Journal of INTELLIGENT CONSTRUCTION

Research Article
https://doi.org/10.26599/JIC.2026.9180128

Exoskeleton locomotion mode prediction in construction using GPT-4o: Zero-shot learning from vision and speech

Ehsan Ahmadi[a,(✉)], Asif Faisal Chowdhury[a,(✉)], Chao Wang[a,(✉)], Mahmood Jasim[b]

[a] *Bert S. Turner Department of Construction Management, Louisiana State University, Baton Rouge, LA 70803, USA*

[b] *Division of Computer Science and Engineering, Louisiana State University, Baton Rouge, LA 70803, USA*

Received: 2025-11-24 ***Revised:*** 2026-03-14 ***Accepted:*** 2026-03-23

ARTICLE INFO

Keywords
locomotion prediction
exoskeleton
multimodal sensor fusion
computer vision
natural language processing

ABSTRACT

Wearable exoskeletons enhance mobility and support in demanding tasks but face challenges in adapting to dynamic construction environments, particularly in predicting locomotion modes for tasks such as ladder climbing, stair navigation, low-space movement, and obstacle navigation. This study investigates the effectiveness of integrating speech and vision data for locomotion prediction while evaluating the generalization capability of large language models, specifically GPT-4o, through zero-shot learning compared to supervised fine-tuning. Using a multimodal framework with field-of-view frames and speech commands captured by smart glasses, we tested contrastive language-image pre-training (CLIP), ImageBind, and GPT-4o. Fine-tuned CLIP achieved an F1-score of 90.05% , yet GPT-4o's zero-shot of 87.87% closely rivaled it, demonstrating strong adaptability to construction's complex demands without task-specific training, while fine-tuned ImageBind trailed at 78.84%. This comparison underscores GPT-4o's substantial potential to enable scalable exoskeleton control by leveraging multimodal comprehension of vision and speech data in dynamic construction settings.

1 Introduction

The construction industry faces pressing challenges related to labor shortages, high physical demands, and stringent safety requirements, all of which are amplified by the dynamic and often unpredictable nature of construction sites. Exoskeletons, defined as wearable devices that augment, enable, assist, and/or enhance physical activity, have been identified as promising tools to address these challenges [1]. These devices provide substantial physical advantages by reducing muscular effort and minimizing the risk of injuries or musculoskeletal disorders associated with tasks such as heavy material handling and repetitive movements [1, 2]. Despite their potential, the evaluation of exoskeletons, especially lower limb devices, for advanced motor skills remains underrepresented in the literature, with most studies focusing on daily gait tasks such as flat ground walking [3]. This gap appears even more pronounced in construction, which involves more advanced and diverse locomotion tasks.

In recent years, exoskeletons have been increasingly explored for their ability to enhance human capabilities across various sectors, including rehabilitation, military, and industrial applications. These wearable mechanical devices assist movement and often incorporate sensors to provide real-time feedback and control. Initially, research in exoskeleton development focused on optimizing mechanical aspects, such as improving power efficiency,

✉ Address correspondence to Ehsan Ahmadi, eahmad2@lsu.edu; Asif Faisal Chowdhury, achowd6@lsu.edu; Chao Wang, chaowang@lsu.edu

Citation: Ehsan Ahmadi, Asif Faisal Chowdhury, Chao Wang, et al. Exoskeleton locomotion mode prediction in construction using GPT-4o: Zero-shot learning from vision and speech. *J Intell Constr*, 2026, 4, 9180128.

© The Author(s) 2026. The articles published in this open access journal are distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

reducing weight, and refining actuation systems to provide effective support, particularly for users with physical impairments [4]. However, as these systems evolved, there has been a growing interest in integrating more sophisticated control mechanisms that rely heavily on sensor input. These sensors, ranging from inertial measurement units (IMUs) to force and vision sensors, are essential for allowing exoskeletons to adapt dynamically to the user's movement and the environment [5, 6].

One of the most significant advancements in exoskeleton technology has been the development of sensor fusion techniques that integrate multimodal data to create more robust and adaptive control strategies, enhancing user interaction and responsiveness [5]. Sensor fusion enables real-time decision-making, allowing exoskeletons to accurately predict and respond to user intent, which is crucial for seamless transitions between different locomotion modes, such as walking, climbing stairs, or navigating ramps [5]. Concurrently, research has progressed towards more interactive and intelligent systems capable of adapting to users' needs. Human–robot interaction (HRI) has become a focal point, emphasizing the integration of human intent with robotic assistance. Recent studies have shown that interactive control mechanisms can significantly improve responsiveness by predicting user intentions in real time, dynamically adjusting support based on user intent and environmental factors, and offering adaptive assistance during various locomotion activities [6, 7].

Despite significant advancements in exoskeleton control and locomotion prediction, several challenges persist, especially in complex and dynamic environments like construction settings. These environments introduce unpredictability, making it difficult for exoskeletons to adapt seamlessly to changing conditions and tasks. Construction and industrial settings are marked by complex tasks, such as ladder climbing, obstacle navigation, and low-space movement, which require exoskeletons to handle a broader range of demanding movements beyond daily activities like walking. To address these challenges, multimodal sensor fusion is essential for improving responsiveness to user needs; however, the integration of human speech and vision inputs for locomotion prediction has not been fully explored. Speech conveys user intent as a natural form of communication, while vision provides critical environmental context, and their combination offers a framework for more intuitive and accurate predictions of user locomotion.

The objective of this research is to address the challenges posed by dynamic environments through a novel multimodal framework that integrates both speech and visual data. Specifically, this study aims to enhance the accuracy of locomotion mode prediction by processing human speech commands and field-of-view (FOV) frames captured by smart glasses, exploring the multimodal understanding and generalization capability of large language models (LLMs) through a comparative analysis of zero-shot (ZS) learning versus supervised fine-tuning.

To achieve these objectives, a comprehensive methodology was implemented, utilizing smart glasses equipped with a camera and microphone to collect multimodal data during various locomotion tasks. The visual data from FOV frames were synchronized with corresponding speech commands, providing a robust dataset for analysis. Three state-of-the-art models, contrastive language-image pre-training (CLIP), ImageBind, and GPT-4o, were employed to process the multimodal inputs. These models were evaluated in both zero-shot and fine-tuned (FT) settings, with data generation and augmentation techniques used to address class imbalances and improve model robustness. By applying this methodology, the study demonstrated how the multimodal data fusion of speech and vision could synergistically enhance the prediction accuracy for complex locomotion tasks in dynamic environments and whether LLMs like GPT-4o can effectively tackle such tasks without task-specific training.

This research contributes to the development of more intuitive and interactive wearable exoskeletons, capable of real-time adaptation in the unpredictable and demanding environments found in construction settings, thereby improving safety and overall user experience.

In safety-critical lower-limb exoskeleton control, prediction quality is not an abstract metric: errors and late transitions translate into incorrect assistance profiles or delayed mode switching, which can be destabilizing during stairs, ladders, and other transitions. Prior exoskeleton and assistive-locomotion studies emphasize that real-time locomotion mode recognition and timely transition detection are critical for safe operation, and report that transition detection can incur substantial delay when mode recognition is not sufficiently respon- sive [8]. While this study does not directly measure injuries avoided, we interpret our model performance as a practical proxy for reducing the rate of wrong-mode assistance events. If a deployed controller makes N mode decisions per shift and the probability of a wrong decision is p_{err}, the expected number of wrong-mode events is $N{\cdot}p_{\mathrm{err}}$. A reduction in p_{err} therefore scales linearly to fewer hazardous mis-assistance events, especially relevant for ladder and stair tasks, which are among the most safety-critical movements in construction [9].

In this manuscript, "dynamic construction environments" refers to conditions that change rapidly and degrade single-modality observability. At the sensor level, this includes: variable illumination (e.g., indoor/outdoor transitions and low-light/nighttime work), motion-induced blur from head/torso movement, and frequent occlusions from structural elements, tools, and other

workers. Construction computer-vision reviews and recent studies identify illumination variation and occlusion as recurring limitations that degrade recognition accuracy, motivating robustness strategies and multimodal sensing [10]. At the task level, "dynamic" also includes frequent transitions between locomotion modes (e.g., level ground → stairs → obstacles → ladder) and heterogeneous site geometries that cause the same locomotion label to appear with different visual affordances and backgrounds, increasing ambiguity for vision-only prediction [11].

2 Related works

In the context of exoskeleton control, locomotion prediction is central to achieving synchronized interaction between the device and its user. These tasks ensure that the exoskeleton can anticipate changes in movement patterns and adapt dynamically, enhancing usability, comfort, and safety. Locomotion prediction involves identifying different modes, such as walking, stair climbing, and ramp navigation, as well as detecting gait phases and managing transitions between movements [4, 12, 13].

To achieve accurate locomotion prediction, various techniques have been applied, ranging from traditional methods to advanced multimodal systems. Traditional machine learning methods, such as support vector machines (SVMs) and decision trees, are frequently employed for locomotion classification tasks. For example, Long et al. [14] used foot pressure sensors and an attitude and heading reference system (AHRS) to classify locomotion modes, including level-ground walking, stair ascent, descent, ramp ascent, descent, and transitions between these modes, using SVM optimized with particle swarm optimization (PSO) and majority vote algorithms. Similarly, Parri et al. [15] used onboard hip joint encoders and plantar pressure-sensitive insoles for tasks like sitting, standing, and walking, applying a combination of time-based decoding and event-based fuzzy logic for dynamic mode transitions. Kim et al. [16] applied IMUs and load cells to recognize modes such as walking on flat ground, stair ascent and descent, and ramp ascent and descent, using a two-layer decision tree optimized for recognizing foot inclination at midstance. Liu et al. [17] developed an SVM-based hierarchical control system for predicting real-time assistive torque for walking, standing, and transitioning tasks, utilizing IMUs and pressure insoles. Zhao et al. [18] combined electromyography (EMG) and plantar stress distribution (PSD) sensors with an SVM-based classification system for recognizing walking, stair, and ramp ascent tasks, integrated within an Internet of Health Things (IoHT) framework. Al-Dabbagh et al. [19] used depth cameras and SVM models to classify flat-ground walking, ramp ascent and descent, and stair navigation, with a focus on predicting locomotion modes and estimating terrain geometries.

In recent years, advanced neural network-based models have become more prominent due to their capacity to handle complex datasets and improve locomotion mode prediction. Zhu et al. [20] utilized IMUs for classifying level-ground walking, stair ascent and descent, and ramp navigation, relying on a neural network model that integrated historical information for pose and terrain recognition. Laschowski et al. [21] applied deep convolutional neural networks (DCNNs) using EfficientNet, InceptionV3, and MobileNet to classify locomotion environments, focusing on indoor and outdoor walking tasks, stairs, and transitions, relying on camera-based data. Guo et al. [22] developed a genetic algorithm-convolutional neural network (GA-CNN) model for locomotion mode recognition and gait phase estimation, employing speech commands to manage tasks such as sitting, standing, walking, speeding up, and slowing down on various terrains. Martínez-Pascual et al. [23] utilized one-dimensional convolutional neural networks (1D-CNNs) for gait activity classification, specifically for walking, ramp ascent, and stair climbing, using joint angle trajectories, with Grad-CAM and t-SNE employed for interpretability.

