Unsupervised Discovery of Long-Term Spatiotemporal Periodic Workflows in Human Activities

Fan Yang[1], Quanting Xie[3], Atsunori Moteki[2], Shoichi Masui[2],
Shan Jiang[2], Kanji Uchino[1], Yonatan Bisk[3], Graham Neubig[3]
[1] Fujitsu Research of America, USA [2] Fujitsu Limited, Japan [3] Carnegie Mellon University, USA

1. Visualization of Our Hard & Soft Tokenizations

In our study, we utilize two tokenization strategies for representing spatiotemporal human activity sequences: hard tokenization and soft tokenization. Hard tokenization assigns fixed discrete tokens to each frame of an activity sequence, which can sometimes lose nuanced information about the underlying pattern. Soft tokenization, on the other hand, provides a probabilistic or weighted representation that capture richer details about the transitions and gradual changes within periodic activities. Therefore, soft tokens are more suitable for estimating initial period window sizes, while hard tokens are more suitable for identifying discriminative temporal boundaries for period segmentation, remaining portion estimation, and anomaly localization. Figure 1 offers a visual comparison of these two tokenization strategies and their corresponding RGB features.

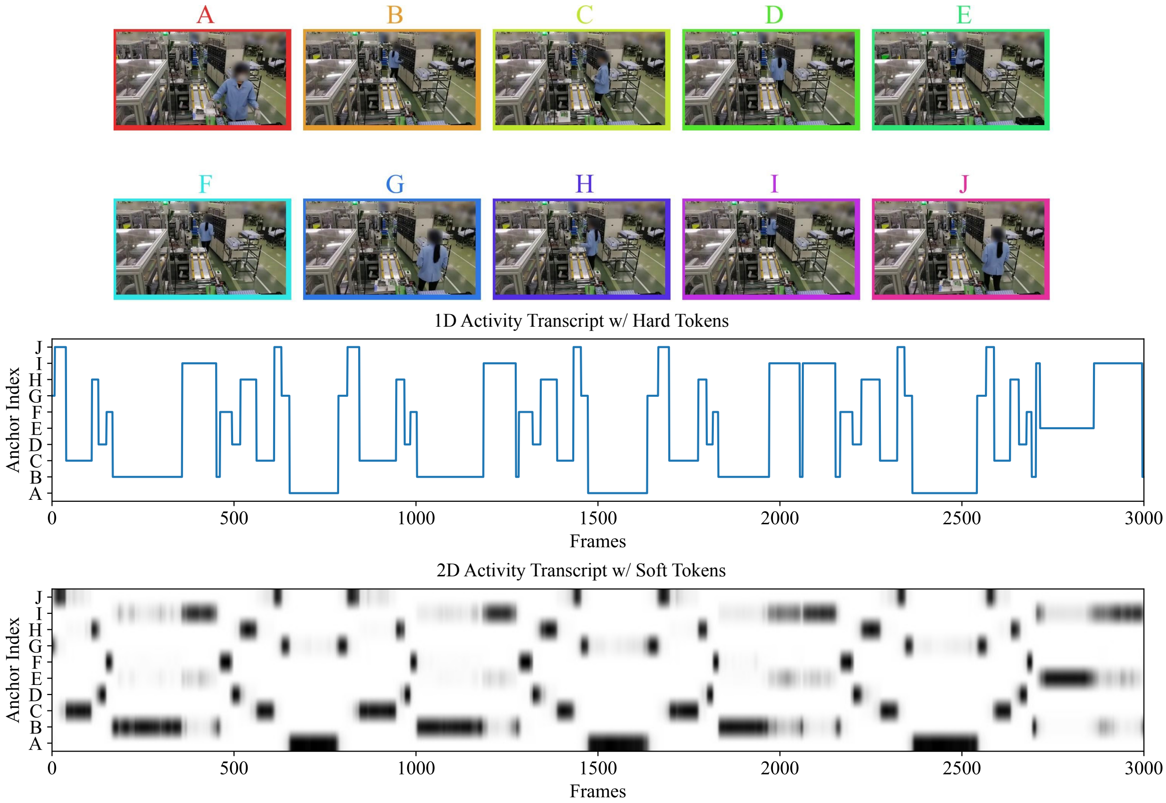


Figure 1. **Visualization of our hard & soft tokenizations for** $3,000$ **frames of an activity sequence.** Compared with hard tokenization, soft tokenization provides richer information about periodic patterns.

2. Visualization of Our Evaluation Metrics

We define three evaluation tasks using the provided spatiotemporal features: **1.** detect period counts and boundaries on normal periods through unsupervised methods; **2.** based on workflows obtained in task 1, predict the remaining phase proportion of an ongoing period when only partial data are available; **3.** based on workflows obtained in task 1, localize procedural anomalies within new periods. Rather than directly assessing the workflow itself, we incorporate workflow flexibility by evaluating its effectiveness through downstream tasks 2 and 3. After executing task 1 offline to derive the workflow, tasks 2 and 3 can be performed online. All three tasks are directly applicable to real-world scenarios. We illustrate how to use our metrics in Figure 2. For task 1, we adopt Mean Absolute Percentage Error (MAPE) to evaluate period counting accuracy. To enhance long-term period assessment, we utilize the average score of Temporal Intersection over Union (tIoU) to quantify boundary accuracy. Given that the number of estimated periods may differ from the GT, we apply Hungarian matching to identify the optimal alignment between predictions and GTs, maximizing the average scores in each sequence. For task 2, we employ Mean Absolute Error (MAE) to compare the estimated remaining proportion value and the GT. For task 3, we currently only assign one anomalous section per test period. Thus, unlike task 1, we do not apply Hungarian matching but instead, rely on tIoU to evaluate the overlap between predicted anomalous regions and GT in each sequence.

