

Toward Sensor Fusion Neuromuscular Interface for Continuous Finger Joint Angle Estimation via Deep Transfer Learning

Yun Chen¹, Graduate Student Member, IEEE, Xinyu Zhang, Graduate Student Member, IEEE, Hongsheng He, Senior Member, IEEE, Lina Pu², Member, IEEE, Wan Shou³, and Qiang Zhang⁴, Member, IEEE

Abstract—Accurate decoding of motor intent from biosignals is an important step toward intuitive upper-limb prosthetic-control interfaces. We propose a novel high-dimensional multimodal deep learning framework that fuses surface electromyography (sEMG) and B-mode ultrasound (US) images to estimate metacarpophalangeal and proximal interphalangeal joint angles continuously. The framework employs a shared Encoder–Decoder–Regression architecture integrating convolutional neural networks (CNNs), transposed convolutions, an action-conditioned multi-head cross-attention module (ATT) that uses the commanded action as a query, and long short-term memory (LSTM) layers to jointly capture spatiotemporal features from both modalities. To improve cross-subject generalization and reduce data requirements for new users, we introduce a transfer learning strategy with parameter freezing. Experiments on data from seven able-bodied subjects show that, compared with sEMG-only and US-only baselines, the fusion model reduces test local root mean square error (RMSE) by 2.187° (23.385%) and 0.890° (11.054%), and increases test local correlation (Pearson’s r) by 0.069 (10.02%) and 0.039 (5.48%) ($p < 0.05$), supporting the potential of multimodal fusion for future prosthetic-control interfaces. A preliminary validation on one amputee participant further supports the feasibility of applying the framework under a residual-limb sensing condition. Ablation studies further confirm that the full CNN+LSTM+ATT model achieves the best performance,

reducing test local RMSE by 0.933° (11.524%) and increasing test local correlation by 0.033 (4.56%) ($p < 0.05$). Furthermore, fine-tuning the pretrained model with only 25% of a new subject’s data yields performance comparable to full retraining, highlighting the framework’s data efficiency.

Index Terms—Surface electromyography, B-mode ultrasound, deep learning, transfer learning, finger motion estimation, multimodal fusion, prosthetic hand.

I. INTRODUCTION

FOREARM amputation has a significant impact on the daily lives and independence of millions of people worldwide, with transradial amputation being one of the most common forms of upper-limb loss [1]. Assistive prosthetic hands have been developed to help amputees regain hand function in everyday tasks, with accurate and reliable identification of user movement intent being key to controlling these devices [2].

Surface electromyography (sEMG) is a primary noninvasive tool for upper-limb rehabilitation and prosthetic control [3]. Early approaches relied on feature engineering and pattern recognition [4], but recent research has shifted toward end-to-end deep learning for kinematics estimation [5], [6]. Convolutional neural network (CNN)-based regression models, for example, have achieved high accuracy in wrist and finger-angle tracking, with correlation coefficients up to 0.9 [7]. Recent models incorporate lightweight temporal architectures and attention mechanisms to improve interpretability [8], [9]. However, sEMG is inherently influenced by surface measurement artifacts, such as electrode-related variability and inter-muscle crosstalk [10]. Although these effects can be reduced, they cannot be entirely eliminated in practice [11]. Furthermore, sEMG cannot directly measure deep muscle and tendon deformation, underscoring the value of ultrasound (US) sensing as a complementary modality [12].

In addition to sEMG, US sensing has received increasing attention as a noninvasive approach for intent decoding and joint-angle estimation. US enables the tracking of internal tissue movement and captures deep muscle and tendon deformations that are not detectable by sEMG [13], [14], [15], [16]. Recent research mainly employs either A-mode

Received 8 September 2025; revised 15 February 2026 and 24 May 2026; accepted 16 June 2026. Date of publication 22 June 2026; date of current version 3 July 2026. Associate Editor: Xiaogang Hu. This work was supported by the Startup Fund at The University of Alabama. (Corresponding author: Qiang Zhang.)

Yun Chen is with the Department of Mechanical Engineering, The University of Alabama, Tuscaloosa, AL 35487 USA.

Xinyu Zhang is with the Luddy School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN 47405 USA.

Hongsheng He and Lina Pu are with the Department of Computer Science, The University of Alabama, Tuscaloosa, AL 35487 USA.

Wan Shou is with the Department of Mechanical Engineering, University of Arkansas, Fayetteville, AR 72701 USA.

Qiang Zhang is with the Department of Mechanical Engineering and the Department of Chemical and Biological Engineering, The University of Alabama, Tuscaloosa, AL 35487 USA (e-mail: qiang.zhang@ua.edu).

Data is available on-line at <https://github.com/ashen-yun/Multimodal-sEMG-US-training-code>

This article has supplementary downloadable material available at <https://doi.org/10.1109/TNSRE.2026.3706311>, provided by the authors.

Digital Object Identifier 10.1109/TNSRE.2026.3706311

[17], [18], [19] or B-mode [20], [21] acquisition methods. Wearable A-mode armbands have demonstrated high-precision joint tracking, achieving a root mean square error (RMSE) of approximately 6° [22]. In comparison, B-mode imaging offers a two-dimensional representation of forearm muscle cross-sections and their movement, thus preserving more comprehensive spatial information than single-scanline A-mode. Prior studies have shown the effectiveness of B-mode imaging for metacarpophalangeal (MCP) angle estimation using both shallow neural networks [23] and CNN pipelines [12]. B-mode is selected in this study because continuous multi-joint finger tracking requires coordinated spatial context from multiple muscle compartments, which is frequently lost in single-scanline A-mode.

Reliance on a single sensing modality may reduce robustness during continuous movements. Measurement artifacts, such as electrode displacement and US probe shifts, can compromise accuracy. These limitations highlight the need to explore multimodal fusion strategies to address the limitations of individual modalities. Multimodal neuromuscular interfaces have incorporated supplementary modalities such as accelerometers (ACC) [24] and inertial measurement units (IMUs) [25], which provide kinematic and inertial information, mechanomyography (MMG) [26] and force myography (FMG) [27], which capture near-surface mechanical activity. While these modalities offer valuable information, they measure physical quantities distinct from the deep-tissue deformation targeted in this study. To advance a muscle-centric interface, the present work focuses on fusing sEMG, which reflects neuromuscular activation, with US sensing, which captures deep muscle and tendon deformation as well as spatial context.

Multimodal methods that combine sEMG and US signals have been applied to tasks such as upper-limb gesture classification and lower-limb continuous joint-torque estimation. For instance, Wei et al. proposed a multimodal multi-level attention network that fuses sEMG and A-mode US, achieving offline gesture-classification improvements of 14.31% and 3.80% relative to unimodal CNNs using only sEMG or only A-mode US [28]. Additionally, Zhang et al. integrated sEMG signals and B-mode US imaging into a Hill-type neuromuscular model for continuous ankle plantarflexion-torque estimation, reducing the normalized root mean square error (NRMSE) by 14.58% compared to the sEMG-only model and by 36.79% compared to the US-only model [29]. In the present study, B-mode US imaging is employed because continuous multi-joint finger tracking requires coordinated spatial context from multiple muscle compartments, which is often lost in single-scanline A-mode. Because sEMG reflects neuromuscular activation with high temporal resolution, whereas US provides tissue-structure and deformation cues, the two modalities offer complementary information. Specifically, US can reduce ambiguity when sEMG is affected by surface measurement artifacts, while sEMG can serve as a stable reference when US-derived features drift during cyclic or repetitive movements [29].

Despite these advances, previous multimodal studies have typically combined heterogeneous biosignals using low-dimensional fusion, extracting a limited set of handcrafted scalar features and integrating them through simple

concatenation or model-based approaches. For example, Xi et al. fused a small set of handcrafted scalar features extracted from sEMG and MMG in a state-space model for continuous multi-joint angle estimation [26], while Lu et al. combined ultrasound-derived muscle deformation and thickness measures with inertial measurement unit (IMU)-derived joint angles for Hill-type muscle-force estimation [25]. Although these approaches are effective for specific tasks, low-dimensional designs may not capture the rich spatiotemporal structure present in sEMG time-frequency patterns and US image sequences. Furthermore, most existing approaches are subject-specific, requiring substantial per-user data and limiting cross-subject generalization. The literature also lacks sEMG-US approaches tailored to continuous hand joint-angle estimation [30]. To address the above gaps in continuous hand-joint angle estimation and high-dimensional multimodal fusion, we designed a fusion architecture for sEMG STFT and B-mode US optical-flow inputs. CNN-based models can extract local time-frequency or spatial-motion features [12], [31], and CNN+LSTM structures can further model short-term dynamics in biosignal-based motion decoding [28], [32]. However, they do not explicitly select movement-relevant features for different finger movements or include a dedicated trainable input-remapping component for new-subject adaptation. Therefore, building on the CNN+LSTM baseline and motivated by attention-based feature selection and transfer learning [9], [33], we introduced learnable upsampling as a modality-specific input-remapping front end and action-conditioned feature selection to adapt to target-subject input distributions, emphasize movement-relevant information, and estimate continuous MCP/PIP joint angles from fused sEMG activation and US tissue-motion features.

The contributions are summarized as follows:

- We developed and evaluated an sEMG-B-mode US fusion neuromuscular interface for continuous multi-joint finger-angle estimation, using temporally aligned sEMG STFT spectra, US optical-flow features, and motion-capture-derived MCP/PIP labels to regress MCP/PIP angles of four fingers.
- We proposed a multimodal sensor-fusion framework that builds on a CNN+LSTM baseline by adding learnable upsampling as a transfer-adaptable input-remapping front end, action-conditioned feature selection, and parameter-freezing transfer learning for data-efficient cross-subject adaptation.
- We evaluated the framework through modality comparisons, ablation studies, statistical testing, and cross-subject transfer learning across seven able-bodied subjects, and further conducted preliminary validation on one amputee participant to assess feasibility under a residual-limb sensing condition.

This paper extends our preliminary study in [32] by expanding the experimental protocol to cover eleven finger actions and seven able-bodied subjects, improving feature extraction for sEMG and B-mode ultrasound images, and upgrading the CNN+LSTM design with action-conditioned cross-attention, alongside more results, preliminary amputee validation, and a comprehensive discussion.

Figure 1 provides a roadmap of the overall multimodal pipeline and its correspondence to the manuscript structure.

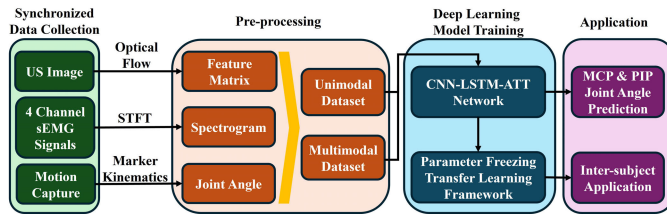


Fig. 1. The workflow of transfer learning-based multimodal fusion with parameter freezing.

