LOCAL VS CLOUD VISION–LANGUAGE MODELS FOR FALL DETECTION IN SURVEILLANCE VIDEOS

Noor Jamal Alkhateeb[1], Aguares Maveron[1], Prateek Mishra[1], Luqman Ali[2]

[1] School of IT, Murdoch University Dubai, Dubai P.O. Box 500700, United Arab Emirates
[2]Emirates Centre for Mobility Research, United Arab Emirates University, Al Ain 15551, United Arab Emirates
E-MAIL: noor.alkhateeb@murdoch.edu.au

Abstract:

Falls among the elderly pose significant health risks and require timely detection to prevent serious consequences. Recent Vision–Language Models (VLMs) demonstrate strong multimodal reasoning capabilities and may offer a flexible alternative to traditional fall-detection pipelines. However, their effectiveness in video-based safety monitoring remains largely unexplored. This paper investigates the preliminary application of both local and cloud-based VLMs for fall detection from surveillance videos. We compared a locally deployed state-of-the-art model, Qwen2.5-VL-7B, with two API-based models, GPT-4o and Claude Sonnet 4.5, and examined their ability to identify fall-related patterns. Experiments were conducted on the UR Fall Detection dataset, consisting of 70 labeled video sequences (30 falls and 40 non-fall activities). Although the models operated on sampled frames rather than on full-motion data, our results indicated that VLMs can effectively detect fall scenarios, yielding interpretable outputs that could support healthcare professionals. Results showed that the local model achieved the highest accuracy (0.86), achieving the highest performance under extended temporal frame coverage. We further analyzed the impact of frame sampling on performance and demonstrated that temporal frame coverage significantly affects model accuracy. The study also discussed potential limitations, including computational demands and reduced accuracy compared to specialized models.

Keywords:

Vision-Language Models; Fall Detection; Explainable AI; Elderly Care

1. Introduction

Incidents of unintended ground-level collapses in older populations remain a significant global public health concern, contributing to substantial physical harm, diminished autonomy, and rising health care expenditures [1]. The World Health Organization reports that nearly 37.3 million such incidents necessitating medical intervention occur each year, with individuals aged 65 and above being particularly vulnerable [2]. These episodes frequently cause bodily harm, including bone fractures and traumatic brain injuries. They can also trigger psychological issues such as a persistent fear of falling, which further impairs mobility and overall well-being. As the global population continues to age, there is an increasing demand for intelligent automated systems that enable independent living while ensuring timely intervention during emergencies. Traditional fall detection technologies have commonly utilized body-mounted instruments, such as accelerometers and gyroscopic sensors [3], or employed camera-based approaches that incorporate visual observation and pose analysis techniques [4]. While these approaches have demonstrated effectiveness, they are often hindered by challenges such as user non-compliance (especially with wearable devices), elevated false alarm rates, and limited ability to interpret complex environmental contexts. Recent progress in the development of expansive VLMs has opened new frontiers in human activity recognition by enabling contextual reasoning from visual scenes. In healthcare, particularly in elder care, fall detection is a critical task, as timely and accurate identification can prevent serious injuries or fatalities. VLMs combine visual perception with language-based reasoning, enabling nuanced scene interpretation and natural language explanations. These models can process images and generate structured descriptions that explain their decisions, which is particularly valuable in healthcare applications where interpretability is critical for clinical acceptance [5]. In this paper, we investigate the application of VLMs for fall detection within smart elder care environments. By leveraging RGB video sequences from the UR Fall Detection Dataset [6] and prompting

different VLMs with structured prompts, we assess the models' ability to detect falls and human activity accurately. Our goal is to evaluate not only the accuracy of fall detection but also the transparency and usefulness of the model outputs for human care givers and developers. This aligns with ongoing efforts in healthcare AI to develop systems that are not only performant but also trustworthy and aligned with human values [7].

Unlike prior fall detection studies that primarily focus on specialized deep learning architectures, this work investigates the practical deployment trade-offs of general-purpose VLMs for surveillance-based healthcare monitoring. In particular, we analyze how temporal frame coverage affects fall recognition performance and compare local versus API-based deployment scenarios under realistic operational constraints. The main contributions of this work are: Evaluation of modern VLMs for fall detection in surveillance videos. Comparison between local and API-based VLM deployment scenarios, and analysis of the impact of temporal frame coverage on fall detection performance. The code of the evaluation and full experiment is available on the following link: https://github.com/ashkree/VLM-Based-Fall-Detection

2. Related Work

Fall detection has been extensively explored within smart healthcare, with a significant body of work focusing on three primary areas: sensor-based methods, vision-based techniques, and hybrid systems that combine both. Each paradigm has contributed substantially to intelligent monitoring in eldercare while also exposing key limitations.

