Nurse Activity Recognition in Gastrostomy Tube Feeding Using Video-Based Pose with Large Language Model-Guided Features

Lingfeng Zhao* [1], Christina Garcia [2], Shunsuke Komizunai [3], Noriyo Colley [4], Atsuko Sato [5], Mayumi Kouchiyama [6], Toshiko Nasu [7], Sozo Inoue [8]
[128]Kyushu Institute of Technology, [3]Kagawa University, [4]Hokkaido University, [567]Hiroshima Bunka Gakuen University

Abstract

In this paper, we improve nursing activity recognition in gastrostomy tube feeding (GTF) with temporal variations and sequential errors by integrating activity context to Large Language Model (LLM) for guided feature selection and post-processing. GTF is a delicate nursing procedure that allows direct stomach access in children for supplemental feeding or medication, but it is underrepresented in datasets, posing challenges for accurate detection. Manual feature engineering may overlook subtle but important motion cues, particularly in opening and closing the gastrostomy cover, where changes are minimal and localized to the hands. Additionally, sequence inconsistencies and missed activities limit the effectiveness of pose estimation methods in healthcare. Leveraging the contextual adaptability of LLMs, we generate new features suggested by the language model, combining them with hand-crafted features to optimize the model. For post-processing, a sliding window smoothing method based on majority voting is applied. To mitigate duration-based discrepancies, a priority handling is incorporated for short-duration activities to preserve their recognition accuracy while addressing repeated labels caused by long-duration actions. Particularly, we applied activity recognition to our unique GTF dataset collected from recorded video of two nurses, two students, and two staff members for three days with 17 labeled activities.

[1]zhaolin46366@gmail.com
[2]alvarez7.christina@gmail.com
[3]komizunai.shunsuke@kagawa-u.ac.jp
[4]noriyo@med.hokudai.ac.jp
[5]a-sato@hbg.ac.jp
[6]kochiyama@hbg.ac.jp
[7]nasu@hbg.ac.jp
[8]sozo@brain.kyutech.ac.jp

Keypoints are extracted using YOLO11. Compared to the baseline, the application of LLM to GTF nurse activity recognition with pose estimation improved the Random Forest performance of F1-score from 54% to 57%. Additionally, incorporating the sliding window smoothing approach based on majority voting with short-term action priority, resulted in a 3% further increase.

1 Introduction

Nursing activity recognition in clinical settings is a critical research area aimed at improving healthcare quality and operational efficiency [1, 2, 3, 4]. Within this domain, gastrostomy tube feeding (GTF) stands out as a procedure that, while common in certain contexts, is infrequent across large datasets—thus making it challenging to detect accurately [5]. GTF requires precise and timely actions to ensure patient safety and comfort, yet its sub-activities exhibit considerable variety in duration and subtlety. For instance, critical tasks like opening/closing the gastrostomy cover involve minimal hand motion, often overlooked in feature engineering [6].

Flexible activity recognition systems are crucial for on-site deployment, ensuring seamless integration and data adaptation [4, 7]. Advances in computer vision have enabled video-based pose estimation for action recognition, offering a non-invasive method to capture skeletal motion [6, 8]. YOLO efficiently tracks keypoints, supporting nursing analysis [9]. However, recognizing complex clinical actions requires robust feature extraction that captures both spatial distribution and dynamic movement patterns. Furthermore, in real-world clinical environments, datasets are frequently imbalanced and contain short-duration activities that are essential but underrepresented [5, 10, 7]. These factors complicate the classification task, as models may struggle to identify brief yet significant actions within the context of longer procedures [1, 3].

To address these challenges, we propose a framework integrating video-based pose estimation, LLM-guided feature engineering, and post-processing techniques [11, 12, 13, 14]. LLMs generate domain-relevant features from training data, complementing hand-crafted features for improved pose extraction [12]. Our method refines skeletal keypoints relevant to GTF while mitigating sequence inconsistencies using sliding-window smoothing with majority voting [13,14]. To enhance recognition of brief yet critical actions, we implement a priority mechanism that prevents short-duration activities from being overshadowed by longer ones, ensuring accurate identification of key nursing tasks [1, 10].

In this paper, we improve GTF activity recognition by addressing the challenges on duration-based discrepancies and sequence inconsistencies leveraging knowledge from LLMs considering temporal and sequential context in prompting. Specifically, our contributions are:

1. Pioneering the application of activity recognition in complex Gastro Tube Feeding with extracted pose data using YOLO11.

2. Utilizing LLM with temporal and sequence contexts for features relevant to short and long-term GTF actions.

3. Comparison of post-processing techniques for GTF activity and introduced a targeted approach with priority handling addressing sequence errors.

To evaluate our framework, we conduct experiments on a dataset comprising multiple participants (Nursing Educators, Nursing Students, and Registered Nurses) performing 17 GTF activities over several sessions. As part of the training process, we apply SMOTE to balance the uneven label distribution before feeding the data into a Random Forest classifier [11]. In our experiments, the proposed method demonstrated clear quantitative advantages. Compared to the baseline, our framework improved accuracy in GTF nurse activity recognition from 55% to 58% and the F1-score from 54% to 57%. By integrating LLM-guided feature engineering and post-processing techniques, an additional 3 % improvement in F1-score was achieved. These findings underscore the promise of integrating LLMs into pose-based activity recognition pipelines, particularly within the resource-constrained environment of clinical nursing care [5, 12, 14].

By combining LLM-guided feature engineering and robust post-processing, our framework not only addresses the intrinsic challenges of GTF recognition but also provides a blueprint for extending similar strategies to other specialized or underrepresented healthcare procedures. Among the prompting strategies, the temporal and sequential context combined in a few-shot prompting approach yielded the highest performance, as it effectively captured both short and long-duration activity nuances. Additionally, the post-processing strategy with a priority protection mechanism for short-duration actions proved to be the most effective, ensuring that brief yet critical tasks were not overshadowed by longer actions. This synergy between context-aware prompting and adaptive post-processing established the foundation for our improved results.

This paper is structured as follows: Section 2 reviews related work on GTF nurse activity recognition, video-based pose estimation, and LLMs. Section 3 describes the dataset and its relevance to temporal and sequential challenges. Section 4 outlines the proposed method, exploring LLM-driven feature extraction and post-processing for GTF activity recognition. Sections 5 and 6 elaborates and discuss the results, comparing baseline models with LLM-assisted approaches. Finally, Section 7 concludes with key findings and future research directions.

2 Related Work

In this section, we review the existing literature and identify the gaps that motivate our proposed framework for GTF activity recognition. First, we describe general approaches for *nursing activity recognition*, highlighting key studies on endotracheal suctioning and gastrostomy tube feeding. Next, we discuss the *challenges associated with video-based pose* estimation in clinical environments,

such as occlusions, multi-person interactions, and short-duration activities. We then survey *feature optimization* strategies, including recent work on Large Language Model (LLM)-driven feature engineering. Following that, we examine *post-processing* techniques that have been employed to refine temporal predictions in action recognition tasks. Finally, we conclude by positioning our own *paper proposal*, illustrating how it integrates LLM-suggested features and adaptive post-processing to address the aforementioned difficulties.

2.1 Nursing Activity Recognition Approaches

Nursing activity recognition has garnered considerable attention due to the critical role nurses play in patient care, especially in procedures like endotracheal suctioning and gastrostomy tube feeding. Existing studies commonly employ machine learning or deep learning pipelines to detect specific actions and assess nursing skills. For instance, Islam et al. [1] investigated a video-based pose estimation framework to recognize nurse actions during endotracheal suctioning, while Ngo et al. [2, 3] focused on analyzing multi-view skeleton data to improve recognition accuracy and provide skill assessment. More recently, Ngo et al. [5] organized a nurse care activity recognition challenge, encouraging participants to fuse generative AI with skeleton-based approaches to boost performance. These works highlight the feasibility of leveraging pose estimation in nurse activity recognition, yet also reveal notable hurdles, such as imbalanced datasets and wide variability in nurse movements.

2.2 Challenges with Video and Pose Data

Video-based pose estimation introduces distinct challenges in clinical environments, where occlusions, motion blur, and varying camera perspectives can degrade keypoint detection [6, 8]. In nurse activity recognition, the situation is further complicated by frequent posture changes, the presence of multiple people in the frame, and the short duration of certain critical actions (e.g., adjusting a medical device). Dobhal et al. [10] underscore that limited data and subtle differences in nurse hand movements can significantly affect model performance. Thus, achieving robust recognition requires not only accurate pose extraction but also carefully designed features to account for motion nuances and sporadic labeling errors.