In addition to standalone models, multimodal approaches have gained traction, combining multiple sensor types to enhance prediction accuracy. Li et al. [24] integrated eye-tracker and depth camera data to predict locomotion modes, including walking on flat ground, stair ascent and descent, and ramp navigation, using deep learning models (MobileNetV1 and PointNet) with dynamic time warping (DTW) for sensor fusion. Qian et al. [11] combined RGBD camera data with IMUs, using a depth sensor-based locomotion mode recognition (LMR) method and an adaptive oscillator-based gait phase estimation (GPE) technique for locomotion and transitions. Sharma et al. [25] used long short term memory (LSTM) networks to predict knee and ankle joint angles, relying on IMUs and egocentric optical flow data from smart glasses, focusing on walking tasks in environments such as classrooms, atriums, and stairs. Tsepa et al. [26] used the data from Ref. [25], applying KIFNet, which fuses inertial sensors and smart glasses data. MobileOne-S0 serves as the image encoder, and a multilayer perceptron (MLP) encodes the kinematic data for continuous knee and ankle joint kinematics prediction. Wang et al. [27] adopted a GA-CNN model that used IMUs and six-axis force sensors for real-time locomotion mode recognition, optimized using Bayesian methods for transitions between tasks such as walking, standing, sitting, and squatting. Lastly, Li et al. [28] developed a multimodal deep belief network (DBN) that fuses accelerometer, gyroscope, and surface electromyography (sEMG) data, applying three fusion strategies (data, feature, and deci-

sion level) to predict both steady-state locomotion and transitions.

However, despite advancements in this field, current multimodal approaches often overlook intuitive modalities like speech and vision, which provide valuable insights into user intent and environmental context. Instead, many methods rely primarily on sensors such as IMUs or pressure insoles, often employing supervised learning on structured and repetitive tasks, which, while effective for specific locomotion modes, may struggle to address the diversity of tasks in dynamic environments like construction. Tasks such as ladder climbing, obstacle navigation, and low-space movement require a richer understanding of both user intent and the surrounding environment. Incorporating speech and vision data offers a more holistic approach, enabling more accurate and adaptable locomotion mode predictions in these complex settings, while examining the potential of large language models to address such tasks could shed light on their ability to generalize and scale in such settings.

Multimodal locomotion recognition literature in rehabilitation and prosthetics commonly assumes comparatively struc tured environments (e.g., standardized stairs/ramps, fewer occluders, and more consistent sensing) and often emphasizes IMU/EMG fusion to detect transitions. In contrast, construction environments introduce heavier occlusion, greater illumination variability, and heterogeneous affordances that can make visually similar scenes correspond to different locomotion risks (e.g., stair-like patterns vs. ladder rungs). Construction computer-vision reviews highlight occlusion and illumination variation as frequently reported limitations; these factors are more prevalent on jobsites than in many lab-based mobility benchmarks [9]. Therefore, models validated in rehabilitation-like settings may degrade in construction unless they explicitly address these domain-specific distribution shifts and include uncertainty-aware safeguards for ambiguous observations [10]. We selected CLIP, ImageBind, and GPT-4o to span three distinct capability points relevant to deployable construction exoskeleton intent recognition: (i) contrastive vision-language embedding that supports lightweight fine-tuning and edge deployment (CLIP), (ii) multimodal embedding alignment beyond vision-text (ImageBind), and (iii) general-purpose multimodal reasoning in a powerful zero-shot model (GPT-4o). CLIP is widely used for transferable vision-language representations trained via natural-language supervision, enabling strong zero-shot baselines and efficient adaptation [29]. GPT-4o was chosen for benchmarking because it is a flagship multimodal model with strong few/zero-shot capability documented in its system card [30], and it is practically accessible via application programming interface (API) for rapid iteration and reproducible prompting. We explicitly considered alternatives such as LLaVA [31] and Flamingo [32]: LLaVA is a major open research baseline for image-language instruction tuning, but its standard form is image–text only (not native audio) and typically requires substantial local compute for stable deployment; Flamingo is a seminal few-shot vision-language model architecture but is not generally distributed as a turnkey on-device baseline in the same way prac-titioners can access GPT-4o via API.

3 Proposed research methodology

Figure 1 illustrates the multimodal locomotion mode prediction framework developed for this study. The framework is designed to operate within a simulated environment, where participants, equipped with smart glasses, perform various construction-related tasks. These smart glasses capture both the participant's FOV frames and voice commands during locomotion activities.

These verbal commands issued by the users trigger the sampling of corresponding FOV frames, ensuring that the system captures both the linguistic and visual aspects of the task. The spoken commands are then processed through a speech-to-text (STT) system, which converts them into textual data. The processed text data is combined with the visual information from the FOV frames and fed into a multimodal prediction model. The model leverages both the linguistic input and the visual data to predict the participant's intended locomotion mode,

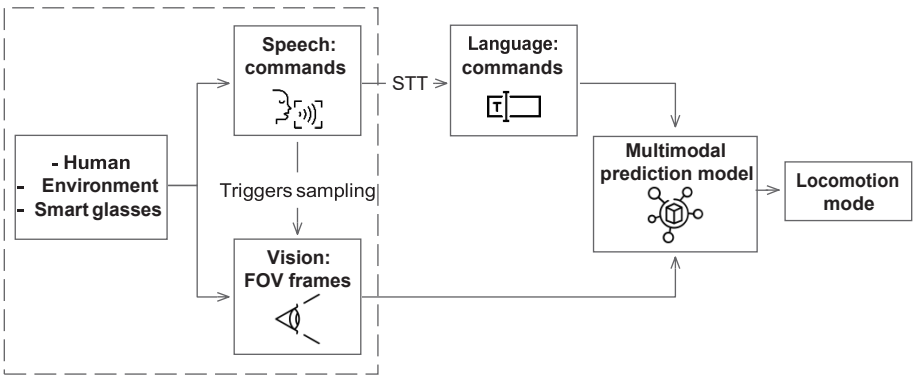


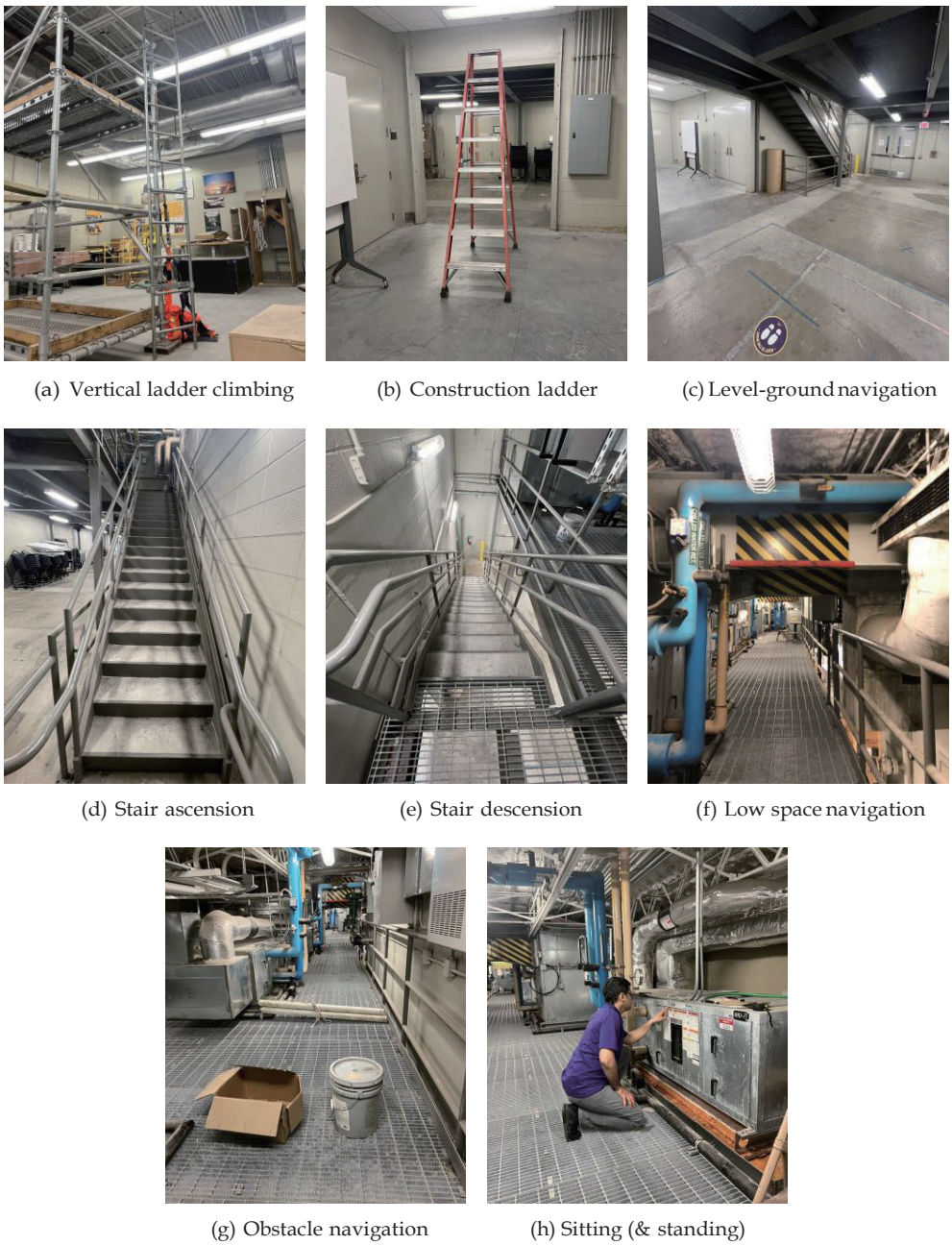
Fig. 1 Locomotion mode prediction framework.

such as ladder climbing, obstacle navigation, or stair ascension.

3.1 Data collection

The data was collected in a controlled setting designed to mimic real-world construction activities, such as vertical and construction ladder climbing, level-ground walking, stair ascension and descension, low-space navigation, obstacle navigation, sitting, and standing (Fig. 2). These tasks represent common movements encountered in construction workflows. The environment had noise levels akin to typical construction sites, including a mechanical mezzanine area with ambient noise levels reflective of machinery and site operations.

The selection of locomotion activities in this study is grounded in construction safety standards and occupational injury data. Falls remain the leading cause of death in the construction industry, and the Fall Prevention Campaign of Occupational Safety and Health Administration (OSHA) identifies ladders and stairs as major contributors to fatal falls [8]. The OSHA standard for ladders in construction (29 CFR 1926.1053) further specifies safety requirements for ladder use on construction sites [33]. Accordingly, our locomotion class list was designed to cover the most safety-critical movement transitions encountered on construction sites: vertical and construction ladder climbing (up and down), stair ascension and descension, level-ground and low-space



(a) Vertical ladder climbing (b) Construction ladder (c) Level-ground navigation

(d) Stair ascension (e) Stair descension (f) Low space navigation

(g) Obstacle navigation (h) Sitting (& standing)

Fig. 2 Environment setting.

navigation, obstacle navigation (stepping over boxes and pipes), and stationary postures (sitting and standing). These classes represent the primary locomotion modes that an exoskeleton controller must reliably distinguish to provide appropriate assistance and prevent hazardous mis-assistance events during mode transitions.

