constructing accurate recognition model that incorporates the contextual nuances of movements. Specifically, the contributions of this paper are:

- Collection of simulated abnormal activities of developmentally disabled individuals, performed by normal participants with extracted pose data from video.
- Generating temporal parameters from LLM by leveraging context and information of each abnormal activity comparing Zero-Shot and Few-Shot.
- Investigating the application of language models towards recognizing abnormal activities of developmentally disabled individuals. While we did not develop a new pose estimation algorithm, we are the first to apply LLMs for optimizing temporal parameters tailored to this dataset. Additionally, our work extends beyond comparing prompting techniques, which also successfully recognized abnormal behavior.

The proposed system showed significant improvements in recognizing abnormal behaviors. The results demonstrated highest performance with Few-Shot prompting where the model achieved F1-scores improvement of 1.24% for biting nails, 4.68% for head banging, 7.69% for throwing things, and 7.31% for attacking compared with baseline. Zero-Shot prompting similarly demonstrated robust recognition, achieving F1 scores exceeding 96% for all abnormal behaviors. Overall, the findings suggest the potential of this system to help caregivers efficiently identify abnormal behaviors in real-world settings.

The rest of the paper is structured as follows: Section 2 discusses the related work in this field comparing the existing methods on abnormal activity recognition with pose estimation data including the applications of LLM. Section 3 describes the collection of simulated recorded abnormal activities and Section

4 details the propose recognition approach incorporating LLM for temporal parameters. Section 5 and 6 covers the performance and evaluation of resulting parameters per activity, concluding with future work in Section 7.

2 Related literature

This section reviews the existing works highlighting the gaps that inspire our proposed framework for recognizing abnormal behavior. We examine general approaches in computer vision in identifying abnormal behavior, challenges in video-based pose estimation, and the importance of time window optimization for improving recognition accuracy. We also explore the potential of Large Language Models (LLMs) in Human Activity Recognition (HAR) and their limitations. Finally, we present our framework as a solution to address the challenges by integrating LLM-suggested temporal parameters.

2.1 Abnormal Behavior Recognition

Computer vision and deep learning have been widely used in the field of human activity recognition [13]. In the area of abnormal behavior recognition, research has been conducted to identify actions and behavioral patterns in large-scale crowds and facility environments. For example, a new dataset named HAJJv2 proposed in [14] includes manually annotated abnormal behaviors and employs a method combining CNN and Random Forest (RF). Specifically, ResNet-50 is used for small-scale crowds, and YOLOv2 is utilized for large-scale crowds, achieving abnormal behavior recognition tailored to the characteristics of each dataset. Small-scale crowds are defined as groups consisting of tens of individuals, while large-scale crowds refer to environments where hundreds or thousands of individuals are present, leading to challenges such as occlusion, blurring, and increased movement variability. This study achieved an AUC of 76.08%, providing a new foundation for abnormal behavior recognition. However, this approach requires switching models based on crowd size, which limits its adaptability for individual-level or complex behavior recognition. In contrast, our framework focuses on utilizing skeleton-based pose estimation to achieve fine-grained individual recognition and improved adaptability across diverse environments.

Furthermore, improvements in deep learning models for abnormal behavior recognition have garnered attention. In [15], a method combining an improved ResNet model and YOLOv3 was proposed, achieving high accuracy using the UTI dataset. This approach enables fast and high-precision real-time abnormal behavior recognition, contributing to enhanced recognition accuracy. However, as this method relies on RGB video data, it is sensitive to the visual characteristics of the environment, which may limit its generalizability. The HAJJv2 dataset initially employed YOLOv2 for large-scale crowd detection due to its balance between accuracy and computational efficiency. However, later studies adopted YOLOv3 to improve object detection in densely packed environments, allowing for more precise identification of overlapping individuals and enhancing

abnormal behavior recognition. In contrast, our approach reduces this dependency by leveraging skeleton-based features, enabling robust recognition across various conditions.

2.2 Video and Pose Estimation in Activity Recognition

In abnormal behavior recognition, methods utilizing RGB video and skeleton information are commonly used. Skeleton-based pose estimation is simpler and more effective than video data; however, LSTM models have limitations in extracting spatial information, posing challenges in dimensionality reduction and time-series data analysis [16]. Additionally, data imbalance caused by the operating environment and the physical characteristics of workers is a critical issue that affects model accuracy [16]. Skeleton-based approaches have been shown to be less affected by lighting conditions and background variations, making them more adaptable to environmental changes compared to RGB-based methods [26]. Furthermore, skeleton data does not rely on appearance information, reducing sensitivity to camera angles and clothing variations, which makes it more stable for action recognition than convolutional neural networks (CNNs) or optical flow-based methods [27]. Based on these findings, skeleton-based features are considered suitable for abnormal behavior recognition in this study.

Therefore, [17] proposed an approach that combines Spatio-Temporal Graph Convolutional Networks (ST-GCN) and a Normalized Attention Mechanism (NAM) to effectively model the spatio-temporal relationships in skeleton data, improving feature extraction accuracy for abnormal behavior recognition. However, limitations remain in recognizing short-duration or contextually ambiguous behaviors. Our research aims to address these gaps by designing parameters for the model tailored to such behaviors.

2.3 Optimizing Time Windows with Abnormal Behavior

The setting of time windows is a crucial factor in time-series data analysis, and its optimization poses challenges, particularly in abnormal behavior recognition using LSTM models. A study using EEG data [18] proposed a method that evaluates various time window sizes with LSTM to select the optimal time window that maximizes classification accuracy. This research highlights that the selection of time window size heavily depends on the characteristics of the data, suggesting the necessity of dynamically choosing optimal time windows based on the target actions and data distribution in abnormal behavior recognition. Additionally, other studies have proposed dynamic sliding window methods, demonstrating the effectiveness of data analysis that considers periodicity and long-term dependencies [19]. While these findings underline the importance of time window optimization, their real-world applicability remains underexplored.

2.4 Application of Large Language Models in HAR

The application of Large Language Models (LLMs) in Human Activity Recognition (HAR) has shown significant advancements in recent years. Unlike tradi-

tional approaches, LLMs can process raw sensor and time-series data directly, enabling efficient and precise activity classification through zero-shot learning and prompt engineering.

Ji et al. [20] introduced HARGPT, utilizing GPT-4 to process accelerometer and gyroscope data without the need for extensive training data. Their method achieved over 80% classification accuracy on datasets such as Capture24 and HHAR, demonstrating the potential of zero-shot learning in HAR. In comparison, the TCN-attention-HAR model proposed by Shao et al. [28] reported a classification accuracy of 95.69% on the UCI HAR dataset using handcrafted features and time-series convolution networks.

Additionally, Shoumi and Inoue [11] emphasized the role of Chain-of-Thought (CoT) prompting in improving the transparency and interpretability of HAR models. By sequentially explaining the reasoning process, CoT prompting allows users to better understand model predictions, enhancing trust and usability.

Furthermore, LLMs have proven effective in integrating multimodal data by converting sensor information into natural language representations [16]. This approach eliminates the need for manual feature engineering, enabling flexible classification of both abnormal and daily activities. Specifically, Shoumi and Wei [21] highlighted the effectiveness of LLMs in recognizing complex movement patterns and contextual information, leading to highly accurate classifications. Despite these advancements, challenges remain, including the computational cost of LLMs and their dependency on data format consistency. Future research must address these issues by optimizing prompt designs and improving the efficiency of LLM-based models.

