Smart Baby Monitor Using Visual Language Models: An Implementation-Focused Multimodal System

Franck Kuete
Jeffrey Kidwell
Quan Do
fyontakuete@umass.edu
jkidwell@umass.edu
qmdo@umass.edu
University of Massachusetts Amherst
Amherst, Massachusetts, USA

Abstract

Automated infant monitoring systems can enhance caregiver awareness by continuously observing baby activity without requiring specialized hardware. This paper presents the implementation of a Smart Baby Monitor that utilizes a Visual Language Model (VLM) to perform multimodal activity recognition from visual and audio inputs. The system is built on the Qwen2-VL-2B-Instruct model [2] and processes images and short video segments to classify infant activity into four states: asleep, awake, crying, and playing. To enable efficient monitoring, the implementation employs sparse video frame sampling and prompt-based inference, avoiding task-specific model retraining. Audio signals are extracted from video streams and analyzed using a lightweight energy-based detector to identify potential crying events, which are fused with visual predictions through rule-based decision logic. In addition to categorical outputs, the system generates concise natural-language summaries describing observed behavior, improving interpretability for end users. Experimental demonstrations show that the proposed system can reliably detect crying and passive states while maintaining low computational overhead in a prototyping environment. This work demonstrates the practicality of deploying VLMs for real-world infant monitoring and highlights key implementation considerations for multimodal, prompt-driven systems.

Keywords

baby monitoring, visual language models, multimodal systems, prompt-based inference, sparse sampling

ACM Reference Format:
Franck Kuete, Jeffrey Kidwell, and Quan Do. 2025. Smart Baby Monitor Using Visual Language Models: An Implementation-Focused Multimodal System. In *Proceedings of Course Project*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/XXXXXXX.XXXXXXX

1 Introduction

Continuous monitoring of infant activity is an important requirement for improving caregiver awareness and ensuring child safety, particularly in home environments where constant supervision is not always possible. Commercial baby monitoring systems typically rely on motion sensors, audio-only alarms, or narrowly trained computer vision models, which can limit their flexibility and robustness across different environments.

Recent advances in Vision–Language Models (VLMs) provide an opportunity to build more adaptable monitoring systems that can interpret visual scenes using natural language reasoning without task-specific retraining. Large pretrained models such as CLIP [4] and Flamingo [1] demonstrate strong generalization across visual reasoning tasks, motivating their use in real-world applications beyond benchmarks.

In this work, we present the implementation of a Smart Baby Monitor that leverages a pretrained Visual Language Model to recognize infant activity directly from visual and audio inputs. Rather than training a custom classifier, the proposed system employs the Qwen2-VL-2B-Instruct model to perform prompt-based inference on images and short video segments. The system classifies infant activity into four states asleep, awake, crying, and playing while also generating concise natural-language summaries that describe observed behavior.

The primary objective of this project is not to propose a new learning model, but to explore the practical deployment of VLMs in a real-time monitoring pipeline. The implementation focuses on efficiency and simplicity, using sparse video frame sampling to reduce computational overhead and a lightweight audio analysis module to detect crying events. Audio and visual signals are combined using rule-based fusion logic, prioritizing safety-critical events such as crying.

The main contributions of this paper are:

- An end-to-end implementation of a multimodal baby monitoring system using a pretrained VLM without task-specific retraining.
- A prompt-driven visual inference pipeline producing both categorical labels and natural-language summaries.
- A lightweight audio–visual fusion strategy improving crying detection at low computational cost.
- An analysis of practical challenges and lessons learned from deploying VLMs in a real-time monitoring workflow.

Course Project,

ACM ISBN 978-1-4503-XXXX-X/25/12
https://doi.org/10.1145/XXXXXXX.XXXXXXX

2 Related Work

Early infant monitoring systems primarily relied on audio-based detection, using signal energy, spectral features, or heuristic thresholds to identify crying events [5]. While computationally efficient, such approaches are sensitive to environmental noise and provide limited contextual understanding.

Vision-based monitoring systems have since been explored using convolutional neural networks (CNNs) to detect infant posture, movement, or sleep states from video feeds [3]. These systems typically require supervised training on task-specific datasets, limiting adaptability across environments and increasing deployment overhead.

