Trustworthy Human-Exoskeleton Collaboration: A Memory-Enhanced LLM Agent for Construction Locomotion Prediction

Ehsan Ahmadi[1] and Chao Wang[2]

***Abstract*— This research presents a multimodal framework for locomotion prediction to enhance high-level exoskeleton control in dynamic construction settings by integrating speech commands and visual data from smart glasses. The study comprises two stages: the first stage evaluates the zero-shot generalization of GPT-4o against fine-tuned CLIP and ImageBind models, achieving weighted F1-scores of 88%, 90%, and 79%, respectively; and the second stage introduces a large language model-based agent with short-term and long-term memory systems to improve context awareness and robustness to command ambiguity. Tested under stringent conditions with challenging, ambiguous, and high-risk scenarios, the agent attains a 90% F1-score compared to a 73% no-memory baseline.**

I. INTRODUCTION

The construction industry confronts significant challenges, including labor shortages, intense physical demands, and heightened safety risks, with work-related musculoskeletal disorders a common issue [1], [2]. Exoskeletons, wearable devices designed to augment, assist, or enhance physical activity, provide a promising solution to alleviate these problems [1]. Yet, their effectiveness depends on sophisticated control systems capable of accurately interpreting user intent across diverse locomotion activities in unpredictable environments, a cornerstone of seamless human-exoskeleton collaboration [3], [4]. Traditional approaches often rely on supervised learning methods, and their integration into industrial applications warrants further exploration [5]. While prior research has focused on routine gait tasks [6], construction tasks such as ladder climbing and obstacle navigation require adaptation to dynamic conditions.

Recent advances in artificial intelligence, particularly Large Language Models (LLMs), offer a transformative opportunity to overcome these limitations, demonstrating significant progress in multimodal understanding [7], [8]. Pre-trained on extensive datasets, LLMs support few-shot and zero-shot generalization, enabling flexible adaptation to varied tasks without requiring extensive retraining [9], [10], [11]. Building on these strengths, LLM-based agents excel in natural language interaction, environmental comprehension, reasoning, planning, and tool usage, performing complex tasks with remarkable efficacy while leveraging memory systems to store and retrieve contextual information for consistent performance [12], [13].

This study develops a multimodal framework that integrates speech commands and visual data from smart glasses to enable adaptive exoskeleton control. The research is structured in two stages. The first stage investigates the generalization capabilities of zero-shot learning with GPT-4o, comparing its performance to fine-tuned CLIP and ImageBind models. The second stage introduces a LLM-based agent augmented with short-term and long-term memory systems to address limitations, particularly regarding context awareness and command ambiguity. By employing diverse commands with increased vagueness and safety-criticality, this stage tests the agent's robustness under demanding conditions. Through these efforts, the study aims to advance human-exoskeleton collaboration, contributing to safer and more efficient construction workflows.

II. METHODOLOGY

The methodology outlines the approaches for both stages, detailing model configurations and agent architecture for locomotion prediction using speech commands and visual data.

A. Stage 1: Multimodal Locomotion Prediction

The first stage developed a framework to predict locomotion modes by combining speech commands and field-of-view (FOV) visual data, evaluating GPT-4o's zero-shot performance against fine-tuned CLIP and ImageBind models.

1) CLIP and ImageBind Supervised Fine-Tuning: The CLIP model [14], implemented as the clip-vit-large-patch14-336 variant, and the ImageBind model [15], configured as the imagebind_huge variant, were fine-tuned for locomotion prediction. Both models processed images through a vision transformer to generate high-dimensional embeddings and used a transformer-based text encoder for command embeddings. The embeddings from each modality were concatenated to form a unified representation, processed by a classification head to predict locomotion modes. The fine-tuning utilized cross-entropy loss and the Adam optimizer, with the resulting model mapping the fused representation to locomotion labels.

2) GPT-4o Zero-Shot Learning: GPT-4o [16], particularly using gpt-4o-2024-05-13, operated without task-specific training, employing Chain-of-Thought (CoT) prompting [17] to facilitate detailed reasoning. The prompt guided the model to analyze FOV sequences for movement patterns, interpret commands for intent, and provide a reasoning trace culminating in a locomotion mode prediction.