Sensor-based approaches, particularly those relying on wearable devices such as accelerometers and gyroscopes, remain prevalent due to their low cost and ease of integration [8]. These devices inferred falls based on sudden variations in velocity or orientation [3]. However, they are often limited by user non-compliance, especially among elderly individuals who may forget or resist wearing them, and their inability to capture contextual information about the environment.

Vision-based methods emerged to overcome these challenges, employing RGB or depth cameras to monitor body posture and motion. The work of Nait-Charif and McKenna [4] is one of the earliest works in this domain. They detected anomalies in home environments by utilizing activity summarization. Subsequent models also incorporated deep learning architectures such as convolutional neural networks(CNNs) and long short-term memory(LSTM) networks [9], demonstrating strong performance in classifying fall-related events. However, these models are often regarded as black-box systems, lacking transparency in decision-making, a major barrier in clinical applications where explainability is crucial.

Hybrid systems attempted to merge the strengths of both paradigms by combining sensor and visual data. These approaches enhanced detection accuracy and contextual awareness, especially in dynamic environments [10]. Nevertheless, they introduced additional system complexity and integration costs, which may hinder deployment in large-scale or resource-constrained settings.

More recently, the emergence of VLMs VLMs such as CLIP [11], Flamingo [12], and Gemini has transformed human activity recognition. These models combine visual perception with natural language reasoning, enabling interpretable scene understanding and human-readable justifications. Their capacity to jointly process visual and textual information makes them attractive for healthcare, where interpretability and context awareness are critical for trust and clinical acceptance [5].

Initial applications of VLMs in healthcare include medical image captioning [13], radiology report generation [2], and skin disease recognition [14]. In addition to the authors in [2], who evaluated State-of-the-art VLMs in fall detection. However, they used the RGB images and input, not the RGB Videos. Despite their limited application to fall detection so far, their ability to identify postures, detect human presence, and generate natural language justifications makes them well-suited for intelligent eldercare monitoring systems.

Our work builds on these foundations by evaluating structured prompting strategies with large-scale VLMs on static RGB videos from the UR Fall Detection Dataset. We assess not only their fall classification accuracy but also the interpretability and limitations of their outputs. This represents a first step toward understanding the trade-offs between general-purpose VLMs and domain-specific fall detection architectures.

2.1. Dataset

The experiments in this study were conducted using the UR Fall Detection Dataset, a publicly available benchmark dataset widely used for evaluating video-based fall detection systems. The dataset contains RGB video recordings of simulated fall events and normal daily activities captured in indoor environments. The dataset consists of 70 video sequences, including 30 fall events and 40 non-fall activities (activities of daily living, ADL). The fall sequences represent various types of falls such as forward falls, backward falls, and lateral falls, while the non-fall activities include common movements such as walking, sitting, bending, and lying down. These activities are intentionally included to create realistic scenarios that may visually resemble falls and therefore challenge automated detection

systems. Each video sequence is labeled at the video level with a binary class indicating whether a fall occurred. The videos were recorded in controlled indoor environments with different subjects performing the activities, providing variability in posture, motion patterns, and scene conditions.

TABLE 1. UR Fall Detection Dataset distribution

Class	Number Of Videos
Fall	30
No Fall (ADL)	40
Total	**70**

To evaluate VLMs, frames were sampled from each video sequence and provided to the models as visual inputs. The models were then prompted to determine whether the sequence corresponds to a fall event or a non-fall activity, producing a binary classification output. This dataset is suitable for evaluating VLM-based fall detection because it contains realistic human activities that can visually resemble falls, enabling assessment of the models' ability to distinguish between hazardous events and normal movements.

2.2. Methodology

We carried out a structured evaluation leveraging the UR Fall Detection Dataset [7], which includes over 70 RGB videos capturing diverse human postures and activities within indoor settings. Each sequence was independently assessed using a local VLM model: Qwen2.5-VL-7B, and two API-based models, GPT-4o and Claude Sonnet 4. The proposed evaluation pipeline processes video sequences through three main stages: frame sampling, VLM inference, and decision extraction. The overall workflow is illustrated in Figure 1. Given a video sequence, a subset of frames is sampled and provided to the VLM along with a prompt requesting the model to determine whether a fall occurred in the scene. The model produces a textual response, which is then converted into a binary classification output.

