IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

Multi-Modal Locomotion Mode Recognition in the Real World for Robotic Hip Complex Exoskeletons

Hyesoo Shin[1,2], Sangdo Kim[1,3], Sunwoo Kim[1,3], Jongwon Lee[1*], Jinkyu Kim[2], and KangGeon Kim[1*]

***Abstract*—Lower limb exoskeletons assist users by supporting joint movements. Since joint motion patterns vary depending on how the user moves, accurately recognizing the type of movement (locomotion mode) is crucial for controlling the exoskeleton and ensuring user safety. Inspired by how humans use multiple types of sensory information to control movement, we developed a multi-modal locomotion mode recognition (LMR) system that uses both mechanical and visual sensor data to identify locomotion modes. Our approach utilizes two fusion methods: intermediate fusion, which combines the data in the form of features, and late fusion, which integrates the sensor data by averaging the recognition results from each sensor. By fusing these two different modalities, the prediction accuracy improved by an average of 11.7% with the test data. Through comparisons with uni-modal LMR systems that rely on a single type of sensor data for locomotion mode recognition, we found that the improved performance of the multi-modal LMR system is due to the visual information's ability to generalize different gait patterns across users and the mechanical sensor data's consistency within the same classes.**

***Index Terms*—Wearable Robotics; Sensor Fusion; Embedded Systems for Robotic and Automation**

Fig. 1. Lower-limb wearable hip complex assist robot MoonWalk.Omni with integrated RGB camera for visual perception

I. INTRODUCTION

LOWER limb exoskeletons have the potential to significantly enhance gait motion in individuals by providing additional support. To maximize the effectiveness of this assistance, it is essential to apply the appropriate magnitude of support at the correct time [1], [2]. To achieve this, control methodologies have been developed that are tailored to the joint kinetic and kinematic patterns of each locomotion mode [3], [4]. For these control systems to be effective and enhance the exoskeleton's overall assistive performance, immediate and accurate locomotion mode recognition (LMR) must occur before the system is applied [5].

To date, researchers have recognized locomotion modes using mechanical sensor data, such as inertial measurement units (IMUs), which capture user's proprioceptive movement. Early methods applied classical machine learning models like Support Vector Machines (SVM) and Random Forests using handcrafted features [6], whereas more recent work has leveraged deep learning architectures such as Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks to automatically learn spatiotemporal features [7], [8]. Although both approaches have shown good LMR performance, methods relying on mechanical sensor data often struggle to generalize across individuals due to gait variability [9], and their recognition performance could degrade when assistive torque from exoskeletons alters joint kinematics [10], [11].

Recently, inspired by the visual guidance of human locomotion, some studies have utilized environmental information to recognize locomotion modes. [12] introduced the "ExoNet" database, the largest and most diverse open-source dataset of walking environments for environment-adaptive locomotion mode recognition systems. Tricomi et al. [13] demonstrated that modulating assistance based on walking terrains improves energy efficiency compared to systems that apply static assistance magnitudes for all locomotion modes. Despite their potential, visual sensors primarily capture environmental context—where the user is moving—rather than the user's

Manuscript received: February 6, 2025; Revised May 29, 2025; Accepted July 21, 2025

This paper was recommended for publication by Editor K.-U. Kyung upon evaluation of the Associate Editor and Reviewers' comments. This work was supported by the Korea Institute of Science and Technology Institutional Programs (Project No. 2E33602, 2V10190)

[1]Hyesoo Shin, Sangdo Kim, Sunwoo Kim, Jongwon Lee and KangGeon Kim are with Artificial Intelligence and Robotics Institute, Korea Institute of Science and Technology, Seoul, South Korea hyesoo030@kist.re.kr, sangdo322@kist.re.kr, sunu0915@kist.re.kr, jwlee@kist.re.kr, danny@kist.re.kr

[2]Hyesoo Shin and Jinkyu Kim are with Department of Computer Science and Engineering, Korea University, Seoul, South Korea hyesoo030@korea.ac.kr, jinkyukim@korea.ac.kr

[3]Sangdo Kim and Sunwoo Kim are with School of Electrical Engineering, Korea University, Seoul, South Korea ysmc1440@korea.ac.kr, sunu1231@korea.ac.kr

*Corresponding Authors

Digital Object Identifier (DOI): see top of this page.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

movement dynamics, limiting their ability to reflect internal states.

While most prior studies on LMR have focused on single-sensor modalities, humans rely on multiple sensory inputs, such as proprioceptive feedback and visual cues from the environment, to adapt their locomotion seamlessly [14]. Inspired by this multi-sensory integration in human movement, we propose multi-modal LMR methods combining proprioceptive and visual perceptions to enhance recognition performance.

Previous studies have explored the combination of mechanical sensors and bioelectrical sensors such as electromyography sensor (EMG), leveraging their complementary roles: EMG reflects motor intent, while mechanical sensors capture movement kinematics [15]. However, EMG signals are sensitive to electrode placement and vary across users and conditions, limiting their use in long-term or outdoor settings [16]. In contrast, mechanical and visual sensors are easier to integrate into exoskeleton systems, making them more suitable for real-time recognition in assistive applications. Based on these advantages, we implement a multi-modal LMR system that integrates mechanical and visual sensors for practical use in wearable robotics.

