Multimodal LLM Framework for Context-Aware Locomotion Prediction in Construction Exoskeletons

Asif Faisal Chowdhury[1], Ehsan Ahmadi, Ph.D.[2], and Chao Wang, Ph.D.[3]

[1] Ph.D. Student, Bert S. Turner Department of Construction Management, College of Engineering, Louisiana State University, Email: achowd6@lsu.edu (corresponding author).
[2]VDC Coordinator, Lease Crutcher Lewis, Portland, OR, Email: aehsvn@gmail.com
[3]Associate Professor and Graduate Program Advisor, Bert S. Turner Department of Construction Management, College of Engineering, Louisiana State University. Email: chaowang@lsu.edu

ABSTRACT

Real-time prediction of human locomotion is vital for adaptive human-machine systems such as assistive exoskeletons. We present a multimodal pipeline that combines a lightweight local vision-language model for continuous first-person scene captions with a prompt-based language model that fuses three signals, user voice command, recent visual context, and short-term history of locomotion states to infer the next mode. The model outputs a confidence score; when this score falls below a threshold, the system requests brief user clarification, improving safety and reducing silent errors. We tested the approach in a simulated live setting. The system reached 94.3% classification accuracy with a median end-to-end latency of 0.5s, showing that low latency is compatible with high accuracy. These results indicate that the method offers a practical path to robust, context-aware, and interactive locomotion assistance, and support its use as a building block for exoskeleton control and other time-critical applications.

INTRODUCTION

The ability to accurately and promptly recognize human locomotion is foundational to creating effective human-machine systems, from assistive exoskeletons that adapt to a user's intent (Laschowski et al. 2020) to intelligent robots that can safely navigate human-centric environments (e.g., construction sites). Traditional approaches often rely on data from wearable sensors (e.g., IMUs, EMG) (Laschowski et al. 2020), which can be intrusive (Beddiar et al. 2020) and may lack environmental awareness. Such sensor-only methods lack contextual reasoning; they cannot account for external factors, such as obstacles or terrain changes, that may influence a user's locomotion intent. While vision-based systems offer a non-invasive alternative, they historically struggle to interpret complex, dynamic scenes with the speed and reliability required for real-time applications, often failing in cases of occlusion or visual ambiguity(Beddiar et al. 2020). In a chaotic construction setting, with moving machinery, uneven terrain, and frequent occlusions, these single-modality models tend to perform poorly.

The recent emergence of powerful Vision-Language Models (VLMs) and Large Language Models (LLMs) has opened a new frontier for solving this problem (Sapkota et al. 2025). These models can translate raw pixels into rich semantic descriptions, enabling a more nuanced and context-aware understanding of human actions. However, their application in real-time systems presents a significant trade-off: the most capable models often have high computational demands and

network/API latencies, making them unsuitable for control loops that require immediate feedback (Jang and Morabito 2025). Conversely, smaller models that can run locally on edge devices may lack the sophisticated reasoning abilities needed to handle novel or complex scenarios (Jang and Morabito 2025).

This paper addresses the performance trade-off by proposing a novel, multimodal agentic system that strategically integrates models of varying scales. Unlike generic multimodal frameworks, our architecture is explicitly tailored to the safety-critical constraints of construction. Our primary contribution is an architecture in which a lightweight, on-device VLM provides continuous, "always-on" environmental awareness by generating a stream of textual scene descriptions. This visual feed is then synthesized by an efficient LLM-based reasoning engine, which is triggered by user voice commands. The reasoning engine makes the final locomotion mode prediction by integrating three key data streams: the user's explicit intent (voice command), the immediate visual context (VLM captions), and short-term memory (the two most recent locomotion modes). To ensure robustness and safety, our system actively manages uncertainty. The reasoning engine calculates a confidence score for each prediction, enabling it to recognize when its assessment is ambiguous. If the confidence falls below a predefined 80% threshold, which is selected to prioritize safety by favoring clarification over potential error, the system initiates a human-in-the-loop clarification dialogue (Gómez-Carmona et al. 2024).

