LATENT MANIFOLD FRAGMENTATION IN WEARABLE sEMG: AN EMPIRICAL AUDIT OF CROSS-SUBJECT GENERALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Surface electromyography (sEMG) gesture recognition literature frequently reports high offline accuracy ($> 80\%$), yet these claims often mask methodological pitfalls such as temporal window leakage and rest-class metric inflation. In this paper, we introduce **sEMG-Bench**, a standardized, leakage-free benchmark framework, and conduct an empirical audit of four representative deep architectures alongside feature-engineered baselines on NinaPro DB5. Our empirical findings reveal three key insights: (1) Under strict Leave-One-Subject-Out cross-validation (LOSOCV), zero-shot active gesture recognition collapses across all deep backbones ($\sim$20–25% Active F1), showing a pronounced disparity with reported offline claims. (2) Signal rectification ($|x(t)|$) provides a vital physical energy prior, recovering $+6\%$ to $+20\%$ Active Macro F1 over raw signed inputs. (3) Paired t-tests ($p > 0.05$) demonstrate statistical parity between lightweight 2D CNNs and complex Transformers for consumer-grade, low-sampling-rate dry electrode setups, indicating that architectural over-engineering yields no significant generalization advantage under inter-subject distribution shifts. Linear probing reveals the underlying mechanism: frozen latent representations encode biometric subject identity (up to 86% accuracy) rather than gesture-invariant topologies, proving that cross-subject adaptation is geometrically essential for consumer-grade wearable interfaces. Code is available at `https://anonymous.4open.science/r/sEMGBench-1507/`.

1 INTRODUCTION

Hand and finger gesture recognition using non-invasive sEMG biosignals (Asif et al., 2020; Soroushmojdehi et al., 2022; Tchantchane et al., 2023) has emerged as a foundational technology for modern Human-Computer Interfaces (HCI). By passively sensing neuromuscular activation signals generated during muscle contractions, sEMG-based decoding enables natural, continuous control across diverse applications, ranging from myoelectric prosthetic limb control (Stival et al., 2019) to robot teleoperation (Vogel et al., 2011; Meattini et al., 2018) and immersive virtual or augmented reality (VR/AR) (Dwivedi et al., 2020) interaction without requiring direct line-of-sight visual tracking.

Despite rapid advancements, evaluating and summarizing State-of-the-Art (SOTA) progress in sEMG gesture recognition remains fraught with discrepancies. Published studies report performance across fragmented experimental choices, creating an optimistic appearance of high offline accuracy that masks practical limitations (Zhong et al., 2025; Ketykó et al., 2019). Specifically, two major methodological flaws pervade the current literature:

1. **Label Scheme Ambiguity and Metric Poisoning:** Published literature inconsistently alternates between Rest-Included (Chen et al., 2020; Sri-Iesaranusorn et al., 2021; Cao et al., 2026), Rest-Excluded (Mekruksavanich et al., 2026; Wu et al., 2025), and personalized subsets (Yang et al., 2024; Cao et al., 2026; Sri-Iesaranusorn et al., 2021) evaluation schemes without explicit demarcation. Because Class 0 (Rest) accounts for over 50% of raw recording duration, raw overall accuracy is heavily inflated, creating an apples-to-oranges comparison across published works while obscuring true active gesture decoding capability.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2. **Validation Split Mismatch:** Existing literature diverges in how dataset splits are constructed. While earlier studies (Chen et al., 2020; Shin et al., 2025) apply random window-level cross-validation across 75%-overlapping sliding frames, others adopt trial-based repetition splits (Sri-Iesaranusorn et al., 2021; Mekruksavanich et al., 2026; Cao et al., 2026; Wu et al., 2025). Random shuffling across overlapping temporal frames introduces strong temporal correlations between training and test sets, producing optimistic offline accuracy metrics that degrade when evaluated on unseen gesture repetitions.

To address these field-wide challenges, this paper presents **`sEMG-Bench`**, a standardized, leakage-free benchmark framework and critical empirical investigation into sEMG gesture recognition. Our main contributions are summarized as follows:

1. **Open-Source Unified Benchmark Pipeline:** We release a standardized, fully reproducible PyTorch benchmark codebase for finger and wrist gesture recognition on the NinaPro benchmark dataset (Pizzolato et al., 2017; Stival et al., 2019), incorporating strict repetition-grouped 2-fold cross-validation that eliminates temporal window leakage.
2. **Literature Audit:** We evaluate representative model configurations spanning 4 literature across multi architecture: Signal2IMG (Chen et al., 2020), Feature extraction + DNN (Sri-Iesaranusorn et al., 2021), CNN-ResBiGRU-SE(Mekruksavanich et al., 2026), and MultiHeadAttention (Cao et al., 2026) and benchmark with 2 handcrafted feature and 1 deep CNN-uniGRU network baselines under identical, leakage-free conditions across Pooled, Per-Subject Intrapersonal, and Interpersonal Zero-Shot (LOSOCV) protocols.
3. **Zero-Shot Architectural Comparison and Prediction Stability:** Cross-subject evaluation indicates that differences in mean zero-shot accuracy between compact convolutional baselines (Chen et al., 2020) and complex 2026 Transformers (Cao et al., 2026) remain statistically marginal ($p > 0.05$) on consumer-grade, low-rate dry electrode setup in Ninapro DB5 (Pizzolato et al., 2017; Stival et al., 2019), indicating that increasing structural complexity alone provides limited benefit against inter-subject physiological variance in this case.

2 Related Work

Hand and finger gesture decoding using sEMG involves high-dimensional kinematic space, leading to significant fragmentation in how published studies configure gesture vocabularies, signal preprocessing, and validation schemes. Even when evaluating on the exact same public benchmark dataset, such as NinaPro DB5 (Pizzolato et al., 2017; Stival et al., 2019), prior literature adopts disparate experimental setups, rendering reported accuracy metrics non-comparable across studies.

Label Space Heterogeneity and Rest Class Inflation. A primary source of evaluation discrepancy lies in the definition of classification targets and the treatment of the neutral rest state. Benchmark datasets like NinaPro DB5 contain 12 basic finger movements (Exercise E1), 17 isotonic and wrist movements (Exercise E2), and 23 grasping postures (Exercise E3). As summarized in Table 1, published works select widely divergent class subsets. For instance, (Sri-Iesaranusorn et al., 2021) merged Exercises B and C into a 41-movement task, whereas (Cao et al., 2026) evaluated on a 22-gesture subset.

Importantly, studies differ in whether the unconstrained rest state (Class 0) is included during training and evaluation. Works like (Chen et al., 2020) and (Cui et al., 2026) evaluate under rest-included schemes (e.g., 13 classes for E1, 18 classes for E2). Because participants spend over 50% of total recording duration in the rest position (Pizzolato et al., 2017), Class 0 heavily dominates test samples. Consequently, standard top-1 accuracy becomes artificially inflated by easy rest classifications, masking severe error rates on fine active gestures. Conversely, (Mekruksavanich et al., 2026) filtered out rest frames entirely, evaluating strictly on 12 (E1) and 17 (E2) active gestures. Comparing a 96.93% accuracy under rest-excluded 5-fold CV (Mekruksavanich et al., 2026) directly against a 69.62% accuracy under rest-included evaluation (Chen et al., 2020) presents an apples-to-oranges comparison that obscures true architectural capability.

Table 1: Summary of published sEMG gesture recognition baselines on NinaPro DB5, detailing validation protocols, dataset exercise subsets, class count definitions, data splitting strategies, and reported accuracies on Exercise E1 and Exercise E2.

