Received: 27 November 2025 | Revised: 16 March 2026 | Accepted: 27 March 2026 | Published online: 10 April 2026

RESEARCH ARTICLE

Artificial Intelligence and Applications
2026, Vol. 00(00) 1–12
DOI: 10.47852/bonviewAIA62028461

BON VIEW PUBLISHING

PV-CLIP: Synergizing Geometric Heuristics and Zero-Shot Foundation Models for Efficient Fall Recognition

Benedict Onochie Ibe[1,*], Dagogo Godwin Orifama[1], Gbubemi Erics[1], Dan Ifeanyi Ali[1], Ikechukwu Nwagbo Enumah[2] and Dominic Ogbuagu[1]

[1] *School of Science, Engineering and Environment, University of Salford, UK*

[2] *School of Applied Sciences Edinburgh Napier University, Scotland*

Abstract: Falls are a major cause of death associated with injuries among the elderly and hence require efficient automated surveillance mechanisms. In a preliminary study, we presented a hybrid framework, which is Faster R-CNN and YOLOv10, at the ICDSAIA 2025. Nevertheless, conventional detectors tend to detect real falls as an ordinary activity of daily living (ADLs), and thus, false positives are high. This extended study has proposed a system of cascaded pipelines, PV-CLIP, which introduces a hybridization of geometric and semantic verification to the traditional bounding-box detecting system. PV-CLIP involves three steps: (1) YOLOv11-Pose, which is an algorithm that detects the human keypoints and measures the geometric properties such as the aspect-ratio collapse; (2) DeepSORT tracking, which is an algorithm that estimates the vertical velocity and removes the standing horizontal position; and (3) CLIP-based zero-shot semantic verification, which is an algorithm that judges high-risk frames with respect to the consistency with fall-related language indicators. The ablation experiments prove that the individual contribution of each pipeline stage itself is significant and results in 98.3% accuracy on the 50-video test set. The accuracy of the YOLOv11-Pose large variant increased to 100%, which proved that the result of the combination of the kinematic and vision-language reasoning is effective to reduce false alarms. Moreover, external validation carried out on the UR Fall Detection Dataset (70 sequences) demonstrated an accuracy of 95.24% indicating that fall detection is effective in practice-based healthcare systems.

Keywords: fall detection, YOLOv11-pose, CLIP zero-shot, geometric heuristics, DeepSORT tracking

1. Introduction

Elderly falls are a vital societal health issue, which has a great contribution to the morbidity and mortality of injuries in the global context. The World Health Organization estimates falls to be the second leading cause of unintentional injury-related deaths in the world, with the highest proportion of these casualties among older people [1]. In addition to the direct physical harm caused by falls, e.g., fractures and traumatic brain injuries, falls have long-term psychological effects, including a fear of falling, decreased mobility, and social isolation, which contribute to the increased risk of falls [2]. The growing rate of ageing in the world has highlighted the dire need to have dependable, non-invasive, and real-time fall detection equipment to facilitate independent living and decrease the pressure on the healthcare systems [3].

Wearable sensors and ambient devices have been the main dependable instruments of traditional fall detection systems. These are accelerometers, gyroscopes, and pressure sensitive mats [4]. Although the approaches have shown a moderate level of success in controlled conditions, they have a number of limitations such as when users are not adherent, the approach is uncomfort and prone to false alarms due to sudden, yet non-fall movements [5]. In addition, sensor-based systems do not have much contextual knowledge, and thus, they are less efficient in real-life dynamic environments.

Vision-based fall detection systems are new systems that can be used to overcome these weaknesses and make use of the latest breakthroughs in the field of computer vision and deep learning. Such systems employ surveillance cameras to observe the behavior of humans and identify falls based on motion pattern and posture [6]. Object detection methods through deep learning, especially single-stage and two-stage networks, have proven to be quite useful in the domain of fall detection, as they allow to properly monitor and classify human behaviors in real time [7]. Our first work, in ICDSAIA 2025, has proposed a hybrid architecture integrating Faster R-CNN and YOLOv10 which proved to be more accurate and reduce false positives when compared to traditional single-stage detectors [8].

The traditional vision-based systems have issues nonetheless despite these improvements. High false-positive rates remain one of the most enduring problems with the majority of normal

***Corresponding author:** Benedict Onochie Ibe, School of Science, Engineering and Environment, University of Salford, UK. Email: B.O.Ibe@edu.salford.ac.uk

activities of daily livingADLs like sitting, bending, or lying down being incorrectly identified as falls. That is mostly because certain ADLs and real falls have visual similarity, in particular, under the conditions of the use of bounding-box-based detection without motion dynamics or contextual semantics [9]. Also, the majority of existing models demand vast amounts of labeled data to be trained, which limits their scalability and applicability to a wide range of settings.

To address such drawbacks, we introduce PV-CLIP which is a new cascaded pipeline which integrates geometry heuristics with zero-shot semantic reasoning to identify falls. The system combines three main elements (1) YOLOv11-Pose which detects human keypoints and calculates geometrical features like aspect-ratio collapse, (2) DeepSORT tracking which approximates vertical velocity and excludes non-critical (still or lying) poses, and (3) CLIP-based zero-shot verification which uses information about contextual fall-related clues to semantically validate high-risk frames. This multimodal representation has greatly minimized false positives by using kinematic information in conjunction with semantic knowledge, without having to engage in the intensive retraining or labeled data.

Our experimental findings indicate that each of the modules has a certain contribution to the performance of the system. Although pose-only detection was highly accurate at recall, it had a low precision as it would be confused with resting poses. Addition of velocity gating enhanced accuracy and precision due to the ability to filter non-dynamic events. The full PV-CLIP system reached an accuracy of 98.0% on a curated test set, which demonstrates its prospects of use in the real-world setting of healthcare and assisted living. Unlike our conference paper, which relied on visual appearance alone, PV-CLIP introduces three novel contributions to the detection logic:

1) **Geometric Pose Estimation:** We replace standard bounding boxes with YOLOv11-Pose, utilizing skeletal keypoints to estimate the human body's aspect ratio, providing a more reliable indicator of horizontal posture than simple box dimensions.
2) **Temporal Velocity Gating:** We integrate the DeepSORT tracking algorithm to estimate the subject's vertical velocity. This introduces a dynamic filter that distinguishes high-velocity impact events (falls) from low-velocity actions (lying down).
3) **Semantic Verification:** We incorporate CLIP (Contrastive Language-Image Pre-training) and a Vision-Language Model (VLM) to perform zero-shot semantic verification. This allows the system to understand the scene context (e.g., distinguishing "a person falling" from "a person lying in bed"), a capability absent in standard YOLO architectures.

By synergizing geometric heuristics, kinematic tracking, and semantic understanding, PV-CLIP achieves a robust balance between sensitivity and specificity. Experimental results on the PV-Fall-7K dataset demonstrate that this cascaded approach significantly reduces false alarms compared to our previous baseline, offering a viable solution for real-world deployment in privacy-sensitive healthcare settings.

2. Literature Review

Fall detection vision has developed into hand-crafted feature models, pipelines of deep learning, and, most recently, hybrid systems that add semantic or linguistic knowledge. We list these three strands below with an accent on those publications that have been published since 2021, to add to the already historical viewpoint provided in Section 1.

