https://doi.org/10.1038/s44182-025-00033-4

Transformer-based multitask assist control from first-person view image and user's kinematic information for exoskeleton robots

Jun-ichiro Furukawa[1,2] ✉ & Jun Morimoto[1,2,3]

Exoskeleton robots necessitate the capacity to promptly generate appropriate assistance for the user's needs across a range of motion scenarios. In this study, we developed a multitask assistance control strategy using a transformer that generates control commands to the exoskeleton robot according to the user's status and the environment. Our approach captured the user's joint and trunk motion and a first-person view image as input to the control system. A series of motion tasks were employed to validate the implemented AI with the proposed approach, including walking, squatting, and a step-up movement. Healthy subjects participated and the application of AI with our method to an exoskeleton robot reduced muscle loads. Moreover, the learned assist strategy was found to generalize, reducing muscle activity in another participant. These findings represent a first step toward achieving exoskeleton robot control that assists diverse movements across individuals in various environments using our transformer-based approach.

Exoskeleton robots attract considerable interest due to the aging of the population and have been instrumental in assisting individuals with movement disorders[1,2]. These robots are frequently utilized for the purpose of rehabilitation training at locations that are equipped with additional safety systems[3,4]. The controllers for these assistive robots are often manually designed, and the robot movements are pre-determined[5]. To overcome this situation, human-in-the-loop optimization techniques are proposed to customize the assist control strategy for each user[6,7]. These techniques employ the user's metabolic cost[8,9], the user's preference[10], or muscle activity[11] as the objective function. This technology enables personalized control. However, current methodologies have only been employed to facilitate periodic movements such as walking. Furthermore, the optimization process often requires a significant investment of time and effort from the user to identify optimal parameters, and proposals to overcome these problems are being tackled[12].

On the other hand, when using exoskeleton robots in everyday life, such as in non-medical applications, the ability to respond to a broader range of movements in accordance with the user's intentions and the surrounding environment is of greater importance. To capture the user's motion intention, electromyography (EMG) is often used, and EMG-based control approaches are proposed[13] using either a finite state machine[14,15], event-based approach[16], regression-based approach[17,18] or data-driven approaches[19]. These EMG-based approaches enable the user to control the exoskeleton robot intuitively. However, they require time-consuming sensor attachment and careful calibration for each user and each day. This cumbersome prevents the daily use of assistive devices. Consequently, these EMG-based conventional methods have only been employed in limited situations, such as in rehabilitation sites. This paper examines the potential of utilizing visual data to assess the surrounding environment, enabling the exoskeleton robot to select appropriate assistive commands in response to the environment without sensors, which are difficult to handle, such as EMG.

To date, only a few studies employ vision sensors to control the exoskeleton robot. For instance, it is proposed to detect an obstacle, such as a small box on the walking path, from depth images and to change the swing leg height to provide sufficient foot clearance to overcome the obstacle[20]. On the other hand, this study focuses on efficient assist generation for exoskeleton robots using image information, which has a different purpose and task from other vision-based approaches (e.g., obstacle detection and object recognition). To the best of our knowledge, no assist algorithm has yet been

[1]Man-Machine Collaboration Research Team, Guardian Robot Project, RIKEN, Seika-cho, Soraku-gun, Kyoto, Japan. [2]Department of Brain Robot Interface, ATR Computational Neuroscience Labs, Seika-cho, Soraku-gun, Kyoto, Japan. [3]Graduate School of Informatics, Kyoto University, Kyoto-Shi, Kyoto, Japan. ✉e-mail: junichiro.furukawa@riken.jp

developed that can generate assist actions to cope with the various situations by utilizing the visual input.

In this study, we seek to leverage recent advancements in computer vision technology, namely deep neural networks, to capture a human movement sequence over a specified period rather than at a single point in time[21]. To this end, we employed a transformer[22], which is capable of accurately generating the output sequence from the input sensor information history. The outputs of the transformer were then used to generate the action sequence of the assistive devices. Specifically, we propose a transformer-based motion generation method for controlling an exoskeleton robot with the objective of assisting user movements in daily life. The sensor inputs to the transformer are composed of the visual information captured by the first-person view camera and the kinematic information, including the user's joint angle and angular velocities. Recent studies have also shown that RNN-based methods such as LSTM and GRU are effective in predicting human motion[23,24]. These methods are effective for time series information from a single sensor. On the other hand, this study integrates different types of sensors, first-person view images, and kinematic information. The transformer-based approach can effectively learn long-range dependencies between different data elements using a self-attention mechanism. It can consistently integrate complex kinematic information and visual data, which is expected to demonstrate superior performance in our assisted robot control task.

A motion sequence generation task was conducted to evaluate the efficacy of our proposed method. The task involved squatting down to pick up an object on the floor and climbing a step (Fig. 1a), which exemplifies movement in everyday situations. The participant was instructed to generate a motion sequence freely, without the necessity of following a predetermined order of motions. For example, the subject could squat in front of an object, then walk to a step and ascend it. The above procedure for generating motion sequences was employed to collect labeled data from the sensors attached to the participants (Fig. 1c). The data was labeled by the participants themselves in a manner that involved direct interaction with the system. The participants were instructed to press a button whenever they felt the need for assistance with the exoskeleton robot. Subsequently, the system was trained with the labeled data to enable the prediction of the user's command sequences. Each element of the sequence was either a button press or not, corresponding to the situation in which assistance was needed or not.

In the experiments, we used our carbon-frame exoskeleton robot[25] to assist the knee joint movements (Fig. 1b) of the participants. The assist forces to the left and right knees of the exoskeleton robot were provided based on the predicted assist action sequence generated by the trained transformer-based assist controller. The derived control sequence was the pressure command to the pneumatic air muscles (PAM) attached to the exoskeleton robot, which drives the knee joints. The assist force command was either "active" or "free." In the event that the "active" command, which predicts the button-press period by the user, was selected, a constant pressure command was transmitted to the exoskeleton robot. Conversely, if the "free" command was selected, the air pressure to the PAM was not supplied, allowing the user who is wearing the exoskeleton robot to move their knee joint freely without being disturbed by the force generated by the PAM.

The rest of this article is organized as follows. Section 2 shows our experimental results. In Section 3, we describe the discussion. Finally, we explained the methods in Section 4.

Results

Assist control performance

The efficacy of the assistive control performance of our proposed method is demonstrated in the sequence of motion generation tasks, which comprise squatting down to pick up an object on the floor and climbing a step (Fig. 1a).

In this experiment, the movement sequence was set to last three minutes, and the participants were asked to perform it at a rhythm of 40 beats per minute. Consequently, the number of squats and left and right climbing steps was identical in both conditions. The squat was performed 30 times, and the step-climb with the right and left leg was executed 15 times. The first participant (P1) was tasked with labeling the data to train the transformer model. Furthermore, a second participant (P2), who was not involved in the data labeling for model training, was asked to perform the movement task to assess the movement generalization performance of the acquired transformer model using the data from the first participant (P1). At the same time, EMG data from muscles involved in knee joint movements (Fig. 1d) and heart rate data were collected during the experiments.

