SAME GESTURE, DIFFERENT SENSORS: CROSS-DATASET REPRESENTATION LEARNING FOR HETEROGENEOUS sEMG GESTURE RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Surface electromyography (sEMG) is a promising sensing modality for wearable hand gesture recognition. Existing datasets are still analyzed separately due to differences in electrode configurations, channel counts, subjects, and acquisition conditions. We propose treating heterogeneous sEMG datasets as complementary sensor observations linked by shared gesture identity. Our cross-dataset representation learning framework handles variable-channel inputs through shared per-channel encoding and dataset-specific adapters. During pre-training, channel masking with channel-representation reconstruction reduces dependence on individual sensors. Supervised cross-dataset contrastive learning aligns same-gesture representations across datasets. The learned representation transfers to an unseen target dataset through masked target adaptation and supervised fine-tuning on the target training split alone. Across five sEMG datasets spanning 2–16 channels, our source-only transfer approach improves average balanced accuracy from 62.57% to 74.15% over single-dataset training, with improvements on all five target datasets. Our results show that shared gesture identity provides useful semantic correspondence for representation transfer across heterogeneous sEMG sensor configurations.

1 INTRODUCTION

Surface electromyography (sEMG) provides a wearable interface for decoding hand and wrist movements. sEMG has been widely studied for human–computer interaction, rehabilitation, and prosthetic control (Ni et al., 2024; Jaramillo-Yánez et al., 2020; Li et al., 2021). Deep learning has substantially improved sEMG gesture recognition (Atzori et al., 2016; Montazerin et al., 2023). However, most models are developed and evaluated within individual datasets and sensing configurations.

The dataset-specific paradigm limits the reuse of existing sEMG data. Independently collected datasets differ in electrode placement, channel count, acquisition hardware, subjects, and recording protocols. sEMG recognition is also sensitive to electrode configuration and sensing location (Young et al., 2011a;b; Pan et al., 2015; Ameri et al., 2019; He et al., 2023). Differences in sensing conditions make knowledge transfer across datasets challenging.

Despite this heterogeneity, many sEMG datasets contain overlapping gesture classes such as hand opening, hand closing, and wrist movements. Their signal observations can differ substantially across sensor configurations while representing the same underlying movement. We use this shared gesture identity as semantic correspondence across datasets. This leads to our central question: *Can shared gesture identity enable representation transfer across heterogeneous sEMG datasets with different sensor configurations?*

Recent work has demonstrated the value of heterogeneous EMG pre-training (Olsson et al., 2026; Huang et al., 2026). We build on this direction by exploiting shared gesture identity across datasets. When gesture vocabularies overlap, the gesture labels provide explicit correspondence between otherwise heterogeneous sensor observations.

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.

We propose a cross-dataset representation learning framework for variable-channel sEMG. A shared per-channel encoder processes each available channel, while dataset-specific adapters account for sensing differences. During source pre-training, channel masking and representation reconstruction reduce dependence on individual sensor observations. Supervised contrastive learning preserves gesture-level structure. A cross-dataset objective further aligns same-gesture observations from different datasets.

We transfer the learned representation to a previously unseen target sensor configuration. The target dataset is completely excluded from heterogeneous source pre-training. A new target adapter is initialized for the unseen configuration. The model then undergoes masked target adaptation and end-to-end supervised fine-tuning using only the target training split. The source datasets are not revisited during this transfer. Our contributions are as follows:

- We formulate cross-dataset sEMG learning using shared gesture identity as correspondence across heterogeneous sensor configurations.
- We propose a variable-channel framework with shared channel encoding, dataset-specific adaptation, masked representation learning, and cross-dataset gesture alignment.
- We introduce source-only transfer to an unseen sensor configuration without revisiting the source datasets.
- We evaluate the framework on five sEMG datasets spanning 2–16 channels and analyze its key components and learned representations.