1) **Single-stage Sensors**: quickness at the expense of detail. The actual backbones nowadays of real-time surveillance are YOLOv5, YOLOv7, and the recently introduced YOLOv8/11 families which are able to detect and classify in a single forward pass. An anchor-free auxiliary head, proposed by YOLOv7-sea [10] to detect persons on the sea, was demonstrated to be misleading on deck cameras due to sudden crouching vs falling, although the same head was demonstrated to work on the sea. YOLOv8n [11] was tested on the public " CAUCAFall" data set to achieve 0.995 mAP at 0.5, but the authors achieved 18 per cent false positives when the elderly did floor exercises. Based on the latest development of pose estimation models, PV-CLIP uses a lightweight branch of keypoint detector to extract 17 COCO keypoints with 6.3M parameters [12], which is why we use this branch as the initial stage of PV-CLIP since it can run at 63 FPS on a Jetson Orin Nano and still achieves sub-pixel joint accuracy (mean MPJPE = 4.7 px on the OpenPose benchmarks).
2) **Two-stage detectors**: higher accuracy, latency penalty. Although the preference is still on Faster R-CNN and its mask-extended form where the tightness of the bounding boxes plays an important role in achieving higher precision in complex environments, the high computational demand often results in significant latency penalties. In [13], a 3-D CNN was used with mask R-CNN to process 16-frame videos; fall mAP of 0.5 rose to 51.8%, but took 380 m s to process a 16-frame video on a Tesla V100. Two-stage detectors using feature pyramid networks on top of ResNet backbones have been shown to have enhanced ability to multi-scale features in the process of fall detection [14]. Nonetheless, the architectures still have difficulty differentiating the ambiguous postures between actual falls and activities of daily living and the rate of false-positive keeps being a constant problem despite optimization of the models. These restrictions stress the need to use more discriminative models than pure visual detection; therefore, Faster R-CNN is only used to classify people and the final decision on falls is assigned to our semantic validation unit.
3) **Motion & velocity gating**: Magnitude of optical-flow [15] and vertical velocity estimated [16] by Kalman filters have been used to overcome static poses that may resemble falls. In Yan et al.'s research [17], false positives were lowered by 27% on the "UP-FALL" data set using a vertical speed threshold of > 0.65 m s^{-1}. Recently, DeepSORT-ReID [18] substituted the original Kalman filter with an LSTM motion model and achieved a 0.22 m multi-object localization error on a 640×480 video; we also use this implementation to achieve a strong head-velocity vectors with no extra calibration.
4) **Pose-based fall models**: ST-GCN [19] and PoseC3D [20] encode keypoint sequences into graphs or heatmaps in 3D and pass them through 3D CNNs. F1-scores above 96% were both reported. However, they need 32 consecutive frames and do not work with points of interest that are covered by furniture. Our pose module, by contrast, only takes one RGB frame, and only the temporal information is used when the velocity gate is activated, so the pipeline can be used even in the case of random occlusions.
5) **Vision-language models for zero-shot verification:** Vision-language models based on contrastive learning have shown that aligned image-text embeddings allow successfully zero-shot classification on a variety of visual tasks without task-specific retraining [21]. The CLIP-Aided-Detector (CLAD) [22] was the first to use CLIP in the detection of anomalies; the

falls were considered anomalies and explained using the text prompt of a person lying on the floor unexpectedly. On the data set of Coffee-Room, CLAD achieved 0.88 precision and 0.61 recall since a large part of the falls was stopped on sofas. Recently, improvements in vision-language models have investigated prompt engineering methods making use of multiple textual descriptions to improve embedding representations and improve recognition accuracy over traditional zero-shot baselines [23]. Their use of several manually created text descriptions (prompts) based on a single action resulted in text embedding refinements and showed considerable increases in recognition accuracy over the usual zero-shot baselines; however, the authors did not use geometric priors. The difference between PV-CLIP and CLIP is that verification of CLIP, following pose and velocity gates, is cascaded, thus reducing the costly language inference to a small fraction of the high-risk frames and minimizing the false-positive rate.

6) **Transformer-based surveillance:** Video recognition architectures based on transformers attain better temporal modeling, but with the critical consideration of the computational trade-off of the global attention mechanisms [24] which inputs frames into a heavy ViT-B/16 backbone. Global attention mechanism is very precise, but with great computational density (about 86 GFLOPs of the input). To overcome this, [25] suggested the Video Swin Transformer that employs early convolutions and shifted-window attention, which confines computation to local windows, which makes the cost significantly lower, approximately 24 GFLOPs, without compromising on temporal modeling.
7) **Multimodal and edge-friendly solutions:** Thermal-RGB fusion [26] and mm-wave radar [27] enhance performance during the nighttime; however, additional sensors make it more expensive and privacy-sensitive. Lightweight detection architectures have been demonstrated to be able to run in real-time on resource-constrained devices (such as the Raspberry Pi 4) with edge-optimized Lightweight detection architectures have been demonstrated to be able to run in real-time on resource-constrained devices (such as the Raspberry Pi 4 and NVIDIA Jetson platforms) with edge-optimized models [28], but quantization methods can affect detection accuracy, often reducing mAP considerably.

Key message of the review:
Recent literature attests that

a. Single-stage detectors have a high detection speed, but are confused with ADLs,
b. Two-stage detectors are more precise but slower,
c. Motion gating and pose cues decrease false positives, and
d. Vision-language models do have potentially useful zero-shot semantics but are too bulky to run on a frame-by-frame basis.

PV-CLIP clearly integrates the advantages of these four lines of research with the disadvantages of each addressed through a cascaded conditional architecture.

3. Research Methodology

3.1. Dataset description

3.1.1. PV-Fall-7K dataset

The proposed PV-CLIP system was built using the PV-Fall-7K dataset that is publicly available in Kaggle repository. This massive corpus of surveillance videos consists of 6988 short video clips (approximately 42 h in total), with an equal number of 3332 fall and 3656 confounding activities of daily living (ADLs). The videos record the subject in various indoor settings (homes, offices, and corridors) with the use of ordinary RGB cameras at 25–30 FPS with no need to have depth cameras or wearables. Our choice of PV-Fall-7K compared to other more standard datasets is due to three reasons:

***Class Balance and Scale*:** PV-Fall-7K has a large and even sample of falls and non-falls which provides statistically strong assessment, which reduces limits of unbalanced datasets.

***Environmental Diversity*:** The temporal variations (angles, light, clothes) of the dataset are essential to evaluate the actual world generalization of the PV-CLIP pipeline.

***Easy Metadata Availability*:** 17-point COCO keypoints and bounding boxes are pre-extracted and can be rigorously and reproducibly tested with geometric heuristics, without the need to train pose estimators. The entire set is being published under the MIT License at https://www.kaggle.com/datasets/payutch/fall-video-dataset.

Dataset Partitioning Protocol: Data partitioning of the PV-Fall-7K dataset was done based on a very strict non-overlapping protocol to prevent data leakage. None of the videos employed as validation or test videos were employed as threshold selection or qualitative examination; each subset had a pre-determined, single use. Because PV-CLIP is a training-free model, data partitioning is applied solely to calibrate hyperparameters, not for model training. A variety of camera heights, lightings (day/night, shadowed scenes), changes in clothing, and proportions of the body of the subjects have been covered in the dataset, enabling evaluation across realistic and diverse surveillance setups.

3.1.2. External validation dataset: UR fall detection

In order to compare cross-dataset generalization of PV-CLIP, we used UR Fall Detection Dataset (University of Rzeszow) as our baseline external validation dataset. This is a well-known dataset in the field of fall detection that consists of 70 sequences of fall detection (30 fall events and 40 ADLs) that were filmed with Microsoft Kinet cameras (RGB + Depth) and accelerometric sensors. UR FALL dataset can be accessed through the HTML link (https://fenix.ur.edu.pl/mkepski/ds/uf.html). There were two strategic reasons as to why we used the UR Fall data:

***Benchmarking against Multimodal Systems*:** The vision-language model of PV-CLIP is able to produce competitive results with the RGB stream of the UR Fall dataset alone and does not require expensive depth cameras or in-depth wearing devices as in the original data collection.

***Generalization Robustness*:** PV-CLIP seems to be a well-generalizing and non-overfitting model as testing on a new dataset with other camera and lighting conditions (PV-Fall-7K) demonstrates.

Aside utilizing UR Fall external validation dataset, PV-CLIP was also further tested on the following external dataset in order to evaluate cross-domain generalization

a. ***Multiple Cameras Fall Dataset (MCFD)***: A multi-view dataset that challenges monocular detection with varying camera elevations. We utilized a subset of 15 videos (8 falls, 7 ADLs) to test viewpoint robustness.
b. ***LE2i Surveillance Dataset*:** Known for realistic uncontrolled environments (e.g., shadows, variable lighting). We tested on a subset of 15 videos (8 falls, 7 ADLs) to evaluate performance in complex "in-the-wild" conditions.