TABLE I
TARGET MUSCLES AND CORRESPONDING
ACTIONS BY SEMG CHANNEL

Channel	Target Muscle	Target Movement
1	Extensor digitorum communis Extensor indicis proprius Extensor digiti minimi	Index, Little finger extension (Corr. Actions: 2, 7, 9, 11)
2	Extensor digitorum communis Extensor indicis proprius	Middle finger extension (Corr. Actions: 4, 9, 11)
3	Flexor digitorum superficialis Flexor digitorum profundus	Index, Middle, Little finger flexion (Corr. Actions: 1, 3, 6, 8, 10)
4	Flexor digitorum profundus Flexor carpi ulnaris	Ring finger flexion (Corr. Actions: 5, 10)

Note: Corr. denotes corresponding. The 11 actions are: 1: Index flexion, 2: Index extension, 3: Middle flexion, 4: Middle extension, 5: Ring flexion, 6: Little flexion, 7: Little extension, 8: Index+Middle flexion, 9: Index+Middle extension, 10: All-four-finger flexion, 11: All-four-finger extension.

Specifically, Sec. II describes the acquisition protocol and feature-label construction, Sec. III presents the model design and parameter-freezing transfer-learning strategy, Secs. IV and V report the able-bodied and amputee validation results, respectively, and Sec. VI discusses the main findings, limitations, and future directions.

II. EXPERIMENT PROTOCOL

A. Experimental Implementation

All experiments used the same synchronized acquisition platform, including four Trigno sEMG sensors (Delsys, Inc., USA), a diagnostic B-mode US system (ArtUs EXT-2H, Teleded, Ltd., Lithuania), and a Vicon Nexus 12.0 motion capture system. The sampling frequency was 1000 Hz for sEMG and 60 Hz for US and motion capture. All data streams were synchronized via Nexus 12.0 and a hardware trigger for the US images. All experimental protocols, including the able-bodied experiment and the preliminary amputee validation, were approved by the Institutional Review Board at the University of Alabama (Protocol ID: 24-01-7231), and informed consent was obtained from all participants. The setup photographs were cropped to minimize identifiable information.

1) **Able-Bodied Experiment:** For the main able-bodied experiment, we collected a multimodal dataset from seven healthy adult subjects, including synchronously recorded sEMG, B-mode US image sequences, and joint angle kinematics of the MCP and proximal interphalangeal (PIP) joints during the execution of eleven specified finger movements (details in Table I). Participants were instructed to perform natural finger flexion/extension, aiming to move the MCP and PIP joints together; however, the relative MCP/PIP motion amplitudes can vary across subjects. Figures 2(a)–(c) show the hand motion-capture marker placement, side-view marker geometry

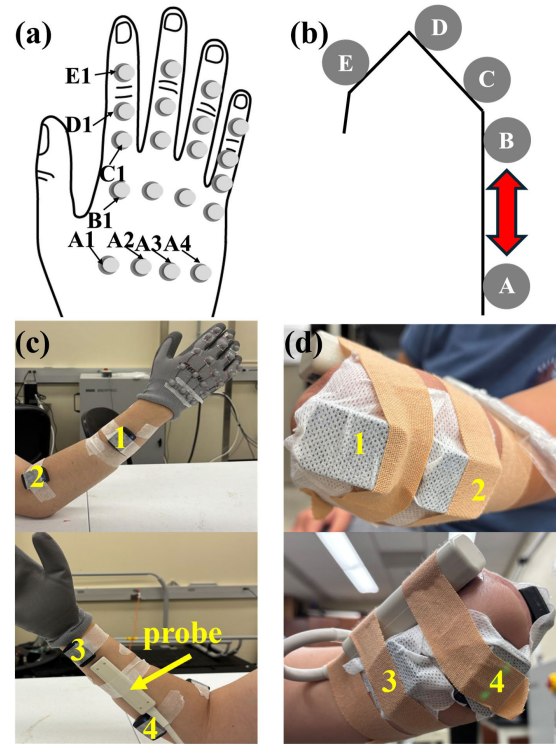


Fig. 2. Experimental setup. (a) Motion-capture marker placement on the hand. (b) Side-view marker geometry used for finger-joint angle calculation. A–E denote the marker points used to define the metacarpal and phalangeal segment vectors, and the red arrow indicates the flexion–extension direction. (c) Experiment setup of synchronized data acquisition. (d) Sensor placement on the amputee participant's residual limb. Labels 1–4 indicate the four sEMG sensors.

for joint-angle calculation, and synchronized data-acquisition setup, respectively. The sEMG sensors were strategically placed over forearm muscles that control individual finger articulations (see Table I). The US transducer was positioned parallel to the radius bone on the medial aspect of the forearm, providing an optimal viewing plane for longitudinal contraction and stretching of the muscle. Kinematic markers were applied to the proximal phalanges and metacarpals of the four fingers, excluding the thumb. The thumb was excluded because its motion is primarily caused by intrinsic hand muscles, not within the imaging field of our B-mode US probe [12], [34]. Specifically, Fig. 2(b) illustrates the side-view marker geometry used for MCP/PIP angle calculation, where marker points A–E define the metacarpal and phalangeal segment vectors and the red double arrow indicates the flexion–extension direction.

Each trial commenced with the hand in a relaxed posture. A single movement cycle comprised moving the target finger to its maximal flexion or extension and then returning it to the relaxed start position. Timing was governed by a 60 beats per minute metronome (one cycle every two seconds), and each cycle was repeated 13 times during a single trial. For analysis, each trial was temporally trimmed to a 20-second effective data segment. Data from seven subjects, each completing three trials, were collected. The collected data were then partitioned for model development: the first two trials of each subject were used for training and validation, and the third trial served as a hold-out set for the final evaluation of model generalization.

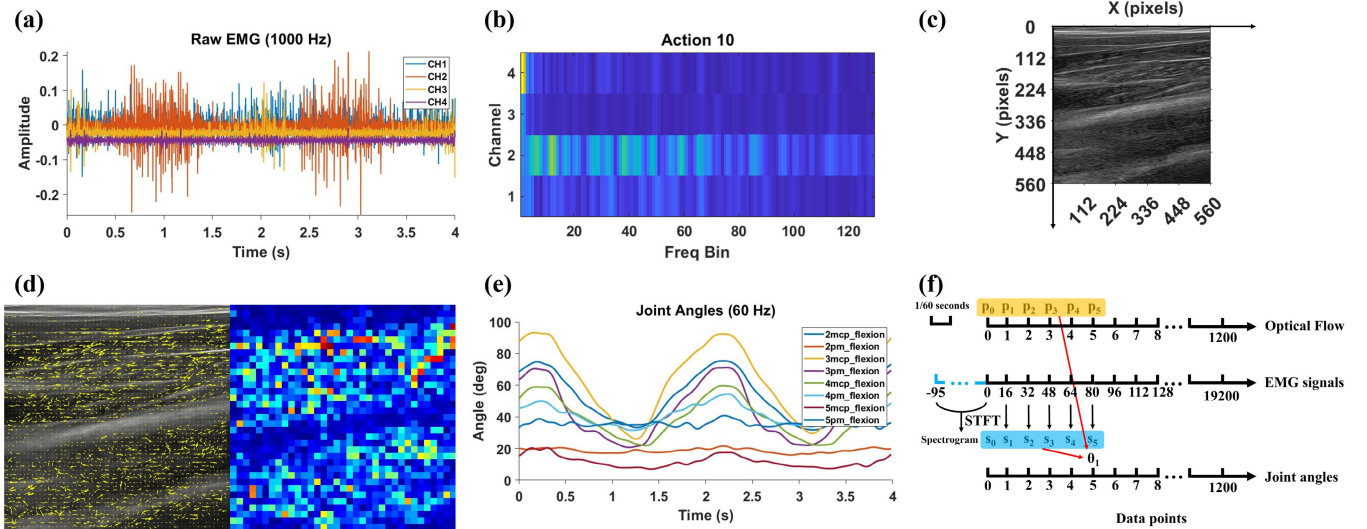


Fig. 3. (a) Raw sEMG signals. (b) One window of sEMG signals from action 10 after STFT. (c) Raw US image from one able-bodied subject. (d) Result of the optical flow method. (e) Joint angle for the eight finger joints. (f) The synchronization of different sensors with joint angles.

2) Preliminary Amputee Validation: To assess feasibility under a residual-limb sensing condition, we conducted an additional preliminary experiment with one adult female participant with unilateral left proximal transradial limb loss since early life and a residual forearm approximately 20% as long as the contralateral forearm (AT1). As shown in Fig. 2(d), the B-mode US probe was positioned on the residual limb over a region where repeatable soft-tissue deformation could be observed during attempted hand opening/closing. Because the residual-limb sensing area was limited and the sEMG sensors could not be placed over the same finger-specific muscle compartments as in the able-bodied experiment, four sEMG sensors were distributed over the dorsal and lateral aspects of the residual limb to maintain the same number of sEMG channels as in the able-bodied setup.

AT1 performed a mirrored bilateral movement protocol. Specifically, she was instructed to attempt Action 10 (all-four-finger flexion) and Action 11 (all-four-finger extension) with the residual limb while simultaneously performing the same movements with the contralateral intact hand. Motion-capture markers were placed on the intact hand using the same marker configuration as in the able-bodied experiment, and the intact-hand MCP and PIP joint-angle trajectories were used as synchronized kinematic labels for the attempted residual-limb movements. Reliable isolated single-finger attempted movements could not be elicited during preliminary testing; therefore, the amputee validation was restricted to attempted gross all-finger flexion and extension. For this preliminary validation, the regression targets consisted of two representative joint-angle trajectories, MCP and PIP, rather than the eight-joint output vector used in the able-bodied experiment.

B. Data Pre-Processing and Dataset Formation

1) sEMG Signals: To extract spectral features from the raw sEMG signals, as shown in Fig. 3(a), we first applied a series of preprocessing steps. The signal was band-pass filtered between 20 and 450 Hz using a fourth-order Butterworth filter, then full-wave rectified and detrended.

To match an integer multiple of the B-mode US sampling rate for our sensor-fusion framework, the signal was resampled from 1000 Hz down to 960 Hz. Unlike temporal features that only capture amplitude variations, spectral features reveal frequency-domain patterns of sEMG that are closely tied to muscle activation. Using the short-time Fourier transform (STFT) provides a compact time–frequency representation that preserves both amplitude dynamics and spectral information. For feature extraction, we applied the STFT to map the time series into the time–frequency domain. This was performed with a 96-sample (100 ms) sliding window, a 16-sample hop length, and zero-padding to 256 points; the visualized STFT result of a single window is shown in Fig. 3(b). The resulting STFT magnitude spectra from six consecutive windows were stacked to form a sequential feature tensor.

2) B-Mode US Signals: To address the speckle noise and high dimensionality inherent in B-mode US images (Fig. 3(c)) [35], [36], we designed a preprocessing pipeline to extract robust features. The pipeline first applied speckle-reducing anisotropic diffusion (SRAD) [37] to denoise each frame while preserving muscle fiber boundaries. Subsequently, the Farneback algorithm [38] computed dense optical flow between frames to extract muscle motion features. To reduce dimensionality, the flow field was divided into a 35×35 grid, and vectors within each cell were averaged, as shown in Fig. 3(d). This time series of flow features was then smoothed with a second-order zero-phase Butterworth low-pass filter. Finally, filtered features mapped from six consecutive frames are concatenated along the channel dimension, creating a $35 \times 35 \times 2 \times 6$ input tensor for the network.

