Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.0429000

Multimodal Personalized Assistive System with Natural Language Explanations for Cognitive Support and Work-Life Balance in Academic Settings

SHOIB AHMED SHOURAV[1], NAHID HOSSAIN[1], SWAKKHAR SHATABDA[2]
[1]Department of Computer Science and Engineering, United International University, United City, Madani Ave, Dhaka 1212, Bangladesh
Email: shoib@cse.uiu.ac.bd, nahid@cse.uiu.ac.bd
[2]Department of Computer Science and Engineering, BRAC University, Merul Badda, Dhaka 1212, Bangladesh
Email: swakkhar.shatabda@bracu.ac.bd

Corresponding author: Shoib Ahmed Shourav (e-mail: shoib@cse.uiu.ac.bd).

ABSTRACT Sustaining faculty wellbeing and cognitive effectiveness during routine office work is a growing concern in academic institutions, yet Human Activity Recognition (HAR) has largely overlooked this setting in favor of student behavior, construction-site safety, and generic productivity. We present TeacherActivityNet, a lightweight, real-time HAR framework that recognizes faculty office activities and translates them into personalized, evidence-grounded wellness guidance through a natural language generation layer. We make three contributions. First, we introduce PETRA, the first faculty-centric video benchmark, with nine annotated office activity classes from 19 participants across diverse rooms and viewing conditions. Second, we develop a fine-tuned YOLOv8n detector with a final-layer residual modification for indoor, low-compute deployment; it attains 0.841 mAP@50 on PETRA, surpassing YOLOv11 (0.743) and YOLOv12 (0.715) at 1.9 ms per image. Third, central to interpretability, we design an on-device Natural Language Explanation Module built on Phi-4-mini-instruct that converts per-session statistics into human-readable wellness summaries and cognitive-ergonomics recommendations. Coupling structured role-separated prompting with a post-hoc numerical audit yields no inconsistent claims in 96.7% of reports. Against a rule-based baseline and Llama-3.2-3B-Instruct, the module achieves the strongest automatic scores (ROUGE-L 0.573, BERTScore-F1 0.891) and the highest blind expert ratings, averaging 4.4 or above (out of 5) across factual consistency, recommendation relevance, and professional tone. Together, these establish a non-intrusive, end-to-end framework linking visual recognition to interpretable language feedback for faculty work-life balance.

INDEX TERMS TeacherActivityNet, PETRA, Natural Language Generation, Multimodal Assistive Systems, Cognitive Support, Activity Recognition, Work-life Balance

I. INTRODUCTION

Sustaining academic productivity and faculty wellbeing depends on a healthy work-life balance and the preservation of cognitive effectiveness. Yet despite growing awareness of occupational health challenges in academic settings, systematic tools capable of intelligently monitoring and supporting faculty during routine office duties remain scarce. Our work builds on Human Activity Recognition (HAR), which offers a principled framework for understanding and optimizing workplace behavior. Crucially, we move beyond activity classification alone: we couple recognition outputs with a natural language explanation layer that translates quantitative session statistics into personalized, actionable wellness guidance. The central problem we address is how to help teachers analyze and refine their office-hour routines so as to foster balanced, cognitively sustainable work practices, and, in turn, strengthen educational quality and student-teacher interaction. The *TeacherActivityNet* system introduced in this paper is designed to meet this need in educational institutions while remaining applicable to broader knowledge-work contexts.

HAR has attracted considerable attention across sports,

healthcare, agriculture, fitness, and surveillance [1]–[18], and educational organizations are increasingly adopting AI-based automation to streamline operations and improve efficiency [19]. Dimitriadou et al. [20] critically examined the use of AI and emerging technologies in the classroom, arguing that computer vision-based surveillance can safeguard classroom safety while tracking student participation and attendance. Related work has applied smart surveillance to plagiarism detection and student supervision in online settings [21], and to detecting abnormal activity on school premises from CCTV footage [22]. While computer vision has thus opened new avenues for task automation in education, analyzing and interpreting teacher behavior in office settings, with the goal of improving office-hour efficacy and faculty wellbeing, remains a largely unaddressed and technically challenging problem. Conventional self-management and time-tracking practices are labor-intensive, error-prone, and offer no path to automated, personalized feedback. A system that closes this loop, from automated video-based recognition through to interpretable, evidence-grounded recommendations, would mark a meaningful advance in faculty support.

Existing HAR solutions rely predominantly on machine learning and deep learning architectures, including CNNs [23], [24], RNNs [25], [26], LSTMs [27], [28], SVMs [29], Naive Bayes [30], and earlier You Only Look Once (YOLO) variants [31]. Each offers distinct trade-offs among accuracy, computational cost, and scalability, but none has been tailored to the faculty office domain. Intelligent surveillance in educational organizations remains confined to classroom and student activity tracking, while existing workplace activity-analysis systems [32]–[39] are largely restricted to construction-site worker safety, and generic employee time-usage systems [40], [41] stay performance-oriented without addressing the needs of academic faculty. Taken together, prior HAR and surveillance approaches suffer from four compounding limitations: (i) domain mismatch stemming from reliance on generic public datasets; (ii) the absence of faculty-centric office activity annotations; (iii) computationally heavy architectures ill-suited to real-time indoor deployment; and (iv) the lack of any mechanism to translate recognition outputs into interpretable, actionable guidance for end users.

To address these limitations, this paper makes the following contributions:

- A new dataset (*PETRA*). We introduce the Personalized Teacher Activity Recognition Dataset, comprising nine activity classes of teachers in office settings. This resource was not previously available in the academic HAR literature.
- A lightweight detector (TeacherActivityNet). We develop a modified and fine-tuned YOLOv8n model optimized for indoor office activity recognition, which outperforms YOLOv11 and YOLOv12 across most evaluation scenarios.
- A natural language explanation module. We introduce an on-device LLM component that synthesizes per-session activity statistics into concise, evidence-grounded wellness summaries and targeted cognitive-ergonomics recommendations.
- A modality analysis. We systematically compare image-based and video-based inference, offering actionable insights into their respective strengths for workplace activity recognition.
- A comprehensive evaluation. We benchmark the full TeacherActivityNet pipeline against multiple recent YOLO variants and established activity recognition models in academic office settings.

The remainder of this paper is organized as follows: Section II presents related work; Section III covers dataset creation, ethics, collection, and annotation; Section IV covers data augmentation, data preparation, and the proposed methodology, including the Natural Language Explanation Module; Section V describes the experimental setup; Section VI analyzes the results and presents ablation studies alongside a qualitative evaluation of the generated wellness reports; and Section VII concludes the paper.

II. RELATED WORKS

HAR is a critical research area with applications across security, surveillance, sports, education, and mental wellbeing. Existing approaches can broadly be categorized along two orthogonal axes: the modality used (image vs. video) and the learning paradigm employed (machine learning vs. deep learning). We survey representative works along these dimensions and discuss their implications for our task.

A. MACHINE LEARNING APPROACHES FOR HAR

Classical machine learning (ML) methods offer interpretability and lower computational overhead [42], though they may be limited in modeling complex spatial-temporal patterns. A wireless method named the Zigbee technique was proposed to track employee activity using computer vision, where CCTV-captured workplace videos were compared against biometric entry logs to evaluate system efficiency [41]; however, the system was designed to function only within a specific region. Sikder et al. [43] introduced the KU-HAR dataset comprising 90 participants performing 18 distinct actions and demonstrated that a Random Forest classifier could achieve a precision of 90%. Krishan Kumar et al. [44] proposed a two-viewpoint based real-time hand gesture recognition method, demonstrating that using multiple perspectives significantly enhances recognition performance by capturing more robust spatial features. Their work highlights the importance of multi-view representation in improving real-time vision-based activity recognition systems. While such approaches remain competitive on constrained settings, their generalization to complex, real-world temporal dynamics is limited.

B. IMAGE-BASED DEEP LEARNING FOR HAR

Image-based deep learning (DL) methods offer strong automatic feature extraction and can encode temporal

information, albeit at the cost of more complex modeling or larger labeled datasets [45]. Rashmi et al. [46] collected 688 image frames and over 54,000 samples from CCTV cameras in computer laboratories, employing YOLOv3 for real-time student action detection. Kumar [47] leveraged Multi-task Cascaded Convolutional Networks (MTCNN) to assess student engagement via facial expression recognition, underscoring the effectiveness of DL for fine-grained activity understanding.

In the fitness domain, Yang et al. [48] compared Pose-Based Branch (PBB) and RGB-Based Branch (RBB) features using CNN, ResNet152, and 3D-CNN across several benchmarks, concluding that pose-based representations consistently outperform RGB-based counterparts. This finding motivates the use of skeleton or pose features in appearance-invariant HAR, which offer long-term robustness but depend critically on reliable pose estimation.

C. VIDEO-BASED DEEP LEARNING FOR HAR

Video-based methods benefit from strong inter-frame correspondence, making labeling more tractable and enabling the capture of temporal dynamics [45], [49]. Singh et al. [50] applied CNN and RNN with Inception V3 on the UCF crime dataset (1,800 videos) for crime scene prediction, highlighting the importance of large datasets and data augmentation. Shreyas et al. [51] proposed a 3D CNN for anomalous activity detection on the same corpus, outperforming SVM and binary classifiers. Latha et al. [52] proposed CNN and LSTM models evaluated on the UCF-50 video dataset for general activity recognition.

More recent transformer-based architectures have pushed the state of the art further. Liu et al. [22] proposed Temporal Segment Transformers and Video Swin Transformer models to recognize abnormal behavior on the CABR50 dataset, demonstrating the scalability of attention mechanisms to long-range temporal reasoning. Nale et al. [53] combined pose estimation with LSTM on the NTU-D 60 dataset to detect suspicious activity by evaluating the geometrical relations of skeletal joints, bridging video-based and skeleton-based paradigms.

D. DOMAIN-SPECIFIC APPLICATIONS

1) Sports Activity Recognition

Host et al. [54] provided a comprehensive overview of ML and DL techniques applied to ball-based sports using video datasets. Xiao et al. [55] introduced deformable convolution and adaptive multi-scale feature methods to analyze sports videos from UCF Sports, UCF-50, and YouTube (UCF 11) benchmarks. Goh et al. [56] focused on fault detection in badminton using YOLOv5 trained on 1,900 images, achieving accuracy surpassing human judges.

2) Security and Surveillance

Security-oriented HAR has attracted sustained attention. Selvi et al. [57] proposed an Enhanced CNN for suspicious action prediction evaluated across multiple datasets including CCTV footage, CAVIAR, and DCSASS. Menaka et al. [58] applied YOLOv3 for real-time suspicious activity detection in ATM surveillance videos. Nandhini et al. [59] employed a stacked hourglass method with a Gaussian classifier for crime scene detection, achieving over 90% accuracy. Rahman et al. [60] developed a system to predict possible anomalies in smart home environments. Ankalaki et al. [61] introduced the OP-Tanish activation function within a 1D-CNN, outperforming traditional activations such as ReLU and SWISH for anomaly detection.

E. WORK-LIFE BALANCE AND TEACHER WELLBEING

Compared to the breadth of activity recognition research, studies addressing work-life balance and mental wellbeing through a computational lens remain scarce. Mulyani et al. [62] investigated children's behavior and teachers' work-life balance within school climates. Abdulaziz et al. [63] collected data from 278 teachers and demonstrated that work-life balance positively affects organizational commitment, while work overload exerts a negative influence. These findings motivate the development of automated tools capable of monitoring and analyzing teacher activity in office environments.

F. COMPARATIVE SUMMARY AND RESEARCH GAP

Figure 1 provides a structured overview of the surveyed literature, organized along the image/video and ML/DL axes. In general, ML approaches are less data-hungry and more interpretable but struggle with complex spatial-temporal patterns [42]. DL methods yield superior performance on large, complex datasets [49] but demand substantial labeled data. Videos facilitate labeling through inter-frame correspondence yet lack explicit temporal metadata, whereas image-based methods can provide temporal cues but require more sophisticated modeling or data [45]. Pose and skeleton-based features are appearance-insensitive and exhibit long-term stability, but depend on reliable pose estimation.

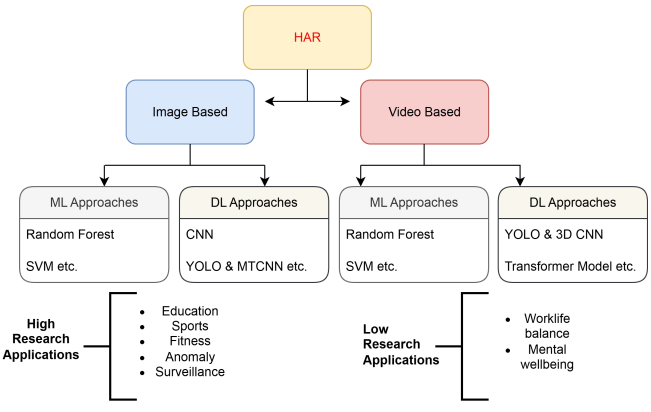


FIGURE 1: Overview of the related work. Image-based and video-based HAR applications with respect to ML and DL approaches.

Despite the extensive body of HAR literature, no publicly

available personalized dataset exists for analyzing the activities of faculty or teachers in office settings. Existing datasets focus on general actions, sports, or security scenarios, leaving workplace cognitive overload, time mismanagement, and mental wellbeing monitoring largely unaddressed. We aim to bridge this gap by constructing a dedicated dataset for teacher activity recognition in office environments.

III. THE *PETRA* DATASET

A key contribution of the work is the creation of *PETRA* (**P**ersonalized **Te**acher Ac**t**ivity **R**ecognition D**a**taset), a novel video dataset specifically designed for teacher activity recognition in office environments, a resource that has not previously existed in the academic HAR literature. *PETRA* includes a wide range of activities captured across realistic office settings, comprising videos of 19 participants performing 9 action classes. The action classes are ''Arriving'', ''Counselling'', ''Eating'', ''Idle'', ''Leaving'', ''Sleeping'', ''Talking'', ''Using_Phone'', and ''Working''.

To ensure diversity and ecological validity, the *PETRA* videos were recorded across 8 different rooms using a smartphone (iPhone 11 Pro Max) at High Definition (HD) quality with 60 Frames Per Second (FPS). Recording conditions were varied with respect to camera distance (adapted to room size), capture angle, and the number of participants visible in frame. The cameraman's average height was 5'6''. *PETRA* includes 19 participants (3 female), mostly aged 20–22 with some participants around 30. All actions were controlled and guided to ensure label accuracy.

PETRA contains a total of 147 videos, partitioned into 110 for training, 27 for validation, and 10 for testing, yielding an approximate train, validation, and test split ratio of 75% : 18% : 7%, respectively. We deliberately allocate a larger share to the validation set (18%) compared to conventional splits, as robust hyperparameter tuning and temporal threshold optimization are critical for action recognition tasks to prevent overfitting to specific action sequences. This distribution ensures sufficient diversity in the validation set to evaluate model performance across varied office activities and employee behaviors, while maintaining adequate test data for final evaluation. The larger validation split is particularly beneficial given the class imbalance and temporal variability inherent in office monitoring scenarios. The dataset used in this study has been publicly released and is available at: https://tinyurl.com/42hvr8n3.

A. ETHICAL CONSIDERATIONS AND PRIVACY PROTOCOLS

The participants were required to provide informed consent prior to making videos, and since we intend to make the data public for research purposes, there will be no opt-out policy. Below are some key points outlining how privacy will be maintained.

1) Written Informed Consent

Written informed consent was obtained from all participants during the data collection process. Participants were clearly informed about the intended use of the data for research purposes, the manner in which it would be shared, the parties who would have access to it, and the duration of its retention.

2) No Opt-out Policy

As we intend to publish the data for all independent researchers to use, we believe that participants are not allowed to opt-out.

3) Privacy Protection In Model Deployment

There are no privacy concerns associated with the faculty activity analysis system and its deployment. Teachers can deploy our proposed model on real-time CCTV footage (as shown in Section IV-D) to determine the total duration of each action class. There is no need to store the CCTV recordings, and this approach helps avoid privacy issues in office rooms. For example, if a faculty member in their office performs only the 9 action classes listed in Table 1, they will be notified only of the total time spent on those nine actions. If a faculty member engages in an activity that is not included in the defined action classes, our model will classify such actions as part of the background class. As a result, the specific action performed during that time cannot be identified.

Therefore, if a faculty member is praying or engaging in any private activity, it will be categorized under the background class, leaving the exact action unidentifiable. In this way, we can maintain their privacy while still allowing the model to operate effectively in office settings. Since we are only focused on the nine defined action classes, we do not track or analyze any other activities the faculty member may be performing in their room. This ensures that the system remains strictly limited to predefined behaviors without violating personal privacy.