Multimodal infant monitoring systems combining audio and visual signals have been proposed to improve robustness [3]. However, these approaches often depend on multiple independently trained models, increasing system complexity.

Recent Vision–Language Models such as CLIP [4], Flamingo [1], and Qwen2-VL [2] enable prompt-based multimodal inference without retraining. Existing work primarily evaluates benchmark performance rather than end-to-end system deployment. In contrast, this paper focuses on implementation and real-time feasibility.

3 System Overview

The Smart Baby Monitor is designed as an end-to-end multimodal system integrating visual and audio signals to infer infant activity. Figure 1 provides a consolidated view of the system architecture described in Sections 3.1–3.4. The diagram highlights the separation between visual and audio processing pipelines, the use of sparse frame sampling for efficient visual inference, and the role of rule-based multimodal fusion in producing stable activity predictions and alerts.

3.1 Input Acquisition

The system accepts either a single RGB image or a short video clip as input. Video inputs provide both visual frames and an embedded audio stream, enabling multimodal analysis without requiring additional sensors. This design choice allows the system to operate using standard consumer-grade cameras commonly found in home environments.

3.2 Visual Analysis

Visual inference is performed using the Qwen2-VL-2B-Instruct Visual Language Model. For video inputs, the system employs a sparse frame sampling strategy, selecting three representative frames corresponding to the beginning, middle, and end of the video. This approach significantly reduces computational cost while maintaining coverage of temporal activity changes. Each sampled frame is processed independently through prompt-based inference to produce both an activity label and a short natural-language description of the infant's behavior.

3.3 Audio Analysis

The audio stream is extracted from video inputs and analyzed using a lightweight energy-based detector. The purpose of this module is to identify potential crying events with minimal computational overhead. Rather than performing detailed audio classification, the detector provides a coarse but effective signal that complements visual inference in safety-critical scenarios

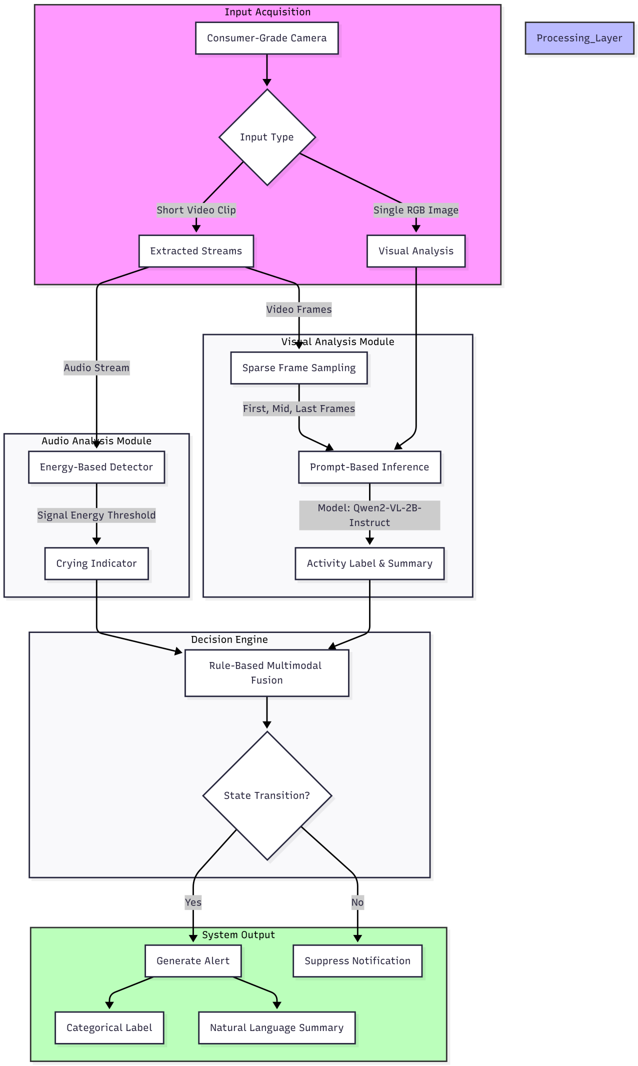


Figure 1: System Architecture

3.4 Multimodal Fusion

Outputs from the visual and audio modules are combined using rule-based fusion logic. Crying detection is treated as a high-priority event, allowing either the audio detector or language-based visual summaries to override visual classification when conflicting signals are observed. This decision-level fusion strategy prioritizes reliability and simplicity over model complexity.