3) Hand Motion Capture: Hand kinematics were recorded with high precision using an eight-camera three-dimensional motion capture system (Vicon Motion Systems, Oxford, UK). Twenty retro-reflective markers were placed at anatomical landmarks on the dorsum of the hand and forearm, as shown in Fig. 2(a). The raw 3D marker coordinates were first low-pass filtered at 6 Hz using a fourth-order Butterworth filter.

To compute the flexion–extension angles of the MCP and PIP joints, we followed the cross-product method

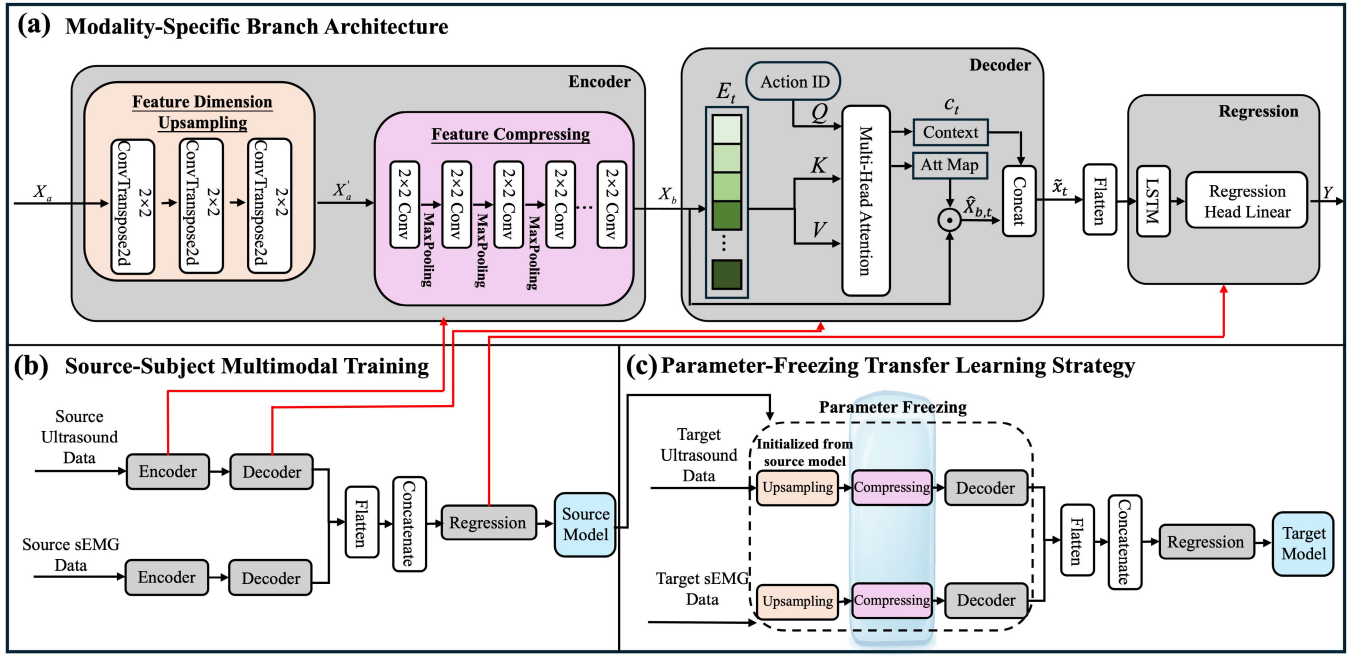


Fig. 4. Proposed model architecture and transfer-learning strategy. (a) Detailed modality-specific branch architecture for sEMG and B-mode US; (b) source-subject multimodal training; (c) parameter-freezing transfer learning for target-subject adaptation. Panels (b) and (c) use simplified blocks corresponding to the detailed modules in (a).

recommended by the International Society of Biomechanics [39]. Segment vectors were defined for the metacarpals and the proximal/middle phalanges using the marker coordinates. The angle between the vectors of adjacent segments was then calculated to determine the corresponding joint angle.

Finally, the calculated joint-angle data were smoothed using a 6 Hz moving-average filter, as shown in Fig. 3(e), and each joint angle was normalized based on the minimum and maximum values from the subject's training set for further analysis. All data modalities (B-mode US, sEMG, and finger joint angles) were temporally aligned for synchronous training, as illustrated in Fig. 3(f). For each aligned data point, the input sequence contained six consecutive time steps; each sEMG STFT feature was computed using a 96-sample causal window ending at the corresponding 60-Hz time point, i.e., the current sample and the preceding 95 samples at 960 Hz, and was paired with the synchronized US optical-flow feature map and motion-capture-derived joint-angle label.

III. SENSOR FUSION MODEL DESIGN

Hereafter, unless otherwise specified, “US” refers to the B-mode ultrasound imaging modality used in this study; prior work using A-mode ultrasound is explicitly denoted as A-mode US. Figure 4 provides a hierarchical overview of the proposed model architecture and transfer-learning strategy. Fig. 4(a) first details the modality-specific branch architecture, which consists of an encoder combining a learnable upsampling module (ConvTranspose2d) and a feature-compression module (Conv2d+MaxPool), a decoder implemented as an action-conditioned cross-attention block, and a regression module for joint-angle estimation. Fig. 4(b) then illustrates the source-subject multimodal training stage, where source ultrasound and sEMG inputs are processed by two modality-specific branches, their decoded features are flattened and

concatenated, and the resulting source model is trained. Fig. 4(c) shows the parameter-freezing transfer-learning stage, where the target-subject model is initialized from the source model and adapted using limited target-subject data. The following subsections describe each module and the transfer-learning strategy in detail.

A. Learnable Upsampling and Feature Compression

For both sEMG and US, we first upsample the input with three transposed-convolution layers and then compress features with stacked 2×2 convolutional layers and max-pooling; we refer to this two-stage pipeline as the encoder. Each convolution and transposed-convolution layer is followed by batch normalization and a LeakyReLU activation; LeakyReLU uses a small fixed negative slope to maintain gradient flow in the negative domain. Given an input sequence X_a , the upsampling block increases spatial resolution and produces X'_a . The feature-compressing block then aggregates local patterns and reduces spatial size to yield a compact representation X_b for downstream processing.

Transposed convolution provides a trainable and adaptive mechanism for restoring task-relevant spatial detail, improving downstream feature fidelity for joint-angle estimation [40]. For sEMG (STFT input), it refines channel–frequency patterns to provide finer structures for subsequent fusion; for US, it preserves fine speckle and fascicle-related textures before compression. Besides upsampling, the block also serves as a subject-adaptation front end in our transfer-learning scheme (Fig. 4(c)). After one transposed convolution layer, the output size is

$$O = S(I - 1) + K - 2P + O_p, \quad (1)$$

where O is the output size, I the input size, K the kernel size, S the stride, P the padding, and O_p the output padding.

In our setting ($K = 2, S = 2$ per layer, $P = 0, O_p = 0$), each layer exactly doubles the spatial resolution before feature compression by the stacked convolution+max-pooling block.

B. Action-Conditioned Cross-Attention

We implement an action-conditioned cross-attention block as the decoder in our encoder-decoder-regression backbone. For brevity, we denote the action-conditioned cross-attention decoder as ATT throughout the manuscript (e.g., CNN+ATT and CNN+LSTM+ATT). In our finger flexion-extension dataset with 11 commanded actions and short temporal windows ($T = 6$), ambiguity frequently arises: adjacent finger flexors (e.g., FDS/FDP compartments) tend to co-contract, yielding overlapping channel-frequency responses in the STFT sEMG, while ultrasound frames exhibit coupled tendon/fascicle motion and speckle patterns that vary with probe pose and gain. To disambiguate such cases, the commanded action label serves as a task-specific prior that steers feature selection toward structures most relevant to the intended movement (labels are available offline during evaluation; deployment without labels is discussed in *Limitations*). Tokens in the US branch correspond to locations on the image grid, whereas tokens in the sEMG branch correspond to elements on the channel-frequency lattice of the STFT representation. This decoder yields a head-averaged, interpretable weighting over spatial tokens (US) and channel-frequency tokens (sEMG) and improves angle estimation.

Let the encoder output be

$$X_b \in \mathbb{R}^{B \times T \times C_c \times H_c \times W_c}, \quad T = 6, \quad N \triangleq H_c W_c \quad (2)$$

We reshape the 2D grid into N tokens and project the channel dimension to d :

$$E = \text{proj}(\text{reshape}(X_b)) \in \mathbb{R}^{B \times T \times N \times d} \quad (3)$$

The commanded action $a \in \{1, \dots, 11\}$ is embedded once per sample and reused across its T steps:

$$z = \text{act_emb}(a) \in \mathbb{R}^{B \times d} \quad (4)$$

For each time step t , multi-head scaled dot-product cross-attention [41] is computed over the N tokens with the action embedding as Query and E_t as Key/Value. Writing the single-head form for clarity (with $d = h d_k$),

$$Q_t = z W_q, \quad K_t = E_t W_k, \quad V_t = E_t W_v \quad (5)$$

$$\alpha_t = \text{softmax}\left(\frac{Q_t K_t^\top}{\sqrt{d_k}}\right) \in \mathbb{R}^{B \times N}, \quad c_t = \alpha_t V_t \in \mathbb{R}^{B \times d} \quad (6)$$

where W_q, W_k, W_v are learned projections and d_k is the per-head dimension. We use head-averaged attention weights as returned by Multi-head Attention:

$$\bar{\alpha}_t = \frac{1}{h} \sum_{i=1}^h \alpha_t^{(i)}, \quad \text{att_map}_t = \text{reshape}(\bar{\alpha}_t) \in \mathbb{R}^{B \times 1 \times H_c \times W_c} \quad (7)$$

The encoder feature is then modulated by element-wise gating (shown for one branch; the sEMG branch uses its channel-frequency lattice for reshaping):

$$\hat{X}_{b,t} = X_{b,t} \odot \text{att_map}_t \quad (8)$$

TABLE II
ARCHITECTURE OF THE MULTIMODAL FULL MODEL

Layer Type	US	sEMG
— Branch-Specific Feature Extractors —		
Input Shape	$6 \times 2 \times 35 \times 35$	$6 \times 1 \times 129 \times 4$
ConvTranspose2d-1	$2 \rightarrow 8, k=2, s=2$	$1 \rightarrow 8, k=2, s=2$
ConvTranspose2d-2	$8 \rightarrow 3, k=2, s=2$	$8 \rightarrow 3, k=2, s=2$
ConvTranspose2d-3	$3 \rightarrow 1, k=2, s=2$	$3 \rightarrow 1, k=2, s=2$
Conv2d-1	$1 \rightarrow 3, k=2, s=1$	$1 \rightarrow 3, k=2, s=1$
Conv2d-2	$3 \rightarrow 8, k=2, s=1$	$3 \rightarrow 8, k=2, s=1$
Conv2d-3	$8 \rightarrow 16, k=2, s=1$	$8 \rightarrow 16, k=2, s=1$
Conv2d-4	$16 \rightarrow 32, k=2, s=1$	$16 \rightarrow 32, k=2, s=1$
Conv2d-5	$32 \rightarrow 64, k=2, s=1$	$32 \rightarrow 64, k=2, s=1$
Embedding Projection	in=64, out=64	in=64, out=64
Multi-Head Attention	heads=4, dim=64	heads=8, dim=64
Branch Output Dim.	18,560	6,208
— Fusion and Regression Head —		
Feature Concatenation	$18,560 + 6,208 \rightarrow 24,768$	
LSTM	in=24,768, h=64, l=1	
Regression Head	in=64, out=8	

Note: Unless otherwise noted, padding $p = 0$ and pooling 2×2 . US Conv2d-4 uses $p = 2$. sEMG Conv2d-3/4 use $p = 1$. sEMG Conv2d-1/2 use 2×1 pooling. ‘in’, ‘h’, and ‘l’ denote input size, hidden size, and layers, respectively.