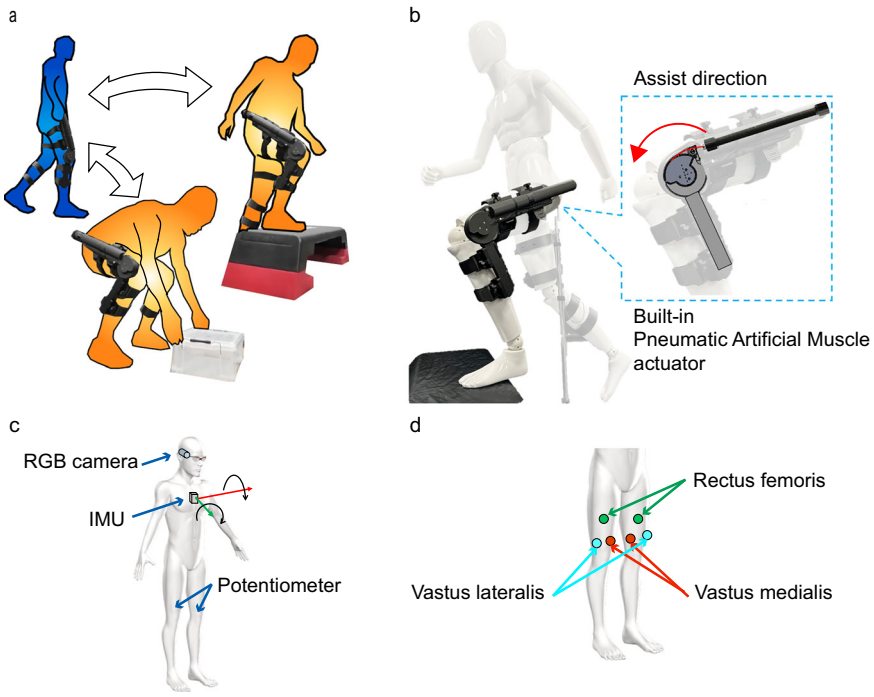


Fig. 1 | Lightweight knee exoskeletons and motion task and data collection. a Motion tasks composed of squatting down to pick up an object on the floor and climbing a step. Each motion was asked to be conducted at a different place, and the participants walked there. **b** Knee exoskeletons. This robot consists of a carbon fiber structure and features a highly responsive joint driven by a pneumatic artificial muscle actuator. **c** Sensors for a model of assist action sequence and their measurement locations. **d** EMG locations for measuring muscle load around the knee joint in a real-time exoskeleton robot control experiment. These EMGs are also used for the EMG-based methods.

Table 1 | Average of EMG signals from three muscles and increased heart rate ratio

	Participant 1		Participant 2	
	Assist	w/o Exo	Assist	w/o Exo
EMGs				
Vastus lateralis ($\times 10^{-2}$) [v]	4.31	5.88	6.39	6.50
Vastus medialis ($\times 10^{-2}$) [v]	3.35	5.22	7.09	9.64
Vastus femoris ($\times 10^{-2}$) [v]	0.854	1.38	1.69	3.92
Increased heart rate ratio				
From 0 to 1.5 min [%]	11.0	20.4	16.1	15.3
From 1.5 to 3 min [%]	17.9	28.0	32.0	38.1

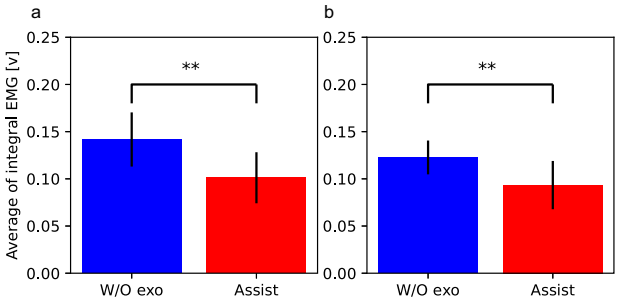


Fig. 2 | Average of integral EMG. a Participant 1, (**b**) Participant 2. Blue bars show the condition without an exoskeleton robot, and red bars show the assist condition with our proposed approach.

Subsequently, the assistive performance of the system was verified by evaluating the EMG amplitudes and increased heart rate to ascertain whether the muscle activities related to the knee movements and the heart rate were reduced with the use of the proposed approach in the multitask condition.

Table 1 shows the mean EMG amplitude of the right and left legs from three muscles and the increased heart rate ratio. The data were obtained under two conditions: 1) wearing the exoskeleton controlled by our proposed approach and 2) not wearing the exoskeleton. In each condition, the participants performed the three-minute movements. The EMG signals were extracted during the periods of rising from a crouched position in the squatting movement or climbing up the step. A total of 45 EMG recording trials, 30 for squat movements, and 15 for climb-up motions were obtained from the right leg, and the same number of recordings were also obtained from the left leg.

As illustrated in Table 1, the muscle activity of the assist condition with our approach is consistently lower than that of the condition without the exoskeleton. Figure 2 depicts the mean integral EMG of all the measured muscles under the two conditions. In this study, we set up the null hypothesis that there is no difference in physical burden between the condition of assistance by the exoskeleton robot using the proposed method and the normal condition of not wearing the exoskeleton robot for the same motion task. The Wilcoxon signed-rank test was applied to verify this, and significant differences were found between the proposed approach and the condition without the exoskeleton ($p < 0.01$) for both P1 and P2 participants. This comparison of muscle activity resulted in rejecting the null hypothesis and supporting the alternative hypothesis.

Table 1 also shows the results of the heart rate ratio, which was averaged over the period from 0 to 1.5 min and from 1.5 to 3 min. It can be observed that the increased heart rate ratio with our approach is lower than without the exoskeleton, apart from the P2 data observed from 0 to 1.5 min. These findings demonstrate that our system is capable of effectively assisting users in multitasking by observing the surrounding environment.

Assist sequence generation

Here, we present the predictive performance of the desired action command sequence. In this analysis, we limited our investigation to data from a single participant, as the other participant was only engaged in the assessment of assist control performance. The participant (P1) completed the set compound motion task nine times, with each trial lasting one minute. Eight of the nine trials were utilized for model training, with the remaining trial serving as the test.

The proposed method incorporates an approach to extracting image features for input, which are combined with kinematic information. In the experimental setup, there were two control commands: "active" and "free". Therefore, the estimation problem of the user's motion intention can be treated here as a binary classification problem. Figure 3a shows the assist action sequences generated by our proposed method and the joint torques. The joint torques are the output of the exoskeleton robot calculated from Eqs. (1) and (2) when using the predicted commands. The blue dashed line shows the ground truth sequences annotated by the participant for each leg. From these results, our approach can accurately generate the assist action sequence.

We compared our proposed approach with the widely used EMG-based movement intention estimation methods. First, we prepared an EMG-based model using the same transformer architecture as the proposed approach. This model does not use ResNet and was trained only on EMG and kinematic information. In addition, as suggested in the previous studies, we adopted a Support Vector Machine (SVM) with a radial basis function kernel and a Linear Discriminant Analysis (LDA) to classify the muscle activity as either "active" or "free" command. The SVM is often used in the upper limb[26,27] and lower limb[28] motion recognition with EMG. Similarly, the LDA is also often used in hand gesture recognition[29,30] and hand muscle recognition for impairment[31] with EMG. Here, we refer to our proposed model as VK-TR and the EMG-based model using the same transformer architecture as the proposed model as EK-T. The EMG-based models are referred to as EK-S when using SVM and EK-LD when using LDA. The input variables to the three EMG-based models were composed of the EMG signals and kinematic information.