2.3 Feature Optimization Methods

A variety of feature engineering approaches have been applied to enhance model performance. Beyond simple keypoint coordinates, researchers have introduced additional descriptors such as joint angles, distances, velocities, and accelerations. Kaneko and Inoue [12] employed Large Language Models (LLMs) to suggest new sensor placements and feature sets in human activity recognition, reducing reliance on expert-driven feature design. Similarly, Ronando and Inoue [13] explored LLM-based feature engineering to improve fatigue detection, while

Shoumi and Inoue [14] provided a comprehensive overview of how LLMs can refine activity recognition pipelines in different application domains. These studies suggest that incorporating LLM-guided insights into existing workflows can significantly reduce engineering overhead and uncover subtle, domain-specific features that might be overlooked by traditional methods [11, 12].

2.4 Post-processing Techniques

Even with optimized features, misclassifications often arise at temporal boundaries or during short-duration actions. Post-processing, therefore, plays a key role in smoothing predictions and correcting boundary errors [17]. Methods vary from simple voting-based smoothing to more advanced techniques that adjust temporal localization. For instance, Nag et al. [15] introduced a post-processing approach for temporal action detection that addresses quantization errors, whereas Tran et al. [16] adopted a data-driven method for refining 3D action recognition results in untrimmed videos. These strategies confirm the potential of post-processing to further refine classification outputs, especially where short or overlapping actions are frequent.

2.5 Application of LLMs in Machine Learning

Motivated by these findings, our work focuses on integrating LLM-driven feature engineering and a dedicated post-processing module into a comprehensive pipeline for GTF nursing activity recognition. Recent advancements in LLMs have driven current research to adopt mixed-method approaches for context-aware applications in medical assistant systems [19]. While previous nurse-centered studies have primarily emphasized either pose extraction or specialized feature engineering, few have combined LLM-suggested features with robust temporal smoothing and short-action handling in the context of underrepresented nursing activities. Prompt design is essential to optimizing the response from LLMs agent frameworks [18]. Temporal context is crucial for understanding and localizing actions [21]. Our approach systematically exploits pose-based features (e.g., joint angles, velocities, center-of-mass changes) alongside LLM-guided feature recommendations, supplemented by an adaptive post-processing mechanism to handle short-duration tasks. Through this fusion, we aim to address the imbalances and intricacies observed in clinical data, ultimately contributing to a more accurate and interpretable nurse activity recognition.

3 Dataset with Participants Performing GTF

This section details the GTF data collection including preprocessing techniques. Gastrostomy tube feeding, a critical and common nursing procedure. Due to the declining birth rate and aging population, the mortality rate from aspiration pneumonia is rising, leading to an increasing number of elderly individuals requiring tube feeding, is underrepresented in datasets due to its complexity,

precision requirements compared to routine tasks. This study pioneers data collection for gastrostomy tube feeding using YOLO11 pose estimation. Information on participant demographics, recording setup, and the annotated activity classes in gastrostomy tube feeding procedures are elaborated. An explanation of the 2D pose estimation method using YOLO11 is explained, highlighting the extraction and preprocessing of skeletal keypoints. Special attention is given to handling challenges like short-duration actions, data imbalance, and multiperson detection issues. Additionally, we discuss the application of data smoothing with interpolation to address missing keypoints, ensuring data consistency and continuity. These steps collectively provide a robust foundation for modeling and analyzing nursing activities.

3.1 Data Collection

The dataset consists of recorded video of 6 participants comprising of two Nursing Educators, two Nursing Students, and two Registered Nurses performing gastrostomy feeding with 17 total activities for three days. Table 1 presents the corresponding labels for each activity, facilitating subsequent classification and analysis. Some tasks, such as head elevation, drag administration have been omitted because a simulator was used.

The video recordings were captured using a UMIDIGI A3X smartphone with a dual-lens camera (1920x1080 pixels, 33 fps) and utilized the ESTE-TF simulator to standardize the activities.

Table 1: Gastro Tube Feeding Nurse Activities

Activity ID	Activity Performed
0	Explanation to patient
1	Confirm necessary items
2	Disinfect hands
3	Wearing gloves
4	Prepare the nutrition solution
5	Check the gastronomy site
6	Open the gastronomy cap
7	Inject lukewarm water
8	Connect the nutrition tube
9	Adjust the infusion rate
10	Removal of gloves
11	Prepare lukewarm water
12	Close the clamp
13	Disconnect the nutrition tube
14	Close the gastronomy cap
15	Clean up used items
16	others

Among the GTF actions, short-duration classes are particularly challenging

as they involve minimal motion localized to small body areas, like the hands, and often occur rapidly. Furthermore, these brief actions are prone to being overshadowed or misclassified when analyzed within longer time frames. This happens because the temporal aggregation of motion patterns from longer actions can dominate the signal, making it difficult for models to differentiate the subtle transitions indicative of short-duration actions. Examples of such short-duration actions are illustrated in Fig. 1.

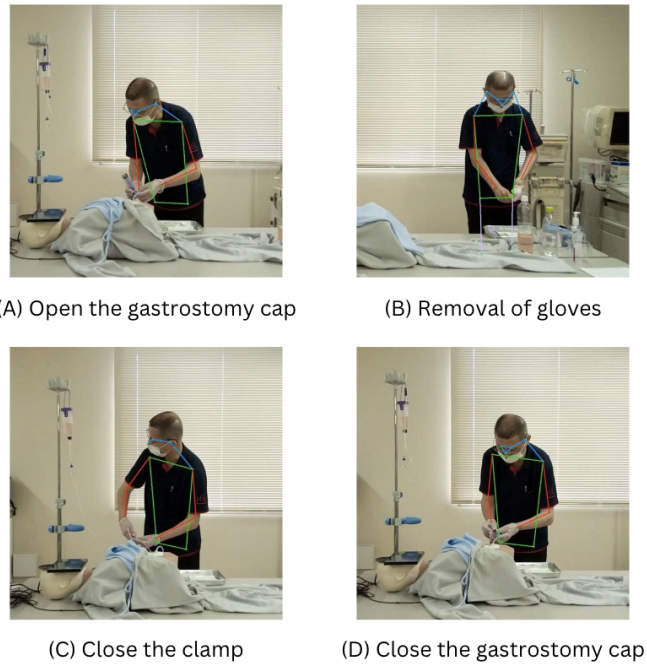

Figure 1: Examples of short-duration actions.

3.2 2D Pose Estimation

To analyze gastrostomy tube feeding activities, we employed the YOLO11 model [9] to extract skeletal keypoints from recorded videos. This model identifies and tracks 17 key body landmarks, including 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. The pose estimation output of YOLO11 for each frame is represented as a vector comprising the X and Y coordinates for each keypoint.

During preprocessing, we excluded body parts obscured by the operating table, such as knees and ankles, which were assigned zero values by YOLO11, indicating that these keypoints were not detected in the video. Furthermore, to address the issue of data imbalance, we applied the Synthetic Minority Oversampling Technique (SMOTE) to balance the dataset. By using SMOTE, we

augmented the data to ensure fair representation of both frequent, long-duration activities and rare, short-duration actions reflected in Fig. 2.

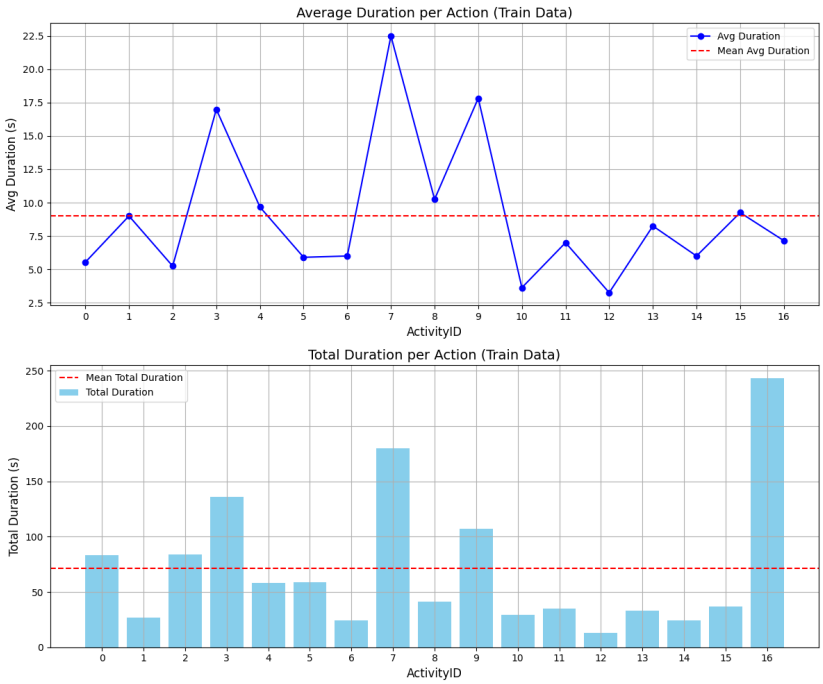

Figure 2: Average and Total Duration per Action in Training Dataset.