4) Local, Single-User Deployment and Privacy Scope

The proposed system is designed solely as a personalized assistive framework to support teachers' cognitive well-being and work-life balance, not as a tool for centralized deployment or institutional monitoring. It operates locally on educator-owned devices (e.g., CCTV, webcams, or mobile devices) for lightweight, short-term behavioral assessments within privately controlled environments, where the user is the sole data subject, controller, and operator. This local, single-user model clearly differentiates the framework from centralized surveillance or institutional data collection systems. Because all processing and feedback remain entirely on the user's personal device with zero external data transmission, the regulatory and adversarial concerns associated with centralized or multi-user deployments fall outside the scope of this work. These include formal consent and withdrawal procedures, remote encrypted transmission, server-side storage policies, and privacy-preserving distributed learning, as well as

threats such as transmission interception, caching, screen capture, model inversion, and access control. For the same reason, additional privacy-preserving mechanisms such as anonymous identifiers or face masking are unnecessary, since the data never leaves the user's direct control and is used solely for personal benefit. The framework therefore prioritizes individual autonomy and local data ownership by design.

B. COLLECTION

The dataset creation process was carried out in two primary stages: first, we collected training and validation videos, followed by test videos. In the first stage, we selected a sample of 12 participants out of 19, each asked to perform one action at a time. Each participant recorded at least nine videos, covering all nine action classes. After collecting the training videos, we selected 3 participants from the remaining 7 and asked them to perform the same actions. Recordings of these three participants across the nine classes were used as the validation set.

In the second stage, we recorded 10 videos of the remaining 4 participants who were not involved in the previous stage of data collection. Since we are interested in measuring how accurate our models are in real-time, unlike the previous stage where each action was performed in separate videos, we asked the participants to perform the actions continuously for a specific period. To rigorously test model performance in real-time settings, most of the videos were recorded in rooms that were not used during the earlier phase of data collection. Table 1 provides descriptions of the action classes, including their duration and the number of instances after annotation.

1) Challenges of Data Collection

We faced many challenges during the data collection process. We have divided these challenges into two sections, as outlined below.

2) Background Problem

We initially hypothesized that the actions would be independent of the background in the videos, as ideally expected. Operating under this assumption, we recorded videos of approximately 10 participants within a single room at the same teacher's desk. Subsequently, we initiated the annotation and augmentation processes to prepare the dataset for model training. Following model training, we evaluated its performance on a new dataset recorded in different rooms. However, the model failed to accurately classify the actions, primarily due to the uniformity of backgrounds in the training set compared to the diverse backgrounds in the test set. This outcome highlighted the necessity of incorporating greater background variability within the training dataset. To address this challenge, we undertook the reconstruction of the dataset, ensuring enhanced background diversity to improve model generalization.

3) Actions Dependent on Objects

A key challenge in constructing *PETRA* was preventing action classes from becoming dependent on the presence or absence of specific objects, a form of dataset bias that can cause models to rely on contextual shortcuts rather than motion-based cues. Two such cases were identified and systematically addressed during data collection.

The first case involved mobile phones. Since two action classes, ''Using_Phone'' and ''Talking'', naturally involve mobile phone usage, an unconstrained recording protocol would result in the phone being absent across all remaining classes, inadvertently making its presence a discriminative cue. To mitigate this, participants were instructed to keep their phones visible on the desk regardless of the action being performed. This ensures that the model cannot exploit object presence as a proxy for action class membership.

The second case arose during the recording of the ''Eating'' action class, where food items, water bottles, and tiffin boxes were initially present only in that class and absent in others. Following the same object-invariance principle, these items were intentionally included in the scenes of other action classes as well. As illustrated in Figure 31 (Appendix A), a participant performing the ''Idle'' action is depicted in a scene that includes a tiffin box, a water bottle, a laptop, and a mobile phone, demonstrating this deliberate effort to decouple object presence from action identity. Additionally, to reflect realistic office interactions, the ''Counselling'' action class was recorded with multiple students dynamically joining and leaving the session.

C. ANNOTATION

After compiling the dataset, we generate the image frames from the training and validation videos using Roboflow[1] through manual annotation. Three annotators performed manual labeling, with a verifier ensuring the accuracy of all annotations. From each frame, a bounding box is drawn to annotate the object with an action class. Extra caution is exercised while annotating the images that look the same but fall into different action classes. For example, in ''Leaving'' and ''Arriving'' action classes, there is a moment in the videos when the participants leaving the seat and taking the seat, look almost identical. Another case is while performing ''Idle'' action, the participants closed their eyes which potentially conflicts with ''Sleeping'' action. Taking these ambiguous scenarios into consideration, such frames are discarded from the training dataset. After annotating, the numbers of instances in the training, validation, and test sets are 6498, 1337, and 808, respectively. In Figure 2 we have shown the annotated images for training.

1) Challenges of Annotation

We employed three annotators and one reviewer for the dataset labeling process. Regular meetings were held to discuss the details of the action classes, ensuring consistent

[1]https://roboflow.com

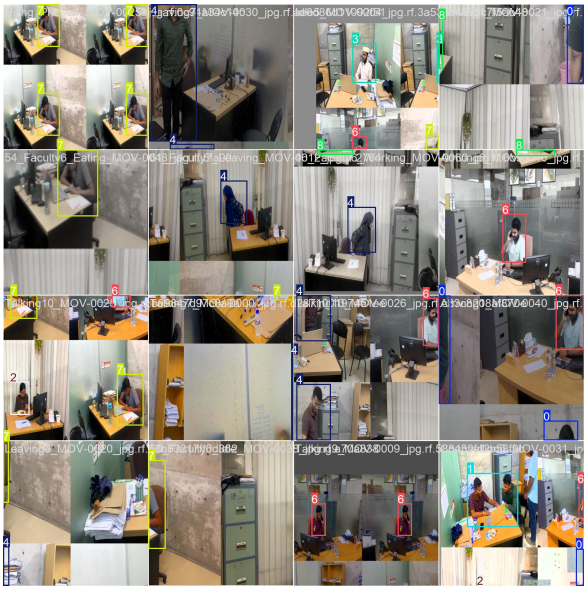


FIGURE 2: Annotated images from the training dataset. The figure illustrates how bounding boxes and class labels are applied to different actions for supervised learning.

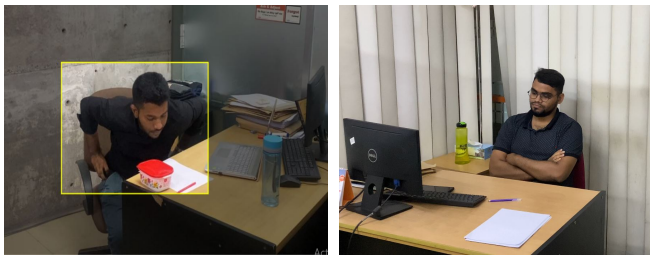
TABLE 1: **Action class definition with duration and total instances in training dataset**

Class Name	Class ID	Definition of the Action Class	Duration Per Video	Total Instances
Arriving	0	Person arriving in the room and taking his/her seat.	5 seconds	665
Counselling	1	Teacher giving counselling time to multiple students.	30/60 seconds	728
Eating	2	Person eating or drinking.	10 seconds	739
Idle	3	Idle for some time, doing none of the other mentioned actions.	30/60 seconds	755
Leaving	4	Person leaving his/her seat and going out of the room.	5 seconds	654
Sleeping	5	Person sleeping in two positions: laying his/her head on the chair or putting their head down on the table.	10 seconds	739
Talking	6	Talking via mobile phone (not considering talking to persons).	30/60 seconds	727
Using_Phone	7	Using his/her phone while sitting in the chair or standing.	30/60 seconds	748
Working	8	Focused on the computer/laptop while using mouse or keyboard.	30/60 seconds	743

The duration per video is approximate and may vary slightly depending on activity instances.

interpretation and annotation across all annotators. After completing the initial annotation, we evaluated our model's performance across the nine action classes and observed that certain classes performed poorly due to inconsistent labeling.

In the "Leaving" and "Arriving" action classes, participants often assumed similar postures while either taking a seat or preparing to leave the room, which caused confusion for the model. Initially, such ambiguous frames were discarded. As shown in Figure 3a, it is challenging to determine whether the participant is leaving or arriving. Subsequently, we decided to retain these frames, anticipating that the issue could be addressed by incorporating spatio-temporal focus during model training.



(a) *Leaving* and *Arriving* (b) *Sleeping* and *Idle*.

FIGURE 3: Ambiguity between the actions classes. The figures highlight how the model confuses similar postures, emphasizing the need for finer feature discrimination

Other action classes also exhibited overlap. For instance, during the "Idle" action, some participants closed their eyes, making it resemble the "Sleeping" action class. Such frames negatively affected model performance for the "Idle" class and were therefore discarded. In Figure 3b, the participant is looking down while performing "Idle", but the closed-eye appearance can confuse the model regarding the correct action class.

IV. METHODOLOGY

Our work consists of three major phases - creating the dataset which already has been explained in Section III, modifying the YOLOv8 model architecture to obtain a better performing detection model, and proposing a method for real-time prediction from videos. Figure 4 presents the architectural pipeline of the proposed TeacherActivityNet. Initially, we curate our dataset followed by preprocessing it. After that, we make modifications to the YOLOv8n model architecture to introduce the TeacherActivityNet. We train and fine tune the TeacherActivityNet model to predict the action classes from the videos of the faculty members. Besides, we measure action class-wise time for each faculty member through face recognition.

A. DATASET AUGMENTATION AND PREPARATION

To be compatible with YOLOv8, the images are reshaped to 640×640 pixels. Two different augmentations are also applied to increase the generalizability of the prediction models. Initially, the images are flipped horizontally. And to give the models the ability to predict from CCTV footage, a noise of 0.3% is added. Both augmentation techniques are employed in accordance with the recommendations made by [50]. In Figure 5a and Figure 5b, a sample before augmentation and another one after augmentation are presented. Only the training images are augmented and increased from 6498 to 19494 instances.

After annotation, each image has two files saved in two directories - one is an image and another is the label file that contains bounding box coordinates. We use the YOLOv8-oriented bounding box dataset. In this dataset, the label file has 9 values instead of 5 which is common in rectangular

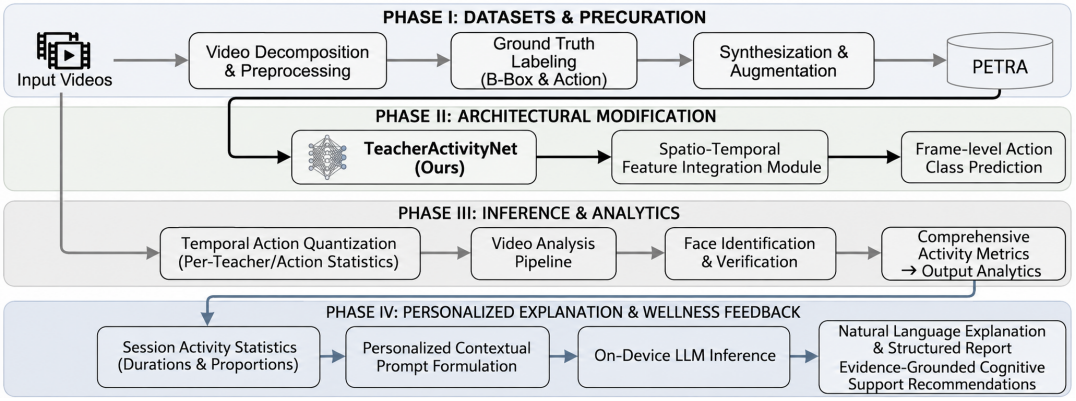


FIGURE 4: End-to-End Architectural Pipeline of TeacherActivityNet

(a) Before augmentation. (b) After augmentation.

FIGURE 5: Training images comparison: (a) shows the original dataset samples, while (b) demonstrates enrichment through horizontal flipping and noise addition for better generalization.

bounding box coordinates. The label of our images has quadrilateral coordinates. The first value is the class ID and the remaining values are coordinates of four corners like $(x_1, y_1), (x_2, y_2), (x_3, y_3)$, and (x_4, y_4).

B. TEACHERACTIVITYNET

YOLOv8 is considered state-of-the-art for object detection in real-time. This version of YOLOv8 has more speed and accuracy than the previous versions. A few key components of YOLOv8 are anchor-free detection, multi-scale predictions, and decoupled head architecture. We adopt the YOLOv8n as the base model for our detection task. In YOLOv8n there are three main parts in the architecture - Backbone, Neck, and Head.

In our proposed model, we modify the backbone of the YOLOv8n architecture. In the backbone, there are five feature pyramids (P) and four stages with each having a c2f (Cross-Stage Partial Network with 2 Convolutions and Fusion) module and a convolution module. We modify the last layer of the convolution module with a residual connection. The input of the convolution module of the last layer adds to the output of the same layer. This process is done for every convolution module. In Figure 6, the internal mechanism of

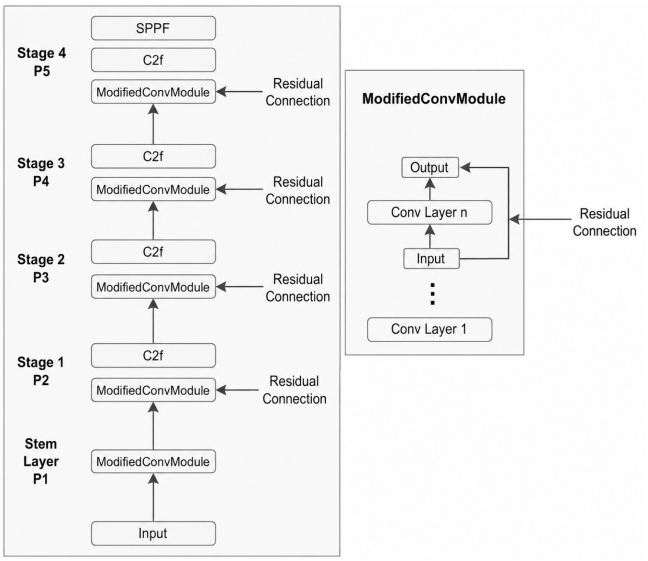


FIGURE 6: Modified backbone of our proposed *TeacherActivityNet*. The figure highlights the key layers and structural changes made to improve feature extraction for activity recognition.

the modified convolution module with residual connection is presented. It adds a residual connection when the input and output channels of the convolution layer are the same with a stride of (1,1). This can help mitigate the vanishing gradient problem and potentially improve the learning of deeper features. Residual connections offer a direct pathway for gradients to propagate backward, helping to alleviate the vanishing gradient problem. Mathematically, considering the gradient of the loss L with respect to the input x, we have in Equation 1:

$$\frac{\partial L}{\partial x} = \frac{\partial L}{\partial y} \cdot \left(\frac{\partial F(x)}{\partial x} + 1 \right) \tag{1}$$

Here, where F(x) denotes the residual mapping learned by the convolution block. The addition of the $+1$ term ensures that the gradients can flow back directly, even when $\frac{\partial F(x)}{\partial x}$ becomes very small. The added residual connections might introduce a small computational overhead, but it is likely to

be minimal given that they are only added under specific conditions. For benchmark analysis, in addition to creating detection models with our proposed TeacherActivityNet, we use pre-trained and fine-tuned YOLOv8 and FasterRCNN Resnet50 FPN model. All the models are trained on annotated images and labels with bounding box coordinates.

1) Rationale for Final-Layer Residual Connection

The residual connection is placed at the final convolutional layer because this layer is closest to the loss function during backpropagation, and the connection helps the gradients flow back more easily, reducing the risk of vanishing gradients. The last layer also contains the most abstract and discriminative features needed for recognizing complex activities. Adding a residual connection here strengthens these features and improves training stability and convergence, without forcing the model to relearn earlier representations. In contrast, adding residual connections in many earlier layers would increase complexity, computation cost, and might disturb useful pretrained features in lower layers. Therefore, using a residual connection only in the final layer provides a good balance between stable learning, strong feature refinement, and minimal interference with the existing YOLO architecture.

C. SPATIO-TEMPORAL FEATURE

During the initial evaluation, the model exhibited poor performance for action classes such as "Arriving", "Leaving", and "Eating". To address this limitation, we incorporated spatio-temporal features during the training phase to improve the mAP50 for these classes. This approach captures both spatial and temporal information by leveraging multiple consecutive frames to identify action patterns more effectively. Figure 7 illustrates the framework for temporal inference and quantitative validation. For the "Eating" action class, we used the pose-based YOLOv8 model to detect the wrist and head positions with movement while focusing on multiple image frames. The YOLO model does not predict the position of the mouth or hand, but it can detect the wrist and head. By using Spatio-Temporal features, we focused on more than one frame and checked that the wrist and head distance changed over time. This helped us determine whether the user was actually consuming food or not. Although it didn't improve much in predicting the "Eating" action class, the Spatio-Temporal feature, where we focused on multiple frames to detect the action class, helped increase the mAP50 for the "Arriving" and "Leaving" action classes. In Figure 8a and in Figure 8b, we can see that YOLOv8 can predict wrist and head positions in image frames using pose-based YOLOv8 models. At the same time, we tried to detect the food item in the same image frame but failed in this instance. Despite testing multiple wrist-to-head distance thresholds and frame-window sizes, Eating detection did not improve appreciably. Drinking motions produced wrist movement with little head displacement, and the base YOLOv8 model frequently failed to localize the food item itself (Figure 8a). This combination of weak motion cues and missed object detection remains the primary cause of the low "Eating" mAP@50.