Our research builds on these insights by integrating dynamic time-window mechanisms optimized for skeleton-based features, ensuring precise recognition of short-duration and complex abnormal behaviors. We collected simulated abnormal activities from normal participants and utilized YOLOv7 to extract pose data from video. We selected YOLOv7 due to its superior balance of accuracy and speed in real-time detection, and outperforming models like YOLOR, YOLOX, YOLOv5, and DETR across various FPS ranges [22]. We compared and evaluated generated temporal parameters from Zero-Shot and Few-Shot optimizing context and example from our collected data.

3 Dataset from Simulated Activities with Normal Subjects

In this section, we elaborate on the data collection process, including the pose estimation method applied to the recorded video capturing abnormal and normal activities.

3.1 Data Collection in the Laboratory

The dataset was collected in a laboratory-controlled environment with a total of five participants in a span of four days [30]. In this study, participants' actions

were recorded using two different cameras, and pose estimation techniques were used to extract skeleton data. The target actions included four types of abnormal behaviors such as "throwing objects" and "hitting the head," as well as four types of normal behaviors such as "walking" and "sitting," as shown in Table 1. The selection of these activities was based on observations and interviews conducted in a facility for individuals with developmental disabilities. Facility staff identified these behaviors as frequently occurring and requiring attention to ensure the safety and well-being of individuals. The abnormal activities were chosen to reflect patterns of self-injurious or aggressive behavior commonly observed in the facility, including "head banging", "throwing objects", "attacking others", and "biting hands or fingers". Similarly, the normal activities represent routine daily actions observed in the same environment, such as "sitting," "walking," "using a phone," and "eating or drinking." This selection ensures that the dataset aligns with real-world behavioral patterns and supports practical application in care settings.

Table 1: Target Activities and Behavior Types

Activity	**Behavior Type**
Sitting quietly	Normal
Using phone	Normal
Walking	Normal
Eating snacks	Normal
Head banging	Abnormal
Throwing things	Abnormal
Attacking	Abnormal
Biting nails	Abnormal

A planned data collection procedure was implemented with time restrictions for each action. To conduct an experiment on abnormal behavior recognition representing facilities for individuals with developmental disabilities, we recruited five adult participants. Participants were selected based on the criteria of being healthy adults capable of accurately reproducing the required actions. From an ethical standpoint, all participants were provided with a Certificate of Participation to acknowledge their contribution to the study, and light refreshments were offered at the end of the experiment. The participants were five graduate students aged 20 to 30 years old. Each session lasted 60 minutes, with a total of 4 sessions, amounting to 240 minutes of recording.

Wide-angle data was captured using the GoPro 9, while detailed data was collected using the iPhone 15 Pro, enabling multi-angle data acquisition. In this study, keypoints for both abnormal and normal behaviors were extracted from the recorded video data and labeled manually. The labeling process was conducted using Vrew, with each frame being annotated manually to ensure accuracy and consistency. Skeleton data included key points such as the head, shoulders, wrists, and knees, and each action was labeled.

Considering the feasibility of this study, the necessity of installing new cam-

eras depends on whether the existing cameras are appropriately positioned and capable of capturing the required data. If the existing cameras provide sufficient data, additional installations may not be required; however, privacy concerns must be addressed through appropriate anonymization and data management measures.

Each participant underwent a one-hour session, during which they performed eight types of activities (four normal activities and four abnormal activities) within a specified timeframe. The experiment was divided into two 30-minute sessions, with a five-minute break between them. It is important to note that this study did not involve actual subjects, such as individuals with disabilities or specific conditions. Instead, the research was conducted using normal participants to simulate abnormal activities of developmentally disabled individuals. The collected data was used to build an action recognition model and improve classification accuracy. The dataset was utilized for the validation and optimization of an abnormal behavior recognition system.

3.2 Action Distribution and Pose Estimation

In this study, participant movements were recorded at a sampling rate of 30 frames per second using GoPro 9 and iPhone 15 Pro cameras. Pose estimation techniques were applied to extract skeleton data from the recorded video. The data was processed using YOLOv7, which generated images and extracted keypoints from nine specific body parts: the nose, eyes, ears, shoulders, elbows, wrists, hips, knees, and ankles. This approach ensured accurate and detailed motion capture for subsequent analysis.

Specifically, the 17 keypoints extracted from YOLOv7 include the nose, left eye, right eye, left ear, right ear, left shoulder, right shoulder, left elbow, right elbow, left wrist, right wrist, left hip, right hip, left knee, right knee, left ankle, and right ankle. Figure 5 illustrates examples of abnormal activities performed by participants as captured during the study. These activities are visually represented to highlight the distinct motion patterns associated with each behavior. Subfigure (a) depicts the activity "Attacking," characterized by short, rapid, and dynamic movements. Subfigure (b) shows "Nail biting," a repetitive and rhythmic motion. Subfigure (c) illustrates "Head banging," which involves rhythmic and forceful head movements. Finally, subfigure (d) represents "Throwing things," characterized by short, discrete, and dynamic arm and hand gestures.

From these results, we observe that different types of abnormal activities exhibit distinct movement characteristics, which impact recognition performance. High-energy activities such as "Attacking" and "Throwing things" involve large, dynamic motions that are more easily captured by pose estimation models. In contrast, behaviors like "Nail biting" and "Head banging" involve subtle, repetitive movements primarily in smaller joints such as the fingers or head, making them more challenging to detect accurately. These observations suggest that pose-based recognition models must account for variations in movement intensity and scale. Future improvements may involve developing weighting mechanisms that prioritize certain keypoints based on the movement patterns