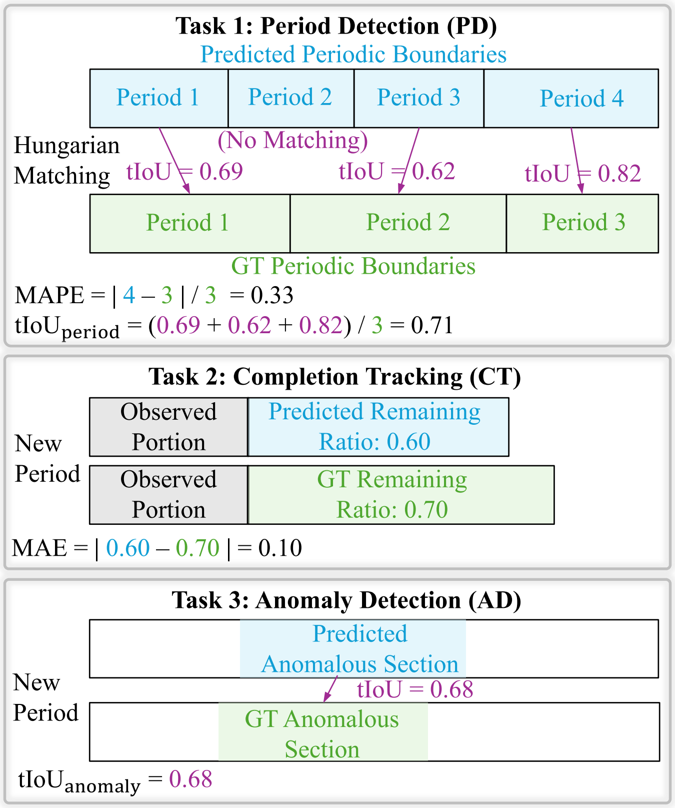

Task 1: Period Detection (PD)
Predicted Periodic Boundaries
Period 1
Period 2
Period 3
Period 4
Hungarian Matching
(No Matching)
tIoU = 0.69
tIoU = 0.62
tIoU = 0.82
Period 1
Period 2
Period 3
GT Periodic Boundaries
MAPE = | 4 – 3 | / 3 = 0.33
tIoU_period = (0.69 + 0.62 + 0.82) / 3 = 0.71
Task 2: Completion Tracking (CT)
New Period
Observed Portion
Predicted Remaining Ratio: 0.60
Observed Portion
GT Remaining Ratio: 0.70
MAE = | 0.60 – 0.70 | = 0.10
Task 3: Anomaly Detection (AD)
New Period
Predicted Anomalous Section
tIoU = 0.68
GT Anomalous Section
tIoU_anomaly = 0.68

Figure 2. **Illustrations of evaluation metrics.**

3. Visualization of Workflows Generated by Process Mining (PM) and Our Baseline Method

In our Related Work Section, we pointed out that many current periodic spatiotemporal mining methods inadvertently reflect Process Mining (PM) [2, 7] principles. In these approaches, identical activity tokens appearing in different positions within a workflow can create self-loops in the graph, which may obscure long-term temporal dependencies in the procedural analysis. Figure 3 illustrates a case where workflows are generated from a **toy transcript** using both the widely adopted PM toolbox PM4PY [1] and our baseline method. Notably, the workflow produced by the PM approach fails to capture long-term temporal dependencies, rendering it unsuitable for applications such as periodic boundary detection, completion tracking, and anomaly detection. In contrast, our baseline method captures these dependencies by ensuring that the workflow is linearly ordered, while also using a multi-branch structure to accommodate spatiotemporal variations.

Figure 4 further illustrates more complex workflows extracted from our benchmark **realistic dataset**. The workflows generated by our method not only reflect the correct temporal ordering of activities but also effectively handle cases where multiple spatial or temporal paths exist, making it a robust tool for long-term activity analysis. This representation structure of our workflow is critical for subsequent tasks like periodic boundary detection, completion tracking, and anomaly detection.