To make this system feasible for embedded devices of the exoskeleton, we utilize two lightweight fusion strategies: intermediate fusion and late fusion. These approaches strike a balance between computational efficiency and recognition accuracy, making them suitable for embedded systems. Moreover, although multi-modal fusion shows intuitive benefits, the individual contributions of proprioceptive and visual modalities in LMR remain underexplored. To address this, we perform a systematic comparison of uni-modal and multi-modal models. Additionally, to support this analysis, we constructed a novel multi-modal outdoor dataset using MoonWalk.Omni, a lower-limb wearable hip-assist robot (Fig. 1). This dataset captures synchronized mechanical and visual sensor data across diverse real-world environments and under varying assistive conditions, enabling comprehensive evaluation of both uni-modal and multi-modal LMR models.

The main contributions of this work are as follows:

- A lightweight multi-modal LMR system is proposed, combining proprioceptive and visual inputs inspired by human locomotor control.
- A systematic comparison of uni-modal and multi-modal recognition models is conducted to analyze the contributions of each modality.
- A novel outdoor dataset is developed, comprising synchronized mechanical and visual data collected across diverse environments and assistive conditions, enabling robust evaluation of LMR performance.

II. Method

A. Uni-modal Locomotion Mode Recognition

To evaluate the effect of two modalities: 1) mechanical sensor data from proprioceptive perception, 2) visual sensor data from visual perception, we adopted a deep neural network specified to each single modality data.

1) Mechanical LMR: We employed the LSTM-CNN architecture (Fig. 2(a)) to train a deep neural network that predicts locomotion modes using mechanical sensor data. The LSTM-CNN architecture is commonly used for human activity recognition [8]. It combines both LSTM and CNN components: the LSTM component automatically learns temporal patterns from sequential data, while the convolutional layers extract spatial features. By integrating both temporal and spatial features, this model achieves high accuracy and generalization capability with a lightweight architecture. For mechanical LMR, we implemented the LSTM-CNN architecture with two layers in both the LSTM and CNN components, following the design described in [8]. Three fully connected (FC) layers were added to the end of the LSTM and CNN components to generate the final prediction.

2) Visual LMR: For the lower-limb exoskeleton robot to use the LMR system in real time, the recognition network must have low computational cost. While videos provide more information including temporal data, leading to potentially higher accuracy, they also require more computation, making them unsuitable for real-time use. Therefore, we used an image classification model to process visual sensor data. The MobileNetV2 network (Fig. 2(b)) was selected for its high accuracy and low computational demand, making it ideal for mobile and resource-constrained environments [17]. The network consists of bottleneck blocks with inverted residuals. Inverted residuals are based on residual layers, which use connections between the input and output of the layer to facilitate efficient training of deep neural networks. Unlike standard residual layers, the bottleneck inputs in inverted residual blocks are expanded into a high-dimensional space to learn more representative features, which are then projected back to the input dimension to be connected to the input. The MobileNetV2 structure described in [17] was used for the image classification model, followed by the three-layer FC prediction head.

Additionally, to assess whether temporal information was useful enough to justify the higher computational cost, we trained a video classification network. We implemented the ResNet architecture with 18 layers, using R(2+1)D convolution (Fig. 2(c)) [18]. R(2+1)D convolution decomposes 3D convolutions into 2D spatial and 1D temporal convolutions, increasing the number of nonlinearities and improving optimization performance while maintaining high accuracy.

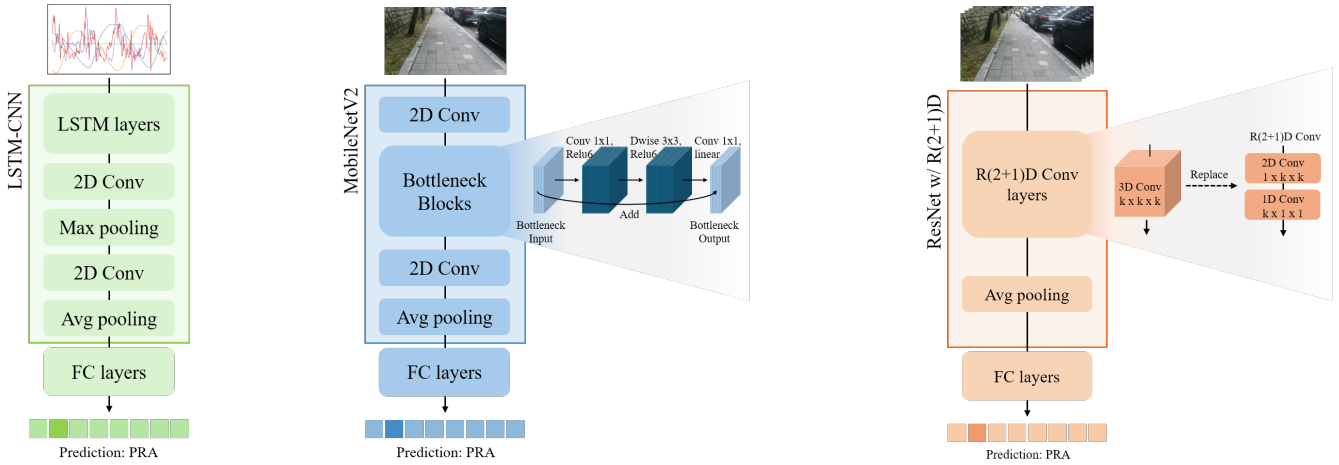
B. Multi-modal Locomotion Mode Recognition

Three levels of fusion—Early, Intermediate, and Late—are commonly used to integrate multi-modal data [19], [20], [21].