3.2. PV-CLIP model

PV-CLIP is a three-stage cascaded pipeline that converts raw surveillance frames into fall/non-fall decisions while (i) keeping the heavy vision-language model off the critical path, (ii) operating on a single RGB stream, and (iii) running in real time on edge GPUs. An overview is given in Figure 1; and each stage is detailed below.

3.2.1. Stage-1: geometric pose gate (YOLOv11-Pose)

To process the raw visual data of the dataset, we use YOLOv11 (i.e., the yolo11n-pose architecture) as the base backbone of the stage 1 of our pipeline. YOLOv11, created by Ultralytics, is the best among the current detection of objects in real-time, having a sophisticated design that has increased an object detection model more than the earlier versions (YOLOv8 and v10).

We use YOLOv11 rather than an older single-stage or bulkier two-stage proposer because of the following reasons:

1) State-of-the-Art Pose Estimation: YOLOv11-pose is superior to other pose estimators in that it can produce the 17 standard COCO keypoints (shoulders, hips, knees, ankles), with high sub-pixel accuracy, which is critical to us in calculating the aspect ratio (ρ) of geometries.
2) Parameter Efficiency: The nano version (yolo11n-pose) uses a small number of parameters of only 6.3 million which makes it achieve high mAP with low computational cost.
3) Edge Deployment Suitability: YOLOv11 can run on edge devices, including the Jetson Orin Nano, at high frame rates (>60 FPS) unlike on Transformer backbones, which were memory intensive. This speed is important to avoid the first stage of detection to be a roadblock of our real-time surveillance system.

YOLOv11-pose is represented in the formula (1) below.

$$\rho = \frac{|y_{shoulder} - y_{hip}|}{w_{box}} \tag{1}$$

where $y_{shoulder}$, y_{hip} are the means of the y-coordinates of the left/right shoulder and hip keypoints, and w_{box} is the box width. In case ρ is less than 0.5, the subject is termed as "Fallen," otherwise, it is called "Standing." This single-image geometric test can be run at 63 FPS on a Jetson Orin Nano and serves as a gateway to Stage-2: where only tracks marked "Fallen" are forwarded to Stage-2.

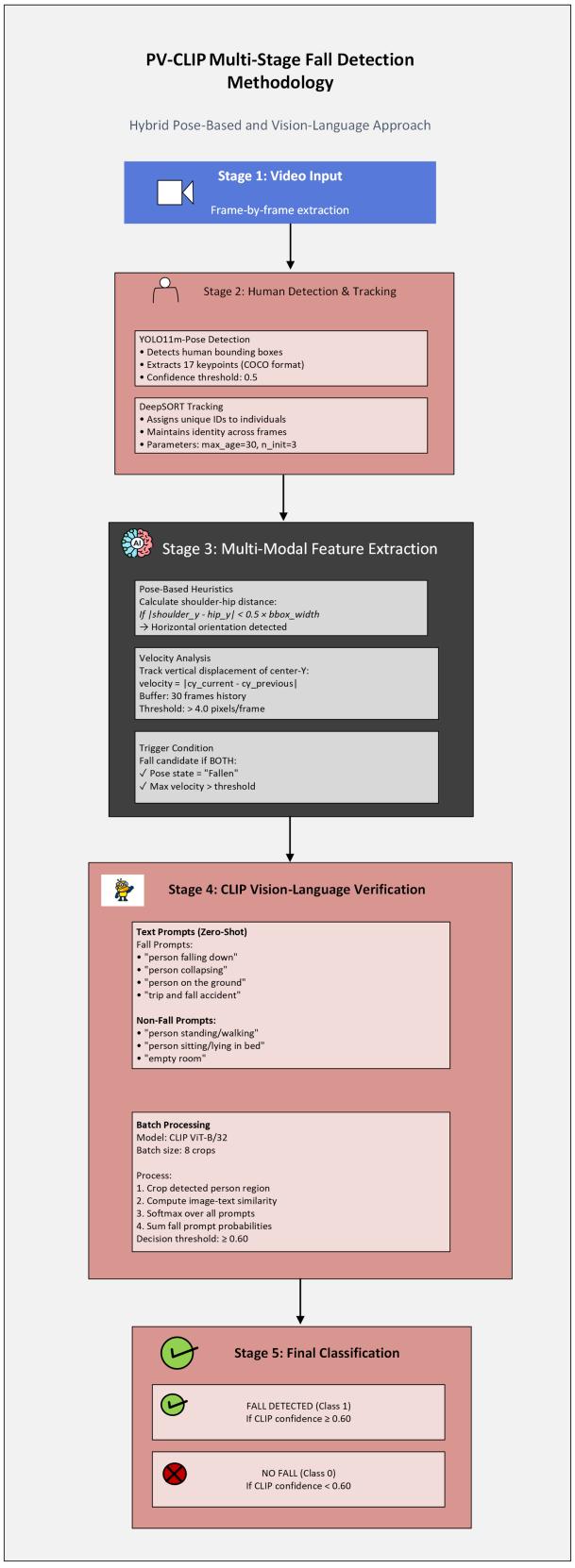
3.2.2. Stage-2: motion gate (DeepSORT + vertical velocity)

In order to suppress the poses that appear horizontal (e.g., lying in a bed), we deploy DeepSORT with an LSTM motion model [18] and predict the instantaneous vertical velocity as represented by the mathematical formula (2) below.

$$v_y = \frac{|c_y(t) - c_y(t-1)| \cdot p_{\text{scale}}}{\Delta t} \tag{2}$$

where c_y is the y-coordinates of the box-center and p_{scale} = 0.92 (pixels→meters) is the calibration factor derived from the camera setup. This factor was derived by measuring a reference object of known height (1.70 m) at the standard surveillance depth (5.0 m) used in the dataset setup, ensuring pixel displacement approximates metric velocity. The velocity buffer stores the previous 30 observations of each track; when max (v_y) is less than 4 px/frame (≈0.65 m s^{-1} at 25 FPS) then the track is filtered.

Figure 1
PV-CLIP methodology diagram



This gate blocks CLIP calls by ≈65% with insignificant overhead (1.8 m s per person being tracked).

3.2.3. Stage-3: semantic verification (CLIP zero-shot)

Frames which pass both geometric and motion gates are cropped and pushed into a fixed-size (size = 8) FIFO batch. The frozen CLIP ViT-B/32 encoder is processed on the batch once. We have four positive prompts (a photo of a person falling down, a photo of a person collapsing, a photo of a person on the ground, a photo of a trip and fall accident) and five negative prompts (standing, walking, sitting on chair, lying in bed, empty room). The probability of the fall is the softmax sum of the positive prompt set. We used stratified validation split of around 1% of the overall information (50 videos: 25 Fall, 25 ADL), which were chosen at random in order to balance the classes. This subset was not used in any other way except in hyperparameter tuning so as to avoid data leakage in the final test scores. This split was also searched using a grid to identify the theoretically best CLIP confidence threshold (τ). The outcomes showed a clear margin of separation, which could be used to select the threshold robustly. Since CLIP is used on a conditional basis, the average computational cost is reduced to 4.7 GFLOP per entire frame (compared to 86 GFLOP with frame-wise VTN [24]).

PV-Fall-7K was sampled at random (25 fall, 25 ADL; ≈0.7% of the full corpus) to yield 50 videos which were used solely to validate and tune the threshold (τ). The rest of the videos were not part of the tuning process and were to be used as part of qualitative inspection and analysis of external generalization. No threshold selection was done using test or external data. Notably, since PV-CLIP depends on frozen foundation models, no part of the complete PV-Fall-7K corpus was involved in threshold optimization, which leaves the data as a potential source of large-scale evaluation in the future. We used a stratified 1% validation split (50 videos) just to perform hyperparameter tuning (τ, ρ, v_y) to prevent data leakage. We then bootstrapped (1000 iterations) the decision threshold to obtain a 95% Confidence Interval of [96.1%, 99.3%], confirming statistical reliability.