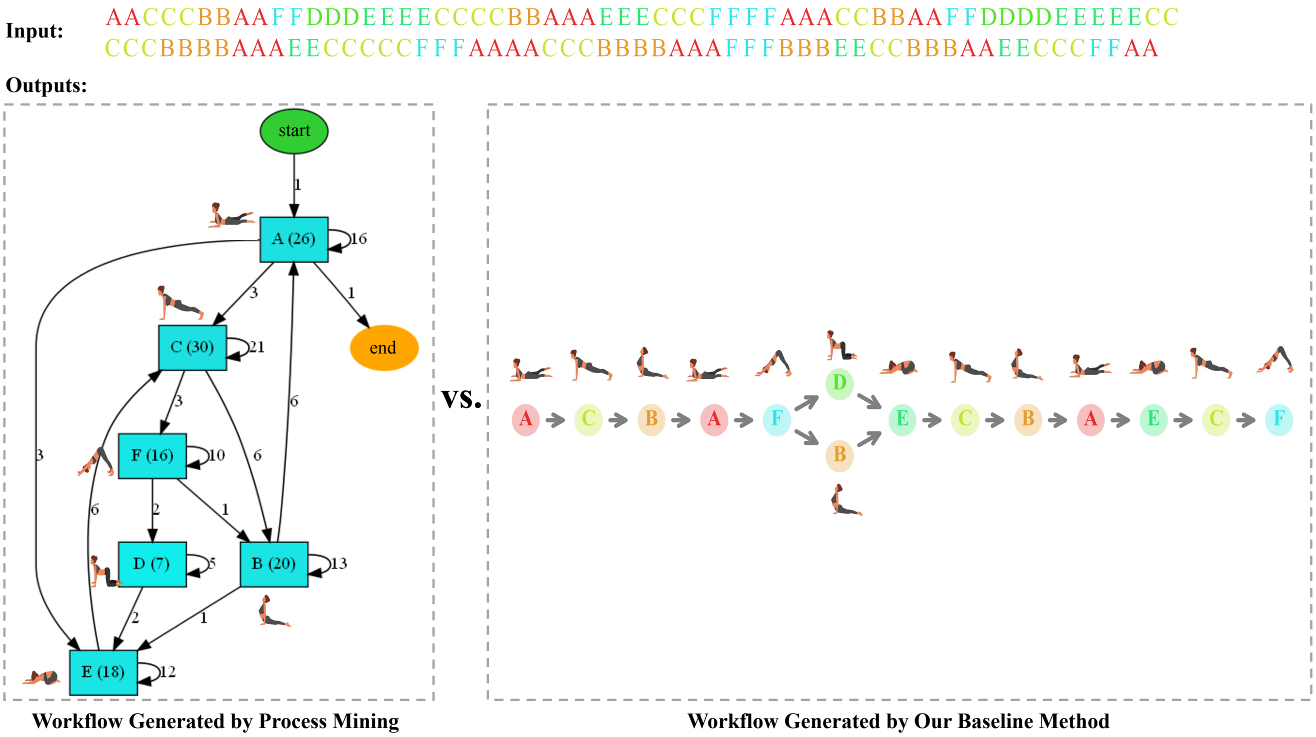
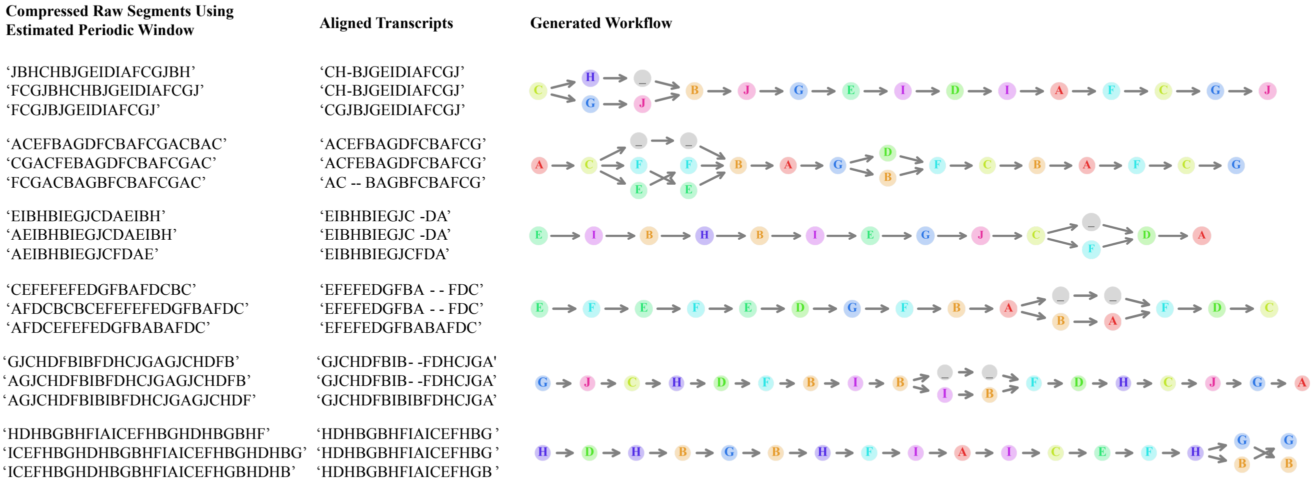

Input:
AACCCBBAAFFDDDEEEECCCCBBAAAEEECCCFFFFAAACCBBAAFFDDDDEEEEECC
CCCBBBBAAAEECCCCCFFFAAAACCCBBBBAAAFFFBBBEECCBBBAAEECCCFFAA
Outputs:
start
1
A (26)
16
3
1
end
C (30)
21
3
6
F (16)
10
6
3
6
2
1
D (7)
5
B (20)
13
2
1
E (18)
12
vs.
A
C
B
A
F
D
B
E
C
B
A
E
C
F
Workflow Generated by Process Mining
Workflow Generated by Our Baseline Method

Figure 3. **Comparison of workflows generated by Process Mining (PM) [2, 7] and our baseline method**. In PM, identical activity tokens positioned at different locations within a workflow can create self-loops in the graph. This may lead to the loss of long-term temporal dependencies, making it difficult to analyze long-term periodic workflow of human activities.