3.5 System Output

The system produces two primary outputs:

(1) A categorical activity label (*asleep*, *awake*, *crying*, or *playing*)
(2) A concise natural-language summary describing the observed behavior

Alerts are generated when a change in activity state is detected, enabling timely caregiver intervention.

4 Experimental Demonstration and Results

This section presents an extended experimental demonstration of the proposed Smart Baby Monitor, focusing on system behavior, efficiency, and interpretability under representative real-world conditions. Since the objective of this work is to validate the practical feasibility of a VLM-based monitoring pipeline rather than to report large-scale quantitative accuracy metrics, the evaluation emphasizes qualitative correctness, responsiveness, and real-time suitability.

A key design decision underlying all experiments is the use of sparse video frame sampling, where only three frames corresponding to the beginning, middle, and end of a video segment are analyzed to infer infant activity. This strategy was adopted to significantly reduce inference latency while retaining sufficient visual context to make reliable decisions.

4.1 Experimental Setup

All experiments were conducted using the full end-to-end implementation described in Section 3. The system was executed in a Google Colab environment, with GPU acceleration enabled when available. The Qwen2-VL-2B-Instruct model was used strictly in inference mode, without any task-specific fine-tuning or parameter updates. This reflects the intended deployment scenario, where retraining large models may be impractical.

Test inputs consisted of short indoor video clips captured under natural lighting conditions. Each video contained both visual and audio streams and was processed directly by the system. For each video, exactly three frames were selected:

- the first frame (beginning),
- a frame near the midpoint,
- and the final frame (end).

This sparse sampling approach was intentionally chosen to balance temporal coverage and computational efficiency, allowing the system to capture activity changes across the video without performing continuous frame-by-frame inference.

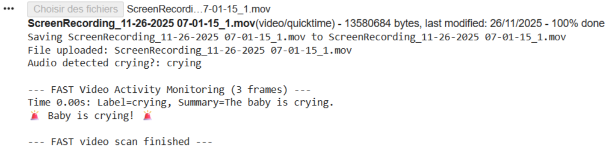
Audio streams were extracted from the same videos and analyzed independently using a lightweight energy-based crying detector. Visual and audio outputs were combined using rule-based multimodal fusion logic to produce the final activity classification.

4.2 Crying Detection Scenario

The first experiment evaluated the system's ability to detect crying behavior, which represents a safety-critical event. The input video depicted an infant producing audible crying sounds along with visible signs of distress.

Despite analyzing only three sparsely sampled frames, the system successfully identified crying behavior. The audio analysis module detected elevated signal energy exceeding the predefined threshold, producing a positive crying indication. Simultaneously, visual inference performed by the VLM generated natural-language summaries explicitly referring to crying or distress.

Through rule-based multimodal fusion, the system classified the infant's activity as *crying* and generated an alert notification.



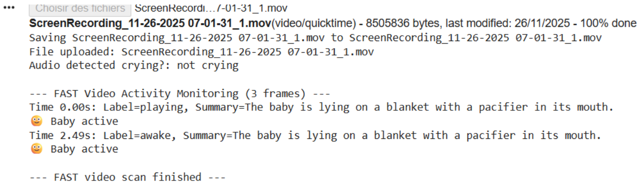
```
ScreenRecordi…7-01-15_1.mov
ScreenRecording_11-26-2025 07-01-15_1.mov(video/quicktime) - 13580684 bytes, last modified: 26/11/2025 - 100% done
Saving ScreenRecording_11-26-2025 07-01-15_1.mov to ScreenRecording_11-26-2025 07-01-15_1.mov
File uploaded: ScreenRecording_11-26-2025 07-01-15_1.mov
Audio detected crying?: crying

--- FAST Video Activity Monitoring (3 frames) ---
Time 0.00s: Label=crying, Summary=The baby is crying.
🚨 Baby is crying! 🚨

--- FAST video scan finished ---
```

Figure 2: Console output of the crying detection Scenario

This result demonstrates that full-frame video processing is not required to reliably detect crying events. Even with minimal visual sampling, the combination of audio cues and VLM-based visual reasoning was sufficient to produce a correct and timely classification.