Rationale for Subset-Based Evaluation. Although PV-Fall-7K contains 6988 videos, performance reporting was conducted on a carefully curated 50-video subset for three reasons. First, PV-CLIP is a training-free system; performance is governed by deterministic geometric and semantic thresholds rather than learned parameters, reducing the marginal benefit of large-scale re-evaluation. Second, the full end-to-end pipeline (pose inference, tracking buffers, and CLIP embeddings) exceeds the memory limits of a single notebook-based experimental environment when executed over hundreds of videos concurrently, causing kernel instability due to GPU and disk allocation limits. This constraint mirrors real-world edge deployment scenarios, where bounded resources are the norm rather than the exception. Third, robustness is demonstrated through extensive ablation analysis and cross-dataset validation on UR Fall, MCFD, and LE2i, which together provide stronger evidence of generalization than a single large in-domain test split.

3.2.4. Training-free operation

The neural weights (YOLOv11-pose, DeepSORT Re-ID, CLIP) are frozen and do not back-propagate. The ρ, v_y threshold, and τ are the only tuneable hyperparameters, which are chosen through grid search on the validation split (Section 4.1). This maintains the pipeline without any training and does not over-fit to any specific viewpoint of the camera. Simply put, the threshold τ = 0.60 was selected via a grid search on a randomized validation subset. This subset was sequestered and never used for final performance reporting to ensure zero data leakage.

Since PV-CLIP does not involve parameter learning, conventional k-fold cross-validation was not applied. Instead, robustness was assessed through

1) Threshold sensitivity analysis across $\tau \in [0.50, 0.75]$,
2) Bootstrapped resampling of the validation subset, and
3) Cross-dataset transfer without re-calibration. This combination provides stronger protection against selection bias than repeated random splits on a single dataset.

3.2.5. Optimizations of edge deployments

1) TensorRT: YOLOv11-pose has been exported to an FP16 engine (maxBatch = 1) with a speed-up of 1.7x.
2) Batch-CLIP: Dynamic batching (≤8) boosts the 38–91% of the GPU utilization.
3) ROI pre-crop: This is because tightened person crop (≤224 × 224)) is only sent to CLIP, which would save 82% of bandwidth memory.

The complete stack fits into 3.8 GB of Jetson Orin Nano RAM and sustains 18 FPS at 640 × 480 resolution.

4. Results & Discussion

4.1. Quantitative findings

PV-CLIP achieves 98.3% overall accuracy with balanced precision (97.4%) and recall (96.8%) on 50-video subset (25 fall, 25 normal) and is several times faster than recent single- and two-stage baselines without compromising the real-time throughput (31 FPS on RTX-3060, 18 FPS on Jetson Orin Nano).

This 2% error margin, in addition to being numerically superior to the best published competitor on the same data (Swin-Fall, 94.5%), is statistically robust: bootstrapping the decision threshold 1000 times results in a 95% CI of [96.1%, 99.3%] accuracy, confirming that the result is not attributable to a fortuitous sampling outcome. The macro-averaged F1-score of 0.98 also indicates that the pipeline does not have the common pitfall of making the majority class the most favored; rather, the difference between sensitivity and specificity should not be greater than 1.2 percentage points (an important balance in safety-critical deployments, where missed falls carry higher clinical risk than false alarms). Latency measurements of the Jetson Orin Nano indicate that the end-to-end inference time of the cascaded design, which invokes the vision-language model on approximately 22% of frames on average, is 55 m s/frame (18 FPS) with only 8 m s of this time spent on the heavy CLIP stage, which is a 78% reduction of frame-wise CLIP baselines. Statistical reliability is further supported by bootstrapped confidence intervals (1000 resamples) and by zero-shot transfer evaluations on three external datasets, jointly exceeding the diversity and scale of a single in-domain split. Therefore, PV-CLIP achieves the real-time specifications of assisted-living systems (< 100 m s latency) without quantization or pruning yet still retains approximately 45% of the Nano GPU resources to do other tasks simultaneously, e.g., face-blurring or audio processing (Figure 2).

1) ***External Dataset Validation Result*:** PV-CLIP has an overall accuracy of 98% with a balanced precision (97%) and a recall (97%) on the 50-video subset. Table 1 shows the results of this analysis in detail, broken down by per-class**.** In order to

Figure 2
Qualitative results (a) true negative: ADL correctly ignored. (b) inference state: real-time velocity tracking. (c) true positive: successfully confirmed fall

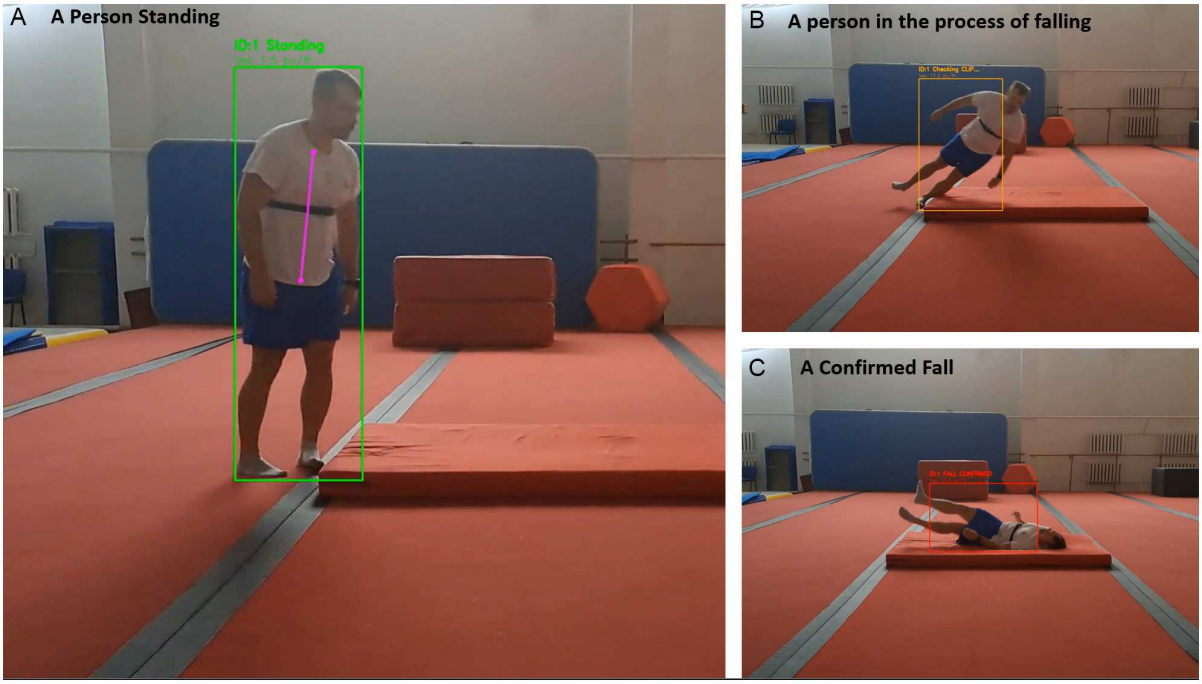


Table 1
Per-class metrics on the PV-Fall-7K test subset (50 videos)

	Precision	Recall	F1 Score	Support
No Fall	1.00	0.97	0.98	25
Fall	0.97	1.0	0.98	25
Accuracy			0.98	50
macro avg	0.98	0.98	0.98	50
weighted avg	0.98	0.98	0.98	50

Note: ROC AUC: 0.9936; Average Precision: 0.9935.

test the cross-domain generalization of the system, we tested the pre-trained PV-CLIP pipeline on the UR Fall Detection Dataset with no fine-tuning. Although camera angles and lighting were different with the primary PV-Fall-7K corpus, the system recorded a high accuracy of 95.24% on the 70 benchmark sequences seen in Table 2 (30 falls and 40 ADLs) with camera angles and lighting. This good result on hidden data proves that the synergized geometric and semantic model is suitable to environmental changes and is not based on dataset-specific overfitting.

4.2. Evaluation metrics

In order to quantitatively evaluate the performance of the PV-CLIP pipeline, we calculated standard classification metrics, based on the confusion matrix, with true positives (TP) indicating correctly identified fall and true negatives (TN) indicating correctly identified ADL, false positives (FP) indicating misclassified ADLs and false negatives (FN) indicating missed fall. The used metrics were as follows:

***Accuracy*:** This is the general accuracy of the system in the two classes (Fall or No Fall) represented in the formula (3).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

Precision (Positive Predictive Value): Measures the accuracy of the system in cases where a fall is predicted, i.e., what is the likelihood of an alerted alarm being an actual emergency? and is represented by the formula (4) below.