Compressed Raw Segments Using Estimated Periodic Window	Aligned Transcripts
‘JBHCHBJGEIDIAFCGJBH’ ‘FCGJBHCHBJGEIDIAFCGJ’ ‘FCGJBJGEIDIAFCGJ’	‘CH-BJGEIDIAFCGJ’ ‘CH-BJGEIDIAFCGJ’ ‘CGJBJGEIDIAFCGJ’
‘ACEFBAGDFCBAFCGACBAC’ ‘CGACFEBAGDFCBAFCGAC’ ‘FCGACBAGBFCBAFCGAC’	‘ACEFBAGDFCBAFCG’ ‘ACFEBAGDFCBAFCG’ ‘AC -- BAGBFCBAFCG’
‘EIBHBIEGJCDAEIBH’ ‘AEIBHBIEGJCDAEIBH’ ‘AEIBHBIEGJCFDAE’	‘EIBHBIEGJC -DA’ ‘EIBHBIEGJC -DA’ ‘EIBHBIEGJCFDA’
‘CEFEFEFEDGFBAFDCBC’ ‘AFDCBCBCEFEFEFEDGFBAFDC’ ‘AFDCEFEFEDGFBABAFDC’	‘EFEFEDGFBA - - FDC’ ‘EFEFEDGFBA - - FDC’ ‘EFEFEDGFBABAFDC’
‘GJCHDFBIBFDHCJGAGJCHDFB’ ‘AGJCHDFBIBFDHCJGAGJCHDFB’ ‘AGJCHDFBIBIBFDHCJGAGJCHDF’	‘GJCHDFBIB- -FDHCJGA' ‘GJCHDFBIB- -FDHCJGA’ ‘GJCHDFBIBIBFDHCJGA’
‘HDHBGBHFIAICEFHBGHDHBGBHF’ ‘ICEFHBGHDHBGBHFIAICEFHBGHDHBG’ ‘ICEFHBGHDHBGBHFIAICEFHGBHDHB’	‘HDHBGBHFIAICEFHBG ’ ‘HDHBGBHFIAICEFHBG ’ ‘HDHBGBHFIAICEFHGB ’

Figure 4. **Illustration of our extracted workflows in our benchmark**. Our extracted workflows are linearly ordered and capture long-term temporal dependencies. We also use the multi-branch structure to handle spatiotemporal diversities.

4. Visualization of Applications Using Our Tokenizations

In Figure 5, we illustrate that explicit semantic labels for activity tokens are unnecessary for effective workflow analysis. Instead of relying on action-to-text tokenization, which requires manual and often ambiguous labeling of complex human activities, we represent actions using arbitrary symbols (e.g., “A–Z”). This abstraction allows us to detect and monitor workflows in real time and apply them to practical scenarios. From a broader perspective on anomaly detection, which encompasses time overruns, omissions, and procedural errors, our system processes a live video stream, matches the detected tokens to a learned workflow, and identifies deviations, all without knowledge of the actual activity names.

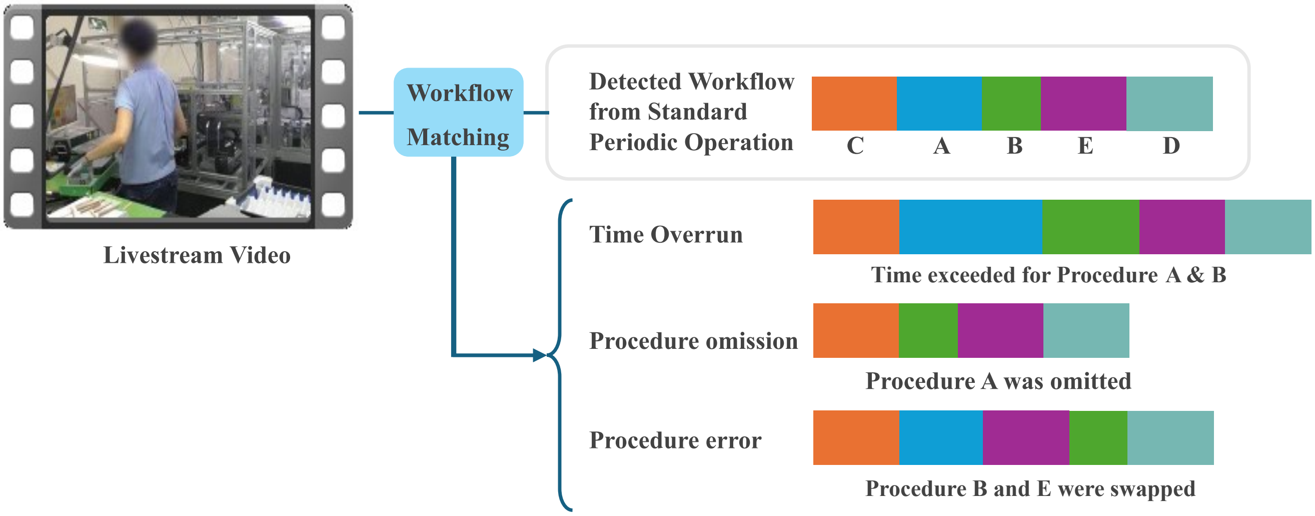

Livestream Video
Workflow Matching
Detected Workflow from Standard Periodic Operation
C
A
B
E
D
Time Overrun
Time exceeded for Procedure A & B
Procedure omission
Procedure A was omitted
Procedure error
Procedure B and E were swapped

Figure 5. **Unsupervised workflow detection enables downstream applications without semantic labels**.

