An End-to-End Deep Learning Pipeline for Child Detection and Activity Recognition in Surveillance Systems

Samad Riaz[1], Shayan Riaz[2], Abdul Razzaq[2], Umar Sadique[3], Shahid Bashir[4], Jehad Ur Rahman[4]
[1]Department of Electrical Engineering, UET Peshawar, Pakistan.
[2]Center for Intelligent Systems and Network Research, UET Peshawar, Pakistan.
[3]Department: Computer Systems Engineering UEAS Swat, Pakistan.
[4]Department of Electrical Engineering, UET Peshawar, Pakistan.
*Correspondence: samadriaz@uetpeshawar.edu.pk

Citation | Riaz. S, Riaz. S, Razzaq. A, Sadique. U, Bashir. S, Rahman. J. U, "An End-to-End Deep Learning Pipeline for Child Detection and Activity Recognition in Surveillance Systems", IJIST, Vol. 8 Issue. 3 pp 1029-1048, June 2026
Received| April 11, 2026 **Revised|** May 14, 2026 **Accepted|** May 21, 2026 **Published|** June 03, 2026.

Children in developing countries often face significant safety risks such as domestic accidents, unsafe environments, and abusive social exposure. The current surveillance systems are able to identify children but, in most cases, they fail to comprehend their activities over time. This creates a critical gap in real-time child safety monitoring. This paper addresses that gap by developing an end-to-end video understanding pipeline for child detection in video frames and recognizes their activities accurately over time. A custom child detection dataset containing 19,890 annotated images was developed and split into 16,107 training and 3,783 validation samples. Several state-of-the-art object detection models, including YOLOv8s, YOLO26s, and RT-DETR-L, were trained and compared based on their performance metrics. DeepSORT is used to preserve the identities of children detected across frames. For activity recognition, a custom video dataset comprising approximately 1,960 clips across 47 activity classes was created. Four deep learning architectures (I3D, ResNet3D, ViViT, and VideoMAE) were trained using standardized configurations with 16-frame clips, 224×224 input resolution, Adam optimization, and cross-entropy loss. YOLO26s achieved the best detection performance with Precision = 97.17%, Recall = 96.18%, mAP@0.50 = 98.51%, and mAP@0.50–0.95 = 87.80%. For activity recognition, VideoMAE outperformed competing models on 448 clips with test accuracy of 86.38% and Macro F1-score of 90.49%, significantly surpassing ViViT (79.24%), ResNet3D (51.34%), and I3D (9.60%). The proposed method provides a scalable and practical solution by combining high-quality spatial localization with temporal activity understanding, enabling real-time monitoring of child activities.

Keywords: Child Safety, Activity Recognition, Deep Learning, Computer Vision, Video Surveillance

Introduction:

Children are one of the most vulnerable groups in society, especially in developing nations where they are highly exposed to a number of risks which hinder healthy development. The recent report by UNICEF [1] revealed that every four minutes a child in the world is killed by an act of violence, which is why child protection mechanisms must be improved. Children in developing nations such as Pakistan are faced with complex problems such as physical, emotional, and sexual abuse, exposure to dangerous conditions, and engaging in unlawful acts such as begging at road crossings and engaging in fights with other children [2].

The development of artificial intelligence and computer vision technologies has provided the new opportunities of automated surveillance and monitoring systems. Although traditional surveillance systems can capture video footage, they are not capable of analyzing and interpret the behaviors of children in real-time, and thus it becomes hard to identify dangerous situations at an early stage [3]. Recent deep learning models, especially convolutional neural networks (CNNs) and vision transformers, have proven to be extremely effective in object detection, tracking, and activity recognition tasks [4].

The recent studies have been conducted on the use of computer vision in monitoring child safety. A computer vision system was developed by [4] as part of the CCTV surveillance system to identify domestic accidents among children with detection accuracies of 87% percent of hazards such as scissors and knives. The system measures the distance of children to hazardous objects to assess risk, and raise alarms when a child is near dangerous tools. In a similar fashion, [5] have employed CNNs and fourier and Radon transforms to detect anomalies by identifying motion in video frames and VGG-Net transfer learning to analyze fight scenes. Transformer-based architectures, especially Vision Transformers (ViT) and Video Vision Transformers (ViViT), have transformed video understanding by effectively learning spatial and temporal relationships in video sequences [5]. With masked autoencoding pretraining schemes, these architectures have demonstrated better video classification performance than traditional CNN-based methods.

Although recent research has made significant progress in child safety systems based on computer vision, the current research is mainly confined to individual tasks like child detection, hazard detection, or limited activity detection under limited conditions. Most existing systems do not integrate detection, tracking, and comprehensive activity understanding, which decreases their efficiency in the context of real-world surveillance. Moreover, some of the current methods are based on general datasets or small sets of behavior, which limit their applicability to child-specific safety monitoring. The lack of large-scale child-centric datasets and end-to-end frameworks that can simultaneously detect, track, and recognize the various activities of children underscores a major gap in the current state of surveillance research.

In this paper, a deep learning-based child safety monitoring system (which includes child detection, multi-object tracking, and activity recognition) is proposed as an end-to-end system. To help with this task, a custom dataset of various child-centered activities was created. The system is a combination of a YOLO-based detector to provide accurate spatial localization, DeepSORT to provide consistent tracking, and VideoMAE to provide robust temporal activity recognition. The experimental results show that the proposed approach achieves high detection accuracy and better activity recognition performance than state-of-the-art methods. The system allows real-time tracking with normal surveillance infrastructure and is a scalable and practical solution to improving child safety.

Research Objectives:

The primary objective of this study is to come up with an end-to-end deep learning system to support real-time child safety monitoring by accurate child detection, tracking, and activity recognition. The research focuses on creating child-specific datasets to be used in

detection and video classification, evaluate several state-of-the-art detection and video classification models, and identify the most effective architectures for surveillance applications. It is also aimed at incorporating powerful tracking systems to ensure time consistency and offer a scalable, practical solution to implement in homes, schools, daycare centers, and in public safety settings.

Novelty:

The novelty of this study is that it develops a fully operational end-to-end child safety surveillance pipeline, which includes child detection, multi-object tracking, and fine-grained activity recognition in a single framework. The proposed system is a combination of precise spatial localization and temporal activity knowledge in 47 different child-specific activities, unlike the existing surveillance systems that are mostly focused on individual tasks such as child detection, proximity to hazards estimation or limited behavior recognition. Also, this study presents domain-specific datasets of both child detection and activity recognition, allowing more realistic testing in child-centric settings. The comparative analysis of the different state-of-the-art object detection and video understanding architectures further contributes to the practical relevance of the work to help identify deployment-efficient solutions to real-world surveillance systems. This integrated solution provides a scalable, reproducible, and application-oriented framework that will take child safety monitoring to a new level beyond the existing fragmented solutions.

Literature Review:

Recent research has explored the use of computer vision and deep learning for analyzing child behavior and improving safety in surveillance environments. Several datasets and models have been proposed to understand specific aspects of child activities.

A benchmark dataset was recently developed by [6] called "Child Aggression Behavior Analysis Dataset" (CABAD). This dataset is designed to recognize the aggressive behaviors of children using deep learning techniques. The dataset includes annotated videos of six aggressive child behaviors including Hitting, throwing objects, kicking surfaces, breaking objects, slamming doors, and pushing collected from different opensource platforms.

This dataset [7] is the movement of an infant that was recorded at neonatal intensive care units. The dataset comprised of 16 depth videos recorded during clinical practice. Each video has 1000 frames and 100s with accurate annotation of the location of the limb-joints, left and right shoulder, elbow, wrist, hip, knee and ankle depicted. The vision system that was developed in the study by [8] detects and classifies autism-related behaviors including flapping of arms, headbanging, and spinning behaviors observed in videos recorded in natural settings. The system obtained a weighted F1-score of 0.83 using Inflated 3D ConvNets (I3D) and Multi-Stage Temporal Convolutional Networks (MS-TCN). This shows that AI-based tools have a great potential to assist clinicians in early ASD assessment and behavioral monitoring.