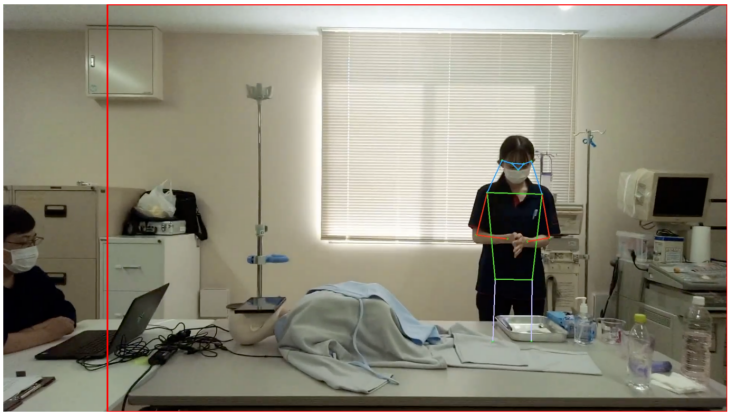

Figure 3: Example frame with unrelated individuals.

One significant challenge in pose estimation arises when multiple individuals appear in a single frame. For instance in the video frames shown in Fig. 3, static and unrelated individuals may appear in the background, leading to

false detections by YOLO11 and the generation of additional skeleton data. To address this issue, we applied a predefined bounding box of size 1620 by 1080 during data processing to restrict the detection area while ensuring that all actions performed by the nurse during the gastrostomy feeding procedure were accurately captured.

3.3 Data Smoothing with Interpolation

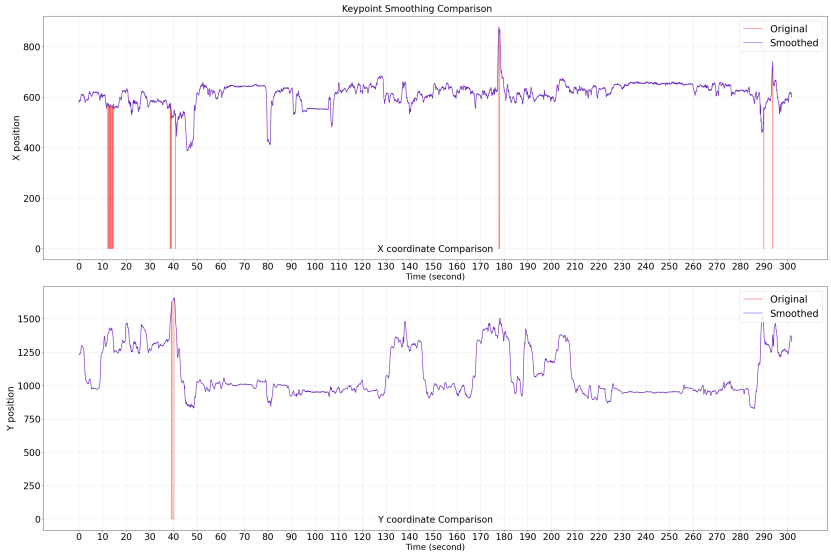

Figure 4: Keypoint Data Smoothing with Interpolation.

To address the issue of missing keypoints in skeleton data, a linear interpolation method was employed wit a smoothness length of 3 seconds. This technique utilized NumPy's interp function to replace zero values in the keypoint coordinates with estimated values based on neighboring non-zero points. By leveraging this approach, missing data points were smoothed effectively, ensuring continuity in the trajectories of the keypoints. Fig. 4 illustrates the effect of smoothing on the X and Y coordinate trajectories, demonstrating the reduction of abrupt changes and improved data consistency.

4 Proposed Method for Gastro Tube Feeding Nurse Activity Recognition

In this section, we detail the proposed method in Fig. 5 for improving activity recognition in GTF using LLM with temporal and sequential context for features and employing a targeted post-processing for class-order discrepancies.

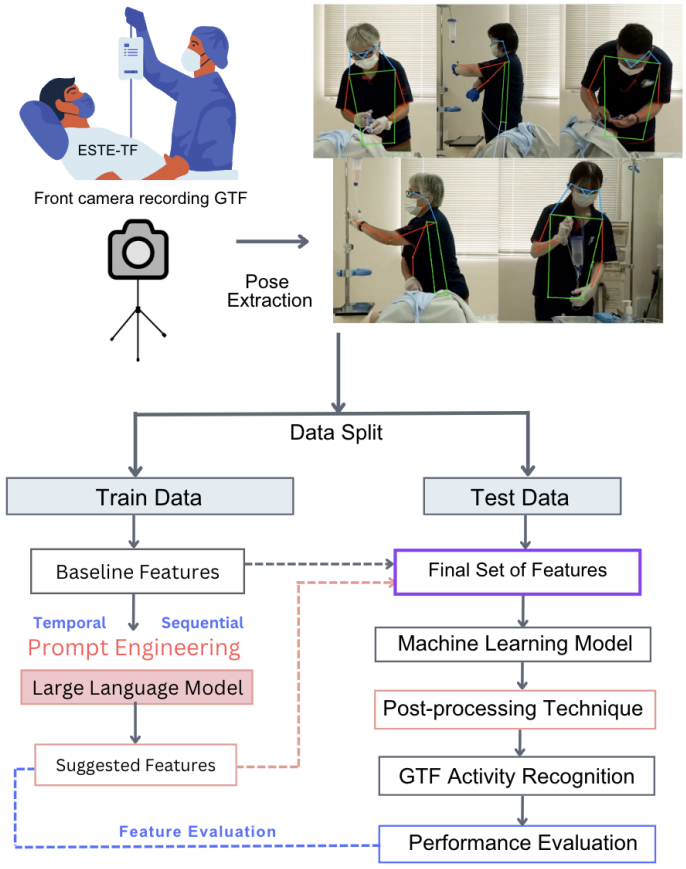

Figure 5: Pipeline for GTF Activity Recognition.

4.1 Prompt Design for LLM Approach

Developing an accurate activity recognition system for gastro-related activities requires identifying both short and long-duration sub-activities. Manual feature engineering may overlook subtle but important motion cues, particularly when opening and closing the gastrostomy cover, where changes are minimal and localized to the hands. Additionally, sequence inconsistencies and missed activities limit the effectiveness of pose estimation methods in healthcare.

Prompting with information and context from the training data, we generate new features suggested by LLM, combining them with hand-crafted features to optimize the recognition model. The output of few-shot prompting is mathematically expressed in equation 1. By considering the key points relevant to the activity, we optimize pose information extracted from the video data.

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

$$P = \{R, C, E, T\} \tag{2}$$

$$F = G(P) = \{\text{feature 1, feature 2, } \ldots, \text{ feature n}\} \tag{3}$$

We applied few-shot prompting with temporal and sequence context in P expressed in eq. 2 to optimize LLM feature suggestions F in eq. 3. The GPT-4o model by OpenAI is used accessed via web interface, having default parameters including a temperature of 0.7, a maximum token limit, and nucleus sampling with a top-p range of 0.9 to 1.0 for diversity. Using the LLM leverages strong contextual understanding to suggest detailed, appropriate feature suggestions where real context is crucial.

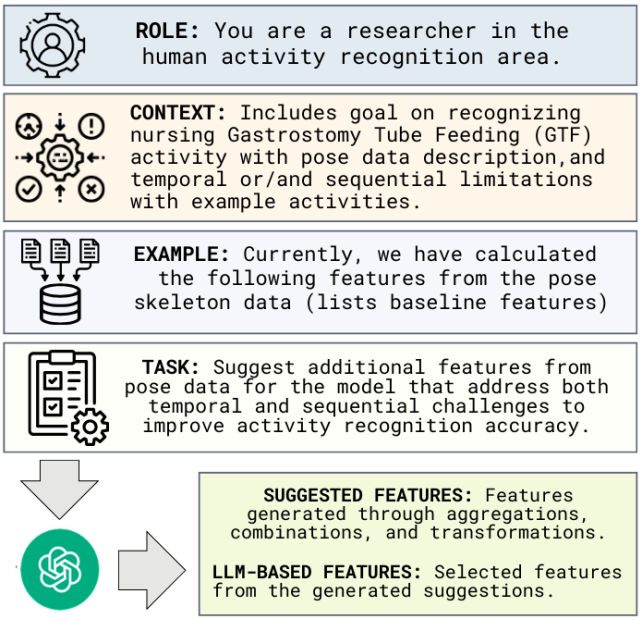

Figure 6: LLM prompting approach for GTF activity features.