Figure 3b shows that our proposed method, even without using bio-signals, was able to achieve comparable or better estimation performance with EMG-based methods, where bio-signals can be used to monitor the user's action command directly. However, measuring these signals needs careful preparation and calibration, which prevents us from adopting EMGs for assistive robot control for daily use. These results indicate that our proposed method allows us to control assistive robots with an accurate estimation of the user's motion intention without cumbersome preparation and calibration procedures.

Ablation testing for image features

We also conducted an ablation study to identify which part of our proposed pipeline contributed to accurately estimating movement intention. In the fully equipped pipeline, the VK-TR model, the image feature is extracted using ResNet[32] followed by Principal Component Analysis (PCA), and these features are plugged into the transformer along with kinematic information to generate assist command sequences.

First, we evaluated the contribution of ResNet to the prediction performance. To implement the model without ResNet, we use the same transformer as the proposed approach, but the image features are extracted in a straightforward way. Concretely, the normalized RGB data are converted to a one-dimensional vector. Then, lower-dimensional image features are extracted using only the PCA. Finally, these image features are plugged into the transformer along with the kinematic information. We refer to this model as VK-T.

Second, to evaluate the usefulness of the visual modality for the assist command classification, we adopted widely used classification methods: Support Vector Machine (SVM) with radial basis function kernel[33] and Logistic Regression (LR)[34]. The image features are extracted in the same way as the VK-T, and the image features are plugged into these classification

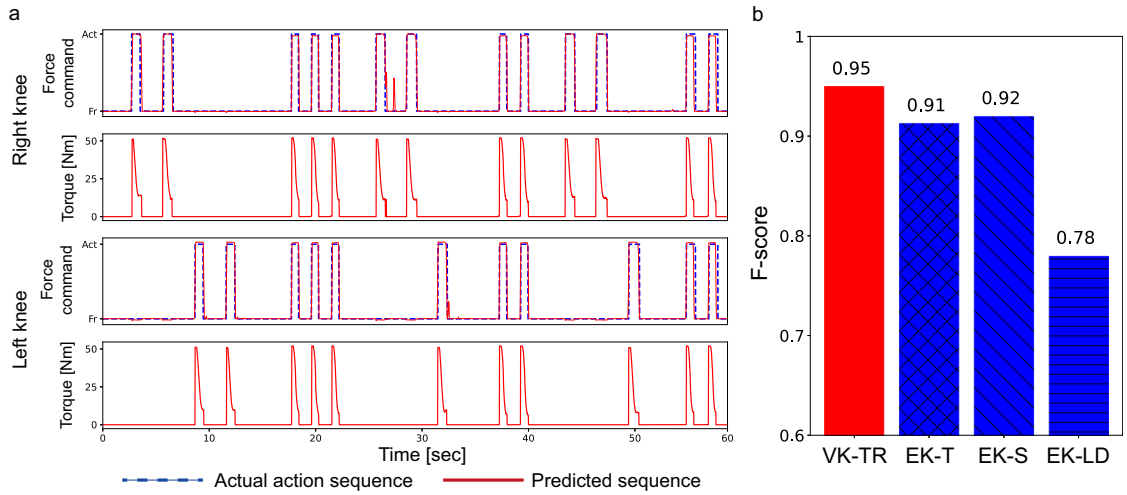
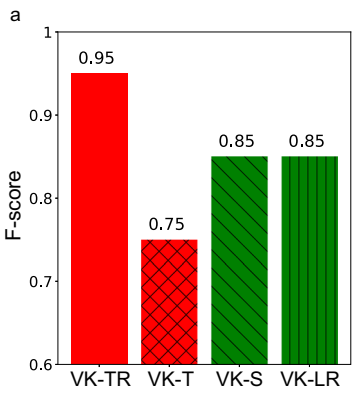


Fig. 3 | Assist action sequence from our approach and performance against EMG-based methods. a The red solid line shows the predicted sequence, and the blue dashed line shows the actual action sequence annotated by the user for each leg motion. Act and Fr of the pressure command indicate "active" and "free," respectively. The corresponding torque to the pressure command is calculated based on Eqs. (1) and (2). **b** F-score of the average of right and left leg motions when the estimations were treated as a binary classification. VK-TR refers to our approach. EK-T uses the same transformer architecture as the proposed approach with EMG and kinematic information, but does not use ResNet. EK-S and EK-LD use a Support Vector Machine with a radial basis function kernel and Linear Discriminant Analysis with EMG and kinematic information, respectively. Our proposed method was able to achieve comparable or better estimation performance with EMG-based methods, even without using any bio-signals. Our method does not need careful preparation and calibration, where these cumbersome procedures prevent us from adopting bio-signals for assistive robot control for daily use.

Fig. 4 | Comparison of performance among different model components. a F-score of the average of right and left leg motions when the estimations were treated as a binary classification.
b Components of each model. VK-TR refers to our approach. VK-T uses the same transformer architecture as the proposed approach but does not use ResNet. VK-S and VK-LR use a Support Vector Machine with a radial basis function kernel and Logistic Regression, respectively. Image features extracted from only PCA are combined with kinematic information and input to VK-T, VK-S, and VK-LR.



b

Components / Methods	Transformer	ResNet	PCA
VK-TR (Proposed)	✓	✓	✓
VK-T	✓		✓
VK-S & VK-LR			✓

methods along with the kinematic information. Here, we refer to these models as VK-S when using SVM and VK-LR when using logistic regression. These models, VK-T, VK-S, and VK-LR, are trained with the same training dataset as the fully equipped pipeline, VK-TR, and the dimensionality of the PCA is determined during model training. All trained models are applied to the same test data.

Figure 4a shows the F-score of the fully equipped pipeline, VK-TR, and other simplified implementations, VK-T, VK-S, and VK-LR. Here, for the method using a transformer, we considered the output as "active" when the predicted value was above the threshold of 0.5 and as "free" when below 0.5. Figure 4b summarizes the components of each method. VK-TR, composed of Transformer and ResNet, showed the highest performance among the four pipelines, while VK-T showed the lowest performance. This result indicates that the object recognition performance of the network to extract image features is an important factor for the proposed transformer-based assist command prediction. Furthermore, VK-S and VK-LR showed intermediate performance. This result suggests the effectiveness of using visual input for the assist command classification, although VK-S and VK-LR approaches cannot be used to predict continuous command trajectories, and VK-TR showed much better prediction performance than these two methods.

Generalization performance in novel situations

We conducted additional analyses to evaluate the performance of the proposed model in situations it had not encountered during training. To create unseen scenarios, we collected data on squatting behavior to pick up two objects with novel colors and shapes, as well as climbing behavior using two types of step platforms that also differed in color and shape. Each dataset includes one minute of motion for each task.