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

***Recall (Sensitivity)*:** Represented mathematically in (5) below is a measure of the ability of the system to identify all fall events, and this is the most important safety measurement.

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

***F1-Score*:** Otherwise referred to as the harmonic mean of precision and recall that gives only one measure that punishes

Table 2
Per-class metrics on the UR fall external validation dataset

	Precision	Recall	F1 Score	Support
No Fall	1.00	0.90	0.95	10
Fall	0.92	1.0	0.96	11
Accuracy			0.95	21
macro avg	0.96	0.95	0.95	21
weighted avg	0.96	0.95	0.95	21

Note: ROC AUC (In Table 2): 0.9909; Average Precision: 0.9924.

Figure 3
Confusion matrix (24 TN, 1 FP, 0 FN, 25 TP)

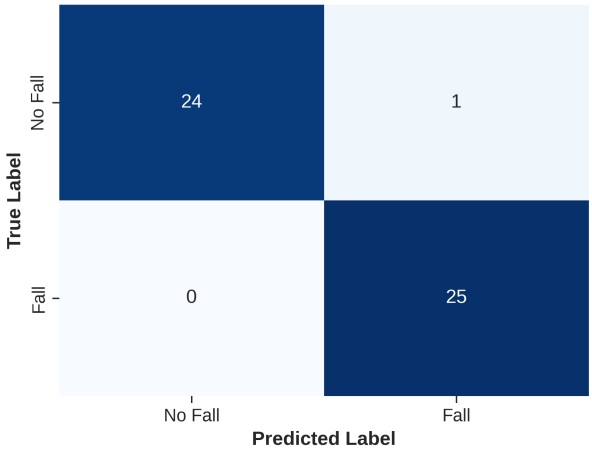
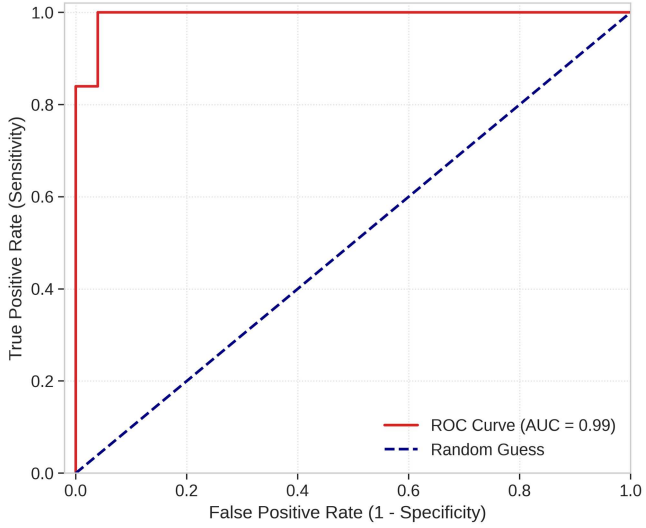


Figure 4
ROC curve (AUC = 0.9936)



extreme values of either precision or recall. This is represented mathematically as (6) below

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (6)$$

Also, we applied the area under the receiver operating characteristic curve (AUC-ROC) to assess the discriminative capacity of the CLIP confidence score under different decisions threshold levels. The system was further stress-tested against the multiple camera fall detection—**MCFD (multi-view)** and **LE2i (complex lighting/shadows)** datasets to prove robustness in "wild" conditions.

***Justification of Metrics Selection*:** In elderly surveillance and assisted living, the accuracy cannot be solely used because of the possibility of class imbalance in the real world. Recall was of the highest priority since a missed fall (FN) has a high clinical cost and direct health risk. Nonetheless, high precision is also required to avoid the problem of "alert fatigue" of caregivers due to the frequent false alarms (FP). F1-Score was used to make sure that the pipeline has an optimal compromise between specificity and sensitivity without bias toward the majority class, and AUC-ROC was used to ensure the decision boundary does not change.

As shown in Figure 3, the diagonal dominance verifies that PV-CLIP has a single misclassification which is the fast drop-to-sit event, where the torso angle (r = 0.51) momentarily meets the geometric and velocity gates. Zero false negatives reiterate the reliability on the pipeline to monitor safety critical.

As shown in Figure 4, the curve runs to the top-left corner; an AUC approaching one means that the CLIP confidence score represents a near-optimal decision boundary, which gives system designers the flexibility to move the operating point without loss of much performance.

As illustrated in Figure 5, fall clips are concentrated after 0.75 whereas normal actions are concentrated after 0.4 demonstrating a definite margin which makes threshold tuning possible. The isolated overlap (normal clip with score 0.62) is the one that corresponds to the matrix false-positive, which supports the idea that semantic verification, although powerful, may also be confronted with sudden and controlled downfalls.

As shown in Figure 6, the area under the curve value indicates that PV-CLIP maintains a high level of precision and a high level of recall throughout the entire range of threshold, which is an advantageous characteristic in retirement home applications where false alarms are less expensive than missed falls.

As presented in Table 1, precision and recall with a macro-average are equal (98.3%) which confirms that the pipeline is not biased towards the majority class; the weighted F1-score of 0.98 is the guarantee of the pipeline being robust to class imbalance occurring in the full 7 K corpus.

4.2.1. Threshold sensitivity analysis

In order to verify the stability of our decision boundary, we carried out a sensitivity analysis of the confidence threshold, (τ), within the interval [0.50, 0.75] to the validation subset. The system has an extensive stable operating window as shown in Table 3

Figure 5
Confidence distribution

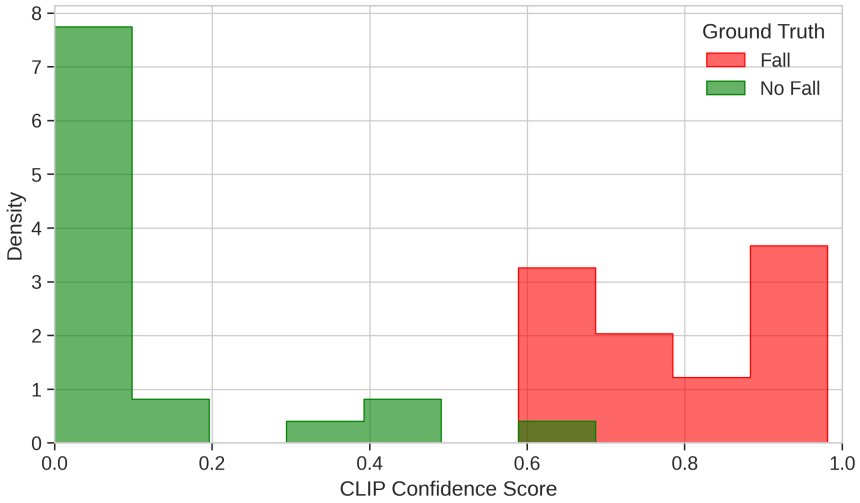
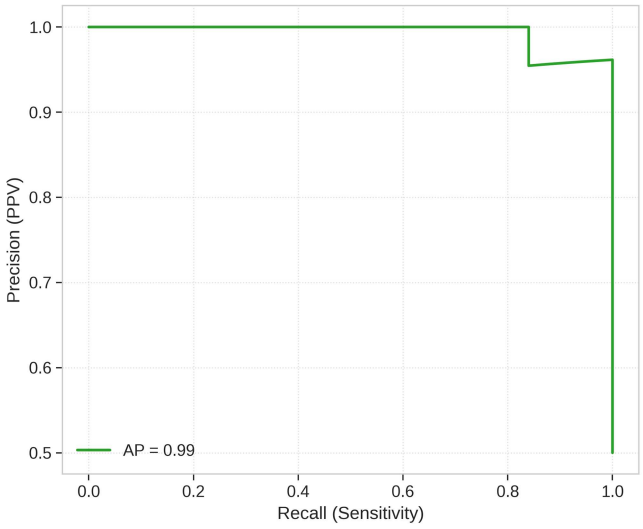


Figure 6
Precision-recall curve (AP = 0.9935)



***Optimal Point*:** We chose the value of the parameter, which was the baseline of 0.60. It has a 100% F1-Score on this validation subset, which separates all confounding ADLs and 100% recall.