The prompting technique employed for GTF pose features detailed in Fig. 6 is made up of the following four structures: (1) The Role, (2) Context, (3) Example of baseline features, and (4) The task.

- **Role** Sets the domain-specific perspective for the model to generate responses ensuring the LLM output is relevant.
- **Context** Detailed background about the task, data, and challenges to ensure the model understands the problem space.
- **Example** Supplies guidance or a baseline to prevent redundancy or irrelevant suggestions.
- **Task** Clearly defines the objective and specific expected output.

4.2 Activity Recognition with Base Features

For classification of GTF activities from skeleton data, a Random Forest Classifier (RFC) is used due to its versatility and robustness against overfitting, ensuring generalization to the data [20]. RFC is adept at handling non-linear data, allowing it to model complex relationships. RFC is effective for both classification and regression tasks, and is capable of handling numerical and categorical features commonly found in human activity recognition datasets.

Building on its robustness, the RFC model was further optimized by tuning its hyperparameters using both GridSearchCV and RandomizedSearchCV to identify the best configuration for the GTF dataset. The selected parameters for the best-performing RFC model include bootstrap set to False, ensuring all features are used without replacement, and max_depth to none, allowing each tree to grow fully. The max features parameter was set to log2, limiting the number of features considered for each split, while min_samples_leaf is 1 and min_samples_split is 2 to promote granular splits for fine-grained classification. Additionally, the number of trees n_estimators is set to 500 to enhance the ensemble's stability, providing a robust and accurate model for classifying GTF activities. These carefully tuned parameters maximize the model's ability to capture complex patterns in the GTF skeleton data.

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{4}$$

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \tag{5}$$

$$\text{max\_ft} = \max_i x_i \tag{6}$$

$$\text{min\_ft} = \min_i x_i \tag{7}$$

$$\text{var\_ft} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \tag{8}$$

$$\text{med\_ft} = \text{median}(X) \tag{9}$$

$$\text{sum\_ft} = \sum_{i=1}^{N} x_i \tag{10}$$

$$\cos(\theta) = \frac{\mathbf{BA}\cdot\mathbf{BC}}{\|\mathbf{BA}\|\|\mathbf{BC}\|}, \qquad \theta = \arccos\left(\cos(\theta)\right) \tag{11}$$

$$v_i = (x_{i+1} - x_i), \qquad v\_seg = \{v_1, v_2, \ldots, v_{N-1}\} \tag{12}$$

The base features extracted from the GTF skeleton data include statistical, geometric, and motion-based descriptors. Statistical features encompass the **mean** (μ, Eq. 4), **standard deviation** (σ, Eq. 5), **maximum** (max_ft, Eq. 6), **minimum** (min_ft, Eq. 7), **variance** (var_ft, Eq. 8), **median** (med_ft, Eq. 9), and **sum** (sum_ft, Eq. 10), capturing central tendencies, data spread, and aggregate values. Geometric features include joint **angles** (θ, Eq. 11), calculated using vector projections to represent relative body orientations. Additionally, motion-based features are derived through **velocity** calculations (v_i, Eq. 12), representing the positional differences between consecutive frames. These features collectively enable detailed analysis of body movements and postures, forming a robust basis for recognizing the collected complex GTF nursing activities. The training dataset consisted of recording from four participants, while the testing dataset included recording from two participants.

4.3 Post-processing prioritizing short-term actions

Fig. 7 details the proposed post-processing technique applied to GTF data featuring a sliding window smoothing based on majority voting. To mitigate duration-based discrepancies, a priority handling is incorporated for short-duration activities S as in Eq. 14 to preserve their recognition accuracy while addressing repeated labels caused by long-duration actions L as in Eq. 13. Given all consecutive frames labeled as activity, *Seg(a)* and predicted classes by the model y_{pred}, the post-processing is expressed mathematically as Eq. 15.

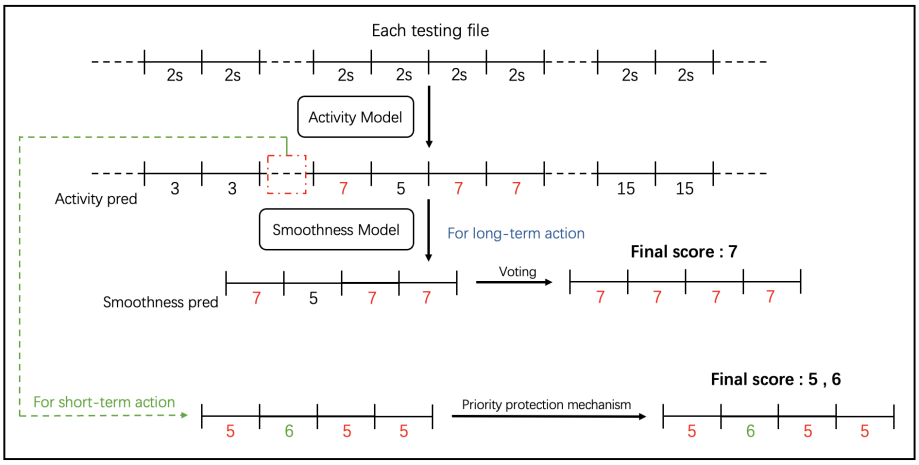

Figure 7: Framework for Post-Processing suggested for GTF activity.

$$L = \{a_1, a_2, \ldots, a_k\} \quad \text{Set of long-duration activities} \tag{13}$$

$$S = \{b_1, b_2, \ldots, b_m\} \quad \text{Set of short-duration activities} \tag{14}$$

$$y_{\text{post}}[i] = \begin{cases} \text{maj\_vote}(y_{\text{pred}}[Seg(a)]), & \text{if } y_{\text{pred}}[i] \in L, \\ y_{\text{pred}}[i], & \text{if } y_{\text{pred}}[i] \in S. \end{cases} \tag{15}$$

5 Results

This chapter analyzes the performance and effectiveness of prompting strategies applied to the LLM for GTF nurse activity recognition. Specifically, it analyzes the contribution of LLM-generated features, the impact of post-processing on recognition accuracy, and improvements in both short and long-duration activity recognition, using comparative metrics against baseline models.

5.1 Features from LLM

The results from the LLM demonstrate the influence of varying prompting contexts on the features proposed. Table 5 details the approach (Few-shot prompt type), the prompting context, and the corresponding features suggested by the LLM. We evaluated three few-shot prompts of: Temporal Context (FS1), Sequential Context (FS2), and Temporal and Sequential Contexts (FS3).

Table 2: Features by Different Prompting Contexts

Approach	Context	Features Proposed
Few-shot (FS1)	Temporal Context	- Acceleration - Jerk - Energy - Angle - Frequency
Few-shot (FS2)	Sequential Context	- Acceleration - Jerk - Joint Distance Ratios - Joint Pair Angle Rates
Few-shot (FS3)	Temporal and Sequential	- Acceleration - Jerk - Energy - Angle - Frequency - Joint Distance Ratios - Joint Pair Angle Rates

From each prompt, we filtered the generated features by relevance resulting in a unique combination of features, with FS1 suggesting 5 features, FS2 suggesting 4 features, and FS3 combining both contexts to suggest 7 features. This comparison underscores the advantage of combining contexts to produce the most comprehensive feature set.

5.2 Activity Recognition with Features

The feature suggestions were evaluated for their effectiveness in addressing the challenges of GTF activity recognition, as shown in Table 3. For comparison, the baseline (Base) results were supplemented with additional features (FS) that were not derived from the prompting structure, providing a point of comparison against the proposed few-shot prompting approaches FS1, FS2, and FS3.