Figure 5 presents the F-scores, illustrating the novel scene recognition performance across different datasets. For the original model trained using our proposed approach, the results indicate high F-scores for squatting to pick up new objects, demonstrating good generalization, whereas the performance is lower for climbing steps with new stairs.

To further investigate the adaptation performance of the proposed model in novel situations with limited additional experience, we fine-tuned the originally proposed model using a small amount of data from the task of climbing steps with new stairs, where the original model had exhibited low

Model \ Data	Original motion	Squatting to pick up a new object A	Squatting to pick up a new object B	Climbing a step with new stair C	Climbing a step with new stair D
Original	0.95	0.92	0.87	0.42	0.088
Finetuned	0.94	0.89	0.85	0.95	0.95

Fig. 5 | Comparison of F-scores between the original motion task and tasks involving new environments across different models. The original motion refers to the task set used when constructing the proposed model. The tasks of squatting to pick up new objects (**A** and **B**) and climbing a step with new stairs (**C** and **D**) involve novel colors and shapes for generalization testing. The original model represents the performance of our proposed model, which experienced only the original motion during training. The finetuned model refers to the model obtained by fine-tuning the original model with a small amount of data from the step-climbing tasks with new stairs (**C** and **D**). Notably, when a separate model was trained from scratch using only the same limited data from climbing steps with new stairs, the F-scores for the two new stair types, (**C** and **D**), were 0.71 and 0.61, respectively, substantially lower than the scores of 0.95 and 0.95 achieved by the finetuned model. These results demonstrate that leveraging the pre-trained model significantly improves performance compared to training from scratch.

generalization performance; we refer to this as the finetuned model. As shown in Fig. 5, the F-scores of the finetuned model are high across all motions. Notably, when a separate model was trained from scratch using only the same limited data from climbing steps with new stairs, the F-scores for the two new stair types were 0.71 and 0.61, respectively, substantially lower than the scores of 0.95 and 0.95 achieved by the finetuned model. These results demonstrate that leveraging the pre-trained model significantly improves performance compared to training from scratch.

These results indicate that the proposed method exhibits a certain degree of generalization capability, even for actions and scenarios not included in the training data. For example, the model showed relatively strong generalization in the squatting task to pick up objects, despite having no prior exposure to this behavior during training. This suggests that the method holds promise for adapting to a broader range of unseen situations. However, we observed a decline in performance under specific conditions, particularly when the training data lacked diversity in colors or shapes. In the step-climbing task, recognition performance deteriorated for previously unseen colors and shapes, indicating that the model relies substantially on visual information. Nevertheless, we found that even in cases of limited generalization, fine-tuning with a small amount of additional data led to significant improvements in accuracy. These findings highlight both the adaptability of the proposed model and its potential for further refinement in handling diverse behavioral scenarios.

Discussion

EMG is commonly used to control exoskeleton robots based on user intentions. However, attaching EMG sensors and calibrating their interfaces is time-consuming and labor-intensive, making daily use impractical. In this study, we aim to generate assistive motions using a first-person camera and motion data, such as the user's joint angles and angular velocities. Achieving high performance without relying on cumbersome sensors like EMG is crucial for enhancing the practicality of the system.

Our developed multitask assistance control strategy, which uses a transformer to generate control commands for the exoskeleton robot according to the status of the user and the environment, was demonstrated on movement tasks consisting of squatting to pick up an object from the floor and climbing a step. This system was trained with labeled data, during which a participant freely labeled the period when he felt he needed assistance in a human-in-the-loop fashion. The assist performance of our proposed approach was evaluated in real-time control of the exoskeleton by measuring the EMGs of the muscles in the knee joint and the heart rate. The evaluation showed that the movement load was reduced in two participants compared to the condition without an exoskeleton robot.

One reason for reducing the physical load was the accuracy of the assist sequence generation, which was achieved through the appropriate use of human and environmental sensor data, as shown in Fig. 3a. This was also supported by the assist sequence generation evaluation, which showed comparable or higher performance than EMG-based methods, as shown in Fig. 3b. In addition, the result of ablation testing for image features showed that effectively combining visual image and kinematic information was important for high performance. This result also showed that our approach is highly capable of handling these sensors' data.

Another reason was the feature of our hardware, which could use two assist commands, "active" with high force and "free" with zero torque. We considered multitasking, which consisted of squatting down to pick up an object on the floor and climbing a step, and requiring participants to walk to each location to do them. Our system substantially dealt with three types of motions. No assistance is required when walking and moving toward decreasing potential energy, so a "free" series was generated. In other words, it is essential not only to provide assistance but also to turn off assistance so as not to impede human movement. Most lower limb exoskeletons are targeted to support single movements such as walking[35,36]. Some of them deal with running, stair ascent, and walking[37]. What they all have in common is the maximization of the assist effect, which is made possible by advances in the lightweight of the hardware. Therefore, their focus is on the constant actuation of an exoskeleton robot. On the other hand, our approach differs in that we address the generation of not only the presence of assist actions but also their absence in a more general situation involving several types of actions.

However, through generalization testing, we found that while the proposed method exhibits a certain level of adaptability to unknown situations, it still faces challenges in specific environments and actions. While improvements can be achieved by incorporating additional training data, our findings confirm that data diversity and training quality are equally essential. Considering these factors, we will explore further enhancements in generalization ability and training data diversity in our future study. Concretely, we plan to proceed with the following steps:

- Collect data for diverse scenarios: Expand the training dataset by incorporating a wider range of environments and actions to enhance the model's generalization ability.
- Apply data augmentation techniques: Use simulations and generative models to create diverse datasets from existing data and integrate them into the training process.
- Enhance model architecture: Explore more adaptive architectures, such as transfer learning and meta-learning, to further improve generalization performance.

- Broaden performance evaluation: Assess the generalization ability of the proposed method under various conditions, including different environments and a diverse set of users, including individuals with movement disorders, to better understand its adaptability across different actions and settings. In addition, applications to uneven terrain and sudden changes in movement intention are also important topics.

In recent years, vision-based approaches have attracted attention in various fields and have demonstrated excellent performance, especially in object recognition and obstacle detection. While a direct comparison with other vision-based methods may not align with the primary objective of this study, we are interested in understanding how assistive control performance varies with different vision models. In future research, we will explore comparisons with state-of-the-art vision methods to further evaluate their impact on our approach.

In addition, the proposed method may be affected by the degradation of visual information accuracy, particularly under challenging environmental conditions such as low light and visual obstructions. In future work, we will assess the impact of these factors on control performance. If performance deterioration is observed, we will explore strategies to complement visual information and integrate additional sensors (e.g., depth cameras and IMUs) to enhance adaptability to environmental changes.

Furthermore, while the transformer-based model demonstrates strong performance, its 'black box' nature raises concerns about interpretability. In future research, we will explore methods to visualize the model's behavior using attention maps to enhance the transparency of its decision-making process. This approach will help identify which aspects of the input data the model prioritizes and how it makes decisions, ultimately improving both its interpretability and reliability.

Methods

Participants

This study conducted the experiments with two participants (P1: male; age = 37 years; mass = 70 kg; height = 1.71 m, P2: male; age = 60 years; mass = 54 kg; height = 1.60 m) with no prior history of movement disorders, after obtaining informed consent from them. The human research ethics committee of RIKEN approved the experiment.