Monitoring toddlers in the indoor environment presents special difficulties because of their erratic movements, similarity, and obstructions. [9] proposed a genetic algorithm-optimized multi-object tracking algorithm called MTTSort, which is a variation of DeepSort. Another finding of the study was a new dataset, called MTTrack, which included annotated indoor video of toddlers aged 2-4 years. MTTSort uses Vision Transformer (ViT)-based attention and feature aggregation in a pool to enhance identity consistency. The model demonstrated better performance on the MTTrack with MOTA 0.98, HOTA 0.68, and IDF1 0.98, which is better than existing MOT techniques in indoor settings. The proposed article [10] introduced a video activity recognition framework based on explainable video activity recognition that integrates the deep learning approach with a tractable probabilistic model to improve the transparency and reliability of the model in decision-making. Their two-layer architecture combines a deep neural network to classify videos with a probabilistic reasoning

layer that produces human-readable explanations. The model was able to enhance recognition and interpretability of various datasets.

The article [5] describes an AI-based daycare safety system that combines IoT sensors, edge-based computer vision, emotion detection, cry detection, and fall detection. Previous research findings indicate that the YOLO-based fall detector, deep-learning emotion recognizer, and cry-analysis models are effective at their own level but remain challenged by noise, limited data, and false alarms. Based on this literature, the proposed system will integrate these methods into a single, real-time, privacy-sensitive platform, and attain high accuracy in all modules. The article [3] suggests a computer-vision algorithm that tracks objects by using YOLOv8 and distance measurements to identify knives, scissors, and estimate distances, thus determining the risk of accidents to toddlers.

Recent advancements in child safety surveillance have increasingly shifted from traditional GPS- and cloud-based tracking systems toward intelligent computer vision-based frameworks that offer real-time behavioral monitoring. [11] proposed an IoT-integrated child detection and counting system using YOLOv8 and DeepSORT, achieving high detection accuracy while enabling scalable smart-city surveillance applications. Similarly, [12] introduced an enhanced YOLOv11n-based child detection framework specifically optimized for noisy surveillance environments, demonstrating strong detection robustness under challenging real-world CCTV conditions.

A comprehensive review by [13] investigates the technologies applied in Edge AI-assisted video analytics in smart cities. This review encompasses the different artificial intelligence models and privacy-saving methods in edge video analytics. This work [14] suggests a vision system based on YOLOv5 that is installed in a smart home to identify the presence of infants and the presence of hazards (e.g. fire, pools) in the environment and notify the caregiver when the child enters a dangerous area. This paper [15] introduces a CNN-based approach to classify simple infant behaviors (crawling, running, sleeping, walking) using video frames. The model was trained on a custom dataset (155 videos, 360 x360 resolution, 8 fps) of children performing these four actions, and it has an overall accuracy of 94.73% on held-out test data.

This work [11] is aimed at detecting and counting children in city video streams. The authors suggest an IoT-vision system based on YOLOv8 (including instance segmentation) to identify children in surveillance images and a DeepSORT tracker to estimate the number of children in crowds. The system was reported to achieve high performance (up to 98% accuracy in detecting / classifying children) and perform better than previous YOLOv5 and YOLOv4-based models in the same task.

This research Introduces [16] ChildACT, a multi-view video dataset of 200 children performing 7 action classes (walking, jumping, etc.), filling a gap in child-specific action-recognition benchmarks. [17] Proposes a two-stage pipeline using enhanced YOLOv8 to detect toddlers and analyze head-body features for fall detection, achieving 96.33% accuracy in toddler fall alerts. CNN-based real-time system [18] (using pretrained AgeNet/GenderNet) that labels people aged 0–12 as children in surveillance video, achieving ~84.7% child-detection sensitivity under varied conditions.

This work Presents [19] ByteTrack, a state-of-the-art tracker that matches all detections (including low-confidence) to tracklets, substantially improving MOTA and reducing ID switches (e.g. +71% MOTA on BDD100K). An end-to-end vision pipeline presented in [20] for detecting infant "non-nutritive sucking" (NNS) events in long videos,

using spatiotemporal CNNs and infant-specific pose features. It achieves ~94.0% AP and 84.9% recall on annotated infant video clips. [21] Introduces **EduNet**, a large classroom video dataset (7,851 clips, 20 teacher/student actions) for action recognition in school settings. Each class (e.g. "Explaining," "Writing on Board") has ≥200 clips, supporting education-domain surveillance research. In this research [22] Releases **ICCWD**, a 10,000-image+caption dataset labeled for child presence (including partially visible or fictional children). ICCWD benchmarking demonstrates even the best detectors only ≈75.3% true-positive rate, indicating the difficulty of automatic child detection.

This paper [23] suggests a multi-agent system based on Raspberry-Pi and combining the abilities of the vision-language reasoning and the YOLOv4 detection. The system recognizes interactions between children and sends real-time abduction alerts through Twilio; reported to be highly accurate on abduction cases. This paper [24] outlines an Edge-AI system of real-time action recognition on ARM processors. Using pruning and quantization, it achieves 87.3% HAR accuracy, with only around 2.1W of power consumption, which is significantly lower than the power consumption of cloud-based baselines in both latency and energy consumption. In video-based action recognition (20202022), surveys recent DNN architectures [25] with emphasis on the shift to spatiotemporal networks to surveillance. This review compares the current datasets and reports that deep spatiotemporal networks have become the most dominant in performance. A comparative summary of existing child safety surveillance systems, including their methodologies, strengths, and limitations, is presented in Table 1.

Table 1. Comparative Analysis of Existing Child Safety Surveillance Systems

Ref	Study / System	Methodology	Dataset / Scope	Strengths	Limitations
[6]	CABAD Dataset + CABA_Net	MobileViT + TCN + Attention LSTM	900 videos, 6 aggression classes	High aggression recognition accuracy (89%), low computational cost	Limited to aggressive indoor behaviors only
[7]	babyPose Dataset	Depth-based infant pose estimation	16 NICU depth videos	Detailed infant limb-joint annotations	Small dataset, no RGB/multimodal support
[8]	Autism Behavior Recognition	I3D + MS-TCN	Natural autism behavior videos	Effective ASD behavior classification (F1 = 0.83)	Limited to autism-specific behaviors
[9]	MTTSort + MTTrack	DeepSORT + ViT + Genetic Optimization	Indoor toddler tracking dataset	Excellent toddler tracking (MOTA 0.98)	Focused only on tracking, no behavior recognition
[10]	Explainable Activity Recognition	Deep learning + Probabilistic reasoning	Generic activity datasets	Improved interpretability	Limited child-specific applications
[5]	Intelligent Daycare System	IoT + CV + Cry/Fall/Emotion Detection	Daycare safety	Multi-module safety monitoring	Integration complexity, noise sensitivity

[3]	Toddler Hazard Detection	YOLOv8 + Distance estimation	Accident prevention scenarios	Hazard proximity monitoring	Limited to object-danger scenarios
[11]	YOLOv8 + IoT Child Detection	YOLOv8 + DeepSORT	Smart city child surveillance	High detection accuracy, scalable	Primarily detection/counting
[12]	YOLOv11n Child Detection	Enhanced YOLOv11n	Noisy CCTV footage	Robust surveillance performance	Focused only on detection
[14]	Smart Home Monitoring	YOLOv5 hazard surveillance	Home safety dataset	Hazard detection + caregiver alerts	Limited behavioral depth
[15]	Child Activity Recognition	CNN	155 videos, 4 actions	High simple activity accuracy	Small dataset, limited activities
[16]	ChildACT Dataset	Child-specific action recognition	200 children, 7 actions	Dedicated child action benchmark	Limited action diversity
[17]	Toddler Fall Detection	Improved YOLOv8	Toddler fall videos	High fall detection accuracy (96.3%)	Single safety event focus
[18]	Child Presence Detection	CNN + AgeNet/GenderNet	Real-time child presence	Strong child detection sensitivity	Limited age estimation reliability
[19]	ByteTrack	Multi-object tracking	General MOT datasets	Superior tracking consistency	Not child-specific
[20]	Infant NNS Recognition	Spatiotemporal CNN + Pose	Infant crib videos	End-to-end infant action analysis	Narrow clinical application
[22]	ICCWD Dataset	Child detection benchmark	10,000 child images	Broad child detection benchmark	Detection challenge remains high
[23]	Child Abduction Detection	YOLOv4 + VLM + Edge AI	Abduction scenarios	Real-time safety alerts	Specialized use case
[24]	Edge-AI HAR	Quantized deep learning	Low-power edge systems	Real-time deployment efficiency	Generic HAR, not child-specific