Table 3: Comparison of Feature Performance, F1-score

Activity	**Baseline**	**FS**	**FS1**	**FS2**	**FS3**	**FS3, DS**
Explanation to patient	0.44	0.38	0.39	0.44	0.40	0.51
Confirm necessary items	0.72	0.76	0.73	0.70	0.63	0.00
Disinfect hands	0.43	0.50	0.36	0.32	0.50	0.31
Wearing gloves	0.62	0.66	0.65	0.57	0.66	0.79
Prepare the nutrition solution	0.64	0.65	0.60	0.58	0.70	0.76
Check the gastronomy site	0.43	0.47	0.44	0.44	0.41	0.71
Open the gastronomy cap	**0.00**	**0.11**	**0.10**	**0.00**	**0.29**	**0.00**
Inject lukewarm water	0.73	0.71	0.66	0.75	0.75	0.90
Connect the nutrition tube	0.41	0.43	0.32	0.30	0.47	0.81
Adjust the infusion rate	0.94	0.94	0.91	0.97	0.93	0.94
Removal of gloves	**0.00**	**0.00**	**0.00**	**0.12**	**0.19**	**0.00**
Prepare lukewarm water	0.62	0.69	0.69	0.62	0.71	0.00
Close the clamp	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Disconnect the nutrition tube	0.42	0.41	0.45	0.34	0.80	0.00
Close the gastronomy cap	**0.00**	**0.00**	**0.06**	**0.00**	**0.15**	**0.00**
Clean up used items	0.32	0.44	0.33	0.34	0.45	0.00
others	0.42	0.44	0.30	0.38	0.45	0.40
Total F1-Score	**0.54**	**0.55**	**0.50**	**0.51**	**0.57**	**0.55**

FS3 significantly improved recognition for short-duration activities specificallt *"Open the gastronomy cap"* (F1 = 0.29), *"Removal of gloves"* (F1 = 0.19), and *"Close the gastronomy cap"* (F1 = 0.15) compared to other approaches. Certain short-duration activities, like *"Close the clamp"*, remained difficult to classify accurately, with F1-scores of 0.00 across all methods. Overall, FS3 achieved the highest total F1-score of 0.57, demonstrating the advantages of combining temporal and sequential contexts in feature generation.

To validate the significance of our improvements, we conducted a paired t-test comparing the per-activity F1-scores between the baseline and FS3. The analysis revealed a statistically significant difference (t = -7.0994, p = 0.0021) with a very large effect size (Cohen's d = 1.8178). This statistical evidence confirms that the improvements achieved through FS3 are not merely due to random variation but represent meaningful enhancements in the model's ability to recognize nursing activities. The low p-value ($p < 0.01$), coupled with the large effect size, demonstrates that the consistent gains across multiple activity categories represent a substantial advancement in nursing activity recognition

performance.

Additionally, we used the same prompts to compare guided feature extraction using DeepSeek (DS). However, compared to FS3, its performance did not show significant improvement, and some activities even exhibited lower F1-Scores.

Table 4: Comparison of Model Performance with FS3 of Few-Shot, F1-score

Activity ID	**RF**	**SVM**	**LSTM1**	**LSTM2**	**BiLSTM1**	**BiLSTM2**
0	0.40	0.55	0.51	0.00	0.54	0.55
1	0.63	0.73	0.76	0.00	0.80	0.88
2	0.50	0.37	0.19	0.20	0.18	0.00
3	0.66	0.63	0.47	0.41	0.49	0.54
4	0.70	0.45	0.46	0.28	0.50	0.61
5	0.41	0.29	0.35	0.03	0.30	0.00
6	**0.29**	0.34	0.00	0.00	0.04	0.00
7	0.75	0.76	0.66	0.69	0.66	0.75
8	0.47	0.43	0.18	0.07	0.31	0.00
9	0.93	0.28	0.43	0.71	0.62	0.81
10	**0.19**	0.00	0.19	0.00	0.11	0.00
11	0.71	0.72	0.39	0.00	0.47	0.00
12	**0.00**	0.00	0.00	0.00	0.00	0.00
13	0.80	0.67	0.20	0.00	0.28	0.00
14	**0.15**	0.14	0.22	0.00	0.15	0.00
15	0.45	0.29	0.16	0.00	0.22	0.00
16	0.45	0.39	0.24	0.49	0.27	0.45
Total F1-Score	**0.57**	**0.51**	**0.42**	**0.33**	**0.46**	**0.49**

In contrast to our approach, advanced models like LSTM and BiLSTM, although powerful in capturing temporal relationships, proved less effective for the specific characteristics of our GTF activity recognition task. The core issue lies in the fuzzy boundaries between actions and the fast-switching nature of activities, which pose significant challenges for accurate segmentation. When the window size is set too small (window_size = 2), the integrity and contextual information of the action are lost, resulting in fragmented captures that miss key aspects of the action. Additionally, this configuration generates noise at action switching points, leading to a large number of erroneous samples. On the other hand, using a larger window size (window_size = 33 (1s)) causes short-term actions to be masked, as the window may encompass multiple actions, further reducing the data sample size and diluting the precision of the model's predictions.

While BiLSTM and LSTM models excel at capturing complex temporal relationships in some contexts, they are not ideal for our dataset where rapid action transitions and short-duration actions are common. Instead, our method that combines LLMs with Random Forest models better addresses these challenges. LLMs bring domain-specific knowledge into the feature extraction pro-

cess, helping to capture subtle, overlapping actions typical in medical procedures like GTF. Additionally, Random Forest models effectively handle the high-dimensional feature space and are resilient to noise, making them less prone to overfitting than time-series models like LSTM. By leveraging prior medical knowledge, our method strikes a unique balance between machine learning capabilities and domain expertise, making it well-suited for GTF activity recognition, especially in scenarios with limited labeled data.

Table 5: Features from Other Prompting Techniques

Context	**Few-Shot**	**Task Decomposition**	**Chain-of-Thought**
Temporal	Acceleration, Jerk, Energy, Angle, Frequency	Velocity, Acceleration, Cumulative velocity, Joint trajectory curvature, Frame-to-frame pose change, Time-lagged joint angles, Velocity ratio across windows	Joint Angle Velocity, Velocity Ratio, Acceleration, Movement Curvature, Multi-Scale Statistics, Joint Distance Change, Movement Magnitude
Sequential	Acceleration, Jerk, Joint Distance Ratios, Joint Pair Angle Rates	Pose changes between frames, Joint position continuity index, Degree of simultaneous movement, Movement consistency, Sudden movement indicators	Pose Change Rate, Movement Direction Consistency, Joint Coordination Index, Movement Continuity Metrics
Combined	Acceleration, Jerk, Energy, Angle, Frequency, Joint Distance Ratios, Joint Pair Angle Rates	Velocity, Acceleration, Cumulative velocity, Joint trajectory curvature, Pose change rate, Time-lagged angles, Velocity ratio, Pose changes, Continuity index, Simultaneous movement, Consistency score, Sudden movement indicators	Joint Angle Velocity, Velocity Ratio, Acceleration, Movement Curvature, Multi-Scale Statistics, Joint Distance Change, Movement Magnitude, Pose Change Rate, Direction Consistency, Joint Coordination Index, Continuity Metrics

We compared three different prompting techniques: Few-Shot, Task Decomposition, and Chain-of-Thought for guided feature extraction. Few-Shot focuses on extracting basic features and capturing the core temporal and sequential dynamics of actions. It is effective for initial experiments and when computational resources are limited. On the other hand, Task Decomposition breaks actions down into smaller, more manageable components, helping isolate specific movements and transitions. Compared to Few-Shot, this method increases the complexity of the feature set and may enhance the model's ability to capture more detailed aspects of behavior. Lastly, Chain-of-Thought employs more advanced reasoning techniques, such as considering joint angle velocity and movement continuity. This method provides a deeper understanding of the coordination and intensity of movements, making it the most comprehensive approach.

5.3 Post-processing

To handle sequence errors in GTF activity recognition accuracy, we evaluated three post-processing methods: logical rule correction, sliding window smoothing based on majority voting, and a short-term action priority method (P3). Logical rule correction (P1) addresses common sequence errors such as action

interruptions, overlaps, and inserted actions, ensuring a more consistent action flow. Sliding window smoothing (P2), based on majority voting, mitigates noise in predictions but can obscure short-term actions, particularly those with IDs 6, 10, 12, and 14. To balance this, the short-term action priority method (P3) defines short-term action IDs and minimum durations, allowing the system to retain these brief actions while leveraging the smoothing method to improve the recognition of long-term activities. Combined, these approaches aim to address the unique challenges of recognizing both short- and long-duration actions in GTF procedures. Table 6 summarizes the resulting F1-score of the three methods applied on GTF pose data.