The primary data source was the Tobii Pro Glasses 2, which recorded FOV frames and voice commands during locomotion tasks. These combined data streams formed the core dataset for locomotion mode prediction.

Table 1 details the sensing hardware and sampling parameters used in this study to ensure reproducibility. The Tobii Pro Glasses 2 scene camera captures full-HD video (1920 pixels (px) × 1080 px) at 25 frames per second (fps), providing sufficient spatial resolution and temporal sampling to observe locomotion-relevant environmental features [34]. Audio was recorded via the integrated microphone of the Tobii Pro Glasses 2; the audio sample rate was extracted from the recorded media container metadata and resampled to 16,000 Hz during preprocessing for compatibility with the Whisper speech-to-text model.

3.1.1 Voice commands

Participants provided short voice commands (e.g., "I'm climbing the ladder" or "I'm walking") before initiating tasks, which were considered to reflect their intent and trigger vision sampling.

3.1.2 FOV frames

A 5-s window of FOV frames was sampled for each command, starting 2 s before and extending 3 s after the command onset to capture the anticipatory phase and the initial engagement with the task. This window size was determined empirically based on the average duration of intention formation and action initiation observed during pilot trials: The 2-s pre-command interval captures preparatory visual cues (e.g., the user orienting toward the target terrain), while the 3-s post-command interval spans the initial engagement with the locomotion task. This asymmetric design ensures that the model receives both anticipatory context and early execution cues, while keeping the computational burden manageable for real-time applications. Shorter windows would risk missing the transition context, while substantially longer windows would introduce unrelated visual frames and increase inference latency. Variability in participant height, pace, and head movement sometimes introduced diversity in vision data, providing realistic variations.

Table 1 Sensor reproducibility parameters

Parameter	Value
Scene camera resolution	1920 px × 1080 px (full HD)
Scene camera frame rate	25 fps
FOV window duration	5 s (2 s pre-command and 3 s post-command)
Frames per window	9 (3 × 3 grid)
Audio recording	Integrated microphone
Audio preprocessing rate	16,000 Hz
STT model	Whisper medium (769M params)

The collected data were annotated using ELAN [35] for precise time-aligned video and audio annotations. Both start and end time for each locomotion activity and speech command were marked. Table 2 summarizes the studied locomotion activities. Figure 3 shows examples of FOV frames captured during activities, illustrating participants' perspectives through the Tobii Pro Glasses 2.

3.2 Data generation and augmentation

3.2.1 Speech command generation

To enrich the dataset and address the variability in natural language communication, command generation was performed using GPT-4 [36], a state-of-the-art large language model, to create 100 commands per locomotion activity through few-shot learning [37]. From these, 20–35 commands per class were selected for diversity and relevance, including examples of ambiguous communication.

Speech synthesis was conducted using the XTTS v2 [38] model to convert selected text commands into speech, generating multiple versions with varied voices to simulate different users. Environmental noise, extracted from silent portions of the original recordings, was added to the synthesized speech to simulate real-world conditions. A final selection of 15–20 audio samples per command class was made.

Table 2 Studied locomotion activities

#	Locomotion activity
1	Construction ladder down climbing
2	Construction ladder up climbing
3	Vertical ladder down climbing
4	Vertical ladder up climbing
5	Level-ground navigation
6	Low space navigation
7	Sitting
8	Standing
9	Stair ascension
10	Stair descension
11	Stepping over box
12	Stepping over pipe



(a) Low space navigation (b) Sitting (c) Stepping over box (d) Vertical ladder down climbing

Fig. 3 Examples of FOV frames for locomotion activities captured during the experiment.

3.2.2 *Data augmentation*

To address class imbalances, the RandomOverSampler technique from the Imbalanced-learn library [39] was used. Classes with fewer than 20 samples were oversampled after selecting the test set, as shown in Fig. 4. FOV frame augmentation involved applying color jitter to introduce variations in brightness, contrast, saturation, and hue. Generated speech commands were used alongside augmented FOV frames to ensure a balanced representation of both audio and visual data. Figure 4 shows the distribution of the training and test sets. We acknowledge that oversampling classes with fewer than 20 samples carries a risk of overfitting, particularly for the fine-tuned CLIP and ImageBind models, because the model may memorize duplicated or near-duplicate training examples rather than learning generalizable features. To mitigate this risk, the oversampling was performed strictly on the training set after a stratified train-test split, ensuring that the held-out test set contains only original (non-duplicated) samples with a distribution representative of the true class proportions (Fig. 4). Additionally, the data augmentation pipeline (color jitter, generated speech commands) introduced controlled variation into the oversampled examples, reducing the likelihood of exact memorization. We recommend that future work report per-class performance with confidence intervals and consider cross-validation strategies to further assess the robustness of models trained on oversampled data.

3.3 *Data preprocessing*

Raw data modalities (text, image, and audio) were prepared for subsequent analysis to ensure compatibility with the models. For visual data, a 3 × 3 grid of FOV frames was generated, spanning a 5-s window, 2 s before and 3 s after the voice command, as shown in Fig. 3. These grid images were standardized to 550 pixels × 550 pixels. Temporal sequence analysis was also performed, processing individual frames sequentially. Voice commands were transcribed into text using OpenAI's Whisper model (medium size) [40]. When audio data were used directly, they were resampled to 16,000 Hz for consistency.

3.3.1 *Acoustic environment characterization*

Construction acoustics are a first-order confounder for

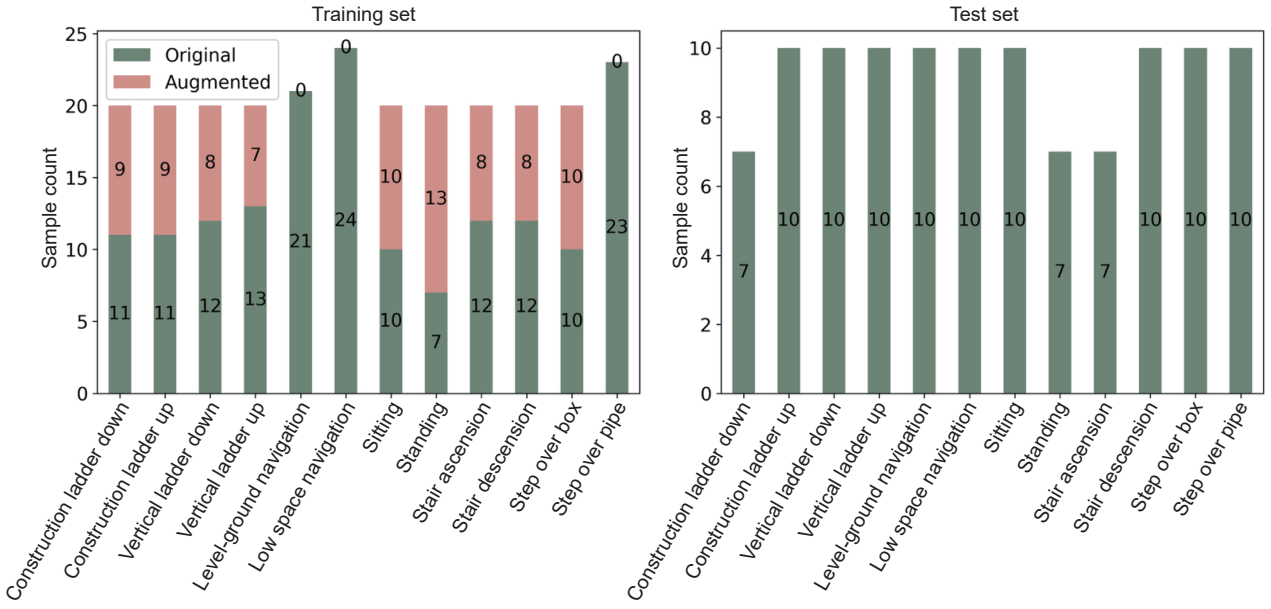


Fig. 4 Class distribution of locomotion modes in training and test sets.

speech-driven intent recognition because common jobsite noise frequently meets or exceeds hazardous exposure thresholds. According to NIOSH guidance, repeated exposures at or above 85 dBA are considered hazardous, and OSHA construction hearing conservation requirements apply when exposures exceed 90 dBA time-weighted average (TWA) [41]. We therefore characterize the recording environment using both sound level and frequency content, following occupational-noise measurement guidance [42]. Ambient noise levels in the mezzanine laboratory were measured using an Extech 407730 digital sound level meter (measurement range: 40–130 dB; accuracy: ±2 dB; resolution: 0.1 dB; frequency response: 300 Hz–8 kHz; and 0.5-inch electret condenser microphone with foam wind-screen). The meter was set to A-weighting and FAST response mode to approximate human hearing sensitivity and capture short-duration transient events. The representative ambient sound pressure level recorded during data collection was L_{Aeq} = 73.6 dBA, which falls within the moderate industrial noise range, typical of indoor construction training facilities where dominant sources include HVAC systems, footfall on steel grating, and intermittent tool handling. At 73.6 dBA, spoken voice commands issued at normal-to-raised vocal effort (approximately 65–80 dBA at the source) yield a signal-to-noise ratio (SNR) near 0–6 dB at the body-worn microphone. This represents a moderately challenging but realistic operating condition for the Whisper medium speech-to-text model. This acoustic characterization contextualizes the STT accuracy reported in Section 5 and informs the noise-robustness requirements discussed in the deployment considerations in Section 6. This noise profile is representative of construction sites where speech-based exoskeleton control must operate, and it informed the design of our speech preprocessing pipeline, including the addition of extracted environmental noise to synthesized speech commands during data augmentation.

3.3.2 Speech-to-text latency and real-time feasibility

For real-time exoskeleton control, end-to-end latency must be explicitly budgeted. We decompose total decision latency as

$$t_{total} = t_{STT} + t_{vision} + t_{model} + t_{comm} \quad (1)$$

where t_{STT} is the speech-to-text conversion time, t_{vision} is the frame extraction and preprocessing time, t_{model} is the inference time of the prediction model, and t_{comm} is the communication overhead. The Whisper medium model (769M parameters) represents a common accuracy–speed compromise, requiring approximately 5 GB VRAM with documented relative speed characteristics varying by hardware configuration [40, 43]. To validate the feasibility of real-time command processing, we benchmarked the speech-to-text pipeline on a workstation equipped with an NVIDIA GeForce RTX 4070 SUPER (12 GiB VRAM) and an Intel Core i9-class CPU (8 cores, 5.05 GHz). For the Whisper Medium model (769M parameters), the average inference latency for transcribing a 5-s audio buffer was estimated at approximately 200 ± 50 ms using FP16 precision. This low-latency profile (real-time factor ≈ 0.04) confirms that the STT component consumes less than 15% of the available processing budget in our 0.5 Hz high-level control loop, ensuring that voice command parsing does not introduce prohibitive delays into the exoskeleton's decision-making process.

3.4 Multimodal prediction modeling

This section outlines the models used for locomotion mode classification, leveraging FOV frames, voice commands, and audio data. We employed three state-of-the-art models: CLIP [29], ImageBind [44], and GPT-4o [36].