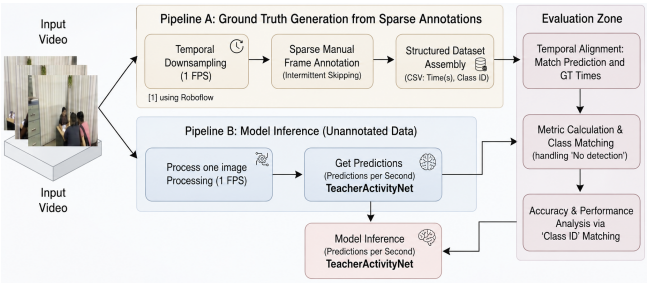


FIGURE 7: Temporal Inference and Quantitative Validation Framework.

1) Action Class Wise Advantages

For the "Eating" action class, although the mAP50 increased slightly, for the action classes "Arriving" and "Leaving", we gained a significant advantage. There were some image frames where the model was really confused between these two action classes, as shown in Figure 3a.

D. PREDICTION FROM VIDEOS

A major challenge for any computer vision model is evaluating performance in real-world scenarios. In addition to reporting validation and test accuracy of the proposed TeacherActivityNet model, we also measure its performance directly from videos. First, we detect the face of each participant in the video so that their activity duration can be counted individually. The face recognition module uses dlib's deep neural network, a 29-layer convolutional ResNet, which takes $150 \times 150 \times 3$ face images and produces 128-dimensional feature embeddings, clustering images of the same person together while separating different identities. We then take the input video and extract one frame per second. After extraction, we annotate frames in which participants clearly perform one of the action classes listed in Table 1 and skip transition frames where no defined action is being performed (for example, when a participant is switching between "Talking" and "Eating").

Once annotation is completed, we prepare a CSV file with two columns: "Time in seconds" and "Class ID," which serves as ground truth. The original test video is then processed by TeacherActivityNet at one frame per second to match the annotation rate. For each second, the predicted class is aligned with the corresponding timestamp and compared to the annotated class label. TeacherActivityNet may also output background (No detection). We finally compute how many seconds were predicted correctly, providing the real-world detection performance in terms of how many annotated frames were accurately classified.

E. NATURAL LANGUAGE EXPLANATION MODULE

To bridge the gap between quantitative activity predictions and actionable faculty insights, we design a lightweight

(a) Raw input frame. (b) Keypoint prediction.

FIGURE 8: Keypoint detection for the *Eating* action class: (a) shows the original frame, and (b) illustrates the predicted head and wrist positions, demonstrating the model's localization accuracy.

post-processing module that synthesizes per-class duration estimates into personalized, human-readable wellness summaries. Rather than exposing raw statistics, this layer produces structured natural language reports grounded in cognitive ergonomics guidelines [64], making system outputs interpretable by non-technical end users.

1) Model Selection and On-Device Deployment

We adopt Microsoft's `Phi-4-mini-instruct` [65], a 3.8B-parameter dense decoder-only transformer optimized for instruction-following and structured generation. We select Phi-4-mini over comparably-sized open alternatives (e.g., Llama-3.2-3B [66], Qwen2.5-3B [67]) based on its superior instruction-following performance at sub-4B scale [65] and its native support for structured role-separated prompting via a well-defined chat template. The model runs entirely on-device, achieving inference latency of ≈0.3–0.5 s per report on a standard consumer GPU, a deliberate design choice that preserves institutional data privacy and eliminates dependency on external API services. The model requires no task-specific fine-tuning and is loaded via the Hugging Face `transformers` library (v≥4.49.0) with automatic dtype and device assignment. Decoding is governed by the hyperparameters summarized in Table 2, balancing linguistic naturalness with output stability.

TABLE 2: Inference hyperparameters for the Natural Language Explanation Module.

Parameter	Value
`max_new_tokens`	120–150
`temperature`	0.7
`top_p`	0.9
Termination token	`<|end|>` / EOS
Quantization	Auto (device-native)

2) Prompt Engineering and Input Formulation

We use a structured system–user chat template consistent with Phi-4-mini-instruct's instruction format. The *system* prompt instructs the model to act as a professional wellness assistant, produce a neutral time-allocation overview across the nine activity classes, and append a single evidence-grounded recommendation (e.g., micro-break scheduling when low-engagement fractions exceed a configurable threshold, digital boundary-setting for disproportionate phone use). The *user* turn supplies per-class durations d_i (in minutes) and their session-relative proportions $p_i = 100 \cdot d_i \,/\, \sum_j d_j$ (%), where $i \in \{1, \ldots, 9\}$ indexes the nine activity classes (Table 1) and j ranges over the same nine classes in the summation. The full prompt template, following Phi-4-mini-instruct's role-separated chat format, is reproduced below.

System Prompt:

`<|system|>`

You are a professional academic wellness assistant specializing in faculty cognitive health and sustainable work-life balance. You will be given the durations (in minutes) and session-relative proportions of nine activity classes recorded during a faculty member's workday: Arriving, Counselling, Eating, Idle, Leaving, Sleeping, Talking, Using_Phone, and Working. Produce a concise, non-judgmental daily summary structured as follows: (1) a neutral overview of time allocation across the reported activity classes; (2) one targeted, practical recommendation grounded in evidence on cognitive ergonomics and desk-based knowledge work. If idle or sleeping activity collectively exceeds a low-engagement threshold, recommend structured micro-break scheduling. If phone use appears disproportionate relative to the session, recommend digital boundary strategies. Maintain a supportive, professional tone throughout.

`<|end|>`

User Turn Template:

`<|user|>`

Today's session activity profile (durations in minutes and approximate session-relative proportions): Working: d_1 min (p_1%), Counselling: d_2 min (p_2%), Idle: d_3 min (p_3%), Sleeping: d_4 min (p_4%), Using_Phone: d_5 min (p_5%), Talking: d_6 min (p_6%), Arriving: d_7 min (p_7%), Leaving: d_8 min (p_8%), Eating: d_9 min (p_9%).

`<|end|>`

`<|assistant|>`

Generation terminates at the model's `<|end|>` / EOS token upon completion of the structured report.

3) Factual Grounding and Hallucination Mitigation

A core reliability requirement for any wellness-facing language module is that generated text must remain *factually anchored* to the quantitative inputs that triggered it. We enforce this through three complementary mechanisms.

a: Structural Factual Anchoring.

Every user-turn prompt explicitly enumerates the nine per-class durations d_i and session-relative proportions p_i as the sole empirical basis for the generated report. The system prompt constrains the model to *describe* these provided statistics rather than infer or extrapolate unseen values, thereby eliminating the principal source of open-ended hallucination in free-form generation. Because the input statistics are deterministic outputs of the upstream classifier rather than free-text claims, the model has no factual ambiguity to resolve.

b: Constrained Decoding.

Generation is capped at `max_new_tokens` = 120–150 (Table 2), limiting the report to two functional segments (time-allocation

overview and a single recommendation) and preventing the model from elaborating beyond the evidence supplied. A temperature of 0.7 with `top_p` $= 0.9$ reduces stochastic variability while preserving linguistic naturalness. The structured role-separated chat template of Phi-4-mini-instruct further suppresses instruction drift across sessions.

c: Post-Hoc Numerical Consistency Audit.

To quantify residual factual errors, we implemented an automated auditor that (i) extracts all numeric tokens (durations in minutes and percentage figures) from each generated report using a regex-based parser, and (ii) verifies each extracted value against the corresponding input statistic, tolerating rounding to the nearest integer. Applied to all 30 expert-evaluated reports, the auditor found that 29 out of 30 reports (96.7%) contained no numerically inconsistent claims. The single flagged instance involved a rounded percentage that differed by one percentage point due to floating-point truncation in the prompt serialization step; this has since been corrected by rounding p_i to one decimal place before injection. These results corroborate the high expert factual-consistency score of 4.7 / 5 reported in Section IV-E6 and confirm that the structured prompting strategy effectively bounds hallucination within the wellness-reporting domain.

4) Baseline Systems for Comparative Evaluation

To ground our evaluation comparatively, we implement a *rule-based template system* (RB) as a strong non-learning baseline. RB iterates over the nine activity classes and fills a fixed sentence template for each class whose duration exceeds a one-minute threshold (e.g., *"You spent {d_i} minutes ({p_i}%) on {class}."*), followed by a hard-coded recommendation selected via a simple priority ladder: if Idle + Sleeping $> 20\%$ of session $\Rightarrow$ micro-break recommendation; else if Using_Phone $> 15\%$ $\Rightarrow$ digital boundary recommendation; else $\Rightarrow$ a generic *"maintain your current productive routine"* statement. RB requires no model inference and serves as both a latency lower bound and a content quality ceiling for purely template-driven approaches. We additionally compare against Llama-3.2-3B-Instruct [66], a competitive open-weight model at the same parameter scale, using an identical system and user prompt to ensure evaluation parity.

5) Automatic Metric-Based Evaluation

a: Reference Summary Construction.

Two occupational-health researchers independently authored reference wellness summaries for each of the 30 session profiles used in the expert rating study, following the same two-part structure (time-allocation overview and one recommendation) specified in the system prompt. Inter-rater agreement on reference content was $\kappa = 0.79$ (Cohen's κ on recommendation category), indicating substantial consensus. The final reference for each session was obtained by selecting the summary rated higher by the third expert (the academic administrator).

b: Evaluation Metrics.

We report ROUGE-1, ROUGE-2, and ROUGE-L [68] using unigram/bigram recall-oriented overlap against the reference summaries, and BERTScore-F1 [69] (`microsoft/deberta-xlarge-mnli` backbone) as a semantic similarity measure robust to paraphrase variation.

c: Quantitative Results.

Table 3 reports mean scores over the 30 sessions. Phi-4-mini outperforms both the rule-based baseline and Llama-3.2-3B across all four metrics. The pronounced ROUGE-2 gap between Phi-4-mini and RB (0.341 vs. 0.187) reflects the rule-based system's inability to form coherent multi-word phrases that align with expert-authored language, whereas the smaller gap with Llama-3.2-3B (0.341 vs. 0.276) highlights that instruction-following quality at sub-4B scale is non-trivially differentiated. BERTScore-F1 confirms that Phi-4-mini's outputs are semantically closest to human references (0.891 vs. 0.863 for Llama-3.2-3B and 0.821 for RB), validating the model-selection rationale stated in Section IV-E1.

TABLE 3: Quantitative NLP evaluation of the three systems against expert-authored reference summaries ($N = 30$ sessions). Higher is better for all metrics. †Statistically significant improvement over RB ($p < 0.05$, paired Wilcoxon signed-rank test).

System	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore-F1
Rule-Based (RB)	0.431	0.187	0.392	0.821
Llama-3.2-3B-Instruct	0.537†	0.276†	0.498†	0.863†
Phi-4-mini (Ours)	**0.614**†	**0.341**†	**0.573**†	**0.891**†

6) Human Expert Evaluation

We extend the original expert rating protocol to include the rule-based baseline and Llama-3.2-3B-Instruct alongside Phi-4-mini, enabling direct quality comparison across systems. The same three domain experts rated all three systems' outputs for the 30 session profiles on three Likert criteria (factual consistency, recommendation relevance, professional tone; 1–5 scale), blind to system identity. Results are reported in Table 4.

Phi-4-mini achieves the highest scores on all three criteria. Notably, the rule-based baseline attains near-parity on factual consistency (4.2 / 5), which is expected since it directly interpolates input numbers into fixed sentence templates, but scores markedly lower on recommendation relevance (2.8 / 5) due to its rigid priority-ladder logic, which frequently produces contextually mismatched suggestions. Llama-3.2-3B-Instruct performs competitively on professional tone (4.5 / 5) but lags behind Phi-4-mini on factual consistency (4.3 vs. 4.7), consistent with weaker instruction-following at this parameter scale. Inter-rater agreement (Krippendorff's α) is reported per criterion; values above 0.67 are considered substantial agreement following established conventions [70].

TABLE 4: Expert Likert ratings (mean $\pm$ SD, scale 1–5) and Krippendorff's α for three systems across $N = 30$ session reports. Three raters; higher is better.

System	Factual Cons.	Rec. Relevance	Prof. Tone	Avg.
Rule-Based (RB)	4.2 ± 0.4	2.8 ± 0.7	3.5 ± 0.6	3.5
Llama-3.2-3B-Instruct	4.3 ± 0.5	4.1 ± 0.6	4.5 ± 0.3	4.3
Phi-4-mini (Ours)	$\mathbf{4.7 \pm 0.3}$	$\mathbf{4.4 \pm 0.5}$	$\mathbf{4.8 \pm 0.2}$	**4.6**
Krippendorff's α	0.74	0.68	0.81	N/A

7) End-User Evaluation Study

a: Study Design.

To complement expert assessment with end-user ecological validity, we recruited 4 academic faculty members from one university under informed consent. Each participant reviewed five generated reports sampled from the 30 session profiles, comprising one report per system (RB and Phi-4-mini) and three additional Phi-4-mini reports covering diverse activity distributions, rating each on three user-oriented criteria:

- **Perceived Usefulness**: "This summary would help me reflect on my workday."
- **Clarity**: "The summary is easy to understand."

- **Actionability**: ''The recommendation is something I could realistically act on.''

Ratings used a 5-point Likert scale (1 = Strongly Disagree to 5 = Strongly Agree). Systems were presented in randomized order and participants were blind to system identity.

b: User Ratings and Qualitative Feedback.

Table 5 summarizes mean user ratings. Phi-4-mini outperforms RB on Perceived Usefulness (4.3 vs. 3.2) and Actionability (4.1 vs. 2.9), with the widest practical gap on Actionability, reflecting faculty dissatisfaction with RB's generic, context-insensitive recommendations. Clarity ratings are closer (4.5 vs. 4.1), as the RB system's direct numerical interpolation is locally straightforward to parse. Qualitative comments collected via a free-text field indicated that Phi-4-mini reports were perceived as *"empathetic and non-judgmental"* and *"actionable without being prescriptive,"* whereas RB outputs were described as *"mechanical"* and *"too generic to act on."* These findings corroborate the quantitative scores and confirm the practical value of the LLM-based approach for real-world faculty deployment.

TABLE 5: Faculty user-study ratings (mean $\pm$ SD, scale 1–5, $N = 4$ participants).

System	Usefulness	Clarity	Actionability
Rule-Based (RB)	3.2 ± 0.8	4.1 ± 0.6	2.9 ± 0.9
Phi-4-mini (Ours)	$\mathbf{4.3 \pm 0.6}$	$\mathbf{4.5 \pm 0.4}$	$\mathbf{4.1 \pm 0.7}$

8) Integration and Role in the Inference Pipeline

By translating classifier outputs into interpretable, contextually grounded narratives, this module operationalizes the system's potential for self-regulated routine optimization. Flagging disproportionate idle, sleeping, or phone-use patterns enables faculty to make informed adjustments that mitigate occupational cognitive overload [71] and support sustainable work-life balance, without requiring any familiarity with the underlying computer vision pipeline.

V. EXPERIMENTAL SETUP

In this section, we discuss our experimental results. We use a single NVIDIA RTX 4090 24GB GPU to train our TeacherActivityNet models, as well other YOLO based models along with Faster Region-based Convolutional Neural Network (R-CNN) based models.

A. HYPER PARAMETER SETTINGS

All the YOLO versions were trained using carefully chosen hyperparameters and augmentation strategies tailored to the requirements of office activity tracking tasks. The training did not allow a certain number of epochs but used early stopping with a level of patience of 5, so that when the performance of the model did not improve further, it was suitable to stop automatically. The input image files were resized to 640x640 pixels and a batch size of 16 was selected that allows the separating training as well as optimization of GPU memory usage. The large amount of data augmentation was used to increase the robustness of the model and generalization. Geometric augmentations encompassed mosaic augmentation, mixup, constrained rotation, translation, shearing and slight perspectival alteration in order to reflect upon the natural range of the office human activities and camera angles. The method of photometric adjustments was done through the regulated changes made in all the aspects of hue, saturation and brightness and they were used to achieve varied lighting conditions. Copy-paste augmentation was only recourse to diversify objects but disable vertical-flipping operation to maintain real human orientations in the office scenario.

The hyperparameter tuning process was performed iteratively using default settings, focusing on augmentation parameters, training configurations, and domain-specific adjustments. This optimization comprised nine specific classes of actions, which had a difference in terms of temporal duration and space characteristics. Spontaneous movement based activities like arriving and leaving which are highly transitional in nature as compared to other classes like sleeping and working needed stronger geometric augmentations in order to extract dynamics of movement whereas classes which have a lot of posture-based such as sleeping and working meant that there was the need to be highly conservative in transformations so that the spatial critical details were not lost. Object-interaction movements required triviality of color and shape deformities, so as to preserve significant hand object interactions. Moreover, the optimizer, initial learning rate, momentum, weight decay and other parameters shown in Table 6.