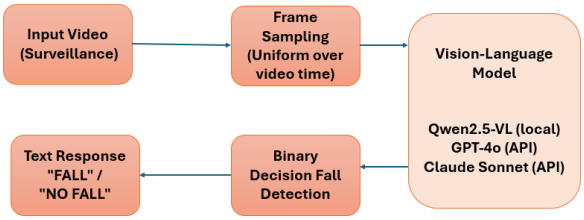

FIGURE 1. Proposed pipeline for fall detection using VLMs. Video sequences are sampled into representative frames, which are provided to the VLM together with a textual prompt.

2.3. Vision–Language Models

Three state-of-the-art VLMs were evaluated:

- Qwen2.5-VL (7B) – deployed locally for offline inference.
- GPT-4o – accessed through an API-based inference service.
- Claude Sonnet 4.5 – evaluated using a cloud API interface.

These models were selected to represent two deployment scenarios:

- Local models, which provide privacy-preserving inference and allow processing of larger numbers of frames.
- API-based models, which provide powerful reasoning capabilities but may impose limits on the number of visual inputs per request.

2.4. Frame Sampling Strategy

Since VLMs typically operate on images rather than continuous video streams, frames must be sampled from each video sequence before inference. For each video in the UR Fall Detection Dataset, frames were extracted and uniformly sampled across the entire duration of the video. A high-frame evaluation setting was also considered. The local Qwen2.5-VL model allows processing a larger number of visual tokens. Therefore, an extended evaluation was conducted where up to 768 frames were provided to the models to capture more temporal information from the video. This comparison allows us to analyze the impact of temporal frame coverage on fall detection performance.

2.5. Evaluation Metrics

To evaluate and compare the performance of all models, we compute five standard classification metrics: accuracy, precision, sensitivity (recall), and F1-score. These metrics are derived from the confusion matrix consisting of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

Accuracy Accuracy measures the overall correctness of the model and is defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Precision Precision measures the proportion of correctly identified falls among all predicted falls. High precision reduces the likelihood of false alarms in real-world scenarios:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Sensitivity (Recall) Sensitivity, also known as recall, measures the proportion of actual falls that are correctly detected by the model. High sensitivity is critical in fall detection systems, as missed falls may be life-threatening:

$$Sensitivity = \frac{TP}{TP + FN} \tag{3}$$

F1-score The F1-score is the harmonic mean of precision and sensitivity, capturing the balance between false positives and false negatives:

$$F1 - score = 2 \times \frac{Precision \times Sensitivity}{Precision + Sensitivity} \tag{4}$$

In addition to these metrics, we also measure the average inference time for the Qwen model and the average cost per sample for ChatGPT and Claude. where (True positive) refers to the number of correctly identified falls (True negative) refers to the number of correctly identified ADLs (False positive) refers to the number of incorrectly identified falls i.e. the model classifies the video as a fall but the actual class is ADL (False negative) refers to the number of incorrectly identified ADL i.e. the model classifies the video as an ADL but the actual class is fall To ensure consistent evaluation across models, a standardized prompt was used during inference.

3 Results and Discussion

3.1 Overall Performance

We evaluated three VLMs for fall detection using the UR Fall Detection Dataset, which contains 70 videos consisting of 30 fall events and 40 normal daily activities. Performance was measured using accuracy, precision, recall, and F1-score derived from the confusion matrix. Table 2 and Figure 2 summarize the performance of the evaluated models. The locally deployed Qwen2.5-VL model achieved the highest overall accuracy of 0.86, demonstrating strong capability for recognizing fall events from video sequences. The model correctly identified 28 out of 30 fall events, achieving a recall of 0.93 for the fall class. However, it produced 8 false positives among non-fall activities, resulting in a slightly lower recall for the NO-FALL class (0.79). The API-based GPT-4o model achieved an accuracy of 0.83. Notably, GPT-4o correctly detected all fall events (recall = 1.00), which is desirable for safety-critical monitoring systems where missed falls must be minimized. However, the model produced 12 false positives, reducing its recall for non-fall activities. Claude Sonnet 4.5 demonstrated lower overall performance, achieving an accuracy of 0.70. Although the model also achieved perfect fall recall (1.00), it misclassified a significant number of non-fall activities as falls, producing 21 false positives and resulting in a low NO-FALL recall of 0.47. These results indicate that VLMs tend to prioritize fall detection sensitivity, sometimes at the cost of higher false alarm rates. For the confusion matrix analysis, Qwen demonstrated balanced performance across both classes, with relatively few missed falls and moderate false alarms. GPT-4o exhibited conservative behavior by classifying ambiguous activities as falls, leading to perfect fall recall but increased false positives, and Claude demonstrated difficulty distinguishing fall events from certain normal activities, particularly movements such as sitting or bending. tables 3 4 and 5 demonstrate the confusion matrix for all models.