The gated feature is flattened and concatenated with the attention context to form the per-step decoder output.

$$\tilde{x}_t = [\text{Flatten}(\hat{X}_{b,t}); c_t] \quad (9)$$

Stacking $\{\tilde{x}_t\}_{t=1}^T$ yields the sequence for downstream temporal modeling and regression

$$[\tilde{x}_1, \dots, \tilde{x}_T] \xrightarrow{\text{LSTM}} \mathbf{h}_T, \quad \hat{\mathbf{y}} = W_r \mathbf{h}_T + \mathbf{b}_r \in \mathbb{R}^8 \quad (10)$$

C. High-Dimensional Fusion, Temporal Modeling, and Regression Head

After extracting high-dimensional features from sEMG and US in the two parallel streams, the outputs of the decoders $X_{c,\text{sEMG}}$ and $X_{c,\text{US}}$ are flattened at each time step and concatenated to form a unified vector. This late-fusion representation exploits complementary features—US provides deep-tissue structural information while sEMG reflects surface neuromuscular activity—and has been shown to benefit joint-motion estimation [42], [43]. By aligning heterogeneous features in a shared latent space, the model captures more complete neuromuscular dynamics, improving robustness and task-specific accuracy. To model short-range temporal dependencies across the fused sequence, we use a shared one-layer LSTM (hidden size 64); an LSTM is a recurrent network with gated memory cells that selectively retain or forget information over time, making it suitable for our short window ($T = 6$) [44]. We take the last hidden state h_T as a compact temporal representation and feed it to a linear regression head to produce the joint-angle output $\hat{\mathbf{y}}$. Complete layer dimensions are summarized in Table II.

D. Transfer Learning Framework Based on Parameter Freezing

We adopt a source-to-target transfer strategy. A base multimodal model is first pretrained on a source-subject dataset

(Fig. 4(b)). To adapt the model to a new participant, only 25% of that subject's labeled data is used (Fig. 4(c)).

During fine-tuning, we proceed in three steps. First, the convolutional feature-compression modules in both branches are frozen to retain reusable feature extractors from the source domain and to limit the number of trainable parameters [45]. Second, the learnable upsampling blocks, implemented with transposed convolution, are reinitialized and retrained so that target-subject inputs are remapped to the operating scale expected by the frozen feature-compression modules. Third, the action-conditioned cross-attention decoder, the shared LSTM, and the regression head are fine-tuned to align sequence dynamics and the final readout with the target subject.

This targeted adaptation updates only the components most sensitive to inter-subject variation—namely, input scale and channel statistics in STFT-based feature and speckle/textural statistics and probe settings in ultrasound—while keeping the feature-compression modules fixed. As a result, the approach reduces overfitting and mitigates catastrophic forgetting [33]. Algorithm 1 summarizes the procedure.

Algorithm 1 Transfer Learning Design for Multimodal Model

Require: Pretrained US branch M_U , sEMG branch M_E

Require: Target fine-tune data $\{U_{ft}, E_{ft}, A_{ft}, Y_{ft}\}$ (25% subset)

Ensure: Fine-tuned branch parameters

```

1: // 1. Load pretrained branches
2:  $M_U \leftarrow \text{LoadPretrainedUSBranch}()$ 
3:  $M_E \leftarrow \text{LoadPretrainedEMGBranch}()$ 
4:
5: // 2. Freeze feature-compressing modules (Fig. 4c)
6: Freeze parameters in  $M_U.\text{conv}$  and  $M_E.\text{conv}$ 
7:
8: // 3. Reinitialize learnable upsampling modules (Fig. 4c)
9: Randomly reinitialize parameters in
    $M_U.\text{ConvTranspose2d}$  and  $M_E.\text{ConvTranspose2d}$ 
10:
11: // 4. Fine-tune on target data
12: for each epoch until convergence do
13:    $S_U \leftarrow M_U(U_{ft}, A_{ft})$  ▷ US branch features
14:    $S_E \leftarrow M_E(E_{ft}, A_{ft})$  ▷ sEMG branch features
15:    $Z \leftarrow \text{Concat}(S_U, S_E)$  ▷ Feature fusion
16:    $(H, \_) \leftarrow \text{LSTM}(Z)$  ▷ Temporal modeling
17:    $\hat{Y} \leftarrow \text{FC}_{\text{reg}}(H_{[:, -1, :]})$  ▷ Regression head
18:    $\mathcal{L} \leftarrow \text{MSE}(\hat{Y}, Y_{ft})$ 
19:   Backpropagate and update unfrozen modules
20: end for

```

Notation: $U_{ft} \in \mathbb{R}^{B \times T \times C_{\text{US}} \times H_{\text{US}} \times W_{\text{US}}}$ (ultrasound inputs); $E_{ft} \in \mathbb{R}^{B \times T \times C_{\text{sEMG}} \times H_{\text{sEMG}} \times W_{\text{sEMG}}}$ (sEMG-STFT inputs); $A_{ft} \in \{1, \dots, 11\}^B$ (action IDs); $Y_{ft} \in \mathbb{R}^{B \times 8}$ (joint-angle labels). S_U, S_E denote decoder outputs; Z is the concatenated feature; H is the LSTM output sequence; $\hat{Y}$ is the regression estimate.

E. Training Setting

All models were trained on a Windows 11 (64-bit) workstation with an Intel® Xeon Gold 6430 CPU (128 GB RAM)

and an NVIDIA RTX 4090 GPU (24 GB) using PyTorch 2.3.0. We used the Adam optimizer and mean-squared error (MSE) loss. Unless otherwise noted, training employed a batch size of 96 and a learning rate of 1×10^{-3} for up to 500 epochs. Early stopping with a patience of 15 epochs was applied based on the validation MSE, and the checkpoint with the lowest validation MSE was retained for evaluation.

F. Evaluation Metrics and Statistical Analysis

We report four metrics on the validation and test sets for three subjects: global RMSE, global Pearson correlation (r), and their action-conditioned local counterparts, as shown in Table III. The results for the other four subjects can be found in the supplementary file. All RMSE values are computed after reversing the per-joint normalization and are reported in degrees ($^\circ$). For the global metrics, we compute per-output RMSE and r over all time steps and then average them equally across the eight outputs to obtain one value per subject. We denote them as Val./Test RMSE and Val./Test r . For the local metrics, we evaluate each commanded action separately by restricting to its time indices and to the subset of output dimensions assigned to that action; per-action values are averaged across their relevant outputs and then across actions to yield one value per subject. We denote these as Val./Test RMSE_L and Val./Test r_L .

Normality across subjects for each model-metric pair was assessed using the Shapiro–Wilk test. When normality held, a repeated-measures analysis of variance (ANOVA) with different models as the within-subject factor (Subject as the blocking factor) was applied. If the ANOVA indicated a significant model effect, paired two-sided t -tests were conducted between models using raw p -values. A comparison was deemed significant at $p < 0.05$, and significant differences were indicated in Fig. 7. For the preliminary amputee validation, pairwise model comparisons were performed across five independent training runs using paired two-sided t -tests.

IV. RESULTS IN ABLE-BODIED PARTICIPANTS

In this section, we evaluate multiple unimodal and multimodal model configurations for modality comparison and ablation studies. For the modality comparison, we evaluate the full CNN+LSTM+ATT under sEMG, US, and Multimodal inputs, and we also evaluate the decoder-removed CNN+LSTM under the same three inputs. For the ablation study (Multimodal input), we evaluate CNN+LSTM, CNN+ATT (LSTM removed), and CNN (LSTM and decoder removed). All results in Table III are reported as mean $\pm$ std over five independent training runs.

A. Performance Comparison Across Single and Multimodal Models

As summarized across all seven subjects (Table III reports three representative subjects), under CNN+LSTM+ATT, the multimodal model outperformed both single-modality counterparts. Averaged across subjects, Test Local RMSE was lower than sEMG-only and US-only by 2.187° (23.385%) and 0.890° (11.054%), respectively ($p = 0.007$ and $p = 0.006$; Fig. 7(a)), while Test Local Correlation increased by 0.069 (10.02%) and 0.039 (5.48%) ($p = 0.005$ and $p = 0.004$; Fig. 7(b)).