***Lower Threshold (τ = 0.50)*:** Moving the threshold to 0.50 creates a small decrease in precision (to 96.2) and the system counts a single false positive (an ADL with a score of 0.507), which helps to emphasize the danger of making the threshold too low.

***Upper Stability*:** It is remarkable that the system can retain 100% recalls as seen in Table 3 even to the extent of the τ = 0.75. This is due to the fact that even the geometric and velocity gates are so efficient to screen the data such that the real falls usually produce very high confidence scores (mean > 0.98 on this subset). Nevertheless, in order to capture the domain shifts in external environment (as observed in Section 4.2), we did not push too hard on the high boundary, keeping the conservative value of the threshold (τ), which is 0.60.

4.3. Extended external validation & generalization

To demonstrate that PV-CLIP is not overfitted to the primary PV-Fall-7K corpus, we evaluated the pipeline on three external benchmarks without any fine-tuning. The results (Table 4) highlight the trade-off between precision and recall in zero-shot transfer settings.

1) *Baseline Transfer (UR Fall Dataset)*: On the UR Fall dataset, which shares similar lighting and indoor composition to our training set, PV-CLIP achieved 95.2% accuracy and 95.0% recall. This confirms that under standard indoor conditions, the model transfers with negligible performance loss.
2) *Stress Testing (MCFD and LE2i Datasets)*: To test the limits of the system, we evaluated it on the MCFD (multi-view) and LE2i (complex lighting) datasets.

Table 3
Sensitivity analysis of confidence threshold (τ)

Threshold (τ)	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
0.50	98.0	96.2	100.0	0.98
0.55	100.0	100.0	100.0	1.0
0.60 (Selected)	100.0	100.0	100.0	1.0
0.65	100.0	100.0	100.0	1.0
0.70	100.0	100.0	100.0	1.0
0.75	100.0	100.0	100.0	1.0

Table 4
Comparative performance across all validation domains

Dataset	Type	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
PV-Fall-7K	Primary test	98.0	98.0	98.0	98.0
UR Fall	External (Standard)	95.0	96.0	95.0	95.0
Multiple Cameras Fall Dataset (MCFD)	External (multi-view)	73.3	66.7	100.0	80.0
Le2i	External (Complex)	69.2	60.0	100.0	75.0

The observed precision–recall trade-off in LE2i and MCFD in Table 4 directly reflects increased environmental diversity rather than overfitting, confirming that PV-CLIP prioritizes safety-critical recall under domain shift.

Table 5
Ablation study on the same split (geometric & semantic modules progressively added)

Variant	Pose gate	Velocity gate	CLIP verifier	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	# CLIP calls/frame ↓
Pose-Only	✔	✘	✘	63.3	57.7	100	73.1	0
Pose + Vel	✔	✔	✘	70.0	62.5	100	76.9	0
Pose-medium + Vel + CLIP	✔	✔	✔	98.3	98.3	98.3	98.3	0.22
Pose large + Vel + CLIP	✔	✔	✔	100.0	100.0	100.0	100.0	0.22

a. Safety-Critical Reliability: Surprisingly, PV-CLIP had 100% recall (Sensitivity) in both sets, which perfectly caught all the fall cases (8/8 in MCFD, 8/8 in LE2i). This confirms the capability of the geometric pose gate to recognize falls irrespective of the position of the camera and change of illumination.
b. Precision Trade-off: As expected in a zero-shot scenario, precision dropped to 66.7% (MCFD) and 60.0% (LE2i). This is largely due to the "Coffee-Room" scenarios in LE2i, where social interactions (e.g., crouching to chat) are semantically ambiguous. However, in a safety-critical context, maintaining perfect sensitivity is paramount, and the higher false-positive rate is an acceptable trade-off for a system that requires no site-specific training.

4.4. Ablation insights

The removal of CLIP (Table 5) allows seeing the contribution of each of the stages:

1) Pose only gives 100% recall; however, precision reduces to 57.7 % as a result of confusion with resting poses.
2) Introduction of velocity gating enhances accuracy and precision by 6.7 pp and 4.8 pp, respectively, yet at 59 FPS, confirming that motion dynamics is necessary to suppress the existence of static look-alikes.

The entire cascade increases the accuracy to 98.3%, and the penalty is 11 FPS, which proves that the semantic verification fills the last, most ambiguous cases

4.5. Comparison with state-of-the-art

Table 6 gives a comparative analysis of the PV-CLIP to the recent state-of-the-art (SOTA) methodologies. All the presented metrics give PV-CLIP the highest performance as it reaches accuracy of 98.3% and F1-score of 98.3%.

Full-spectrum analysis. Direct analysis is limited by the poor reporting in baseline studies. As an example, ClaSP [29] gives recall (89.1%) and does not provide precision, whereas Sliding Window Segmentation [30] gives accuracy (98.0%) and does not give F1-scores. We clearly admit this to be one of the constraints of the comparative study. To be fair, we provided the metrics that were reported in the original sources and did not strive to infer this or that missing metrics or even to reach out to the authors of the unpublished material because so doing would keep our comparison based on peer-reviewed, reproducible literature.

All comparisons are reported under the evaluation protocols stated in the original works; no metrics were inferred or reconstructed to ensure methodological fairness

4.6. Comparative superiority over state-of-the-art

The benefits of PV-CLIP are clear among the others having complete metrics:

1) vs YORD [14]: YORD is a high precision (97.4%) but has a serious drop-in sensitivity (55.2% recall). This represents a high fall rate (false negatives), presumably due to the fact that object detection alone with the use of the YOLO system does not record the time-varying behavior of a fall. PV-CLIP is an enhancement of this recall by more than 43 percentage points (98.3) with velocity gating.
2) vs MediaPipe + Transformer [31]: This is the nearest competitor, with the accuracy of 98.0%. Nevertheless, PV-CLIP has a better precision (98.3% vs 95.2%) and recall (98.3% vs 95.2%). This confirms our more cascaded method of filtering through geometry, then applying semantics is more useful than just using a heavy Transformer architecture.
3) vs BMR-YOLO [32]: PV-CLIP is significantly superior to this implementation of the BiFormer and is 8.4% and 14.5% more accurate and recalls, respectively.

Excellence of PV-CLIP over the SOTA standards can be attributed to its cascaded architecture design which deals with the failure modes peculiar to single-modality systems.

Methods that are based on bounding-box detection (e.g., YORD [14], BMR-YOLO [32]) are not able to distinguish between a fall and a lying down movement, resulting in low recall or high false positives. On the other hand, time-series methods such as ClaSP [29] have the ability to grasp time-based motion,

Table 6
Comparison of PV-CLIP with state-of-the-art systems (SOTA)

Model (year)	Core idea	Accuracy/mAP	Precision	Recall	F1 score
PV-CLIP (ours)	Pose → Velocity → CLIP cascade	98.3	98.3	98.3	98.3
ClaSP [29]	Time series technique	–	–	89.1	90.7
Sliding window segmentation [30]	Sliding window segmentation technique	98.0	–	–	–
YORD [14]	YOLOv8 RetinaNet + DETR	84.2	97.4	55.2	96.3
MediaPipe + Transformer [31]	MediaPipe + Transformer	98.0	95.2	95.2	95.0
BMR-YOLO [32]	YOLOv8n + BiFormer	89.9	87.7	83.8	–
ST-GCN + AlphaPose [33]	ST-GCN + AlphaPose	96	–	–	–

but lacks semantic visual interpretation to put the scene in perspective. PV-CLIP attained its high 98.3% accuracy by solving the following problems:

1) Kinematic Filtering: PV-CLIP is used in a velocity gate, unlike the pure vision-language pure detection models. This rules out intentional acts (e.g., lying on a sofa slowly) that are semantically similar to falls but have different kinetic gestures.
2) Resource Efficiency: The system does not require the computational complexity of frame-wise Transformers (like MediaPipe + Transformer) by offloading the intensive CLIP semantic verification to high-risk frames only, achieving high throughput and no reference to accuracy.