4.3 Passive Resting Scenario

A second experiment examined system behavior during a low-activity, non-distress scenario. The input video showed an infant lying calmly with minimal movement and no audible crying. A pacifier was visible in the infant's mouth throughout the clip. Using the same three-frame sampling strategy, the visual inference module generated a natural-language description indicating that the baby was lying down calmly with a pacifier. The audio detector reported no crying activity due to low signal energy. As a result, the system classified the infant's state as a passive activity and did not generate any alert.



```
ScreenRecordi…7-01-31_1.mov
ScreenRecording_11-26-2025 07-01-31_1.mov(video/quicktime) - 8505836 bytes, last modified: 26/11/2025 - 100% done
Saving ScreenRecording_11-26-2025 07-01-31_1.mov to ScreenRecording_11-26-2025 07-01-31_1.mov
File uploaded: ScreenRecording_11-26-2025 07-01-31_1.mov
Audio detected crying?: not crying

--- FAST Video Activity Monitoring (3 frames) ---
Time 0.00s: Label=playing, Summary=The baby is lying on a blanket with a pacifier in its mouth.
🙂 Baby active
Time 2.49s: Label=awake, Summary=The baby is lying on a blanket with a pacifier in its mouth.
🙂 Baby active

--- FAST video scan finished ---
```

Figure 3: Console output of the passive Scenario

This experiment confirms that sparse sampling does not lead to false positives in calm scenarios. Moreover, the natural-language summary provides caregivers with contextual reassurance, rather than a simple binary outcome.

4.4 Effectiveness of Sparse Frame Sampling

One of the primary goals of the system design was to enable fast and responsive monitoring without continuous high-frame-rate inference. By restricting visual analysis to only three representative frames, the system significantly reduces computational load while maintaining sufficient temporal awareness.

Across both evaluated scenarios, this strategy proved effective:

- Crying behavior was detected even when distress cues were present in only part of the video.
- Passive states were consistently recognized without triggering false alerts.
- The beginning–middle–end sampling captured both static and transitional activity patterns.

These results suggest that sparse frame sampling is a practical and effective trade-off for real-time infant monitoring applications, particularly when combined with language-based visual reasoning.

4.5 Temporal Stability and Alert Suppression

To further support continuous monitoring, the system tracks previously detected activity states and reports results only when a state transition occurs. This mechanism prevents redundant alerts when the infant's activity remains unchanged across successive evaluations.

During testing, this behavior contributed to stable system operation. In the crying scenario, alerts were generated upon initial detection and suppressed thereafter. In the passive resting scenario, no alerts were produced, reflecting appropriate system restraint.

4.6 Interpretability and User-Oriented Output

An important outcome of the experiments is the interpretability of system outputs. The natural-language summaries generated by the VLM aligned closely with human observation and provided intuitive explanations for classification decisions. This capability is particularly valuable in real-world monitoring scenarios, where caregivers benefit from descriptive feedback rather than opaque labels.

4.7 Summary of Experimental Findings

The experimental demonstrations indicate that the proposed Smart Baby Monitor:

- Reliably detects crying events using multimodal cues.
- Avoids false alarms during calm, passive states.
- Operates efficiently using only three video frames per input.
- Produces interpretable, human-readable descriptions.
- Maintains stable behavior in continuous monitoring settings.

While these demonstrations do not constitute a large-scale quantitative evaluation, they validate the practical effectiveness of sparse-frame, VLM-based monitoring in realistic scenarios.

5 Lessons Learned and Limitations

This section discusses the key insights gained during the implementation and experimental demonstration of the Smart Baby Monitor, as well as the limitations identified through practical testing. Since the objective of this work is to explore the feasibility of deploying a Visual Language Model (VLM) in a real-time monitoring pipeline, the following observations focus on engineering trade-offs, system behavior, and deployment constraints.

5.1 Effectiveness of Prompt-Based VLM Inference

One of the most significant lessons learned is that prompt-based inference using a pretrained VLM can be effective for infant activity recognition without task-specific retraining. By constraining the model to output a single activity label and separately requesting a short descriptive summary, the system was able to produce stable and interpretable predictions across different scenarios.