Motion task

To apply our approach in the concrete motion task, we considered blending three motion types: squatting down to pick up an object on the floor, climbing a step, and walking. The squatting and climbing are required to be conducted at different places (Fig. 1a). Therefore, the participant squats at one specific place, walks to another place, and climbs a step at another specific place. A step lift platform (Reebok International Ltd. USA) was used to raise and lower the steps, and the height was set to 25 cm. On the other hand, a white box was used as an object to pick up by squatting down.

In this study, we measured compound motions nine times, each trial lasting one minute. Eight of these nine trials were used in the model training, and one remaining was used for the test. In model training, 80% of the eight trials were used as the training dataset and 20% as the evaluation dataset. Subsequent data-dependent parameters, such as value normalization, are calculated based on this training dataset. During each trial, the participant was allowed to freely choose squats and climb a step, as well as the number and order of these movements. In the real-time exoskeleton robot control experiment, the compound motions were set to three minutes, and the number of squatting and climbing was 30 times, respectively, for consistency across participants. The climbing step was also set to 15 times for the right leg and 15 times for the left leg. The participants were asked to conduct these motions at a rhythm of 40 beats per minute.

Data measurement

To obtain the user state, angles and angular velocities of the right and left knee joints, and accelerations and velocities of the trunk in sagittal and coronal planes were used. Using the potentiometer of the lower limb exoskeleton robot system and IMU sensor (3DM-GX3-25, LOAD Micro-Strain Inc., USA) attached to the user's trunk, we simultaneously obtained the knee joints and the trunk motions. In addition to the user's state, an RGB camera mounted on the eyeglass (DITECT Co.Ltd, Japan) obtained a first-person view image of the surrounding environment. A schematic diagram showing the measurement state is shown in Fig. 1c, and these sensor signals were used as the input for our model.

In comparison with the proposed approach, we adopted the EMG-based method. To obtain the EMG signals, we measured six muscle activities from the vastus lateralis, the vastus medialis, and the rectus femoris in the right and left leg, respectively (Fig. 1d), simultaneously with the above sensor signals. We used Ag/AgCl bipolar surface EMG electrodes with a sampling rate of 1 kHz.

For learning the generative model of the action sequence, the log in which the participant simultaneously recorded whether or not the necessity of assistance was also obtained. We used a handy button interface to record the log. This interface outputs the pressure value for the PAM control while the button is pressed and zero when the button is not pressed. We prepared two button interfaces, one for the right knee and one for the left knee, and gave them to the right and left hands for operation. In this study, we set movements that increase potential energy as areas that require assistance, and other movements, including walking, as areas that do not need assistance, and asked the participant to manipulate them. These records were used as the output of the model.

In real-time exoskeleton robot control, we also measured six muscle activities from the same muscle location mentioned above to evaluate motion burden related to knee joints (Fig. 1d). The electrodes and sampling rate are the same as above. Each EMG signal was subjected to full-wave rectification and low-pass filtering at 10 Hz. This study measured sensor data other than EMG at 50 Hz.

Exoskeleton hardware

We take advantage of the lightweight hardware characteristics of our exoskeleton robots, instead of constant support. Our exoskeleton robot focuses on assisting knee joints[25] and is designed to be lightweight with a carbon fiber structure. This robot is attached by fixing it to each thigh and leg with a band. The skeleton is made of carbon resin material: the thigh part is about 604 g, the shank part is about 206 g, and the total of one leg is about 810 g. It features a highly responsive joint powered by a pneumatic artificial muscle (PAM) actuator provided by FESTO[38] (Fig. 6). Thus, our robot allows the user's movements to proceed smoothly without any hindrance, even when the actuator is not in operation.

The joint torque of the exoskeleton is generated by the PAM as follows:

$$\tau = rF \tag{1}$$

where r is the pulley radius, and F is the PAM force generated by the path contraction of the spiral fibers embedded in a pneumatic bladder. When a constant pressure P is applied to PAM, the relationship between the force and contraction rate of the PAM can be written as a 2nd-order polynomial. On the other hand, pressure increase and force generation at the same contraction rate have a linear relationship. From these, the quadratic model of PAM force can be written as follows:

$$F = \frac{(f_u - f_l)\,P + P_u f_l - P_l f_u}{P_u - P_l} \tag{2}$$

where P is the pressure of PAM. P_u and P_l are constant pressure of 0.7 MPa and 0.6 MPa. f_u is a quadratic force model when pressure is set to 0.7 MPa and f_l is a quadratic force model when pressure is set to 0.6 MPa:

$$\begin{aligned} f_u &= a_u\alpha^2 + b_u\alpha + c_u \\ f_l &= a_l\alpha^2 + b_l\alpha + c_l, \end{aligned}$$

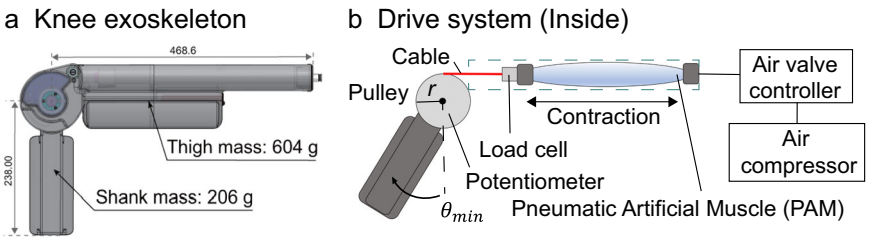


Fig. 6 | Exoskeleton robot system. a Carbon frame knee exoskeleton robot. **b** Joint torque driving system by pneumatic artificial muscle (PAM).

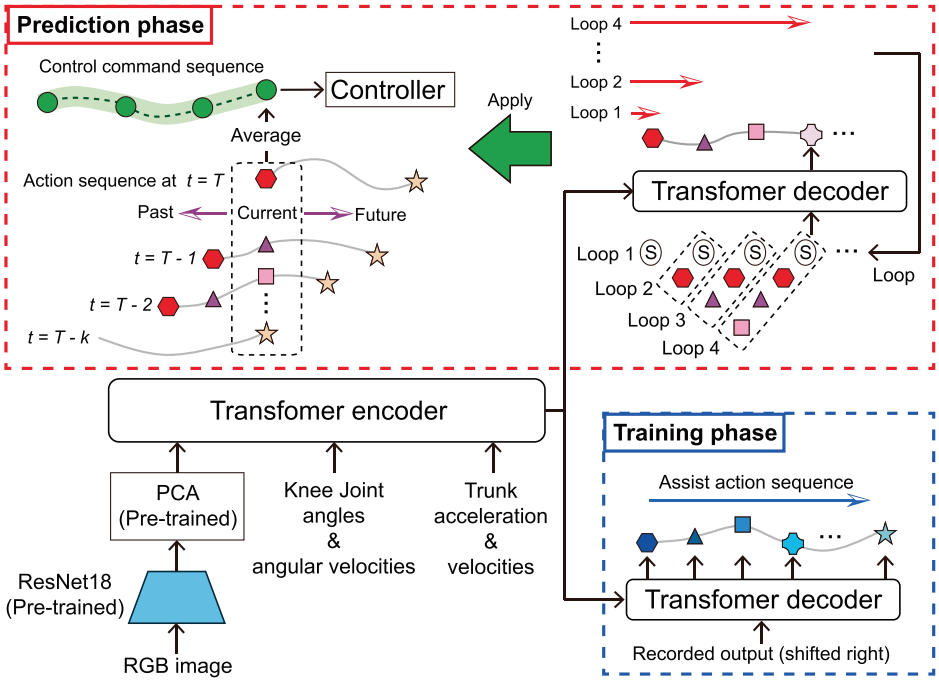


Fig. 7 | Architecture of assist action prediction. Our approach to generating assist action sequences mainly uses ResNet18 and transformer models. This model predicts sequences a few milliseconds ahead of sensor inputs that contain historical information from a few milliseconds before the current time. Predicted assist action sequence is averaged by the temporal ensemble and used for the final control command.