Study / Baseline	Validation Protocol	Class Count & Label Scheme	Data Splitting Strategy	E1 Acc	E2 Acc
(Chen et al., 2020)	Intrapersonal	13 (E1) / 18 (E2) [Rest Included]	Random Window Split	69.62%	67.42%
(Sri-Iesaranusorn et al., 2021)	Intrapersonal	41 Movements [Rest Included]	Repetition Split (R1,3,4,6 / R2,5)	84.25% (41 Combined)	
(Mekruksavanich et al., 2026)	Intrapersonal	12 (E1) / 17 (E2) [Rest Excluded]	5-Fold CV	96.93%	95.58%
(Cao et al., 2026)	Intrapersonal	22 Gestures [Rest Included]	Block/Trial-Level Split	81.96% (22 Combined)	

Temporal Window Leakage and Validation Protocols. A second critical flaw concerns data partitioning strategies. When segmenting continuous sEMG signals using sliding windows (e.g., 200 ms length with 50 ms stride), consecutive frames share up to 75% temporal overlap (150 ms of identical raw signal). Early deep learning baselines (Chen et al., 2020) performed random window-level shuffling prior to splitting into train/validation/test sets. This random partitioning introduces severe temporal data leakage, as near-identical overlapping frames from the same gesture execution appear simultaneously in both training and evaluation sets, yielding artificially elevated offline accuracy estimates (see more in Table 10 of (Yang et al., 2024)). To prevent leakage, more recent architectures (Sri-Iesaranusorn et al., 2021; Mekruksavanich et al., 2026; Cui et al., 2026) enforce repetition-grouped splitting (e.g., training on repetitions 1, 3, 4, 6 and testing on held-out repetitions 2 and 5 per subject).

Inter-Subject Distribution Shift and the Zero-Shot Gap. Finally, virtually all published SOTA baselines focus exclusively on Intrapersonal (Within-Subject) protocols, where a dedicated model is trained and calibrated for each target individual. While intrapersonal models achieve high accuracy ($> 80\%$), they require time-consuming user-specific data collection that hinders practical deployment.

More critically, (Yang et al., 2024) benchmarked model performance in interpersonal scenario by transforming sEMG signals into 2D representations, specifically data heatmaps, spectrograms, and continuous wavelet transforms (CWT), to process them as standard images. Their research indicate that for Ninapro DB5, the LOSO approach result amongst their configuration hover around 40% accuracy. While their benchmark provides useful baseline comparisons, it introduces two major limitations: (1) relying solely on AUROC and accuracy leaves the evaluation vulnerable to class-imbalance bias; and (2) omitting standard deviations obscures whether performance differences across methods are statistically meaningful.

Because of that, the community was pressing the need for a unified, leakage-free benchmark framework (`sEMG-Bench`) that evaluates representative SOTA architectures under standardized, uncorrupted conditions across both intrapersonal and zero-shot interpersonal regimes.

3 Model Curation and Baseline

To construct a comprehensive benchmark suite, we curated four representative deep learning architectures (Chen et al., 2020; Sri-Iesaranusorn et al., 2021; Mekruksavanich et al., 2026; Cao et al., 2026), alongside two handcrafted feature baselines (Handcrafted TD + XGBoost (Chen & Guestrin, 2016) and handcrafted feature + MLP) and one DL baseline. Official open-source code was publicly accessible only for (Mekruksavanich et al., 2026)[1]. For the remaining literature baselines where official code was unavailable, we faithfully reimplemented each architecture from scratch in PyTorch following the exact mathematical formulations, layer configurations, and hyperparameter specifications detailed in the original publications. A complete summary of the baseline validation protocols is provided in Table 1, and detailed architectural specifications, parameter counts, and attention mechanisms are audited in Table 2.

For fair comparison, all models were trained from scratch for 30 epochs using the Adam optimizer (Kingma & Ba, 2014), Cross-Entropy Loss, an initial learning rate of 0.001, a ReduceLROnPlateau learning rate scheduler with factor $= 0.5$ and patience $= 3$, and early stopping with patience $= 6$ to avoid overfitting.

[1] `https://github.com/januchit/Hand_Gesture_Recognition`

Table 2: Comprehensive parameter count and architectural technology audit across all evaluated models on NinaPro DB5, detailing trainable parameter counts for 13-class (Exercise E1) and 18-class (Exercise E2) setups, core neural backbones, and embedded attention mechanisms. The handcrafted features include MAV, RMS, WL, ZC, SSC, VAR, and IEMG.

Model / Baseline	**Params (E1)**	**Params (E2)**	**Core Architecture / Backbone**	**Attention Mechanism**
1D CNN-GRU baseline	397.5k	398.8k	1D CNN + GRU + Global Average Pooling (GAP)	1D Channel Excitation Mask
Handcrafted + XGBoost	NA	NA	Handcrafted TD Features (112-dim) + Gradient Boosted Trees	None
MAV + DNN	211.2k	212.5k	Handcrafted TD Features (16-dim) + 4-Layer Deep MLP	None
(Mekruksavanich et al., 2026)	394.9k	396.2k	Strided 1D Conv + Residual BiGRU (128 units) + LayerNorm	Squeeze-and-Excitation (SE) Channel Attention
(Chen et al., 2020)	35.6k	36.0k	Spectrogram transformation + 2DCNN	None
(Sri-Iesaranusorn et al., 2021)	260.4k	261.7k	Handcrafted Window Feature + 4-Layer Deep MLP	None
(Cao et al., 2026)	196.2k	196.5k	Cascaded Transformer + Channel Attention + BiLSTM	Multi-Head Self-Attn + Channel Attn + Temporal Attn

3.1 1D CNN-GRU Baseline Architecture

While recent literature has introduced increasingly complex multi-stage architectures, such as residual BiGRUs with Squeeze-and-Excitation blocks (Mekruksavanich et al., 2026) and cascaded Transformers with dual-path channel-temporal attention (Cao et al., 2026), it remains unclear whether such structural complexity is strictly necessary for decoding short sEMG window sequences. Our 1D CNN-GRU baseline is explicitly designed to evaluate the synergistic combination of channel-wise excitation masking, multi-stage 1D spatial-spectral feature extraction, and unidirectional temporal recurrence. As implemented in our codebase, the network architecture consists of four sequential functional stages:

1. **Input Normalization and 1D Channel Excitation Mask:** Raw sEMG input sequences ($X \in \mathbb{R}^{B\times 16\times T}$) are first normalized across the batch dimension via 1D Batch Normalization (BN_{1D}). The normalized signals are then passed through a 1D mask convolution layer ($\text{Conv1D}(16 \to 16, k = 5, p = 2)$) followed by a Sigmoid activation to generate an adaptive channel-wise excitation mask:

$$M(X) = \sigma\left(\text{Conv1D}_{mask}\left(\text{BN}_{1D}(X)\right)\right), \quad X_{exc} = \text{BN}_{1D}(X) \odot M(X) \tag{1}$$

   This excitation mechanism dynamically scales channel responses to emphasize informative electrode channels prior to feature extraction.

2. **Two-Stage 1D CNN Feature Extractor:** The masked feature representation X_{exc} is processed by a two-stage 1D convolutional backbone equipped with Swish (SiLU) non-linear activations, Batch Normalization, and Max Pooling downsampling:
   - *Stage 1:* $\text{Conv1D}(16 \to 64, k = 5, p = 2) \to \text{BN}_{1D} \to \text{SiLU} \to \text{MaxPool1D}(k = 2)$.
   - *Stage 2:* $\text{Conv1D}(64 \to 128, k = 3, p = 1) \to \text{BN}_{1D} \to \text{SiLU} \to \text{MaxPool1D}(k = 2)$.

This downsamples the sequence along the temporal axis while expanding spatial-spectral representation to 128 feature channels.

3. **Unidirectional GRU Temporal Encoder:** The downsampled feature maps are transposed into sequence format ($X_{seq} \in \mathbb{R}^{B\times T'\times 128}$) and fed into a single-layer unidirectional Gated Recurrent Unit (GRU) with 256 hidden units. The GRU models directional temporal evolution across sequential frames. A temporal mean pooling operation averages the GRU hidden states across time steps, collapsing the temporal sequence into a unified 256-dimensional context vector:

$$H_{gru} = \text{GRU}\left(X_{seq}\right) \in \mathbb{R}^{B\times T'\times 256}, \quad h_{pooled} = \frac{1}{T'}\sum_{t=1}^{T'} H_{gru}[t] \in \mathbb{R}^{B\times 256} \tag{2}$$

4. **Dense Classifier Head:** The 256-dimensional pooled context vector passes through a Dropout layer ($p = 0.4$), a dense linear layer (Linear$(256 \rightarrow 256)$), a Swish activation, a 1D Batch Normalization layer (BN$_{1D}(256)$), and a final linear classification layer (Linear$(256 \rightarrow K)$) to output class probability logits.