3.4.1 CLIP zero-shot

CLIP, developed by OpenAI, is designed to learn joint visual and textual representations from natural language supervision on a vast dataset of 400 million (image and text) pairs [29]. Leveraging the semantic alignment between images and language, CLIP enables zero-shot classification by generalizing to new tasks without additional training data. In this study, we used a pre-trained CLIP model as a baseline to compare against other models and settings. To implement zero-shot classification, cosine similarity was calculated separately between the image and text embeddings against label embeddings. In the multimodal approach, when both image and text embeddings were used, similarity scores from each modality were averaged to yield a combined prediction.

3.4.2 CLIP fine-tuned

CLIP was also fine-tuned by adding a task-specific classification head. Image and text embeddings were mapped to locomotion labels through a fully connected (FC) layer. Both 3 × 3 grid images and sequences of frames were processed to capture temporal dependencies across frames. For sequences of frames, a temporal aggregation mechanism utilizing positional embeddings and multi-head self-attention, similar to the Video Q-Former approach [45], was employed. Image and text embeddings were also concatenated for multimodal classification (Eq. (2)).

$$f_{concat} = \text{concat}(f_{image}, f_{text}), \hat{y} = \text{softmax}(W \cdot f_{concat} + b) \quad (2)$$

where W and b are the weight matrix and bias term of the FC layer, respectively, f_{image} is the image embedding, and f_{text} is the text embedding.

The training was driven by cross-entropy loss and optimized with Adam using learning rate scheduling.

3.4.3 *ImageBind fine-tuned*

ImageBind, developed by Meta AI, expands multimodal learning capabilities by integrating visual, auditory, inertial, and depth sensor data, creating a single joint embedding space for multiple modalities [44]. This model is particularly suited for tasks involving multiple sensor inputs, like locomotion prediction, where visual data alone may not capture the full context. We fine-tuned it with image–text and image–audio modalities by concatenating respective embeddings (Eq. (3)).

$$f_{\text{fusion}} = \text{concat}(f_{\text{image}}, f_{\text{text}}) \text{ or } f_{\text{fusion}} = \text{concat}(f_{\text{image}}, f_{\text{audio}}) \quad (3)$$

where f_{image}, f_{text}, and f_{audio} are the respective embeddings for image, text, and audio modalities, respectively. The fused representation f_{fusion} was then passed to the classification head to predict the locomotion activity label. Fine-tuning was conducted using the same optimization strategy as for CLIP, with cross-entropy loss and Adam optimizer guiding the training process.

3.4.4 *GPT-4o zero-shot*

GPT-4o, based on the GPT-4 architecture developed by OpenAI [36], demonstrates strong performance in tasks requiring contextual reasoning, language comprehension, and multimodal understanding. For locomotion prediction, GPT-4o used FOV frames and user commands, employing chain-of-thought (CoT) prompting [46, 47] to generate step-by-step reasoning, which is crucial for interpreting temporal changes in 3 × 3 grid images and the user intent in voice commands. We specifically asked the model to provide a reasoning path for its final prediction in the system prompt, as shown in Fig. 5. The figure also indicates the prompts used for different settings: image-only, text-only, and image–text.

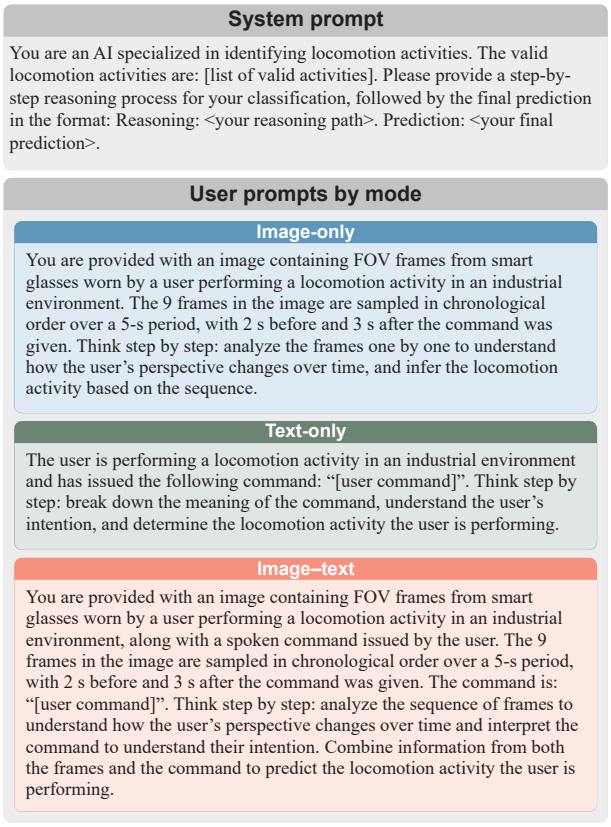


Fig. 5 CoT prompting with GPT-4o.

4 Training setup and model evaluation

This section outlines the model configurations, training setup, hyperparameters, and experiments conducted to evaluate image, text, and audio modalities for locomotion prediction using CLIP, ImageBind, and GPT-4o. Various evaluation approaches were used to compare model performance across these models and modalities.

4.1 *Model configurations and training setup*

For CLIP-based modeling, we utilized the clip-vit-large- patch14-336 model, which incorporates a vision transformer as the image encoder, producing 1024-dimensional embeddings, and a transformer-based text encoder generating 768-dimensional embeddings. For temporal processing, a sequence of 9 frames was used, with temporal positional embeddings added to maintain sequential order. An 8-headed multi-head self-attention mechanism was applied to aggregate temporal features for classification. CLIP was initially evaluated in zero-shot mode for image-only, text-only, and image–text configurations. Following zero-shot evaluation, CLIP was fine-tuned with a batch size of 5, a learning rate of 5×10^{-5}, and over 15 epochs. The fine-tuning process explored both individual and combined modalities, as well as a temporal setup for sequential frame processing (image–text (temporal)).

For ImageBind, the imagebind_huge model was used to integrate visual, auditory, and textual data. Each modality generated 1024-dimensional embeddings, which were concate-nated to form a combined representation. The model was fine-tuned with a batch size of 5, a learning rate of 1×10^{-5}, and over 20 epochs, focusing on the image–text and image–audio modalities.

GPT-4o was evaluated for zero-shot learning using the gpt- 4o-2024-05-13 model in image-only, text-only, and image–text configurations. The OpenAI Chat Completion API was configured with a temperature of 0.7, a top-*p* of 0.9, a fre- quency penalty of 0.0, and a maximum token limit of 200 to facilitate the generation of reasoning paths for locomotion mode prediction. The temperature of 0.7 and top-*p* of 0.9 were selected to balance reasoning diversity with output precision in a safety-critical classification task. Because the model's output space is constrained to a closed set of 12 locomotion labels (enforced via the system prompt), the risk of hallucination, i.e., generating labels outside the valid set, is substantially mitigated. A reasoning-optimized temperature supports

the chain-of-thought reasoning process by allowing the model to consider multiple plausible interpretations of ambiguous inputs before committing to a prediction, rather than deterministically selecting the highest-probability token at each step. We note that OpenAI's Structured Outputs feature [48] can further enforce schema-level constraints on generated labels, which we recommend for production deployment to eliminate residual hallucination risk. Sensitivity analysis of these hyperparameters is recommended for future work to quantify their impact on classification performance.

4.2 Evaluation metrics

The performance of each model was evaluated using the F1-score, the weighted average F1-score, and the confusion matrix. The F1-score was computed for each locomotion activity to measure the balance between precision and recall. The weighted average F1-score, which considers the class imbalance, was used to evaluate overall model performance. Confusion matrices were employed to visualize and analyze classification errors by comparing predicted labels with true labels.

5 Performance evaluation

5.1 General performance overview

This section provides a high-level view of the overall effectiveness of the models based on the weighted average F1-scores, as illustrated in Fig. 6. The performance across models varies significantly, highlighting the differences between zero-shot and fine-tuned approaches, as well as the impact of multimodal data inputs (image, text, and audio).

5.1.1 CLIP zero-shot models

The performance of the CLIP zero-shot models corresponds with our expectations, particularly for the CLIP zero-shot image-only model, which yielded a lower weighted average F1-score of 0.51%. This can be attributed to the inherent misalignment between the vision data (FOV frames) and the locomotion mode labels. The FOV frames represent what participants see during an activity rather than the activity itself. This discrepancy makes it challenging for the model to map these frames directly to specific locomotion modes, which are inherently action-based. Furthermore, the grid structure of images encapsulates spatiotemporal information, introducing more complexity for the model to process.

On the other hand, the CLIP zero-shot text-only and CLIP zero-shot image–text models demonstrate improved performance, with weighted average F1-scores of 31.58% and 32.28%, respectively. This can be attributed to the fact that user commands provide a more explicit description of the activity being performed. This semantic alignment between the user's commands and the locomotion mode labels leads to higher similarity and better performance.

5.1.2 CLIP fine-tuned models

The fine-tuned CLIP models demonstrated significant

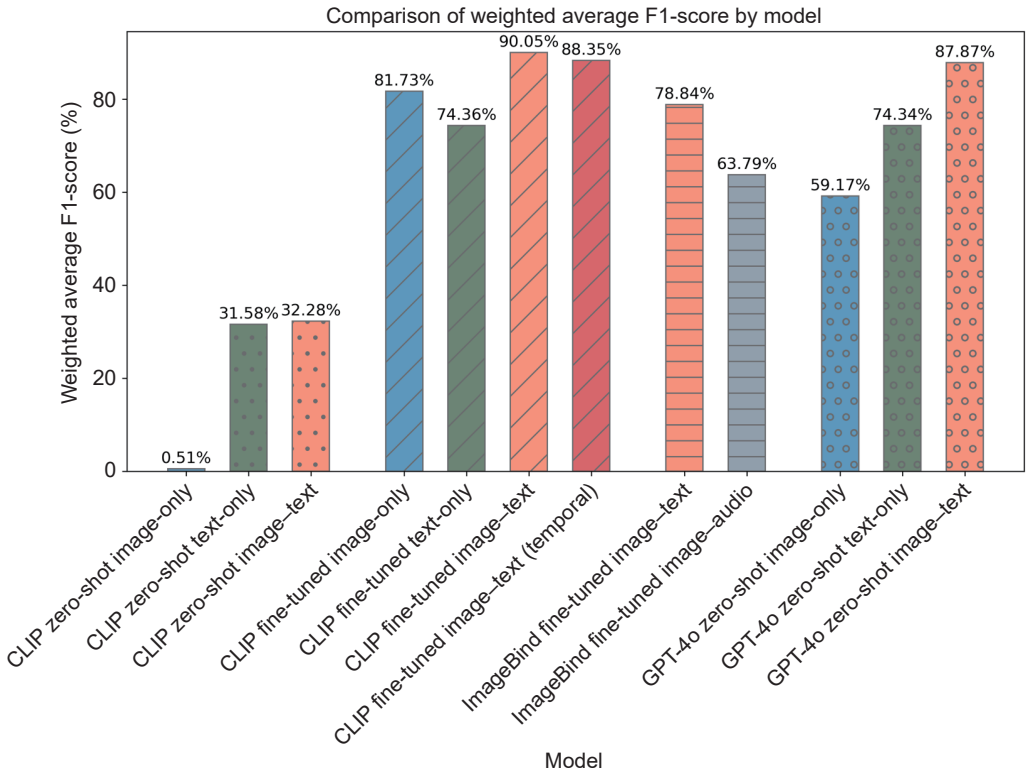


Fig. 6 Comparison of weighted average F1-score.

improvements in performance across all input modalities, indicating the benefits of task-specific adaptation. For the CLIP fine-tuned image-only model, the weighted average F1-score increased substantially to 81.73%, compared to the 0.51% observed in its zero-shot counterpart. This result underscores the model's capacity to learn visual features specific to locomotion activities when fine-tuned on this particular task. Despite the inherent challenges of using FOV frames, where the visual data is indirectly related to the locomotion mode, the model effectively learned to extract more discriminative task-relevant features through fine-tuning.