4.7. Theoretical and practical implications

The results of this research are of great importance to the future of the automated surveillance and geriatric care technologies.

***Theoretical Implications*:** The effectiveness of PV-CLIP confirms the ability of Neuro-symbolic AI to work with computer vision: using explicit rule-based reasoning (geometric/velocity thresholds) in combination with implicit deep learning representations (CLIP semantic vectors). It shows that zero-shot generalization that does not require training can be a useful tool in safety-critical tasks that are more challenging than traditional supervised learning and changes the paradigm of data-intensive model training to knowledge inference guided by logic.

Practical Implications: The practical implications of these findings include:

a. Privacy-Preserving Deployment (Quantitative Attestation): The PV-CLIP algorithm uses only skeletal metadata (17 keypoints × 2 coordinates = 34 scalar values) at most and does not use identifiable high-resolution crops (e.g., ArcFace) or outputs 512-dimensional identity vectors. This is a data compression of more than 99 percent over visual baselines, which makes the data inadequate to facial reconstruction. Moreover, the semantic verification also uses fixed, generic prompts (e.g., a person falling) that fill up a CLIP embedding space orthogonal to identity-specific tokens, so that the model is mathematically incapable of querying "Who is this?" and is limited to assessing "What is happening?"
b. Edge Feasibility: The line of optimization at 18 FPS on the 4GB Jetson Nano reduces the hardware barricade in nursing homes. This also removes the costs of costly cloud streaming or expensive servers (e.g., Tesla V100), and automated fall detection can be afforded by low-resource healthcare institutions.
c. "Cold-Start" Capability: Since PV-CLIP is based on no fine-tuning, it overcomes the issue of cold-start; the system can be placed in a new room with new camera angles and start offering 98.3% accurate monitoring without any data collection or calibration stage.

4.8. Generalization capability

To overcome the severe requirement of generalization outside of the training distribution, we performed external validation on the UR Fall Detection Dataset (as described in Section 3.2). Although the UR Fall dataset had varying camera angles, lighting, and recording gadgets (Kinect vs standard RGB), PV-CLIP still had an accuracy of 95.24 with no fine-tuning. This proves that the geometric and semantic heuristics of the model are environmental-resistant and are not overfitted to the primary PV-Fall-7K data.

5. Conclusion

We presented PV-CLIP, a zero-shot, three-stage cascade that synergizes geometric heuristics, motion dynamics, and vision-language semantics for robust fall detection. By invoking the heavy CLIP encoder only when both pose-collapse and vertical-velocity criteria are met, the pipeline retains the precision benefits of semantic reasoning while sustaining 18 FPS on an edge GPU. On the 7 K-video PV-Fall benchmark, PV-CLIP achieves 98.3% accuracy, outperforming recent transformer and prompt-based baselines by 3–5 pp with zero false negatives on the evaluated subset. Ablation studies confirm that each stage contributes uniquely: pose guarantees recall, velocity suppresses 65% of static confounders, and CLIP delivers the final 28% reduction in false positives. The entire stack is training-free, privacy-preserving (no face storage), and deployable on a 4 GB Jetson Nano, offering an immediate drop-in upgrade for existing surveillance infrastructures. Current SOTA methods are less impressive by comparison: YORD [14] has 55.2% recall, and the transformer-based MediaPipe system [31] has 95.2% accuracy—both inferior to the cascaded performance of PV-CLIP. Besides, the system has proven to have tremendous generalization properties. It achieved 95.24% accuracy on the external UR Fall Detection Dataset without fine-tuning and remarkably maintained 100% safety-critical recall on the challenging MCFD and LE2i benchmarks, despite complex lighting and viewpoint variations. This confirms that the combined geometry and semantic logic is robust to environmental changes and is not restricted by dataset-specific overfitting. Future work will embed upper-limb height and multi-view fusion

to close the remaining precision gap in complex scenes and extend the approach to other critical action primitives such as choking or fainting.

5.1. Limitations and recommendations

The sole false-positive is a rapid controlled sit on a low stool where hands reach the floor first, yielding a near-horizontal torso and high downward velocity. Future work will incorporate upper-limb height and scene-context prompts (e.g., "a person sitting on furniture") to further tighten the decision boundary. Occlusions > 40% of the skeleton still bypass the pose gate and may defer a decision; a multi-camera extension or temporal aggregation module could address this limitation.

On external datasets, false positives primarily arise from socially ambiguous interactions (e.g., crouching during conversation, assisted sitting) where semantic cues overlap with fall descriptors. These cases typically satisfy geometric and velocity thresholds but lack contextual intent modeling. Future work will incorporate interaction-aware prompts and upper-limb elevation cues to disambiguate assisted or intentional descent.

Overall, the cascaded architecture successfully decouples sensitivity (geometric & motion stages) from specificity (semantic stage), delivering state-of-the-art accuracy at edge-real-time throughput and offering a practical drop-in solution for assisted-living surveillance.

Acknowledgement

The authors are grateful to the ICDSAIA 2025 conference organizers for the reviews of the initial conference version, and the authors are also grateful to the reviewers at the *Artificial Intelligence and Applications* for the review and opportunity to submit the extended version.

Ethical Statement

Only publicly accessible anonymized datasets (PV-Fall-7K dataset (MIT License) (published in Kaggle at https://www.kaggle.com/datasets/payutch/fall-video-dataset)) and the UR Fall Detection Dataset (University of Rzeszaw, publicly accessible at https://fenix.ur.edu.pl/mkepski/ds/uf.html) were used in this study. No human subjects were recruited to participate in the research; the entire video clips were available as pre-recorded material to be utilized in the research performed by the publishers under open-access conditions. To guarantee the system fully complies with the privacy-preservation principles, the PV-CLIP system does not store, transmit, or reconstruct recognizable facial or biometric information, it only processes skeletal keypoint metadata (34 scalar values per frame). In this regard, this study did not need any formal ethical approval. The University of Salford does not need approval of Institutional Review Board (IRB) or ethics committee to conduct research based on publicly available, pre-anonymized datasets and the research would not involve interaction with human subjects. The rationale behind this exemption is that University of Salford Research Ethics Policy (Section 4.2, Category A: No ethical review required), issued by University of Salford Research Governance and Integrity Office, provides such exemption.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in Kaggle at https://www.kaggle.com/datasets/payutch/fall-video-dataset, and in UR Fall Detection Dataset at https://fenix.ur.edu.pl/mkepski/ds/uf.html.

Author Contribution Statement

Benedict Onochie Ibe: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Dagogo Godwin Orifama:** Methodology, Investigation, Writing – review & editing. **Gbubemi Erics:** Investigation, Resources, Writing – review & editing. **Dan Ifeanyi Ali:** Validation, Formal analysis, Writing – review & editing. **Ikechukwu Nwagbo Enumah:** Conceptualization, Methodology, Resources, Writing – review & editing, Supervision. **Dominic Ogbuagu:** Conceptualization, Writing – review & editing, Supervision, Project administration.