TABLE III
EVALUATION RESULTS FOR SUBJECTS 1–3

	Val. RMSE (°)	Val. r	Test RMSE (°)	Test r	Val. RMSE _L (°)	Val. r_L	Test RMSE _L (°)	Test r_L
sEMG CNN+LSTM	9.633±0.000	0.784±0.002	8.748±0.221	0.842±0.009	6.644±5.315	0.775±0.193	8.748±0.332	0.705±0.008
sEMG CNN+LSTM+ATT	7.640±0.111	0.873±0.002	6.976±0.111	0.900±0.005	6.422±1.661	0.801±0.087	7.087±0.111	0.761±0.016
US CNN+LSTM	5.647±0.221	0.902±0.007	5.869±0.332	0.902±0.011	5.647±0.332	0.823±0.023	5.647±0.221	0.776±0.007
US CNN+LSTM+ATT	5.315±0.221	0.906±0.006	6.090±0.554	0.891±0.019	5.758±1.550	0.829±0.096	5.979±0.554	0.779±0.018
Multimodal CNN	6.865±0.221	0.883±0.009	7.419±0.332	0.897±0.008	6.201±0.221	0.784±0.007	7.197±0.332	0.755±0.013
Multimodal LSTM	6.188±0.034	0.899±0.001	6.302±0.071	0.923±0.002	5.484±0.049	0.798±0.005	6.651±0.102	0.779±0.003
Multimodal ATT	9.395±0.054	0.797±0.004	9.086±0.108	0.826±0.004	9.535±0.086	0.682±0.008	9.765±0.188	0.659±0.008
Multimodal CNN+LSTM	5.758±0.000	0.871±0.004	5.758±0.111	0.893±0.001	4.761±0.111	0.774±0.005	5.647±0.111	0.754±0.004
Multimodal CNN+ATT	6.754±0.000	0.867±0.001	6.422±0.000	0.898±0.001	5.979±0.000	0.782±0.007	6.422±0.111	0.766±0.003
Multimodal CNN+LSTM+ATT	4.983±0.443	0.918±0.003	5.315±0.111	0.943±0.003	4.540±0.111	0.836±0.004	5.315±0.221	0.815±0.009
	Val. RMSE (°)	Val. r	Test RMSE (°)	Test r	Val. RMSE _L (°)	Val. r_L	Test RMSE _L (°)	Test r_L
sEMG CNN+LSTM	8.686±0.000	0.865±0.001	11.536±0.136	0.779±0.007	9.500±0.136	0.622±0.009	12.079±0.136	0.644±0.013
sEMG CNN+LSTM+ATT	6.514±0.136	0.928±0.002	8.686±0.271	0.883±0.008	6.921±0.136	0.697±0.001	9.229±0.407	0.682±0.024
US CNN+LSTM	7.057±0.136	0.899±0.004	9.229±0.136	0.832±0.007	7.464±0.136	0.684±0.006	9.772±0.136	0.695±0.010
US CNN+LSTM+ATT	6.514±0.814	0.916±0.022	8.957±0.543	0.861±0.019	6.786±0.814	0.700±0.022	9.636±0.814	0.652±0.043
Multimodal CNN	7.057±0.136	0.907±0.007	8.279±0.136	0.879±0.006	7.736±0.136	0.672±0.010	9.093±0.136	0.734±0.008
Multimodal LSTM	6.611±0.174	0.918±0.005	8.373±0.124	0.880±0.004	7.212±0.206	0.687±0.006	9.025±0.153	0.712±0.006
Multimodal ATT	10.969±0.136	0.771±0.009	11.709±0.290	0.767±0.008	11.856±0.136	0.604±0.010	12.822±0.331	0.597±0.025
Multimodal CNN+LSTM	5.700±0.136	0.891±0.002	7.600±0.136	0.853±0.006	5.971±0.136	0.647±0.016	8.143±0.271	0.695±0.007
Multimodal CNN+ATT	7.057±0.000	0.891±0.001	8.550±0.000	0.862±0.001	7.329±0.000	0.674±0.003	8.957±0.000	0.694±0.006
Multimodal CNN+LSTM+ATT	5.293±0.000	0.949±0.000	7.193±0.136	0.912±0.004	5.429±0.000	0.733±0.006	7.464±0.136	0.734±0.008
	Val. RMSE (°)	Val. r	Test RMSE (°)	Test r	Val. RMSE _L (°)	Val. r_L	Test RMSE _L (°)	Test r_L
sEMG CNN+LSTM	9.961±0.640	0.749±0.040	11.057±0.366	0.698±0.030	9.504±0.640	0.639±0.013	11.606±0.457	0.601±0.008
sEMG CNN+LSTM+ATT	7.037±0.091	0.886±0.003	8.499±0.091	0.839±0.003	6.671±0.274	0.721±0.019	8.681±0.183	0.693±0.012
US CNN+LSTM	5.757±0.091	0.927±0.005	9.504±0.000	0.798±0.002	5.849±0.183	0.752±0.002	10.144±0.091	0.668±0.008
US CNN+LSTM+ATT	5.483±0.091	0.935±0.003	8.407±0.091	0.854±0.002	5.300±0.091	0.767±0.004	8.316±0.091	0.694±0.007
Multimodal CNN	6.488±0.091	0.900±0.002	8.864±0.091	0.824±0.005	6.488±0.091	0.758±0.007	9.413±0.183	0.720±0.004
Multimodal LSTM	6.471±0.102	0.905±0.003	8.637±0.112	0.832±0.006	6.304±0.107	0.740±0.006	8.994±0.137	0.687±0.004
Multimodal ATT	10.850±0.198	0.744±0.009	12.015±0.183	0.692±0.008	10.448±0.239	0.648±0.011	12.754±0.142	0.555±0.013
Multimodal CNN+LSTM	5.483±0.274	0.881±0.006	8.316±0.274	0.796±0.011	5.392±0.183	0.718±0.011	8.590±0.183	0.668±0.014
Multimodal CNN+ATT	6.305±0.000	0.884±0.001	8.407±0.000	0.829±0.002	6.123±0.000	0.719±0.001	8.499±0.091	0.704±0.006
Multimodal CNN+LSTM+ATT	4.752±0.091	0.953±0.003	7.219±0.091	0.890±0.005	4.478±0.091	0.790±0.005	6.945±0.091	0.748±0.011

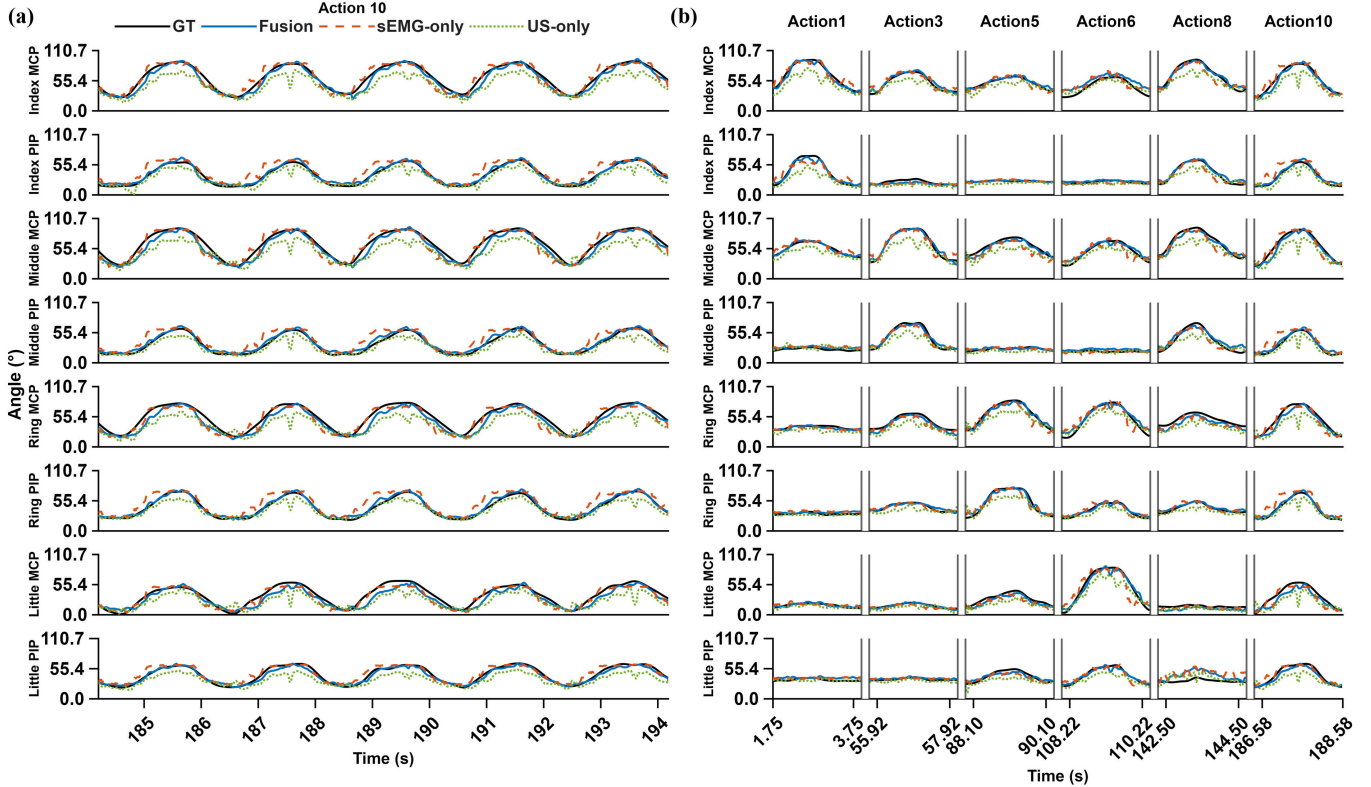


Fig. 5. Estimation results on held-out test data for unimodal and multimodal CNN+LSTM+ATT models. GT denotes the ground-truth joint angle. (a) Estimation results for Action 10 over a 10-s segment; (b) estimation results for representative 2-s segments from each of the six selected actions (Action 1/3/5/6/8/10; see Table I). Vertical dividing lines indicate transitions between selected action segments.

A repeated-measures ANOVA across models confirmed a significant model effect for all reported metrics ($p < 0.001$).

We further examined the CNN+LSTM backbone. In this setting, the multimodal model significantly outperformed the

TABLE IV

TEST SET RMSE_L (°) FOR CROSS-SUBJECT TRANSFER LEARNING

	Subj. 1	Subj. 2	Subj. 3	Subj. 4	Subj. 5	Subj. 6	Subj. 7
Subj. 1	—	8.550	8.499	8.205	11.095	7.981	14.393
Subj. 2	11.627	—	8.590	8.287	10.652	8.390	16.086
Subj. 3	9.412	8.143	—	7.713	9.764	7.572	13.969
Subj. 4	11.516	8.279	7.676	—	9.209	7.674	14.498
Subj. 5	8.637	7.736	7.859	7.713	—	8.186	13.969
Subj. 6	10.298	8.279	8.316	8.287	9.431	—	13.758
Subj. 7	8.305	8.279	8.407	8.205	10.874	7.674	—
Δ RMSE	4.651	0.814	1.279	1.395	2.552	-0.102	6.350

Note: '—' denotes not applicable (same-subject pair). Δ RMSE is the difference between the mean RMSE across six transfer experiments and the baseline RMSE obtained by training on the full dataset for each subject.

US-only model on Validation RMSE ($p = 0.022$), Test RMSE ($p = 0.004$), Test Correlation ($p = 0.045$), Validation Local RMSE ($p = 0.002$), and Test Local RMSE ($p = 0.016$). The multimodal CNN+LSTM also showed a lower mean Test Local RMSE than the sEMG-only CNN+LSTM; see Table III. Overall, across both CNN+LSTM+ATT and CNN+LSTM backbones, sensor fusion consistently reduced dynamic-movement error and improved correlation relative to single-modality models.

B. Ablation Experiment Results

We compared four multimodal variants—CNN, CNN+LSTM, CNN+ATT, and CNN+LSTM+ATT. Relative to CNN+LSTM, CNN+LSTM+ATT reduced Test Local RMSE by 0.933° (11.524%, $p = 0.007$; Fig. 7(c)) and increased Test Local Correlation by 0.033 (4.56%, $p = 0.030$). Relative to CNN+ATT, CNN+LSTM+ATT also improved Test Local RMSE and Test Local Correlation ($p = 0.002$ and $p = 0.0001$).

Both single-component variants reduced Test Local RMSE versus CNN ($p < 0.05$), while their gains in Test Local Correlation over CNN were not significant ($p > 0.05$). Differences between CNN+LSTM and CNN+ATT on the test set were not significant for either dynamic metric ($p > 0.05$). Validation metrics followed the same ordering; CNN+LSTM+ATT remained superior to the other variants ($p < 0.05$). Table III lists results for three representative subjects; all statistics reported in this section are averaged over seven subjects.

These results show that CNN+LSTM+ATT is the strongest multimodal configuration in this study, indicating that combining temporal modeling with action-conditioned attention yields the most effective fusion for joint-angle estimation.