Figure 1: a. Activity Attacking

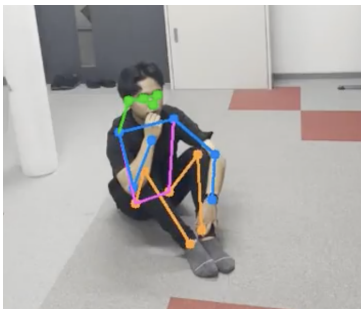


Figure 2: b. Activity Nail biting

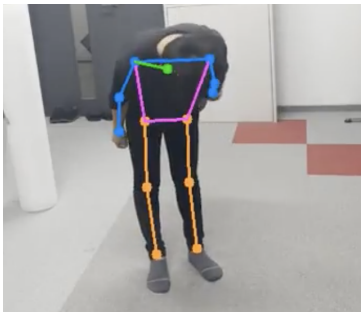


Figure 3: c. Activity Head banging

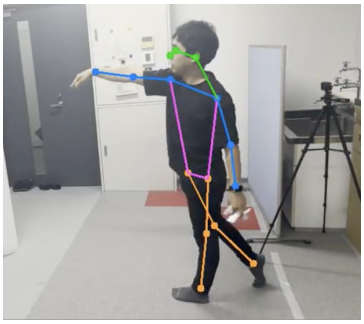


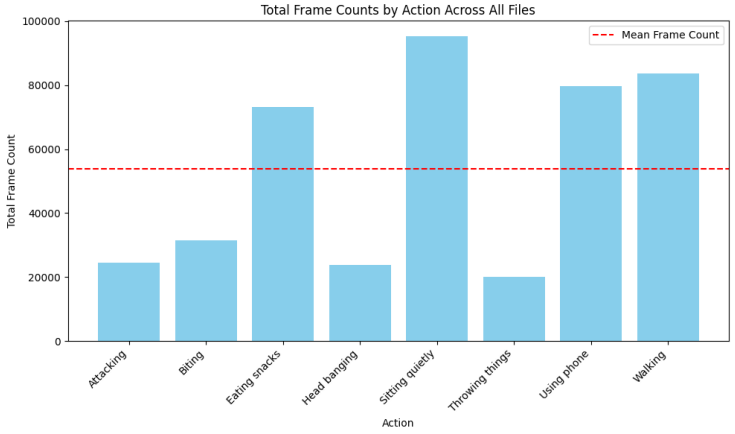
Figure 4: d. Activity Throwing things

Figure 5: Abnormal Activities performed by Normal Participants

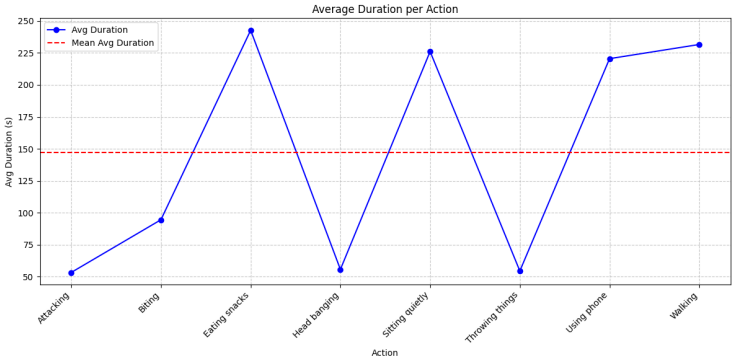
of different activities. Additionally, integrating temporal context by analyzing motion trajectories over multiple frames could further enhance recognition accuracy, especially for repetitive behaviors.

The total frame counts for each action category are visualized in Figure 6. This figure highlights the distribution of data across the eight action categories, including both abnormal and normal behaviors. Understanding this distribution is critical for assessing the balance of the dataset and identifying potential biases that could impact the performance of the recognition model. For instance, an imbalance in the number of frames for different actions could lead to overfitting or underrepresentation of certain categories during training, which would subsequently affect the generalizability of the model.

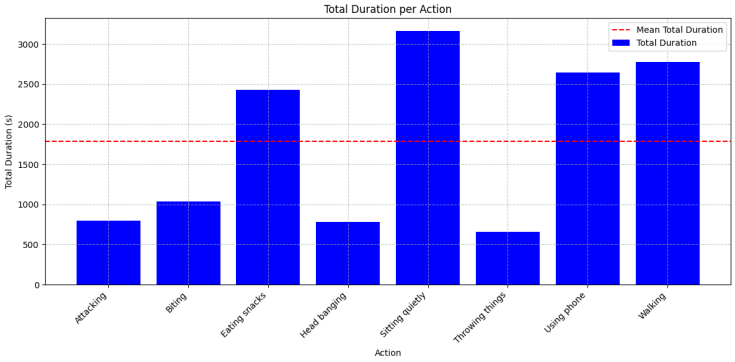
Figure 6 (a) presents the total frame counts for each action category. It is evident that normal actions, such as "Sitting quietly" and "Walking," have significantly higher frame counts compared to others, such as "Throwing things" and "Attacking." This discrepancy indicates that normal behaviors tend to dominate the dataset in terms of duration. The red dashed line representing the mean frame count serves as a reference point, making it easier to visualize which



(a) Frame counts by action category across all files.



(b) Average duration by action category across all files.



(c) Total duration by action category across all files.

Figure 6: Data Distribution and Duration per Action Category

categories are above or below the average distribution.

Figure 6 (b) depicts the average duration per action category, offering further insights into the dataset composition. For example, actions like "Eating snacks" and "Sitting quietly" have much longer average durations compared to "Attacking" or "Throwing things," which are shorter and more abrupt behaviors. These differences emphasize the variability in temporal patterns across the dataset, which the model must account for during training and evaluation.

Figure 6 (c) illustrates the total duration per action category across all files. As expected, normal actions such as "sitting quietly" and "walking" have the highest total duration, while abnormal behaviors like "throwing things" and "attacking" have significantly shorter durations. This distribution is intentional, as abnormal behaviors naturally occur less frequently in a typical day. The red dashed line represents the mean total duration, serving as a reference for comparison. This data composition ensures that the dataset realistically reflects the distribution of daily activities, which is essential for training a model capable of detecting rare but significant abnormal behaviors.

These results indicate that both the total frame count and the duration of each action significantly impact the learning process of the recognition model. The overrepresentation of normal activities could lead to a bias where the model favors detecting common behaviors while struggling to correctly classify rarer abnormal activities. Additionally, the variation in action duration suggests that the model needs to be trained with a time-sensitive approach, incorporating mechanisms such as adaptive windowing or weighted sampling to ensure balanced representation. Addressing these factors is crucial for improving the robustness and fairness of the recognition system.

4 Proposed Method for LLM-guided Abnormal Activity Recognition

Figure 7 shows the proposed workflow for abnormal activity recognition using pose estimation and LLM-guided temporal parameters. Abnormal activities are recorded using a front camera, and YOLOv7 Pose Estimation extracts skeletal keypoints from the video. After keypoints are extracted using YOLOv7 Pose Estimation, a pre-processing step is applied to standardize the skeletal data for consistency and accuracy. To ensure a stable reference throughout the video, the target individual is tracked across frames. To remove redundant or low-confidence detections, non-maximum suppression (NMS) is applied, retaining only the most reliable keypoints. The processed keypoints are then stored in a structured format, preserving temporal coherence across frames for later feature extraction. These steps ensure that the keypoint data is clean, consistent, and suitable for abnormal activity recognition. These keypoints are pre-processed, and a Large Language Model (LLM) optimizes temporal parameters including window size, overlap rate, and length sequence

Since the dataset contains sequential motion data, we employ an LSTM

model to capture temporal dependencies effectively. LSTMs retain important information over long sequences by using gating mechanisms while mitigating the vanishing gradient problem [23]. Compared to attention-based models like Transformers, LSTMs require fewer computational resources and are less prone to overfitting with small datasets [24]. Additionally, LLMs assist in parameter tuning by suggesting optimal time-window sizes, reducing the need for extensive manual searches [25].

The optimized data undergoes feature extraction and is fed into an LSTM model designed to learn and classify activity patterns. The system identifies abnormal behaviors, and its performance is evaluated for accuracy and reliability. Performance with temporal parameters from different prompting techniques are then compared.

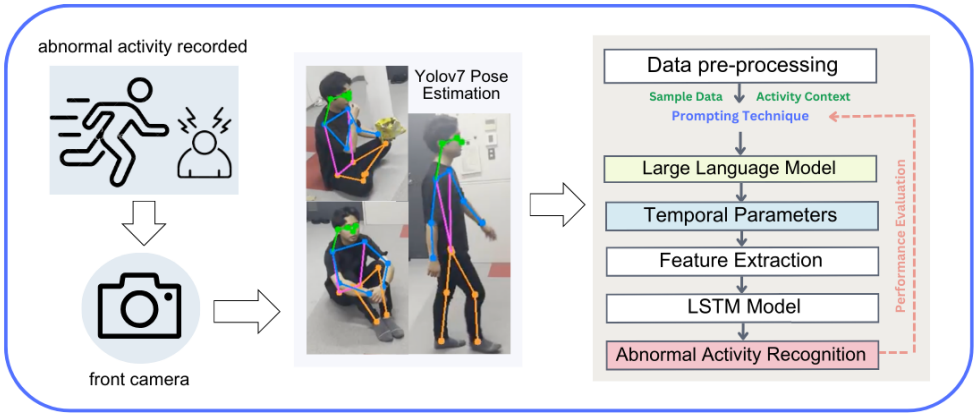


Figure 7: Overview of Abnormal Behavior Recognition Framework

4.1 Prompting Techniques for Temporal Parameters

The comparison illustrated in Figure 8 showcases the prompting strategies for Zero-Shot and Few-Shot learning used to generate temporal parameters for recognizing abnormal activities.