4 Dataset Preprocessing and Repetition Indexing

Our empirical evaluation utilizes Exercises E1 (Exercise A, 12 basic finger movements) and E2 (Exercise B, 17 functional hand and wrist movements) from the NinaPro DB5 benchmark dataset (Pizzolato et al., 2017; Stival et al., 2019). In the standard acquisition protocol, each participant performs K distinct active movement classes across 6 discrete repetition trials, with each 5-second active gesture separated by a 3-second neutral rest interval.

To enforce strict, leakage-free repetition-level data partitioning, we structure temporal segmentation as follows. In raw recordings, rest intervals preceding active movements are frequently unindexed (repetition $= 0$). We map each preceding rest interval to the repetition index ($r \in \{1, \ldots, 6\}$) of its corresponding active gesture trial. This converts the raw sequence into $K \times 6$ distinct, self-contained movement-rest repetition cycles per participant.

By grouping each active gesture alongside its adjacent rest interval into unified repetition units, sliding-window extraction (200 ms window, 50 ms stride) is structured to avoid bridging across repetition boundaries. Consequently, when partitioning data for cross-validation (e.g., training on repetitions $\{1, 3, 4\}$ and testing on $\{2, 5, 6\}$), both active gesture windows and their associated rest windows remain isolated, preventing temporal cross-fold data leakage.

5 Validation Protocols

To systematically evaluate model performance across different calibration regimes and data distributions, we evaluate all baseline architectures under two distinct validation protocols:

- **Pooled Multi-Participant Intrapersonal Benchmark:** Samples from all 10 participants are aggregated into a unified dataset, partitioned by repetition indices into two equal training and testing folds across 3 independent random repetition-shuffling seeds. This setup tests the capacity of models to learn shared gesture representations across multiple users simultaneously while maintaining strict repetition-level temporal isolation.

- **Interpersonal Zero-Shot LOSOCV Benchmark:** Evaluated under a Leave-One-Subject-Out cross-validation (LOSOCV) protocol. In each fold, a model is trained on data from 9 source subjects and evaluated on the 10th held-out subject without any target-user fine-tuning or calibration. This protocol rigorously measures zero-shot generalization under severe inter-subject physiological distribution shift.

Table 3: Quantitative performance comparison under the **Pooled Multi-Participant Protocol** on NinaPro DB5 Exercise E1 ($K = 13$ classes). Compares Raw Signed $x(t)$ vs. Rectified $|x(t)|$ input modes, reporting mean $\pm$ standard deviation across 3-repeated 2-fold splits and the net Active F1 gain (Δ_{ActiveF1}).

	Raw Signed Input $x(t)$		**Rectified Input $\lvert x(t)\rvert$**				**Rectification Impact**	
Model / Baseline	**Acc (%)**	**F1$_{\text{active}}$ (%)**	**Acc (%)**	**BalAcc (%)**	**F1$_{\text{macro}}$ (%)**	**F1$_{\text{active}}$ (%)**	**Δ_{Acc} (%)**	**Δ_{ActiveF1} (%)**
1D CNN-GRU baseline	75.40 ± 1.08	47.84 ± 1.69	81.04 ± 1.72	63.54 ± 1.79	64.86 ± 2.29	62.50 ± 2.43	+5.64	**+14.66**
Handcrafted + XGBoost	NA	NA	79.80 ± 1.75	58.12 ± 1.74	61.20 ± 2.56	58.55 ± 2.76	NA	NA
MAV + DNN baseline	NA	NA	82.96 ± 1.87	67.20 ± 2.55	68.86 ± 2.93	66.80 ± 3.13	NA	NA
(Mekruksavanich et al., 2026)	72.19 ± 1.15	41.48 ± 0.96	79.05 ± 1.41	58.57 ± 1.95	60.15 ± 1.85	57.42 ± 1.96	+6.86	**+15.94**
(Chen et al., 2020)	72.55 ± 0.74	40.75 ± 1.05	80.61 ± 1.17	60.73 ± 2.01	63.09 ± 2.06	60.59 ± 2.21	+8.06	**+19.84**
(Sri-Iesaranusorn et al., 2021)	NA	NA	$\mathbf{83.03 \pm 1.88}$	$\mathbf{67.76 \pm 2.56}$	$\mathbf{69.06 \pm 3.08}$	$\mathbf{67.02 \pm 3.30}$	NA	NA
(Cao et al., 2026)	78.79 ± 1.31	57.09 ± 1.99	81.19 ± 1.81	64.45 ± 2.23	65.53 ± 2.51	63.23 ± 2.67	+2.40	**+6.14**

6 RESULTS

6.1 DB 5 E1 DATASET EXPERIMENTAL ANALYSIS

6.1.1 POOLED MULTI-PARTICIPANT INTRAPERSONAL BENCHMARK AND RECTIFICATION ANALYSIS

The pooled multi-participant protocol evaluates model capacity to learn a generalized, shared feature space across heterogeneous user recordings simultaneously. Table 3 summarizes quantitative performance under both Raw Signed input ($x(t)$) and Rectified input ($|x(t)|$) regimes, detailing the exact mean gain (Δ) unlocked by non-linear rectification.

Under raw signed sEMG inputs ($x(t)$), as feature-engineered baselines inherently compute magnitude transformations (such as $\frac{1}{N}\sum |x_i|$) within their feature extraction pipelines, rendering raw signed vs. absolute value ablation not applicable (marked as NA in Table 3). However, as reported in Table 3, end-to-end models, 1D CNN-GRU baseline, reach 47.84% Active F1 and (Mekruksavanich et al., 2026)'s model achieve 41.48% Active F1. Applying signal rectification ($|x(t)|$) as a physical inductive bias yields a substantial performance recovery across all end-to-end architectures. Nonlinear rectification delivers a net Active F1 gain of $+14.66\%$ for 1D CNN-GRU baseline, $+19.84\%$ for (Chen et al., 2020), $+15.94\%$ for (Mekruksavanich et al., 2026), and $+6.14\%$ for (Cao et al., 2026).

This supports the hypothesis that raw 1D convolutions struggle to extract consistent activation profiles from zero-mean oscillatory biosignals, whereas rectified envelopes provide an informative energy representation that aligns feature distributions across participants.

6.1.2 INTERPERSONAL ZERO-SHOT LOSOCV BENCHMARK

Table 4: Quantitative performance comparison under the **Interpersonal Zero-Shot LOSOCV Protocol** on NinaPro DB5 Exercise E1 ($K = 13$ classes). Values report mean $\pm$ standard deviation across all 10 held-out test folds.

Model / Baseline	**Acc (%)**	**BalAcc (%)**	**F1$_{\text{macro}}$ (%)**	**F1$_{\text{active}}$ (%)**
1D CNN-GRU baseline	$\mathbf{67.72 \pm 4.75}$	29.92 ± 5.16	29.18 ± 5.93	24.08 ± 6.43
Handcrafted + XGBoost	65.42 ± 5.55	29.19 ± 5.34	28.11 ± 7.02	22.99 ± 7.80
MAV + DNN baseline	65.40 ± 6.73	30.52 ± 6.49	28.65 ± 8.48	23.59 ± 9.24
(Mekruksavanich et al., 2026)	64.90 ± 6.66	29.93 ± 5.43	27.86 ± 6.78	22.73 ± 7.30
(Chen et al., 2020)	66.09 ± 6.07	28.99 ± 6.91	27.32 ± 8.12	22.06 ± 8.75
(Sri-Iesaranusorn et al., 2021)	65.82 ± 6.69	$\mathbf{31.82 \pm 6.69}$	$\mathbf{29.88 \pm 9.41}$	$\mathbf{24.87 \pm 10.21}$
(Cao et al., 2026)	65.92 ± 6.87	30.18 ± 7.26	28.99 ± 8.24	23.92 ± 8.91

Under zero-shot inter-subject shift, active gesture decoding capability experiences a severe collapse across all evaluated architectures, dropping to a low performance band of F1$_{\text{active}} \approx 22.06\%$–$24.87\%$. As evident in Table 4, despite achieving $> 82\%$ intrapersonal accuracy under subject-specific calibration, no baseline architecture maintains functional decoding utility when transferred to an unseen participant without fine-tuning.