The comparison indicates the differences in the scope of tasks, model architecture, complexity of the dataset, performance measures and practical constraints. As shown, despite some of the past studies being effective in specialized fields such as aggression detection, autism behavior analysis, or limited activity recognition, most of the past studies are only effective in limited areas of behavior or single surveillance activities. By contrast, the proposed system exhibits competitive performance and tackles a much broader real-world challenge by providing unified child detection, tracking, and fine-grained activity recognition across 47 diverse activity classes. This broader applicability, and the possibility of being scaled up, makes the proposed framework a more holistic child safety surveillance system. The benchmark comparison between the proposed framework and existing state-of-the-art child detection and activity recognition systems is shown in Table 2.

Table 2. Benchmark Comparison of the Proposed Framework with Existing State-of-the-Art Child Detection and Activity Recognition Systems

Study	Task	Model	Performance	Limitation
[6]	Aggression recognition	CABA_Net	89% Accuracy	Only 6 behaviors
Autism Behavior [8]	ASD behavior	I3D + MS-TCN	F1 = 0.83	Autism-specific
Child Activity Recognition [15]	Simple actions	CNN	94.73% Accuracy	Only 4 classes
Proposed System	47-class child activity recognition	VideoMAE	86.38% Accuracy, F1 = 90.49%	Larger real-world scope

Although there has been significant advancement in child safety surveillance, the current studies focus on individual aspects of child safety surveillance like child detection, toddler tracking, hazard recognition, or narrow behavior classification. The majority of systems do not provide integrated of spatial localization, temporal identity preservation, and recognition of fine-grained activities within a single framework. Also, most of the previous studies are based on limited datasets or behavioral categories, which restricts their application to real-world scenarios. Thus, the need for a scalable, end-to-end child surveillance solution which can effectively monitor in real time all the diverse safety conditions and this study seeks to provide this solution.

The analyzed literature sources have greatly contributed to the development of the proposed surveillance framework by highlighting the progress and shortcomings in the fields of child detection, tracking, and activity recognition. The recent YOLO-based child detection systems, hazard monitoring frameworks and smart home surveillance solutions demonstrated the effectiveness of deep learning to detect children accurately. Likewise, tracking systems like MTTSort and ByteTrack highlighted the need to maintain child identity consistency in dynamic environments, which motivated the inclusion of DeepSORT in this work.

Moreover, the studies of child activity recognition such as CABAD, autism behavior analysis, and ChildACT highlighted the necessity to have a strong temporal understanding of child behaviors. All these studies together have provided the necessary methodological backgrounds and have also revealed the lack of a single unified end-to-end framework that integrates child detection, tracking, and comprehensive activity recognition, which is the aim of this study.

Methodology:

The system proposed is a two-stage pipeline (child detection and activity recognition). In the first stage, state-of-the-art YOLO-based object detection models are employed to accurately detect and localize children in video frames. These models are designed to achieve real-time performance and high detection accuracy across different environmental conditions. The second stage involves processing the identified child regions with video classification models that identify activities over time. The complete operational workflow of the proposed framework is illustrated in Figure 1, which presents the sequential stages from input video acquisition to final alert generation.

As shown in Figure 1, the pipeline starts with the input surveillance video that can be captured by cameras, smart home systems, daycare centers or other public surveillance sources. The first step involves frame extraction of the input video where continuous video streams are separated into sequential image frames for further spatial analysis.

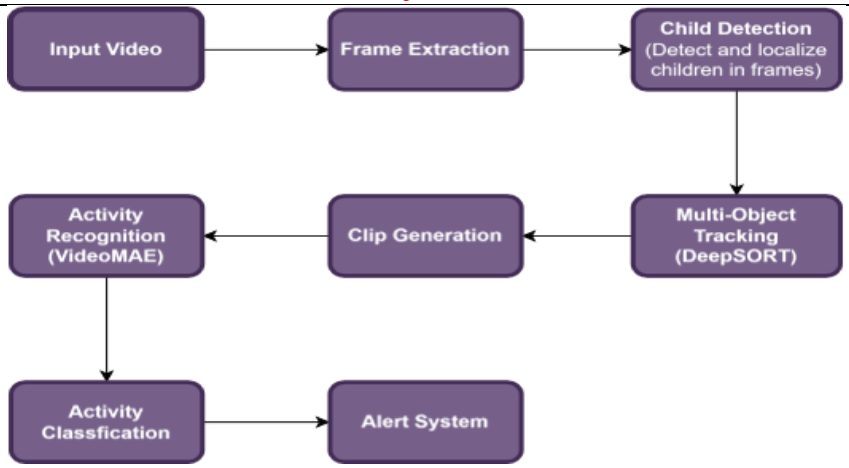

Figure 1. Flowchart of the proposed child detection and activity recognition pipeline.

Afterward, all extracted frames are put through state-of-the-art YOLO-based child detection models, which effectively detect and localize children in individual frames by producing bounding boxes around child instances. These object detection models are chosen particularly due to their real-time operation, high accuracy, and robustness under different environmental conditions.

After detection, the system uses DeepSORT-based multi-object tracking as shown in Figure 1 to ensure identity consistency of each detected child across successive frames. This tracking step makes sure that the children are constantly tracked over time even in situations involving movement, temporary occlusion or dynamic changes in the scene.

Once stable tracking has been accomplished, clip generation is done by bundling sequentially tracked child frames into fixed-length video clips. These clips maintain temporal continuity and offer coherent behavioral sequences which can be used for downstream activity recognition.

To analyze and classify activities, four state-of-the-art deep learning models were used that represent four different ways of learning spatio-temporal features. Such models are the Inflated 3D Convolutional Network (I3D), 3D Residual Networks (3D ResNet), Video Masked Autoencoders (VideoMAE)) [26], and the Video Vision Transformer (ViViT) [27]. All these models represent convolutional and transformer-based video-understanding paradigms. I3D model is an extension of 2D CNN models into the time dimension, making it possible to learn on RGB and optical flow streams. The 3D ResNet model implements residual learning to deep 3D convolutional networks that make it possible to effectively train spatio-temporal representations. VideoMAE and ViViT on the other hand, are recent transformer-based approaches of transformer-based methods, with VideoMAE using self-supervised masked reconstruction to learn representations and ViViT using pure transformer networks to classify videos. The research examines different approaches for modeling motion, appearance and temporal relations in video data through these varied architectures.

Overall, Figure 1 demonstrates how the proposed framework integrates child detection, tracking, clip generation, activity recognition, classification, and alert generation into a unified child safety surveillance system capable of robust real-time deployment.

Dataset Composition:

The data set has two main parts i.e. child detection images and activity classification videos. The child-detection dataset was formatted in YOLO format where it contained 19,890 images divided into 16,107 training and 3,783 validation samples. The dataset contains 17,821

child instances. Although the dataset has 17821 instances of children. Figure 2 shows the statistical distribution of the child detection dataset. In particular, Figure 2 shows: (a) the overall frequency of classes of child instances, (b) the spatial distribution of annotated bounding boxes, (c) the normalized center-point density of child locations within frames, and (d) the width-height distribution of bounding boxes. Analysis shows that child instances are highly concentrated within image frames, and the sizes of bounding boxes differ significantly due to variations in child pose, camera angle, and the context of the environment. This diversity enhances model robustness by subjecting detection architectures to diverse real-world surveillance scenarios.