C. Transfer Learning Comparisons

As shown in Table IV, we compared all cross-subject transfer-learning conditions with a baseline model trained from scratch on the full dataset of each target subject. After fine-tuning, Δ RMSE ranged from -0.102° to 6.350° , and was $\leq 2.6^\circ$ for most subjects, indicating that our method recovers most of the accuracy with substantially fewer labels. By contrast, applying the model without fine-tuning for cross-subject joint-angle estimation produced local RMSE values that exceeded 7° for all cross-subject pairs, highlighting the key role of the proposed transfer strategy in cross-subject generalization.

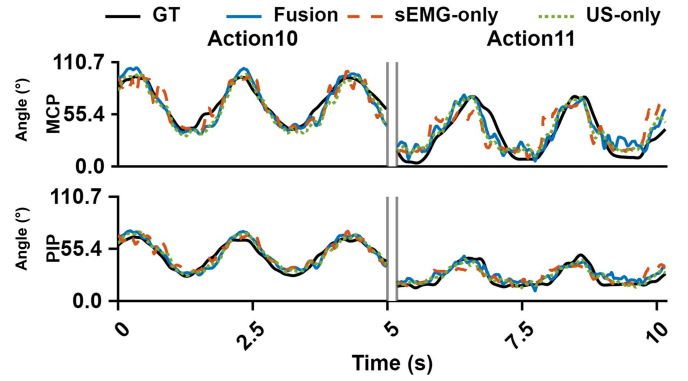


Fig. 6. Representative prediction trajectories on the amputee test set for attempted Action 10 (all-four-finger flexion) and Action 11 (all-four-finger extension). The two rows show MCP and PIP angle predictions. GT denotes the synchronized intact-hand kinematic label.

V. RESULTS IN THE AMPUTEE PARTICIPANT

Figure 6 shows representative prediction trajectories on the amputee test set. Results are reported as mean $\pm$ std over five independent training runs. The sEMG-only model achieved a mean RMSE of $10.445 \pm 0.448^\circ$ and a mean Pearson's r of 0.886 ± 0.008 . The US-only model reduced the mean RMSE to $8.760 \pm 0.448^\circ$ and increased r to 0.921 ± 0.007 , while Multimodal achieved a comparable mean RMSE of $8.802 \pm 0.550^\circ$ and r of 0.921 ± 0.008 . Both US-only and Multimodal significantly improved RMSE and correlation compared with sEMG-only ($p < 0.05$), whereas no significant difference was observed between US-only and Multimodal ($p > 0.05$). Relative to sEMG-only, Multimodal reduced the mean RMSE by 1.643° (15.73%) and increased the mean correlation by 0.035 (4.00%). As shown in Fig. 6, US-only and Multimodal closely tracked the major MCP and PIP flexion-extension trajectories, while sEMG-only showed larger deviations.

VI. DISCUSSION

A. Why Multimodal Fusion Works: Mechanism and Evidence

Figure 5 provides a qualitative comparison of representative estimation trajectories for the proposed fusion model and unimodal baselines. In Fig. 5(a), a long continuous segment from Action 10 is shown, where the fusion model estimations remain stable over time and exhibit no apparent temporal drift relative to the ground truth. Fig. 5(b) further presents representative short-window segments from six selected actions, with each segment covering one typical movement cycle. Overall, the fusion model better preserves the cycle-to-cycle waveform consistency across joints compared with unimodal sEMG-only and US-only results. Its estimation curve closely matches the ground truth, consistent with the quantitative gains in RMSE and correlation reported in Sec. IV (e.g., global test NRMSE = 0.068; see Table V).

Specifically, the single-modality models exhibit clear, opposing weaknesses. The sEMG-only model, while responsive to motion dynamics, suffers from noise and signal crosstalk, leading to inaccurate magnitude estimates (e.g., at the little finger PIP joint). Conversely, the US-only model provides a stable estimate but attenuates peaks and can exhibit temporal lag, as seen at the middle finger PIP joint.

TABLE V
COMPARISON OF CONTINUOUS FINGER JOINT ANGLE ESTIMATION METHODS

Reference	Sensors	DL model	Recognition type	Results type	(N)RMSE	r
[31]	sEMG (5 ch)	TF2AngleNet (CNN on time/freq)	MCP (5 fingers)	NRMSE; r	0.095	0.947
[46]	sEMG (8 ch)	Deep Regression Forest+ANN	MCP (5 fingers)	RMSE; r	7.322°	0.844
[9]	sEMG (14 ch)	Encoder–Decoder+GRU+Attention	MCP+PIP+DIP	NRMSE; r	0.124	0.600
[12]	Ultrasound, B-mode (1 probe)	CNN (mod. VGG16)+SVM	MCP (4 fingers)	RMSE	7.35°	–
[23]	Ultrasound, B-mode (1 probe)	ANN	MCP (5 fingers)	NRMSE; r	0.15	0.89
[47]	sEMG (8)+FMG (8)	MMMFNet (GRU+CSI+Attention)	MCP (5 fingers)	RMSE	6.225°	–
[27]	sEMG (8)+FMG (8)	BioCNN-T (CNN+Transformer)	MCP (5 fingers, 6 actions)	NRMSE	0.046	–
[24]	sEMG (12)+ACC (12×3-axis)	Fusionformer (CNN–Transformer)+MW-ADDA	MCP+PIP+DIP	RMSE; r	3.76°	0.939
Ours	sEMG (4)+Ultrasound, B-mode (1 probe)	Multimodal CNN+LSTM+ATT	MCP+PIP (4 fingers, 8 joints)	NRMSE; RMSE; r	0.068; 7.036°	0.905

Note: – indicates the metric was not reported in the corresponding reference. **Abbreviations:** MCP = metacarpophalangeal; PIP = proximal interphalangeal; DIP = distal interphalangeal; FMG = force myography; ACC = accelerometer; ANN = artificial neural network; GRU = gated recurrent unit; SVM = support vector machine; VGG16 = VGG-16 network; CSI = cross-stream interaction; MW-ADDA = Modality-Wise Adversarial Discriminative Domain Adaptation.

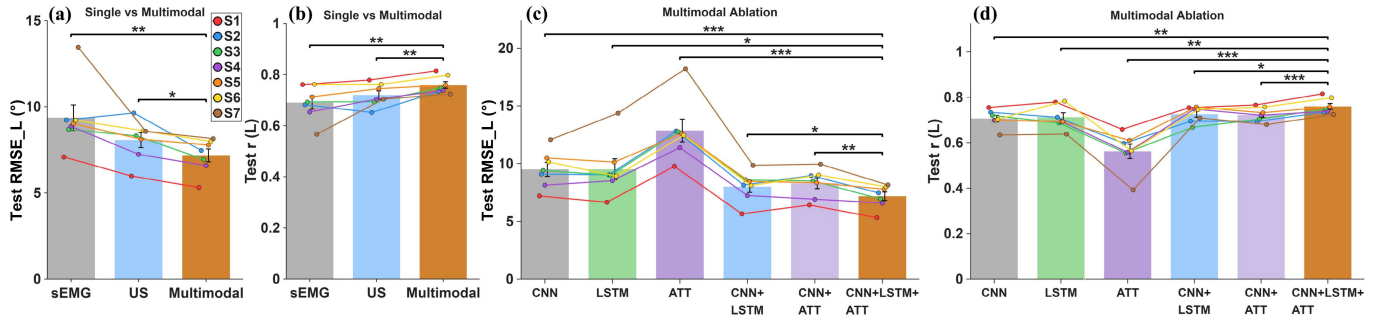


Fig. 7. Local test performance. (a–b) Unimodal (sEMG, US) vs Multimodal under CNN+LSTM+ATT—(a) RMSE, (b) r ; (c–d) Multimodal ablations (CNN, CNN+LSTM, CNN+ATT, CNN+LSTM+ATT)—(c) RMSE, (d) r . *Asterisks indicate statistically significant differences based on paired two-sided t -tests following repeated-measures ANOVA (* $p < 0.05$, ** $p < 0.01$, and *** $p < 0.001$).

The proposed fusion framework therefore benefits from modality complementarity rather than simply from adding another sensor. In the able-bodied experiment, the sEMG sensors were placed over finger-related flexor and extensor muscle groups, providing distributed activation patterns across actions, whereas US captured internal tissue deformation and spatial motion cues. The fusion model combines these two sources of information to reduce the ambiguity of either single modality, leading to faster dynamic response and more accurate magnitude estimation.

B. Fusion Feasibility in the Amputee Participant

The amputee experiment provides a preliminary feasibility assessment under a residual-limb sensing condition. In this participant, Multimodal and US-only achieved comparable performance, and both were significantly better than sEMG-only. This pattern suggests that B-mode US captured the dominant movement-related soft-tissue deformation in the residual limb, while the incremental information provided by sEMG was limited in this setup.

This result is consistent with the sensor-placement differences between the two experiments. In the able-bodied experiment, the four sEMG sensors were placed over finger-related muscle regions, allowing different actions to produce distinct channel-level activation patterns. In the amputee experiment, the short residual forearm limited the available sensing area, and the sEMG sensors could not be placed over the same finger-specific muscle compartments. Thus, sEMG likely captured more global residual-muscle activation, whereas US directly measured the tissue deformation associated with the mirrored flexion–extension movements. Under

this ultrasound-dominant condition, the Multimodal model maintained performance close to US-only and improved over sEMG-only, indicating that the framework can operate under a constrained residual-limb setup, although the fusion gain depends on the amount of independent sEMG information available.

C. Comparison With Existing Methods

Compared with sEMG-only estimators, our model achieves lower error and higher correlation on the global test set (NRMSE = 0.068, RMSE = 7.036°, r = 0.905) than the sEMG-only approach in [46] (7.322°, 0.844), and it compares favorably with the sEMG-only multi-joint estimator in [9] reported in normalized error (NRMSE = 0.124, r = 0.60). These observations are consistent with the design rationale that B-mode ultrasound provides deep-tissue kinematic cues complementary to sEMG, while the encoder–decoder with action-conditioned cross-attention and an LSTM captures coupled spatial–temporal dependencies.

Against ultrasound-only pipelines trained on B-mode images, our global RMSE in degrees is comparable to the B-mode ultrasound method in [12] (7.35°), suggesting that fusing surface activation with internal tissue motion may reduce error under comparable settings.