Table 6: Performance of Post-Processing Techniques with Few-Shot, F1-score

Activity	**Baseline**	**P1**	**P2**	**P3**
Explanation to patient	0.44	0.41	0.45	**0.45**
Confirm necessary items	0.72	0.63	0.70	**0.70**
Disinfect hands	0.43	0.50	0.32	**0.32**
Wearing gloves	0.62	0.67	0.69	**0.69**
Prepare the nutrition solution	0.64	0.70	0.71	**0.71**
Check the gastronomy site	0.43	0.36	0.57	**0.60**
Open the gastronomy cap	0.00	0.10	0.00	**0.25**
Inject lukewarm water	0.73	0.42	0.75	**0.76**
Connect the nutrition tube	0.41	0.43	0.46	**0.46**
Adjust the infusion rate	0.94	0.93	0.97	**0.97**
Removal of gloves	0.00	0.14	0.00	**0.17**
Prepare lukewarm water	0.62	0.71	0.76	**0.88**
Close the clamp	0.00	0.19	0.00	**0.00**
Disconnect the nutrition tube	0.42	0.19	0.84	**0.88**
Close the gastronomy cap	0.00	0.17	0.00	**0.15**
Clean up used items	0.32	0.30	0.40	**0.40**
others	0.42	0.47	0.49	**0.49**
Total F1-Score	0.54	0.49	0.59	**0.60**

P3 significantly improved the recognition of short-duration activities such as *"Open the gastronomy cap"* (F1 = 0.25) and *"Removal of gloves"* (F1 = 0.17), which had an F1-score of 0.00 in the baseline. This also include *"Prepare the nutrition solution"* and *"Disconnect the nutrition tube"*, where P3 achieved higher F1-scores of 0.71 and 0.88, respectively, indicating handling performance across multiple GTF activities. On the other hand, persistent challenging activities like *"Close the clamp"* continued to have low F1-scores, suggesting that further improvements may be needed for specific short-duration tasks.

The overall results indicate that P3 consistently achieved the highest F1-scores across most activities, leading to the highest total F1-score of 0.60, compared to 0.54 in the baseline. The confusion matrices highlight the improvements: Fig. 8 shows the baseline model's struggles with short-duration activities, while Fig. 9 demonstrates P3 post-processing's significant improvements.

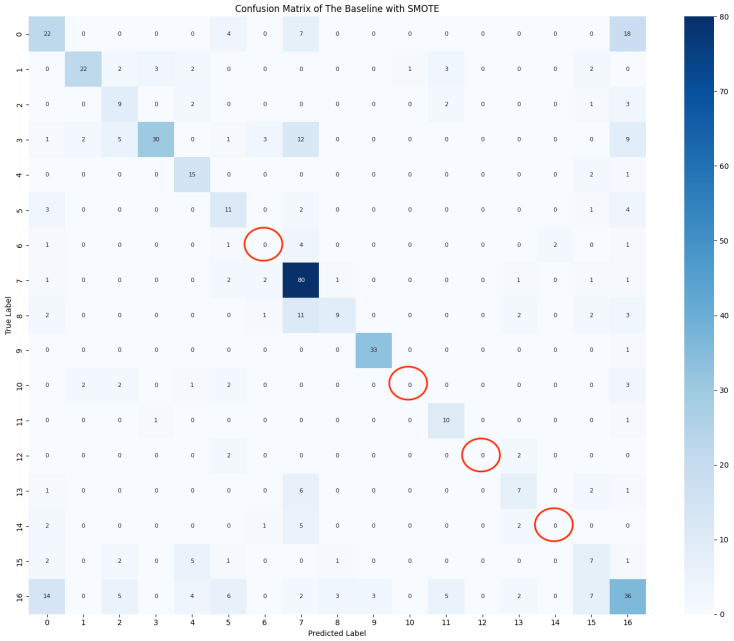

Figure 8: Confusion Matrix of Baseline.

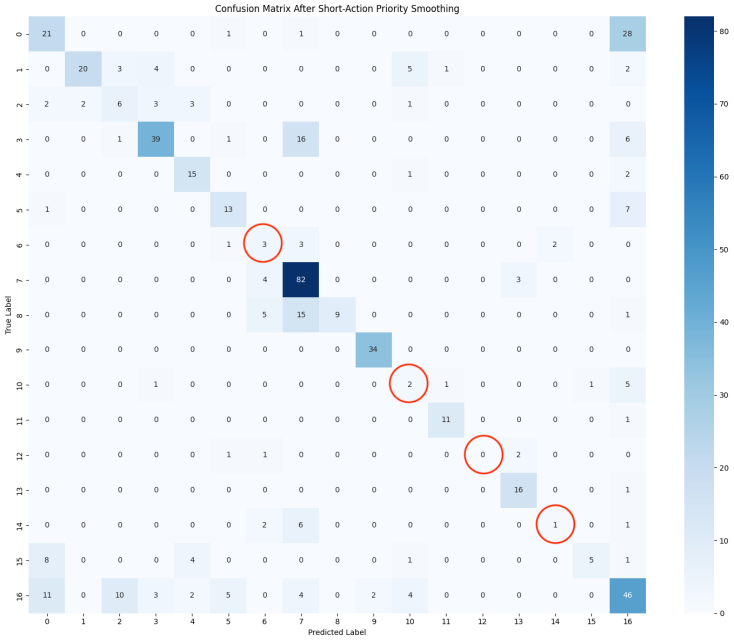

Figure 9: Confusion Matrix of P3 post-processing, Few-Shot.

Table 7: Performance of Post-Processing Techniques with CoT, F1-score

Activity	**TD***	**CoT***	**P1**	**P2**	**P3**
Explanation to patient	0.14	0.48	0.48	0.46	0.47
Confirm necessary items	0.00	0.79	0.79	0.81	0.81
Disinfect hands	0.41	0.35	0.35	0.18	0.18
Wearing gloves	0.77	0.64	0.64	0.57	0.57
Prepare the nutrition solution	0.82	0.59	0.59	0.56	0.56
Check the gastronomy site	0.00	0.00	0.00	0.00	0.00
Open the gastronomy cap	0.84	0.40	0.15	0.00	0.00
Inject lukewarm water	0.97	0.93	0.94	0.89	0.90
Connect the nutrition tube	0.24	0.81	0.81	0.83	0.83
Adjust the infusion rate	1.00	0.96	0.97	0.97	0.97
Removal of gloves	0.00	0.00	0.00	0.00	0.00
Prepare lukewarm water	0.59	0.80	0.80	0.92	0.92
Close the clamp	0.00	0.00	0.15	0.00	0.00
Disconnect the nutrition tube	0.94	0.36	0.17	0.21	0.21
Close the gastronomy cap	0.00	0.33	0.33	0.00	0.33
Clean up used items	0.54	0.80	0.57	0.78	0.78
others	0.41	0.45	0.46	0.49	0.50
Total F1-Score	**0.54**	**0.62**	**0.61**	**0.59**	**0.60**

*resulting performance using sequential and temporal context (FS3)

Based on the results in Table 7, we evaluated the impact of different prompting techniques and post-processing methods on GTF activity recognition. The comparison between Task Decomposition (TD) and Chain-of-Thought (CoT) demonstrates that CoT prompting yields superior feature representations, improving the F1-score from 0.57 to 0.62. This suggests that the features generated using CoT prompts capture more relevant contextual and sequential information, leading to better recognition performance.

To further refine the predictions, we applied three post-processing methods: Logical Rule Correction (P1), Sliding Window Smoothing (P2), and the Action Priority Method (P3). However, none of these approaches improved the overall F1-score, with all leading to slight performance degradation. The Logical Rule Correction method (P1) slightly decreased the F1-score to 61%, likely due to rigid predefined rules that failed to generalize across real-world variations in nursing behavior. The Sliding Window Smoothing method (P2) resulted in a more pronounced decline to 59%, as it smoothed out short but critical actions, reducing their distinguishability. The Action Priority Method (P3) improved upon P2 by preserving short-duration activities, leading to a slightly better F1-score of 60%, but it still suffered from the drawback of overlooking contextual dependencies.

6 Discussion on dealing with GTF activities

6.1 Gastrostomy Tube Feeding Nurse Activity

Gastrostomy Tube Feeding(GTF) is a complex and highly variable nursing activity that poses unique challenges for activity recognition compared to routine nursing tasks. Firstly, GTF activities are typically sequential, consisting of a series of clearly ordered steps such as disinfection, injecting nutrition solution, and adjusting the infusion rate. This sequential nature underscores the critical importance of temporal dependencies between actions, as any omission or misordering of steps can significantly impact patient safety and care quality. Certain steps lasting only a few seconds, with subtle movements localized to the hands. In contrast, other steps may last several tens of seconds or even longer. This variation in action duration imposes additional demands on recognition models, requiring them to capture subtle signals of short-duration actions within the context of longer actions.