Figure 3 was created with the LabelImg annotation tool, where each child instance in an image was manually annotated with exact spatial coordinates. In order to enhance the reliability of annotation, several rounds of manual verification were conducted. Primary annotations were done by trained annotators followed by secondary cross-validation to identify labeling inconsistencies, missing detections or inaccurate bounding boundaries. The quality of annotation was achieved by repeatedly reviewing the data and randomly checking the samples to ensure the data are uniform throughout the dataset.

To create the child activity recognition component, a custom video dataset containing about 1,960 clips across 47 child-specific activity classes was created using real-world data collection methods on mobile devices and surveillance cameras. The dataset covers indoor and outdoor settings, a variety of child behaviors at homes, playgrounds, daycare settings, and in the public. Videos mainly depict children between infancy and about six years of age, covering a wide range of developmental stages and behaviors.

The dataset was specifically designed to include a broad range of child activities, including normal daily activities, social interactions, play activities, movement patterns, and safety-critical events such as falls, hazardous interactions, or abnormal behaviors. This variety increases the realism and practical applicability of the dataset to surveillance-based child safety applications.

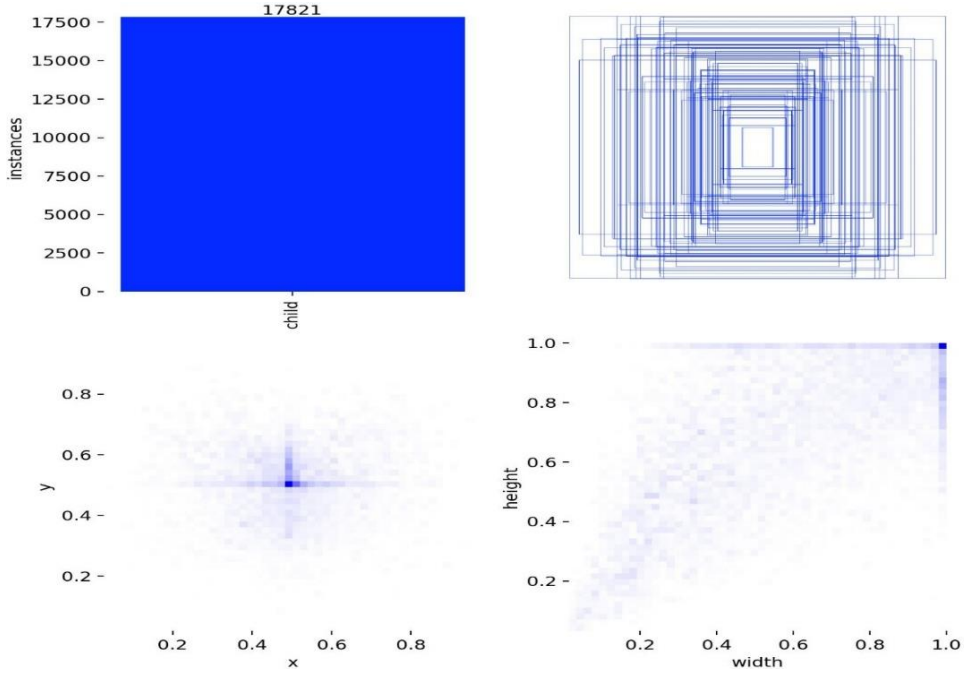

Figure 2. Detection Dataset Distribution

All the video samples were manually checked and categorized into predefined behavioral classes based on observable temporal behaviors. The definitions of standardized activities were created before annotation to guarantee uniformity in all classes. Several evaluators were involved in the annotation process, and ambiguous samples were annotated twice and validated by consensus to reduce labeling bias and enhance the reliability of the dataset.

To support robust model evaluation, the dataset was partitioned using a class-balanced strategy to preserve representative behavioral distributions across training, validation, and testing subsets.

The current study employs a detailed video dataset that is used to identify child activities, including 47 different classes of activities and about 1,960 video clips. The data is inclusive of a wide variety of child behaviors and activities including daily activities such as eating, drinking, sleeping and playing with toys and even more specific such as stair climbing, children fighting and interaction with guardians. The data has been logically arranged into folders of activities, and each folder has video clips of one activity type as represented in Figure 4.

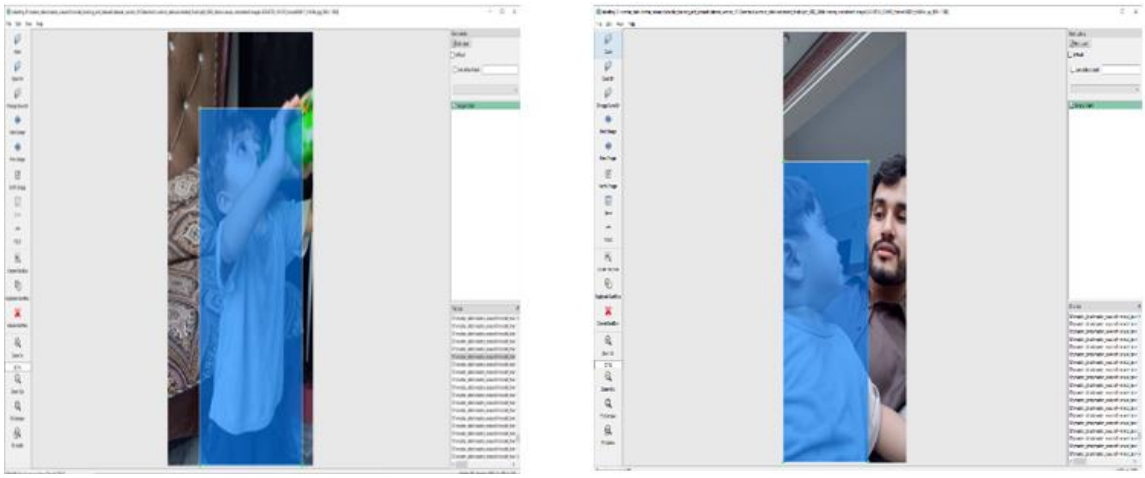

Figure 3. Sample Annotated Image using LabelImg tool for data annotation

Figure 4 provides a comprehensive statistical overview of the activity recognition dataset. The upper section illustrates the full video count distribution across all 47 activity categories, demonstrating significant behavioral diversity within the dataset. Frequently represented activities include normal activities, guardian-assisted behaviors, playing with toys, walking/running, and eating, while less frequent categories capture rarer but safety-relevant behaviors such as accidents, hazardous interactions, and unusual child activities.

The lower section of Figure 4 further summarizes dataset characteristics through multiple statistical perspectives, including: (a) the top 20 most represented activity classes, (b) histogram-based distribution of video frequency per class, (c) cumulative activity coverage, and (d) dataset summary statistics. The dataset exhibits moderate class imbalance, with an average of approximately 41.7 videos per class and a standard deviation of 50.9, reflecting realistic real-world behavioral frequency variations.