Statistical hypothesis testing indicates that zero-shot performance across deep architectures remains constrained within a narrow range (64.90%–67.72% Overall Acc, 22.06%–24.08% Active

F1). Paired t-tests between our 1D CNN-GRU model (24.08% $F1_{active}$) and Chen et al. (Chen et al., 2020) (22.06% $F1_{active}$) or the cascaded Transformer of Cao et al. (Cao et al., 2026) (23.92% $F1_{active}$) show no statistically significant difference ($p = 0.594$ and $p = 0.077$, respectively). While our baseline yields a statistically detectable improvement over Mekruksavanich et al. (Mekruksavanich et al., 2026) (22.73% $F1_{active}$, $p = 0.003$), the modest absolute margin demonstrates that scaling structural complexity alone provides limited practical benefit against physiological domain shift on NinaPro DB5. This also suggests the need for larger multi-subject cohorts to evaluate scaling behaviors thoroughly.

Nevertheless, 1D CNN-GRU baseline achieves the highest cross-subject prediction accuracy stability, exhibiting the lowest standard deviation across cross-subject evaluation folds ($\pm 4.75\%$ std. dev., compared to $\pm 6.87\%$ for (Cao et al., 2026) and $\pm 6.73\%$ for MAV + DNN baseline). This indicates that the compact recurrent-convolutional architecture preserves greater cross-subject stability across evaluation folds on this benchmark.

6.2 DB5 E2 Dataset Experimental Analysis

To evaluate model scalability and generalization across more complex isometric and isotonic wrist movements, we extend our comparative audit to Exercise E2 of NinaPro DB5 ($K = 18$ classes: 17 active wrist/hand postures plus the neutral rest state). The experimental outcomes on E2 mirror the overarching performance trends observed in Exercise E1, confirming that these empirical behaviors represent fundamental characteristics of sEMG signal processing rather than task-specific artifacts.

6.2.1 Pooled Multi-Participant Intrapersonal Benchmark

Table 5 summarizes quantitative performance under the Pooled Multi-Participant Protocol on Exercise E2 ($K = 18$ classes), reporting mean $\pm$ standard deviation across 3-repeated 2-fold cross-validation splits.

Table 5: Quantitative performance comparison under the **Pooled Multi-Participant Protocol** on NinaPro DB5 Exercise E2 ($K = 18$ classes). Values report mean $\pm$ standard deviation across 3-repeated 2-fold splits.

Model / Baseline	**Acc (%)**	**BalAcc (%)**	**$F1_{macro}$ (%)**	**$F1_{active}$ (%)**
1D CNN-GRU baseline	82.59 ± 0.91	60.11 ± 1.48	61.47 ± 1.57	59.42 ± 1.66
Handcrafted + XGBoost	80.60 ± 0.63	53.57 ± 1.12	56.25 ± 1.11	53.94 ± 1.17
MAV + DNN baseline	84.65 ± 1.01	$\mathbf{65.20 \pm 1.77}$	66.65 ± 1.79	64.91 ± 1.89
(Mekruksavanich et al., 2026)	80.53 ± 0.60	55.63 ± 0.87	56.78 ± 0.67	54.48 ± 0.71
(Chen et al., 2020)	80.80 ± 0.55	55.16 ± 1.03	56.61 ± 0.98	54.29 ± 1.03
(Sri-Iesaranusorn et al., 2021)	$\mathbf{84.69 \pm 1.07}$	65.14 ± 2.03	$\mathbf{66.77 \pm 2.02}$	$\mathbf{65.04 \pm 2.13}$
(Cao et al., 2026)	82.48 ± 0.99	60.84 ± 1.58	61.68 ± 1.66	59.65 ± 1.76

Under the 18-class wrist movement setting, feature-engineered baselines ((Sri-Iesaranusorn et al., 2021) and MAV + DNN) maintain a statistically significant advantage ($p < 0.001$) over raw deep learning backbones. (Sri-Iesaranusorn et al., 2021) achieves top pooled performance (84.69% Acc / 65.04% Active F1), outperforming end-to-end deep learning baselines such as (Mekruksavanich et al., 2026) (80.53% Acc / 54.48% Active F1, $p = 6.46 \times 10^{-6}$) and (Chen et al., 2020) (80.80% Acc / 54.29% Active F1, $p = 2.14 \times 10^{-7}$).

Our 1D CNN-GRU baseline achieves competitive performance among deep learning models (82.59% Acc / 59.42% Active F1), outperforming (Chen et al., 2020) by $+1.79\%$ Overall Accuracy and $+5.13\%$ Active F1, and outperforming (Mekruksavanich et al., 2026) by $+2.06\%$ Overall Accuracy ($p = 4.30 \times 10^{-5}$) and $+4.94\%$ Active F1. Statistical testing shows no significant difference between our 1D CNN-GRU model and the cascaded Transformer of (Cao et al., 2026) (82.48% Acc / 59.65% Active F1, $p = 0.195$), indicating that GRU recurrence effectively matches attention mechanisms.

6.2.2 Interpersonal Zero-Shot LOSOCV Benchmark

Under zero-shot cross-subject transfer, Exercise E2 exhibits the exact same performance collapse observed in Exercise E1. As shown in Table 6, active gesture recognition drops to a low range

Table 6: Quantitative performance comparison under the **Interpersonal Zero-Shot LOSOCV Protocol** on NinaPro DB5 Exercise E2 ($K = 18$ classes). Values report mean $\pm$ standard deviation across all 10 held-out test folds.

Model / Baseline	Acc (%)	BalAcc (%)	$F1_{macro}$ (%)	$F1_{active}$ (%)
1D CNN-GRU baseline	68.64 ± 6.51	26.71 ± 8.60	24.73 ± 9.22	20.70 ± 9.66
Handcrafted + XGBoost	68.87 ± 6.90	29.29 ± 6.79	28.22 ± 6.59	24.42 ± 6.90
MAV + DNN baseline	67.33 ± 8.60	27.69 ± 7.59	25.97 ± 7.13	22.10 ± 7.43
(Mekruksavanich et al., 2026)	67.15 ± 8.11	28.07 ± 6.83	26.45 ± 6.11	22.60 ± 6.33
(Chen et al., 2020)	67.18 ± 7.42	26.81 ± 5.63	25.14 ± 5.85	21.19 ± 6.11
(Sri-Iesaranusorn et al., 2021)	67.58 ± 8.67	27.28 ± 7.21	26.18 ± 7.14	22.31 ± 7.41
(Cao et al., 2026)	67.05 ± 8.65	26.50 ± 7.00	25.70 ± 7.30	21.79 ± 7.59

Table 7: **Subject ID Linear Probe Performance on Unseen Repetitions** (NinaPro DB5 E1, 10 subjects). Evaluated on held-out repetitions $\{2, 4, 6\}$ across all subjects using frozen penultimate embeddings h (Chance: 10.00%).

Model Backbone	Embedding Dim (h)	Probe Top-1 Acc (%)	Probe Macro F1 (%)
MAV + DNN baseline	256	**86.35**	**86.24**
(Sri-Iesaranusorn et al., 2021)	256	**83.05**	**83.00**
1D CNN-GRU baseline	256	**81.22**	**81.09**
(Cao et al., 2026)	64	64.47	64.00
(Mekruksavanich et al., 2026)	256	62.45	62.47
(Chen et al., 2020)	64	59.93	59.61

of $F1_{active} \approx 20.70\%$–$24.42\%$ across evaluated architectures, confirming that uncalibrated model transfer degrades equally on complex wrist configurations as it does on basic finger movements.

Formal paired t-tests reveal no statistically significant difference in zero-shot accuracy between several baseline models on Exercise E2 ($p > 0.05$), mirroring the convergence observed in Exercise E1. Paired comparisons reveal no statistically significant performance difference between our handcrafted+XGB (24.42% $F1_{active}$), our 1D CNN-GRU model (20.70% $F1_{active}$) and (Sri-Iesaranusorn et al., 2021) (22.31% $F1_{active}$, $p = 0.273$), (Chen et al., 2020) (21.19% $F1_{active}$, $p = 0.344$), or MAV + DNN (22.10% $F1_{active}$, $p = 0.194$). These matching cross-dataset trends demonstrate that physiological domain shifts, such as subject-specific muscle crosstalk and electrode placement variations, impose a substantial bottleneck across sEMG gesture recognition tasks.