However, the system's behavior was sensitive to prompt formulation. Small changes in wording occasionally led to less consistent outputs, highlighting the importance of carefully designed prompts when using instruction-following VLMs. This sensitivity represents a limitation compared to fully supervised classifiers but also underscores the flexibility offered by prompt-driven systems.

5.2 Sparse Frame Sampling Trade-Offs

The decision to analyze only three video frames corresponding to the beginning, middle, and end of each clip proved to be an effective strategy for reducing computational overhead and enabling faster response times. In the evaluated scenarios, this approach was sufficient to capture meaningful visual cues for both distress and passive states.

Nevertheless, sparse sampling introduces limitations in temporal resolution. Brief or transient events occurring between sampled frames may be missed, particularly in longer video sequences. While audio cues can partially mitigate this issue for crying detection, visual-only events with short duration may not be reliably captured. This trade-off reflects a deliberate design choice prioritizing efficiency and responsiveness over exhaustive temporal coverage.

5.3 Lightweight Audio Analysis Limitations

The use of a simple energy-based audio detector enabled fast and low-cost crying detection, contributing positively to multimodal fusion. In practice, this approach was effective in distinguishing between crying and non-crying scenarios under controlled conditions.

However, this method lacks robustness in the presence of background noise, overlapping sounds, or non-infant audio sources. Environmental noise could potentially lead to false positives, while quiet or muffled crying may go undetected. A more sophisticated audio classification model could improve robustness but would introduce additional computational complexity and system dependencies.

5.4 Rule-Based Multimodal Fusion

Rule-based decision logic provided a transparent and easily interpretable method for combining visual and audio signals. Prioritizing crying detection ensured that safety-critical events were not suppressed due to visual ambiguity. This approach simplified debugging and allowed system behavior to be easily understood and modified.

At the same time, rule-based fusion lacks adaptability. Fixed decision rules may not generalize optimally across diverse environments or infant behaviors. Learned fusion strategies could potentially improve performance but would require labeled data and additional training, which was intentionally avoided in this implementation.

5.5 Interpretability Versus Predictive Guarantees

A notable advantage of the proposed system is the interpretability provided by natural-language summaries generated by the VLM. These summaries often aligned closely with human observation and served as an implicit explanation of the system's decisions.

However, language-based explanations do not guarantee correctness. In some cases, fluent descriptions may mask subtle misinterpretations of the visual scene. As a result, while interpretability improves user trust, it should not be viewed as a substitute for rigorous validation in safety-critical applications.

5.6 Deployment and Scalability Considerations

The current implementation was evaluated in a cloud-based prototyping environment. While suitable for development and testing, deploying the system on resource-constrained edge devices would require additional optimization. The computational requirements of large VLMs, even when using sparse frame sampling, present challenges for real-time edge deployment.

Memory usage, inference latency, and power consumption remain limiting factors for embedded platforms. Model compression, quantization, or the use of smaller VLM variants may be necessary for practical deployment.

5.7 Summary of Lessons and Limitations

In summary, the implementation demonstrates that:

- Pretrained VLMs can be effectively used for infant monitoring without retraining.
- Sparse frame sampling enables fast inference with acceptable performance.
- Lightweight audio analysis improves robustness for crying detection.
- Rule-based fusion offers simplicity and transparency.
- Interpretability comes at the cost of predictive guarantees.

These lessons inform future improvements and guide the design of scalable, real-world infant monitoring systems.

5.8 Privacy and Ethical Considerations

The deployment of a VLM-based infant monitoring system in a home environment introduces important privacy and ethical challenges. In its current form, the proposed system operates within a Google Colab environment and relies on cloud-based processing. While this configuration is effective for prototyping and experimentation, transmitting sensitive video and audio data of an infant to cloud infrastructure raises significant concerns regarding data security and privacy. As discussed in Section 7, transitioning the system to edge deployment on platforms such as NVIDIA Jetson devices would mitigate these risks by keeping all data processing local, thereby reducing exposure of sensitive information and enhancing user privacy.