In general, prompting the LLM to generate temporal parameters can be expressed mathematically as Eq. 1, where T represents the temporal parameters elaborated in Eq. 2, and P is the prompting approach.

$$T = \arg\max_{T} p_G(T \mid P) \tag{1}$$

$$T = \{\text{window size, overlap rate, length sequence}\} \tag{2}$$

The type of input prompt P depends on the prompting method. In Zero-Shot prompting, the input to P_{ZeroShot} includes activity description, baseline parameters, activity context, and optimization goal, as defined in Equation 3.

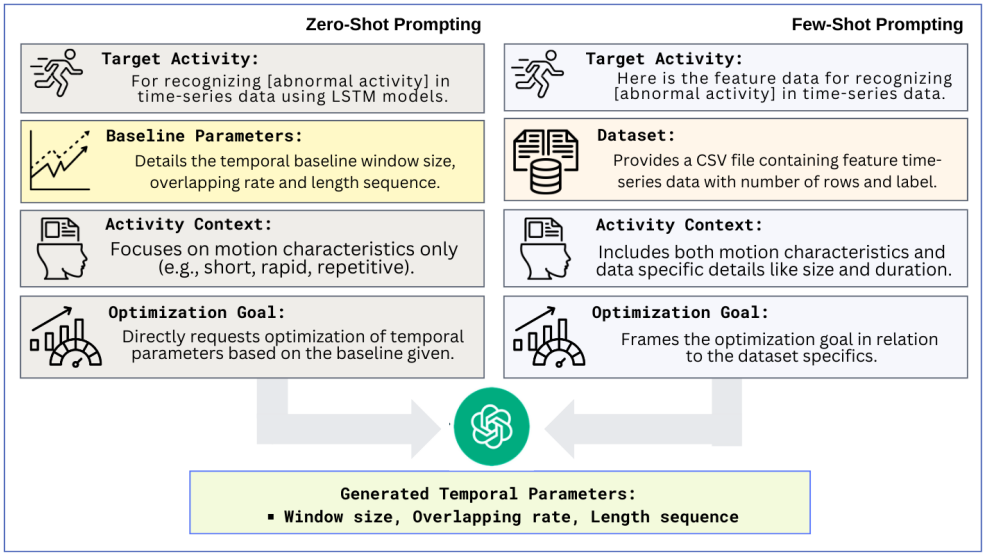


Figure 8: Prompt Templates for Temporal Parameter Optimization

In Few-Shot prompting, P_{FewShot} expands on this by incorporating dataset-specific examples, as described in Equation 4.

$$P_{\text{ZeroShot}} = \{A, B, C, G\} \tag{3}$$

$$P_{\text{FewShot}} = \{D, A, C, G\} \tag{4}$$

By following these formulations, LLMs adaptively generate temporal parameters tailored to the characteristics of activities and datasets, which we take advantage to have targeted model catering to the data.

The use of LLMs for temporal parameter optimization was chosen because they provide a flexible and efficient way to determine parameters that are challenging to optimize using traditional methods like Grid or Random Search. The ability to incorporate activity-specific motion characteristics through natural language prompts enables fine-tuned and context-aware optimization

In Zero-Shot prompting, the focus is on providing baseline parameters, activity context, and optimization goals directly to the LLM without prior dataset knowledge. It emphasizes characteristics like short, rapid, or repetitive motion to guide the optimization of temporal parameters such as window size, overlap rate, and sequence length. The implemented template for Zero-Shot is:

- *"Target Activity (A)":* Specify the target abnormal activity to be recognized and specific recognition model.
- *"Baseline Parameters (B) ":* Provide the initial configuration for temporal parameters, including window size, overlap, and sequence length. These serve as the starting point for optimization.

- *"Activity Context (C)":* Describe the unique characteristics of the target activity in motion type, repetition or rhythm to guide the model toward activity-specific optimization. Attacking focus on short, rapid, and discrete movements. Nail biting focus on repetitive and rhythmic patterns. Head banging focus on rhythmic, repetitive head movements. Throwing things focus on short, dynamic movements with distinct arm/hand gestures.
- *"Optimization Goal (G)":* Define the objective, such as improving recognition performance for the activity while maintaining computational efficiency.

In contrast, Few-Shot prompting incorporates specific dataset features, including a CSV file with labeled time-series data, allowing the LLM to frame its optimization objectives in the context of dataset-specific details. Both approaches leverage the LLM's ability to derive activity-specific temporal parameters efficiently, ensuring precise recognition of abnormal activities while considering computational efficiency. The template for Few-Shot is as follows:

- *"Target Activity (A)":* Similarly, this part specify the target abnormal activity.
- *"Dataset (D)":* Details about the dataset used for analysis including example data with information on the file type, number of data points, and action labels.
- *"Activity Context (C)":* This includes both motion characteristics and dataset-specific details on total duration of the action across all videos.
- *"Optimization Goal (G)":* Frames the optimization goal in relation to dataset specifics, balancing recognition accuracy with efficiency.

This study used GPT-4-turbo as the LLM to generate temporal parameters for abnormal activity recognition. First, the temperature was set to 0.7 to balance diversity and consistency in outputs. Preliminary testing showed that values below 0.7 resulted in overly uniform outputs, while values above 0.8 led to unstable generation. Next, top-p was set to 0.9 to constrain token selection to probabilistically significant choices, thereby reducing the risk of selecting extremely unlikely tokens while maintaining generation flexibility. Additionally, the maximum token limit was set to 25,000 to ensure sufficient contextual information and detailed descriptions of abnormal activities. Lower token limits resulted in insufficient context, making it difficult to generate the long-form outputs required for behavior analysis. A limit of 25,000 was found to provide an optimal balance between information retention and resource efficiency. Furthermore, frequency penalty and presence penalty were set to 0 to prevent the LLM from unnecessarily repeating specific expressions or introducing unrelated new topics. Pre-experimental evaluations indicated that increasing or decreasing these values led to inconsistencies in the definition of abnormal activities.

Thus, setting both parameters to 0 was determined to be the most effective approach.

4.2 Features Extracted from 2D Pose Estimation

From the original keypoints, the following are used for feature extraction: right hand, left hand, right shoulder, left shoulder, right eye, left eye, right foot, and left foot. Using these keypoints, 14 features were calculated from the combination keypoints from targeted body parts relevant to the activities in the datset. The extracted features from pose data include various motion and spatial metrics to enhance abnormal activity recognition.

The use of 2D pose estimation was chosen for its computational efficiency and resilience to noise compared to raw RGB video or depth-based methods. Skeleton data provide a concise representation of motion, enabling the model to focus on the dynamic and spatial aspects of actions without being influenced by visual or environmental artifacts.

Speed and acceleration are captured for key body parts, including the right and left hands and feet, measuring both speed (v, Equation 5) and acceleration (a, Equation 6) during movements. These features highlight dynamic and rapid motion patterns crucial for recognizing abnormal activities. Angles (θ, Equation 7) are computed between key joints, such as the angle formed by the right shoulder and right wrist, as well as the left shoulder and left wrist, to capture limb positioning and orientation. Eye movement is analyzed through vertical (Δy_{eye}, Equation 8) and horizontal (Δx_{eye}, Equation 9) displacements of the right and left eyes, providing insights into directional gaze and motion patterns. Lastly, distance features (d, Equation 10) measure the spatial relationships between specific keypoints, such as the right hand and right eye or the left hand and left eye, to track the relative positioning of body parts. These features collectively provide a robust representation of motion and spatial dynamics, enhancing the model's ability to analyze and recognize abnormal activities effectively.