We implemented and evaluated this complete pipeline in a simulated live environment, rigorously testing its end-to-end performance across a series of realistic locomotion tasks. We demonstrate that this hierarchical approach achieves both high classification accuracy and low latency, presenting a viable framework for deploying advanced AI in real-world, interactive human-assistive technologies.

RELATED WORKS

Vision-based human locomotion or activity recognition has been extensively studied in computer vision. Early approaches often combined object detection and pose estimation to classify human actions from video; for example, one approach detected construction workers with YOLO (Redmon et al. 2016) and then used pose joints to recognize their activities (Jiang and Messner 2023; Son and Kim 2021). Deep learning has since revolutionized this field. Two-stream convolutional networks and 3D CNNs enabled end-to-end video action recognition, while gait-specific models achieved high accuracy by learning unique motion patterns (Ruiz-Barroso et al. 2025). However, these methods primarily rely on pattern recognition and often lack the semantic reasoning required to understand why a transition is occurring or to handle unseen scenarios.

Beyond traditional vision models, there is growing interest in combining vision-language models (VLMs) with large language models (LLMs) for embodied perception and reasoning. VLMs (e.g. image captioning or CLIP-based models (Radford et al. 2021) provide semantic scene understanding by linking visual content to text, while LLMs offer advanced reasoning, commonsense knowledge, and sequential decision-making. Recent surveys note that LLMs can significantly improve robots' planning and execution by incorporating high-level knowledge and reasoning over visual inputs (Jeong et al. 2024).

Modern embodied AI systems increasingly integrate multiple modalities: for example, Google's PaLM-E model (Driess et al. 2023) combines visual sensor input with a language model, enabling a single 562-billion parameter model to perform robotic manipulation planning, visual Q&A, and

captioning within one unified policy. Another paradigm is to keep perception and language modules separate but connected via prompting. A work presented a modular framework where pre-trained VLMs and LLMs communicate in natural language to exchange information and perform zero-shot multimodal reasoning (Zeng et al. 2022). This approach leverages the complementary knowledge of each model without requiring joint training. Along similar lines, other work introduced a coordinator agent (CoLA) that uses an LLM to orchestrate multiple VLM "experts" for complex visual reasoning tasks. The LLM can mediate among the outputs of different vision models via language, achieving state-of-the-art results on vision-language benchmarks such as visual question answering and visual entailment (Chen et al. 2023). Overall, the use of VLMs for constant visual awareness coupled with LLM-based reasoning is gaining traction in robotics and embodied AI, as it marries perceptual context with powerful general reasoning abilities (Jeong et al. 2024).

Our work builds upon these modular approaches but differentiates itself by focusing on the specific latency and reliability constraints of construction exoskeletons. While works like M3-Agent (Long et al. 2025) build long-term memory for complex reasoning, our system optimizes for immediate, short-horizon context suitable for real-time control. Furthermore, we integrate a confidence-gated Human-in-the-Loop mechanism, a critical safety feature often absent in purely autonomous reasoning agents (Khan et al. 2025).

SYSTEM OVERVIEW

Architecture

The proposed system employs a multi-stage pipeline that integrates real-time visual perception, language-based user intent, and short-term memory to predict locomotion modes.

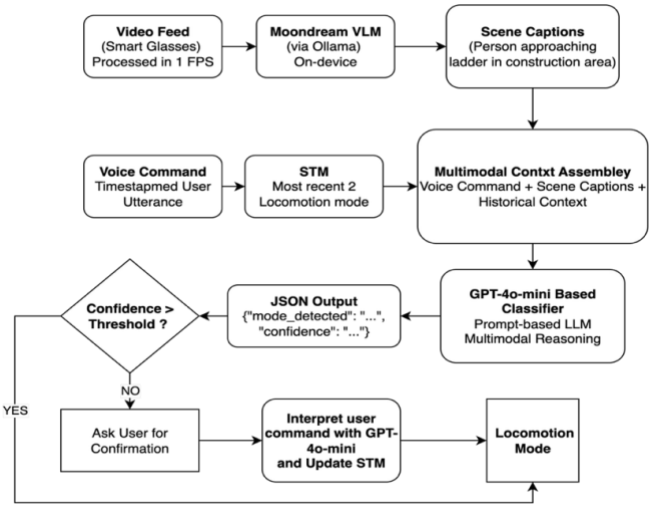


Figure 1: The proposed system architecture for multimodal locomotion mode detection. The model processes video and voice inputs through a VLM and LLM-based classifier, incorporating a confidence-based user confirmation loop to update the system's state.