5. Visualization of Our Normalization for Spatiotemporal Features

One of the challenges in periodic activity detection (PAD) is handling feature variance caused by camera movement. As demonstrated with the RepCount benchmark [4], even with a high-contrast periodic pattern, camera motion can obscure the underlying periodicity. In our experiments, we normalize the 3D body poses by:

1. Tracking the target person,
2. Removing camera movement,
3. Applying pose smoothing, and
4. Performing interpolation.

Figure 6 shows a test sample from RepCount, where the normalization steps effectively reduce feature variance. After normalization, our baseline method and LLM-based approaches are able to detect the short-term periodic patterns in an unsupervised manner, achieving superior performance compared to some supervised methods.

Nonetheless, instead of claiming a new state-of-the-art solution, we mainly want to show a common characteristic of existing short-term PAD benchmarks: an overemphasis on the feature engineering rather than a focus on modeling periodic patterns. In our benchmark, we provide unified normalized spatiotemporal features in the hope of shifting the focus from intensive feature engineering towards a comprehensive analysis of long-term periodic patterns. Moreover, Figure 7 illustrates samples of the normalized spatiotemporal features provided in our benchmark.

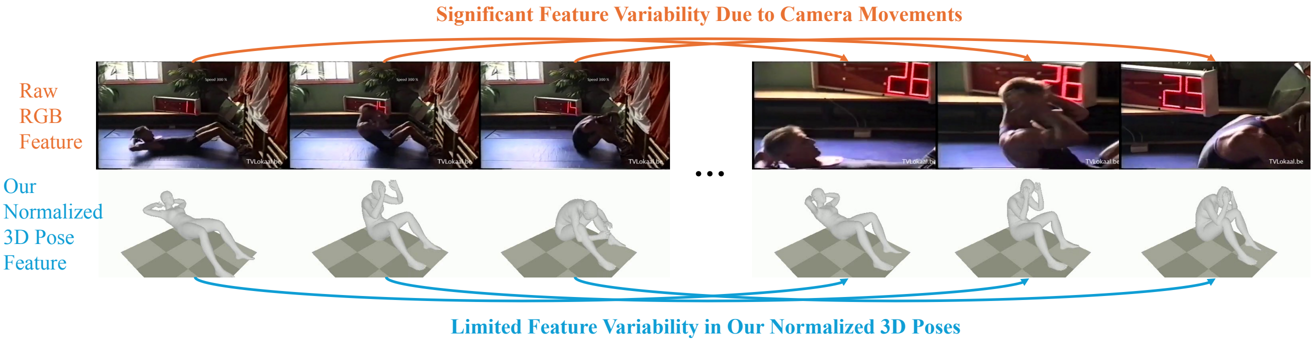

Significant Feature Variability Due to Camera Movements
Raw RGB Feature
Our Normalized 3D Pose Feature
...
Limited Feature Variability in Our Normalized 3D Poses

Figure 6. **Illustration of our normalization for spatiotemporal features in prior short-term PAD benchmark.** Camera movements pose a significant challenge to periodic analysis in conventional short-term PAD benchmark RepCount [4]. After applying our normalization—removing camera movement, tracking target person, applying pose smoothing, and performing interpolation—short-term periodic patterns can be easily detected in an unsupervised manner. However, we advocate for a shift in focus from feature engineering to the analysis of long-term periodic patterns by providing our benchmark.