The CLIP fine-tuned image–text model further improved performance, achieving a weighted average F1-score of 90.05%. The addition of text, where user commands provided contextual descriptions of the activity, greatly enhanced the model's ability to predict locomotion modes. Fine-tuning allowed the model to integrate visual and textual inputs more effectively, leveraging the complementary information from both modalities. However, the CLIP fine-tuned image–text (temporal) model, which utilized sequential frames instead of grid images, showed a marginally lower F1-score of 88.35%. This result suggests that, for this task, the temporal information provided by sequential frames may not sufficiently enhance performance compared to the static grid representation. It also indicates potential room for model improvement in handling sequential frames.

5.1.3 *ImageBind fine-tuned models*

The fine-tuned ImageBind models also demonstrated strong performance, particularly when utilizing both image and text inputs. The ImageBind fine-tuned image–text model achieved a weighted average F1-score of 78.84%, a solid result that, while lower than the CLIP fine-tuned image–text model's performance (90.05%), still underscores the efficacy of combining visual and textual inputs.

Additionally, the ImageBind fine-tuned image–audio model achieved a weighted average F1-score of 63.79%, highlighting the limited distinctiveness of the audio modality for this specific task. Many of the locomotion activities occurred within a similar environment, which reduces the distinct auditory patterns necessary for fine-tuning to significantly enhance performance. Additionally, the audio data comprises both user speech and environmental noise, complicating the model's task for understanding informative signals. As a result, the fine-tuning process yields comparatively lower improvements when the audio modality is utilized, particularly in contrast to the textual data derived from speech transcriptions generated by the Whisper medium model. The lower performance of the image–audio modality (63.79%) can be further attributed to a domain gap between the synthesized and real-world speech conditions. The XTTS v2 text-to-speech model [8] produces relatively "clean" speech signals that lack the prosodic variability, hesitations, and environmental masking characteristic of natural speech on construction sites. Although environmental noise extracted from silent recording segments was added during augmentation, this static noise profile may not fully replicate the dynamic, non-stationary acoustic conditions of an active jobsite, where noise levels commonly reach or exceed 85–90 dBA [41]. Furthermore, certain locomotion tasks such as sitting and standing are inherently non-discriminative in the audio domain: They produce minimal distinctive sound signatures compared to tasks like stair climbing or ladder contact, making audio cues uninformative for these classes. Future work should investigate augmenting synthesized speech with measured noise profiles at varying SNRs and exploring noise-robust audio encoders to improve the audio modality's contribution.

5.1.4 *GPT zero-shot models*

The performance of the GPT-4o zero-shot models is partic-ularly impressive, given the model's lack of task-specific fine-tuning. The GPT-4o zero-shot image–text model achieved a weighted average F1-score of 87.87%, a result remarkably close to the CLIP fine-tuned image–text model (90.05%) and surpassing the performance of the ImageBind fine-tuned image–text model (78.84%).

The strength of the GPT-4o zero-shot image–text model can be attributed to its ability to integrate multiple data modalities effectively, using a combination of FOV frames and user-issued commands to infer locomotion activities. By leveraging CoT prompting, the model follows a step-by-step reasoning process, analyzing both visual and textual cues simultaneously. This guided reasoning enhances the model's capacity to understand user intentions and changes in the user's perspective, making it effective even in a zero-shot context.

In contrast, the GPT-4o zero-shot text-only model achieved a solid F1-score of 74.34%, slightly below the performance of its image–text counterpart, but still competitive. The GPT-4o zero-shot image-only model, however, recorded a lower score of 59.17%, indicating the inherent difficulty of relying solely on FOV data without the command in a zero-shot setting.

5.2 *Class-wise performance analysis*

In this section, we provide a comprehensive class-wise analysis of model performance across various locomotion modes, based on the F1 scores presented in Table 3 and the confusion matrices shown in Fig. 7.

5.2.1 *Ladder up/down climbing (construction and vertical)*

CLIP zero-shot models underperformed in ladder-related tasks. The CLIP zero-shot image–text model

recorded F1-scores of 0.00% for construction ladder down climbing and 16.67% for vertical ladder down climbing, with slight improvements for construction ladder up climbing (18.18%) and vertical ladder up climbing (47.37%). The confusion matrix (Fig. 7a) showed frequent misclassifications, often defaulting to vertical ladder up climbing. The CLIP zero- shot text-only model faced similar challenges, struggling to identify specific ladder types and directions, likely due to the lack of semantically distinct context in the commands. The vision modality did not contribute at all, making language input the primary factor influencing predictions.

Fine-tuning greatly enhanced performance. The CLIP fine-tuned image–text model achieved perfect classification for down climbing (100.00%) and near-perfect scores for up climbing (95.24% for construction and 90.00% for vertical). The confusion matrix also shows minimal misclassifications (Fig. 7b).

The ImageBind fine-tuned image–text model performed adequately, achieving 50.00% for down climbing and 84.21% for up climbing in construction ladders, and 78.26% for down climbing and 80.00% for up climbing in vertical ladders. The confusion matrix (Fig. 7c) shows that many of the misclassifications occurred between the two types of ladders, but the model was consistently able to classify the direction (up or down) correctly.

The GPT-4o zero-shot image–text model achieved strong results for ladder-related activities, though its performance was slightly lower compared to its performance in other classes. It recorded F1-scores of 76.92% for construction ladder down climbing and 83.33% for up climbing, 85.71% for vertical ladder down climbing, and 62.50% for up climbing. Errors mainly pertained to ladder type identification, though the direction was correctly predicted (Fig. 7d). When specific ladder details were unclear, the model defaulted to general reasoning paths, as seen in example #1, Table 4. Reasoning paths were generated most frequently in the image–text setting, underscoring the value of combining visual and command data for improved predictions, as shown in examples #2 and #3 in Table 4.

Table 3 Class-wise F1-scores across models (in %)

Class	CLIP ZS image-only	CLIP ZS text-only	CLIP ZS image–text	CLIP FT image-only	CLIP FT text-only	CLIP FT image–text	CLIP FT image–text (temporal)	ImageBind FT image–text	ImageBind FT image-audio	GPT-4o ZS image-only	GPT-4o ZS text-only	GPT-4o ZS image–text
Construction ladder down climbing	0.00	0.00	0.00	46.15	54.55	100.00	100.00	50.00	40.00	85.71	55.56	76.92
Construction ladder up climbing	0.00	50.00	18.18	90.91	90.91	95.24	100.00	84.21	47.06	76.92	80.00	83.33
Vertical ladder down climbing	0.00	0.00	16.67	81.82	75.00	100.00	95.24	78.26	57.14	66.67	58.82	85.71
Vertical ladder up climbing	0.00	50.00	47.37	100.00	80.00	90.00	94.74	80.00	50.00	52.63	66.67	62.50
Level-ground navigation	0.00	0.00	18.18	84.21	70.59	66.67	75.00	88.89	64.00	47.06	100.00	100.00
Low space navigation	0.00	0.00	11.76	94.74	84.21	88.89	84.21	88.89	81.82	66.67	75.00	85.71
Sitting	0.00	75.00	90.00	78.26	90.00	94.74	88.89	90.00	70.59	0.00	57.14	88.89
Standing	8.08	53.85	60.00	44.44	92.31	85.71	93.33	92.31	76.92	0.00	85.71	92.31
Stair ascension	0.00	0.00	0.00	95.24	86.96	86.96	86.96	95.24	62.50	90.91	95.24	95.24
Stair descension	0.00	54.05	42.55	86.96	94.74	94.74	94.74	100.00	94.74	62.07	95.24	90.91
Stepping over box	0.00	53.85	30.77	76.92	26.67	82.35	70.59	28.57	33.33	92.31	36.36	92.31
Stepping over pipe	0.00	46.15	50.00	77.78	31.58	94.74	76.19	50.00	75.00	69.23	72.73	100.00

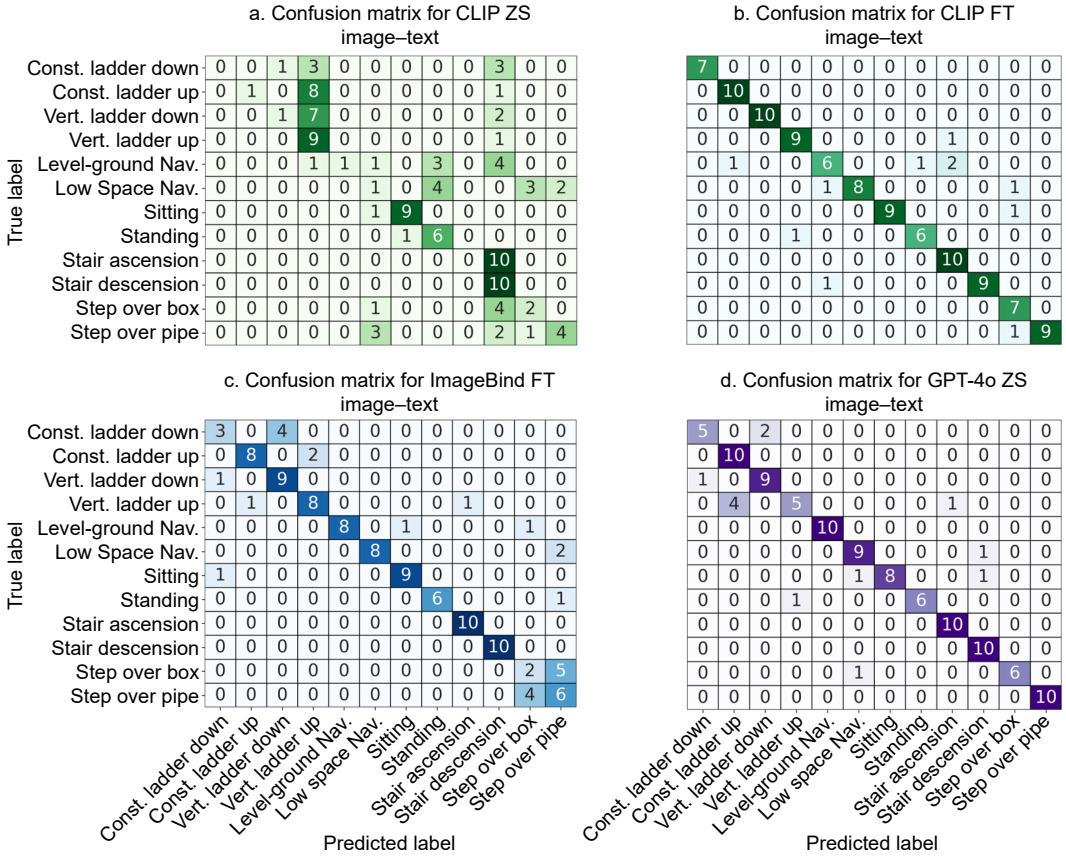


Fig. 7 Comparison of confusion matrices based on different models: (a) CLIP zero-shot image–text, (b) CLIP fine-tuned image–text, (c) ImageBind fine-tuned image–text, and (d) GPT-4o zero-shot image–text.