The data flows through the following steps:

1. **Continuous Visual Perception:** The system processes a live video feed from a first-person perspective (e.g., smart glasses). To balance context frequency with computational load, every second frame is fed into a lightweight, on-device vision model (Roboflow Blog 2025) VLM, run via Ollama). We selected Moondream for its balance of inference speed and descriptive capability, allowing it to generate concise, natural-language descriptions (e.g., "a person approaching a ladder in a construction area") without the prohibitive latency of cloud-based VLMs. This provides a constant stream of visual context, updated every two seconds.
2. **Triggering and Multimodal Context Assembly:** The core reasoning process is triggered by a timestamped voice command from the user. Upon receiving a command, the system assembles a multimodal context that includes three components: (a) the user's voice command (transcribed to text), (b) the most recent scene caption(s) from the VLM (providing immediate visual context, including a window of three most recent scene captions), and (c) historical context from a Short-Term Memory (STM) module, which stores the user's two most recent locomotion modes. By combining these, the system captures the user's explicit intent, current surroundings, and recent state transitions.
3. **LLM-Based Reasoning and Classification:** The assembled multimodal context is formatted into a prompt and passed to an efficient LLM classifier (GPT-4o-mini). We chose GPT-4o-mini over larger variants to minimize API latency while retaining sufficient reasoning capability for intent classification. This model performs multimodal reasoning to infer the user's intended locomotion mode. Its output is constrained to a structured JSON format containing two fields: *mode_detected* and *confidence_score* .
4. **Confidence-Gated Human-in-the-Loop (HIL) Clarification:** The system then assesses the *confidence_score* provided by the LLM. If the score exceeds a set threshold of 0.8 (i.e., ≥80% confidence), the prediction is regarded as reliable and is acted upon immediately. If the confidence drops below this threshold, the system initiates a HIL verification step, prompting the user to confirm or clarify the intended action before moving forward (Gómez-Carmona et al. 2024). This approach prevents the system from making decisions based on uncertain predictions, thereby improving safety and robustness. We selected the 0.8 threshold empirically; preliminary tests showed correct inferences consistently scored above 0.9, while misclassifications fell below 0.8. This balance prevents silent errors without causing frequent, unnecessary interruptions for moderately confident predictions. In practice, this mechanism enables the system to handle ambiguity smoothly by deferring to human judgment only when necessary.
5. **State Update:** The final, confirmed locomotion mode, either from a high-confidence prediction or from the user's clarification, is used to update the STM. This ensures the system's memory remains accurate for subsequent predictions, completing the agentic loop.

Model Choices and Baseline Comparison: The architecture balances accuracy, latency, and robustness by integrating multiple modalities. Our models, Moondream 2 and GPT-4o-mini, are lightweight enough for on-device or real-time use while being capable of handling rich visual and textual data. Single-modality approaches have limitations: IMU-based classifiers, like (Coser et al. 2025), achieve about 94% accuracy in labs with 1-2 ms inference but lack environmental context, which risks misinterpretation in new scenarios. Vision-only methods, such as (Yang et al. 2024) achieve over 95% accuracy in controlled settings but fall below 80% in complex environments or during sudden transitions. Fusion approaches that combine

sensors and predictive algorithms, like (Carvalho et al. 2025) 's LSTM, reach around 98% accuracy before movement but rely on low-level data and lack semantic understanding or uncertainty handling. Our framework combines vision, voice, and context, including a reasoning step that knows when to ask for help, thereby improving robustness while maintaining high accuracy.

PILOT TEST

Video Scenarios and Environment

Our evaluation utilized five video-recorded scenarios (~3 minutes each) simulating real-world locomotion in a varied indoor environment. Each scenario features a participant navigating a continuous sequence of tasks through an environment including a level hallway for walking, a staircase for ascent/descent, a chair for sitting/standing, and a low-clearance obstacle for crouching. This design ensures that a range of locomotion modes and, critically, the transitions between them were captured realistically and systematically.