where α is the contraction rate of the PAM, which is varied according to the joint angle. The contraction rate is calculated by joint angle θ and minimum joint angle θ_{min}: $\alpha = \frac{r(\theta-\theta_{min})}{l_{pam}}$, where l_{pam} is the effective length of the PAM. Each parameter a_u, b_u, c_u, a_l, b_l and c_l is calculated by the calibration using the contraction rate and load cell. Note that Eq. (2) indicates that when the contraction rate of PAM is large, the force that can be generated decreases, while when PAM is at its natural length, a large force can be generated. By setting the pressure P to be applied, the joint torque is generated according to the above relational expression. In this study, the constant pressures $P = 0.7$ MPa and $P = 0$ were set and input for "active" with high force for the assist and "free" with zero pressure for the no-assist state, respectively. This action command is conducted independently in the right and left leg, and the 0.7 MPa is the maximum pressure that can be generated in our system. The control frequency of this exoskeleton robot was set to 50 Hz.

Prediction of assist action sequence

To predict the assist action sequence for driving the exoskeleton robot, we introduced a transformer framework[22] and ResNet18[32]. Figure 7 shows the architecture. We refer to our assist generative model from vision and kinematic information as VK-TR when comparing it with other methods. This model is trained using input and output datasets obtained from human movement measurements described above. The input state is defined by knee joint angle θ and angular velocity $\dot{\theta}$ and trunk acceleration $\ddot{\theta}$ and velocities in the sagittal and coronal plane and first-person view image features $\mathbf{C}$ as $\mathbf{x}=[\theta_{kr},\theta_{kl},\dot{\theta}_{kr},\dot{\theta}_{kl},\ddot{\theta}_{ts},\dot{\theta}_{ts},\ddot{\theta}_{tc},\dot{\theta}_{tc},\mathbf{C}]^{\top}$, where the subscripts kr, kl, ts, and tc represent the right and left knee joint, and trunk motion in sagittal and coronal plane, respectively. This $\mathbf{x}$ is normalized to a value between zero and one before being put into the model. The output $\mathbf{P}$ for model learning stores the right and left leg values obtained by normalizing the 0 and 0.7 MPa pressure value series for the PAMs control recorded by the participant using the button interface to zero and one.

The feature of the first-person view RGB image $\mathbf{C}$ is extracted by applying Principal Component Analysis (PCA) after obtaining a feature map from the global average pooling layer of ResNet18. The PCA and ResNet18 model was trained in advance using the training dataset. The ResNet18 was learned by inputting 100 × 100 RGB data and having it solve the problem of discriminating between 0 and 1 in the output data. Regarding the number of dimensions in PCA, we selected from the numbers that explain a contribution rate between 95% and 99% during the transformer model learning with the hyperparameter search described below. The number of components was 10, and the contribution rate was 97%. In our proposed model, ResNet18 and PCA are used to extract essential features $\mathbf{C}$ from first-person view images. ResNet18, known for its strong performance in image classification, was chosen for its robust feature extraction capabilities. To further enhance computational efficiency and reduce noise, PCA was applied to reduce the dimensionality of the extracted image features. By using an orthogonal transformation that maximizes data variance, PCA ensures effective feature compression. Directly inputting high-dimensional features into the Transformer can lead to excessive computational costs and instability during training. Therefore, dimensionality reduction promotes more efficient and stable learning.

To predict assist actions, we focus on two things. One is that human movements, as seen at a particular moment, have been initiated several tens to hundreds of milliseconds in the past[39]. Second, focusing on predicting

multiple steps, rather than one step at a time, reduces errors[40]. In other words, we predict the k-step assist action sequence $\mathbf{P}_{t+k}$ from $\mathbf{x}_{t-1}$, including the l-step history, where t represents the current time. In this study, we set l = 10 and k = 5, and the predictions are generated at 50 Hz. The assist action sequence predicted up to the k-step is averaged by the temporal ensemble and used for the final control command in Eqs. (1) and (2), as shown in Fig. 7.

The accuracy of the transformer model increases as the number of layers increases, but the inference time also increases. Therefore, we determined the number of layers through prior testing to strike a balance between this and the robot's real-time controllability, and we set two layers for both the encoder and decoder.

The hyperparameters of the ResNet18 and transformer models related to the learning were determined using Optuna (Preferred Networks, Inc., Japan), an automatic hyperparameter optimization framework. In this study, hyperparameter optimization was conducted over 170 trials using Optuna. From this optimization, the embedding dimension and the number of attention heads are set to 64 and 8, respectively. The optimal number of dimensions in the PCA mentioned above was also searched in this framework. Here, the ResNet18 and the transformer model ran on a PC equipped with an Intel(R) Xeon(R) CPU at 3.6 GHz and a double RTX 3090 GPU.

EMG-based movement intention estimation methods

We prepared three EMG-based methods to estimate movement intention. One is the EMG-based model using the same transformer architecture as the proposed model, and the others are traditional classification methods. To estimate "active" or "free" commands in time series classification using EMG signals, we adopted a support vector machine (SVM) with radial basis function kernel and linear discriminant analysis (LDA), which are widely used in EMG-based classification studies. For example, the SVM is used in the upper limb[26,27] and lower limb[28] motion recognition. Similarly, the LDA is also used in hand gesture recognition[29,30] and hand muscle recognition for impairment[31].

The input process for the EMG-based method is common to all three methods and is as follows. The processed six EMG signals e, knee joint angle θ, and angular velocity $\dot{\theta}$ were used as the current state $\mathbf{x}^{EK}=[e_1,\cdots,e_6,\theta_{kr},\theta_{kl},\dot{\theta}_{kr},\dot{\theta}_{kl}]^\top \in R^{10}$, where the subscripts kr and kl represent the right and left knee joint. The EMG signals were measured from around the knee joint muscles, as shown in Fig. 1d. The i-th processed EMG signal e_i is normalized as follows $e_i = \epsilon_i/\epsilon_i^{mvc}$, where the EMG signal ϵ_i is derived as full-wave rectified and low-pass filtered value of raw EMG signals. The ϵ_i^{mvc} indicates the maximum voluntary contraction (MVC) output. We used the EMG signals, angles, and angular velocities collected simultaneously when measuring the nine trials mentioned in the motion task section. The participant's MVC was observed before the actual data measurement. The EMG signals were measured by Ag/AgCl bipolar surface EMG electrodes with a sampling rate of 1 kHz, and the full-wave rectifying and low-pass filtering were processed. Then, they were down-sampled to 50 Hz to match the angular information.