This material is based upon work supported by the National Science Foundation under Grant No. 2222881.

[1]Ehsan Ahmadi is a Ph.D. Candidate with the Bert S. Turner Department of Construction Management, Louisiana State University, Baton Rouge, LA, USA. eahmad2@lsu.edu

[2]Dr. Chao Wang is an Associate Professor with the Bert S. Turner Department of Construction Management, Louisiana State University, Baton Rouge, LA, USA. chaowang@lsu.edu

B. Stage 2: LLM-Based Agent with Memory Integration

The second stage introduced an agent to overcome Stage 1's limitations, particularly in ambiguous and safety-critical command processing, designed for rapid task transitions and robust performance. The agent integrates a LLM with a structured workflow comprising the Perception Module, Memory Modules, and Refinement Module to enhance context-awareness and safety (Figure 1).

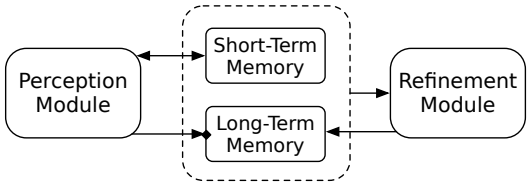


Fig. 1: Overview of the LLM-based agent architecture; the gated arrow denotes clarity-filtered storage to long-term memory

1) Perception Module: The Perception Module serves as the initial processing unit, interpreting multimodal inputs—spoken commands and FOV frames captured over a defined period—to generate an initial locomotion mode prediction. It processes the inputs using GPT-4o (gpt-4o-2024-05-13), integrating context from short-term memory structured as recent locomotion modes, environments, and primary objects users interact with, which informs safety and consistency checks, such as preventing unsafe transitions like "sitting" atop a ladder by prioritizing visual evidence over conflicting commands. The module employs a highly structured prompt, expanding upon Stage 1's general approach, to guide frame-by-frame analysis for movement direction, command interpretation with ambiguity checks, discrepancy quantification between visual and linguistic inputs, and safety analysis of transitions based on short-term memory. The output is a JSON response, including the predicted locomotion mode, a detailed CoT reasoning trace, clarity scores for vagueness, discrepancy, and confidence, environmental or scene context (e.g., indoor or outdoor settings), and identified objects or obstacles, providing a foundation for subsequent processing.

2) Memory Modules: Short-term memory functions as a transient buffer, retaining a dynamic record of recent events, including prior locomotion modes, environmental context, and objects or obstacles users interact with, within a defined temporal window. It is continuously updated with each perception output and employs a pruning mechanism to remove outdated entries, ensuring relevance for immediate context needs, such as validating a ladder descent following an ascent. Long-term memory operates as a persistent repository, storing all locomotion events as vector embeddings within a ChromaDB database, with text and image embeddings generated using text-embedding-ada-002 and clip-vit-large-patch14, respectively. Retrieval employs a multivector similarity search, calculating cosine similarity between current and stored embeddings, with weights adjusted based on discrepancy score: text similarity is weighted by $1-\text{discrepancy}$, and image similarity by discrepancy, prioritizing visual or textual cues accordingly, selecting the top-K most relevant events. Events are categorized as safety-critical or routine, with retrieval guided by a composite score:

$$(w_s \cdot \text{similarity}) + (w_i \cdot \text{importance}) + (w_c \cdot \text{confidence}) - (w_d \cdot \text{discrepancy} + w_v \cdot \text{vagueness}), \tag{1}$$

where a penalty for vagueness and discrepancy is applied, reduced for safety-critical events to ensure that events critical to safety, with lower vagueness and discrepancy, are ranked higher. Each event's importance score decays exponentially over time, with safety-critical events decaying more slowly to preserve their significance, while routine events fade more quickly, and pruning is performed to maintain memory efficiency by removing less significant entries over time; frequent retrieval of an event boosts its importance, enhancing its retention.