7 Representation Entanglement and the Adaptation Imperative

A persistent barrier in sEMG is the collapse of deep neural models during cross-subject deployment: across six benchmarked architectures on NinaPro DB5 (Exercise 1), zero-shot gesture $F1_{active}$ on held-out subject S_1 plateaus at an anemic 22%–25%. While prior studies vaguely attribute this gap to physiological variance, our representation learning probes reveal a concrete geometric cause: deep sEMG models implicitly encode biometric subject identity rather than subject-invariant gesture manifolds.

Latent Space Probing: Models Learn Subject Identifiers. To diagnose this failure mode, we trained the gesture recognition models on Exercise 1 (excluding the rest class) using odd repetitions $\{1, 3, 5\}$, subject ID $(2 \dots 10)$. We then froze their penultimate representations ($h \in \mathbb{R}^d$) and trained a linear probe $W_s \in \mathbb{R}^{10 \times d}$ to classify Subject ID $(1 \dots 10)$. When evaluated on unseen, temporally isolated attempts (even repetitions $\{2, 4, 6\}$), the probe achieved 81.22% accuracy on our 1D CNN-GRU and up to 86.35% on handcrafted time-domain networks, vastly exceeding the uniform chance baseline of 10.00% (Table 7).

This substantial gap demonstrates that even without domain labels, deep backbones preserve individual anatomical signatures (e.g., muscle geometry, subcutaneous fat, and skin impedance) across independent motor executions. Notably, handcrafted features (MAV+DNN at 86.35%) act as direct biometric fingerprints, while deep temporal encoders (e.g., Cao et al. Transformer at 64.47%) partially filter static amplitude biases but still retain dominant subject identity.

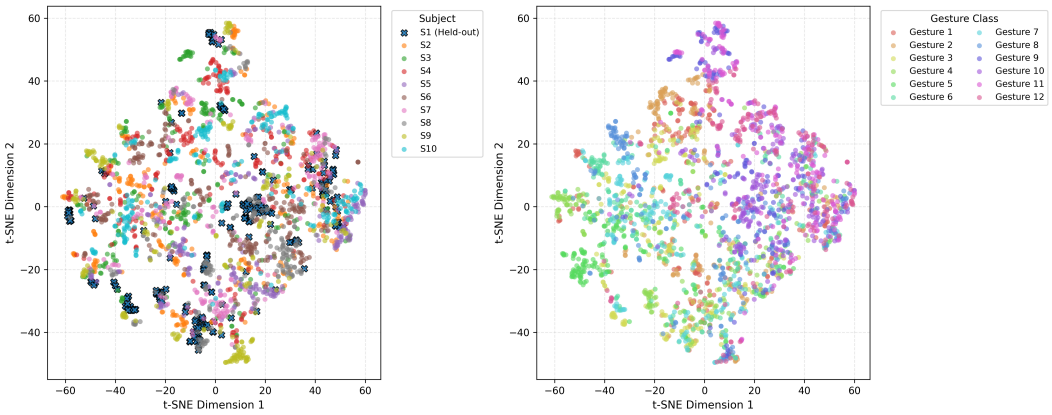

Figure 1: **Dual-Perspective t-SNE of Penultimate Embeddings** ($h \in \mathbb{R}^{256}$, our 1D CNN-GRU, $N = 2{,}400$ unseen test points). **Left:** Colored by Subject ID (S_1 held out marked with $\times$). **Right:** Colored by Gesture Class, showing severe intra-class fragmentation across subjects.

Geometric Verification via Manifold Projections. The dual-perspective t-SNE projections of the penultimate embeddings ($h \in \mathbb{R}^{256}$, 1D CNN-GRU baseline) in Figure 1 corroborate this geometric pathology (see Appendix D for additional architectures). When partitioned by Subject ID (Figure 1, Left), the latent space fragments into disjoint, subject-governed sub-manifolds rather than a shared, identity-invariant distribution. Within each individual, including the held-out subject S_1, embeddings further fractionate into approximately 10 to 12 distinct micro-clusters corresponding to discrete gesture trials. This demonstrates that while the convolutional backbone preserves strong intrapersonal class separability, it anchors these gesture representations to user-specific anatomical baselines.

Conversely, when colored by Gesture Class (Figure 1, Right), identical gestures across subjects fail to coalesce into unified decision clusters. At a macro level, the global manifold organizes along broad anatomical synergies, differentiating into coarse topological zones for index finger flexion (Gestures 1–2), the middle-ring-little digit group (Gestures 3–8), and thumb articulations (Gestures 9–12). At the micro level, however, each gesture class remains dispersed across disconnected, subject-specific islands.

Taken together, this geometry demonstrates that deep encoders implicitly prioritize biometric subject identity over subject-invariant kinematics, rendering zero-shot inference vulnerable to domain shift. Critically, because the held-out subject S_1 still exhibits distinct intra-subject gesture clusters, the underlying gesture topology is preserved rather than collapsed. This indicates that zero-shot transfer fails due to inter-subject manifold misalignment, establishing that lightweight calibration or test-time domain adaptation is well-suited to align these pre-structured sub-manifolds into a unified decision space.

8 Conclusion

This study presents a rigorous empirical audit and standardized benchmark of deep learning architectures for consumer-grade sEMG gesture decoding. Our findings demonstrate that while deep models achieve strong recognition accuracy under intra-subject calibration, generalization collapses under zero-shot cross-subject deployment. Through latent representation probing and manifold projection, we uncover the geometric mechanism behind this failure: deep backbones implicitly encode biometric subject identity rather than fine grain gesture-invariant topologies, rendering static zero-shot transfer fundamentally ill-posed. These empirical insights highlight the limitations of increasing architectural complexity alone and provide essential grounding for developing future adaptive and subject-aligned gesture recognition interfaces.

AI USE STATEMENT

In this work, we used generative AI tools for code generation and polishing language purpose. We have reviewed all AI-assisted work. For example, we checked LLM-generated research ideas for potential plagiarism through a manual literature survey and LLM-generated code was verified and tested for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Ali Raza Asif, Asim Waris, Syed Omer Gilani, Mohsin Jamil, Hassan Ashraf, Muhammad Shafique, and Imran Khan Niazi. Performance evaluation of convolutional neural network for hand gesture recognition using emg. *Sensors*, 20(6):1642, 2020.

Marco E. Benalcazar, Lorena Barona, Leonardo Valdivieso, Xavier Aguas, and Jonathan Zea. Emg-epn-612 dataset, November 2020. URL `https://doi.org/10.5281/zenodo.4421500`.

Tian-ao Cao, Huixin Liu, Jionghang Wu, Tong Zhu, Junhong Zheng, Zhengkui Chen, and Jijun Tong. Deep learning model cascade in electromyography decoding for gesture classification. *Scientific Reports*, 2026.

Patrick PK Chan, Qiuxia Li, Yinfeng Fang, Linyi Xu, Kairu Li, Honghai Liu, and Daniel S Yeung. Unsupervised domain adaptation for gesture identification against electrode shift. *IEEE Transactions on Human-Machine Systems*, 52(6):1271–1280, 2022.

Lin Chen, Jianting Fu, Yuheng Wu, Haochen Li, and Bin Zheng. Hand gesture recognition using compact cnn via surface electromyography signals. *Sensors*, 20(3):672, 2020.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

Luoqi Cui, Yong Liu, Hadi Fathollahi Abdar, Xinqin Gao, Mingshun Yang, and Mohammad Reza Chalak Qazani. Sensor topology-aware three-branch fusion for semg gesture recognition. *Sensors*, 26(15):4733, 2026.

Anany Dwivedi, Yongje Kwon, and Minas Liarokapis. Emg-based decoding of manipulation motions in virtual reality: Towards immersive interfaces. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 3296–3303. IEEE, 2020.