Tasks and Goals

Each scenario was designed to include seven distinct locomotion modes representative of actions a worker might perform in an industrial or construction setting:

- **Ground walk**: Normal walking on a level surface.
- **Crouch walk**: Navigating under a low-clearance obstacle.
- **End crouch walk**: Transitioning from crouching to standing.
- **Stair ascension**: Walking up stairs.
- **Stair descension**: Walking down stairs.
- **Sitting down**: Transitioning from standing to sitting to rest.
- **Standing up**: Rising from a seated position.

The primary goal was to evaluate the agent's ability to accurately recognize each mode and seamlessly handle transitions by integrating voice commands with visual cues. By ensuring every mode and context switch appears in each scenario, we created a comprehensive testbed to verify the system's performance in a continuous, multi-step operation, a critical requirement for practical deployment.

Simulated "Live" Setup

We used a "simulated live" setup, processing pre-recorded video and voice data in real time to balance experimental control with realistic timing. The simulation fed video frames to the system at 1 FPS and injected scripted voice commands at their precise timestamps from the original recordings. This approach replicated the timing and concurrency of a live field test while ensuring perfect repeatability across runs.

Hardware and Software Environment

Simulations ran on an Ubuntu 25.04 workstation (8-core CPU, 62 GB RAM, RTX 4070 SUPER)

using a Python stack (Flask, OpenCV, PyTorch, and OpenAI API). The complete agentic pipeline, including the on-device Moondream model (via Ollama) and the GPT-4o-mini classifier, ran concurrently using background workers to mimic a parallelized live deployment.

Triggering Mechanism and Timing

Each locomotion task in the recordings was preceded by a scripted voice command (e.g., "I am going upstairs"). These commands were transcribed and timestamped into a JSON file that served as a trigger schedule for the simulation. When a timestamp was reached, the system received the transcribed text as a live input, prompting it to analyze the current visual and historical context to make a prediction. This synchronized triggering ensured a realistic sequence of user intent (voice) followed by physical action (vision), which is crucial for evaluating the system's contextual reasoning.

Test Runs and Metrics

We conducted three independent runs for each of the five scenarios (15 total runs) to ensure the reliability and consistency of our results. To assess system stability, performance metrics were calculated for each run independently. The results discussed below reflect the average accuracy across these three trials, while the detailed confusion matrix and event counts correspond to a single representative run to illustrate specific classification behaviors. We evaluated performance using two primary metrics:

1. **Latency**: Measured as the total time elapsed between a voice command's timestamp in the simulation and the moment the system produced its corresponding prediction. Crucially, this measurement includes the network round-trip time for the OpenAI API call, providing a conservative estimate of real-world performance.
2. **Accuracy**: Calculated as the percentage of correct mode predictions against the ground truth labels associated with each command. A prediction was deemed correct if the agent's highest-confidence output matched the actual activity performed by the user. While we also observed confusion patterns between modes (e.g., crouch walk vs. end crouch), the primary reported metric is overall classification accuracy.

This evaluation protocol provides a robust assessment of the system's real-time performance in terms of speed (latency) and reliability (accuracy) during complex locomotion tasks, giving us confidence in how it would perform in an actual deployment.

PRELIMINARY RESULTS

Overall Performance

The system demonstrated high accuracy and low latency across all test scenarios. Across the three independent runs, the system achieved a consistent average classification accuracy of 94.3%. In the representative run detailed in Table 1 (consisting of 35 command events across 7 tasks), the system correctly identified the locomotion mode in 33 instances. The results, summarized in Table 1 , and Figure 2 , validate the effectiveness of combining continuous visual understanding with voice cues for robust, real-time classification.

Table 1. Locomotion mode classification performance per class (Representative Run). The agent's accuracy for each class is shown as the percentage of correctly recognized instances.