3) Refinement Module: The Refinement Module enhances decision-making by re-evaluating ambiguous inputs, activating when the Perception Module's clarity score—computed as:

$$w_v \cdot (1-\text{vagueness}) + w_d \cdot (1-\text{discrepancy}) + w_c \cdot \text{confidence}, \tag{2}$$

falls below a dynamic threshold escalating over evaluation cycles. It reprocesses inputs with enriched context from short-term memory and long-term memory, using a structured prompt that extends the original Perception Module prompt with insights derived from long-term memory, including the retrieved locomotion mode and a summary of command and visual details, to generate a refined response with an updated prediction, revised scores, and a reasoning trace, ensuring accuracy and safety in ambiguous or high-risk scenarios.

III. EVALUATION

The evaluation was conducted in an environment (Figure 2) designed to emulate several construction activities, utilizing Tobii Pro Glasses 2 to capture speech commands and FOV frames. All commands were transcribed using OpenAI's Whisper (medium size) [18]. The dataset encompassed twelve locomotion modes: construction ladder up climbing, construction ladder down climbing, vertical ladder up climbing, vertical ladder down climbing, level-ground navigation, low-space navigation, sitting down, standing up, stair ascension, stair descension, stepping over a box, and stepping over a pipe. In Stage 1, the dataset employed a 5-second FOV window, capturing 2 seconds before and 3 seconds after command onset, structured into 3x3 grid images, with 248 samples for training and 111 samples for testing. In Stage 2, the dataset comprised 226 samples and utilized a shorter 1.5-second FOV window (0.25 seconds before and 1.25 seconds after command onset) to minimize overlap during faster continuous transitions, while maintaining the 3x3 grid structure. To rigorously test robustness, Stage 2 includes three command types:

clear commands, such as "I'm walking" for level-ground navigation, which provided unambiguous instructions; vague commands, such as "I'm heading up" for stair ascension, which introduced interpretive challenges; and safety-critical commands, such as "I'm walking forward" when positioned atop a construction ladder, which posed significant risks if misinterpreted as level-ground navigation rather than ladder descent. These commands targeted high-risk tasks, including ladder climbing, obstacle navigation, low-space movement, and stair descension. The test data included 164 clear, 44 vague, and 18 safety-critical commands.

A. Stage 1 Performance

The evaluation of Stage 1 focused on weighted precision, recall, and F1-scores, assessing the effectiveness of multimodal inputs across the three models, as summarized in Table I. The fine-tuned CLIP model achieved the highest weighted F1-score of 90%, with a precision of 91% and a recall of 90%, reflecting robust performance due to its task-specific optimization and effective integration of visual and linguistic embeddings. The fine-tuned ImageBind model recorded a weighted F1-score of 79%, with a precision of 81% and a recall of 78%, indicating moderate performance but lower effectiveness compared to CLIP. GPT-4o, evaluated in a zero-shot setting, achieved a weighted F1-score of 88%, with a precision of 89% and a recall of 88%, demonstrating strong generalization without training, closely rivaling CLIP's performance.

TABLE I: Stage 1 Performance Metrics (Weighted Average)

Model	Precision	Recall	F1-Score
CLIP Fine-Tuned	91%	90%	90%
ImageBind Fine-Tuned	81%	78%	79%
GPT-4o Zero-Shot	89%	88%	88%

These results highlight GPT-4o's competitive performance compared to supervised fine-tuning, underscoring the potential of multimodal LLMs for zero-shot learning from users' vision and speech in complex and diverse locomotion prediction scenarios typical of construction environments. However, its effectiveness was constrained by difficulties in resolving ambiguous commands and the lack of temporal context, which impacted overall robustness at this stage. These limitations underscore the need for further investigation into adaptive and context-aware LLM agents, which were explored in Stage 2.

B. Stage 2 Performance

Stage 2's evaluation was significantly more demanding, designed to test the agent's robustness under conditions that more closely mimicked the complexities of real-world construction environments. The 1.5-second FOV window challenged the agent to interpret rapid task transitions, while the inclusion of clear, vague, and safety-critical commands introduced varying levels of ambiguity and risk. Performance was assessed using weighted precision, recall, and F1-scores, supplemented by Brier Score and Expected Calibration Error (ECE) to evaluate prediction reliability.