István Ketykó, Ferenc Kovács, and Krisztián Zsolt Varga. Domain adaptation for semg-based gesture recognition with recurrent neural networks. In *2019 international joint conference on neural networks (IJCNN)*, pp. 1–7. IEEE, 2019.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Roberto Meattini, Simone Benatti, Umberto Scarcia, Daniele De Gregorio, Luca Benini, and Claudio Melchiorri. An semg-based human–robot interface for robotic hands using machine learning and synergies. *IEEE Transactions on Components, Packaging and Manufacturing Technology*, 8 (7):1149–1158, 2018.

Sakorn Mekruksavanich, Narit Hnoohom, and Anuchit Jitpattanakul. Deep residual network with channel attention for improving hand gesture recognition with surface electromyography signal. *IEEE Access*, 2026.

Stefano Pizzolato, Luca Tagliapietra, Matteo Cognolato, Monica Reggiani, Henning Müller, and Manfredo Atzori. Comparison of six electromyography acquisition setups on hand movement classification tasks. *PloS one*, 12(10):e0186132, 2017.

Fernando Quivira, Toshiaki Koike-Akino, Ye Wang, and Deniz Erdogmus. Translating semg signals to continuous hand poses using recurrent neural networks. In *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pp. 166–169. IEEE, 2018.

Sasha Salter, Richard Warren, Collin Schlager, Adrian Spurr, Shangchen Han, Rohin Bhasin, Yujun Cai, Peter Walkington, Anuoluwapo Bolarinwa, Robert Wang, et al. emg2pose: A large and diverse benchmark for surface electromyographic hand pose estimation. *Advances in Neural Information Processing Systems*, 37:55703–55728, 2024.

Jungpil Shin, Abu Saleh Musa Miah, Sota Konnai, Shu Hoshitaka, and Pankoo Kim. Electromyography-based gesture recognition: Hierarchical feature extraction for enhanced spatial-temporal dynamics. *arXiv preprint arXiv:2504.03221*, 2025.

Rahil Soroushmojdehi, Sina Javadzadeh, Alessandra Pedrocchi, and Marta Gandolla. Transfer learning in hand movement intention detection based on surface electromyography signals. *Frontiers in Neuroscience*, 16:977328, 2022.

Panyawut Sri-Iesaranusorn, Attawit Chaiyaroj, Chatchai Buekban, Songphon Dumnin, Ronachai Pongthornseri, Chusak Thanawattano, and Decho Surangsrirat. Classification of 41 hand and wrist movements via surface electromyogram using deep neural network. *Frontiers in bioengineering and biotechnology*, 9:548357, 2021.

Francesca Stival, Stefano Michieletto, Matteo Cognolato, Enrico Pagello, Henning Müller, and Manfredo Atzori. A quantitative taxonomy of human hand grasps. *Journal of neuroengineering and rehabilitation*, 16(1):28, 2019.

Rayane Tchantchane, Hao Zhou, Shen Zhang, and Gursel Alici. A review of hand gesture recognition systems based on noninvasive wearable sensors. *Advanced intelligent systems*, 5(10): 2300207, 2023.

Jörn Vogel, Claudio Castellini, and Patrick van der Smagt. Emg-based teleoperation and manipulation with the dlr lwr-iii. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 672–678. IEEE, 2011.

Chuheng Wu, S Farokh Atashzar, Mohammad M Ghassemi, and Tuka Alhanai. An lstm feature imitation network for hand movement recognition from semg signals. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Jehan Yang, Maxwell Soh, Vivianna Lieu, Douglas J Weber, and Zackory Erickson. Emgbench: Benchmarking out-of-distribution generalization and adaptation for electromyography. *Advances in Neural Information Processing Systems*, 37:50313–50342, 2024.

Wenjuan Zhong, Xinyu Jiang, Katarzyna Szymaniak, Milad Jabbari, Chenfei Ma, and Kianoush Nazarpour. Deep feature learning from electromyographic signals for gesture recognition systems. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2025.

A Evaluation Metrics

To provide a comprehensive and unpoisoned assessment of gesture decoding capability, we report four complementary performance metrics:

1. **Overall Accuracy (Acc):** Standard top-1 classification accuracy across all evaluation windows.
2. **Balanced Accuracy (BalAcc):** The unweighted mean of recall across all individual gesture classes, mitigating overall class frequency disparities.
3. **Macro F1 Score ($\mathbf{F1_{macro}}$):** The arithmetic mean of $F1$ scores computed across all classes, including the neutral rest state (Class 0).
4. **Active Macro F1 Score ($\mathbf{F1_{active}}$):** The mean $F1$ score calculated strictly across active gesture classes ($c \in \{1, \ldots, K\}$), excluding Class 0 (Rest).

As established in Section 2, because the neutral rest state accounts for over 50% of raw signal duration, standard Accuracy and Macro F1 metrics are susceptible to majority-class inflation. The **Active Macro F1 score** serves as our primary benchmark metric, isolating true active neuromuscular decoding performance from rest-class bias.

B Ablation study

B.1 Intrapersonal in DB5 E1

B.1.1 Per-Subject Intrapersonal Benchmark

Table 8: Quantitative performance comparison under the **Per-Subject Intrapersonal Protocol** on NinaPro DB5 Exercise E1 ($K = 13$ classes, `ABS` input mode). Values report mean $\pm$ standard deviation across all 10 subject-specific models.

Model / Baseline	Acc (%)	BalAcc (%)	$\mathbf{F1_{macro}}$ (%)	$\mathbf{F1_{active}}$ (%)
1D CNN-GRU baseline	83.27 ± 2.97	67.01 ± 4.10	68.59 ± 4.30	66.45 ± 4.67
Handcrafted + XGBoost	81.93 ± 2.65	64.69 ± 3.24	66.13 ± 3.55	63.83 ± 3.86
MAV + DNN baseline	84.71 ± 2.72	70.19 ± 3.64	71.79 ± 3.75	69.91 ± 4.06
(Mekruksavanich et al., 2026)	82.24 ± 3.16	65.85 ± 3.77	67.07 ± 4.19	64.86 ± 4.56
(Chen et al., 2020)	84.49 ± 2.96	69.38 ± 4.04	71.25 ± 4.19	69.34 ± 4.52
(Sri-Iesaranusorn et al., 2021)	**84.72 ± 2.66**	**70.31 ± 3.73**	**71.88 ± 3.81**	**70.02 ± 4.13**
(Cao et al., 2026)	82.91 ± 2.67	68.13 ± 3.21	68.87 ± 3.54	66.81 ± 3.83

When model training is restricted exclusively to an individual's neuromuscular data, performance improves consistently across all architectures, yielding accuracy gains of $+2.0\%$ to $+5.0\%$ over multi-participant pooling. As detailed in Table 8, 1D CNN-GRU baseline improves from 81.04% to 83.27% accuracy (+2.23%). Eliminating inter-subject physiological variance allows neural network filters to specialize to individual muscle activation topologies and electrode alignments.

However, the per-subject results reveal a persistent metric discrepancy between Overall Accuracy and Active Macro F1. Across all evaluated architectures, Overall Accuracy (81.93%–84.72%) consistently exceeds Active F1 (63.83%–70.02%) by more than 14 percentage points. Because the neutral rest position (Class 0) accounts for over 50% of raw signal duration, models achieve high classification accuracy on rest frames, which skews standard accuracy metrics upward while masking misclassification rates on fine active finger movements. This empirical margin highlights the importance of adopting Active Macro F1 as the primary benchmark metric for prosthetic decoding.

B.2 Intrapersonal in DB5 E2

B.2.1 Per-Subject Intrapersonal Benchmark

Under subject-specific training, performance trends on Exercise E2 closely align with Exercise E1. As detailed in Table 9, restricting training to individual participant data improves classification accuracy across all architectures by $+1.2\%$ to $+4.0\%$ over multi-participant pooling. MAV + DNN

Table 9: Quantitative performance comparison under the **Per-Subject Intrapersonal Protocol** on NinaPro DB5 Exercise E2 ($K = 18$ classes). Values report mean ± standard deviation across all 10 subject-specific models.