Locomotion Mode	Instances (Total)	Correctly Predicted	Accuracy
Ground walk	5	5	100%
Crouch walk	5	4	80%
End crouch walk	5	4	80%
Stair ascension	5	5	100%
Stair descension	5	5	100%
Sitting down	5	5	100%
Standing up	5	5	100%
Overall	**35**	**33**	**94.3%**

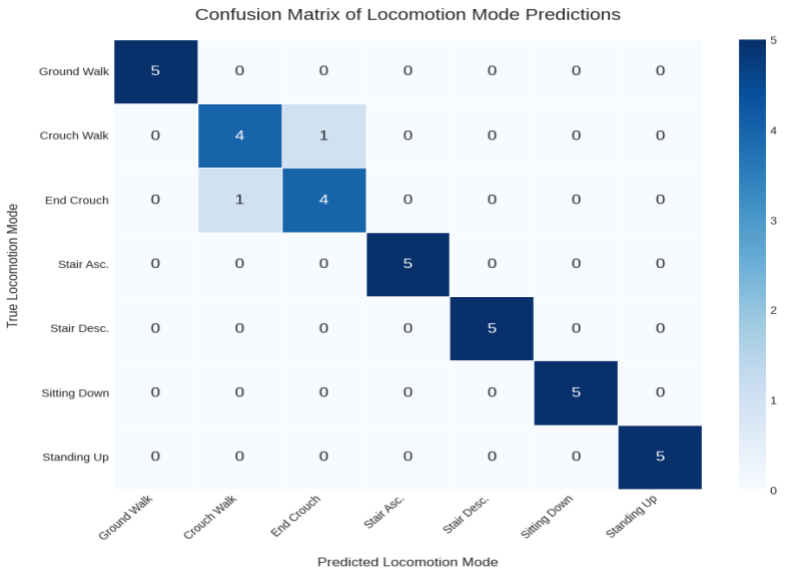


Figure 2. Confusion matrix of the agent's locomotion mode predictions vs. ground truth labels for a representative run. Each cell shows the number of events (out of 5 per class) for a given true mode (rows) that were classified as the predicted mode (columns).

A crucial finding is that the model's confidence was correlated with its correctness. The average confidence score for correct predictions was high at 0.92, whereas the two misclassifications received lower confidence scores (between 0.6 and 0.7). This demonstrates that the system can reliably estimate its own uncertainty.

Latency Analysis

The system met real-time requirements, with a median latency of 0.5 seconds and an average of approximately 0.7 seconds. As shown in the cumulative distribution in Figure 3, roughly 80% of all predictions were delivered in under 0.6 seconds. The tail latency was primarily driven by a small number of events that required more complex reasoning on novel visual scenes, with the

95th percentile latency at approximately 2.4 seconds. Even this worst-case performance remains within a tolerable range for interactive use. The fast median time is a direct result of the system's architecture, where lightweight, on-device models handle the majority of the continuous processing load.

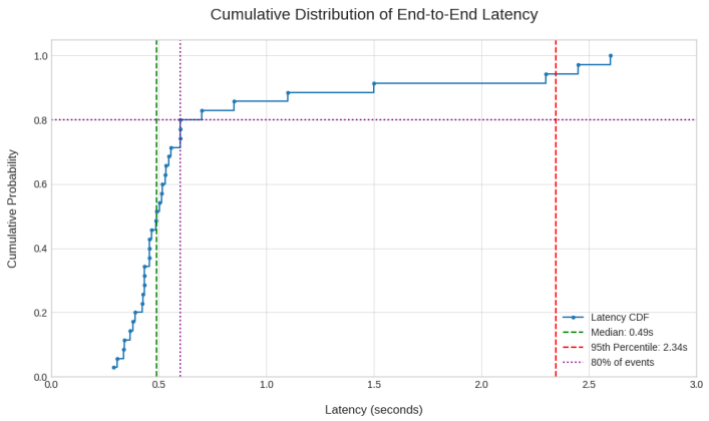


Figure 3.Cumulative distribution function (CDF) of the system's end-to-end latency. The plot shows a median latency of 0.49 seconds, with 80% of events processed within approximately 0.6 seconds and a 95th percentile latency of 2.34 seconds

Impact of Human-in-the-Loop (HIL) Clarification

The system's ability to detect its own uncertainty was central to safe operation. Low confidence scores activated the human-in-the-loop step in 2 of 35 cases. In those cases, the user's clarification was sent to GPT-4o-min together with the relevant context. It returned the correct prediction label, resolving the ambiguity. This shows that the system flags uncertainty and requests input instead of failing silently, which is important for real-world, safety-focused use.