Early Fusion integrates raw sensor data directly into a joint representation. Since sensor data are merged before being processed through the network, it can effectively learn the correlations between different data types. However, due to the heterogeneous nature of raw data, early fusion is difficult to implement as direct combination of disparate modalities poses significant challenges.

Intermediate Fusion merges features extracted from multiple modalities and learns a joint representation using an additional

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.



(a) Mechanical LMR (b) Visual LMR - Image (c) Visual LMR - Video

Fig. 2. Uni-modal locomotion mode recognition network architecture

network. This approach allows the model to utilize data from different sources to make predictions automatically. Since features are easier to merge than raw data, various integration methods can be applied with this technique.

Late Fusion combines data at the decision level, allowing the model to be trained separately on each modality while still being indirectly influenced by the other. Although this approach does not directly combine the information from different modalities, it can still perform well, especially in cases where one sensor fails.

Since early fusion of heterogeneous sensor data could be challenging, intermediate and late fusion methods were employed. For intermediate fusion (Fig. 3(a)), the same network architectures used in the uni-modal models were adopted as feature extractors. Features from each modality were concatenated at the feature level to learn correlative information and passed through a three-layer fully connected network, followed by a softmax function that converts the output logits into a probability distribution:

$$Z = FC\left(\, concat\left(\, z_{mec}, z_{vis} \,\right)\right) \tag{1}$$

$$LM_{MID} = argmax\left(\, softmax\left(Z\right)\right) \tag{2}$$

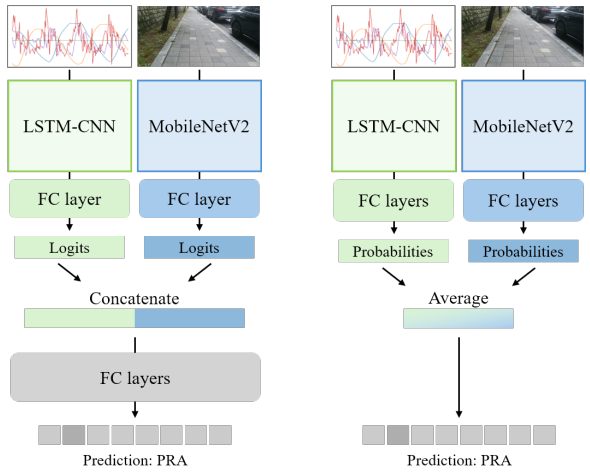
where LM denotes the recognition result of the method, and $argmax$ represents the argmax function, which returns the index of the largest value. $softmax, FC$ and $concat$ denote the softmax function, the final fully connected layer, and the concatenation function, respectively. z_{mec} and z_{vis} are the output logits from the LSTM-CNN and MobileNetV2, which have the same architecture as used in the uni-modal methods.

For late fusion (Fig. 3(b)), the architecture used in the uni-modal methods was applied. The scores from the final fully connected layer of each model were averaged to produce the final result as follows:

$$LM_{LATE} = argmax\left(\, \alpha\, z'_{mec} + (1-\alpha)\, z'_{vis} \,\right) \tag{3}$$

where $z' = softmax\left(z\right)$ and α were set to 0.5 to obtain the average of the two values. These fusion strategies were chosen not only for their conceptual simplicity but also for their compatibility with embedded systems in exoskeletons, where memory and latency constraints are critical.

Table I showed more detail of the architectures such as number of parameters, input sizes and average inference latency over 100 runs on the NVIDIA GeForce RTX 4090 GPU. Due to the inherent differences in input structure and the architectural demands of each modality, the capacity of the models is different. Rather than matching models by size, we aimed to use architectures that are representative and effective for their respective data types. While the network structures differ depending on the modality, the same FC structure is consistently used across all models. This design choice helps to ensure that performance differences can be primarily attributed to the modality or fusion strategy, rather than to differences in prediction head capacity.



(a) Intermediate fusion (b) Late fusion

Fig. 3. Multi-modal locomotion mode recognition network architecture

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

TABLE I
COMPARISON OF DIFFERENT MODELS WITH VARIOUS INPUT MODALITIES AND ARCHITECTURES. FE+MAG(IMU) CONSISTED OF HIP FLEXION-EXTENSION ANGLES (FE) AND THE MAGNITUDES OF ACCELEROMETER, GYROSCOPE, AND QUATERNION DATA (MAG(IMU)). FC STRUCTURE REPRESENTS THE NUMBER OF OUTPUT UNITS IN EACH OF THE THREE FULLY CONNECTED LAYERS, IN ORDER.

Model	Input Modality	Input Size	Backbone	# Params	Latency (ms)	FC Structure
MEC	FE+MAG(IMU)	200×5	LSTM+CNN	73,928	1.8	[128,64,8]
IMG	Image	640×360×3	MobileNetV2	2,396,616	1.4	[128,64,8]
VID	Video	16×640×360×3	ResNet18 with R(2+1)D	31,553,553	103.5	[128,64,8]
MID	FE+MAG(IMU), Image	200×5, 640×360×3	LSTM+CNN, MobileNetV2	2,478,152	3.2	[128,64,8]
LATE	FE+MAG(IMU), Image	200×5, 640×360×3	LSTM+CNN, MobileNetV2	2,487,056	3.3	[128,64,8]

C. Multi-modal Outdoor Dataset

To evaluate and compare the performance of various LMR network architectures, we constructed a comprehensive multi-modal outdoor dataset integrating mechanical and visual sensor data. The dataset was used to assess three uni-modal LMR networks: a mechanical sensor LMR network (MEC), an image LMR network (IMG), and a video LMR network (VID) as well as two multi-modal networks: an intermediate fusion network combining mechanical sensor and image data (MID), and a late fusion network integrating mechanical sensor and image data (LATE).