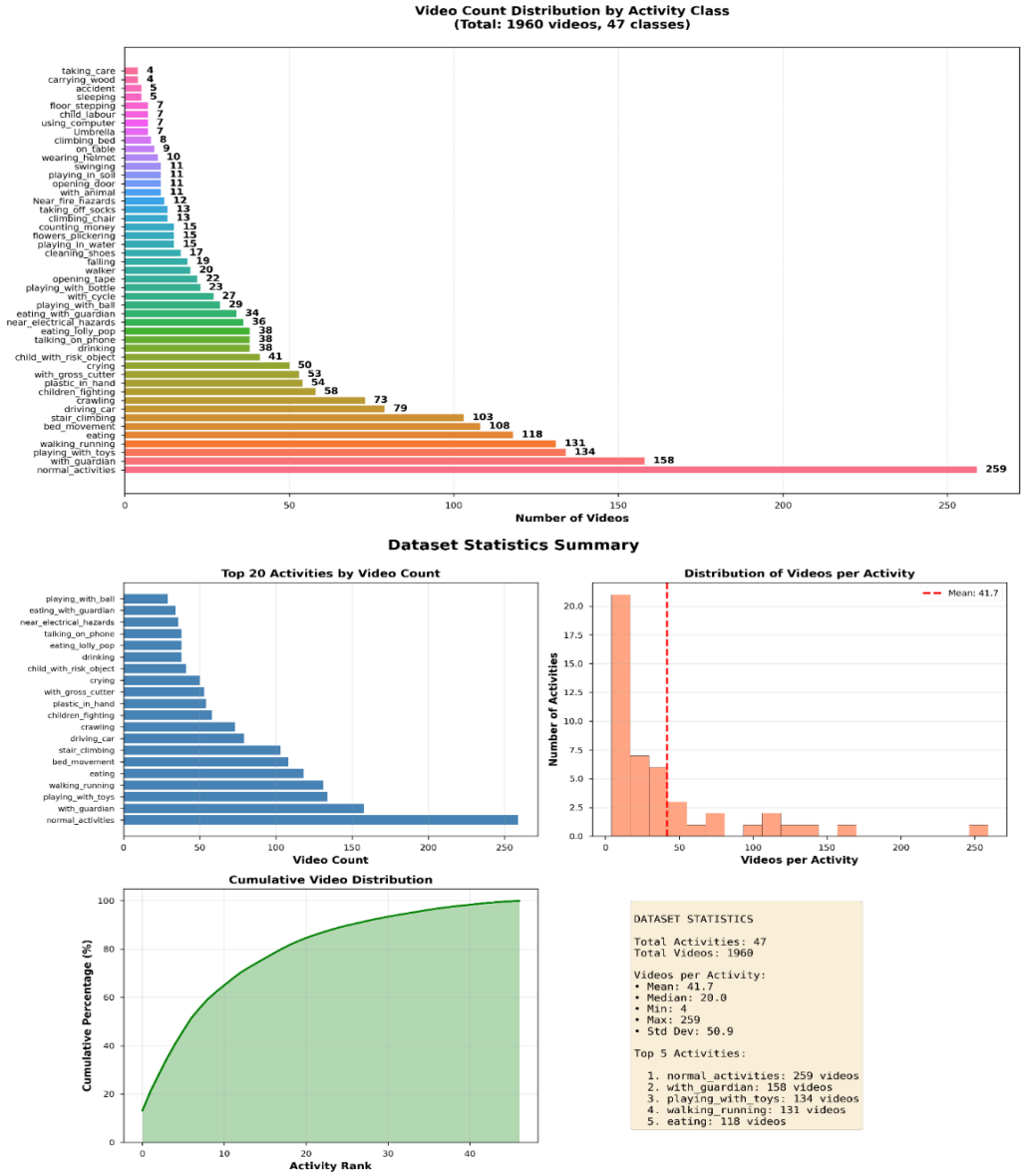

Figure 4. Video classification dataset statistics summary

Training Configuration:

All the models were trained to 100 epochs to reach convergence and the best performance. Training was done with the Adam optimizer with learning rate scheduling and cross-entropy loss was used to perform multi-class classification. The video clips were made up of 16 sampled frames each with an input resolution of 224 x 224 pixels. All the configuration mentioned in Table 3. Generalization was enhanced by using standard data augmentation methods, such as random cropping, horizontal flipping, and temporal sampling.

Table 3. Training Configuration for each model

Parameter	Configuration
Training Epochs	100
Optimizer	Adam (with learning rate scheduling)
Loss Function	Cross-entropy loss
Frames per Clip	16
Input Resolution	224 × 224
Data Augmentation	Random cropping, horizontal flipping, temporal sampling

A detailed summary of the experimental protocol and validation strategy employed for both child detection and activity recognition tasks is presented in Table 3. The standardized hold-out validation approach, consistent hyperparameter settings, and balanced dataset partitioning ensured fair comparative benchmarking across all evaluated models while improving reproducibility and methodological transparency. The detailed experimental protocol, dataset split strategy, training configuration, and evaluation metrics used in this study are summarized in Table 4.

Video Classification Models Training:

All models were trained on an NVIDIA GeForce RTX 3060 GPU with 12GB VRAM. Training and validation used consistent dataset splits, with final evaluation performed on a held-out test set. Both training and validation metrics were tracked during training to monitor convergence Shown in Figure 5.

Child Detection Models:

Child detection is performed using a YOLO-based object detection model trained on a custom dataset. The dataset consists of 19,890 annotated images in YOLO format, collected from diverse indoor and outdoor environments. Each image is labeled with a single class ("child"), enabling the model to learn child-specific visual features. YOLO26s was also trained for 100 epochs under the same dataset and image-size settings to enable a fair comparison. The training curves demonstrate strong convergence and slightly improved localization performance compared to YOLOv8s and RT-DETR-L. Final validation performance reached Precision = 0.9717, Recall = 0.9618, mAP@0.50 = 0.9851, and mAP@0.50–0.95 = 0.8780. The best mAP@0.50–0.95 (0.8787) occurred near the end of training (epoch 99), suggesting that most gains were achieved late in the optimization and that the final checkpoint is close to optimal. Figure 6 shows the overall performance of yolov26s model.

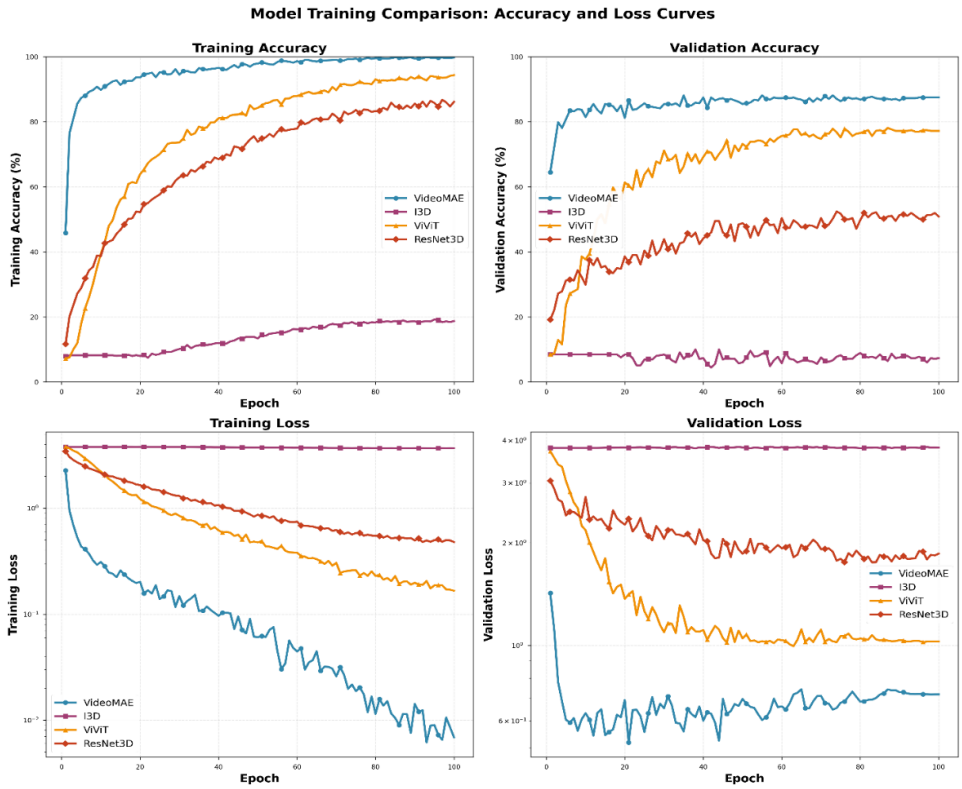

Figure 5. Training and validation performance matrixes of video classifiers

Table 4. Experimental Protocol and Validation Strategy

Component	Dataset Size	Split Strategy	Training Set	Validation Set	Models Evaluated	Training Configuration	Evaluation Metrics
Child Detection	19,890 images	Hold-out validation	16,107 images (81%)	3,783 images (19%)	YOLOv8s, YOLO26s, RT-DETR-L	100 epochs, identical hyperparameters	Precision, Recall, mAP@0.50, mAP@0.50–0.95
Activity Recognition	~1,960 video clips, 47 classes	Class-balanced hold-out	~70%	~15%	I3D, 3D ResNet, ViViT, VideoMAE	16-frame clips, 224×224 resolution, Adam optimizer, Cross-Entropy Loss	Accuracy, Precision, Recall, Macro F1-score