The sliding window smoothing method effectively reduces prediction noise through a majority voting mechanism, which helps mitigate the issue of short-duration actions being overshadowed by long-duration actions. However, the complex temporal structure of GTF activities—such as hierarchical step dependencies and nonlinear temporal relationships—presents more profound challenges for existing methods. Specifically, the sliding window smoothing assumes consistency of action labels within a local time window, but this assumption may fail in the following scenarios:

1. Overlapping Steps and Parallel Operations

   Some steps in GTF may overlap briefly, such as a nurse adjusting the infusion rate (a long-duration action) while simultaneously closing the clamp (a short-duration action). The majority voting mechanism of the sliding window tends to select the label of the longer action, which occupies a larger portion of the window, thus ignoring the short-duration action. This frequency-based smoothing strategy struggles to capture the instantaneous nature of simultaneous actions.

2. Hierarchical Temporal Dependencies

   The sequential nature of GTF activities not only appears in linear step sequences but may also involve nested sub-steps (e.g., “preparing the nutrition solution” involves several micro-operations). The sliding window only focuses on local time segments and cannot model the global hierarchy or long-term dependencies across windows. For example, the correct identification of “closing the gastrostomy cap” may require reference to the completion status of the preceding step (e.g., “opening the gastrostomy cap”), which current methods lack the ability to process contextually.

3. Instantaneous Keyframes in Long-duration Actions

   Some long-duration actions (e.g., “checking the gastrostomy site”) may contain critical instantaneous sub-actions (e.g., pressing a specific area

with fingers). The smoothing process of the sliding window may dilute the signal of these keyframes, causing the model to fail to distinguish subtle but important changes in action.

These limitations highlight the shortcomings of traditional temporal modeling methods in complex clinical scenarios. While the sliding window smoothing improves the overall F1 score, it remains a heuristic rule, difficult to adapt to the dynamic temporal logic in GTF tasks. Future research should explore more advanced temporal modeling architectures, such as Transformer-based models, whose self-attention mechanism can explicitly capture long-range dependencies and hierarchical structures. Additionally, incorporating Graph Neural Networks (GNN) to model logical relationships between actions or introducing temporal logic constraints (e.g., action state machines) may further optimize recognition accuracy, especially when handling overlapping actions and nested steps. Such methods are expected to address the shortcomings of sliding window smoothing and provide more granular temporal understanding for GTF activity recognition.

A further limitation of the current research is the inability to consistently estimate nurses' intentions or clinical judgments based solely on their observed activities. While activity recognition models can identify the physical aspects of GTF, understanding the rationale or decision-making processes behind these actions remains an open challenge.

6.2 Comparison with General Nursing and Non-Healthcare Activities

Compared to most conventional nursing activities (such as repositioning and medication administration), Gastrostomy Tube Feeding (GTF) presents three distinct challenges.

First, GTF follows a strict sequential order, where each step must adhere to specific clinical protocols. Any deviation from the correct sequence may increase the risk of infection. This requirement for strict procedural adherence is particularly crucial in the design of temporal models, as conventional classification or detection methods often struggle to incorporate such rigid sequential logic.

Second, the duration of GTF activities varies significantly, encompassing both extremely short actions (lasting less than 3 seconds) and prolonged actions (lasting much longer). This large variation in action duration poses additional challenges for the model. In short-duration actions, movements are often subtle and easily overlooked, while long-duration actions require enhanced temporal sequence processing capabilities to accurately track and recognize transitions between different actions.

Finally, the issue of localized signal occlusion is particularly prominent in the GTF scenario. Many critical actions, such as "Close the clamp," primarily involve hand movements, which are often occluded by medical tools or parts of the patient's body. Traditional pose estimation methods, especially those based on full-body skeleton detection, tend to suffer from missing key points

or accumulated errors under such conditions. For example, occlusion of arm or hand joints near the patient may result in incomplete or unstable skeleton detections, further affecting the model's ability to distinguish short-duration critical actions.

These three characteristics distinguish GTF activity recognition from other nursing tasks, which generally involve larger movements and have lower temporal dependencies [26]. In theory, GTF recognition requires models to integrate strict temporal logic, adaptability to variable time scales, and fine-grained tracking and correction of local limb key points.

Unlike general nursing activities, each category of GTF activities is often further subdivided into finer-grained micro-labels. The presence of these micro-labels provides more detailed supervisory signals for the model but simultaneously increases the complexity of annotation and recognition. This multi-category, sequential nature, coupled with the use of micro-labels, places higher technical demands on temporal data processing, feature extraction, and post-processing optimization in GTF activity recognition.

Although GTF activities and endotracheal suctioning(ES) activities[2, 3] share nearly identical characteristics, the challenges in GTF lie in the presence of instantaneous actions and overlapping contexts. For example, actions such as "close the clamp" in GTF are typically instantaneous, involving minimal motion, which inherently makes them difficult to recognize. Furthermore, the clamp is small and often positioned along the feeding tube in locations prone to occlusion by other objects or the nurse's own body. Additionally, such actions may occur simultaneously with other ongoing tasks, such as adjusting the infusion rate or repositioning equipment, leading to overlapping visual and motion cues. In contrast, ES actions, such as "open/close the cap," although relatively short in duration, are less likely to be occluded or performed within overlapping contexts, thereby simplifying their recognition.

6.3 Contribution of Using LLM for GTF

The analysis of feature importance underscores the significant contribution of features suggested by large language models (LLMs) in improving the efficacy of nursing activity recognition. By integrating domain-specific expertise with the flexibility of LLMs, we were able to identify nuanced and previously unrecognized motion patterns that are essential for differentiating between similar actions. For instance, features pertaining to joint velocities and angular variations around the hands proved to be vital for the identification of short-term tasks, such as the opening or closing of a gastrostomy cap. These features not only enhanced the model's capacity to detect subtle actions but also played a role in decreasing misclassification rates in intricate scenarios. The incorporation of LLM-guided feature suggestions facilitated the optimization of the feature space, resulting in a more robust representation of nursing activities compared to the exclusive use of handcrafted features.

Unlike approaches that rely solely on human expertise or end-to-end deep learning models, large language models (LLMs) leverage their contextual rea-

soning and cross-domain knowledge generalization capabilities during feature extraction and model design [22, 23]. Given the challenges posed by GTF activities, including strict temporal dependencies, variations in action duration, and hand occlusions, traditional methods often struggle to capture subtle short-duration movements or maintain global temporal consistency for long-duration actions when lacking sufficient domain priors.

By utilizing natural language prompting, LLMs can automatically generate more targeted features across different time scales. For instance, they can extract hand acceleration and angular variation features for short-duration actions or sequence dependency measures for long-duration actions. Additionally, LLMs can incorporate relevant cross-domain insights learned from pretraining, aiding the model in recognizing critical steps in GTF activities.

Moreover, the generalization ability of LLMs allows them to provide reasonable feature suggestions even when data availability is limited or domain-specific samples are scarce [24, 25]. This reduces reliance on manually designed features while improving recognition accuracy. Therefore, in the complex scenario of GTF activity recognition—characterized by high temporal dependencies, uneven motion amplitudes, and localized occlusions—leveraging LLMs to guide feature selection can significantly enhance the capture of short-duration movements and maintain temporal consistency in long-duration actions, thereby addressing multiple challenges associated with GTF recognition.

The application of LLMs in this study represents a pivotal advancement for GTF activity recognition, addressing challenges unique to this domain. Unlike traditional feature engineering, LLMs dynamically adapted to the nuances of GTF tasks by leveraging contextual knowledge to suggest domain-relevant features. For instance, LLM-generated features focusing on hand movements, such as joint angles and velocities, were instrumental in distinguishing subtle GTF actions. This approach not only optimized the feature space but also demonstrated the transformative potential of LLMs in addressing the intricate requirements of GTF activity recognition.

7 Conclusion

In this study, we explored how context in LLM Prompting affect generating features for improving nurse activity recognition. Results show that the generated features from LLMs is influenced by the context in the prompting strategy and affects the model performance.

