Assessing Video LLM Performance in Detecting Child Safety Risks

Amaya Keys, Saurav K. Aryal, PhD, Gloria Washington, PhD
AI4PC Lab, Institute for Human-Centered Artificial Intelligence
Howard University

INTRODUCTION

- Over 4 million children are treated for unintentional household injury per year [2], with many likely due to the limits of human attention.
- Traditional child monitoring devices possess critical limitations, namely:
  1) They require constant and scrutinous supervision.,
  2) They are restricted to static, pre-defined infant behaviors.
  3) They act as passive rather than proactive alert systems.
- Increases in remote work, childcare costs, and the adoption of smart home technology suggests the existence of a demand for improved child safety tools.
- Advancements in multimodal large language models has yielded enhanced video understanding, behavior recognition and complex reasoning capabilities.

OBJECTIVE

- Evaluate two multimodal LLMs for their efficacy in detecting child safety risks.
- Determine the feasibility of developing an AI-enabled child safety monitor that matches or enhances human perception.

METHODS

Models Evaluated

The study tested two vision-language models on T4 GPUs:

- **Qwen2-VL-2B-Instruct** [3] (using INT8 quantization)
- **Video-LLaVa-7B** [1] (using 4-bit quantization)

Data Collection

Both models were passed **45 use-case videos** of children engaging in either dangerous (e.g., climbing unstable surfaces or nearing electrical hazards) or age-appropriate activity (e.g., playing with toys or eating whilst supervised).

Video Analysis

Videos were processed using **overlapping 5-second windows with a 2-second stride** to **simulate live video streaming**. For each window, 4-8 frames were sampled and compiled into a temporary video clip for analysis.

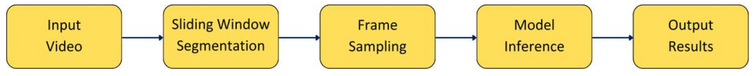

Figure 1. Diagram Illustrating the Video Analysis Approach

METHODS

Each clip was analyzed using the following prompt:

> **Context**: "You are a child safety monitor. Assess this situation from a CHILD SAFETY perspective and evaluate if the child is engaging in unsafe behavior right now based on what you observe. Focus on: actively dangerous actions/behaviors, potential hazards, or immediate threats that will likely cause injury, especially if unsupervised."

> **Question**: "Is this child in any immediate danger? Consider what specific risks you see and why this is dangerous for a CHILD specifically and then begin your response with either 'Yes, the child is in danger' or 'No, the child is not in danger'."

Output

- Full model responses to prompt
- An annotated video with danger overlays
- **Record of the first instance of danger detected** (in both seconds and frame number)

Evaluation Metrics

The models' performance was compared against researcher perception at both **temporal** (second-level) and **granular** (frame-level) precision.

RESULTS

As shown in Figures 2 and 3:

- On **31/45 videos** evaluated at second-level granularity and **29/45 videos** evaluated at frame-level precision, Qwen2-VL's predictions of the first instance of danger detected most closely matched the human perceptual baseline.
- On **23/45 videos** evaluated at both temporal and frame-level granularity, Video-LLaVa's predictions most closely matched the human perceptual baseline.

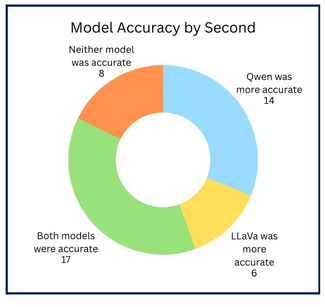

Figure 2. Detection Accuracy Comparison Measured in Seconds

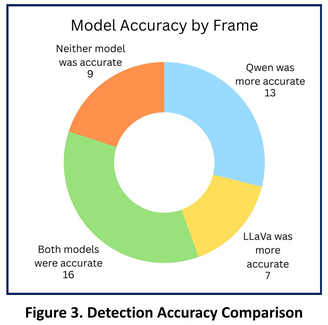

Figure 3. Detection Accuracy Comparison Measured in Frames

RESULTS

- Video-LLaVa outperformed Qwen2-VL in only 6 cases.
- Both models successfully identified danger simultaneously with humans in 17/45 instances.

Overall, as shown in Figure 4, Qwen2-VL demonstrated a **67% average accuracy rate**, while Video-LLaVa achieved a **lower accuracy rate of 51%**.

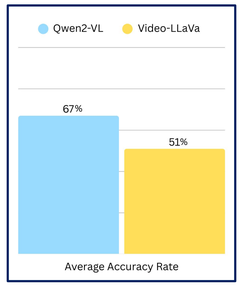

Figure 4. Average Detection Accuracy Percentage of Both Models

CONCLUSION

The results of this preliminary study demonstrate strong potential for the use of multimodal AI for automated child danger detection, bridging the gap left by traditional monitoring devices. Qwen2-VL's performance in handling diverse and dangerous scenarios suggests its superiority for this application.

Limitations

- Lack of temporal context
- Difficulty with nuanced risks and judgement
- Single human rater

Future Work

- Fine-tune the models and newer models.
- Supply context from previous video windows to improve situational awareness
- Collect additional data to refine precision for eventual deployment

REFERENCES

[1] Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., & Yuan, L. (2023). Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. arXiv. https://doi.org/10.48550/ARXIV.2311.10122

[2] Phelan, K. J., Khoury, J., Kalkwarf, H., & Lanphear, B. (2005). Residential Injuries in U.S. Children and Adolescents. Public Health Reports®, 120(1), 63–70. https://doi.org/10.1177/003335490512000111

[3] Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., & Lin, J. (2024). Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. arXiv. https://doi.org/10.48550/ARXIV.2409.12191

ACKNOWLEDGEMENTS

This research project was supported in part by the Office of Naval Research grant N00014-22-1-2714. The work is solely the responsibility of the authors and does not necessarily represent the official view of the Office of Naval Research.