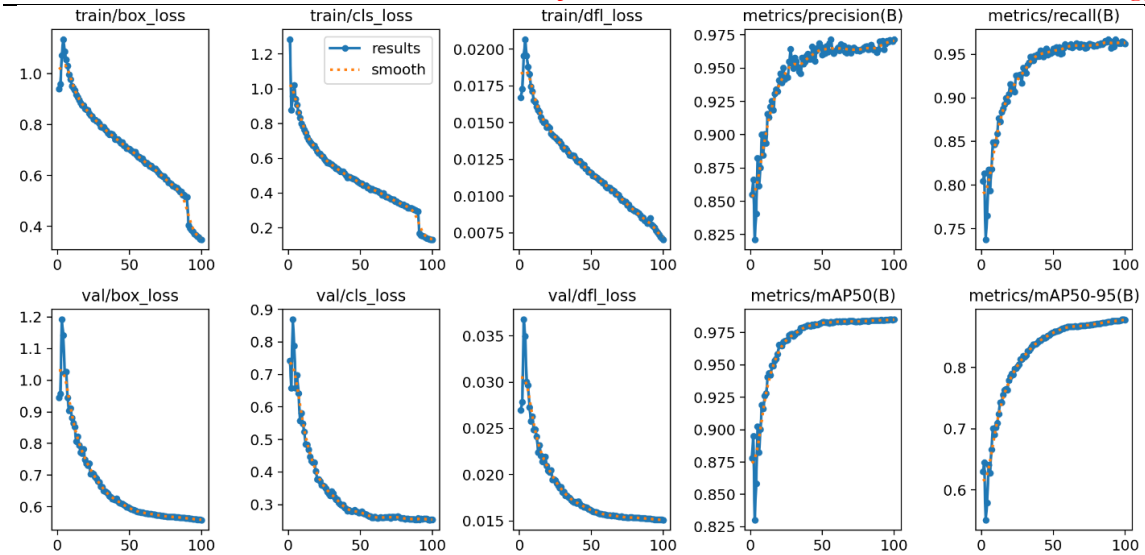

Figure 6. Training and Validation Performance of Yolov26s Model on child detection

Interfacing Pipeline:

The entire pipeline of child activity recognition is shown in Algorithm 1. The structure combines child detection based on the YOLO model, child tracking based on the detected children, and temporal activity recognition based on the VideoMAE model. The system works in a sequence of video frames and produces annotated video frames with predicted activities.

The first stage employs YOLOv8s (You Only Look Once version 8 small) for real-time child detection. For each input frame $I_t \in \mathbb{R}^{H\times W\times 3}$ at time t, the YOLO model produces a set of detections $\mathcal{D}_t = \{d_1, d_2, \ldots, d_n\}$, where each detection d_i is characterized by:

$$d_i = \{b_i, c_i, s_i\} (1)$$

Here, $d_i$represents the *i-th detected object* in a given video frame. Each detection is defined by three components:

b_i: Bounding box coordinates representing the spatial location of the detected child in the frame

c_i: Class label of the detected object (in this study, "child")

s_i: Confidence score indicating the probability of correct detection

Thus, each detection $d_i$encapsulates both spatial and semantic information required for downstream tracking and activity recognition tasks.

where $b_i = x_1^{(i)}, y_1^{(i)}, x_2^{(i)}, y_2^{(i)}]$ represents the bounding box coordinates, c_i is the class ID, and $s_i \in [0,1]$ is the detection confidence score.

The detection process applies two thresholds:

Confidence threshold: $\tau_{conf} = 0.3$ - Only detections with $s_i \geq \tau_{conf}$ are retained

IoU threshold: $\tau_{IoU} = 0.45$ - non-maximum suppression (NMS) removes overlapping detections. The filtered detections are given by:

$$\mathcal{D}_t^{filtered} = \{d_i \in \mathcal{D}_t : s_i \geq \tau_{conf} \text{ and } \text{NMS}(d_i, \mathcal{D}_t, \tau_{IoU}) = \text{keep}\} (2)$$

Here, D_t^{filtered}represents the set of valid detections at time $t$after applying confidence thresholding and Non-Maximum Suppression (NMS).

D_t: Set of all detected objects in frame t

d_i: Individual detection defined as $\{b_i, c_i, s_i\}$

s_i: Confidence score of detection d_i

τ_{conf}: Confidence threshold used to filter low-confidence detections

$\text{NMS}(\cdot)$: Non-Maximum Suppression function to remove overlapping bounding boxes

τ_{IoU}: Intersection-over-Union threshold used in NMS
A detection $d_i$is retained only if:
Its confidence score exceeds the threshold τ_{conf}, and
It is not suppressed during the NMS process (i.e., selected as "keep").

To handle imperfect detection where children may not be detected in every frame, the pipeline implements temporal smoothing. When no detections are found in frame t ($|\mathcal{D}_t| = 0$) but previous detections exist ($|\mathcal{D}_{t-1}| > 0$), the system uses the previous detections with reduced confidence:

$$\mathcal{D}_t = \begin{cases} \mathcal{D}_t^{filtered} & \text{if } |\mathcal{D}_t^{filtered}| > 0 \\ \{d' : d' = \{b, c, s \times 0.9\} \text{ for } d = \{b, c, s\} \in \mathcal{D}_{t-1}\} & \text{if } |\mathcal{D}_t^{filtered}| = 0 \text{ and } |\mathcal{D}_{t-1}| > 0 \\ \emptyset & \text{otherwise} \end{cases} \quad (3)$$

This equation defines the final detection set D_t^*at time t, incorporating a fallback mechanism to handle missed detections.
D_t^{filtered}: Filtered detections after thresholding and NMS
D_{t-1}: Detections from the previous frame
$d_i = (b_i, c_i, s_i)$: Detection tuple (bounding box, class, confidence)
d_i': Adjusted detection with reduced confidence
$|\cdot|$: Cardinality (number of detections)

Case-wise behavior:

Normal case:

If detections exist in the current frame, use them directly.

Fallback case:

If no detections are found in the current frame but previous detections exist, reuse previous detections with reduced confidence ($0.9 \times s_i$) to maintain temporal continuity.

Empty case:

If no detections exist in both current and previous frames, return an empty set.