TABLE 6: **Training Hyperparameter Settings for YOLO-based Models**

Hyperparameter	Value	Notes
Optimizer	Stochastic Gradient Descent (SGD)	Momentum-based gradient descent
Initial Learning Rate	0.01	Cosine annealing schedule
Momentum	0.937	SGD momentum coefficient
Weight Decay	0.0005	L2 regularization
Warmup Epochs	3	Linear warmup phase
Batch Size	16	Optimized for GPU memory
Input Image Size	640×640	Resized from original frames
Early-Stopping Patience	5	Epochs without improvement
Maximum Epochs (cap)	100	Upper bound; early stopping applied
Hardware	NVIDIA RTX 4090 24GB	Single GPU training

B. EVALUATION METRICS: MAP@50 AND MAP@50-95

The evaluated results are reported using mAP@50 and mAP@50-95 metrics.

The quantitative measure of the YOLO model performance was the mean Average Precision (mAP), which is one of the most commonly used metrics in object detection tasks. It describes the trade-off between recall and precision by calculating the average precision (AP) over a range of confidence thresholds.

The Average Precision for each action class c, denoted as AP_c, is defined as the area under the precision-recall curve:

$$\mathrm{AP}_c = \int_0^1 \mathrm{P}_c(r)\, dr \tag{2}$$

where $P_c(r)$ represents the precision at recall level r for class c.

1) mAP@50

The mAP@50 metric calculates the mean of the Average Precision values across all classes at a fixed Intersection over Union (IoU) threshold of 0.5. It is computed as:

$$\text{mAP@50} = \frac{1}{N}\sum_{c=1}^{N} \text{AP}_c^{\text{IoU}=0.5} \tag{3}$$

where N is the total number of classes.

2) mAP@50-95

The mAP@50-95 metric provides a more comprehensive evaluation by averaging the Average Precision over multiple IoU thresholds ranging from 0.5 to 0.95 in steps of 0.05:

$$\text{mAP@50-95} = \frac{1}{N}\sum_{c=1}^{N}\frac{1}{T}\sum_{t=1}^{T} \text{AP}_c^{\text{IoU}=i_t} \tag{4}$$

where $T = 10$ represents the number of IoU thresholds $i_t \in \{0.5, 0.55, 0.6, \ldots, 0.95\}$.

This evaluation strategy ensures that the model's detection performance is assessed both at the common IoU threshold of 0.5 and across a range of stricter thresholds, providing a comprehensive understanding of localization accuracy and robustness across different action classes.

TABLE 7: **Box precision, recall, mAP@50 and mAP@50-95 results for test dataset using TeacherActivityNet**

Class	Images	Instances	P	R	mAP@50	mAP@50-95
all	808	808	0.738	0.827	0.841	0.491
Arriving	25	25	0.454	0.800	0.819	0.558
Counselling	183	183	0.952	0.965	0.990	0.593
Eating	59	59	0.610	0.371	0.413	0.270
Idle	64	64	0.663	0.688	0.700	0.406
Leaving	30	30	0.478	0.933	0.871	0.534
Sleeping	106	106	0.908	0.953	0.980	0.572
Talking	96	96	0.881	0.849	0.873	0.454
Using_Phone	119	119	0.789	0.941	0.940	0.404
Working	126	126	0.908	0.943	0.981	0.630

In Table 7, the achieved results using TeacherActivityNet for different action classes are shown separately for the test dataset. Box precision, recall, mAP@50 and mAP@50-95 value for each class is shown. The data instances are considered as images with bounding box to generate these test results. From this Table 7, we can observe that the "Eating" action class has the lowest performance in mAP@50 whereas the prediction of "Counselling" and "Working" is showing higher prediction performance. Though the dataset used contains fewer instances of Leaving and Arriving action classes, the prediction worked well for these action classes. The overall mAP@50 for the dataset using TeacherActivityNet is around 0.841 on the test dataset considering all the action classes. In Table 13 (Appendix A) the performances for each action class is shown for validation dataset using TeacherActivityNet model.

In Figure 32 (Appendix A), the confusion matrix using the TeacherActivityNet pre-trained model for the validation dataset is provided. Predictions for most of the action classes are very good on the validation dataset. The prediction of the action class "Idle" is average. The main reason behind average accuracy is due to some image frames of "Idle" action class where participants were looking down. It looks like their eyes are closed. So, the model predicts "Sleeping".

In Figure 9, prediction accuracy drops in the test dataset. Counselling class has very high accuracy. The Eating class has the lowest accuracy. The main reason behind this is actually the Using_Phone action class. In some of the image frames, we can see the participant holding the food like they are holding the mobile phone. That is why in 14 instances the model is predicting Using_Phone instead of Eating. To get a better view of the confusion matrix we have given the normalized version of the confusion matrix here in Figure 10 and in Figure 33 (Appendix A). As our test dataset

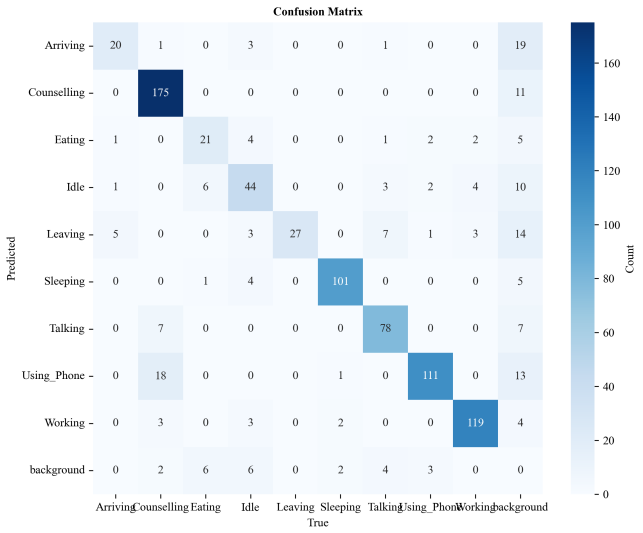


FIGURE 9: Confusion matrix for the test dataset of *TeacherActivityNet*. This figure shows how well the model generalizes to unseen test data.

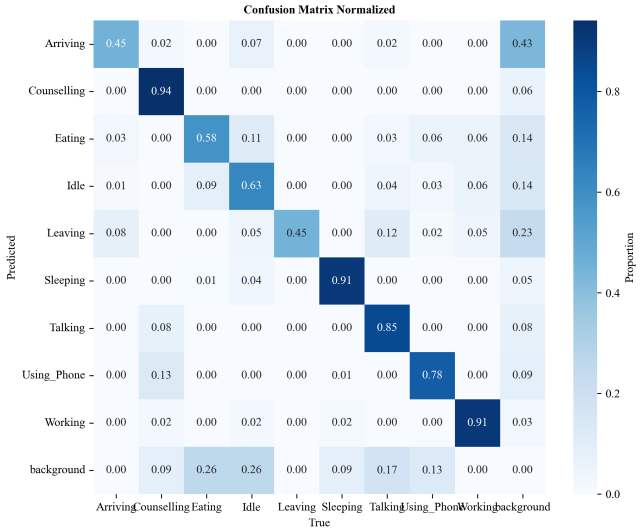


FIGURE 10: Normalized confusion matrix for the test dataset of *TeacherActivityNet*. Shows model generalization across all action classes.

is imbalanced, it will help us better understand the prediction of the model. In Figure 11 we can see the predictions for each class by the Faster R-CNN based pretrained model. The best performing action classes are "Counselling" and "Talking". However, the action classes "Arriving" and "Eating" perform poorly with a low mAP@50 of 0.3685 and 0.0382 respectively.

Based on the confusion matrices, it is evident that TeacherActivityNet outperforms the Faster R-CNN–based pre-trained models. Subsequently, we will compare it with multiple YOLO versions to evaluate whether TeacherActivityNet achieves superior performance across all models.

C. SPATIO-TEMPORAL FEATURE TUNING

While testing the dataset we have used Spatio-Temporal feature with focusing on the temporal tracking. We have found that due to dataset

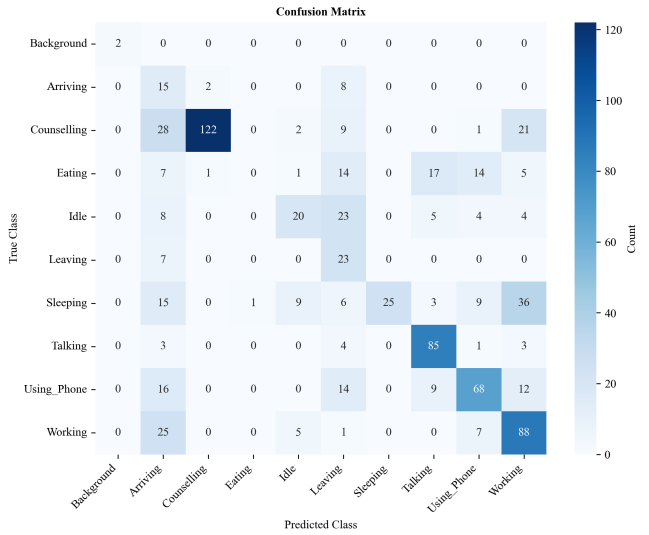


FIGURE 11: Confusion matrix for the *Faster R-CNN* model on the test dataset. This illustrates class-level detection performance of the object detection model.

imbalance even with a action class specific windows set, mAP@50 for any of the action classes were not improving. We propose an adaptive temporal smoothing consistency of actions recognition using our proposed TeacherActivityNet. Contrary to fixed windows, the system uses class-specific window size according to action duration: 3 frames of an action with a small duration (Arriving, Leaving), 5–12 frames of actions with moderate shift duration (Eating, Talking, Idle, Using_Phone, and Working) and 15–20 frames of stable actions at the same duration (Counselling, Sleeping).

The four implemented and compared smoothing strategies are (i) Adaptive Majority Voting, (ii) Confidence Weighted Voting, (iii) Weighted Averaging with Recency Bias, and (iv) Adaptive Hybrid Ensemble. These four methods achieved mAP@50 of 0.7638, 0.7749, 0.7662 and 0.7522 respectively. Comparison of the four smoothing strategies given on Appendix D. In contrast to motion-vector based algorithms, object tracking and class-specific time-buffers using IoU are used on our approach to capture the prediction consistency over time without optical flow. The results indicate that adaptive smoothing did not improve the mAP@50 performance compared to the baseline value of 0.841, with the highest mAP@50 achieved using adaptive smoothing being 0.7749.

VI. RESULTS AND ANALYSIS

A. TEST AND VALIDATION RESULTS WITH CONFUSION MATRIX

As we use TeacherActivityNet as a prediction model for our study, it has two different sets, one is used for training and the other for validation. This method takes images labeled with a bounding box. Our proposed model achieved 0.929 as the mean Average Precision(mAP@50) on the validation dataset and 0.841 on the test dataset.

B. CONFIDENCE CURVES WITH TRAIN AND VALIDATION LOSS

We analyze the performance of the confidence curves of our model, TeacherActivityNet, and evaluate the training and validation loss to assess model performance.

The confidence curves for F1 score, precision, precision-recall, and recall over the test dataset are presented. As shown in Figure 12, the confidence threshold corresponding to the maximum F1 score across all classes is 0.279, with an F1 score of 0.76. Figure 13 illustrates the average precision for all classes, reaching a maximum of 1.00 at a confidence threshold of 0.914. The precision-recall curve in Figure 14 indicates a mean Average Precision (mAP@0.5) of 0.841. Finally, Figure 15 presents the recall curve, with an average recall value of 0.95. Notably, except for precision, the "Sleeping" action class exhibits lower average performance compared to the other action classes.

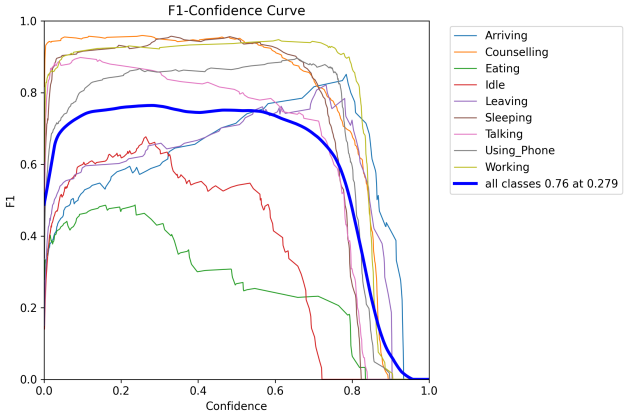


FIGURE 12: F1 curve for all activities. The figure shows the harmonic mean of precision and recall across the dataset.

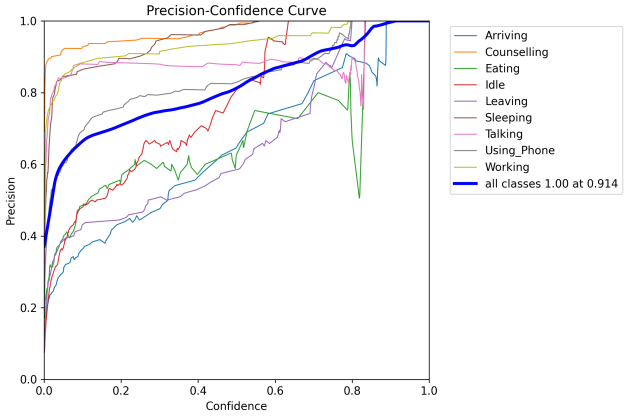


FIGURE 13: Precision curve for all activities. This illustrates the proportion of true positives among predicted positives for each class.

In Figure 16 train and validation results are shown for bounding box loss(box_loss), class prediction loss(cls_loss), distribution focal loss(dfl_loss), precision - recall for the training dataset and mAP for validation dataset. For the loss functions, it reflects a decrement of loss. It is also observed that the loss function for the training dataset is showing decreasing behavior, on the other hand, for the validation set, the result fluctuates.

C. MODEL PREDICTIONS ON VALIDATION AND TEST IMAGES

This section presents a comparison between the labeled images and the predictions made by the proposed TeacherActivityNet model. All comparisons are performed on batches of 16 images, with each batch of labeled images evaluated against the corresponding model

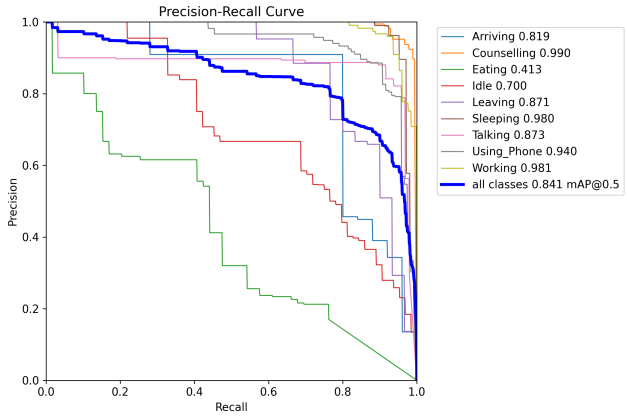


FIGURE 14: Precision-Recall curve for all activities. Shows the trade-off between precision and recall for the classifier.

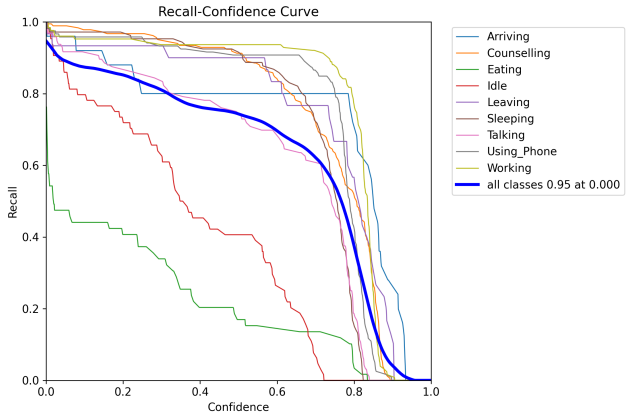


FIGURE 15: Recall curve for all activities. Illustrates the proportion of true positives correctly identified for each class.

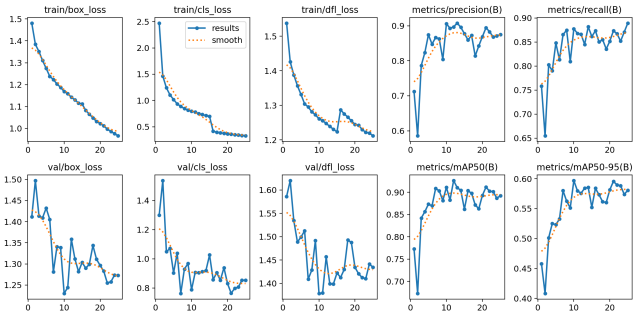


FIGURE 16: Training and validation loss along with metrics evaluation. The figure illustrates the model's convergence and performance trends over epochs.

predictions. Detailed model predictions on the validation dataset are provided in Appendix A.

In Figure 17 we can see the labels of the test images and in Figure 18 we can see the predictions made by our proposed model TeacherActivityNet. All predictions here are correct, except for one image. We can find this image in the second row first column.