We proposed a comprehensive framework integrating video-based pose estimation, LLM-guided feature engineering, and post-processing techniques to address challenges in gastrostomy tube feeding (GTF) activity recognition. Pose estimation data was extracted using YOLO11, followed by balancing the dataset with SMOTE to mitigate class imbalances. LLMs were employed to suggest novel, domain-specific features, which were integrated with handcrafted features. A Random Forest classifier was used for recognition, with post-processing applied to refine predictions, particularly for short-duration actions.

The experimental results validate the effectiveness of the proposed framework. Compared to the baseline model, the Random Forest model improved accuracy in nurse activity recognition using pose estimation in GTF from 55% to 58%, and the F1-score from 54% to 57%. Furthermore, incorporating post-processing techniques with priority handling enhanced short-term action detection, leading to an additional 3% gain in performance metrics.

Future research should focus on exploring advanced temporal modeling approaches, such as Transformers or hybrid neural architectures, to better capture temporal dependencies in nursing activities. The integration of generative adversarial networks (GANs) or other advanced data augmentation techniques could further enhance data diversity and improve model robustness. Expanding the framework to encompass additional clinical procedures and validating its performance across diverse healthcare settings would provide greater insights into its generalizability and utility.

Moreover, while domain-specific Large Language Models (LLMs), such as BioGPT and MedPaLM, have shown great potential in improving feature extraction for specialized tasks like nurse activity recognition in gastroenterostomy tube feeding (GTF), their integration into our current framework presents certain challenges. These models are often limited by access constraints, such as API or pre-trained model availability, and resource limitations in real-time processing environments. Furthermore, incorporating these models would require substantial computational resources and task-specific fine-tuning, adding complexity to the pipeline and increasing the time-to-deployment. Thus, investigating the integration of domain-specific LLMs will be an important avenue for future work, aiming to refine feature extraction for GTF activity recognition and improve the overall framework.

References

[1] Islam, S., Hossain, S. M. H., Uddin, M. Z., Hossain, S., and Ahad, M. A. R. "Enhancing Nursing Activity Recognition During Endotracheal Suctioning Through Video-based Pose Estimation and Machine Learning". International Journal of Activity and Behavior Computing, 2024(3), pp. 1–15. https://doi.org/10.60401/ijabc.36

[2] Ngo, H. A. V., Vu, Q. N. P., Colley, N., Ninomiya, S., Kanai, S., Komizunai, S., Konno, A., Nakamura, M., and Inoue, S. "Toward recognizing nursing activity in endotracheal suctioning using video-based pose estimation". International Journal of Activity and Behavior Computing, 2024(1). https://doi.org/10.60401/ijabc.1

[3] Ngo, H. A. V., Colley, N., Ninomiya, S., Kanai, S., Komizunai, S., Konno, A., Nakamura, M., and Inoue, S. "Nurses' Skill Assessment in Endotracheal Suctioning Using Video-based Activity Recognition". International Journal of Activity and Behavior Computing, 2024(2), pp. 1–24. https://doi.org/10.60401/ijabc.20

[4] Garcia, C., and Inoue, S. "Challenges and Opportunities of Activity Recognition in Clinical Pathways." In Human Activity and Behavior Analysis. 2024. CRC press. https://doi.org/10.1201/9781003371540-8

[5] Ngo, H. A. V., Kaneko, H., Hassan, I., Ronando, E., Shoumi, M. N., Munemoto, R., Hossain, T., and Inoue, S. "Summary of the nurse care activity recognition challenge using skeleton data from video with generative AI". International Journal of Activity and Behavior Computing, 2024(3), pp. 1–20. https://doi.org/10.60401/ijabc.31

[6] Liu, J., Mu, X., Liu, Z. et al. "Human skeleton behavior recognition model based on multi-object pose estimation with spatiotemporal semantics". Machine Vision and Applications 34, 44(2023). https://doi.org/10.1007/s00138-023-01396-0

[7] 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 (September 2019), 24 pages. https://doi.org/10.1145/3351244

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

[9] Khanam, R. and Hussain, M. "YOLOv11: An Overview of the Key Architectural Enhancements". 2024. doi: 10.48550/arXiv.2410.17725.

[10] Dobhal, U., Garcia, C., and Inoue, S. "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). Association for Computing Machinery, New York, NY, USA, pp. 493–499. https://doi.org/10.1145/3675094.3678445

[11] Miyake, N., Kaneko, H., Ronando, E., Garcia, C., Inoue, S. "Toward Detecting and Explaining Stress of Nurses Using Wearable Devices and LLMs". Proceedings of the International Conference on Ubiquitous Computing and Ambient Intelligence (UCAmI 2024). UCAmI 2024. Lecture Notes in Networks and Systems, vol 1212. Springer, Cham. https://doi.org/10.1007/978-3-031-77571-0_28

[12] Kaneko, H., and Inoue, S. "Toward Pioneering Sensors and Features Using Large Language Models in Human Activity Recognition". In Adjunct

Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing (UbiComp/ISWC '23 Adjunct). Association for Computing Machinery, New York, NY, USA, pp. 475–479. https://doi.org/10.1145/3594739.3610741

[13] Ronando, E., and Inoue, S. "Improving Fatigue Detection with Feature Engineering on Physical Activity Accelerometer Data Using Large Language Models". International Journal of Activity and Behavior Computing, 2024(2), pp. 1–22. https://doi.org/10.60401/ijabc.18

[14] Shoumi, M. N., and Inoue, S. "Leveraging the Large Language Model for Activity Recognition: A Comprehensive Review". International Journal of Activity and Behavior Computing, 2024(2), pp. 1–27. https://doi.org/10.60401/ijabc.21

[15] Nag, S., Zhu, X., Song, Y-Z., and Xiang, T. Post-processing temporal action detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18837-18845, 2023. https://doi.org/10.1109/CVPR52729.2023.01806

[16] Tran, M. T., Vu, M. Q., Hoang, N. D., and Bui, K-H. N. "An Effective Temporal Localization Method with Multi-View 3D Action Recognition for Untrimmed Naturalistic Driving Videos," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), New Orleans, LA, USA, 2022, pp. 3167-3172, doi: 10.1109/CVPRW56347.2022.00357.

[17] C. Wang, Y. Xu, H. Liang, W. Huang and L. Zhang, "WOODY: A Post-Process Method for Smartphone-Based Activity Recognition," in IEEE Access, vol. 6, pp. 49611-49625, 2018, doi: 10.1109/ACCESS.2018.2866872.

[18] Amatriain, X. "Prompt Design and Engineering: Introduction and Advanced Methods." arXiv preprint arXiv:2401.14423, 2024. https://arxiv.org/abs/2401.14423

[19] Javaheri, H., Ghamarnejad, O., Lukowicz, P., Stavrou, G. A., and Karolus, J., "LLMs Enable Context-Aware Augmented Reality in Surgical Navigation", Art. no. arXiv:2412.16597, 2024. doi:10.48550/arXiv.2412.16597.

[20] Gangwal, R. Types of Sampling and Sampling Techniques. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2019/09/data-scientists-guide-8-types-of-sampling-techniques/. Accessed July 10, 2024.

[21] Alwassel, H., Heilbron, F. C., Escorcia, V., and Ghanem, B. Diagnosing Error in Temporal Action Detectors. In Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part III. Springer-Verlag, Berlin, Heidelberg, 264–280. https://doi.org/10.1007/978-3-030-01219-9_16

[22] Li, Z., Deldari, S., Chen, L., Xue, H., and Salim, F. D. SensorLLM: Aligning large language models with motion sensors for human activity recognition. OpenReview. 2025. https://openreview.net/forum?id=cDd7kg9mkP

[23] Xi Chen, Julien Cumin, Fano Ramparany, Dominique Vaufreydaz. Towards LLM-Powered Ambient Sensor Based Multi-Person Human Activity Recognition. The 30th International Conference on Parallel and Distributed Systems, Oct 2024, Belgrade, Serbia. https://hal.science/hal-04619086v2

[24] Qu, H., Cai, Y., and Liu, J.LLMs are good action recognizers. arXiv. 2024. https://arxiv.org/abs/2404.00532

[25] Ray, L. S. S., Zhou, B., Suh, S., and Lukowicz, P. Initial findings on sensor-based open vocabulary activity recognition via text embedding inversion. arXiv. 2025. https://arxiv.org/abs/2501.07408

[26] Backman E, Granlund M, Karlsson AK. Parental Perspectives on Family Mealtimes Related to Gastrostomy Tube Feeding in Children. Qual Health Res. 2021;31(9):1596-1608. doi:10.1177/1049732321997133