TABLE 2. Performance comparison of evaluated VLMs for fall detection on the UR Fall Detection Dataset

Model	Frames	Accuracy	Fall Recall	No-Fall Recall	Fall Precision	No-Fall Precision
Qwen 2.5-VL	native video (768)	0.86	0.93	0.79	0.82	0.92
GPT-4o	8	0.83	1.00	0.70	0.71	1
Claude Sonnet 4.5	8	0.70	1.00	0.47	0.59	1

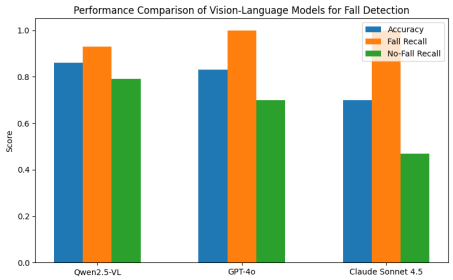

FIGURE 2. Performance comparison of evaluated VLMs for fall detection on the UR Fall Detection Dataset. The local Qwen2.5-VL model achieves the highest overall accuracy, while GPT-4o and Claude Sonnet 4.5 demonstrate perfect recall for fall events but produce higher false positives for non-fall activities

TABLE 3. Classification metrics for Qwen2.5-VL, including precision, recall, and F1-score for the FALL and NO-FALL classes.

Qwen2.5-VL		
	Predicted Fall	Predicted No Fall
Actual Fall	28	2
Actual No Fall	8	32

TABLE 4. Classification metrics for GPT-4o. The model achieves perfect recall for fall events but produces several false positives for non-fall activities.

GPT-4o		
	Predicted Fall	Predicted No Fall
Actual Fall	30	0
Actual No Fall	12	28

TABLE 5. Classification metrics for Claude Sonnet 4.5. The model detects all fall events but struggles to distinguish non-fall activities, resulting in a high number of false positives.

Claude Sonnet 4.5		
	Predicted Fall	Predicted No Fall
Actual Fall	30	0
Actual No Fall	21	19

TABLE 6. Performance comparison under different frame budgets. Limiting the number of frames significantly affects the performance of the local VLM model.

Model	Frames	Accuracy
Qwen2.5-VL	768	0.86
Qwen2.5-VL	8	0.60
GPT-4o	8	0.83
Claude Sonnet 4.5	8	0.70

3.2 Impact of Frame Budget

An important observation in this study concerns the effect of temporal frame coverage on model performance. The API-based models were evaluated using 8 uniformly sampled frames per video, due to practical constraints such as input limits and inference cost. In contrast, the local Qwen2.5-VL model supports native video input and can process a significantly larger temporal context (approximately 768 frames). To analyze the impact of temporal information, we conducted an additional experiment where Qwen was restricted to the same 8-frame input used for the API-based models. Under this constraint, the model's accuracy decreased to approximately 0.60, indicating a substantial performance drop when temporal information is limited. This experiment highlights the importance of temporal context in fall detection. Unlike static image classification, fall events often depend on motion progression across multiple frames, and limited frame sampling may remove critical cues required for accurate recognition. These findings suggest that models capable of processing longer video segments or higher frame budgets may provide improved performance in real-world monitoring systems. From a deployment perspective, the results reveal trade-offs between local and API-based VLMs. Local models such as Qwen2.5-VL provide the advantage of processing longer video sequences and maintaining data privacy through offline inference. In contrast, API-based models offer strong reasoning capabilities and easy integration but may be constrained by input limits and operational cost. For applications such as elderly fall monitoring, where both accuracy and privacy are critical, locally deployed VLMs may offer a practical and scalable solution.

Although specialized fall detection architectures may achieve higher accuracy on benchmark datasets, they typically require task-specific training and lack natural-language interpretability. In contrast, VLMs offer zero-/few-shot reasoning capabilities together with explainable textual outputs, making them attractive for flexible healthcare monitoring applications. Table 7 compares the proposed VLMs' performance with some similar existing studies.