5.2.2 Navigation and low space tasks

The CLIP zero-shot image–text model performed poorly in level-ground navigation (F1-score of 18.18%), with frequent misclassifications involving activities like standing and stair descension (Fig. 7a). For low space navigation, the model also struggled, achieving an F1-score of 11.76%, likely due to weak contextual alignment between user commands and locomotion labels. Commands like "I'm bending" lacked the clarity needed to map accurately to low space navigation. Both image-only and text-only models recorded F1-scores of 0.00%, indicating insufficient discriminative power in these modalities individually. A marginal improvement is observed in the image–text model.

The CLIP Fine-tuned image–text model achieved 66.67% for level-ground navigation and 88.89% for low space navigation. For level-ground navigation, which had the lowest F1-score among the fine-tuned classes, misclassifications were frequent with activities like standing, stair ascension, and construction ladder up climbing (Fig. 7b).

The ImageBind fine-tuned image–text model showed notable improvements after fine-tuning, achieving 88.89% for both level-ground navigation and low space navigation. The model outperformed the CLIP fine-tuned image–text model in both tasks.

GPT-4o models showed strong results, with 100.00% for level-ground navigation in both image–text and text-only settings, and 85.71% for low space navigation in the image–text setting. However, one significant misclassification arose in low space navigation, where the model incorrectly predicted stair descension due to ambiguous command input and subtle visual cues, as illustrated in example #4 in Table 4. Although the model accurately detected the downward movement in the frames, it failed to recognize the context of low-space navigation. This highlights the difficulty in interpreting vague commands like "I should be careful here again", which can mislead the model's reasoning.

5.2.3 Sitting and standing

The CLIP zero-shot text-only model performed reasonably well in sitting and standing tasks, with F1-scores of 75.00% and 53.85%, respectively. The CLIP zero-shot image–text model achieved higher scores of 90.00% for sitting and 60.00% for standing, due to a higher degree of semantic alignment between commands and labels. The vision modality did not contribute significantly, as evidenced by the low F1-scores in the image-only setting (0.00% for sitting and 8.08% for standing).

Fine-tuning improved performance across modalities. The CLIP fine-tuned text-only model reached 90.00% for sitting and 92.31% for standing, while the CLIP fine-

Table 4 Examples of GPT-4o zero-shot reasoning paths and predictions

#	True label	Reasoning path and prediction		
		GPT-4o ZS image-only	GPT-4o ZS text-only	GPT-4o ZS image–text
1	Construction ladder down climbing	Reasoning: Based on the sequence of frames showing the user descending a ladder. Prediction: Construction ladder down climbing ✓	Reasoning: — Prediction: Vertical ladder down climbing ✗	Reasoning: Based on the frames and the spoken command "I'm at the top of the other ladder. I am going back down", it is clear that the user is descending a ladder. Prediction: Vertical ladder down climbing ✗
2	Vertical ladder up climbing	Reasoning: — Prediction: Construction ladder up climbing ✗	Reasoning: — Prediction: Vertical ladder up climbing ✓	Reasoning: Based on the provided frames and the spoken command "I am climbing the straight ladder", the user is performing the activity of climbing a vertical ladder. The sequence of images shows the user's perspective moving up the rungs of a ladder, consistent with vertical ladder climbing. Prediction: Vertical ladder up climbing ✓
3	Vertical ladder down climbing	Reasoning: Based on the frames provided, the user appears to be climbing down a vertical ladder in an industrial environment. The sequence of images shows a downward movement with hands grasping the rungs of a ladder, indicating a descent. Prediction: Vertical ladder down climbing ✓	Reasoning: — Prediction: Stair descension ✗	Reasoning: Based on the provided frames and the spoken command "Now I'm coming down", it is clear that the user is descending. The frames show a progression of descending movements on a ladder, as indicated by the user's hands and feet positions. Prediction: Vertical ladder down climbing ✓
4	Low space navigation	Reasoning: Based on the provided frames, it appears that the user is moving along a path with a visible change in elevation. The downward angle and the presence of railings suggest that the user is likely descending the stairs. Prediction: Stair descension ✗	Reasoning: — Prediction: Construction ladder down climbing ✗	Reasoning: Based on the frames and the command, "I should be careful here again", it appears the user is navigating an environment where caution is necessary. The frames show a consistent descent down a grated metallic structure, indicative of staircases often found in industrial settings. The perspective and angle change in the frames suggest movement downwards, and the presence of railings and the structure of the stairs are typical of industrial staircases. Prediction: Stair descension ✗
5	Stepping over box	Reasoning: Based on the provided frames, it appears that the user is approaching and then stepping over an obstacle (the box) on the ground. The sequence of frames shows the user's movement towards the box and then the box becoming less visible, suggesting the user is stepping over it. Prediction: Stepping over box ✓	Reasoning: — Prediction: Stepping over pipe ✗	Reasoning: Based on the command "I'm about to go over the obstacle" and the provided frames showing the user's field of view over time, it is evident that the user is preparing to step over an obstacle (the box in this case). Prediction: Stepping over box ✓

tuned image–text model achieved 94.74% and 85.71%, respectively. The image-only model's lower F1-score of 44.44% for standing likely affected the combined image–text model's overall score. Using temporal FOV frames, the image–text (temporal) model improved to 93.33% for standing.

The ImageBind fine-tuned image–text model performed well, with F1-scores of 90.00% for sitting and 92.31% for standing. The ImageBind fine-tuned image–audio model had lower performance (70.59% for sitting and 76.92% for standing). These results, while lower than the image–text model, are still relatively strong for the image–audio model and rank just behind tasks such as low space navigation (81.82%) and stair descension

(94.74%). This pattern aligns with the strong performance seen in the image–text counterparts, suggesting that the improved accuracy in these tasks is primarily driven by visual information, with the audio modality contributing minimally.

GPT-4o's zero-shot image–text model achieved strong results, with F1-scores of 88.89% for sitting and 92.31% for standing. The model's reasoning paths demonstrate its ability to infer user intent by integrating minimal visual cues with language inputs. The GPT-4o zero-shot text-only model performed moderately, achieving 57.14% for sitting and 85.71% for standing. The image-only model, however, struggled significantly with 0.00% F1-scores for both tasks. A review of reasoning paths indicated that recurring metallic structures often led to confusion between sitting and standing, particularly with other vertical tasks such as ladder or stair climbing.

This highlighted the importance of language input for less dynamic tasks where visual cues alone can be ambiguous.

5.2.4 *Stair ascension/descension*

CLIP zero-shot models showed poor performance for stair ascension, with image–text, image-only, and text-only settings all yielding an F1-score of 0.00%. The confusion matrix (Fig. 7a) indicated consistent misclassification of stair ascension as stair descension, suggesting difficulty in distinguishing between upward and downward movement, defaulting to the latter. For stair descension, the text-only model performed moderately better with an F1-score of 54.05%, while the image–text model achieved 42.55%. Overprediction of stair descension was common across tasks, implying that the semantic context in zero-shot commands was insufficient for precise activity differentiation, leading to higher similarity scores favoring Stair Descension.

Fine-tuning improved CLIP model performance significantly. The CLIP fine-tuned image–text model achieved 86.96% for stair ascension and 94.74% for stair descension, with minimal misclassifications seen in the confusion matrix (Fig. 7b). The image-only model also performed well, recording 95.24% for stair ascension and 86.96% for stair descension. However, the image–text and image–text temporal models did not surpass the text-only model, maintaining identical F1 scores. This suggests that concatenation-based fusion might not have fully utilized visual information for these tasks.

The ImageBind fine-tuned image–text model performed exceptionally, reaching an F1-score of 95.24% for stair as- cension and 100.00% for stair descension with accurate classifications shown in the confusion matrix (Fig. 7c).

GPT-4o's results in stair-related tasks were strong. Both zero-shot image–text and text-only models achieved 95.24% for stair ascension, while the image-only model scored 90.91%. For stair descension, the image–text model reached 90.91%, with the text-only setting at 95.24%. The image-only model, however, struggled with stair descension (F1-score of 62.07%), emphasizing the importance of command input for this task.

The confusion matrix (Fig. 7d) highlighted some misclassifications: Sitting and low space navigation were sometimes predicted as stair descension, and vertical ladder up climbing was occasionally misclassified as stair ascension. This underscores that, while the model correctly identified the general movement direction, specific task recognition remained challenging. The confusion between vertical ladder up climbing and stair ascension can be attributed to strong visual similarities in the egocentric FOV frames: Both tasks produce repetitive horizontal edge patterns (rungs vs. stair treads) with similar upward egocentric motion cues and metallic/industrial backgrounds. When the FOV captures only a partial view of the climbing structure, particularly under occlusion from the user's hands or structural elements, the discriminative visual features (e.g., rung spacing, handrail geometry, and the presence of a stair landing) may be insufficient for reliable differentiation. This failure mode is consistent with findings in construction action recognition literature, where occlusion is identified as a primary source of classification error [49]. Potential mitigations include incorporating depth cues (e.g., from stereo or structured-light cameras), fusing with IMU data to capture distinct kinematic signatures of ladder vs. stair climbing, and extending temporal context through memory-augmented architectures to leverage longer-horizon sequential patterns.

5.2.5 *Obstacle navigation: Stepping over box and pipe*

CLIP zero-shot models displayed moderate performance in obstacle navigation tasks. The CLIP zero-shot image–text model achieved F1-scores of 30.77% for stepping over box and 50.00% for stepping over pipe, while the text-only model performed slightly better, with scores of 53.85% and 46.15%, respectively.

Fine-tuning improved CLIP model outcomes, with the image-only model achieving 76.92% for stepping over box and 77.78% for stepping over pipe. The CLIP fine-tuned image–text model integrated visual and textual features effectively, scoring 82.35% for stepping over box and 94.74% for stepping over pipe. Minimal misclassifications were observed in the confusion matrix (Fig. 7b).

The ImageBind fine-tuned image–text model performed less effectively on obstacle tasks compared to other classes, with F1-scores of 28.57% for stepping over box and 50.00% for stepping over pipe. The confusion matrix (Fig. 7c) revealed challenges in differentiating between these obstacle classes. The ImageBind fine-tuned image–audio model showed bet ter results for stepping

over pipe (75.00%) but continued to struggle with stepping over box (33.33%).

GPT-4o performed exceptionally well in these tasks. The zero-shot image–text model achieved an F1-score of 92.31% for stepping over box and a perfect 100.00% for stepping over pipe. The text-only model performed moderately, with 36.36% for stepping over box and 72.73% for stepping over pipe. The image-only model scored 69.23% for stepping over pipe but struggled with stepping over box (36.36%). Example #5 in Table 4 underscores the importance of both modalities. In cases where the command was generic (e.g., "I'm about to go over the obstacle"), visual frames provided essential context for identifying the specific obstacle, such as the box.