Relative to multimodal fusion baselines, some multimodal systems—the sEMG+FMG system in [47] and the sEMG+ACC system in [24]—report lower RMSEs (6.225° and 3.76°, respectively), but these results are obtained with substantially heavier instrumentation—typically 8 sEMG + 8 FMG sensors or 12 sEMG + 12 ACC modules. In contrast, our configuration uses only 4 sEMG channels and

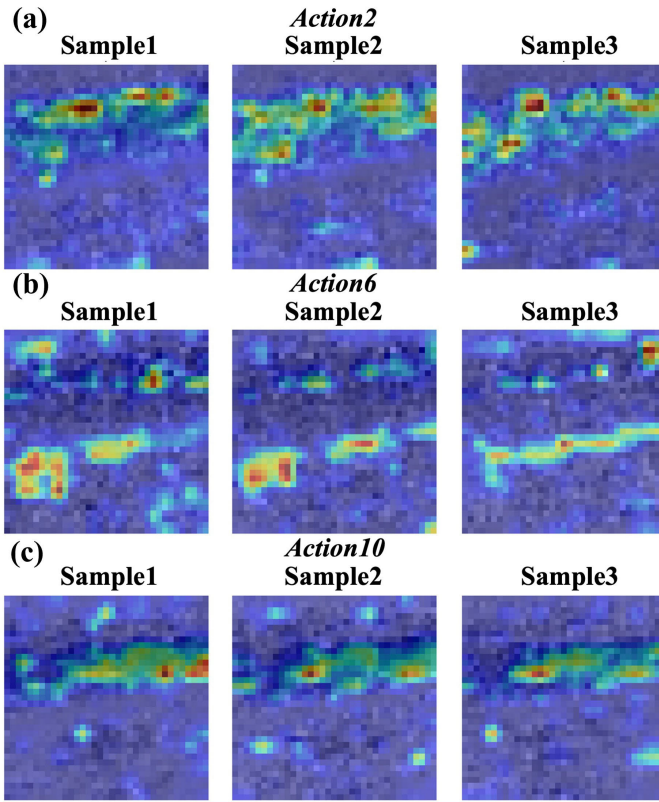


Fig. 8. US-branch attention maps. (a) Action 2 (index extension), (b) Action 6 (little-finger flexion), and (c) Action 10 (all-finger flexion); warmer colors denote higher normalized attention.

a single B-mode probe while achieving comparable accuracy and correlation, yielding a favorable balance between performance and hardware complexity for continuous finger-joint angle estimation.

D. Interpretable Attention for B-Mode US

The limited interpretability of deep learning models is a common problem in their application. Some studies have addressed this issue. For example, Lee et al. visualized the attention weights of different sEMG signal channels to explain how their model focuses on sensor channels that are highly related to specific finger movements [9]. The model in this study is based on multimodal inputs of sEMG and B-mode US, with separate attention modules designed for each modality. Unlike research that focuses on a single electrical signal modality, this section analyzes the spatial attention mechanism for the B-mode US modality. In our implementation, the spatial attention maps correspond to the head-averaged softmax weights from the US-branch attention (att_map in Eq. (7)), upsampled to the original US frame size and min-max normalized to [0, 1]; they gate encoder features (Eq. (8)) and are concatenated with the context c_t before temporal modeling (Eq. (9)).

Figure 8 indicates that the attention module is action-specific, with its spatial attention distribution adapting to different action commands. Specifically, when performing Action 2 in Fig. 8(a) and Action 10 in Fig. 8(c), the attention maps focus on muscle tissues at different depths, suggesting that the model can capture morphological changes in different muscle groups as they contract or stretch. During the execution

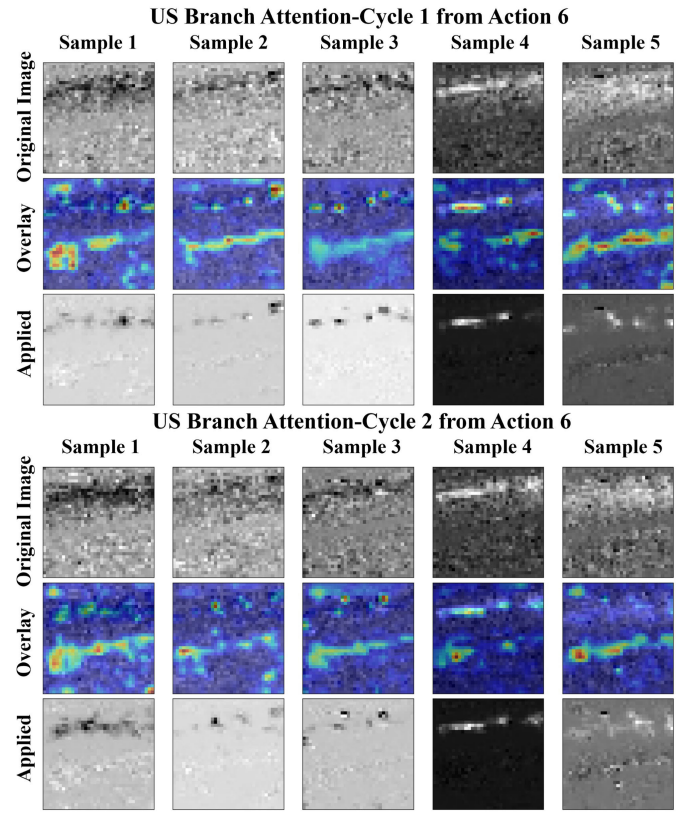


Fig. 9. US-branch attention visualization for Action 6 (Little-finger flexion): Cycle 1–2 with Samples 1–5 sampled every 0.3 s (± 0.6 s around Sample 3).

of Action 6 in Fig. 8(b), part of the attention concentrates on hyperechoic boundaries, potentially corresponding to bone or fascia. This suggests that for this specific task, the displacement of rigid contours caused by muscle activity may be used by the model as an informative distinguishing feature. Thus, the model is not limited to analyzing single muscle activity signals but can integrate deep tissue structure information provided by US images and adaptively extract task-relevant dynamic and structural features.

Figure 9 further supports the stability and periodic nature of this attention pattern. Over two consecutive movement cycles for Action 6, the attention distribution follows a reproducible “disperse–focus–disperse” evolution aligned with movement phases. For example, at Sample 3 in both cycles (near maximal flexion), the attention maps are similar in location and intensity. From the “Attention Overlay” results, it can be inferred that the model periodically tracks a horizontal hyperechoic structure (which could be bone or fascia). This reproducible pattern supports robust, non-transient attention dynamics based on deep-structure changes rather than momentary artifacts.

In summary, the interpretability provided by this attention mechanism enhances transparency by supporting the physiological plausibility behind the model’s decisions, which is important for building trust and moving toward practical applications.

E. Real-Time Feasibility and Latency Analysis

To assess real-time feasibility, we tested the inference latency on a portable device Jetson Nano 2GB with fixed

input shapes and a batch size of one. We report both the mean latency and the 99th-percentile latency (p99), where p99 indicates that 99% of runs finish within the reported time. Using TensorRT with FP16 precision, the mean end-to-end inference latency was 114.6 ms with a p99 of 120.8 ms, corresponding to an output update rate of 8.70 Hz.

For EMG preprocessing (filtering + STFT), the mean latency was 19.6 ms with a p99 of 20.8 ms. For ultrasound feature compression (downsampling the optical flow field from 560×560 to 35×35 with normalization and stacking), the mean latency was 18.4 ms with a p99 of 19.5 ms. For the dense optical flow algorithm applied on raw 560×560 pixels images, the mean runtime was approximately 880 ms per frame pair, which is the most time-consuming step of our pipeline.

F. Limitations and Future Work

Although advances have been made in multimodal fusion and interpretability, several limitations persist that hinder practical deployment.

The proposed action-conditioned cross-attention mechanism (Eqs. (5)–(6)) currently depends on prior action labels and does not yet include a real-time intent recognizer. For future online deployment, integration of an online motion intent classifier is essential, as demonstrated by recent studies [48], [49]. Additionally, redesigning queries to be derived directly from streaming measurements would remove the reliance on offline annotation.

In terms of computational efficiency, the model achieves an inference rate of approximately 8.7 Hz on a Jetson Nano 2GB with pre-processed inputs. However, optical flow-based feature extraction from ultrasound images remains a bottleneck on the current device due to hardware limitations. Accelerating this preprocessing step is critical for real-time application and may be addressed through hardware acceleration. Recent studies have demonstrated millisecond-level optical flow latency on edge devices [50], [51].

Beyond algorithmic speed, system-level stability in real-time scenarios has not been systematically quantified. Critical safety determinants such as continuous throughput, memory footprint, and latency jitter caused by thermal constraints must be evaluated within the closed control loop [52]. In addition, we did not systematically evaluate cross-session (day-to-day) repeatability, where electrode/probe re-positioning, changes in probe pressure/angle or imaging settings, and daily physiological variations may introduce distribution shifts and affect performance. Although a preliminary amputee validation was included, it involved a single participant under a constrained residual-limb sensing condition; therefore, larger amputee cohorts with varied residual-limb lengths, amputation histories, and sensor-placement conditions are needed to evaluate the generalizability of the sensor-fusion framework.

To tackle these challenges, future work will prioritize online integration and deployability. The development of a complete online pipeline is planned to evaluate true sensor-to-actuation latency, with particular emphasis on accelerated ultrasound feature extraction. In parallel, lightweight online intent recognition will be implemented to enable label-free operation. Comprehensive system-level analyses of resource usage and robustness will also be conducted in future closed-loop feasibility experiments with amputee participants.

VII. CONCLUSION

We have presented a high-dimensional multimodal deep-learning framework that fuses sEMG and B-mode US signals to achieve continuous finger joint-angle estimation toward upper-limb prosthetic-control interfaces. Central to our approach is a shared encoder–decoder architecture augmented with multi-head cross-attention and LSTM modules, which jointly capture the rich spatiotemporal dynamics inherent in both biosignal modalities. To address the challenge of cross-subject variability and minimize new-user data requirements, we introduced a transfer learning strategy with selective parameter freezing, enabling rapid adaptation to unseen subjects with only 25% of their data.

Extensive experiments on seven able-bodied subjects demonstrate that our sensor-fusion model significantly outperforms single-modality baselines, achieving a global test RMSE of 7.036° and Pearson correlation of 0.905. Ablation studies confirm that integrating both attention and temporal modeling yields the largest gains in dynamic movement estimation. Notably, fine-tuning with only a quarter of a target subject's data recovers performance comparable to full-data retraining. A preliminary amputee validation further showed that the proposed framework can operate under a residual-limb sensing condition, with Multimodal achieving performance comparable to US-only and better than sEMG-only.