In the EMG-based model using the transformer architecture, we predict the k-step assist action sequence $\mathbf{P}_{t+k}$ from $\mathbf{x}_{t-1}$, including the l-step history, where t represents the current time. The parameters l and k are the same as the proposed model, and the assist action sequence predicted up to the k-step is also averaged by the temporal ensemble and used for the final control command.

To estimate the current assist action sequence $\mathbf{P}_t$ in time series classification, the average value from the current state $\mathbf{x}_t^{EK}$ to m milliseconds past $\mathbf{x}_{t-m}^{EK}$ was used as the input $\bar{\mathbf{x}}^{EK}$ since the EMG signals are activated before the actual limb movements[39]. The m, which indicates how much past information to include, and the regularization parameters of SVM were determined with the hold-out using the same training and evaluation dataset. The learned models were applied to the same test dataset as we used for our proposed approach. Here, we refer to these models by EMG and kinematic information as EK-S when using SVM and EK-LD when using LDA.

Ablation testing

We prepare a method that removes ResNet from our proposed pipeline to examine the effect of image feature extraction methods on performance. In this method, we use the same transformer architecture as the proposed approach, but the image features are extracted without ResNet. We refer to this model as VK-T.

In addition, we implemented two classification methods: Support Vector Machine (SVM)[33] with radial basis function kernel and Logistic Regression[34] with $L2$ penalty. Here, we refer to these baseline methods by vision and kinematic information as VK-S when using SVM and VK-LR when using logistic regression.

In these three methods, VK-T, VK-S, and VK-LR, we extracted image features in a standard and straightforward manner: For the RGB data obtained at 100 × 100, each pixel value is normalized by dividing it by 255, and then, those values are made into a one-dimensional vector, after which features are extracted using the PCA. Regarding the number of dimensions in PCA, it was selected to have the same contribution rate of 97% as the proposed approach. The image feature $\mathbf{C}_b$ extracted in this way is used. In other words, the input state is $\mathbf{x}^{ablation}=[\theta_{kr},\theta_{kl},\dot{\theta}_{kr},\dot{\theta}_{kl},\ddot{\theta}_{ts},\dot{\theta}_{ts},\ddot{\theta}_{tc},\dot{\theta}_{tc},\mathbf{C}_b]^\top$, and the only difference from our approach is $\mathbf{C}_b$. In the VK-T, the state $\mathbf{x}^{ablation}$ is input to the transformer architecture. In the VK-S and VK-LR, the state is input to the SVM and Logistic Regression, respectively. The regularization parameters of the SVM and Logistic Regression models were determined using the hold-out method. These three models were learned using training and evaluation datasets, and the learned models were applied to the same test dataset.

Generalization testing

We collected data on the behavior of squatting to pick up two kinds of objects, A and B, which are new in color and shape, and on climbing behavior using two types of step platforms, C and D, which are also new in color and shape, to see how the proposed model can work in situations it has not experienced during training, as shown in the image of Fig. 5. In each A and B, squatting to pick up motions are conducted for one minute, and in each C and D, climbing step motions are performed for one minute. These data are used for generalization testing.

To further train the proposed model on new stair-climbing data, we use 40 s of newly acquired movements. In other words, the 40 s of data includes a climbing step motion with two kinds of stairs with different shapes and colors. We refer to this as a finetuned model. On the other hand, we also prepared a completely new model that was trained using only 40 s of data from climbing a step with new stairs, in which the data is the same as the additional training in the finetuned model. We refer to this as a separate model. In the finetuned model and the separate model, the hyperparameters are fixed the same as the original model.

Data analysis

During the three-minute exercise, EMGs were segmented between the participants' lowest and highest body positions during squatting and climbing a step. In other words, the state is from the lowest to the standing position in the squat motion, and the state is from when the participants place their one leg on the step lift platform to the time of finishing climbing in the step-up exercise. Therefore, 45 target movements that provide assistance for each leg are extracted, including squats and step lifts. In other words, same-name muscles (Vastus lateralis, Vastus medialis, and Rectus femoris) have a total of 90 time series data, including the left and right leg. For the 90 extracted data, we calculated the average for each muscle. To see the overall burden, we calculated the integral EMG (iEMG) in each squat and step-up motion. The iEMG is calculated using a trapezoidal approximation in each muscle of 90 time-series data. We then combined the iEMG of all muscles and calculated the average and standard error from the 90 data.

Statistics

We reported the means and standard errors calculated within the participant for the muscle burden of EMGs. We compared the assist condition

using our approach to the condition without an exoskeleton robot (normal exercise). These EMG data were not normally distributed, and we applied the Wilcoxon signed-rank, two-tailed test, and the significance level was 0.05.

Data availability

The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

Code availability

The underlying code for this study is not publicly available but may be made available to qualified researchers on reasonable request from the corresponding author.