TABLE 7. Comparison with Existing Fall Detection Studies in UR Fall dataset

Method	Reference	Accuracy
CNN-LSTM	[15]	0.94
Modular DL	[8]	0.98
Proposed VLMs		0.86

4. Conclusions

This study evaluated the effectiveness of modern VLMs for video-based fall detection using the UR Fall Detection Dataset. Three models were examined: the locally deployed Qwen2.5-VL, and two API-based models, GPT-4o and Claude Sonnet 4.5. The results demonstrate that VLMs can effectively recognize fall events directly from video frames without requiring specialized fall-detection architectures. Among the evaluated models, Qwen2.5-VL achieved the highest overall accuracy (0.86) when processing the full video input, while GPT-4o achieved slightly lower accuracy (0.83) but demonstrated perfect recall for fall events. Claude Sonnet 4.5 showed lower overall performance due to a higher number of false positives in

non-fall activities. These findings indicate that current VLMs tend to prioritize fall detection sensitivity, which may increase false alarm rates in practical deployments. An important observation from this study is the strong influence of temporal context on model performance. When the Qwen2.5-VL model was restricted to the same 8-frame input used for API-based models, its accuracy dropped to approximately 0.60. This result highlights the importance of temporal information in video-based activity recognition tasks such as fall detection. From a deployment perspective, the results suggest a trade-off between local and API-based models. Local VLMs allow processing of longer video sequences and preserve data privacy through offline inference, whereas API-based models offer convenient integration but may be limited by input constraints and operational cost. Future work will explore larger fall detection datasets, improved frame-sampling strategies, and real-time deployment of VLM-based fall detection systems for smart healthcare and assisted living environments.

References

[1] M. M. Islam, O. Tayan, M. R. Islam, M. S. Islam, S. Nooruddin, M. N. Kabir, and M. R. Islam, "Deep learning based systems developed for fall detection: A review," *Ieee access*, vol. 8, pp. 166 117–166 137, 2020.

[2] O. Guendoul, A. Bahaj, Y. Tabii, and R. Oulad Haj Thami, "Preliminary evaluation of vision-language models for fall detection," in *International Conference on Applied Informatics*. Springer, 2025, pp. 220–234.

[3] M. Mubashir, L. Shao, and L. Seed, "A survey on fall detection: Principles and approaches," *Neurocomputing*, vol. 100, pp. 144–152, 2013.

[4] H. Nait-Charif and S. J. McKenna, "Activity summarisation and fall detection in a supportive home environment," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, vol. 4. IEEE, 2004, pp. 323–326.

[5] F.-L. Chen, D.-Z. Zhang, M.-L. Han, X.-Y. Chen, J. Shi, S. Xu, and B. Xu, "Vlp: A survey on vision-language pre-training," *Machine Intelligence Research*, vol. 20, no. 1, pp. 38–56, 2023.

[6] L. Martínez-Villaseñor, H. Ponce, J. Brieva, E. Moya-Albor, J. Núñez-Martínez, and C. Peñafort-Asturiano, "Up-fall detection dataset: A multimodal approach," *Sensors*, vol. 19, no. 9, p. 1988, 2019.

[7] W. Samek and K.-R. Müller, "Towards explainable artificial intelligence," in *Explainable AI: interpreting, explaining and visualizing deep learning*. Springer, 2019, pp. 5–22.

[8] M. Dutt, A. Gupta, M. Goodwin, and C. W. Omlin, "An interpretable modular deep learning framework for video-based fall detection," *Applied Sciences*, vol. 14, no. 11, p. 4722, 2024.

[9] R. Amari, Z. Noubigh, S. Zrigui, D. Berchech, H. Nicolas, and M. Zrigui, "Deep convolutional neural network for arabic speech recognition," in *International Conference on Computational Collective Intelligence*. Springer, 2022, pp. 120–134.

[10] I. A. Abro, S. S. Alharbi, N. S. Alshammari, A. Algarni, N. A. Almujally, A. Jalal, and H. Liu, "Multimodal intelligent biosensors framework for fall disease detection and healthcare monitoring," *Frontiers in Bioengineering and Biotechnology*, vol. 13, p. 1544968, 2025.

[11] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[12] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: a visual language model for few-shot learning," *Advances in neural information processing systems*, vol. 35, pp. 23 716–23 736, 2022.

[13] E. Sacoransky, B. Y. Kwan, and D. Soboleski, "Chatgpt and assistive ai in structured radiology reporting: A systematic review," *Current Problems in Diagnostic Radiology*, vol. 53, no. 6, pp. 728–737, 2024.

[14] A. Ovsyannikov, M. Airapetian, D. , N. J. Alkhateeb, and L. Ali, "Comparative evaluation of vision-language models in skin diseases recognition," in *2025 International Conference on Computer and Applications (ICCA)*. IEEE, 2025, pp. 1–7.

[15] A. M. Alashjaee, "Deep learning for network security: an attention-cnn-lstm model for accurate intrusion detection," *Scientific Reports*, vol. 15, no. 1, p. 21856, 2025.