Algorithm 1 Child Activity Recognition Pipeline

Require: Video stream or file V, YOLO model M_{yolo}, VideoMAE model $M_{videomae}$, Tracker T
Ensure: Annotated video frames

Initialize track buffers $\mathcal{B} = \{\}$, activities $\mathcal{A} = \{\}$
for each frame I_t in V **do**
  $\mathcal{D}_t \leftarrow \text{DetectChildren}(I_t, M_{yolo})$
  if $|\mathcal{D}_t| = 0$ and temporal smoothing enabled **then**
    $\mathcal{D}_t \leftarrow \text{TemporalSmooth}(\mathcal{D}_{t-1})$
  end if
  $\mathcal{T}_t \leftarrow \text{UpdateTracker}(\mathcal{D}_t, T)$
  $\mathcal{G} \leftarrow \text{GroupIntersecting}(\mathcal{T}_t)$
  for each group $g \in \mathcal{G}$ **do**
    if $|g| = 1$ and $|\mathcal{T}_t| = 1$ **then**
      $F \leftarrow I_t$ ▷ Full frame
    else if $|g| = 1$ and $|\mathcal{T}_t| > 1$ **then**
      $F \leftarrow \text{Mask}(I_t, b_g, \mathcal{T}_t \setminus g)$
    else
      $b_{union} \leftarrow \text{UnionBox}(g)$
      $F \leftarrow \text{Crop}(I_t, b_{union})$
    end if
    Append F to buffer $\mathcal{B}[\text{id}_g]$
  end for
  if $t \bmod 5 = 0$ **then**
    $\mathcal{A} \leftarrow \text{PredictActivities}(\mathcal{B}, M_{videomae})$ ▷ Parallel
  end if
  $I_{annotated} \leftarrow \text{DrawDetections}(I_t, \mathcal{T}_t, \mathcal{A})$
  Output $I_{annotated}$
end for

Results and Discussion:

It is an experimental section where findings of the developed child activity recognition framework are provided, and the aspects of detection and activity recognition in the system are evaluated. The goal of the analysis is to estimate the system on the level of the correct localization of the children in the video frame and detection of their activities within the time frame.

Three detectors that were trained on the same single-class child detection dataset over 100 epochs, all models performed extremely well, with mAP@0.50 being near show in Table 5 at each method, suggesting that child presence can be detected reliably under the conditions evaluated. However, the disparities are even more pronounced when we look at the more stringent localization measure mAP@0.50 0.95 that is more representative of the quality of bounding-boxes at different levels of IoU. In this comparison, YOLO26s delivered the best overall localization (mAP@0.50–0.95 = 0.8780), followed by YOLOv8s (0.8647) and RT-DETR-L (0.8525).

Table 5. performance metrics of the trained detection algorithms

Model	Precision (P)	Recall (R)	mAP50	mAP50-95
RT-DETR-L	0.9680	0.9644	0.9818	0.8525
YOLO26s	0.9717	0.9618	0.9851	0.8780
YOLOv8s	0.9736	0.9608	0.9827	0.8647

Table 6 summarizes the test-set performance of the four trained video activity recognition models using overall accuracy and class-balanced metrics. VideoMAE achieves the best results (Accuracy = 0.8638) and the strongest class-consistent performance (Macro F1 = 0.9049, Macro Precision = 0.9158, Macro Recall = 0.9075), indicating robust recognition across the full set of activities. ViViT ranks second (Accuracy = 0.7924, Macro F1 = 0.8314), showing competitive performance but with a clear gap to VideoMAE. The 3D-CNN baselines perform substantially worse: ResNet3D reaches moderate performance (Accuracy = 0.5134, Macro F1 = 0.5544), while I3D fails to learn meaningful discrimination under the current setup (Accuracy = 0.0960, Macro F1 = 0.0108). Overall, the metrics show a consistent advantage of transformer-based approaches for this dataset, with VideoMAE providing the best balance between overall accuracy and class-level generalization.

Table 6. Performance comparison of video activity recognition models on the 47-class test set (Accuracy, Micro F1, Macro/Weighted F1, and Macro/Weighted Precision/Recall).

Model	Accuracy	Macro F1	Weighted F1	Macro Precision	Macro Recall	Weighted Precision	Weighted Recall	Micro F1
VideoMAE	0.8638	0.9049	0.8616	0.9158	0.9075	0.8736	0.8638	0.8638
I3D	0.0960	0.0108	0.0322	0.0066	0.0325	0.0196	0.0960	0.0960
ViViT	0.7924	0.8314	0.7902	0.8494	0.8301	0.8029	0.7924	0.7924
ResNet3D	0.5134	0.5544	0.5231	0.7244	0.5154	0.6592	0.5134	0.5134

The relative comparison of the object detection models shows that all three models performed extremely well in detecting objects on the custom validation dataset. Among them, YOLO26s achieved the best performance with a mAP@0.50 of 0.9850 and mAP@0.50–0.95 of 0.8787, closely followed by YOLOv8s, while RT-DETR-L showed slightly lower performance. The close performance between YOLO26s and YOLOv8s indicates that both architectures are highly effective for child detection when trained on a domain-specific dataset.

Computationally, YOLO-based models had better real-time performance than transformer-based detectors. YOLOv8s reached around 77–107 FPS, and YOLO26s reached 66–80 FPS, which is very appropriate to be used in real-time. Meanwhile, RT-DETR-L reached around 2324 FPS, which, despite the competitiveness in accuracy, can restrict its

application in real-time systems. These findings point out the accuracy-speed trade-off, with lightweight architectures being more acceptable to real-world use.

The performance of activity recognition models shows a clear distinction between transformer-based and convolution-based approaches. VideoMAE achieved the highest performance with an accuracy of 86.38% and a Macro F1-score of 90.49%, followed by ViViT, while ResNet3D and I3D exhibited significantly lower performance.

VideoMAE outperforms its rivals due to its transformer-based architecture and masked autoencoding that allows efficient learning of both spatial and temporal representations. VideoMAE can better model complex activity patterns than more traditional convolution-based models because it can capture long-range dependencies between video frames.

Statistical Analysis and Error Discussion:

To test the validity of the received results, all experiments were performed under the same training and evaluation conditions using fixed dataset splits. The stated performance measures constitute the results on held-out validation and test sets, which makes the results of all the evaluated models fairly compared. Though several independent experimental runs were not carried out to calculate statistical variance and confidence intervals, an error analysis was done to assess model behavior and robustness.

The analysis shows that the majority of misclassifications are between the visually and temporally related activities, i.e., between walking and running, playing and active movement. The challenges come about because of the small differences in motion and a lack of temporal context in short video clips.

Moreover, small errors during the detection phase sometimes carry over to the activity recognition phase, with a minor impact on end predictions. Nevertheless, transformer-based models, especially VideoMAE, are always better than other architectures in all evaluation metrics, which shows that they have a strong generalization capacity and are very robust. Subsequent work will involve several experimental runs, estimation of a confidence interval, and hypothesis testing to further confirm the statistical significance and model stability.

Recommendations and Future Work:

Although the suggested end-to-end child safety surveillance system proves to be highly effective in terms of child detection, tracking, and activity recognition, there are still several opportunities to be further enhanced. Future research should be directed towards expanding the scale and scope of child-centric data, to capture more complex real world conditions, more behavior types, and more variations in the environment.

Additional increase in the diversity of datasets would lead to further increases in model generalization in different surveillance environments. In technical terms, future work may look into optimization of lightweight models, edge deployment, and real-time embedded implementations to enhance scalability to resource constrained environments like smart homes, daycare centers, and wearable safety systems.

Safety awareness, not founded solely on visual surveillance, may also be improved through the integration of multimodal sensing techniques, such as audio analysis, physiological monitoring and sensor integration of context. Also, additional enhancements can be made, such as anomaly detection to identify rare or dangerous child behaviors, self-supervised learning to enhance domain adaptation, and adaptive personalization to monitor child behavior in a specific age. Larger-scale cross-validation, hypothesis testing, and longitudinal deployment studies may also be used to enhance the statistical robustness.

Ethically and socially, future research needs to focus on privacy-saving AI, federated learning, and secure surveillance protocols in order to make responsible deployment and protect child data at the same time.

Conclusion:

This study introduced an in-depth deep learning-based child safety monitoring and activity recognition system based on video surveillance. The proposed architecture integrates object detection, multi-object tracking, and the video activity recognition into a single end-to-end-architecture that can analyze the children's activities in real-time. To address the shortcomings of the available datasets of activity recognition, an activity recognition dataset was created that contains more child-specific behaviors and safety-related activities data. To conduct activity recognition, a dataset comprising of 47 classes of activities carried out by children was trained and tested using different deep learning networks, including VideoMAE, ViViT, ResNet3D, and I3D. Overall, this research study demonstrates the feasibility of using contemporary methods of deep learning to monitor child safety automatically. The system offers a scalable and practical solution that can be incorporated in existing surveillance systems in homes, daycare centers, schools, and other public places by integrating a high-quality detection, time tracking, and advanced activity recognition capabilities.