2 Related Work

2.1 sEMG-based Hand Gesture Recognition

sEMG has been widely studied for hand gesture recognition and human–machine interaction (Jaramillo-Yánez et al., 2020; Li et al., 2021; Xiong et al., 2021; Xu & Xiong, 2021; Ni et al., 2024). Deep architectures have substantially improved sEMG gesture recognition. Prior work includes convolutional approaches (Geng et al., 2016; Atzori et al., 2016) and more recent Transformer-based models (Zhang et al., 2023; Montazerin et al., 2023).

sEMG recognition remains sensitive to changes in subjects, sessions, and electrode configurations. Prior work has studied robustness to electrode shift and sensing placement (Young et al., 2011a;b; Pan et al., 2015; Ameri et al., 2019; He et al., 2023). Transfer learning and domain adaptation have also been explored for cross-subject generalization (Côté-Allard et al., 2019; Ketykó et al., 2019; Hoshino et al., 2022; Wang et al., 2023; 2024; Colot et al., 2026) and cross-session robustness (Du et al., 2017; Zhang et al., 2024; Guo et al., 2024; Li et al., 2025). Most of these studies address subject, session, or placement shifts. Our work instead considers representation learning across heterogeneous datasets with heterogeneous sensor configurations.

2.2 Representation Learning for sEMG

Recent work has explored contrastive and masked learning for transferable sEMG representations. sEMGPoseMIM learns trial-invariant representations through cross-trial mutual information maximization. Hand-motion supervision further refines the representations through cross-modal knowledge distillation (Dai et al., 2023). STCNet uses subject-aware contrastive learning to address inter-subject variability (Yang et al., 2025). Self-supervised sEMG learning has also adopted redundancy-reduction objectives such as VICReg (Bardes et al., 2021; Raghu et al., 2025). Masked modeling provides another approach to robust sEMG representation learning (Wang et al., 2025).

EMBridge aligns EMG with synchronized hand-pose representations through cross-modal representation learning (Cui et al., 2026). AEMG learns generalizable representations from heterogeneous EMG recordings through large-scale self-supervised pre-training (Huang et al., 2026). These approaches build on broader advances in contrastive and masked representation learning (Oord et al., 2018; Chen et al., 2020; Khosla et al., 2020; Bardes et al., 2021; He et al., 2022). Our work also targets transferable representations. In contrast, shared gesture identity provides explicit semantic correspondence between heterogeneous sEMG datasets.

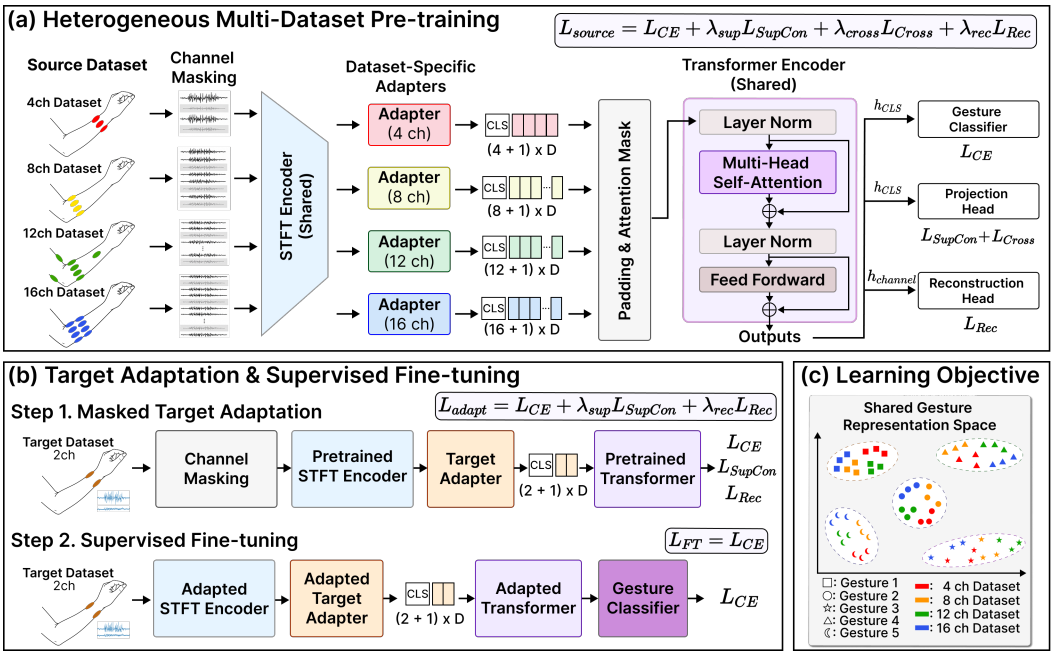

Figure 1: Overview of the proposed cross-dataset sEMG representation learning framework. Heterogeneous source datasets are jointly used to learn shared gesture representations across different sensor configurations. The framework combines shared per-channel encoding, dataset-specific adapters, cross-dataset gesture alignment, and masked channel-representation learning. The source-pretrained representation is transferred to an unseen target through masked adaptation and end-to-end supervised fine-tuning.