Model / Baseline	Acc (%)	BalAcc (%)	F1_{macro} (%)	$\text{F1}_{\text{active}}$ (%)
1D CNN-GRU baseline	83.83 ± 2.36	63.42 ± 3.56	64.17 ± 3.52	62.26 ± 3.70
Handcrafted + XGBoost	83.96 ± 2.47	62.46 ± 4.18	64.37 ± 4.14	62.49 ± 4.36
MAV + DNN baseline	**86.27 ± 2.27**	**69.32 ± 3.59**	**70.29 ± 3.44**	**68.74 ± 3.60**
(Mekruksavanich et al., 2026)	83.12 ± 2.65	61.80 ± 4.04	62.79 ± 4.10	60.82 ± 4.31
(Chen et al., 2020)	84.83 ± 2.48	65.69 ± 4.27	66.60 ± 4.10	64.84 ± 4.31
(Sri-Iesaranusorn et al., 2021)	86.12 ± 2.21	68.89 ± 3.65	69.96 ± 3.38	68.39 ± 3.53
(Cao et al., 2026)	84.09 ± 2.27	64.64 ± 3.67	65.11 ± 3.44	63.26 ± 3.59

Table 10: Benchmark evaluation across all 306 participants on EMG-EPN612. Performance is contrasted between pooled intra-personal and interpersonal zero-shot setups, reporting the resulting generalization gap (Δ = Zero-Shot − Pooled).

Model / Architecture	Pooled Intra-Personal		Interpersonal Zero-Shot		Generalization Gap	
	Acc (%)	$\text{F1}_{\text{active}}$ (%)	Acc (%)	$\text{F1}_{\text{active}}$ (%)	ΔAcc	$\Delta\text{F1}_{\text{active}}$
1D CNN-GRU baseline	55.30	47.16	44.48	36.05	−10.82	−11.11
Handcrafted + XGBoost	48.86	38.71	43.50	32.82	−5.36	−5.89
MAV + DNN baseline	52.12	42.85	42.62	32.57	−9.50	−10.28
Sri-Iesaranusorn et al. (2021)	52.16	42.93	42.85	32.72	−9.31	−10.21
Chen et al. (2020)	52.40	43.55	**44.75**	35.46	−7.65	−8.09
Mekruksavanich et al. (2026)	56.06	48.18	44.26	35.71	−11.80	−12.47
Cao et al. (2026)	**57.31**	**49.55**	44.37	**36.10**	−12.94	−13.45

and (Sri-Iesaranusorn et al., 2021) maintain top intrapersonal performance (86.27% Acc / 68.74% Active F1 and 86.12% Acc / 68.39% Active F1, respectively).

In addition, the persistent margin between Overall Accuracy (83.12%–86.27%) and Active Macro F1 (60.82%–68.74%) is consistently observed in Exercise E2. This 17+ percentage point divergence reinforces the insight drawn from E1: standard accuracy metrics are heavily inflated by the prevalence of rest frames (Class 0), demonstrating that Active Macro F1 must serve as the primary evaluation metric for prosthetic gesture control regardless of gesture complexity.

B.3 Additional Result on EPN612 Dataset

To evaluate whether our empirical findings generalize beyond the 10-subject cohort of NinaPro DB5, we benchmark all curated architectures on the training set of **EMG-EPN612** dataset (Benalcazar et al., 2020), comprising $N = 306$ intact participants recorded with an 8-channel consumer-grade Thalmic Myo armband across 5 active functional gestures.

Experimental Methodology. To eliminate label ambiguity and non-causal leakage, we enforce the identical rigorous evaluation constraints established in sEMG-Bench:

1. **Causal Active-Window Extraction:** We sample non-overlapping 200 ms windows ($w = 40$ frames at $f_s = 200$ Hz) strictly during active neuromuscular contraction ($\geq$ `startPointforGestureExecution`), preventing flatline rest poisoning on active gesture tokens.
2. **Hardware-Bound Rectification:** Inputs are causally scaled by hardware dynamic range and rectified ($|x(t)|/128.0$) without whole-trial peak-normalisation.
3. **Protocols:** We evaluate: (i) *Pooled Intra-Personal*, performing a stratified 2:1 train-test partition over all 306 participants; and (ii) *Interpersonal Zero-Shot*, training on 245 source participants ($\approx 80\%$) and evaluating on 61 completely unseen target participants without user calibration.

Findings and Cross-Dataset Convergence. As detailed in Table 10, cross-subject performance degradation findings from Ninapro DB5 dataset is scaled to a higher subject dataset. When evaluated on 61 unseen, uncalibrated subjects, decoding performance across all deep architectures collapses into a remarkably narrow band: 44.26%–44.75% accuracy and 35.46%–36.10% active macro F1. Notably, complex cascaded Transformers with channel attention (Cao et al., 2026) outperform the compact 34k-parameter 2D CNN (Chen et al., 2020) and our 1D CNN-GRU baseline by only $+0.64$ and $+0.05$ percentage points in active F1. This cross-architecture performance ceiling indicates that model scaling and architectural complexity alone cannot resolve inter-subject domain shift.

C Further Discussion

C.1 Finger Gesture Decoding Remains an Open Challenge

Because published literature predominantly emphasizes intrapersonal validation in sEMG finger gesture recognition, this study provides an empirical grounding for evaluating genuine research progress under uncalibrated settings. When evaluated under zero-shot inter-subject transfer, active gesture decoding performance drops to a constrained performance range ($F1_{\text{active}} \approx 22\%$–$25\%$, Balanced Accuracy $\approx 30\%$). This degradation demonstrates that sEMG-based prosthetic finger control remains an open challenge, exposing a substantial gap between high offline intrapersonal claims ($> 95\%$) and real-world uncalibrated utility.

In addition, our benchmark findings indicate that recent over-parameterized deep learning architectures fail to reliably surpass classical time-domain feature baselines on NinaPro DB5. Across both pooled multi-participant and per-subject intrapersonal protocols, classical feature-engineered DNN models (achieving 83.0% pooled and 84.7% per-subject accuracy) match or slightly exceed complex 2026 architectures, including cascaded Transformers (Cao et al., 2026) (81.19% pooled accuracy). This highlights the importance of establishing unified, leakage-free benchmark suites (`sEMG-Bench`) to evaluate future architectural proposals rigorously.

C.2 Discrete Pose Classification vs. Continuous Movement-Centric Decoding

Current literature in sEMG-based gesture recognition predominantly frames hand movement decoding as a discrete classification task over isolated static poses (Mekruksavanich et al., 2026; Chen et al., 2020; Sri-Iesaranusorn et al., 2021). While effective for benchmark comparisons, this discrete formulation oversimplifies human hand function. Natural hand dynamics are inherently continuous, fluid, and composite, governed by fine-grained individual digit kinematics and complex multi-finger synergies that underpin real-world dexterity (Salter et al., 2024; Quivira et al., 2018).

Framing gesture recognition strictly as discrete class selection restricts the system from decoding unconstrained, intermediate hand configurations. Acknowledging this limitation, we advocate for future research to cultivate a dual-track paradigm: maintaining discrete classification for high-confidence discrete commands while advancing continuous, movement-centric regression and joint-angle trajectory decoding. Expanding this line of research will bridge the gap between discrete gesture classification and dexterous, real-time prosthetic and robotic control.

C.3 Limitation

A critical consideration in interpreting our empirical findings is the hardware and cohort scope of the evaluation. Our benchmark evaluates NinaPro DB5, which utilizes dual consumer-grade Thalmic Myo armbands (16 dry electrode channels sampled at 200 Hz across 10 intact participants). While representative of low-cost wearable interfaces, several characteristics must be recognized:

Firstly, typical sEMG spectral energy extends up to 450–500 Hz. The 200 Hz sampling rate of the Myo armband imposes a Nyquist cutoff at 100 Hz, attenuating higher-frequency motor unit action potential dynamics.

Another limitation is that dry electrodes have higher skin-electrode interface impedance and greater motion-artifact vulnerability than clinical wet Ag/AgCl arrays. In addition, 16 sparse circumferential channels offer lower spatial resolution than dense high-density sEMG grids.