Discussions

Our experimental results validate the efficacy of a multimodal approach that integrates temporal context, vision, and voice commands. Unlike IMU-based approaches that struggle with the "semantic gap" between mechanically similar actions, our system achieves 94.3% accuracy by leveraging visual and linguistic context to resolve ambiguity. The system demonstrated high robustness, maintaining accurate predictions even through significant visual occlusions and at abrupt transitions. This resilience stems from the model's use of memory; by reasoning over the last known mode and recent scene descriptions (3-frame window), it can maintain state continuity during brief periods of ambiguous sensory input. The few observed errors were minor, transient misclassifications between similar states (e.g., crouch walk vs. end crouch), stemming from slight timing misalignments between visual evidence and verbal cues.

Limitations and Scalability

Despite promising results, this study has limitations. The evaluation was conducted in a controlled environment with a single participant to establish baseline feasibility (pilot study). Real-world

deployment introduces challenges such as multi-user interference and network instability. While our 0.5 s median latency is promising, it relies on API connectivity. In "dead zones" common on construction sites, reliance on cloud-based LLMs (like GPT-4o-mini) is a vulnerability. Future work will explore local-only fallback models (e.g., Llama 3 (Grattafiori et al. 2024) on edge) to ensure basic functionality when connectivity is lost, and will expand testing to diverse user populations to validate the HIL mechanism's scalability.

CONCLUSIONS

This paper presented a multimodal LLM framework for context-aware locomotion prediction in construction exoskeletons. By combining efficient on-device vision (Moondream) with cloud-based reasoning (GPT-4o-mini) and a safety-first Human-in-the-Loop mechanism, we achieved high accuracy (94.3%) and low latency (0.5 s) in simulated trials. Key contributions include the demonstration of semantic reasoning superior to traditional pattern recognition for complex transitions and a practical architecture for managing uncertainty in safety-critical tasks. Future research will focus on field testing in unscripted environments and addressing network-denied operations to fully realize the potential of intelligent, assistive construction technology.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation (NSF) under Grant No. 2222881. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

REFERENCES

Beddiar, Djamila Romaissa, Brahim Nini, Mohammad Sabokrou, and Abdenour Hadid. 2020. "Vision-Based Human Activity Recognition: A Survey." *Multimedia Tools and Applications* 79 (41–42): 30509–55. https://doi.org/10.1007/s11042-020-09004-3.

Carvalho, Simão P., Joana Figueiredo, João J. Cerqueira, and Cristina P. Santos. 2025. "Locomotion Mode Prediction in Real-Life Walking with and without Ankle–Foot Exoskeleton Assistance." *Applied Intelligence* 55 (7): 546. https://doi.org/10.1007/s10489-025-06416-2.

Chen, Liangyu, Bo Li, Sheng Shen, et al. 2023. "Large Language Models Are Visual Reasoning Coordinators." arXiv:2310.15166. Preprint, arXiv, October 23. https://doi.org/10.48550/arXiv.2310.15166.

Coser, Omar, Christian Tamantini, Matteo Tortora, et al. 2025. "Deep Learning for Human Locomotion Analysis in Lower-Limb Exoskeletons: A Comparative Study." *Frontiers in Computer Science* 7 (August): 1597143. https://doi.org/10.3389/fcomp.2025.1597143.

Driess, Danny, Fei Xia, Mehdi S. M. Sajjadi, et al. 2023. "PaLM-E: An Embodied Multimodal Language Model." arXiv:2303.03378. Preprint, arXiv, March 6. https://doi.org/10.48550/arXiv.2303.03378.

Gómez-Carmona, Oihane, Diego Casado-Mansilla, Diego López-de-Ipiña, and Javier García-Zubia. 2024. "Human-in-the-Loop Machine Learning: Reconceptualizing the Role of the User in Interactive Approaches." *Internet of Things* 25 (April): 101048. https://doi.org/10.1016/j.iot.2023.101048.