Ablation studies compared three configurations: a no-memory baseline relying solely on the Perception Module without memory, a short-term memory-only setup, and the full system integrating both short-term and long-term memory, as summarized in Table II. The no-memory baseline achieved a weighted F1-score of 73%, with a precision of 81% and a recall of 70%, reflecting limitations in processing ambiguous or discrepant inputs without contextual support. The short-term memory-only configuration improved the F1-score to 81%, with a precision of 86% and a recall of 81%, as recent events facilitated smoother transitions. The full system, incorporating both memory types, attained a weighted F1-score of 90%, with a precision of 92% and a recall of 90%, demonstrating the synergistic effect of immediate and historical context in enhancing prediction accuracy.

TABLE II: Stage 2 Ablation Metrics (Weighted Average)

Configuration	Precision	Recall	F1-Score
No Memory	81%	70%	73%
Short-Term Memory Only	86%	81%	81%
Full Memory	92%	90%	90%

Calibration metrics, presented in Table III, further underscored the agent's reliability. The Brier Score decreased from 0.244 in the no-memory condition to 0.169 with short-term memory only, and further to 0.090 with full memory integration, indicating improved prediction calibration. Similarly, the Expected Calibration Error dropped from 0.222 to 0.133 and then to 0.044, reflecting a robust alignment between prediction confidence and actual outcomes.

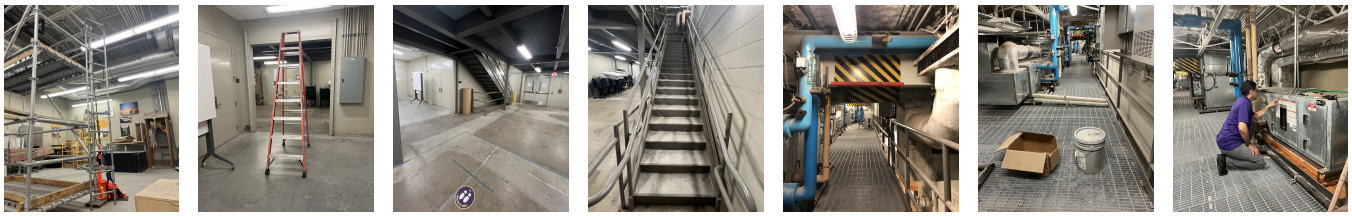
TABLE III: Stage 2 Calibration Metrics

Configuration	Brier Score	ECE
No Memory	0.244	0.222
Short-Term Memory Only	0.169	0.133
Full Memory	0.090	0.044

Performance across command types, as illustrated in Fig. 3, provided insights into the differential impact of memory systems under Stage 2's varied conditions. For clear commands, the full system achieved an F1-score of 94%, approaching perfect accuracy due to their unambiguous nature. Vague commands saw a substantial improvement from 69% in the no-memory condition to 83% with short-term memory and 86% with the full memory system, highlighting short-term memory's effectiveness in resolving most ambiguities by leveraging recent activity context. Safety-critical commands improved from 38% to 72%, indicating substantial progress with the full memory system, though the task remains inherently challenging.

IV. CONCLUSION

This study presents a multimodal framework that significantly advances high-level exoskeleton control through pre-



(a) Vertical Ladder (b) Construction Ladder (c) Level-ground (d) Stairs (e) Low Space (f) Obstacles (g) Sitting (& Standing)