In Figure 19 we can see the second set of labeled images and in Figure 20 we can see predictions of the model. Again, the model is accurate in every image frame, but the problem remains the same

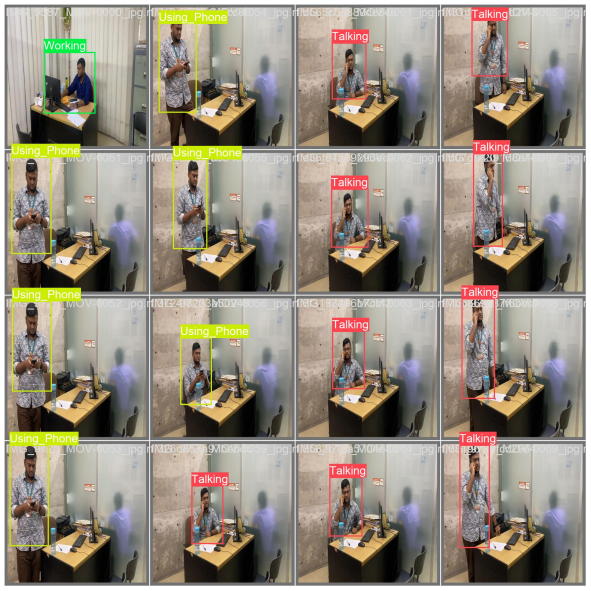


FIGURE 17: Labeled images of the test dataset (batch-0). Shows the ground truth annotations used for model evaluation.

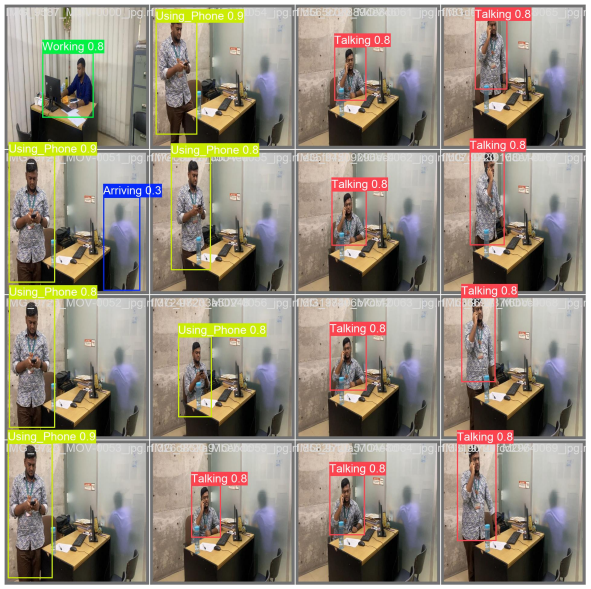


FIGURE 18: Predictions of the test dataset (batch-0). Highlights how accurately the model predicts the annotated actions.

here. The third person participates in the predictions on multiple image frames from the last column and the last two rows. The last row of the two images was merged and is shown in Figure 21. The action classes are perfectly predicted, but the model is counting his part here. That person is not even in the same room.

1) Failure Cases

We will discuss some failure scenarios of our research work. Our dataset does not include any videos recorded in low-light conditions or videos containing multiple teachers. Although the videos include multiple people and we achieved a good mAP@50 across the action classes, there are still some failure cases. In a few instances,

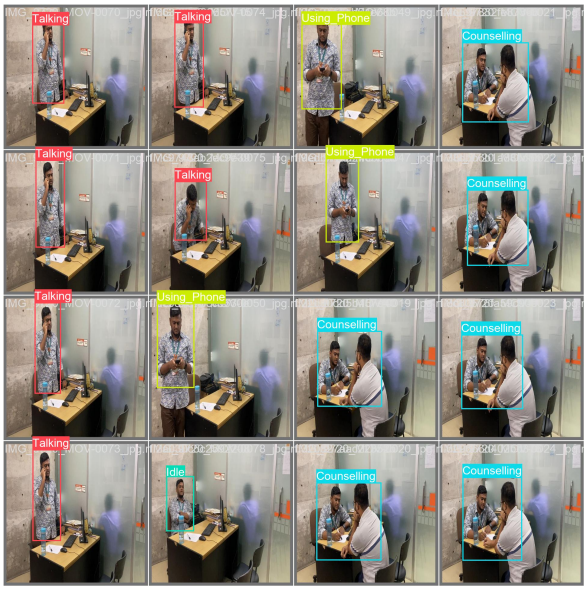


FIGURE 19: Labeled images of the test dataset (batch-1). Shows the ground truth annotations used for model evaluation.

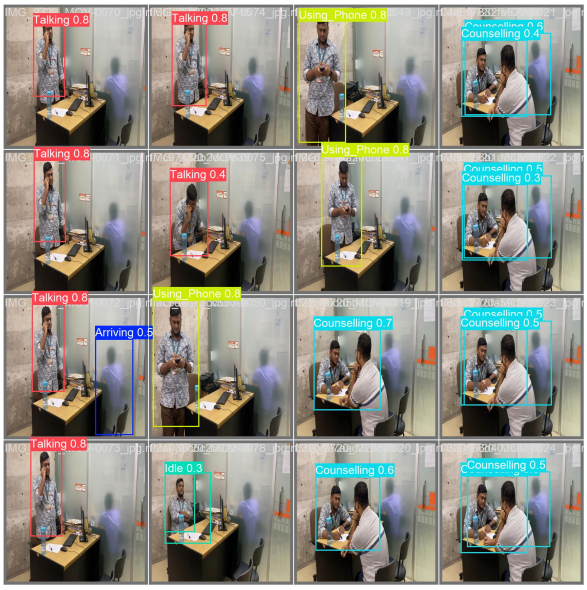


FIGURE 20: Predictions of the test dataset (batch-1). Highlights how accurately the model predicts the annotated actions.

individuals located in another room were mistakenly detected and classified as performing certain actions. In Figure 22. The image frame has action class "Using_Phone", but the model predicts two action classes. One is correct, which is "Using_Phone" but another is an incorrect prediction for the action class "Arriving". Although it looks like a correct prediction, we can see the back of someone who is entering the room maybe. But in reality there was a transparent glass from which we can see the next room, where a person was just sitting in a chair. Our model predicted that and it becomes confused, but at least even if the picture is blur our model can predict action class, that is the only positive sign of our model TeacherActivityNet. The same problem persists in Figure 21 as well. The action class

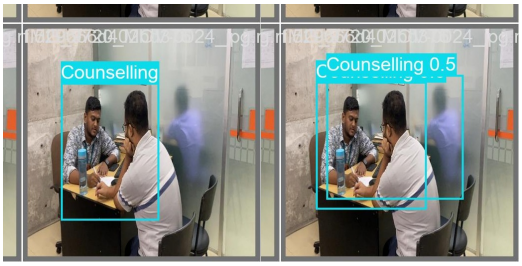


FIGURE 21: Merged view of ground truth labels and model predictions side by side for the test dataset (batch-1). This figure demonstrates the model's performance on different test batches.

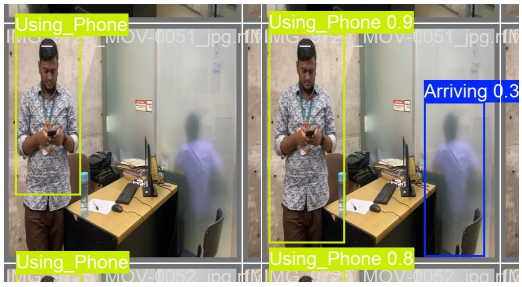


FIGURE 22: Merged view of ground truth labels and model predictions side by side for the test dataset (batch-0). This visualization allows easy comparison between actual annotations and predicted outputs.

is "Counselling", but the model predicts multiple "Counselling" instances because the person sitting in the next room is visible again through a transparent glass door.

To evaluate the model's performance under varying illumination, we augmented the training dataset by adjusting brightness levels by ±25%. The proposed model achieved an mAP@50 of 0.796 on the test set and 0.928 on the validation set. Detailed results are provided in Appendix E.

In summary, the proposed TeacherActivityNet model demonstrates strong performance in predicting individual instances. For several action classes, prediction scores are particularly high, reaching up to 0.9, indicating strong model confidence. Although YOLO-based models are primarily designed for object detection rather than action recognition, *TeacherActivityNet* achieves accurate action predictions for the majority of image instances. Some challenges remain for the "Eating" action class; however, it is expected that, similar to other classes, prediction performance for these instances can be further improved in future work.

D. BENCHMARKING WITH FASTER RCNN MODEL PERFORMANCE

In order to evaluate some other method results using our dataset, we use the Faster R-CNN ResNet50 Feature Pyramid Network (FPN) pre-trained model. Overall mAP@50 is 0.543 and detailed results for the Faster R-CNN ResNet50 FPN model are provided in Appendix B.

E. MODEL PERFORMANCE ON VIDEOS

All the results above were generated using images with specific bounding boxes. To evaluate real-time performance across all nine action classes, we used a 127-second video from the test dataset.

In the initial phase of the task, we perform face recognition to correctly predict the person. We have taken a picture of the

participant who is present in the video to perform the recognition task. Using the third-party face_recognition Python library, which employs the dlib ResNet-29 model, we detected the person. We can easily detect the first and last appearances of the person in a video.

After the face recognition process is performed successfully, we use our proposed model *TeacherActivityNet* to predict the action class for each task. To compare the predicted results and generate the accurately predicted tasks, the video was divided into image frames and annotated using Roboflow. After that, the predicted and annotated image frames are compared to find the correctly predicted instances. In Table 8, the annotated frame numbers and the predicted frame numbers are given for each action class. The total predicted time for each action class is also provided to make a decision on the overall task prediction for each person. The ''No Detection'' class indicates the number of times the model predicted that the person was not involved in any of the 9 available action classes. ''No Detection'' can be correct or incorrect. This action class in real-time videos is very much possible due to the frequent switching of actions by the participant. The time when the participant is switching our proposed model *TeacherActivityNet* will not be able to predict any of the action classes.

TABLE 8: **Predicted frame counts for real-time video from test dataset using the model**

Action Class	Input Frames	Predicted Frames	Accurately Predicted Frames
Arriving	3	3	1
Counselling	29	38	29
Eating	7	0	0
Idle	4	26	4
Leaving	3	8	3
Sleeping	12	12	11
Talking	8	11	5
Using_Phone	10	14	10
Working	12	2	2
No Detection	N/A	15	N/A

F. ABLATION STUDIES

In Table 9, a comparison of different convolutional models in YOLOv8 with *TeacherActivityNet* is shown. We will try to find out if any convolutional module performs better than the *TeacherActivityNet*.

Depthwise Separable Convolutional Module(DSC of YOLOv8): This method uses the DWConv convolutional module and has mAP@50 lower than the other two modules.

Simple Grouped Shuffle Convolutional Module(Simple GS of YOLOv8): This is a combination of Standard Convolutional Module and Depthwise Separable Convolution module for group sampling. We can clearly see that it has the best inference speed of 1.82 ms/image compared to the other two methods.

Standard Convolutional Model(SC Module of YOLOv8): This typically refers to a deep learning model that uses standard convolutional layers for feature extraction. From Table 9, it is observed that this model has outperformed the other two applied methods(DSC and Simple GS). It has mAP@50 on the test dataset of 0.720.

TeacherActivityNet: From Table 9, we can say that our proposed model *TeacherActivityNet* demonstrates superior performance than other convolutional modules with mAP@50 of 0.841 in the test dataset. However, the inference speed of the Simple GS convolutional module is marginally better than that of *TeacherActivityNet*. Although the *TeacherActivityNet* inference speed is slower, compared to other convolutional modules, we can trade a little speed for more accuracy.

In Table 9, we examine the impact of incorporating spatio-temporal features and dataset configuration on model performance. The spatio-temporal feature indicates whether the model processes multiple frames simultaneously, enabling it to capture motion and temporal dynamics across consecutive frames. Regarding dataset configuration, the full dataset consists of 19,494 training images with augmentation applied, while the partial dataset contains only 6,498 training images without augmentation. The results demonstrate that leveraging spatio-temporal features alongside the full dataset yields the best performance, as reflected in the mAP@50 scores of our *TeacherActivityNet* model.

TABLE 9: **Ablation study on model components and dataset configuration**

Module Setting	Detection Backbone	Spatio-Temporal Feature	Dataset Configuration	mAP@50	Inference Speed (ms/image)
DSC	YOLOv8n (DSC)	No	Full	0.473	2.16
SimpleGS	YOLOv8n (SimpleGS)	No	Full	0.503	1.82
SC	YOLOv8n (SC)	No	Full	0.72	2.1
Teacher ActivityNet	Modified YOLOv8n	Yes	Partial	0.704	1.9
Teacher ActivityNet	Modified YOLOv8n	Yes	Full	0.841	1.9

G. YOLO MODELS VS *TEACHERACTIVITYNET*

In this section, we are going to compare some of the YOLO-based models, such as YOLOv5, YOLOv8, YOLOv9, YOLOv10, YOLOv11, and YOLOv12, with *TeacherActivityNet*. We will compare their results based on mAP@50 on the validation and test datasets. In Table 10, we have compared the performance of most of the YOLO-based models with *TeacherActivityNet* on the test and validation datasets. In the test dataset, *TeacherActivityNet* has outperformed all other models with an mAP@50 of 0.841. The second-best model is YOLOv11 with an mAP@50 of 0.743. Detailed results for the different YOLO versions are provided in Appendix C. Table 10 summarizes the comparison of all YOLO versions with the *TeacherActivityNet* model.

TABLE 10: **Summary of the performance comparison of YOLO-based models with *TeacherActivityNet***

Model	Test mAP@50	Validation mAP@50	Inference Speed (ms/image)
YOLOv5	0.697	0.923	3.8
YOLOv8	0.720	0.941	2.1
YOLOv9	0.733	0.932	5.7
YOLOv10	0.736	0.927	3.1
YOLOv11	0.743	0.922	2.0
YOLOv12	0.715	0.932	3.0
TeacherActivityNet	0.841	0.929	1.9

In summary, we can say that our proposed *TeacherActivityNet* model has outperformed the rest of the YOLO models. It also has a better inference speed as well.

H. ACTIVITY RECOGNITION MODEL PERFORMANCE

Our main task is to recognize the activities of the teachers. To achieve this, we have explored several activity recognition models to gain insight into spatial and temporal-based approaches. We have analyzed the performance of three models: two of them, MC3-18 (Mixed Convolution 3D with 18 layers) and R2Plus1D (R(2+1)D, a factorized 3D convolution network), are based on the ResNet-18 architecture, while the third, the Temporal Segment Transformer

(TST), is a custom 3D CNN model built on ResNet3D-18. The TST is enhanced with temporal pooling and a fully connected classifier, designed for efficient spatio-temporal feature learning and action recognition in video data.

TABLE 11: **Comparing activity recognition models**

Metric	MC3-18	R2plus1D	TST
Validation Accuracy	48.15%	40.74%	70.42%
Test Accuracy	37.11%	32.85%	58.34%

In Table 11, we can see that the performance of the activity recognition models is not satisfactory, with the performance of the test dataset being notably poor. These models were trained directly on videos. Based on the results, we can conclude that the *TeacherActivityNet* and FasterRCNN-based pre-trained models outperform the activity recognition models.

1) Causes of Poor Activity Recognition Model Performance

The main reason behind the poor performance of the activity recognition models is that they focus on the entire image frame, rather than the specific region that actually represents the action. In YOLO-based models, we have image annotation with bounding box labels available. These YOLO-based models focus only on the specific regions of the image, not on the unnecessary parts of the image frame. Some may suggest that we should have cropped the image frames to show only the person. However, if we had done that, another problem would have emerged. Some of our defined actions depend on the importance of objects in the image frames, not just the person. For example, if someone is performing the "Working" action class and we crop the person from the image frame, we would no longer see the computer or laptop, which are necessary for the "Working" action class. According to the definition of the class, those objects are essential. Therefore, we cannot crop the videos or extracted frames of the videos. In Figure 34a (Appendix A) and in Figure 34b (Appendix A), we have shown the comparison between the original image and the cropped image. In Figure 34a, we can see that the monitor is visible, which is responsible for the "Working" action class. In Figure 34b, we can no longer see the monitor. These activity recognition models that focus on the entire frame lead to poor performance overall. We can conclude that *TeacherActivityNet*, or any other YOLO model, would definitely perform better in our task of faculty personalised assistive system.

I. REASON BEHIND NO IMPROVEMENT IN ADAPTIVE ANALYSIS

The primary reason behind the poor improvement in the adaptive window size analysis is that, in our test dataset, the class-wise image instances are not balanced. This is because the actions are different and cannot be performed by the participants for the same period of time. For example, the time for the Arriving and Leaving action classes is shorter than for the other action classes. This causes an imbalance in the test dataset. Additionally, multiple participants performed the actions in different ways, taking varying amounts of time. Therefore, we cannot set a single window size for all the classes. Even with different, class-specific window sizes, our model still fails to improve prediction rates. In Figure 23 we can see the window size is set according to action class. And in Figure 24 we can see no action class is getting an improvement.

J. QUALITATIVE ANALYSIS OF GENERATED WELLNESS REPORTS

To illustrate the end-to-end behavior of *TeacherActivityNet*, we present representative natural language reports generated by the explanation module for two contrasting session profiles drawn from the test set. Table 12 shows the input activity profiles alongside the corresponding generated reports.