REFERENCES

- [1] V. Mendez, F. Iberite, S. Shokur, and S. Micera, "Current solutions and future trends for robotic prosthetic hands," *Annu. Rev. Control, Robot., Auto. Syst.*, vol. 4, no. 1, pp. 595–627, May 2021.
- [2] A. T. Takele and D. S. Negash, "sEMG-based control of prosthesis hand using LSTM classification and SMC," *IEEE Access*, vol. 13, pp. 127068–127085, 2025.
- [3] Z. Wei, Z.-Q. Zhang, and S. Q. Xie, "Continuous motion intention prediction using sEMG for upper-limb rehabilitation: A systematic review of model-based and model-free approaches," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 32, pp. 1487–1504, 2024.
- [4] Y. Xue, X. Ji, D. Zhou, J. Li, and Z. Ju, "sEMG-based human in-hand motion recognition using nonlinear time series analysis and random forest," *IEEE Access*, vol. 7, pp. 176448–176457, 2019.
- [5] Y. Xu et al., "A powered prosthetic hand with vision system for enhancing the anthropathic grasp," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 33, pp. 1827–1840, 2025.
- [6] S. Tam, M. Boukadoum, A. Campeau-Lecours, and B. Gosselin, "Intuitive real-time control strategy for high-density myoelectric hand prosthesis using deep and transfer learning," *Sci. Rep.*, vol. 11, no. 1, p. 11275, May 2021.
- [7] Z. Qin, S. Stapornchaisit, Z. He, N. Yoshimura, and Y. Koike, "Multi-joint angles estimation of forearm motion using a regression model," *Frontiers Neurobot.*, vol. 15, Aug. 2021, Art. no. 685961.
- [8] J. Shin, A. S. M. Miah, S. Konnai, I. Takahashi, and K. Hirooka, "Hand gesture recognition using sEMG signals with a multi-stream time-varying feature enhancement approach," *Sci. Rep.*, vol. 14, no. 1, p. 22061, Sep. 2024.
- [9] H. Lee, D. Kim, and Y.-L. Park, "Explainable deep learning model for EMG-based finger angle estimation using attention," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 1877–1886, 2022.
- [10] Z. Liao, A. Golparvar, M. J. Bathaei, F. A. Cardoso, and C. M. Boutry, "Wearable and implantable devices for continuous monitoring of muscle physiological activity: A review," *Adv. Sci.*, vol. 13, no. 15, Mar. 2026, Art. no. e09934, doi: 10.1002/adv.202509934.
- [11] L. Mesin, "Crosstalk in surface electromyogram: Literature review and some insights," *Phys. Eng. Sci. Med.*, vol. 43, no. 2, pp. 481–492, Jun. 2020.
- [12] K. Bimbraw, C. J. Nycz, M. Schueler, Z. Zhang, and H. K. Zhang, "Simultaneous estimation of hand configurations and finger joint angles using forearm ultrasound," *IEEE Trans. Med. Robot. Bionics*, vol. 5, no. 1, pp. 120–132, Feb. 2023.

- [13] J. Wijntjes and N. van Alfen, "Muscle ultrasound: Present state and future opportunities," *Muscle Nerve*, vol. 63, no. 4, pp. 455–466, Oct. 2020.
- [14] F. Sarto, J. Spörri, D. P. Fitze, J. I. Quinlan, M. V. Narici, and M. V. Franchi, "Implementing ultrasound imaging for the assessment of muscle and tendon properties in elite sports: Practical aspects, methodological considerations and future directions," *Sports Med.*, vol. 51, no. 6, pp. 1151–1170, Jun. 2021.
- [15] C. M. Moran and A. J. W. Thomson, "Preclinical ultrasound imaging—A review of techniques and imaging applications," *Frontiers Phys.*, vol. 8, p. 124, May 2020.
- [16] X. Yang, C. Castellini, D. Farina, and H. Liu, "Ultrasound as a neurobotic interface: A review," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 54, no. 6, pp. 3534–3546, Jun. 2024.
- [17] X. Yang, X. Sun, D. Zhou, Y. Li, and H. Liu, "Towards wearable A-mode ultrasound sensing for real-time finger motion recognition," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 6, pp. 1199–1208, Jun. 2018.
- [18] Y. Jin et al., "Estimation of joint torque in dynamic activities using wearable A-mode ultrasound," *Nature Commun.*, vol. 15, no. 1, p. 5756, Jul. 2024.
- [19] Q. Zengyu et al., "A simultaneous gesture classification and force estimation strategy based on wearable A-mode ultrasound and cascade model," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 2301–2311, 2022.
- [20] A. T. Kamatham, M. Alzamani, A. Dockum, S. Sikdar, and B. Mukherjee, "Sparse sonomyography-based estimation of isometric force: A comparison of methods and features," *IEEE Trans. Med. Robot. Bionics*, vol. 4, no. 3, pp. 821–829, Aug. 2022.
- [21] Q. Zhang, N. Fragnito, X. Bao, and N. Sharma, "A deep learning method to predict ankle joint moment during walking at different speeds with ultrasound imaging: A framework for assistive devices control," *Wearable Technol.*, vol. 3, p. e20, Sep. 2022.
- [22] B. G. Sgambato et al., "High performance wearable ultrasound as a human-machine interface for wrist and hand kinematic tracking," *IEEE Trans. Biomed. Eng.*, vol. 71, no. 2, pp. 484–493, Feb. 2024.
- [23] Y. Huang, X. Yang, Y. Li, D. Zhou, K. He, and H. Liu, "Ultrasound-based sensing models for finger motion classification," *IEEE J. Biomed. Health Informat.*, vol. 22, no. 5, pp. 1395–1405, Sep. 2018.
- [24] Y. Zhang and W. D. Wang, "A modality-wise multi-modal transfer learning and dual stream CNN-transformer hybrid network for ACC-sEMG-based continuous estimation of finger joint angles," *Biomed. Signal Process. Control*, vol. 112, Feb. 2026, Art. no. 108445.
- [25] Y. Lu et al., "Investigation of a wearable piezoelectric-IMU multi-modal sensing system for real-time muscle force estimation," *Smart Mater. Struct.*, vol. 32, no. 6, Jun. 2023, Art. no. 065013.
- [26] X. Xi, C. Yang, S. M. Miran, Y.-B. Zhao, S. Lin, and Z. Luo, "SEMG-MMG state-space model for the continuous estimation of multijoint angle," *Complexity*, vol. 2020, pp. 1–12, Feb. 2020.
- [27] Z. Chen, H. Wang, H. Chen, and T. Wei, "Continuous motion finger joint angle estimation utilizing hybrid sEMG-FMG modality driven transformer-based deep learning model," *Biomed. Signal Process. Control*, vol. 85, Aug. 2023, Art. no. 105030. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1746809423004639>
- [28] S. Wei, Y. Zhang, and H. Liu, "A multimodal multilevel converged attention network for hand gesture recognition with hybrid sEMG and A-mode ultrasound sensing," *IEEE Trans. Cybern.*, vol. 53, no. 12, pp. 7723–7734, Dec. 2023.
- [29] Q. Zhang, N. Fragnito, J. R. Franz, and N. Sharma, "Fused ultrasound and electromyography-driven neuromuscular model to improve plantarflexion moment prediction across walking speeds," *J. Neuroeng. Rehabil.*, vol. 19, no. 1, p. 86, Aug. 2022.
- [30] S. Hwang et al., "A multimodal fatigue detection system using sEMG and IMU signals with a hybrid CNN-LSTM-attention model," *Sensors*, vol. 25, no. 11, p. 3309, May 2025.
- [31] H. Jiang et al., "TF2AngleNet: Continuous finger joint angle estimation based on multidimensional time-frequency features of sEMG signals," *Biomed. Signal Process. Control*, vol. 107, Sep. 2025, Art. no. 107833.
- [32] Y. Chen, X. Zhang, H. Li, H. He, W. Shou, and Q. Zhang, "Towards neurobotic interface for finger joint angle estimation: A multi-stage CNN-LSTM network with transfer learning," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2025, pp. 4409–4416.
- [33] M. Iman, H. R. Arabnia, and K. Rasheed, "A review of deep transfer learning and recent advancements," *Technologies*, vol. 11, no. 2, p. 40, Mar. 2023.
- [34] D. Zadok, O. Salzman, A. Wolf, and A. M. Bronstein, "Towards predicting fine finger motions from ultrasound images via kinematic representation," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2023, pp. 12645–12651.
- [35] E. Sassaroli, C. Crake, A. Scorza, D.-S. Kim, and M.-A. Park, "Image quality evaluation of ultrasound imaging systems: Advanced B-modes," *J. Appl. Clin. Med. Phys.*, vol. 20, no. 3, pp. 115–124, Mar. 2019.
- [36] E. J. Snider, S. I. Hernandez-Torres, and E. N. Boice, "An image classification deep-learning algorithm for shrapnel detection from ultrasound images," *Sci. Rep.*, vol. 12, no. 1, p. 8427, May 2022.
- [37] Y. Yu and S. T. Acton, "Speckle reducing anisotropic diffusion," *IEEE Trans. Image Process.*, vol. 11, no. 11, pp. 1260–1270, Nov. 2002.
- [38] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Proc. 13th Scand. Conf. Image Anal.* Cham, Switzerland: Springer, 2003, pp. 363–370.
- [39] G. Wu et al., "ISB recommendation on definitions of joint coordinate systems of various joints for the reporting of human joint motion—Part II: Shoulder, elbow, wrist and hand," *J. Biomech.*, vol. 38, no. 5, pp. 981–992, May 2005.
- [40] X. Zhang et al., "V-shaped dense denoising convolutional neural network for electrical impedance tomography," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–14, 2022.
- [41] A. Vaswani et al., "Attention is all you need," *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, 2017, pp. 5998–6008.
- [42] M. Salvi et al., "Multi-modality approaches for medical support systems: A systematic review of the last decade," *Inf. Fusion*, vol. 103, Mar. 2024, Art. no. 102134.
- [43] H. Wang, "Research on ultrasonic nondestructive testing method based on multimodal sensor fusion and deep learning," in *Proc. 8th Int. Conf. Adv. Algorithms Control Eng. (ICAACE)*, Mar. 2025, pp. 1–6.
- [44] X. Zhang et al., "Deep learning-based malicious energy attack detection in sustainable IoT network," in *Proc. Int. Conf. Comput., Netw. Commun. (ICNC)*, Feb. 2024, pp. 417–422.
- [45] J. Richter-Klug, P. Mania, G. Kazhoyan, M. Beetz, and U. Frese, "Improving object pose estimation by fusion with a multimodal prior—utilizing uncertainty-based CNN pipelines for robotics," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 2282–2288, Apr. 2022.
- [46] W. Hai, T. Qing, S. Na, and Z. Xiaodong, "Simultaneous estimation of hand joints' angles toward sEMG-driven human-robot interaction," *IEEE Access*, vol. 10, pp. 109385–109394, 2022.
- [47] Z. Chen, M. A. Gouda, L. Ji, and H. Wang, "A multimodal multistream multilevel fusion network for finger joint angle estimation with hybrid sEMG and FMG sensing," *Alexandria Eng. J.*, vol. 110, pp. 9–23, Jan. 2025.
- [48] S. Abbaspour, A. Naber, M. Ortiz-Catalan, H. GholamHosseini, and M. Lindén, "Real-time and offline evaluation of myoelectric pattern recognition for the decoding of hand movements," *Sensors*, vol. 21, no. 16, p. 5677, Aug. 2021.
- [49] C. Chen, Y. Yu, X. Sheng, J. Meng, and X. Zhu, "Real-time hand gesture recognition by decoding motor unit discharges across multiple motor tasks from surface electromyography," *IEEE Trans. Biomed. Eng.*, vol. 70, no. 7, pp. 2058–2068, Jul. 2023.
- [50] Z. Zhang, H. Jiang, and H. Singh, "NeuFlow: Real-time, high-accuracy optical flow estimation on robots using edge devices," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2024, pp. 5048–5055.
- [51] T. Romera et al., "Optical flow algorithms optimized for speed, energy and accuracy on embedded GPUs," *J. Real-Time Image Process.*, vol. 20, no. 2, p. 32, Apr. 2023.
- [52] S. M. Engdahl, S. A. Acuña, R. R. Kaliki, and S. Sikdar, "Sonomyography for control of upper-limb prostheses: Current state and future directions," *JPO J. Prosthetics Orthotics*, vol. 36, no. 3, pp. 174–184, 2024.