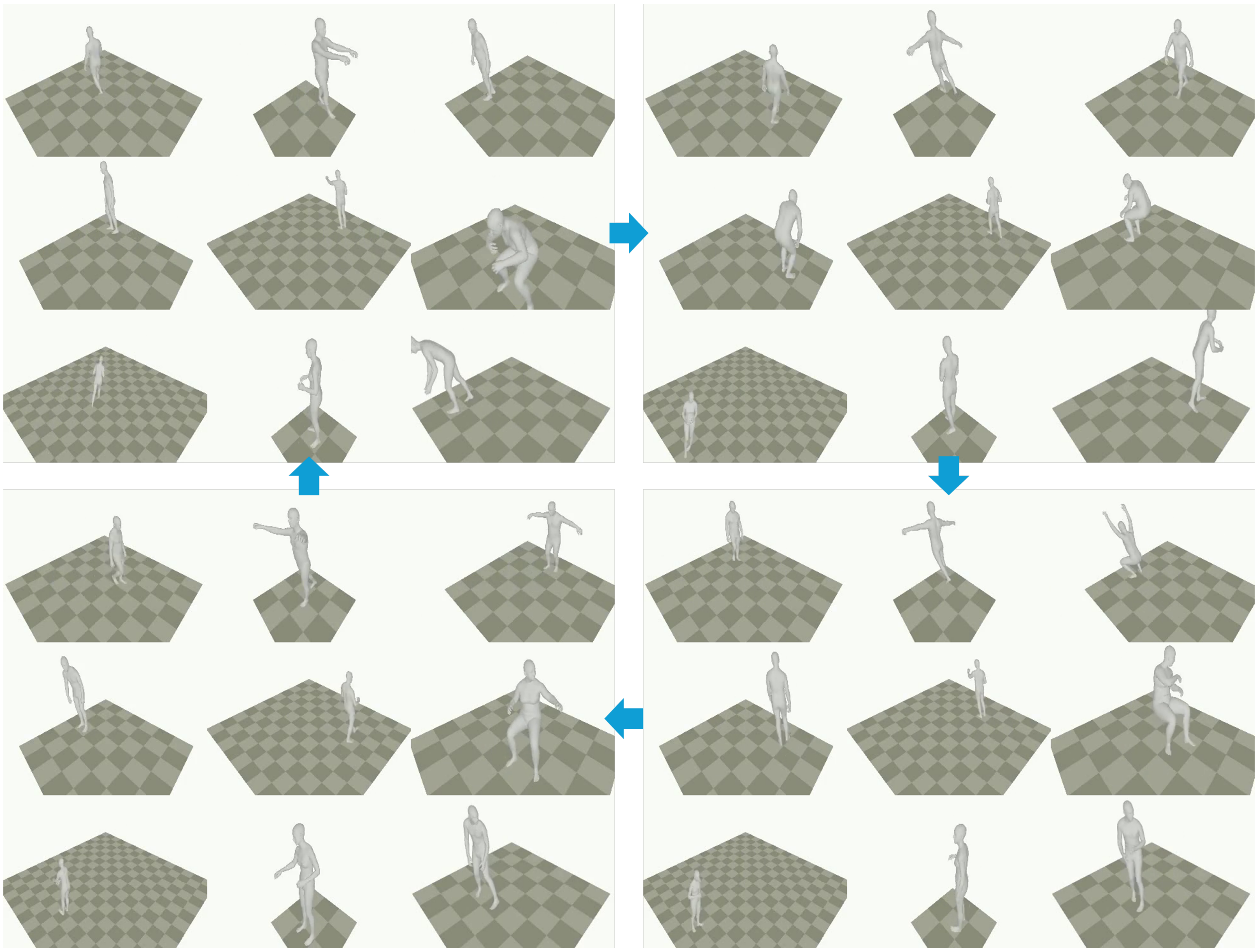


Figure 7. **Illustration of our normalized poses in our benchmark.** We provide normalized poses in our benchmark to eliminate feature variance from camera movement, allowing the evaluation to focus on modeling long-term periodic patterns.

6. Visualization Comparing Our Baseline Method and Gemini on Tasks 2 and 3

In our main paper, our quantitative analysis reveals that LLMs excel at counting periods in task 1; however, they encounter difficulties with task 2 (completion tracking) and task 3 (anomaly detection). Consequently, we infer that while LLMs are capable of recognizing repetitive patterns within sequences of hundreds of low-contrast tokens, they struggle to leverage this ability for the extended workflows demanded by tasks 2 and 3. To illustrate these findings further, Figure 8 and Figure 9 depict the performance of our baseline method alongside that of Gemini [3, 5, 6] on detailed examples of tasks 2 and 3, respectively. These results not only affirm the effectiveness of our baseline method but also expose the limitations of LLMs in addressing these tasks, emphasizing the critical need to develop specialized approaches for modeling long-term periodic workflows effectively.