Session S1 exemplifies a pattern of productive engagement punctuated by significant low-engagement intervals: the combined idle and sleeping fraction (23.9%) triggers a micro-break recommendation aligned with evidence on attentional restoration [64]. Session S2 presents a contrasting profile dominated by phone use, prompting the module to generate a digital boundary recommendation focused on reducing attentional switching costs. Together, these examples demonstrate the module's capacity to produce contextually differentiated, non-generic reports that directly address the cognitive and work-life balance dimensions central to the *TeacherActivityNet* framework.

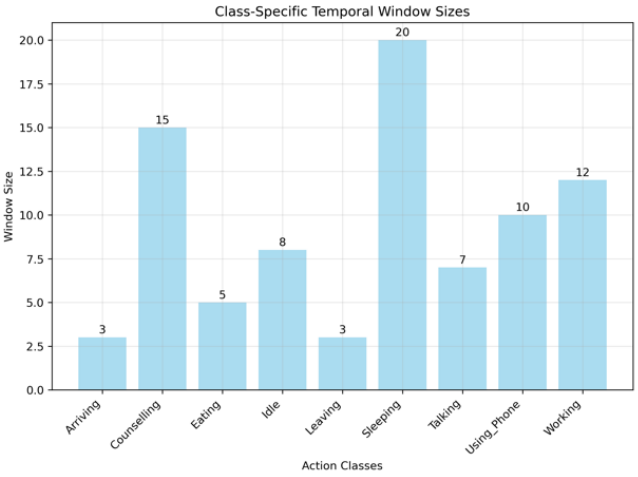


FIGURE 23: Adaptive window size analysis. The figure shows how varying the window size affects the system's performance metrics.

FIGURE 24: Adaptive improvement rate. This figure illustrates the rate of performance improvement as adaptive parameters are applied over time.

VII. CONCLUSION

This paper presented *TeacherActivityNet*, a lightweight, real-time Human Activity Recognition system purpose-built for academic office environments. To support this work, we introduced *PETRA*, the first faculty-centric video benchmark comprising nine annotated office activity classes, addressing a critical gap in the HAR literature.

TABLE 12: Representative input–output pairs from the Natural Language Explanation Module, drawn from two real-world deployment sessions recorded in an actual academic workplace setting. Session proportions are computed relative to total detected activity duration. Reports are generated fully on-device with no post-editing.

Session	Activity Profile	Generated Report
S1	Working: 48 min (30.2%), Counselling: 38 min (23.9%), Idle: 26 min (16.4%), Sleeping: 12 min (7.5%), Using_Phone: 14 min (8.8%), Talking: 11 min (6.9%), Arriving: 5 min (3.1%), Leaving: 5 min (3.1%), Eating: 0 min (0%). Total: 159 min.	*"Today you dedicated approximately 48 minutes to focused work and 38 minutes to student counselling, fulfilling core academic responsibilities. Nonetheless, 26 minutes of idle time and 12 minutes of desk sleeping suggest recurring low-engagement intervals.* ***Recommendation:*** *Introduce 5-minute movement breaks every 60–90 minutes when low-engagement states exceed 20% of session duration, a strategy supported by evidence on attentional restoration and fatigue mitigation in desk-based knowledge work."*
S2	Working: 35 min (22.9%), Counselling: 12 min (7.8%), Idle: 8 min (5.2%), Sleeping: 4 min (2.6%), Using_Phone: 55 min (35.9%), Talking: 20 min (13.1%), Arriving: 6 min (3.9%), Leaving: 6 min (3.9%), Eating: 7 min (4.6%). Total: 153 min.	*"Your session reflects a relatively balanced spread of professional activities, with 35 minutes of focused work and 20 minutes of collegial interaction. However, 55 minutes of phone use constitutes the single largest activity block at 35.9% of the session, which may fragment sustained cognitive engagement.* ***Recommendation:*** *Consider consolidating phone-related tasks into two designated time windows per session to reduce attentional switching costs and protect uninterrupted deep work periods."*

The fine-tuned YOLOv8n-based detector achieves 0.841 mAP@50 on *PETRA*, outperforming YOLOv11, YOLOv12, Faster R-CNN-based systems, and established activity recognition frameworks across most evaluation scenarios, while remaining suitable for real-time edge deployment. Beyond recognition, we conducted the first systematic comparison of image-based and video-based inference modalities for workplace HAR, yielding actionable insights for practical deployment trade-offs. Crucially, the system extends beyond activity classification by integrating a LLM-powered Natural Language Explanation Module that converts per-session activity duration statistics into human-readable wellness summaries and evidence-grounded cognitive ergonomics recommendations. Expert evaluation confirmed the quality of generated reports, with mean scores above 4.4 out of 5 across factual consistency, recommendation relevance, and professional tone. Together, these contributions establish a non-intrusive, end-to-end framework for automated faculty wellness interventions that promote work-life balance and cognitive sustainability in academic institutions, without reliance on manual self-reporting.

A. LIMITATIONS

The current annotation scheme treats activities as mutually exclusive and does not account for concurrent or overlapping actions, such as a teacher simultaneously counseling a student while eating. Moreover, the *PETRA* and *TeacherActivityNet* assume a single teacher per scene; while the underlying YOLO person detector can in principle crop and process multiple individuals independently, this multi-occupant pipeline has not been formally evaluated. A practical path to multi-occupant deployment, in which a multi-object tracking algorithm such as DeepSORT is paired with the existing dlib face recognition module to assign each detected person a persistent track identity and an independent activity timeline, is detailed in the Future Work subsection. These limitations motivate the directions outlined in the following Future Work subsection.

B. FUTURE WORK

The authors are planning to work on mixed-action classes and intend to work with multiple teachers. We will compare the performance of *TeacherActivityNet* with other similar datasets, if available, to justify its overall performance. For action classes with lower average precision, we can work on pose estimation as well. The foundation of our work allows us to create custom assistive systems for diverse office workspaces which seek to boost mental health and maximize work hours while boosting office output. We will try to figure out why the modifications made to the backbone of YOLOv8 do not work well with the rest of the YOLO models. Specifically, with YOLOv12, we will try some additional modifications to improve its performance. As shown by the consistently low Eating mAP@50 of approximately 0.28 across all four adaptive-window strategies (Table 21 to Table 24 in Appendix; e.g., 0.2821 in Table 23), this class remains the most challenging owing to object-invariance and data-imbalance effects identified in this study. We therefore plan to fuse object-aware pose features (explicit food and utensil detection combined with wrist-to-head trajectory cues) with class-balanced resampling and focal-loss weighting to strengthen detection performance for such low-scoring classes. In future deployments, multi occupant offices can be supported by pairing a standard multi object tracking algorithm such as DeepSORT with the existing dlib face recognition module, so that each detected individual is assigned a persistent track identity and an independent per person activity timeline. In addition, we would develop an application with a graphical user interface, that would enhance the usability and practical impact of our work.

ACKNOWLEDGMENT

Funded by the Institute for Advanced Research Publication Grant of United International University (Ref. No.: IAR-2025-Pub-082). The authors also used GPT-5 to review and improve the grammatical consistency of the manuscript.

REFERENCES

[1] B. Wu, M. Xue, Y. Jia, N. Zhang, G. Zhao, X. Wang, and C. Zhang, "Lightweight and efficient skeleton-based sports activity recognition with astm-net," *PLoS One*, vol. 20, no. 7, p. e0324605, 2025.

[2] M. Lu, T. Qi, C. Ci, Z. Ren, S. Zhang, and Y. Lv, "Mtcf-net: Leveraging large models for multimodal time series analysis in sports and fitness consumer electronics," *IEEE Transactions on Consumer Electronics*, 2025.

[3] H. Yin, R. O. Sinnott, and G. T. Jayaputera, "A survey of video-based human action recognition in team sports," *Artificial Intelligence Review*, vol. 57, no. 11, p. 293, 2024.

[4] C. Zheng and Y. Zhou, "Multi-modal iot data fusion for real-time sports event analysis and decision support," *Alexandria Engineering Journal*, vol. 128, pp. 519–532, 2025.

[5] P. Grzegorz, P.-M. Justyna, K. Pascal, R. Matthias, P. Iwona, S. Holger, and D. Martin, "Enhancing inertial sensor-based sports activity recognition through reduction of the signals and deep learning," *Expert Systems with Applications*, vol. 263, p. 125693, 2025.

[6] Z. Zhao, C. He, X. Jiang, and X. Xu, "Self-adaptive uncertainty modeling and relation reasoning for cross-modal egocentric human action recognition in sports," *Computers and Electrical Engineering*, vol. 127, p. 110583, 2025.

[7] M. Toshpulatov, W. Lee, S. Lee, H. Yoon, and U. Kang, "Ddc3n: Doppler-driven convolutional 3d network for human action recognition," *IEEE Access*, 2024.

[8] Y. Cao and T. Li, "Svm action recognition model based on skeletal key point analysis with posture sensors to help sports training," *BMC Sports Science, Medicine and Rehabilitation*, vol. 17, no. 1, p. 220, 2025.

[9] J. Li, Y. Fan, J. Gong, J. Chen, R. Chen, W. Zhang, and Y. Zhao, "A robust and interpretable framework for sports activity recognition based on wearable sensor signals and image representations," *Engineering Applications of Artificial Intelligence*, vol. 167, p. 113500, 2026.

[10] H.-J. Lee and S.-J. Buu, "Wi-fi-enabled vision via spatially-variant pose estimation based on convolutional transformer network," *IEEE Access*, 2025.

[11] Y. Zhang, S. You, S. Karaoglu, and T. Gevers, "3d human pose estimation and action recognition using fisheye cameras: A survey and benchmark," *Pattern Recognition*, p. 111334, 2025.

[12] H. Vijay, S. Mathur, S. Agarwal, T. Perumal, and A. Sharma, "Motionscope ai: Comprehensive human activity recognition through integrated pose analysis and temporal modeling," *IEEE Sensors Journal*, 2026.

[13] M. Skipanes, G. Demartini, K. Franke, and A. B. Nissen, "Information analysis in criminal investigations: methods, challenges, and computational opportunities processing unstructured text," *Policing: A Journal of Policy and Practice*, vol. 19, p. paaf005, 2025.

[14] V. Kshirsagar, N. Pachpor, S. Suryawanshi, T. Chavan, N. J. Nair, P. Agrawal, and T. Shahane, "Artificial intelligence powered crime scene analysis service," *MethodsX*, vol. 15, p. 103430, 2025.

[15] M. Karim, S. Khalid, S. Lee, S. Almutairi, A. Namoun, and M. Abohashrh, "Next generation human action recognition: A comprehensive review of state-of-the-art signal processing techniques," *IEEE Access*, 2025.

[16] Z. Lai, J. Yang, S. Xia, Q. Wu, Z. Sun, W. Yu, and L. Pei, "Smart: Scene-motion-aware human action recognition framework for mental disorder group," *IEEE Internet of Things Journal*, 2024.

[17] T. S. Qureshi, M. H. Shahid, A. A. Farhan, and S. Alamri, "A systematic literature review on human activity recognition using smart devices: advances, challenges, and future directions," *Artificial Intelligence Review*, vol. 58, no. 9, p. 276, 2025.

[18] X. Xu, J. Yang, V. Govindarajan, N. Alturki, G. Kumar, L. Yee, and A. K. Bashir, "Neurosymbolic ai empowered consumer electronics healthcare for wifi-based human activity recognition," *IEEE Transactions on Consumer Electronics*, 2025.

[19] F. M. Ben Williamson and J. Potter, "Re-examining ai, automation and datafication in education," *Learning, Media and Technology*, vol. 48, no. 1, pp. 1–5, 2023.

[20] E. Dimitriadou and A. Lanitis, "A critical evaluation, challenges, and future perspectives of using artificial intelligence and emerging technologies in smart classrooms," *Smart Learning Environments*, vol. 10, no. 1, Feb. 2023. [Online]. Available: http://dx.doi.org/10.1186/s40561-023-00231-3

[21] M. K. Saini and N. Goel, "How smart are smart classrooms? a review of smart classroom technologies," vol. 52, no. 6, 2019. [Online]. Available: https://doi.org/10.1145/3365757

[22] H. C. Liu, J. H. Chuah, A. S. M. Khairuddin, X. M. Zhao, and X. D. Wang, "Campus abnormal behavior recognition with temporal segment transformers," *IEEE Access*, vol. 11, pp. 38 471–38 484, 2023.

[23] W. N. Ismail, H. A. Alsalamah, M. M. Hassan, and E. Mohamed, "Auto-har: An adaptive human activity recognition framework using an automated cnn architecture design," *Heliyon*, vol. 9, no. 2, 2023.

[24] N. Gaud, M. Rathore, and U. Suman, "Mhcnls-har: Multi-headed cnn-lstm based human activity recognition leveraging a novel wearable edge device for elderly health care," *IEEE Sensors Journal*, 2024.

[25] S. Zhu, R. G. Guendel, A. Yarovoy, and F. Fioranelli, "Continuous human activity recognition with distributed radar sensor networks and cnn–rnn architectures," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[26] N. Bansal, A. Bansal, and M. Gupta, "Analyzing the variability of rnn hyperparameters and architectures for har with wearable sensor data," in *2024 3rd International conference on Power Electronics and IoT Applications in Renewable Energy and its Control (PARC)*. IEEE, 2024, pp. 232–237.

[27] N. K. Singh and K. S. Suprabhath, "Har using bi-directional lstm with rnn," in *2021 International Conference on Emerging Techniques in Computational Intelligence (ICETCI)*. IEEE, 2021, pp. 153–158.

[28] Y. Hu, X.-Q. Zhang, L. Xu, F. X. He, Z. Tian, W. She, and W. Liu, "Harmonic loss function for sensor-based human activity recognition based on lstm recurrent neural networks," *IEEE Access*, vol. 8, pp. 135 617–135 627, 2020.

[29] A. Y. Shdefat, N. Mostafa, Z. Al-Arnaout, Y. Kotb, and S. Alabed, "Optimizing har systems: comparative analysis of enhanced svm and k-nn classifiers," *International Journal of Computational Intelligence Systems*, vol. 17, no. 1, p. 150, 2024.

[30] K. Maswadi, N. A. Ghani, S. Hamid, and M. B. Rasheed, "Human activity classification using decision tree and naïve bayes classifiers," *Multimedia Tools and Applications*, vol. 80, no. 14, pp. 21 709–21 726, 2021.

[31] A. P. Kshirsagar and H. Azath, "Yolov3-based human detection and heuristically modified-lstm for abnormal human activities detection in atm machine," *Journal of Visual Communication and Image Representation*, vol. 95, p. 103901, 2023.

[32] A. A. Mohy, H. A. Bassioni, E. O. Elgendi, and T. M. Hassan, "Innovations in safety management for construction sites: the role of deep learning and computer vision techniques," *Construction Innovation*, 2024.

[33] M. Pinnelli, I. D. P. M. Albanil, A. Ledda, E. Schena, R. Setola, and C. Massaroni, "Design and performance evaluation of a smart helmet system for human activity recognition in worker safety applications," *IEEE Sensors Journal*, vol. 26, no. 2, pp. 3206–3215, 2025.

[34] A. Sochopoulos, T. Poliero, J. Ahmad, D. G. Caldwell, and C. Di Natali, "Human activity recognition algorithms for manual material handling activities," *Scientific Reports*, vol. 15, no. 1, p. 10954, 2025.

[35] M. Ayub and E.-S. M. El-Alfy, "Enhanced human activity recognition using mixture of swin transformers with recurrence plots of triaxial sensor data," *IEEE Access*, vol. 13, pp. 195 719–195 734, 2025.

[36] G. Gisha, M. Thangavel, and J. D. Udayan, "Human activity recognition and behavioural prediction: a comprehensive systematic review," *Multimedia Tools and Applications*, vol. 84, no. 40, pp. 48 849–48 893, 2025.

[37] X. Liu, F. Xu, Z. Zhang, and K. Sun, "Fall-portent detection for construction sites based on computer vision and machine learning," *Engineering, Construction and Architectural Management*, vol. 32, no. 3, pp. 1499–1521, 2025.

[38] N. C. Hoang, M. T. Le, K. Nguyen, Q. C. Nguyen *et al.*, "Deep learning-based human activity recognition with fmcw radar: A review," *IEEE Sensors Journal*, 2026.

[39] I. Yousif, J. Samaha, J. Ryu, and R. Harik, "Safety 4.0: Harnessing computer vision for advanced industrial protection," *Manufacturing Letters*, vol. 41, pp. 1342–1356, 2024.

[40] M. Ajay Srivathsan and R. Rajesh, "Work track: An automated employee activity tracking and monitoring system."

[41] P. Yuganthini, A. Vigneswari, S. Jancy, M. AntoPraveena *et al.*, "Activity tracking of employees in industries using computer vision," in *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*. IEEE, 2021, pp. 1321–1329.

[42] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[43] N. Sikder and A.-A. Nahid, "Ku-har: An open dataset for heterogeneous human activity recognition," *Pattern Recognition Letters*, vol. 146, pp. 46–54, 2021.