Received: 1 October 2024; Accepted: 9 May 2025;
Published online: 11 June 2025

References

1. Meuleman, J., van Asseldonk, E., van Oort, G., Rietman, H. & van der Kooji, H. Lopes ii-design and evaluation of an admittance controlled gait training robot with shadow-leg approach. *IEEE Trans. Neural Syst. Rehabil. Eng.* **24**, 352–263 (2016).
2. Hartigan, C. et al. Mobility outcomes following five training sessions with a powered exoskeleton. *Top. Spinal Cord Injury Rehabilit.* **21**, 93–99 (2015).
3. Stauffer, Y. et al. The WalkTrainer–a new generation of walking reeducation device combining orthoses and muscle stimulation. *IEEE Trans. Neural Syst. Rehabil. Eng.* **17**, 38–45 (2009).
4. Riener, R., Lunenburger, L., Maier, I., Colombo, G. & Dietz, V. Locomotor training in subjects with sensori-motor deficits: An overview of the robotic gait orthosis lokomat. *J. Healthcare Eng.* **1**, 197–216 (2010).
5. Banala, S. K., Kim, S. H., Agrawal, S. K. & Scholz, J. P. Robot assisted gait training with active leg exoskeleton (alex). *IEEE Trans. Neural Syst. Rehabil. Eng.* **17**, 2–8 (2009).
6. Poggensee, K. L. & Gollinss, S. H. How adaptation, training, and customization contribute to benefits from exoskeleton assistance. *Sci. Rob.* **6**, eabf1078 (2021).
7. Zhang, J. et al. Human-in-the-loop optimization of exoskeleton assistance during walking. *Science* **356**, 1280–1284 (2017).
8. Ding, Y., Kim, M., Kuindersma, S. & Walsh, C. J. Human-in-the-loop optimization of hip assistance with a soft exosuit during walking. *Sci. Rob.* **3**, eaar5438 (2018).
9. Gordon, D. F. N., McGreavy, C., Christou, A. & Vijayakumar, S. Human-in-the-loop optimization of exoskeleton assistance via online simulation of metabolic cost. *IEEE Trans. Rob.* **38**, 1410–1429 (2022).
10. Ingraham, K. A., Remy, C. D. & Rouse, E. J. The role of user preference in the customized control of robotic exoskeletons. *Sci. Rob.* **7**, eabj3487 (2022).
11. Gleb, K. et al. Human-in-the-loop personalization of a bi-articular wearable robot's assistance for downhill walking. *IEEE Trans. Med. Rob. Bionics* **6**, 328–339 (2024).
12. Zhibo, J., Hong, H., Jianda, H. & Juanjuan, Z. A relationship model between optimized exoskeleton assistance and gait conditions improves multi-gait human-in-the-loop optimization performance. *IEEE Trans. Neural Syst. Rehabil. Eng.* **32**, 4304–4313 (2024).
13. Cheng, S. et al. STMI: stiffness estimation method based on semg-driven model for elbow joint. *IEEE Trans. Instrum. Meas.* **72**, 1–14 (2023).
14. Hayami, N. et al. Development and validation of a closed-loop functional electrical stimulation-based controller for gait rehabilitation using a finite state machine model. *IEEE Trans. Neural Syst. Rehabil. Eng.* **30**, 1642–1651 (2022).
15. Mungai, M. E. & Grizzle, J. W. Feedback control design for robust comfortable sit-to-motions of 3d lower-limb exoskeletons. *IEEE Access* **9**, 122–161 (2021).
16. Rajasekaran, V., Vinagre, M. & Aranda, J. Event-based control for sit-to-stand transition using a wearable exoskeleton. In: *International Conference on Rehabilitation Robotics (ICORR)*, 400–405 (IEEE, 2017).
17. Ao, D., Song, R. & Gao, J. Movement performance of human-robot cooperation control based on emg-driven hill-type and proportional models for an ankle power-assist exoskeleton robot. *IEEE Trans. Neural Syst. Rehabil. Eng.* **25**, 1125–1134 (2017).
18. Furukawa, J., Noda, T., Teramae, T. & Morimoto, J. Human movement modeling to detect biosignal sensor failures for myoelectric assistive robot control. *IEEE Trans. Rob.* **33**, 846–857 (2017).
19. Furukawa, J. & Morimoto, J. Composing an assistive control strategy based on linear bellman combination from estimated user's motor goal. *IEEE Rob. Autom. Lett.* **6**, 1051–1058 (2021).
20. Liu, D. X. et al. Vision-assisted autonomous lower-limb exoskeleton robot. *IEEE Trans. Syst. Man Cybern. Syst.* **51**, 3759–3770 (2021).
21. Lai, L., Huang, A. Z. & Gershman, S. J. Action chunking as policy compression. Preprint at https://osf.io/preprints/psyarxiv/z8yrv (2022).
22. Vaswani, A. et al. Attention is all you need. In: *Proceedings of the 31st international conference on neural information processing systems*, 6000-6010 (NIPS, 2017).
23. Lovanshi, M., Tiwari, V. & Jain, S. 3d skeleton-based human motion prediction using dynamic multi-scale spatiotemporal graph recurrent neural networks. *IEEE Trans. Emerging Top. Comput. Intell.* **8**, 164–174 (2024).
24. Liu, S. et al. Long short-term human motion prediction in human-robot co-carrying. In: *International conference on advanced robotics and mechatronics (ICARM)*, 815–820 (IEEE, 2023).
25. Furukawa, J., Okajima, S., An, Q., Nakamura, Y. & Morimoto, J. Selective assist strategy by using lightweight carbon frame exoskeleton robot. *IEEE Rob. Autom. Lett.* **7**, 3890–3897 (2022).
26. Osk, M. A. & Hu, H. Support vector machine-based classification scheme for myoelectric control applied to upper limb. *IEEE Trans. Biomed. Eng.* **55**, 1956–1965 (2008).
27. Zheng, X., Chen, W. & Cui, B. Multi-gradient surface electromyography (semg) movement feature recognition based on wavelet packet analysis and support vector machine (svm). In: *International conference on bioinformatics and biomedical engineering*, 1–4 (IEEE, 2011).
28. Ceseracciu, E. et al. Svm classification of locomotion modes using surface electromyography for applications in rehabilitation robotics. In: *19th international symposium in robot and human interactive communication*, 165–170 (IEEE, 2010).
29. Zhang, H., Zhao, Y., Yao, F., L. Xu, P. S. & Li, G. An adaptation strategy of using lda classifier for emg pattern recognition. In: *35th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, 4267–4270 (IEEE, 2013).
30. Botros, F. S., Phinyomark, A. & Scheme, E. J. Day-to-day stability of wrist emg for wearable-based hand gesture recognition. *IEEE Access* **10**, 125942–125954 (2022).
31. Adewuyi, A. A., Hargrove, L. J. & Kuiken, T. A. An analysis of intrinsic and extrinsic hand muscle emg for improved pattern recognition control. *IEEE Trans. Neural Syst. Rehabil. Eng.* **24**, 485–494 (2016).
32. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In: *IEEE conference on computer vision and pattern recognition (CVPR)*, 770–778 (IEEE, 2015).
33. Nakano, T. et al. Gaits classification of normal vs. patients by wireless gait sensor and support vector machine (svm) classifier. In: *IEEE/ACIS 15th international conference on computer and information science (ICIS)*, 1–6 (IEEE, 2016).
34. Slade, P., Kochenderfer, M. J., Delp, S. L. & Collins, S. H. Personalizing exoskeleton assistance while walking in the real world. *Nature* **610**, 277–282 (2022).

35. Nuckols, R. W. et al. Individualization of exosuit assistance based on measured muscle dynamics during versatile walking. *Sci. Rob.* **6**, eabj1362 (2021).
36. Peng, X., Acosta-Sojo, Y., Wu, M. I. & Stirling, L. Actuation timing perception of a powered ankle exoskeleton and its associated ankle angle changes during walking. *IEEE Trans. Neural Syst. Rehabil. Eng.* **30**, 869–877 (2022).
37. Ishmael, M. K., Archangeli, D. & Lenzi, T. A powered hip exoskeleton with high torque density for walking, running, and stair ascent. *IEEE ASME Trans. Mechatron.* **27**, 4561–4572 (2022).
38. FESTO. http://www.festo.com (2024).
39. Koike, Y. & Kawato, M. Estimating of dynamic joint torques and trajectory formation from surface electromyography signals using a neural network model. *Biol. Cybern.* **73**, 291–300 (1995).
40. Zhao, T. Z., Kumar, V., Levine, S. & Finn, C. Learning fine-grained bimanual manipulation with low-cost hardwarel. In: *Proceedings of robotics: science and systems* (*RSS*), https://doi.org/10.15607/RSS.2023.XIX.016 (2023).

Acknowledgements

We thank Mr. Akihide Inano for supporting the system development. This work was supported in part by JSPS KAKENHI JP25K01207 and JP23K24925. The funder played no role in study design, data collection, analysis, and interpretation of data, or the writing of this manuscript.

Author contributions

J.F. and J.M. conceived this study and designed the proposed approach. J.F. performed the experiments and analyzed the data. J.F. and J.M. wrote the main manuscript text, and J.F. prepared all the figures. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44182-025-00033-4.

Correspondence and requests for materials should be addressed to Jun-ichiro Furukawa.