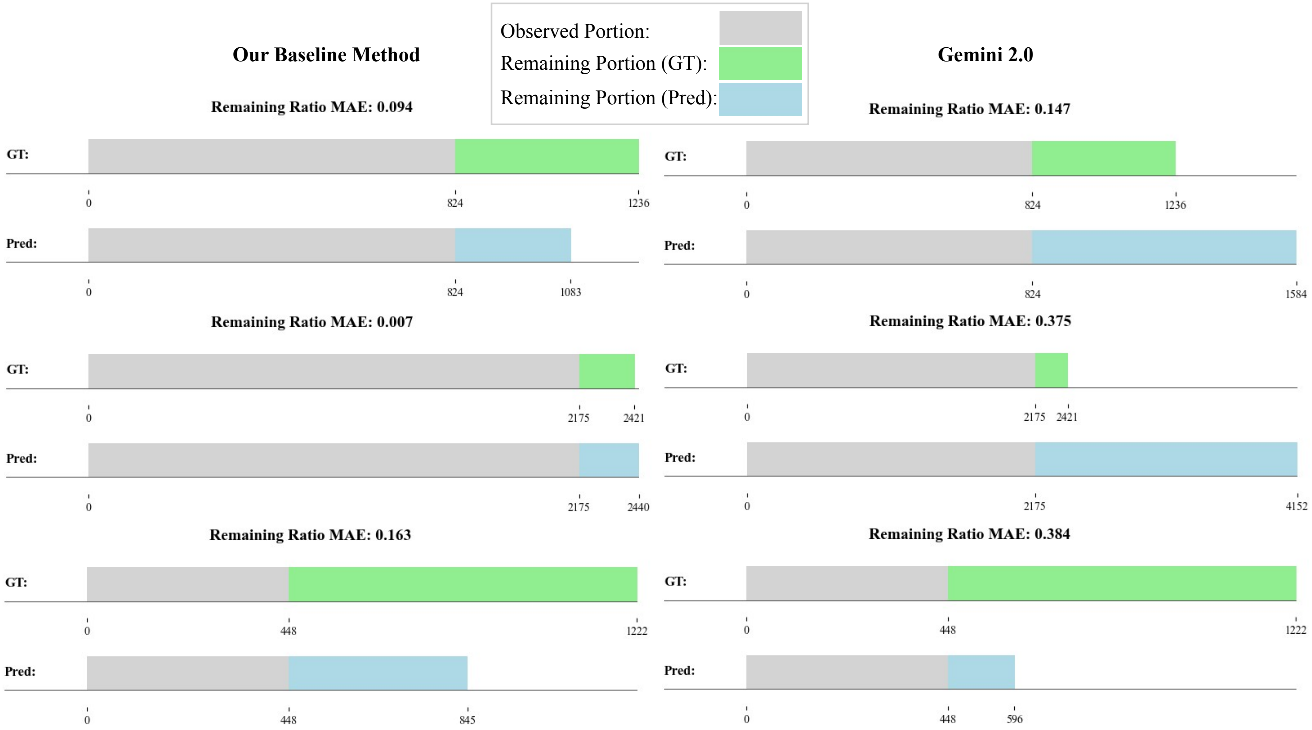


Figure 8. **Examples of our baseline method and Gemini on Task 2 (completion tracking)**.

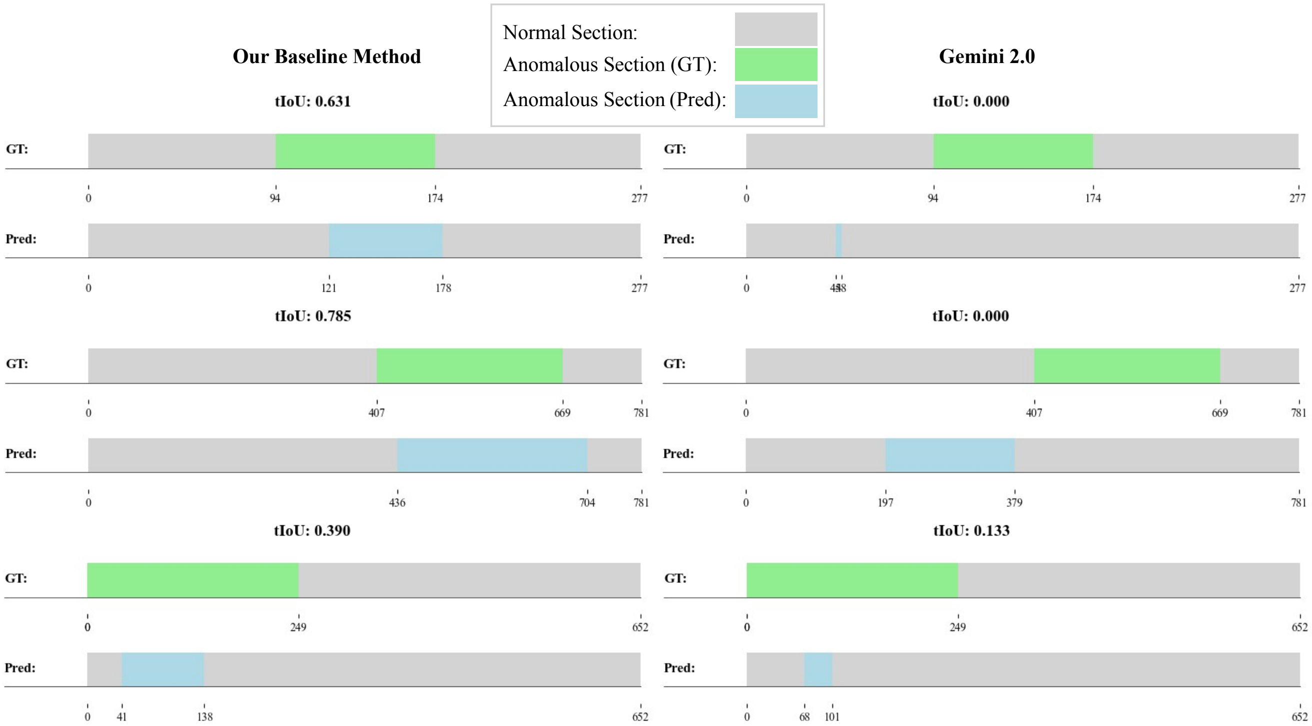


Figure 9. **Examples of our baseline method and Gemini on Task 3 (anomaly detection)**.

7. Visualization of Related Benchmarks and Our Benchmark

Figure 10 provides an overview of how existing periodic activity detection (PAD) benchmarks differ from our proposed benchmark. Most conventional benchmarks focus on short-term periodicity where the transitions between periods are highly contrasting (e.g., jumps in activity). In contrast, many real-world scenarios exhibit long-term periodicity. In these cases, each period is composed of a sequence of ordered activity tokens that together form complex workflows. Although prior work in human activity segmentation has successfully split activities into multiple tokens according to these workflows, such studies often assume that the period boundaries are pre-defined. Our benchmark addresses this gap by focusing not only on segmentation but also on the detection of the period boundaries in long-term activity sequences. This design allows for a more realistic evaluation of methods in capturing temporal dependencies and workflow structures. Note that the long-term period not only refers to an extended duration but, more importantly, to a greater volume of unique activity tokens that collectively form complex workflows.