The extracted features, such as speed, acceleration, and angles, were specifically selected to capture the distinct characteristics of abnormal behaviors. For example, short-duration rapid movements in "Attacking" are captured effectively through acceleration, while repetitive patterns in "Head Banging" and "Nail Biting" are quantified through speed and rhythm metrics.

$$v = \frac{\|\mathbf{p}(t) - \mathbf{p}(t-1)\|}{\Delta t} \tag{5}$$

$$a = \frac{v(t) - v(t-1)}{\Delta t} \tag{6}$$

$$\cos\theta = \frac{(\mathbf{p}_B - \mathbf{p}_A) \cdot (\mathbf{p}_C - \mathbf{p}_B)}{\|\mathbf{p}_B - \mathbf{p}_A\| \cdot \|\mathbf{p}_C - \mathbf{p}_B\|} \tag{7}$$

$$\Delta y_{\text{eye}} = y_{\text{eye}}(t) - y_{\text{eye}}(t-1) \tag{8}$$

$$\Delta x_{\text{eye}} = x_{\text{eye}}(t) - x_{\text{eye}}(t-1) \tag{9}$$

$$d = \|\mathbf{p}_A - \mathbf{p}_B\| = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2} \tag{10}$$

4.3 Abnormal Activity Recognition

The model architecture and the temporal parameters employed to recognize abnormal behavior ensure that the system captures sequential data effectively.

LSTM (Long Short-Term Memory) models are selected for this task due to their capability to learn long-term dependencies in sequential data. Abnormal behavior, which often exhibits subtle and short-duration patterns embedded within normal actions, requires a model that can recognize these temporal dependencies.

Temporal parameters, such as sliding window size and overlap rate, were optimized to balance recognition accuracy with computational efficiency. The high frame rate of 30fps was chosen to ensure even the smallest movements were captured, providing a detailed temporal resolution for the recognition task.

To evaluate model generalization across different individuals, this study employs a leave-one-person-out cross-validation (LOPO-CV) strategy. The dataset consists of time-series features and action labels from five individuals. Each person's data is used as the test set once, while the remaining four are combined as the training set, ensuring evaluation on unseen individuals.

Before training, features and labels are separated and converted to numerical format. The data is then segmented into time-series sequences using a sliding window approach, with the time step length determined based on data characteristics to capture temporal dependencies.

Performance metrics, including accuracy, precision, recall, and F1-score, are recorded after each fold, and the confusion matrices are aggregated to assess overall model performance. Since LOPO-CV inherently tests on unseen individuals, no separate validation set is required.

The configuration of 64 LSTM units, 20 epochs, and a batch size of 32 strikes a balance between computational efficiency and the model's ability to learn complex patterns. These settings enable the model to effectively distinguish between normal and abnormal behaviors, even in challenging real-world scenarios.

5 Results Towards Activity Recognition of Developmentally Disabled Individuals

In this study, we proposed and evaluated a video-based framework that utilizes pose estimation data to recognize both abnormal and normal human behaviors. We evaluated the performance by analyzing normal versus abnormal activities,

with results averaged across five folds using a leave-one-subject-out approach among the five participants

In this section, we detail the results of the implemented framework to our collected dataset. To achieve high-precision recognition of these behaviors, the framework employs an LSTM model, which processes time-series data to capture temporal dependencies inherent in dynamic actions. Key temporal parameters were systematically adjusted to optimize performance. These parameters include the window size, which determines the length of the analyzed time frame; the overlap rate, which specifies the degree of overlap between consecutive windows; and the sequence length, which defines the number of frames processed together.

The adjustment of these parameters enables accurate recognition of short-duration abnormal behaviors while maintaining high accuracy for normal ones, demonstrating the framework's adaptability and potential for caregiving and monitoring applications. These parameters were fine-tuned to ensure precise detection of short-duration abnormal behaviors while maintaining high accuracy for normal activities. The results demonstrate the framework's adaptability and potential for real-world applications, particularly in caregiving and monitoring systems.

5.1 Results from Overall Activity

In this section, we elaborate on the results of the recognition performance of all four normal and all four abnormal activities with varying temporal parameters.

To recognize these behaviors, an LSTM model was employed, incorporating time-series data. The performance was compared by modifying settings for window size, overlap rate, and sequence length. The details of each setting and their respective results are presented below.

Table 2: Recognition Performance with Different Temporal Settings

	Parameters			**Performance (%)**	
Approach	**WS**	**OR**	**SL**	**Accuracy**	**F1 Score**
A	30	50%	30	71	69
B	30	50%	60	72	70
C	**30**	**50%**	**90**	**74**	**75**
D	30	50%	120	74	75
E	30	50%	150	75	75

For model performance evaluation, we adopted five different settings (Approach A, B, C, D, E). These settings were designed to analyze the impact of temporal context on recognition performance by fixing the window size at 30 frames and the overlap rate at 50%, while varying only the sequence length from 30 to 150 frames in steps of 30. The results show that increasing the sequence length up to 90 frames improved performance, with the highest accuracy of 75% and F1 score of 76% achieved at this setting. However, further increases

beyond 90 frames did not lead to any additional performance gains (see Table 2). In particular, when increasing the sequence length beyond 90 frames, the accuracy remained nearly unchanged, indicating that extending the temporal context excessively does not necessarily yield further improvements.

5.2 Results from Per Activity

This section examines the recognition of each abnormal activity against all normal activities to evaluate how the temporal parameters tailored to each abnormal behavior enhance performance. The framework was designed to address realistic scenarios in which abnormal behaviors typically occur over short durations within a continuous stream of normal behaviors. For example, in practical applications such as surveillance systems in facilities or caregiving environments, abnormal behaviors are often embedded within normal behaviors, requiring precise detection of short-duration abnormal events. Each abnormal behavior (e.g., Attacking, Biting) was individually evaluated against all other behaviors classified as normal. Recognition performance was assessed for each combination.

To evaluate the recognition performance of individual abnormal behaviors, the study examined how well the model distinguished each abnormal behavior (Attacking, Biting, Head Banging, Throwing Things) from all normal behaviors (Walking, Sitting Quietly, Using Phone, Eating Snacks). Two evaluation approaches were adopted. The performance metrics for these evaluations are summarized in Tables 3 and 4.

Zero-shot and Few-shot approaches have distinct advantages and limitations in optimizing abnormal behavior recognition. Zero-shot enables rapid estimation without labeled data but relies on pre-trained knowledge, limiting its accuracy for complex or short-duration behaviors. Few-shot improves recognition accuracy by refining parameters with a small dataset but requires additional data, which may not always be feasible. Zero-shot is useful when labeled data is unavailable or for quick estimations but struggles with unique or subtle behaviors. Few-shot, while more accurate, depends on data availability. Neither approach is universally applicable—zero-shot aids initial analysis, whereas few-shot provides better optimization but is constrained by data availability.

Zero-shot recognition performance was evaluated without any prior optimization specific to each behavior, relying on parameters suggested by the LLM to reflect the short-duration and distinct patterns of abnormal behaviors. In Few-shot, small amounts of training data were provided to refine the model's parameters and enhance recognition performance for each behavior category.

5.2.1 Performance Analysis of Zero-shot Results

To analyze the recognition performance in a Zero-Shot scenario, each abnormal behavior was compared against normal behaviors. Zero-shot learning uses only text-based prompts in the LLM to generate parameter recommendations based on pre-trained knowledge. This evaluation aimed to determine how accurately the model could distinguish abnormal behaviors and identify specific challenges

in recognizing certain abnormal actions. The results are summarized in Table 3.