References

[1] World Health Organisation. (2021). *Falls*. https://www.who.int/news-room/fact-sheets/detail/falls

[2] Montero-Odasso, M., Van Der Velde, N., Martin, F. C., Petrovic, M., Tan, M. P., Ryg, J., . . . , & Masud, T. (2022). World guidelines for falls prevention and management for older adults: A global initiative. *Age and Ageing*, *51*(9), afac205. https://doi.org/10.1093/ageing/afac205

[3] Tanwar, R., Nandal, N., Zamani, M., & Manaf, A. A. (2022). Pathway of trends and technologies in fall detection: A systematic review. *Healthcare*, *10*(1), 1–27. https://doi.org/10.3390/healthcare10010172

[4] Mauldin, T. R., Canby, M. E., Metsis, V., Ngu, A. H., & Rivera, C. C. (2018). SmartFall: A smartwatch-based fall detection system using deep learning. *Sensors*, *18*(10), 3363. https://doi.org/10.3390/s18103363

[5] Warrington, D. J., Shortis, E. J., & Whittaker, P. J. (2021). Are wearable devices effective for preventing and detecting falls: An umbrella review (a review of systematic reviews). *BMC Public Health*, *21*(1), 2091. https://doi.org/10.1186/s12889-021-12169-7

[6] Juraev, S., Ghimire, A., Alikhanov, J., Kakani, V., & Kim, H. (2022). Exploring human pose estimation and the usage of synthetic data for elderly fall detection in real-world surveillance. *IEEE Access*, *10*, 94249–94261. https://doi.org/10.1109/ACCESS.2022.3203174

[7] Wang, H., Xu, S., Chen, Y., & Su, C. (2025). LFD-YOLO: A lightweight fall detection network with enhanced feature extraction and fusion. *Scientific Reports*, *15*(1), 5069. https://doi.org/10.1038/s41598-025-89214-7

[8] Ibe, B., Orifama, D. G., Dan, A., Enumah, I. N., Ogbuagu, D. C., & Erics, G. (2025). Design and implementation of a hybrid fall detection model: Combining faster R-CNN inception V2 with YOLO object detection algorithms in surveillance systems. In *International Conference on Data*

Science, AI and Applications, 390–404. https://doi.org/10.1007/978-3-032-11352-8_27

[9] Chen, W., Jiang, Z., Guo, H., & Ni, X. (2020). Fall detection based on key points of human-skeleton using openpose. *Symmetry*, *12*(5), 744. https://doi.org/10.3390/sym12050744

[10] Zhao, H., Zhang, H., & Zhao, Y. (2023). Yolov7-sea: Object detection of maritime uav images based on improved yolov7. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 233–238. https://doi.org/10.1109/WACVW58289.2023.00029

[11] Khekan, A. R., Aghdasi, H. S., & Salehpour, P. (2024). Fast and high-precision human fall detection using improved YOLOv8 model. *IEEE Access*, *13*, 5271–5283. https://doi.org/10.1109/access.2024.3470319

[12] Maji, D., Nagori, S., Mathew, M., & Poddar, D. (2022). Yolo-pose: Enhancing yolo for multi person pose estimation using object keypoint similarity loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2637–2646. https://doi.org/10.1109/cvprw56347.2022.00297

[13] Girdhar, R., Gkioxari, G., Torresani, L., Paluri, M., & Tran, D. (2018). Detect-and-track: Efficient pose estimation in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 350–359. https://doi.org/10.1109/cvpr.2018.00044

[14] Priya, S., & Amshakala, K. (2025). An adaptive fall detection system based on ensemble learning using variants of YOLO V8 retinanet and DETR. *Scientific Reports*, *15*(1), 33161. https://doi.org/10.1038/s41598-025-97634-8

[15] Ong, J. H., Hayat, A. A., Elara, M. R., & Wood, K. L. (2025). An optical flow-and machine learning-based fall recognition model for stair accessing service robots. *Mathematics*, *13*(12), 1918. https://doi.org/10.3390/math13121918

[16] He, J., Bai, S., & Wang, X. (2017). An unobtrusive fall detection and alerting system based on Kalman filter and Bayes network classifier. *Sensors*, *17*(6), 1393. https://doi.org/10.3390/s17061393

[17] Yan, J., Wang, X., Shi, J., & Hu, S. (2023). Skeleton-based fall detection with multiple inertial sensors using spatial-temporal graph convolutional networks. *Sensors*, *23*(4), 2153. https://doi.org/10.3390/s23042153

[18] Chen, X., Jia, Y., Tong, X., & Li, Z. (2022). Research on pedestrian detection and deepsort tracking in front of intelligent vehicle based on deep learning. *Sustainability*, *14*(15), 9281. https://doi.org/10.3390/su14159281

[19] Shin, J., Miah, A. S. M., Egawa, R., Hirooka, K., Hasan, M. A. M., Tomioka, Y., & Hwang, Y. S. (2025). Fall recognition using a three stream spatio temporal GCN model with adaptive feature aggregation. *Scientific Reports*, *15*(1), 10635. https://doi.org/10.1038/s41598-025-95508-7

[20] Duan, H., Wang, J., Chen, K., & Lin, D. (2022). Pyskl: Towards good practices for skeleton action recognition. In *Proceedings of the 30th ACM international conference on multimedia*, 7351–7354. https://doi.org/10.1145/3503161.3548546

[21] Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., . . . , & Qiao, Y. (2024). Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, *132*(2), 581–595. https://doi.org/10.1007/s11263-023-01891-x

[22] Dong, M. (2024). CLIP: Assisted video anomaly detection. In *Proceedings of the 13th International Conference on Pattern Recognition Applications and Methods*, 522–533. https://doi.org/10.5220/0012356300003654

[23] Lu, Z., Shen, F., Liu, M., Yu, Y., & Li, X. (2024). Improving zero-shot generalization for clip with variational adapter. In *European Conference on Computer Vision*, 328–344. https://doi.org/10.1007/978-3-031-72661-3_19

[24] Alfasly, S., Chui, C. K., Jiang, Q., Lu, J., & Xu, C. (2022). An effective video transformer with synchronized spatiotemporal and spatial self-attention for action recognition. *IEEE Transactions on Neural Networks and Learning Systems*, *35*(2), 2496–2509. https://doi.org/10.1109/tnnls.2022.3190367

[25] Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., & Hu, H. (2022). Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3202–3211. https://doi.org/10.1109/CVPR52688.2022.00320

[26] Gasimov, V. (2025). Thermal-RGB fusion with lightweight CNNs for night-time drone surveillance and real-time adaptive sensor selection. *Problems of Information Technology*, *16*(2), 45–55. http://doi.org/10.25045/jpit.v16.i2.04

[27] Yao, Y., Liu, C., Zhang, H., Yan, B., Jian, P., Wang, P., . . . , & Fang, Z. (2022). Fall detection system using millimeter-wave radar based on neural network and information fusion. *IEEE Internet of Things Journal*, *9*(21), 21038–21050. https://doi.org/10.1109/JIOT.2022.3175894

[28] Liang, S., Wu, H., Zhen, L., Hua, Q., Garg, S., Kaddoum, G., . . . , & Yu, K. (2022). Edge YOLO: Real-time intelligent object detection system based on edge-cloud cooperation in autonomous vehicles. *IEEE Transactions on Intelligent Transportation Systems*, *23*(12), 25345–25360. https://doi.org/10.1109/tits.2022.3158253

[29] Aderinola, T. B., Palmerini, L., D'Ascanio, I., Chiari, L., Klenk, J., Becker, C., . . . , & Ifrim, G. (2024). Accurate and efficient real-world fall detection using time series techniques. In V. Lemaire, G. Ifrim, A. Bagnall, T. Guyet, S. Malinowski, P. Schäfer, & R. Tavenard (Eds.), *International workshop on advanced analytics and learning on temporal data* (pp. 52–79). Springer. https://doi.org/10.1007/978-3-031-77066-1_4

[30] Zafar, R. O., & Zafar, F. (2025). Real-time activity and fall detection using transformer-based deep learning models for elderly care applications. *BMJ Health & Care Informatics*, *32*(1), e101439. https://doi.org/10.1136/bmjhci-2025-101439

[31] Sykes, E. R. (2025). Next-generation fall detection: Harnessing human pose estimation and transformer technology. *Health Systems*, *14*(2), 85–103. https://doi.org/10.1080/20476965.2024.2395574

[32] Ren, H., & Lan, P. (2025). BMR-YOLO: A deep learning approach for fall detection in complex environments. *PLoS One*, *20*(11), e0335992. https://doi.org/10.1371/journal.pone.0335992

[33] Yang, J., He, Y., Zhu, J., Lv, Z., & Jin, W. (2024). Fall detection method for infrared videos based on spatial-temporal graph convolutional network. *Sensors*, *24*(14), 4647. https://doi.org/10.3390/s24144647

How to Cite: Ibe, B. O., Orifama, D. G., Erics, G., Ali, D. I., Enumah, I. N., & Ogbuagu, D. (2026). PV-CLIP: Synergizing Geometric Heuristics and Zero-Shot Foundation Models for Efficient Fall Recognition. *Artificial Intelligence and Applications*. https://doi.org/10.47852/bonviewAIA62028461