References

[1] Alessandro Berti, Sebastiaan van Zelst, and Daniel Schuster. Pm4py: A process mining library for python. *Software Impacts*, 17: 100556, 2023. 2

[2] Philippe Fournier-Viger, Tin Truong Chi, Youxi Wu, Jun-Feng Qu, Jerry Chun-Wei Lin, and Zhitian Li. Finding periodic patterns in multiple sequences. *Periodic Pattern Mining: Theory, Algorithms, and Applications*, pages 81–103, 2021. 2, 3

[3] Google. Gemini 2.0 flash thinking, 2025. Accessed: March 1, 2025. 6

[4] Huazhang Hu, Sixun Dong, Yiqun Zhao, Dongze Lian, Zhengxin Li, and Shenghua Gao. Transrac: Encoding multi-scale temporal correlation with transformers for repetitive action counting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19013–19022, 2022. 4, 5

[5] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 6

[6] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 6

[7] Wil Van Der Aalst. Process mining. *Communications of the ACM*, 55(8):76–83, 2012. 2, 3

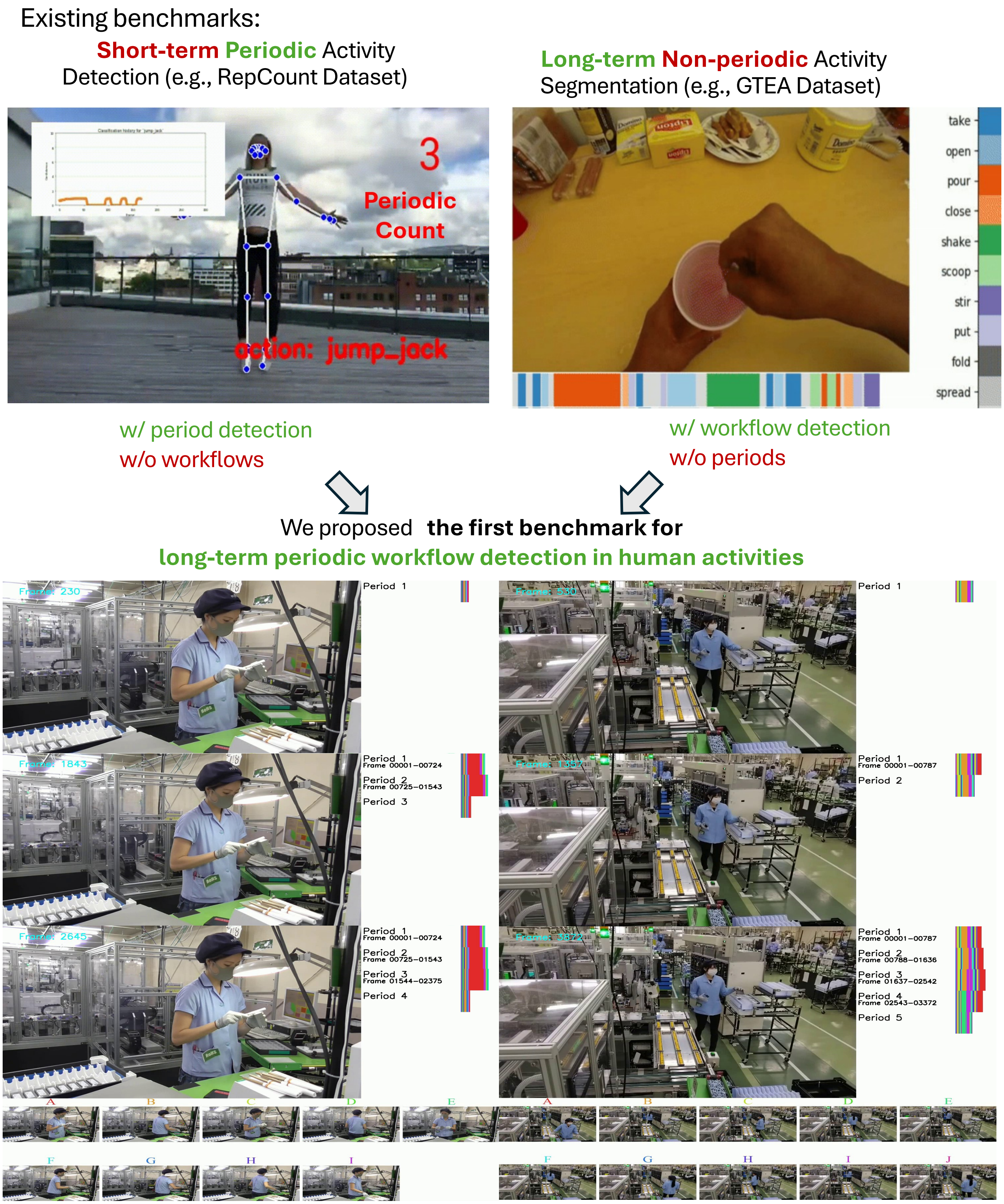


Figure 10. **Visualization of related benchmarks and our benchmark**. Current benchmarks for periodic activity detection generally focus on short-term periods with high-contrast patterns between periods (*e.g.*, continuous jumps). However, many real-world activity periods are long-term, with each period consisting of ordered activity tokens that form workflows. Additionally, existing studies on human activity segmentation can divide activities into multiple tokens according to a workflow. However, these studies typically overlook the aspect of period detection, often assuming that a single period has already been provided for segmentation purposes.