As shown in Fig. 1, data collection was conducted using MoonWalk.Omni, a wearable hip-assist robot designed to support older adults experiencing age-related mobility decline. The robot's sensing system included a 6-axis IMU mounted on the back, which captured trunk movement, and two joint encoders mounted on the hips to measure flexion-extension and abduction-adduction angles for each leg. Consequently, the mechanical sensor data consisted of hip flexion-extension angles, hip abduction-adduction angles, and accelerometer and gyroscope readings from the trunk-mounted IMU.

The visual data consists of egocentric RGB video capturing the environment approximately 2–3 steps ahead, consistent with prior findings on human gaze behavior during walking [22], [23]. This was collected using an RGB-D camera (OAK-D-IoT-40, Luxonis) mounted on the front of the trunk with a pitch angle of 75 degrees, offering a stable and forward-facing egocentric view. Compared to head- or leg-mounted cameras, this placement provides a more stable and consistent visual perspective, integrated with the movement of the robot [24].

In the assist-on condition, where the robot supports muscle strength, differences occur in the joint angles, angular velocity, and posture information during walking compared to the assist-off condition, where no muscle assistance is provided. To examine the generalizability of LMR models across different motor dynamics caused by robotic assistance, the multi-modal sensor data collected under two conditions, assist-on and assist-off. In the assist-on condition, the robot provided muscle support via adaptive torque generation using the Delayed Output Feedback Control (DOFC) method [25]. This controller produces assistive torque proportional to the leg joint's range of motion and dynamically adapts to changes in terrain and speed, without requiring manual parameter adjustment.

To capture the distinct gait requirements such as joint kinematics, timing, and control strategies, induced by environmental conditions, we focused on three environmental factors that substantially alter the hip mechanics: slope direction (ascent vs. descent), vertical-path type (ramp vs. stair), and surface condition (paved vs. unpaved) [26], [27]. Based on these factors, the dataset consisted of eight labeled classes that reflect common and biomechanically distinct conditions encountered in daily walking environments, providing a practical yet comprehensive basis for real-world deployment: unpaved ramp ascent (URA), paved ramp ascent (PRA), unpaved ramp descent (URD), paved ramp descent (PRD), unpaved flat (UF), paved flat (PF), upstairs (US), and downstairs (DS).

III. EXPERIMENTS

A. Experimental Protocol

To collect the dataset described above, we conducted structured walking trials with thirteen healthy adult participants (height: 171.0 ± 9.4 cm; weight: 67.9 ± 12.9 kg; age: 26.1 ± 1.7 years) approved by the Korea Institute of Science and Technology Institutional Review Board (KIST-202310-HR-003). Although the exoskeleton used in this study is designed to assist older adults with age-related mobility decline, healthy participants were selected to evaluate the baseline performance of the proposed LMR system under controlled conditions. Future work will extend this approach to the intended target population to assess generalizability and robustness in more variable real-world scenarios.

Data were collected outdoors in a natural and built environment spanning approximately 40,800 square meters, which included elevation changes of around 60 meters. The walking paths included a mix of paved roads, mountain trails, sloped paths, and stairs (Fig. 4). Three courses (yellow, green, and blue) covering approximately 1.3–1.4 km were used to capture diverse walking conditions.

Participants were divided into three groups, ensuring a range of heights across three walking paths: yellow, green, and blue. Each group walked their assigned path at a self-selected pace while wearing the hip-assist exoskeleton. To account for both assistive conditions (assist-on and assist-off) and to balance ascent and descent classes, each participant walked the route four times: twice in each direction (clockwise and counterclockwise), accompanied by the same labeler. Assistive torque was provided using the DOFC method and the root mean squared (RMS) torque delivered was 3.54 Nm, with a maximum of 11.39 Nm.

Multi-modal sensory data were collected sequentially and time-aligned. All mechanical sensor data were sampled at

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

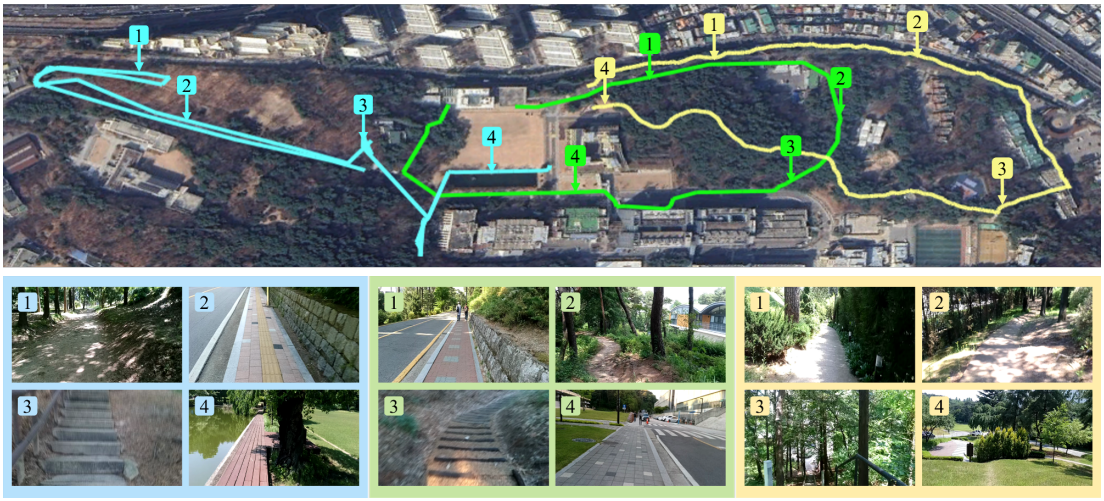


Fig. 4. Real-world locomotion mode recognition experiment course. Map of the 1.3-1.4km course used for real-world data acquisition and validation.