6 Discussion of key findings

6.1 *Multimodality and model performance*

The importance of incorporating multiple modalities, visual and linguistic data, into the prediction of locomotion activities was evident. Although there were cases where individual modalities (e.g., image-only or text-only models) performed better for certain classes, the multimodal models generally achieved higher overall performance across most locomotion activities. For instance, as seen in Fig. 6, the fine-tuned CLIP image–text model achieved a weighted average F1-score of 90.05%, surpassing both the CLIP image-only (81.73%) and CLIP text-only (74.36%) models, followed by the GPT image-only (59.17%), text-only (74.34%), and image–text (87.87%) models. This confirms that the integration of visual and linguistic information enables the model to capture complementary cues, leading to more accurate predictions across a variety of tasks.

However, it is worth noting that the multimodal models did not always outperform individual modalities for all classes. For example, lower accuracies were observed for specific tasks such as level-ground navigation, low space navigation, stair ascension, and standing in the image–text setting compared to the top-performing individual modality in fine-tuned CLIP models, as seen in Table 3. This suggests that for some locomotion activities, the concatenation fusion technique did not fully capitalize on the strengths of each modality. To address this, exploring advanced fusion strategies, such as cross-attention or hierarchical fusion, might better balance the contributions of each modality and enhance overall performance. Moreover, GPT-4o zero-shot models also showed lower performance in a few cases when combining both modalities, particularly in tasks such as construction ladder down climbing, vertical ladder up climbing, and stair descension, as shown in Table 3. These inconsistencies further emphasize the need for refined prompting strategies to handle real-world imperfections in input data, such as ambiguities or delays. By incorporating prompt techniques that instruct the model to recognize ambiguities and focus on clearer, more important information, it could potentially improve the model's accurate predictions when combining both modalities.

The temporal representation of input data was pivotal for model performance. The grid format in fine-tuned models such as CLIP effectively captured temporal context without requiring sequential frame analysis. The GPT-4o zero-shot model also demonstrated strong reasoning with this static grid, utilizing condensed temporal information to perform well across tasks. However, the CLIP fine-tuned image–text (temporal) model, which used a temporal aggregation mechanism with positional embeddings and multi-head self-attention, did not significantly outperform the grid-based approach. This highlights that while the grid representation was sufficient for this task, improving the handling of sequential frames could further enhance performance.

6.2 *Occlusion handling and hallucination avoidance*

A critical concern for deploying vision–language models in construction environments is the risk of hallucination, generating plausible but incorrect predictions, particularly under occlusion or ambiguous visual conditions. Our framework addresses this risk through several complementary mechanisms. First, the GPT-4o system prompt constrains the model's output to a closed set of 12 valid locomotion labels, preventing the generation of arbitrary or out-of-vocabulary predictions. This closed-set constraint significantly reduces the output space and eliminates free-form generation that could lead to hallucinated labels. Second, the CoT prompting strategy requires the model to articulate its reasoning before committing to a prediction, making the decision process transparent and auditable. When visual cues are occluded or ambiguous (as seen in example #4 in Table 4), the reasoning path reveals the model's uncertainty, enabling downstream safety systems to flag low-confidence predictions. For production deployment, we recommend incorporating uncertainty-aware fields (e.g., vagueness scores and discrepancy scores) into the prediction schema, along with structured output enforcement via JSON schema constraints [48], to provide explicit confidence signals that can trigger fallback behaviors, such as maintaining the current locomotion mode, when prediction reliability is insufficient for safe mode switching.

6.3 *Toward quantitative human–robot interaction metrics*

The predictions presented in Tables 3 and 4 are qualitative locomotion labels. However, for practical HRI in exoskeleton control, quantitative metrics are essential to assess system performance beyond classification accuracy. Future extensions of this framework should report:

(i) transition timing error: the temporal offset between the predicted and actual onset of a locomotion mode transition; (ii) mode-switch delay: the end-to-end latency from user intent to exoskeleton response; (iii) stability proxies: kinematic or force-based measures (e.g., center-of-pressure variability and joint torque deviations) that quantify the physical impact of correct vs. incorrect mode predictions on the user; and (iv) calibration metrics: expected calibration error (ECE) and Brier scores to assess whether the model's confidence estimates are reliable for safety-critical decision-making. Prior exoskeleton literature reports transition delays on the order of seconds and emphasizes that reducing delay and wrong switching is directly safety-relevant [8, 50], motivating these quantitative metrics as essential complements to the label-level accuracy reported in this study.

6.4 Generalization vs. task-specific performance

Fine-tuning the models resulted in significant performance improvements. For instance, the weighted average F1-score for the CLIP image–text model increased from 32.28% (zero-shot) to 90.05% (fine-tuned). However, considering the dataset's constrained diversity, the task-specific accuracy gains observed in these models, such as CLIP, may not ensure strong generalization. This limitation is evident in their lower performance in zero-shot settings, which suggests challenges in adapting to new or rapidly evolving environments, such as those commonly encountered in construction.

In contrast, the large model GPT-4o demonstrated promising generalization across various construction-related tasks in our study by effectively interpreting prompts and integrating visual cues with speech commands to discern user intent. Leveraging new paradigms like zero-shot learning, this model and similar large language models can adapt quickly to unfamiliar tasks with minimal input, as evidenced by its 87.87% F1-score closely rivaling fine-tuned CLIP's 90.05%. This adaptability is particularly beneficial in dynamic environments such as industrial and construction sites, where evolving tasks and conditions demand models capable of generalizing effectively across diverse scenarios.

Despite this, as highlighted earlier, most models, with the exception of the CLIP zero-shot models, generally predicted upward and downward directions accurately. However, GPT-4o zero-shot faced challenges in distinguishing specific locomotion modes, such as ladder types (e.g., vertical vs. construction ladders), where it scored 62.50% for vertical ladder up climbing and 76.92% for construction ladder down climbing compared to fine-tuned CLIP's 90.00% and 100%, respectively. This limitation, evident in misclassifications like predicting stair descension for low space navigation due to ambiguous commands, suggests that while GPT-4o excels in general directionality, its zero-shot performance can falter with fine-grained task differentiation and unclear inputs. Enhancing it with few-shot learning or retrieval-augmented systems could bolster its ability to leverage additional context, improving adaptability across diverse construction scenarios.

While fine-tuning enables models to excel in specific tasks, even using smaller models like CLIP, the trade-off between generalization and task-specific performance is a critical consideration for real-world applications, especially considering the wide variety of activities within changing environments. The strong zero-shot performance of large language models like GPT-4o suggests a scalable path forward, offering insights into their ability to generalize across complex construction tasks, as explored in this study.

7 Conclusions

This study developed a multimodal framework to integrate field-of-view frames and speech commands for locomotion prediction, targeting complex tasks in dynamic construction settings, such as ladder climbing, stair navigation, low-space movement, and obstacle navigation, using smart glasses. We investigated two key aspects: the effectiveness of combining vision and speech data to enhance locomotion prediction accuracy and the generalization capability of large language models, particularly GPT-4o, through zero-shot learning compared to supervised fine-tuning.

In the image–text setting, the fine-tuned CLIP model demonstrated the best overall performance, achieving a weighted average F1-score of 90.05%. The GPT-4o model, prompted using a zero-shot CoT approach, followed closely with an F1-score of 87.87%. The fine-tuned ImageBind model performed well with an F1-score of 78.84%, though it was outperformed by fine-tuned CLIP and GPT-4o. This suggests that ImageBind may be less suited for this task compared to the other models.

Models that integrated image and text modalities consistently outperformed single-modality approaches. The combination of visual and linguistic cues improved the models' ability to capture user intent during complex locomotion tasks. The CoT prompting with GPT-4o was particularly effective, allowing the model to infer user intent from commands and understand temporal sequences in FOV frames, producing rea- soning paths that led to accurate predictions.

The inclusion of the audio modality did not enhance performance notably when evaluated with ImageBind, indicating that the visual and linguistic cues were more defining for the studied locomotion tasks.

Class-wise analysis showed that most models, except for zero-shot CLIP, generally predicted upward and

downward directions in vertical movements, such as ladder climbing, and stair ascension and descension, accurately, as well as other tasks involving similar vertical movements like low-space navigation, sitting, and standing.

Overall, this research demonstrates the effectiveness of integrating field-of-view frames and speech commands to enhance locomotion prediction while affirming GPT-4o's substantial potential as a scalable solution for exoskeleton control in dynamic construction settings. By evaluating models like CLIP, ImageBind, and GPT-4o, these findings advance adaptive automation in real-world construction scenarios, though certain challenges remain to fully realize their applicability.

Several limitations emerged in this study, some of which have been highlighted in Section 5. The dataset used in this study did not encompass all possible construction activities and environmental conditions, potentially limiting the models' generalization capabilities. The integration of GPT-4o provided an effective means of leveraging the zero-shot learning to address some of these challenges. However, the model sometimes misinterpreted tasks when commands were ambiguous or lacked sufficient environmental context, leading to prediction errors.

Although our data collection was designed to mimic realistic construction locomotion tasks, it remains closer to a controlled replication than a fully unconstrained jobsite deployment. Known field factors, extreme lighting variability (including nighttime and low-light conditions), denser occlusions from equipment and co-workers, adverse weather, and a broader diversity of obstacles, can reduce recognition performance for vision-based perception unless explicitly accounted for. Construction vision studies specifically report degradation under occlusion and low-light conditions [10, 51], supporting the need for a staged field-validation roadmap. Accordingly, we position this work as a foundational benchmark and propose a follow-on field study with broader environmental variation, incorporating robustness measures such as uncertainty gating and temporal memory to handle distribution shifts encountered on active construction sites.

Another key limitation is the static nature of the models used in this study. The reliance on reasoning derived from a limited temporal context, specifically a 5-s window of FOV frames, restricted the models' ability to infer sequential patterns or leverage broader historical context. This static approach made the models less effective in understanding evolving user intentions over time.

Additionally, construction environments are inherently noisy, and this study leveraged a dataset partially collected in a noisy mechanical mezzanine setting, where Whisper effectively transcribed speech inputs with consistent performance. To further explore real-world applicability, future studies could evaluate model performance across an even broader range of noisy speech conditions and consider denoising techniques to enhance robustness in varied settings.

Future work should address these limitations by exploring memory- and retrieval-augmented LLM agents that incorporate historical and environmental context into the reasoning process and enable context-aware actions. By introducing a retrieval system, such approaches could dynamically align current tasks with prior activities, improving robustness against command ambiguity and enhancing sequential decision-making. Advancements in LLMs like GPT-4o have significantly reduced latency and API costs, making real-time applications increasingly viable in construction environments. However, further optimization is still required to meet the stringent demands of highly dynamic and safety-critical settings. Future iterations of this research should focus on leveraging these advancements while addressing remaining challenges to enhance scalability, performance, and safety.

A critical practical consideration for deploying GPT-4o in construction exoskeleton control is its reliance on cloud-based API infrastructure. Construction sites frequently experience poor or intermittent network connectivity, which introduces unpredictable latency and reliability concerns incompatible with safety-critical real-time control. In contrast, fine-tuned models such as CLIP are lightweight enough for edge de- ployment on embedded hardware, enabling deterministic inference latency without network dependency. We therefore propose a hybrid deployment architecture: An edge-deployed fine-tuned model (e.g., CLIP) serves as the primary real-time controller with guaranteed low-latency predictions, while the cloud-based GPT-4o model operates as a secondary reasoning layer that periodically refines predictions, handles ambiguous cases flagged by uncertainty scores, and provides contextual reasoning when network connectivity permits. This architecture preserves the safety-critical responsiveness of edge inference while leveraging GPT-4o's superior zero-shot reasoning capabilities as an advisory system. Future work should benchmark inference latency and power consumption across representative edge hardware platforms to establish deployment feasibility for construction environments.