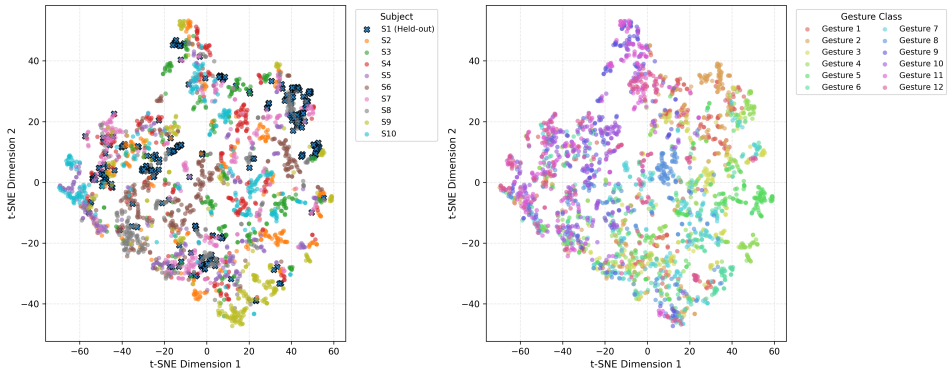

Figure 2: **Dual-Perspective t-SNE of Penultimate Embeddings** ($h \in \mathbb{R}^{256}$, (Sri-Iesaranusorn et al., 2021), $N = 2{,}400$ unseen test points). **Left:** Colored by Subject ID (S_1 held out marked with $\times$). **Right:** Colored by Gesture Class, showing severe intra-class fragmentation across subjects.

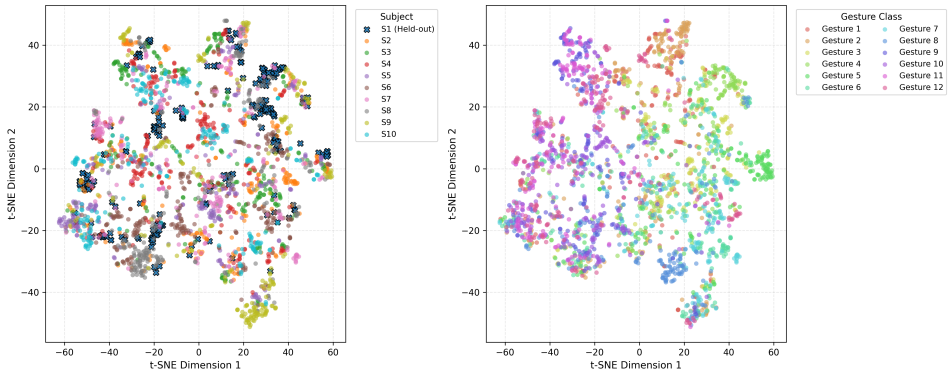

Figure 3: **Dual-Perspective t-SNE of Penultimate Embeddings** ($h \in \mathbb{R}^{256}$, (Chen et al., 2020), $N = 2{,}400$ unseen test points). **Left:** Colored by Subject ID (S_1 held out marked with $\times$). **Right:** Colored by Gesture Class, showing severe intra-class fragmentation across subjects.

Beyond problem formulation, real-world deployment introduces physical and environmental confounders that are frequently omitted in benchmark evaluations. In unconstrained applications, sEMG sensors are susceptible to physical rotation, spatial displacement across sessions, skin-electrode impedance variations, and signal degradation over extended use.

Finally, in multi-device configurations, such as the dual-armband setup evaluated in NinaPro DB5, practical usability is vulnerable to human error. Users may inadvertently don the armbands out of sequence, invert device orientation, or shift electrode positions relative to underlying muscle bellies. Current machine learning architectures rely heavily on static spatial channel ordering and fail to generalize when confronted with such unstructured spatial noise and hardware-shift dynamics. Developing rotation-invariant representations, self-supervised spatial recalibration, and domain-adaptation techniques remains an important open challenge for robust sEMG (Chan et al., 2022).

D More t-SNE Results

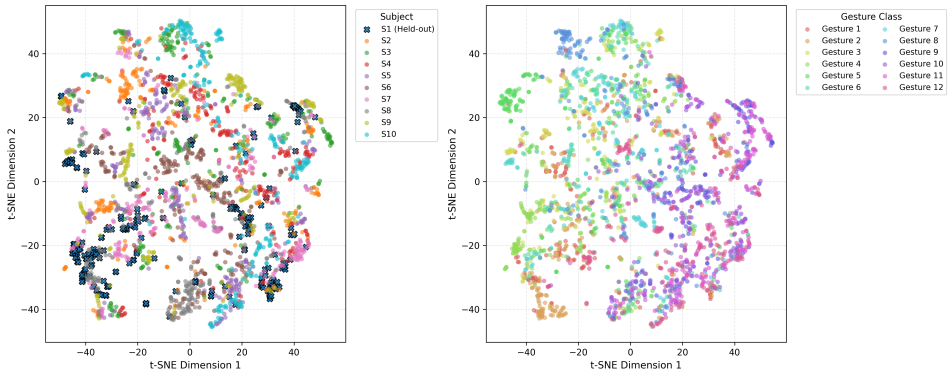

Figure 4: **Dual-Perspective t-SNE of Penultimate Embeddings** ($h \in \mathbb{R}^{256}$, (Cao et al., 2026), $N = 2{,}400$ unseen test points). **Left:** Colored by Subject ID (S_1 held out marked with $\times$). **Right:** Colored by Gesture Class, showing severe intra-class fragmentation across subjects.

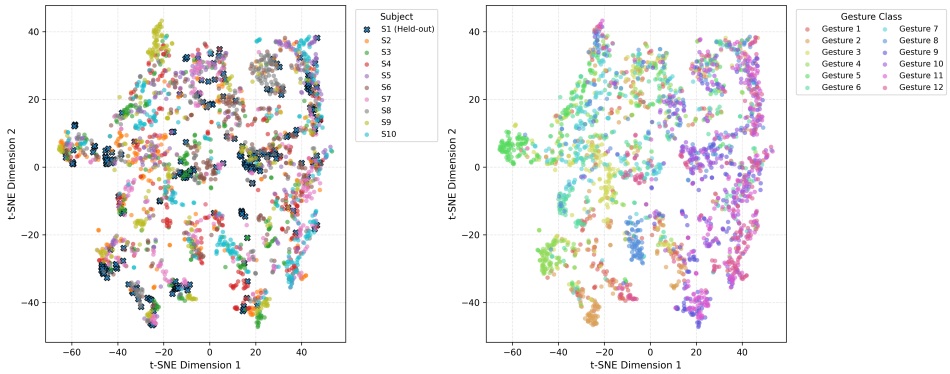

Figure 5: **Dual-Perspective t-SNE of Penultimate Embeddings** ($h \in \mathbb{R}^{256}$, (Mekruksavanich et al., 2026), $N = 2{,}400$ unseen test points). **Left:** Colored by Subject ID (S_1 held out marked with $\times$). **Right:** Colored by Gesture Class, showing severe intra-class fragmentation across subjects.

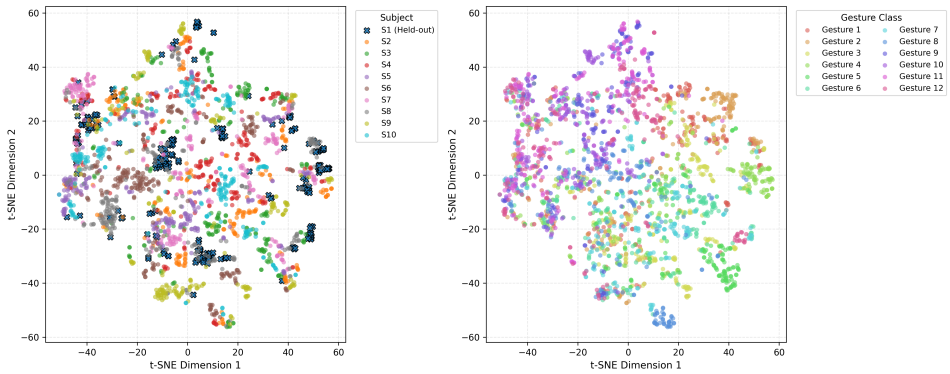

Figure 6: **Dual-Perspective t-SNE of Penultimate Embeddings** ($h \in \mathbb{R}^{256}$, our DNN-MAV, $N = 2{,}400$ unseen test points). **Left:** Colored by Subject ID (S_1 held out marked with $\times$). **Right:** Colored by Gesture Class, showing severe intra-class fragmentation across subjects.