100 Hz, and visual data were recorded as video at 30 Hz with a resolution of 640×360. For training, we processed the sequential mechanical sensor data into two-second segments with a 50% overlap, providing sufficient information while keeping the segments small enough for processing within the robot. For visual sensor data, 16 frames were uniformly sampled from the two-second clip to serve as input for the video network, and one frame from the middle of the clip was used as input for the image network.

B. Training and Evaluation Setup

Although the dataset included both assist-on and assist-off conditions, only the assist-off data were used for training to promote adaptability across control methods. Both conditions were used in the test set to evaluate the impact of assistive torque on LMR performance. The dataset consisted of 2,664 training samples, with 333 samples per class, and 840 testing samples across both torque settings, with 105 samples per class.

Since the collected sensor data were insufficient to create additional validation sets, we employed K-fold cross-validation to fully utilize the dataset and prevent overfitting. We selected a K value of 5 and trained the model with early stopping set to 20 epochs. The visual sensor data were randomly augmented using color jitter, greyscale, and horizontal flip to improve generalization. We fine-tuned all layers using the ImageNet pre-trained model for all models utilizing visual sensor data to enhance training efficiency.

Except for the VID network, which requires more computational memory, we used a batch size of 30 for all methods (8 for the VID network). The initial learning rate was set to 0.000002 for the MEC method and 0.0002 for the other networks, with the cosine annealing algorithm used for learning rate scheduling. All training and evaluation processes were conducted on the NVIDIA GeForce RTX 4090 GPU.

C. Mechanical Sensor Configuration

Both uni-modal and multi-modal LMR networks require consistent training conditions. In our dataset, the mechanical sensor data included a 6-axis IMU and two encoders mounted on the hip joint to measure hip flexion-extension and abduction-adduction angles. We first investigated performance using various mechanical sensor configurations to determine the optimal sensor configuration for LMR.

We integrated the quaternion values computed from the accelerometer and gyroscope data for the IMU sensor data. The configuration included the magnitudes of the accelerometer, gyroscope, and quaternion data. The possible sensor configurations consisted of hip flexion-extension angles (FE), hip abduction-adduction angles (AA), accelerometer, gyroscope, quaternion data (IMU), and the magnitudes of accelerometer,

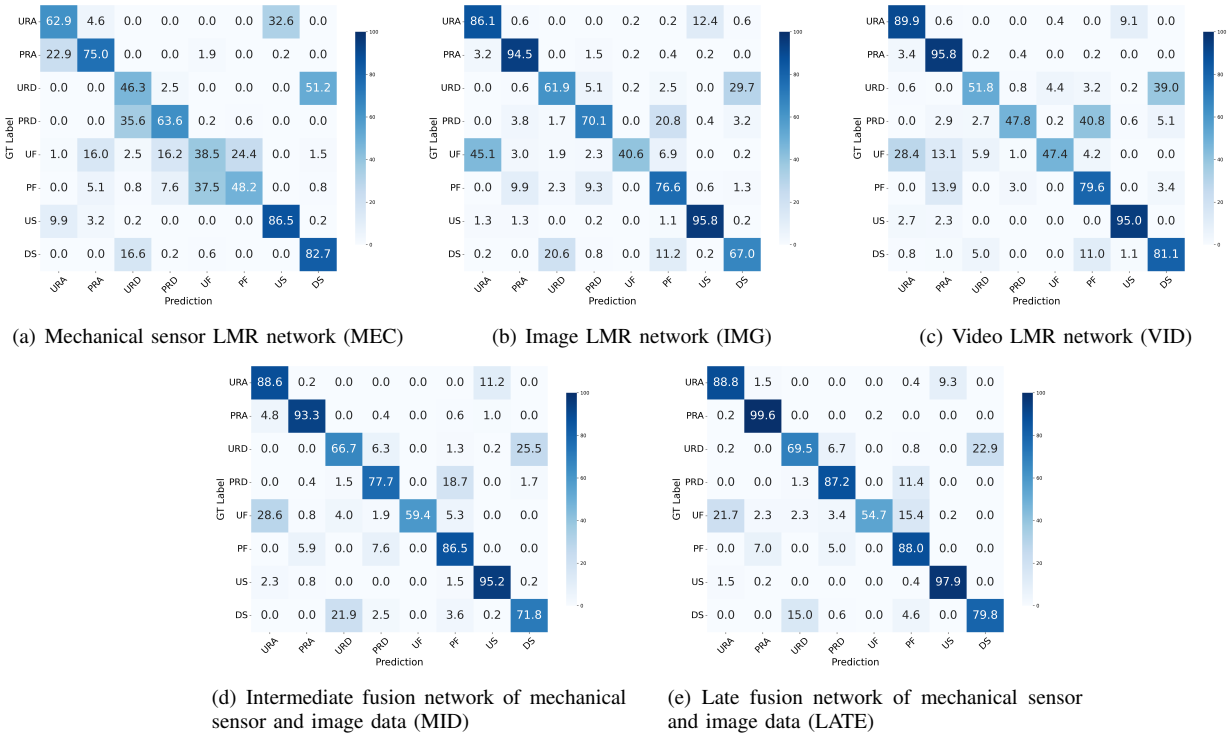
TABLE II
THE LOCOMOTION MODE RECOGNITION ACCURACY OF MEC METHOD WITH VARIOUS SENSOR CONFIGURATIONS. TESTED WITH THE DATA COLLECTED WITH ASSIST-OFF CONDITION