[44] A. Krishan Kumar, A. Kaushal Kumar, and S. Guo, "Two viewpoints based real-time recognition for hand gestures," *IET Image Processing*, vol. 14, no. 17, pp. 4606–4613, 2020.

[45] A. M. Conway, A. M. Durso, M. F. Kinnaird, C. Packer, M. Kosmala, and R. Kays, "Frame-by-frame annotation of video recordings using deep neural networks," *Ecosphere*, vol. 12, no. 3, p. e03384, 2021.

[46] M. Rashmi, T. Ashwin, and R. M. R. Guddeti, "Surveillance video analysis for student action recognition and localization inside computer laboratories of a smart campus," *Multimedia Tools and Applications*, vol. 80, no. 2, pp. 2907–2929, 2021.

[47] C. Pabba and P. Kumar, "A vision-based multi-cues approach for individual students' and overall class engagement monitoring in smart classroom environments," *Multimedia Tools and Applications*, vol. 83, no. 17, pp. 52 621–52 652, 2024.

[48] Y. Yang, F. Angelini, and S. M. Naqvi, "Pose-driven human activity anomaly detection in a cctv-like environment," *IET Image Processing*, vol. 17, no. 3, pp. 674–686, 2023.

[49] S. Wang, J. Cao, and P. S. Yu, "Deep learning for spatio-temporal data mining: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 8, pp. 3681–3709, 2019.

[50] V. Singh, S. Singh, and P. Gupta, "Real-time anomaly recognition through cctv using neural networks," *Procedia Computer Science*, vol. 173, pp. 254–263, 2020.

[51] D. Shreyas, S. Raksha, and B. Prasad, "Implementation of an anomalous human activity recognition system," *SN Computer Science*, vol. 1, no. 3, p. 168, 2020.

[52] K. Latha, M. Musammil *et al.*, "Human action recognition using deep learning methods (cnn-lstm) without sensors," in *2022 1st International Conference on Computational Science and Technology (ICCST)*. IEEE, 2022, pp. 447–451.

[53] R. Nale, M. Sawarbandhe, N. Chegogoju, and V. Satpute, "Suspicious human activity detection using pose estimation and lstm," in *2021 International symposium of Asian control Association on Intelligent Robotics and Industrial Automation (IRIA)*. IEEE, 2021, pp. 197–202.

[54] K. Host and M. Ivašić-Kos, "An overview of human action recognition in sports based on computer vision," *Heliyon*, vol. 8, no. 6, 2022.

[55] L. Xiao, Y. Cao, Y. Gai, E. Khezri, J. Liu, and M. Yang, "Recognizing sports activities from video frames using deformable convolution and adaptive multiscale features. j. cloud comput. 12 (1), 167 (2023)," 2023.

[56] G. L. Goh, G. D. Goh, J. W. Pan, P. S. P. Teng, and P. W. Kong, "Automated service height fault detection using computer vision and machine learning for badminton matches," *Sensors*, vol. 23, no. 24, p. 9759, 2023.

[57] E. Selvi, M. Adimoolam, G. Karthi, K. Thinakaran, N. M. Balamurugan, R. Kannadasan, C. Wechtaisong, and A. A. Khan, "Suspicious actions detection system using enhanced cnn and surveillance video," *Electronics*, vol. 11, no. 24, p. 4210, 2022.

[58] K. Menaka, R. V. Raj, C. V. Aravind, V. Vamsi, S. Fardeen, and G. Yugendar, "Yolo algorithm-based suspicious activity detection in atm surveillance," in *2024 International Conference on Advances in Modern Age Technologies for Health and Engineering Science (AMATHE)*. IEEE, 2024, pp. 1–5.

[59] T. Nandhini and K. Thinakaran, "Deep neural network-based crime scene detection with frames," in *2023 Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)*. IEEE, 2023, pp. 1–8.

[60] M. M. Rahman, D. Gupta, S. Bhatt, S. Shokouhmand, and M. Faezipour, "A comprehensive review of machine learning approaches for anomaly detection in smart homes: Experimental analysis and future directions," *Future Internet*, vol. 16, no. 4, p. 139, 2024.

[61] S. Ankalaki and M. Thippeswamy, "A novel optimized parametric hyperbolic tangent swish activation function for 1d-cnn: application of sensor-based human activity recognition and anomaly detection," *Multimedia Tools and Applications*, vol. 83, no. 22, pp. 61 789–61 819, 2024.

[62] S. Mulyani, A. A. Salameh, A. Komariah, A. Timoshin, N. A. A. N. Hashim, R. S. P. Fauziah, M. Mulyaningsih, I. Ahmad, and S. M. Ul din, "Emotional regulation as a remedy for teacher burnout in special schools: Evaluating school climate, teacher's work-life balance and children behavior," *Frontiers in Psychology*, vol. 12, p. 655850, 2021.

[63] A. Abdulaziz, M. Bashir, and A. A. Alfalih, "The impact of work-life balance and work overload on teacher's organizational commitment: do job engagement and perceived organizational support matter," *Education and Information Technologies*, vol. 27, no. 7, pp. 9641–9663, 2022.

[64] P. Albulescu, I. Macsinga, A. Rusu, C. Sulea, A. Bodnaru, and B. T. Tulbure, ""Give me a break!" a systematic review and meta-analysis on the efficacy of micro-breaks for increasing well-being and performance," *PLOS ONE*, vol. 17, no. 8, p. e0272460, 2022.

[65] Microsoft, :, A. Abouelenin, A. Ashfaq, A. Atkinson, H. Awadalla, N. Bach, J. Bao, A. Benhaim, M. Cai, V. Chaudhary, C. Chen, D. Chen, D. Chen, J. Chen, W. Chen, Y.-C. Chen, Y. ling Chen, Q. Dai, X. Dai, R. Fan, M. Gao, M. Gao, A. Garg, A. Goswami, J. Hao, A. Hendy, Y. Hu, X. Jin, M. Khademi, D. Kim, Y. J. Kim, G. Lee, J. Li, Y. Li, C. Liang, X. Lin, Z. Lin, M. Liu, Y. Liu, G. Lopez, C. Luo, P. Madan, V. Mazalov, A. Mitra, A. Mousavi, A. Nguyen, J. Pan, D. Perez-Becker, J. Platin, T. Portet, K. Qiu, B. Ren, L. Ren, S. Roy, N. Shang, Y. Shen, S. Singhal, S. Som, X. Song, T. Sych, P. Vaddamanu, S. Wang, Y. Wang, Z. Wang, H. Wu, H. Xu, W. Xu, Y. Yang, Z. Yang, D. Yu, I. Zabir, J. Zhang, L. L. Zhang, Y. Zhang, and X. Zhou, "Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras," 2025. [Online]. Available: https://arxiv.org/abs/2503.01743

[66] A. Grattafiori *et al.*, "The llama 3 herd of models," 2024. [Online]. Available: https://arxiv.org/abs/2407.21783

[67] Qwen, :, A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Tang, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, and Z. Qiu, "Qwen2.5 technical report," 2025. [Online]. Available: https://arxiv.org/abs/2412.15115

[68] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013/

[69] T. Zhang*, V. Kishore*, F. Wu*, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=SkeHuCVFDr

[70] K. Krippendorff, "Content analysis: An introduction to its methodology," Thousand Oaks, California, Apr 2019. [Online]. Available: https://methods.sagepub.com/book/mono/content-analysis-4e/toc

[71] E. Demerouti, A. B. Bakker, F. Nachreiner, and W. B. Schaufeli, "The job demands-resources model of burnout," *Journal of Applied Psychology*, vol. 86, no. 3, pp. 499–512, 2001.

APPENDIX A SUPPLEMENTARY DATA

In Figure 25, validation labels are shown. We can see the action class names along with every image. And there will some combined images due to mosaic built-in augmentation in YOLO. In Figure 26, the predicted images of validation by the *TeacherActivityNet* models are shown. The bounding box and the class score are also provided for each activity. The class score reflects how reliable the prediction result *TeacherActivityNet* is for each class.

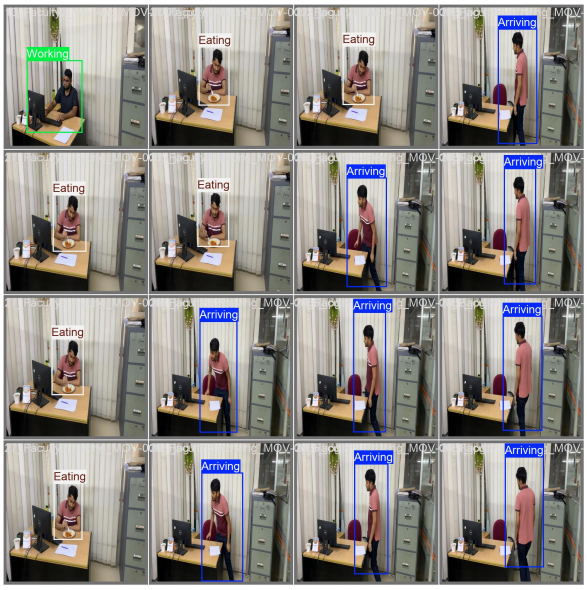


FIGURE 25: Labeled images of the validation dataset (batch-0). Shows the ground truth annotations used for model evaluation.

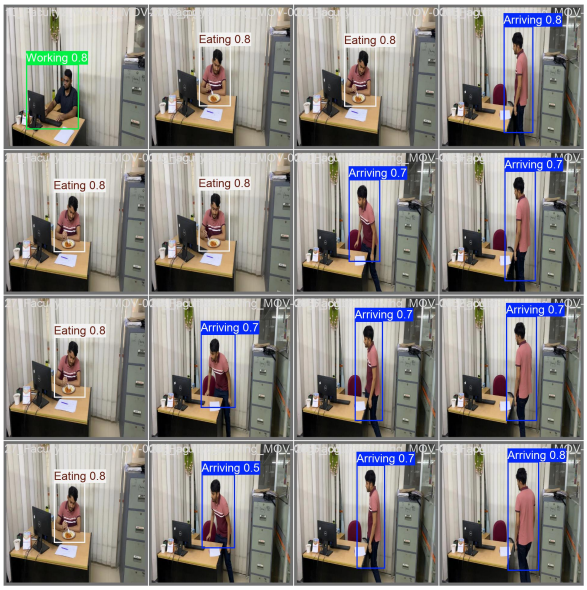


FIGURE 26: Predictions of the validation dataset (batch-0). Highlights how well the model predicts the annotated actions in the dataset.

In Figure 27 and in Figure 28 we can see similar model performance. model predicts each action class perfectly. In Figure 29 and in Figure 30 all the ''Arriving'' actions predicted accurately

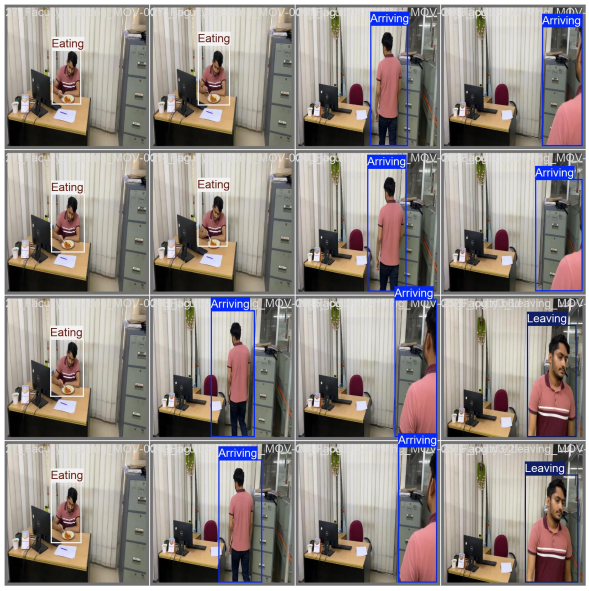


FIGURE 27: Labeled images of the validation dataset (batch-1). Shows the ground truth annotations used for model evaluation.

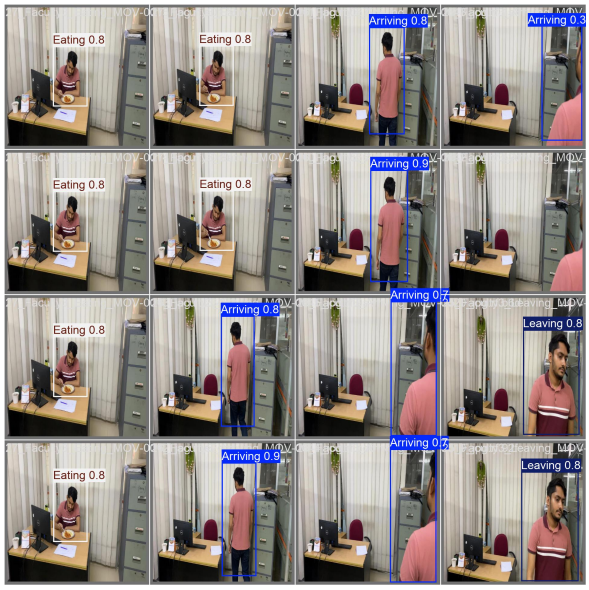


FIGURE 28: Predictions of the validation dataset (batch-1). Highlights how well the model predicts the annotated actions in the dataset.

by our model *TeacherActivityNet* in validation dataset. And the confidence score for each prediction is 0.9.

APPENDIX B FASTER RCNN

In Table 14, the results achieved for the Faster R-CNN ResNet50 FPN model are shown in terms of correctly predicted instances and mAP@50. From the table, we can see that the prediction for the ''Eating'' class is around 0.0382 for this model. However, the prediction rate for the ''Idle'' class is significantly low in our dataset. The ''Counselling'' class shows the highest performance with an mAP@50 of 0.9644. The overall performance of Faster R-CNN ResNet50 is 0.543, which is lower than the overall performance of

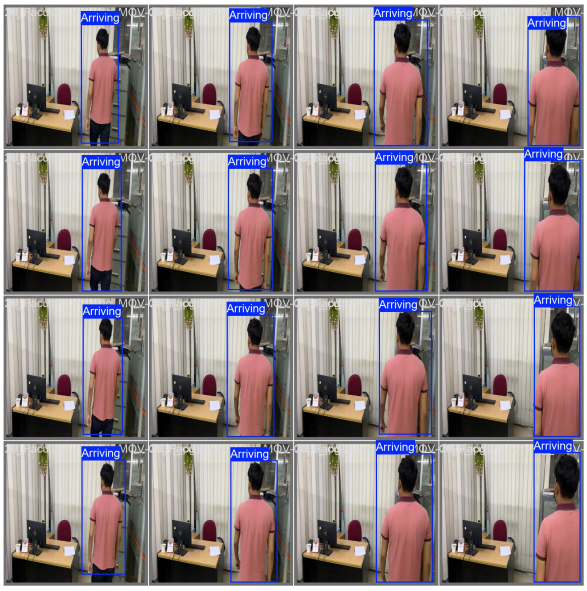


FIGURE 29: Labeled images of the validation dataset (batch-2). Shows the ground truth annotations used for model evaluation.

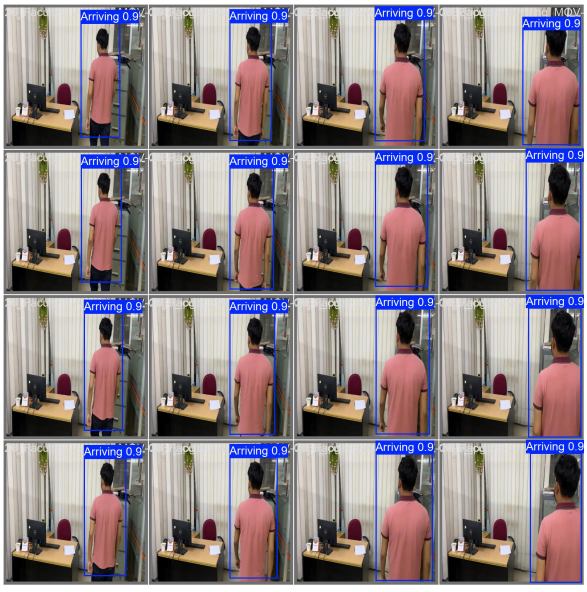


FIGURE 30: Predictions of the validation dataset (batch-2). Highlights how well the model predicts the annotated actions in the dataset.

TeacherActivityNet.

APPENDIX C YOLO MODEL VS *TEACHERACTIVITYNET* DETAILS

First, we show the performance of each action class in the test dataset across different YOLO models.

In Table 15, we can see the performance of the action classes using YOLOv5. In Table 16, the performance is slightly better than YOLOv5. We have used the YOLOv8 model here. In Table 17, we obtained a better result using YOLOv9. Although the "Eating" action class mAP@50 is very poor, the overall performance is good. In Table 18, we got almost the same result as YOLOv9

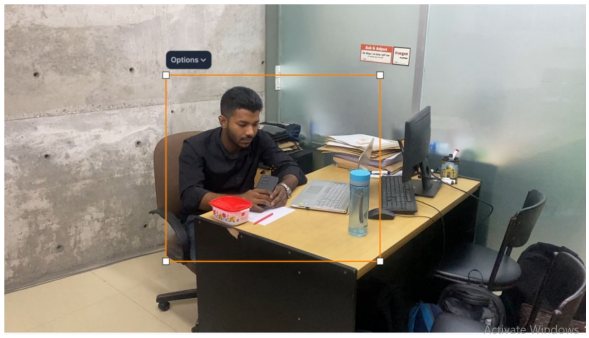


FIGURE 31: Ensuring dataset diversity with multiple nearby objects. The figure highlights how the presence of overlapping or adjacent subjects introduces ambiguity and enriches training examples.