Additionally, the reliance on a 5-s temporal window limits the model's ability to handle long-duration tasks or leverage broader historical context for sequential decision-making. Future iterations should explore memory-augmented architectures, such as short-term memory buffers that retain recent predictions and observations, long-term memory stores (e.g., vector databases) that encode site-specific patterns, and retrieval-augmented generation (RAG) approaches that dynamically retrieve relevant prior experiences during inference. These

mechanisms would enable the model to reason over longer temporal horizons, recognize recurring task sequences (e.g., repeated ladder-to-ground transitions during roofing work), and adapt to site-specific locomotion patterns without retraining. Such memory-augmented systems represent a natural extension of the zero-shot reasoning paradigm demonstrated in this work toward fully adaptive exoskeleton control agents.

Acknowledgements

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

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

Author contribution statement

All authors have given approval to the final version of the manuscript.

Use of AI statement

None.

References

[1] S. Kim, A. Moore, D. Srinivasan, et al. Potential of exoskeleton technologies to enhance safety, health, and performance in construction: Industry perspectives and future research directions. IISE Trans Occup Ergon Hum Fact, 2019, 7: 185–191.

[2] O. Flor-Unda, B. Casa, M. Fuentes, et al. Exoskeletons: Contribution to occupational health and safety. Bioengineering, 2023, 10: 1039.

[3] D. Pinto-Fernandez, D. Torricelli, M. del Carmen Sanchez-Villamanan, et al. Performance evaluation of lower limb exoskeletons: A systematic review. IEEE Trans Neural Syst Rehabil Eng, 2020, 28: 1573–1583.

[4] A. M. Dollar, H. Herr. Lower extremity exoskeletons and active orthoses: Challenges and state-of-the-art. IEEE Trans Robot, 2008, 24: 144–158.

[5] D. Novak, R. Riener. A survey of sensor fusion methods in wearable robotics. Robot Auton Syst, 2015, 73: 155–170.

[6] Y. P. Zhang, G. Z. Cao, L. L. Li, et al. Interactive control of lower limb exoskeleton robots: A review. IEEE Sens J, 2024, 24: 5759–5784.

[7] B. S. Rupal, S. Rafique, A. Singla, et al. Lower-limb exoskeletons: Research trends and regulatory guidelines in medical and non-medical applications. Int J Adv Robot Syst, 2017, 14(6): 1–27

[8] X. Ma, Y. Liu, X. Zhang, et al. Real-time continuous locomotion mode recognition and transition prediction for human with lower limb exoskeleton. IEEE J Biomed Health Inform, 2025, 29: 1074–1086.

[9] Occupational Safety and Health Administration. OSHA's fall prevention campaign [Online]. https://www.osha.gov/stop-falls (accessed 2026-01-30).

[10] B. H. Zhang, B. Yang, C. J. Wang, et al. Computer vision-based construction process sensing for cyber-physical systems: A review. Sensors, 2021, 21: 5468.

[11] S. Yun, S. Hong, S. Hwang, et al. Analysis of masonry work activity recognition accuracy using a spatiotemporal graph convolutional network across different camera angles. Autom Constr, 2025, 175: 106178.

[12] Y. P. Qian, Y. N. Wang, C. H. Chen, et al. Predictive locomotion mode recognition and accurate gait phase estimation for hip exoskeleton on various terrains. IEEE Robot Autom Lett, 2022, 7: 6439–6446.

[13] S. Y. Yu, J. F. Yang, T. H. Huang, et al. Artificial neural network-based activities classification, gait phase estimation, and prediction. Ann Biomed Eng, 2023, 51: 1471–1484.

[14] Y. Long, Z. J. Du, W. D. Wang, et al. PSO-SVM-based online locomotion mode identification for rehabilitation robotic exoskeletons. Sensors, 2016, 16: 1408.

[15] A. Parri, K. B. Yuan, D. Marconi, et al. Real-time hybrid locomotion mode recognition for lower limb wearable robots. IEEE/ASME Trans Mech, 2017, 22: 2480–2491.

[16] H. Kim, Y. J. Shin, J. Kim. Kinematic-based locomotion mode recognition for power augmentation exoskeleton. Int J Adv Robot Syst, 2017, 14: 1–14.

[17] X. H. Liu, Q. N. Wang. Real-time locomotion mode recognition and assistive torque control for unilateral knee exoskeleton on different terrains. IEEE/ASME Trans Mech, 2020, 25: 2722–2732.

[18] Y. F. Zhao, J. X. Wang, Y. F. Zhang, et al. Flexible and wearable EMG and PSD sensors enabled locomotion mode recognition for IoHT-based in-home rehabilitation. IEEE Sens J, 2021, 21: 26311–26319.

[19] A. H. A. Al-Dabbagh, R. Ronsse. Depth vision-based terrain detection algorithm during human locomotion. IEEE Trans Med Robot Bionics, 2022, 4: 1010–1021.

[20] L. Zhu, Z. Wang, Z. G. Ning, et al. A novel motion intention recognition approach for soft exoskeleton via IMU. Electronics, 2020, 9: 2176.

[21] B. Laschowski, W. McNally, A. Wong, et al. Environment classification for robotic leg prostheses and exoskeletons using deep convolutional neural networks. Front Neurorobot, 2022, 15: 730965.

[22] E. Guo, C. Perlette, M. Sharifi, et al. Speech-based human-exoskeleton interaction for lower limb motion planning. In: Proceedings of the IEEE 4th International Conference on Human–Machine Systems, Toronto, Canada, 2024: pp 1–6.

[23] D. Martínez-Pascual, J. M. Catalán, A. Blanco-Ivorra, et al.

Gait activity classification with convolutional neural network using lower limb angle measurement from inertial sensors. IEEE Sens J, 2024, 24: 21479–21489.

[24] M. H. Li, B. X. Zhong, E. Lobaton, et al. Fusion of human gaze and machine vision for predicting intended locomotion mode. IEEE Trans Neural Syst Rehabil Eng, 2022, 30: 1103–1112.

[25] A. Sharma, E. Rombokas. Improving IMU-based prediction of lower limb kinematics in natural environments using egocentric optical flow. IEEE Trans Neural Syst Rehabil Eng, 2022, 30: 699–708.

[26] O. Tsepa, R. Burakov, B. Laschowski, et al. Continuous prediction of leg kinematics during walking using inertial sensors, smart glasses, and embedded computing. In: Proceedings of 2023 IEEE International Conference on Robotics and Automation, London, United Kingdom, 2023: pp 10478–10482.

[27] J. Q. Wang, D. M. Wu, Y. Z. Gao, et al. Integral real-time locomotion mode recognition based on GA-CNN for lower limb exoskeleton. J Bionic Eng, 2022, 19: 1359–1373.

[28] J. Y. Li, J. H. Zhang, K X. Li, et al. A multimodal framework based on deep belief network for human locomotion intent prediction. Biomed Eng Lett, 2024, 14: 559–569.

[29] A. Radford, J. W. Kim, C. Hallacy, et al. Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning, 2021: pp 8748–8763.

[30] OpenAI. GPT-4o system card [Online]. https://openai.com/index/gpt-4o-system-card/ (accessed 2026-01-30). (in Chinese)

[31] H. T. Liu, C. Y. Li, Q. Y. Wu, et al. Visual instruction tuning. In: Proceedings of the 37th International Conference on Neural Information Processing Systems, New Orleans, USA, 2023: p 1516.

[32] J. B. Alayrac, J. Donahue, P. Luc, et al. Flamingo: A visual language model for few-shot learning. In: Proceedings of the 36th International Conference on Neural Information Processing Systems, New Orleans, USA, 2022: p 1723.

[33] Occupational Safety and Health Administration. 1926.1053–ladders [Online]. https://www.osha.gov/laws-regs/regulations/standardnumber/1926/1926.1053 (accessed 2026-01-30).

[34] Tobii Pro. Tobii pro glasses 2–product description [Online]. https://www.tobii.com/products/discontinued/tobii-pro-glasses-2 (accessed 2026-01-30).

[35] H. Sloetjes, P. Wittenburg. Annotation by category: ELAN and ISO DCR. In: Proceedings of the 6th International Conference on Language Resources and Evaluation, Marrakech, Morocco, 2008.

[36] OpenAI, J. Achiam, S. Adler, et al. GPT-4 technical report. 2023, arXiv: 2303.08774. arXiv. org e-Print archive. https://arxiv.org/abs/2303.08774 (accessed 2025-11-20).

[37] T. B. Brown, B. Mann, N. Ryder, et al. Language models are few-shot learners. In: Proceedings of the 34th International Conference on Neural Information Processing Systems, Vancouver, Canada, 2020: p 159.

[38] Coqui. AI. TTS: A deep learning toolkit for text-to-speech, battle-tested in research and production [Online]. https://github.com/coqui-ai/TTS (accessed 2025-11-20).

[39] G. Lemaître, F. Nogueira, C. K. Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. J Mach Learn Res, 2017, 18: 1–5.

[40] A. Radford, J. W. Kim, T. Xu, et al. Robust speech recognition via large-scale weak supervision. In Proceedings of the 40th International Conference on Machine Learning, Honolulu, USA, 2023: pp 1182.

[41] Occupational Safety and Health Administration. Occupational noise exposure–hearing loss in construction [Online]. https://www.osha.gov/noise/construction (accessed 2026-01-30).

[42] Centers for Disease Control and Prevention. Understand noise exposure [Online]. https://www.cdc.gov/niosh/noise/prevent/understand.html (accessed 2026-01-30).

[43] OpenAI. Whisper: Robust speech recognition via large-scale weak supervision [Online]. https://github.com/openai/whisper (accessed 2026-01-30).

[44] R. Girdhar, A. El-Nouby, Z. Liu, et al. ImageBind one embedding space to bind them all. In: Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, 2023: pp 15180–15190.

[45] H. Zhang, X. Li, L. D. Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Singapore, 2023: 543–553.

[46] Z. Y. Li, H. Liu, D. Zhou, et al. Chain of thought empowers transformers to solve inherently serial problems. In: Proceedings of the 12th International Conference on Learning Representations, Vienna, Austria, 2024: 1–38.

[47] J. Wei, X. Z. Wang, D. Schuurmans, et al. Chain-of-thought prompting elicits reasoning in large language models. In: Proceedings of the 36th International Conference on Neural Information Processing Systems, New Orleans, USA, 2022: pp 1800.

[48] OpenAI. Structured outputs–OpenAI API [Online]. https://platform.openai.com/docs/guides/structured-outputs (accessed 2026-01-30).

[49] Z. Q. Li, D. S. Li. Action recognition of construction workers under occlusion. J Build Eng, 2022, 45: 103352.

[50] G. Du, J. C. Zeng, C. Gong, et al. Locomotion mode recognition with inertial signals for hip joint exoskeleton. Appl Bionics Biomech, 2021, 2021: 6673018.

[51] B. Xiao, Q. Lin, Y. Chen. A vision-based method for automatic tracking of construction machines at nighttime based on deep learning illumination enhancement. Autom Constr, 2021, 127: 103721.