	FE	FE+IMU	FE+MAG(IMU)
Acc (%)	55.4	45.9	**63.0**
	AA	AA+IMU	AA+MAG(IMU)
Acc (%)	9.5	21.6	21.5
	FE+AA	FE+AA+IMU	FE+AA+MAG(IMU)
Acc (%)	43.4	38.2	50.6
	-	IMU	MAG(IMU)
Acc (%)	-	26.8	29.0

TABLE III
LOCOMOTION MODE RECOGNITION ACCURACY (%) AND ACCURACY DIFFERENCE BETWEEN ASSIST-OFF AND ASSIST-ON CONDITION OF EACH NETWORK

	MEC	IMG	VID	MID	LATE
assist-off	63.0	74.1	74.9	79.9	**83.2**
assist-on	51.7	77.8	78.9	81.3	**82.5**
accuracy diff	-11.3	+3.7	+4.0	+1.4	**-0.7**

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.



(a) Mechanical sensor LMR network (MEC)

(b) Image LMR network (IMG)

(c) Video LMR network (VID)

(d) Intermediate fusion network of mechanical sensor and image data (MID)

(e) Late fusion network of mechanical sensor and image data (LATE)

Fig. 5. Confusion matrix of each network. Each row means the ground truth label, and each column represents the prediction result of the network. The numbers represent the classification accuracy for each class using test data collected without the assistive torque of the exoskeleton. Eight classes are shown in the confusion matrix: unpaved ramp ascent (URA), paved ramp ascent (PRA), unpaved ramp descent (URD), paved ramp descent (PRD), unpaved flat (UF), paved flat (PF), upstairs (US), and downstairs (DS)

gyroscope, and quaternion data (MAG(IMU)).

We trained the mechanical LMR method using all combinations of these sensor configurations. Table II shows the LMR accuracy for each case using the MEC method. The network showed worse classification performance when hip abduction-adduction angles were included, compared to when these data were excluded. Including the magnitude of IMU sensor data improved the model's ability to learn more representative features for each class. The configuration that combined hip flexion-extension angles with the magnitudes of accelerometer, gyroscope, and quaternion data predicted locomotion modes most accurately, achieving 63.0% accuracy among other configurations.

As a result, we used the mechanical sensor configuration of hip flexion-extension angles and the magnitudes of accelerometer, gyroscope, and quaternion data for all other networks utilizing mechanical sensor data.

D. Recognition Results

The evaluation results of all LMR networks are shown in Table III in terms of accuracy. For the test dataset under the same condition (assist-off) as the training data, the MEC method achieved a recognition accuracy of 63.0%. This performance was lower than that of the other uni-modal LMR methods, with IMG and VID achieving accuracies of 74.1% and 74.9%, respectively.

Moreover, since gait features can be altered by the assistive torque of the exoskeleton, the MEC method experienced the largest performance drop, with an 11.3% decrease when tested on data collected under the assist-on condition. In contrast, when tested with the assist-on dataset, the IMG network showed a 3.7% improvement in accuracy. Since the image classification network does not directly rely on joint movement, it demonstrated better generalization to varying gait characteristics.

By fusing the two modalities, accuracy improved for both test datasets compared to all uni-modal methods. In the assist-off test dataset, intermediate fusion achieved 79.9% accuracy, while late fusion reached 83.2%. For the assist-on test dataset, intermediate fusion achieved 81.3% accuracy, while late fusion reached 82.5%.

IV. DISCUSSION

Since the multi-modal approach yielded superior overall performance, it is essential to interpret the results beyond raw accuracy. Each input modality provides different levels of information richness and structural complexity, influencing learning dynamics and model behavior. Therefore, accuracy alone may not fully capture the relative strengths of each model. To complement quantitative metrics, we include qualitative analyses—such as confusion matrices and case-specific error analysis—to better understand the role of each modality while reducing structural and procedural bias.

1) Mechanical LMR: The classification performance for each class in the test dataset with the assist-off condition is visualized as a confusion matrix (Fig. 5). As shown in Fig. 5(a), the lower performance of the MEC method was

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

(a) The effect of a subject's height on a camera's field of view (L: 169cm, R: 189cm)

(b) The effect of the movement of the trunk (L: Paved Ramp Descent, R: Paved Flat)

(c) The effect of the movement of the trunk (L: Unpaved Flat R: Unpaved Ramp Ascent)

(d) The effect of the slope of the experimental environment (L: Unpaved Ramp Descent, R: Down Stair)

Fig. 6. Visual sensor data from multi-modal outdoor dataset showing the effect of subject's height (a), movement of the trunk (b),(c) and the slope of the experimental environment (d), which lead to misclassification.

primarily due to the misclassification of paved classes as unpaved classes. Distinguishing between paved and unpaved surfaces based solely on joint movement is challenging, which may have disadvantaged the MEC method.