Table 3: Recognition Performance, Zero-shot

	Parameters			**Performance %**	
Target Abnormal Activity	**WS**	**OR**	**SL**	**Accuracy**	**F1-score**
All Normal vs. Attacking	20	50%	40	98.50	96.27
All Normal vs. Biting	**15**	**66%**	**40**	**99.49**	**98.81**
All Normal vs. Head Banging	30	66%	60	99.23	97.63
All Normal vs. Throwing	20	66%	60	99.61	98.62

The Zero-Shot results reveal several key insights into the model's ability to differentiate between abnormal and normal behaviors. Biting demonstrated the highest performance, achieving an F1 score of 98.81%. Its rhythmic and periodic movement patterns likely contributed to its distinctiveness, making it easier for the model to classify accurately. Head Banging also showed strong performance, with an F1 score of 97.63%, suggesting that its unique movement patterns were well-captured by the model. Throwing Things achieved an F1 score of 98.62%, reflecting the model's ability to recognize abrupt movements, although occasional variability led to minor misclassifications. Attacking exhibited the lowest recognition performance among the behaviors, with an F1 score of 96.27%. This was likely due to the short, discrete nature of attacking movements and their similarity to certain normal behaviors, such as hand movements during walking. These results highlight areas for potential improvement, particularly in enhancing the robustness of the model for short and variable actions. The difference between the highest accuracy of 98.81% and the lowest accuracy of 96.27% was considered. A t-test was conducted to examine whether the accuracy difference between Biting (98.81%) and Attacking (96.27%) was due to chance. The result yielded a p-value of 0.03, which is less than the significance level of 0.05. Therefore, the accuracy difference is statistically significant, indicating that the difference in performance between Biting and Attacking is not due to random variation. This conclusion confirms that there is a true performance difference between the two tasks.

5.2.2 Few-shot Recognition Performance Analysis

Few-Shot learning was employed to further refine the model's parameters for each behavior. By leveraging small amounts of labeled data, the model achieved near-perfect performance for most behaviors, as shown in Table 4. In this approach, the number of feature samples was kept constant to ensure experimental consistency. Additionally, two images were consistently used along with feature CSV files, providing additional information to enhance the few-shot learning process. Specifically, the first two images from each action sequence were selected to ensure consistency.

Few-Shot learning led to noticeable improvements in F1 scores and accuracy,

Table 4: Recognition Performance, Few-shot

	Parameters			Performance %	
Target Abnormal Activity	**WS**	**OR**	**SL**	**Accuracy**	**F1-score**
All Normal vs. Attacking	20	50%	80	99.09	97.48
All Normal vs. Biting	**20**	**50%**	**60**	**100.0**	**100.0**
All Normal vs. Head Banging	20	75%	60	99.95	99.90
All Normal vs. Throwing Things	25	80%	75	99.79	99.40

particularly for behaviors like Attacking and Throwing Things. Key observations include. Few-shot learning resulted in notable improvements in F1 scores and accuracy, especially for challenging behaviors like Attacking and Throwing Things. Biting achieved perfect scores in both F1 (100.0%) and accuracy (100.0%), underscoring the effectiveness of the proposed parameters in capturing its rhythmic motion. Throwing Things showed minor improvements in F1 and accuracy; however, the inherent variability of the behavior continued to challenge consistent classification. Attacking demonstrated significant gains, highlighting the critical role of parameter optimization in enhancing the recognition of short, rapid actions. This study focuses on recognizing abnormal behaviors in facilities for individuals with developmental disabilities, adopting a few-shot learning approach to address data collection constraints and ensure model adaptability. Due to ethical and privacy concerns, collecting large-scale labeled data on abnormal behaviors is challenging in such facilities. Additionally, certain abnormal behaviors occur infrequently, making it difficult to acquire sufficient labeled data for training. Therefore, few-shot learning, which enables models to learn effectively from a small number of samples while maintaining high classification performance, was deemed an appropriate method. These findings underscore the potential of few-shot learning to improve model performance across diverse behaviors.

6 Discussion

In this section, we analyze the results obtained from the proposed framework, focusing on its ability to recognize both abnormal and normal behaviors with varying levels of complexity. By evaluating the overall and per-activity performance, we aim to identify the factors influencing recognition accuracy, particularly the effectiveness of temporal parameter settings. Special attention is given to understanding how these parameters contribute to distinguishing short-duration abnormal behaviors embedded within continuous streams of normal activities. This analysis provides critical insights into the framework's adaptability and potential for practical applications in dynamic environments such as caregiving facilities.

6.1 Results from Overall Activity

Recognition Performance by Behavior Category Normal behaviors (e.g., Sitting Quietly, Walking) were recognized with relatively high accuracy. Abnormal behaviors (e.g., Attacking, Throwing Things) exhibited some misclassifications. This suggests that similarities between actions and imbalances in the dataset may have contributed to these errors.

Effect of Time-Series Data In Approach A (Time Steps = 60, referring to sequence length), higher misclassification rates are observed, particularly between similar actions like "Attacking" and "Throwing Things." In Approach B (Time Steps = 60), short-term temporal relationships are captured, resulting in improved accuracy. Finally, Approach C (Time Steps = 90) leverages broader temporal relationships, achieving the highest recognition accuracy, especially for complex actions. Our current approach employs a single time window to capture temporal dependencies. However, the parallel use of multiple context windows, where short-duration activities are captured with short-term windows and long-duration activities with long-term windows, remains an unexplored but promising direction. Future work could investigate this approach to enhance recognition accuracy by incorporating both rapid motion changes and sustained activity patterns. This could be particularly beneficial in distinguishing transient abnormal behaviors embedded within continuous streams of normal activities.

Analysis of the Confusion Matrix Using the confusion matrices shown in Figures 9, 10, and 11, the classification accuracy for each behavior category was evaluated in detail. Figure 9 represents the confusion matrix when using a single frame with the LSTM model (Approach A). Figure 10 corresponds to the results when employing 30 time steps (Approach B), while Figure 11 shows the results with 90 time steps (Approach C).n these confusion matrices, abnormal behaviors are highlighted by red circles. These matrices illustrate the impact of incorporating temporal context on improving recognition accuracy.

Figure 9 shows the confusion matrix for 30 time steps, demonstrating the model's classification performance. The model achieves high accuracy for "Walking" (5192 correct classifications) and "Throwing Things" (910), indicating that these activities have distinct skeletal movement patterns that facilitate accurate recognition. However, "Using Phone" exhibits a notable challenge, with 1681 instances misclassified as "Sitting Quietly" and 550 as "Biting", suggesting difficulty in differentiating between these activities due to their similar postures. Similarly, "Eating Snacks" is often confused with "Biting" (433) and "Sitting Quietly" (1259), highlighting the need for better feature extraction to distinguish these fine-grained movements.

Figure 10 shows the confusion matrix for 60 time steps. This approach achieves very high accuracy for "Walking" (4838) and "Throwing Things" (522), with minimal misclassification. However, "Using Phone" has low accuracy due to frequent misclassification as "Sitting Quietly" (1179), highlighting a key challenge. Similarly, "Eating Snacks" is often confused with "Biting" (487) and "Sitting Quietly" (588), suggesting difficulties in distinguishing actions with