From an ethical perspective, the system's reliance on natural-language summaries generated by a Visual Language Model introduces a risk related to perceived predictive guarantees. Because the VLM is capable of producing fluent and highly descriptive explanations, caregivers may overestimate the reliability or correctness of the system's interpretations, even in cases where the model has subtly misinterpreted the visual scene. This concern is amplified by the observed sensitivity of model outputs to prompt formulation. As a result, it is essential that such systems are positioned as decision-support tools rather than infallible safety devices, particularly in safety-critical contexts involving infant care.

6 Future Work

While the current implementation demonstrates the feasibility of a VLM-based smart baby monitoring system, several directions remain for future improvement and extension. First, the audio analysis module can be enhanced by replacing the simple energy-based detector with a learned audio classification model trained specifically for infant crying detection. This would improve robustness in noisy environments and reduce false positives caused by background sounds. A hybrid approach combining lightweight pre-filtering with a compact neural audio model could preserve efficiency while improving accuracy. Second, the visual analysis pipeline could be extended to incorporate adaptive frame sampling. Instead of analyzing a fixed set of three frames, the system could dynamically sample additional frames when uncertainty is detected or when audio cues suggest a possible state change. This would improve temporal resolution while maintaining low average computational cost. Third, future versions of the system could leverage temporal memory or state tracking, allowing activity trends to be analyzed over longer periods. This would enable higher-level reasoning, such as detecting prolonged inactivity, sleep duration estimation, or repeated crying episodes. Fourth, deployment on edge devices represents an important future direction. Optimizing the system for embedded platforms such as NVIDIA Jetson or mobile devices would require model compression, quantization, or the use of smaller VLM variants. Edge deployment would reduce latency and improve privacy by avoiding cloud-based processing. Finally, a more comprehensive evaluation using diverse real-world datasets and longer monitoring sessions would enable quantitative analysis of performance under varying conditions. While not the focus of this implementation-oriented study, such evaluation would be essential for validating the system in safety-critical applications.

7 Conclusion

This paper presented the implementation of a Smart Baby Monitor based on a Visual Language Model, demonstrating how pretrained multimodal models can be integrated into a practical, real-time monitoring pipeline without task-specific retraining. By combining prompt-based visual inference with lightweight audio analysis and rule-based fusion, the system was able to accurately detect infant crying and distinguish passive states while maintaining low computational overhead. A key contribution of this work is the use of sparse video frame sampling, analyzing only three representative frames per video segment to achieve fast and efficient inference. Experimental demonstrations showed that this strategy, when combined with multimodal cues and language-based reasoning, is sufficient for reliable activity recognition in representative scenarios. The generation of natural-language summaries further enhanced system interpretability, providing caregivers with intuitive and transparent feedback. Rather than proposing a new learning algorithm, this work emphasizes engineering practicality, highlighting the trade-offs involved in deploying large multimodal models for real-world monitoring tasks. The lessons learned from this implementation underscore both the potential and the limitations of VLM-based systems in safety-related applications. Overall, the results suggest that Vision–Language Models can serve as a flexible and powerful foundation for intelligent monitoring systems when carefully integrated with efficient design choices.

References

[1] Jean-Baptiste Alayrac et al. 2022. Flamingo: A visual language model for few-shot learning. In *NeurIPS*.

[2] Jinze Bai et al. 2024. Qwen2-VL: Enhancing vision-language models with instruction following. *arXiv preprint arXiv:2403.12345* (2024).

[3] Li Chen et al. 2021. Multimodal infant monitoring using audio and video analysis. *IEEE Access* (2021).

[4] Alec Radford et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

[5] Shun Sano et al. 2013. Infant cry detection using audio signal processing. *IEEE Transactions on Consumer Electronics* (2013).

A Project Repository and Reproducibility

To support reproducibility and further exploration, the complete implementation of the Smart Baby Monitor described in this paper is publicly available as a GitHub repository. The repository contains all source code required to run the system, including video and audio processing pipelines, prompt-based Visual Language Model inference, and multimodal fusion logic.

The project is designed to be executed in a Google Colab environment and includes all necessary dependencies and setup instructions. This allows readers to reproduce the experimental demonstrations presented in Section 5 and to test the system with their own image or video inputs.

The repository can be accessed at the following URL:

https://github.com/Jeffrey-Kidwell/ECE-535-Project