However, the well-classified results for unpaved ramp classes suggest that unpaved environments exhibit distinctive features, indicating a need for different assistive control strategies. Apart from the difficulty in distinguishing between paved and unpaved surfaces, the network showed strong intra-class consistency. Compared to other uni-modal methods, it displayed less confusion among ramp ascent, ramp descent, and flat classes.

2) Visual LMR: Fig. 5(b) shows the image classification accuracy for each class. Although the IMG network outperforms MEC, it exhibits a high error rate of 45% in misclassifying the unpaved flat class as unpaved ramp ascent. Additionally, it has a 21% error rate in confusing the paved ramp descent class with the paved flat class—errors rarely seen with the MEC method.

Since the visual sensor is mounted on the trunk, differences in user height could cause inconsistencies in visual data within the same class (Fig. 6(a)). Misclassifications could also arise when environmental features from different classes appear similar due to trunk movement (Fig.6(b), Fig.6(c)). In addition, the slopes of unpaved ramps and stairs are often similar in our dataset (Fig.6(d)), resulting in high error rates when distinguishing between these classes across all uni-modal methods.

For the VID method, despite the network having access to more information from the sensor data, it does not demonstrate significant performance improvement over the IMG method. This may be due to the nature of our dataset, which lacks substantial temporal information that could aid in classifying locomotion modes.

3) Multi-modal LMR: Due to the inclusion of consistent within-class data from mechanical LMR, except for confusion between unpaved and paved surfaces, both multi-modal LMR networks demonstrated consistent and robust performance in recognizing locomotion modes in both test settings (assist on and off). The generalization properties of visual LMR across different gait characteristics further contributed to performance gains, with LATE fusion outperforming MEC by approximately 30.8% in the assist-on setting and multi-modal models surpassing uni-modal ones by an average of 11.7% across all test conditions.

However, the observed influence of each modality suggests an imbalance in informational richness between proprioceptive and visual inputs. For example, fusion reduced misclassification of unpaved ramp descent as down stair—a case difficult for both uni-modal models—but visually induced errors, such as confusion between paved ramp descent and paved flat class, remained. These errors were more effectively mitigated by the LATE method (9.4%) than by MID (2.1%), indicating MID's higher sensitivity to visual noise.

This difference could reflect how each fusion strategy handles heterogeneous information. Although intermediate fusion has the potential to capture feature-level correlations between modalities [19], it typically requires more expressive architectures to integrate heterogeneous features effectively [28]. In our implementation, the feature concatenation followed by fully connected layers used in MID may have been insufficient to fully exploit the complementary characteristics of mechanical and visual signals. In contrast, LATE fusion, which aggregates decisions at the score level, likely helps avoid such interference, contributing to greater robustness and better performance.

V. CONCLUSIONS

Inspired by the human locomotor control system, we proposed a lightweight multi-modal locomotion mode recognition (LMR) system that integrates mechanical sensor data for proprioceptive input and visual sensor data for environmental perception.

To support its development and evaluation, we constructed a multi-modal outdoor dataset encompassing eight distinct gait environments, with data collected under both assist-off and assist-on conditions.

The effectiveness of multi-modality was evaluated by comparing three uni-modal LMR networks with two multi-modal

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

fusion models. The multi-modal networks achieved an overall accuracy of 81.7%, outperforming all uni-modal baselines across both assistive settings. Visual modality offered strong generalization across varying gait dynamics, while mechanical signals provided consistent intra-class representations. These complementary characteristics were effectively exploited through fusion, leading to enhanced LMR performance. The results show that lightweight fusion strategies—such as intermediate feature concatenation and late score-level averaging—can substantially improve recognition accuracy without compromising real-time feasibility, making them well-suited for deployment on wearable robots.

While our findings offer valuable insight into the respective roles of proprioceptive and visual modalities in LMR, they are based on a specific sensor configuration and dataset. Given the variability in sensor types and locomotion settings across real-world scenarios, future research is needed to assess the generalizability of our approach. Additionally, as our multi-modal LMR system is intended for integration with a lower-limb exoskeleton, future work will focus on real-time implementation and incorporation into a control framework that dynamically adjusts assistive torque profiles based on predicted locomotion modes.

REFERENCES

[1] A. J. Young, J. Foss, H. Gannon, and D. P. Ferris, "Influence of power delivery timing on the energetics and biomechanics of humans wearing a hip exoskeleton," *Frontiers in bioengineering and biotechnology*, vol. 5, p. 4, 2017.

[2] P. Malcolm, W. Derave, S. Galle, and D. De Clercq, "A simple exoskeleton that assists plantarflexion can reduce the metabolic cost of human walking," *PloS one*, vol. 8, no. 2, p. e56137, 2013.

[3] I. Kang, D. D. Molinaro, S. Duggal, Y. Chen, P. Kunapuli, and A. J. Young, "Real-time gait phase estimation for robotic hip exoskeleton control during multimodal locomotion," *IEEE robotics and automation letters*, vol. 6, no. 2, pp. 3491–3497, 2021.

[4] K. Seo, K. Kim, Y. J. Park, J.-K. Cho, J. Lee, B. Choi, B. Lim, Y. Lee, and Y. Shim, "Adaptive oscillator-based control for active lower-limb exoskeleton and its metabolic impact," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 6752–6758.