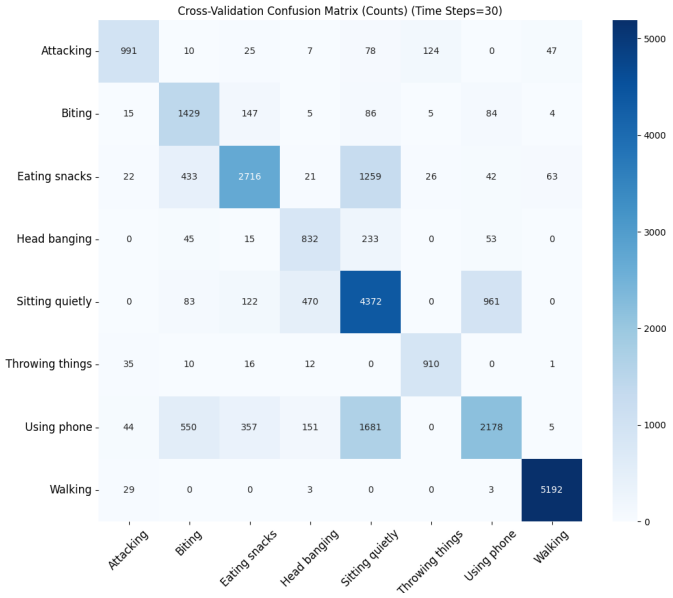


Figure 9: Confusion Matrix of Baseline, Approach A

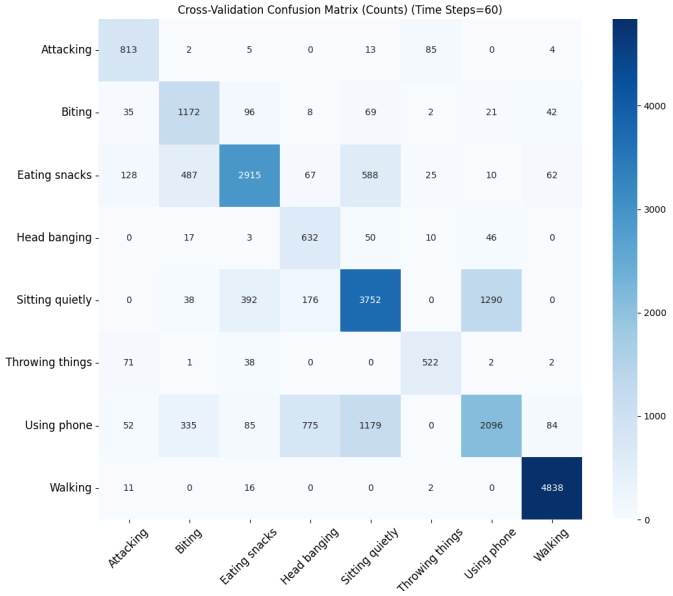


Figure 10: Confusion Matrix for Approach B

similar postures. These results indicate limitations in recognizing fine-grained movements even with longer temporal windows.

Figure 11: Confusion Matrix of Approach C

Figure 11 shows the confusion matrix for 90 time steps, which achieved the best overall performance. This approach achieves very high accuracy for "Walking" (4504) and "Throwing Things" (335), with minimal misclassification. However, "Using Phone" has low accuracy due to frequent misclassification as "Sitting Quietly" (1204) and "Biting" (310), highlighting a key challenge. Similarly, "Eating Snacks" is often confused with "Biting" (548) and "Sitting Quietly" (614), suggesting difficulties in distinguishing actions with similar postures. These results indicate limitations in recognizing fine-grained movements even with longer temporal windows, and challenges remain in accurately differentiating activities with subtle hand movements or similar postural characteristics.

6.2 Discussion on Per-Activity Results

The results emphasize the flexibility and practicality of the proposed framework, particularly for real-world applications that demand accurate recognition of specific abnormal behaviors in dynamic environments. Table 5 summarizes the recognition performance across various abnormal behaviors, highlighting the following key observations

- The rhythmic nature of certain actions (e.g., Biting) makes them easier to distinguish, even in zero-shot scenarios. This demonstrates the framework's ability to recognize patterns that exhibit repetitive and predictable motion characteristics without prior task-specific training.
- Few-shot learning substantially enhances the model's performance for challenging behaviors, such as Throwing Things and Attacking, which often involve complex and less repetitive motion. This underscores the importance of incorporating even a small amount of labeled data to fine-tune the model for such activities..
- Future work could explore hybrid approaches that combine zero-shot and few-shot learning for scenarios with limited labeled data. Leveraging the strengths of both approaches could address the trade-offs between model generalization and precision in real-world settings where data collection is often constrained.

Table 5: Recognition Performance of Normal vs. Abnormal Behaviors

Abnormal Activity	**Prompt**	**F1 Score, %**	**Improvement, %**
Attacking	Baseline	90.17	
	ZeroShot	96.27	6.10
	FewShot	**97.48**	**7.31**
Biting nail	Baseline	98.76	
	ZeroShot	98.81	0.05
	FewShot	**100.00**	**1.24**
Head Banging	Baseline	95.22	
	ZeroShot	97.63	2.41
	FewShot	**99.9**	**4.68**
Throwing things	Baseline	91.71	
	ZeroShot	98.62	6.91
	FewShot	**99.4**	**7.69**

The results demonstrated highest performance with Few-Shot prompting where the model achieved F1-scores improvement of 1.24% for biting nails, 4.68% for head banging, 7.69% for throwing things, and 7.31% for attacking compared with baseline. Zero-Shot prompting similarly demonstrated robust recognition, achieving F1 scores exceeding 96% for all abnormal behaviors. Overall, the findings suggest the potential of this system to help caregivers efficiently identify abnormal behaviors in real-world settings.

7 Conclusion

In this study, we propose to optimize temporal parameters with pose estimation data of simulated abnormal activities of developmentally disabled individuals by

incorporating behavior context to Large Language Models (LLMs). The contributions of this work includes the creation of a unique dataset with labeled abnormal behaviors and the proposed application of LLMs to this dataset comparing results of Zero-Shot and Few-Shot. Our method leverages the context of the collected abnormal activity data to prompt LLMs to suggest window size, overlap rate, and LSTM model's length sequence tailored to the specific characteristics of these activities.

The results of this study demonstrate the proposed framework's capability to recognize both abnormal and normal behaviors with high accuracy. Notably, incorporating temporal parameters generated from prompting with abnormal data context resulted in significant performance improvements compared to traditional single-frame-based recognition methods. This framework has the potential to enhance safety and efficiency in behavioral monitoring within facilities, indicating promising applications in real-world scenarios. By employing various settings for window size and overlap rates, we confirmed that considering longer temporal contexts effectively captures the continuity of actions, emphasizing the importance of time-series data in behavior recognition.

The proposed framework was validated to accurately recognize not only normal behaviors, such as sitting quietly and walking, but also abnormal behaviors, such as biting and throwing objects. Notably, "biting," with its rhythmic and periodic characteristics, achieved the highest recognition accuracy as the model effectively captured its unique patterns. However, "attacking" presented challenges due to the diversity and short duration of its movements, leading to occasional misclassifications. These findings highlight the critical role of tailored parameter settings for each behavior category to maximize model performance.

Furthermore, the ability to quickly and accurately detect abnormal behaviors contributes not only to improving safety within facilities but also to reducing the burden on caregivers. The capability to reliably identify short-duration abnormal behaviors embedded within normal activities is a key feature for real-time response in such environments.

Future challenges include expanding datasets to account for the diversity and individuality of abnormal behaviors. Optimizing time-series models to better capture action contexts offers additional opportunities for performance improvements. Additionally, comparative evaluations with other behavior recognition algorithms and frameworks will further clarify the practicality and versatility of the proposed approach. Alternatively, one direction for future research is the implementation of multiple context windows running in parallel to capture variations in activity durations. This technique could enhance recognition performance by accommodating activities of different lengths more effectively. While time constraints prevented us from testing this approach, its potential impact warrants further investigation and evaluation in future studies.

The findings of this study highlight the potential of behavior monitoring technologies and provide a practical foundation for improving safety and efficiency in facilities for developmentally disabled individuals. Building on these insights, we aim to develop more advanced and versatile systems for abnormal behavior recognition through further research.