Fig. 2: Environment Setting

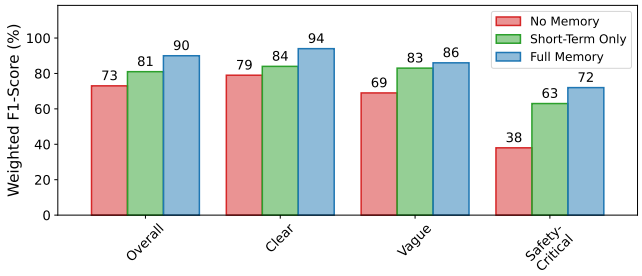


Fig. 3: Weighted F1-Score Across Command Types for LLM Agent

cise locomotion prediction in construction environments. The first stage validated the potential of zero-shot learning with GPT-4o, achieving performance comparable to fine-tuned CLIP. The second stage introduced a memory-augmented large language model-based agent, attaining a weighted F1-score of 90% and robust calibration under rigorous conditions such as complex command scenarios encompassing clear, vague, and safety-critical instructions. Future research should enhance the Perception Module's clarity scoring for high-discrepancy cases, refine intent recognition with user feedback, explore retrieval strategies using knowledge graphs or scene graphs, and conduct real-time evaluations in live construction settings to ensure robustness. These advancements will strengthen the framework's applicability for safe and efficient human-exoskeleton collaboration in dynamic construction workflows.

REFERENCES

[1] S. Kim, A. Moore, D. Srinivasan, A. Akanmu, A. Barr, C. Harris-Adamson, D. M. Rempel, and M. A. Nussbaum, "Potential of exoskeleton technologies to enhance safety, health, and performance in construction: Industry perspectives and future research directions," *IISE Transactions on Occupational Ergonomics and Human Factors*, vol. 7, no. 3-4, pp. 185–191, 2019.

[2] M. I. Al-Khiami, S. M. Lindhard, and S. Wandahl, "Integrating exoskeletons in the construction sector: a systematic review of empirical evaluation tools and future directions," *Engineering, Construction and Architectural Management*, 2024.

[3] M. R. Tucker, J. Olivier, A. Pagel, H. Bleuler, M. Bouri, O. Lambercy, J. d. R. Millán, R. Riener, H. Vallery, and R. Gassert, "Control strategies for active lower extremity prosthetics and orthotics: a review," *Journal of neuroengineering and rehabilitation*, vol. 12, pp. 1–30, 2015.

[4] R. Baud, A. R. Manzoori, A. Ijspeert, and M. Bouri, "Review of control strategies for lower-limb exoskeletons to assist gait," *Journal of NeuroEngineering and Rehabilitation*, vol. 18, pp. 1–34, 2021.

[5] O. Coser, C. Tamantini, P. Soda, and L. Zollo, "Ai-based methodologies for exoskeleton-assisted rehabilitation of the lower limb: a review," *Frontiers in Robotics and AI*, vol. 11, p. 1341580, 2024.

[6] D. Pinto-Fernandez, D. Torricelli, M. del Carmen Sanchez-Villamanan, F. Aller, K. Mombaur, R. Conti, N. Vitiello, J. C. Moreno, and J. L. Pons, "Performance evaluation of lower limb exoskeletons: a systematic review," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 7, pp. 1573–1583, 2020.

[7] OpenAI, "Gpt-4v system card," https://openai.com/index/gpt-4v-system-card/, 2023, accessed: 2025-04-16.

[8] G. Team, P. Georgiev, V. I. Lei, R. Burnell, L. Bai, A. Gulati, G. Tanzer, D. Vincent, Z. Pan, S. Wang *et al.*, "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," *arXiv preprint arXiv:2403.05530*, 2024.

[9] T. B. Brown, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.

[10] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," *Advances in neural information processing systems*, vol. 35, pp. 22 199–22 213, 2022.

[11] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, "Finetuned language models are zero-shot learners," *arXiv preprint arXiv:2109.01652*, 2021.

[12] Z. Xi, W. Chen, X. Guo, W. He, Y. Ding, B. Hong, M. Zhang, J. Wang, S. Jin, E. Zhou *et al.*, "The rise and potential of large language model based agents: A survey," *Science China Information Sciences*, vol. 68, no. 2, p. 121101, 2025.

[13] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin *et al.*, "A survey on large language model based autonomous agents," *Frontiers of Computer Science*, vol. 18, no. 6, p. 186345, 2024.

[14] 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.

[15] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, "Imagebind: One embedding space to bind them all," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 180–15 190.

[16] OpenAI, "Hello gpt-4o," May 2024, accessed April 16, 2025. [Online]. Available: https://openai.com/index/hello-gpt-4o

[17] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[18] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.