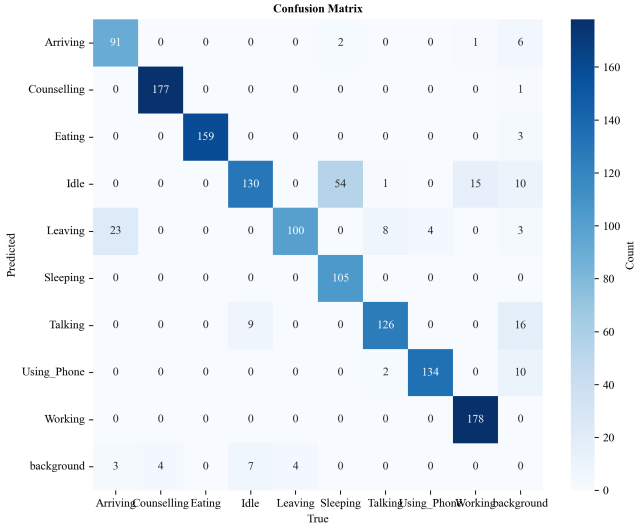


FIGURE 32: Confusion matrix for the validation dataset of *TeacherActivityNet*. This figure highlights the model's classification performance on validation samples.

using the YOLOv10 model. In Table 19, the performance is very good. For the action classes "Arriving" and "Leaving", there are significant gains. In Table 20, using YOLOv12, the performance slightly decreased for many action classes.

APPENDIX D ADAPTIVE WINDOW BASED FOUR SMOOTHING STRATEGIES

In this section we will show the performance of the *TeacherActivityNet* model in four smoothing strategies using adaptive windows for each action classes. Each test sample was evaluated under two augmentation conditions, doubling the instance count from 808 to 1616 while preserving the original class distribution. Table 21 to Table 24 present the mAP@50 results for each action class in different experimental settings. Overall, the detection performance remains high for classes like *Counselling*, *Sleeping*, and *Working*, consistently achieving mAP@50 values above 0.94 across all four datasets.

Some variation is observed in classes such as *Arriving*, *Eating*, *Idle*, and *Leaving*. For example, the *Arriving* class shows a decrease from 0.6984 in Table 21 to 0.6494 in Table 24. Similarly, the *Eating* class consistently remains the lowest across all datasets, ranging from 0.2734 to 0.2838, indicating that this activity is more

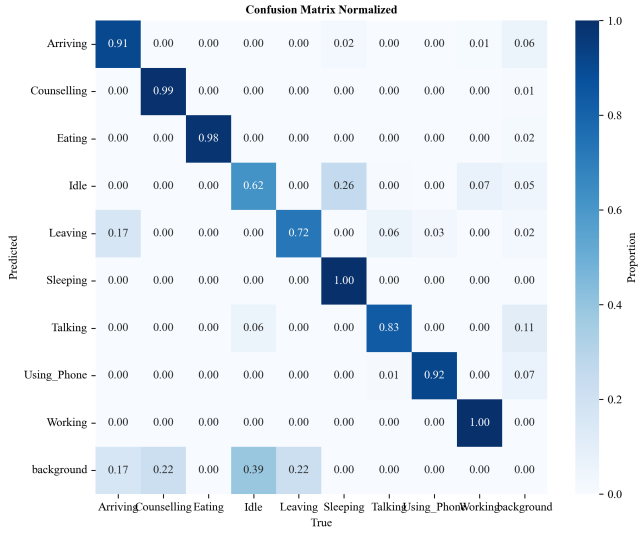


FIGURE 33: Normalized confusion matrix for the validation dataset of *TeacherActivityNet*. This highlights the per-class accuracy in a normalized format.

(a) Original frame. (b) Cropped ROI.

FIGURE 34: Visualizing the *Working* action class: (a) displays the raw frame used for annotation, and (b) shows the extracted region of interest (ROI) optimized for model training.

challenging for the model to detect accurately.

The *Idle* class shows minor fluctuations, slightly increasing from 0.5952 in Table 22 to 0.6123 in Table 24. The *Talking* and *Using_Phone* classes also experience small variations, generally remaining above 0.73 and 0.88, respectively.

Overall, these comparisons indicate that while some classes such as *Eating* and *Arriving* are more difficult for the model, the majority of action classes achieve consistently high mAP@50 values across the four datasets, demonstrating stable detection performance.

APPENDIX E LOW LIGHT PERFORMANCE

This section presents the results for the test and validation datasets after incorporating brightness augmentation. The detailed performance metrics for each action class are summarized in Table 25 and Table 26, respectively.

TABLE 13: **Box precision, recall, mAP@50 and mAP@50-95 results for validation dataset using** *TeacherActivityNet*

Class	Images	Instances	P	R	mAP@50	mAP@50-95
all	1337	1337	0.902	0.894	0.929	0.610
Arriving	117	117	0.947	0.752	0.925	0.654
Counselling	181	181	0.960	0.927	0.954	0.338
Eating	159	159	0.977	1.000	0.995	0.663
Idle	130	130	0.645	0.992	0.868	0.620
Leaving	109	109	0.734	0.908	0.895	0.576
Sleeping	172	172	1.000	0.601	0.764	0.521
Talking	137	137	0.863	0.942	0.974	0.623
Using_Phone	154	154	0.993	0.921	0.992	0.745
Working	178	178	0.997	1.000	0.995	0.746

TABLE 14: **The mAP@50 for test dataset using Faster RCNN ResNet50 FPN**

Action Class	Total	Correct	mAP@50
Arriving	25	15	0.3685
Counselling	183	122	0.9644
Eating	59	0	0.0382
Idle	64	20	0.4518
Leaving	30	23	0.4572
Sleeping	106	25	0.5303
Talking	96	85	0.8073
Using_Phone	119	68	0.6465
Working	126	88	0.6256

TABLE 15: **Box precision, recall, mAP@50 and mAP@50-95 results for test dataset using YOLOv5 for each action class**

Action	Images	Instances	Box (Precision)	Recall	mAP@50	mAP@50-95
all	808	808	0.680	0.706	0.697	0.410
Arriving	25	25	0.384	0.880	0.628	0.397
Counselling	183	183	0.959	0.951	0.984	0.622
Eating	59	59	0.355	0.299	0.331	0.210
Idle	64	64	0.654	0.503	0.537	0.334
Leaving	30	30	0.423	0.633	0.478	0.304
Sleeping	106	106	0.695	0.962	0.810	0.479
Talking	96	96	0.862	0.585	0.772	0.384
Using_Phone	119	119	0.800	0.773	0.814	0.346
Working	126	126	0.990	0.767	0.920	0.615

TABLE 16: **Box precision, recall, mAP@50 and mAP@50-95 results for test dataset using YOLOv8 for each action class**

Action	Images	Instances	Box (Precision)	Recall	mAP@50	mAP@50-95
all	808	808	0.733	0.722	0.720	0.435
Arriving	25	25	0.592	0.720	0.620	0.397
Counselling	183	183	1.000	0.988	0.995	0.645
Eating	59	59	0.399	0.475	0.425	0.289
Idle	64	64	0.512	0.542	0.484	0.275
Leaving	30	30	0.629	0.833	0.690	0.408
Sleeping	106	106	0.915	0.809	0.904	0.549
Talking	96	96	0.808	0.657	0.670	0.341
Using_Phone	119	119	0.769	0.555	0.719	0.365
Working	126	126	0.975	0.919	0.973	0.645

TABLE 17: **Box precision, recall, mAP@50 and mAP@50-95 results for test dataset using YOLOv9 for each action class**

Class	Images	Instances	P	R	mAP@50	mAP@50-95
all	808	808	0.691	0.711	0.733	0.424
Arriving	25	25	0.672	0.840	0.813	0.529
Counselling	183	183	0.986	1.000	0.995	0.601
Eating	59	59	0.253	0.339	0.237	0.124
Idle	64	64	0.685	0.562	0.525	0.265
Leaving	30	30	0.277	0.867	0.676	0.451
Sleeping	106	106	0.876	0.642	0.857	0.482
Talking	96	96	0.772	0.563	0.675	0.350
Using_Phone	119	119	0.713	0.856	0.877	0.370
Working	126	126	0.989	0.733	0.940	0.643

TABLE 18: **Box precision, recall, mAP@50 and mAP@50-95 results for test dataset using YOLOv10 for each action class**

Action	Images	Instances	Box (Precision)	Recall	mAP@50	mAP@50-95
all	808	808	0.728	0.686	0.736	0.437
Arriving	25	25	0.453	0.840	0.729	0.484
Counselling	183	183	0.989	0.955	0.992	0.624
Eating	59	59	0.736	0.189	0.415	0.298
Idle	64	64	0.522	0.500	0.517	0.304
Leaving	30	30	0.422	0.833	0.622	0.385
Sleeping	106	106	0.845	0.774	0.886	0.519
Talking	96	96	0.804	0.438	0.655	0.338
Using_Phone	119	119	0.825	0.849	0.841	0.355
Working	126	126	0.952	0.795	0.971	0.623

TABLE 19: **Box precision, recall, mAP@50 and mAP@50-95 results for test dataset using YOLOv11 for each action class**

Action	Images	Instances	Box (Precision)	Recall	mAP@50	mAP@50-95
all	808	808	0.688	0.729	0.743	0.443
Arriving	25	25	0.609	0.880	0.833	0.564
Counselling	183	183	0.984	0.990	0.994	0.645
Eating	59	59	0.320	0.390	0.340	0.231
Idle	64	64	0.510	0.537	0.488	0.257
Leaving	30	30	0.308	0.833	0.759	0.489
Sleeping	106	106	0.817	0.674	0.816	0.467
Talking	96	96	0.898	0.490	0.689	0.319
Using_Phone	119	119	0.763	0.832	0.781	0.348
Working	126	126	0.983	0.933	0.990	0.664

TABLE 20: **Box precision, recall, mAP@50 and mAP@50-95 results for test dataset using YOLOv12 for each action class**

Action	Images	Instances	Box (Precision)	Recall	mAP@50	mAP@50-95
all	808	808	0.677	0.691	0.715	0.422
Arriving	25	25	0.596	0.709	0.705	0.455
Counselling	183	183	0.950	0.973	0.989	0.620
Eating	59	59	0.396	0.525	0.368	0.248
Idle	64	64	0.464	0.234	0.375	0.212
Leaving	30	30	0.466	0.867	0.715	0.486
Sleeping	106	106	0.680	0.896	0.841	0.483
Talking	96	96	0.766	0.648	0.747	0.357
Using_Phone	119	119	0.808	0.565	0.754	0.362
Working	126	126	0.971	0.804	0.940	0.576

TABLE 21: **Adaptive Majority Voting: mAP@50 results and window sizes for each action class**

Class	Images	Instances	mAP@50	Window
Arriving	50	50	0.6984	3
Counselling	366	366	0.9617	15
Eating	118	118	0.2838	5
Idle	128	128	0.5908	8
Leaving	60	60	0.8051	3
Sleeping	212	212	0.9570	20
Talking	192	192	0.7364	7
Using_Phone	238	238	0.8987	10
Working	252	252	0.9424	12

TABLE 22: **Confidence Weighted Voting: mAP@50 results and window sizes for each action class**

Class	Images	Instances	mAP@50	Window
Arriving	50	50	0.7521	3
Counselling	366	366	0.9617	15
Eating	118	118	0.2734	5
Idle	128	128	0.5952	8
Leaving	60	60	0.8175	3
Sleeping	212	212	0.9570	20
Talking	192	192	0.7548	7
Using_Phone	238	238	0.9063	10
Working	252	252	0.9557	12

TABLE 23: **Weighted Averaging with Recency Bias: mAP@50 results and window sizes for each action class**

Class	Images	Instances	mAP@50	Window
Arriving	-	50	0.6919	3
Counselling	-	366	0.9617	15
Eating	-	118	0.2821	5
Idle	-	128	0.5936	8
Leaving	-	60	0.8228	3
Sleeping	-	212	0.9570	20
Talking	-	192	0.7402	7
Using_Phone	-	238	0.9024	10
Working	-	252	0.9444	12

TABLE 24: **Adaptive Hybrid Ensemble: mAP@50 results and window sizes for each action class**

Class	Images	Instances	mAP@50	Window
Arriving	-	50	0.6494	3
Counselling	-	366	0.9617	15
Eating	-	118	0.2802	5
Idle	-	128	0.6123	8
Leaving	-	60	0.7490	3
Sleeping	-	212	0.9570	20
Talking	-	192	0.7416	7
Using_Phone	-	238	0.8801	10
Working	-	252	0.9390	12

TABLE 25: **Box precision, recall, mAP@50 and mAP@50-95 results for the brightness-augmented test dataset using** *TeacherActivityNet*

Action	Images	Instances	Box (Precision)	Recall	mAP@50	mAP@50-95
all	808	808	0.750	0.769	0.796	0.492
Arriving	25	25	0.637	0.960	0.869	0.601
Counselling	183	183	0.977	0.913	0.988	0.641
Eating	59	59	0.537	0.356	0.448	0.275
Idle	64	64	0.445	0.464	0.434	0.257
Leaving	30	30	0.632	0.867	0.835	0.576
Sleeping	106	106	0.810	0.846	0.886	0.520
Talking	96	96	0.867	0.719	0.846	0.457
Using_Phone	119	119	0.891	0.897	0.914	0.461
Working	126	126	0.958	0.895	0.940	0.641

TABLE 26: **Box precision, recall, mAP@50 and mAP@50-95 results for the brightness-augmented validation dataset using** *TeacherActivityNet*

Action	Images	Instances	Box (Precision)	Recall	mAP@50	mAP@50-95
all	1337	1337	0.925	0.878	0.928	0.621
Arriving	117	117	0.915	0.940	0.972	0.652
Counselling	181	181	0.971	0.972	0.978	0.480
Eating	159	159	0.999	1.000	0.995	0.741
Idle	130	130	0.611	0.992	0.708	0.474
Leaving	109	109	0.972	0.771	0.971	0.651
Sleeping	172	172	1.000	0.380	0.776	0.462
Talking	137	137	0.856	0.942	0.967	0.630
Using_Phone	154	154	1.000	0.903	0.990	0.746
Working	178	178	0.998	1.000	0.995	0.748

SHOIB AHMED SHOURAV is an Assistant Professor in the Department of Computer Science and Engineering at United International University (UIU), Bangladesh. He completed both his Bachelor of Science and Master of Science in Computer Science and Engineering from United International University, where he developed a strong foundation in theoretical and applied computing. With a dedicated focus on advancing the field of computing, his research interests span a diverse range of areas, including the application of Computer Science in Education, Algorithm Design and Analysis, and Computer Vision. He is particularly passionate about leveraging intelligent systems and computational methodologies to address real-world challenges and enhance learning experiences in academic environments.

NAHID HOSSAIN is an Assistant Professor in the Department of Computer Science and Engineering (CSE) and the Undergraduate Program Coordinator at United International University (UIU), Dhaka, Bangladesh. His research centers on Natural Language Processing (NLP) and Large Language Models (LLMs), with a focus on language understanding and generation, language technology and resource development, and model efficiency and optimization. His research contributions span transformer-based grammatical and spelling error correction, dialect conversion systems, and text simplification for low-resource languages, alongside foundational resource development such as corpora and evaluation benchmarks. His recent work investigates context minimization strategies to improve the resource efficiency of LLMs in constrained computational environments. Prior to his current role, he served as a Lecturer in the same department. Before entering academia, he worked as a Software Engineer in the AI Division at eGeneration Ltd., Dhaka, Bangladesh.

SWAKKHAR SHATABDA is a distinguished academic and researcher currently serving in the Department of Computer Science and Engineering at BRAC University. He obtained his Ph.D. from Griffith University, Australia, where he worked at the Queensland Research Laboratory, NICTA, and completed his B.Sc. in Computer Science and Engineering from the Bangladesh University of Engineering and Technology (BUET). His research expertise spans artificial intelligence search, optimization, applied machine learning, algorithms, and computational biology. Professor Shatabda has an extensive publication record in leading international journals and conferences and serves as an Academic Editor for PLOS ONE. Before joining BRAC University, he held several key academic and administrative positions at United International University (UIU), including Professor, Director of the B.Sc. in Data Science program, Director of IQAC, and Undergraduate Program Coordinator in the Department of Computer Science and Engineering. In addition to his academic contributions, he works as an AI and ML Consultant at Nodes Digital Limited and serves as the Director of Quality Assurance at the Board of Accreditation for Engineering and Technical Education (BAETE) under The Institution of Engineers, Bangladesh (IEB).