[5] Y. Qian, Y. Wang, C. Chen, J. Xiong, Y. Leng, H. Yu, and C. Fu, "Predictive locomotion mode recognition and accurate gait phase estimation for hip exoskeleton on various terrains," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 6439–6446, 2022.

[6] F. Labarrière, E. Thomas, L. Calistri, V. Optasanu, M. Gueugnon, P. Ornetti, and D. Laroche, "Machine learning approaches for activity recognition and/or activity prediction in locomotion assistive devices—a systematic review," *Sensors*, vol. 20, no. 21, p. 6345, 2020.

[7] H. T. T. Vu, H.-L. Cao, D. Dong, T. Verstraten, J. Geeroms, and B. Vanderborght, "Comparison of machine learning and deep learning-based methods for locomotion mode recognition using a single inertial measurement unit," *Frontiers in neurorobotics*, vol. 16, p. 923164, 2022.

[8] K. Xia, J. Huang, and H. Wang, "Lstm-cnn architecture for human activity recognition," *IEEE Access*, vol. 8, pp. 56 855–56 866, 2020.

[9] S. Suh, V. F. Rey, and P. Lukowicz, "Tasked: transformer-based adversarial learning for human activity recognition using wearable sensors via self-knowledge distillation," *Knowledge-Based Systems*, vol. 260, p. 110143, 2023.

[10] T. K. Uchida, A. Seth, S. Pouya, C. L. Dembia, J. L. Hicks, and S. L. Delp, "Simulating ideal assistive devices to reduce the metabolic cost of running," *PloS one*, vol. 11, no. 9, p. e0163417, 2016.

[11] S. Song and S. H. Collins, "Optimizing exoskeleton assistance for faster self-selected walking," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 786–795, 2021.

[12] B. Laschowski, W. McNally, A. Wong, and J. McPhee, "Computer vision and deep learning for environment-adaptive control of robotic lower-limb exoskeletons," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2021, pp. 4631–4635.

[13] E. Tricomi, M. Mossini, F. Missiroli, N. Lotti, X. Zhang, M. Xiloyannis, L. Roveda, and L. Masia, "Environment-based assistance modulation for a hip exosuit via computer vision," *IEEE Robotics and Automation Letters*, vol. 8, no. 5, pp. 2550–2557, 2023.

[14] A.-K. Rogge, D. Hamacher, G. Cappagli, L. Kuhne, K. Hötting, A. Zech, M. Gori, and B. Röder, "Balance, gait, and navigation performance are related to physical exercise in blind and visually impaired children and adolescents," *Experimental brain research*, vol. 239, pp. 1111–1123, 2021.

[15] C. Zhao, K. Liu, H. Zheng, W. Song, Z. Pei, and W. Chen, "Cross-modality self-attention and fusion-based neural network for lower limb locomotion mode recognition," *IEEE Transactions on Automation Science and Engineering*, 2024.

[16] M. Boyer, L. Bouyer, J.-S. Roy, and A. Campeau-Lecours, "Reducing noise, artifacts and interference in single-channel emg signals: A review," *Sensors*, vol. 23, no. 6, p. 2927, 2023.

[17] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[18] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459.

[19] S. Y. Boulahia, A. Amamra, M. R. Madi, and S. Daikh, "Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition," *Machine Vision and Applications*, vol. 32, no. 6, p. 121, 2021.

[20] K. Gadzicki, R. Khamsehashari, and C. Zetzsche, "Early vs late fusion in multimodal convolutional neural networks," in *2020 IEEE 23rd international conference on information fusion (FUSION)*. IEEE, 2020, pp. 1–6.

[21] C. G. Snoek, M. Worring, and A. W. Smeulders, "Early versus late fusion in semantic video analysis," in *Proceedings of the 13th annual ACM international conference on Multimedia*, 2005, pp. 399–402.

[22] J. S. Matthis, J. L. Yates, and M. M. Hayhoe, "Gaze and the control of foot placement when walking in natural terrain," *Current Biology*, vol. 28, no. 8, pp. 1224–1233, 2018.

[23] K. Bonnen, J. S. Matthis, A. Gibaldi, M. S. Banks, D. M. Levi, and M. Hayhoe, "Binocular vision and the control of foot placement during walking in natural terrain," *Scientific reports*, vol. 11, no. 1, p. 20881, 2021.

[24] B. Zhong, R. L. Da Silva, M. Li, H. Huang, and E. Lobaton, "Environmental context prediction for lower limb prostheses with uncertainty quantification," *IEEE Transactions on Automation Science and Engineering*, vol. 18, no. 2, pp. 458–470, 2020.

[25] B. Lim, J. Lee, J. Jang, K. Kim, Y. J. Park, K. Seo, and Y. Shim, "Delayed output feedback control for gait assistance with a robotic hip exoskeleton," *IEEE Transactions on Robotics*, vol. 35, no. 4, pp. 1055–1062, 2019.

[26] D. S. Marigold and A. E. Patla, "Age-related changes in gait for multi-surface terrain," *Gait & posture*, vol. 27, no. 4, pp. 689–696, 2008.

[27] R. Bárbara, S. M. Freitas, L. B. Bagesteiro, M. R. Perracini, and S. R. Alouche, "Gait characteristics of younger-old and older-old adults walking overground and on a compliant surface," *Brazilian Journal of Physical Therapy*, vol. 16, pp. 375–380, 2012.

[28] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," 2017. [Online]. Available: https://arxiv.org/abs/1705.09406