Grattafiori, Aaron, Abhimanyu Dubey, Abhinav Jauhri, et al. 2024. “The Llama 3 Herd of Models.” arXiv:2407.21783. Preprint, arXiv, November 23. https://doi.org/10.48550/arXiv.2407.21783.

Jang, SiYoung, and Roberto Morabito. 2025. “Edge-First Language Model Inference: Models, Metrics, and Tradeoffs.” arXiv:2505.16508. Preprint, arXiv, May 29. https://doi.org/10.48550/arXiv.2505.16508.

Jeong, Hyeongyo, Haechan Lee, Changwon Kim, and Sungtae Shin. 2024. “A Survey of Robot Intelligence with Large Language Models.” *Applied Sciences* 14 (19): 8868. https://doi.org/10.3390/app14198868.

Jiang, Zhouqian, and John I. Messner. 2023. “Computer Vision Applications In Construction And Asset Management Phases: A Literature Review.” *Journal of Information Technology in Construction* 28 (April): 176–99. https://doi.org/10.36680/j.itcon.2023.009.

Khan, Nasrullah, Dohyeong Kim, Minju Kim, Daeho Kim, and Dongmin Lee. 2025. “Mp-Mcwt: Message Passing Strategy for Multi-Camera Worker Tracking in Construction.” Preprint, SSRN. https://doi.org/10.2139/ssrn.5122800.

Laschowski, Brock, William McNally, Alexander Wong, and John McPhee. 2020. “ExoNet Database: Wearable Camera Images of Human Locomotion Environments.” *Frontiers in Robotics and AI* 7 (December): 562061. https://doi.org/10.3389/frobt.2020.562061.

Long, Lin, Yichen He, Wentao Ye, et al. 2025. “Seeing, Listening, Remembering, and Reasoning: A Multimodal Agent with Long-Term Memory.” arXiv:2508.09736. Preprint, arXiv, August 15. https://doi.org/10.48550/arXiv.2508.09736.

Radford, Alec, Jong Wook Kim, Chris Hallacy, et al. 2021. “Learning Transferable Visual Models From Natural Language Supervision.” arXiv:2103.00020. Preprint, arXiv, February 26. https://doi.org/10.48550/arXiv.2103.00020.

Redmon, Joseph, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. “You Only Look Once: Unified, Real-Time Object Detection.” arXiv:1506.02640. Preprint, arXiv, May 9. https://doi.org/10.48550/arXiv.1506.02640.

Roboflow Blog. 2025. “Moondream 2: Multimodal and Vision Analysis.” March 11. https://blog.roboflow.com/moondream-2/.

Ruiz-Barroso, Paula, José María González-Linares, Francisco M. Castro, and Nicolás Guil. 2025. “Real-Time Multiple People Gait Recognition in the Edge.” *Scientific Reports* 15 (1): 19276. https://doi.org/10.1038/s41598-025-02351-x.

Sapkota, Ranjan, Yang Cao, Konstantinos I. Roumeliotis, and Manoj Karkee. 2025. “Vision-Language-Action Models: Concepts, Progress, Applications and Challenges.” arXiv:2505.04769. Preprint, arXiv, May 7. https://doi.org/10.48550/arXiv.2505.04769.

Son, Hyojoo, and Changwan Kim. 2021. “Integrated Worker Detection and Tracking for the Safe Operation of Construction Machinery.” *Automation in Construction* 126 (June): 103670. https://doi.org/10.1016/j.autcon.2021.103670.

Yang, Gayoung, Jeong Heo, and Brian Byunghyun Kang. 2024. “Adaptive Vision-Based Gait Environment Classification for Soft Ankle Exoskeleton.” *Actuators* 13 (11): 428. https://doi.org/10.3390/act13110428.

Zeng, Andy, Maria Attarian, Brian Ichter, et al. 2022. “Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language.” arXiv:2204.00598. Preprint, arXiv, May 27. https://doi.org/10.48550/arXiv.2204.00598.