Acknowledgement: This manuscript has not been published previously and is not under consideration elsewhere.

Author's Contribution: Samad Riaz conceptualized the study and developed the methodology. **Shayan Riaz** and **Abdur Razzaq** contributed to dataset preparation and preprocessing. **Umar Saddique** assisted in manuscript review and compilation. **Shahid Bashir** supervised the research and provided critical revisions. All authors reviewed and approved the final manuscript.

Conflict of Interest: The authors declare that there is no conflict of interest regarding the publication of this manuscript in IJIST.

References:

[1] "FAST FACTS: Violence against children widespread, affecting millions globally." Accessed: May 09, 2026. [Online]. Available: https://www.unicef.org/press-releases/fast-facts-violence-against-children-widespread-affecting-millions-globally

[2] K. R. Tanveer, M. S. Luqman, and A. Qureshi, "Ensuring Child Safety: An IoT-Based Surveillance System for Remote Monitoring and Detection of Anomalous Behavior," pp. 1–5, Nov. 2024, doi: 10.1109/ICETST62952.2024.10737980.

[3] J. H. Tan and C. P. Goh, "Enhancing Child Safety: Computer Vision-Based Accident Detection for Infants and Toddlers," *2024 3rd Int. Conf. Digit. Transform. Appl.*, pp. 179–183, 2024, doi: 10.1109/ICDXA61007.2024.10470712.

[4] G. Singh, A. R. Shekhar, X. Yu, and J. Saniie, "Smart Infant Monitoring System Using Computer Vision and AI," *IEEE Int. Conf. Electro Inf. Technol.*, vol. 2023-May, pp. 347–352, 2023, doi: 10.1109/EIT57321.2023.10187295.

[5] A. Prathyanga, P. Shyaminda, P. Chamikara, S. Lakshan, S. Thelijjagoda, and D. Kasthurirathna, "Intelligent Daycare: Enhancing Child Safety with IoT and Machine Learning Innovations," *Proc. 9th Int. Conf. Commun. Electron. Syst. ICCES 2024*, pp. 530–538, 2024, doi: 10.1109/ICCES63552.2024.10859472.

[6] Shehzad Ali, Md Tanvir Islam, "CABAD: A video dataset for benchmarking child aggression recognition," *Alexandria Eng. J.*, vol. 127, pp. 1143–1157, 2025, doi: https://doi.org/10.1016/j.aej.2025.06.035.

[7] Lucia Migliorelli, Sara Moccia, "The babyPose dataset," *Data Br.*, 2020, [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/33083503/

[8] Pengbo Wei, David Ahmedt-Aristizabal, "Vision-based activity recognition in children with autism-related behaviors," *Heliyon*, vol. 9, no. 6, 2023, doi: https://doi.org/10.1016/j.heliyon.2023.e16763.

[9] Somaieh Amraee, Bishoy Galoaa, Matthew Goodwin, Elaheh Hatamimajoumerd, Sarah Ostadabbas, "Multiple Toddler Tracking in Indoor Videos," *arXiv:2311.17656*, 2023, [Online]. Available: https://arxiv.org/abs/2311.17656

[10] Chiradeep Roy, Mahsan Nourani, "Explainable Activity Recognition in Videos using Deep Learning and Tractable Probabilistic Models," *ACM Trans. Interact. Intell. Syst.*, vol. 13, no. 14, 2023, [Online]. Available: https://dl.acm.org/doi/10.1145/3626961

[11] "Integrating YOLOv8 and IoT in a Computer Vision System for Child Detection in Smart Cities." Accessed: May 09, 2026. [Online]. Available: https://thesai.org/Publications/ViewPaper?Volume=16&Issue=9&Code=IJACSA&SerialNo=60

[12] K. L. Tran, M. N. Dang, T. N. Trong, H. N. Quoc, and L. N. Kieu, "Enhancing YOLOv11n for Reliable Child Detection in Noisy Surveillance Footage," Feb. 2026, Accessed: May 09, 2026. [Online]. Available: http://arxiv.org/abs/2602.10592

[13] "Opportunities, Applications, and Challenges of Edge-AI Enabled Video Analytics in Smart Cities: A Systematic Review | IEEE Journals & Magazine | IEEE Xplore." Accessed: May 09, 2026. [Online]. Available: https://ieeexplore.ieee.org/document/10198424

[14] "(PDF) Smart Home Monitoring System for Early Childhood Using Computer Vision Technology." Accessed: May 09, 2026. [Online]. Available: https://www.researchgate.net/publication/395824244_Smart_Home_Monitoring_System_for_Early_Childhood_Using_Computer_Vision_Technology

[15] "(PDF) Child Activity Recognition using Deep Learning." Accessed: May 09, 2026. [Online]. Available: https://www.researchgate.net/publication/354778046_Child_Activity_Recognition_using_Deep_Learning

[16] A. Sandygulova, A. Yershov, A. Zhanatkyzy, and Z. Telisheva, "ChildACT: Child Action Recognition Dataset in RGB Data," *ACM/IEEE Int. Conf. Human-Robot Interact.*, pp. 1088–1092, 2025, doi: 10.1109/HRI61500.2025.10974037.

[17] "Falling Detection of Toddlers Based on Improved YOLOv8 Models - PubMed." Accessed: May 09, 2026. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/39409491/

[18] S. S. Hidayat, D. Aprilia, S. Hadwi, I. Mujahidin, M. C. A. Prabowo, and F. A. Rakhman, "Child Presence Detection for Child Safety with Deep Neural Networks," *J. Inform. J. Pengemb. IT*, vol. 10, no. 2, pp. 370–381, Apr. 2025, doi: 10.30591/JPIT.V10I2.6540.

[19] Y. Zhang *et al.*, "ByteTrack: Multi-Object Tracking by Associating Every Detection Box," Apr. 2022, Accessed: May 09, 2026. [Online]. Available: http://arxiv.org/abs/2110.06864

[20] S. Zhu *et al.*, "A Video-based End-to-end Pipeline for Non-nutritive Sucking Action Recognition and Segmentation in Young Infants," Mar. 2023, Accessed: May 09, 2026. [Online]. Available: http://arxiv.org/abs/2303.16867

[21] "EduNet: A New Video Dataset for Understanding Human Activity in the Classroom Environment." Accessed: May 09, 2026. [Online]. Available: https://www.mdpi.com/1424-8220/21/17/5699

[22] Klim Kireev, Ana-Maria Creţu, Raphael Meier, Sarah Adel Bargal, Elissa Redmiles, Carmela Troncoso, "A Manually Annotated Image-Caption Dataset for Detecting Children in the Wil," *arXiv:2506.10117*, 2025, [Online]. Available:

https://arxiv.org/abs/2506.10117

[23] T. S. Yashwanth, Y. S. Royal, M. Kashyap, V. R. Shreya, and D. K. N, "Real Time Child Abduction and Detection System," pp. 1–6, Dec. 2025, doi: 10.1109/SITA67914.2025.11273371.

[24] V. Godase, "Edge AI for Smart Surveillance: Real-time Human Activity Recognition on Low-power Devices," *SSRN Electron. J.*, 2025, doi: 10.2139/SSRN.5383804.

[25] Zihan Wang, Yang Yang, Zhi Liu, Yifan Zheng, "Deep Neural Networks in Video Human Action Recognition: A Review," *arXiv:2305.15692*, 2023, [Online]. Available: https://arxiv.org/abs/2305.15692

[26] Zhan Tong, Yibing Song, Jue Wang, Limin Wang, "VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training," *arXiv:2203.12602*, 2022, [Online]. Available: https://arxiv.org/abs/2203.12602

[27] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, Cordelia Schmid, "ViViT: A Video Vision Transformer," *arXiv:2103.15691*, 2021, [Online]. Available: https://arxiv.org/abs/2103.15691