References

[1] Spreat, S., and Conroy, J. A Supply and Demand Perspective on the Workforce Crisis in Intellectual Disability. Journal of Intellectual and Developmental Disability, pp. 1–5, 2021. doi: 10.3109/13668250.2021.1975365

[2] Peters, V., Frielink, N., van Leest, C., Heerkens, L., and Embregts, P. J. C. M. Impact pathways: Putting workers front and center in addressing workforce shortages in intellectual disability care. International Journal of Operations and Production Management, 44(13), pp. 251–262, 2024. doi: 10.1108/IJOPM-03-2023-0175

[3] Japan Welfare and Medical Service Agency. Survey on securing human resources for disability welfare services, etc. in 2023. Research Group, Management Support Center, 2024.

[4] Ashok, N., Hughes, D., and Yardley, S. Challenges and opportunities for improvement when people with an intellectual disability or serious mental illness also need palliative care: A qualitative meta-ethnography. Palliative Medicine, pp. 1–16, 2023. doi: 10.1177/02692163231175928

[5] Ke, S.-R., Le, H. U. T., Yoo, J.-H., Lee, Y.-J., Hwang, J.-N., and Choi, K.-H. A Review on Video-Based Human Activity Recognition. Computers, 2(2), pp. 89–93, 2013. doi: 10.3390/computers2020089

[6] Inoue, S., Lago, P., Hossain, T., Mairittha, T., and Mairittha, N. Integrating Activity Recognition and Nursing Care Records: The System, Deployment, and a Verification Study. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., 3(3), Article 86, pp. 1–24, 2019. doi: 10.1145/3351244

[7] Garcia, C., and Inoue, S. Challenges and Opportunities of Activity Recognition in Clinical Pathways. In *Human Activity and Behavior Analysis*, CRC Press, pp. 1–18, 2024. doi: 10.1201/9781003371540-8

[8] Morais, R., Le, V., Tran, T., Saha, B., Mansour, M., and Venkatesh, S. Learning Regularity in Skeleton Trajectories for Anomaly Detection in Videos. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11988–11996, 2019. doi: 10.1109/CVPR.2019.01227

[9] Fikry, M., Garcia, C. A., Vu, Q. N. P., Inoue, S., Oyama, S., Yamashita, K., Sakamoto, Y., and Ideno, Y. Improving Complex Nurse Care Activity Recognition Using Barometric Pressure Sensors. In *Human Activity and Behavior Analysis*, CRC Press, pp. 261–283, 2024. doi: 10.1201/9781003371540-18

[10] Dobhal, U., Garcia, C., and Inoue. "Synthetic Skeleton Data Generation using Large Language Model for Nurse Activity Recognition". In Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '24). Associ-

ation for Computing Machinery, New York, NY, USA, pp. 493–499. https://doi.org/10.1145/3675094.3678445

[11] Ni'ma Shoumi, M., and Inoue, S. Leveraging the Large Language Model for Activity Recognition: A Comprehensive Review. International Journal of Activity and Behavior Computing, pp. 11–12, 2024. doi: 10.60401/ijabc.21

[12] Zhang, S., Li, S., Zhang, S., Shahabi, F., Xia, S., Deng, Y., and Alshurafa, N. Deep Learning in Human Activity Recognition with Wearable Sensors: A Review on Advances. Sensors, 22(4), pp. 6–10, 2022. doi: 10.3390/s22041476

[13] Jobanputra, C., Bavishi, J., and Doshi, N. Human Activity Recognition: A Survey. Procedia Computer Science, Vol. 155, pp. 698–703, 2019. doi: 10.1016/j.procs.2019.08.100

[14] Alafif, T., Hadi, A., Allahyani, M., Alzahrani, B., Alhothali, A., Alotaibi, R., and Barnawi, A. Hybrid Classifiers for Spatio-Temporal Abnormal Behavior Detection, Tracking, and Recognition in Massive Hajj Crowds. Electronics, Vol. 12, No. 1165, pp. 2–3, 2023. https://www.mdpi.com/2079-9292/12/5/1165

[15] Qian, H., Zhou, X., and Zheng, M. Abnormal Behavior Detection and Recognition Method Based on Improved ResNet Model. Computers, Materials and Continua, 65(3), pp. 2153–2167, 2020. https://www.techscience.com/cmc/v65n3/38118

[16] Nassif, A. B., Darweesh, E., Al-Azzawi, A., and Jamous, R. Anomaly Detection in Human-Robot Collaboration Using LSTM Autoencoder. Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction, pp. 334–336, 2023. doi: 10.1109/HRI.2023.1001234

[17] Li, Z., Zhang, A., Han, F., Zhu, J., and Wang, Y. Worker Abnormal Behavior Recognition Based on Spatio-Temporal Graph Convolution and Attention Model. Electronics, Vol. 12, No. 2915, pp. 2–5, 2023. doi: 10.3390/electronics12132915

[18] Martín-Chinea, K., Ortega, J., Gómez-González, J. F., Pereda, E., Toledo, J., and Acosta, L. Effect of Time Windows in LSTM Networks for EEG-Based BCIs. Cognitive Neurodynamics, Vol. 17, pp. 385–398, 2023. doi: 10.1007/s11571-023-09895-2

[19] Dong, L., Fang, D., Wang, X., Wei, W., Scherer, R., and Woźniak, M. Prediction of Streamflow Based on Dynamic Sliding Window LSTM. Water, Vol. 12, pp. 6–11, 2020. doi: 10.3390/w12010006

[20] Ji, S., Zheng, X., and Wu, C. HARGPT: Are LLMs Zero-Shot Human Activity Recognizers? Proceedings of the ACM International Conference on Human Activity Recognition, pp. 1–3, 2024. URL: https://example.com/hargpt-2024

[21] Shoumi, N., and Wei, L. What Do Sensor-Based Human Activity Recognition Studies Tell Us About Multimodal Integration? Sensors, 24(1), pp. 190–192, 2024. doi: 10.3390/s24010190

[22] Wang, C.-Y., Bochkovskiy, A., Liao, H.-Y. M.: YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. arXiv preprint arXiv:2207.02696, pp. 1–15, 2022. http://arxiv.org/abs/2207.02696

[23] Hochreiter, S., and Schmidhuber, J.: Long Short-Term Memory. Neural Computation, 9(8), pp. 1735–1740, 1997. doi: 10.1162/neco.1997.9.8.1735

[24] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., and Jones, L.: Attention Is All You Need. Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS), pp. 1–15, 2017. http://arxiv.org/abs/1706.03762

[25] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., and Kaplan: Language Models are Few-Shot Learners. arXiv preprint arXiv:2005.14165, pp. 6–7, 2020. http://arxiv.org/abs/2005.14165

[26] Yan, S., Xiong, Y., and Lin, D.: Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI-18), pp. 6–7, 2018. http://arxiv.org/abs/1801.07455

[27] Zhou, K., Wu, T., Wang, C., Wang, J., and Li, C.: Skeleton-Based Abnormal Behavior Recognition Using Spatio-Temporal Convolution and Attention-Based LSTM. Procedia Computer Science, 174, pp. 424–432, 2020. doi: 10.1016/j.procs.2020.06.111

[28] Wei, X., and Wang, Z. TCN-attention-HAR: Human Activity Recognition Based on Attention Mechanism Time Convolutional Network. Scientific Reports, 14(7414), pp. 1–11, 2024. doi: 10.1038/s41598-024-64412-3

[29] Mori, S. Understanding Autism Through Cultural Models and Its Implications. Doctoral Thesis, Kobe University, Japan, pp. 118–122, 2021. doi: 10.18910/78901

[30] Fujioka, T., Garcia, C., and Inoue, S. (2025, March 17). Challenge: Abnormal Activity Detection in Individuals with Developmental Disabilities. IEEE Dataport. https://dx.doi.org/10.21227/qfkw-sa40.s