2.3 Heterogeneous and Multi-Dataset sEMG Learning

Recent work has explored unified pre-training across heterogeneous EMG datasets. SAFT uses large-scale supervised pre-training and transfers the learned model to held-out sensing settings (Olsson et al., 2026). AEMG uses self-supervised pre-training to learn representations across heterogeneous devices and sensor configurations (Huang et al., 2026). Both studies demonstrate the potential of reusing heterogeneous EMG datasets despite differences in acquisition settings.

Our work focuses on explicit gesture-level correspondence across datasets. Recordings with the same gesture identity from different datasets form cross-dataset positive pairs. The resulting objective aligns same-gesture representations across sensor configurations. Masked channel-representation learning further reduces dependence on individual sensor observations. For source-only transfer, the target dataset is completely excluded from heterogeneous source pre-training. The learned representation is then adapted to the unseen target configuration without revisiting the source datasets.

3 Methodology

We propose a cross-dataset representation learning framework for heterogeneous sEMG datasets with different sensor configurations. The framework uses shared gesture identity as semantic correspondence across heterogeneous datasets. Figure 1 illustrates the overall framework. Heterogeneous source datasets first provide a shared representation across sensor configurations. The source-pretrained representation is then transferred to an unseen target using only the target training split.

3.1 Problem Formulation

We consider K heterogeneous source sEMG datasets. The datasets may differ in channel count, electrode placement, acquisition hardware, subjects, and recording protocol. The gesture vocabularies may also differ across datasets. We define a shared label space $\mathcal{Y}_{\text{shared}}$ using gestures that can be

consistently matched across the participating datasets. Samples with the same shared label represent the same underlying gesture despite differences in sensor configuration.

Our goal is to learn gesture-relevant representations while reducing dependence on a particular sensing configuration. The learned representation is transferred to a target dataset that is completely excluded from heterogeneous source pre-training. The target may contain a previously unseen sensor configuration. Formal notation is provided in Appendix A.

3.2 Variable-Channel sEMG Representation

We preserve the native channels of each dataset rather than converting heterogeneous recordings into a common fixed-channel layout. Each available sEMG channel is treated as an individual sensor observation. The variable-channel design avoids artificial channel correspondence across heterogeneous datasets.

Each channel is transformed into a log-magnitude STFT and encoded by a shared 2D convolutional encoder. The same encoder parameters are used for every channel and dataset. A lightweight dataset-specific residual adapter accounts for systematic differences between sensor configurations. The adapted channel representations are combined with learnable channel-index embeddings and a CLS token:

$$\mathbf{T}_i = [\mathbf{t}_{\mathrm{CLS}}; \tilde{\mathbf{e}}_{i,1} + \mathbf{p}_1; \ldots; \tilde{\mathbf{e}}_{i,C_i} + \mathbf{p}_{C_i}]. \tag{1}$$

Each token has embedding dimension D. The number of channel tokens varies with the native channel count C_i. Mini-batches are padded to the maximum channel count within each batch. A padding mask excludes padded positions from self-attention.

A shared Transformer (Vaswani et al., 2017) processes the resulting token sequence. The contextualized CLS output $\mathbf{h}_{i,\mathrm{CLS}}$ provides the sample-level gesture representation. The contextualized channel outputs $\mathbf{h}_{i,c}$ support channel-representation reconstruction.

3.3 Heterogeneous Source Pre-training

Source pre-training jointly uses gesture classification, supervised contrastive learning, cross-dataset gesture alignment, and masked channel-representation reconstruction. The objectives encourage gesture discrimination and robustness to different sensor observations and configurations.

Channel masking. Channel masking is applied to the raw sEMG waveform before STFT encoding. A subset of available channels is masked while at least one real channel remains visible. Masked channels are prevented from contributing as visible key/value tokens in Transformer self-attention. Their output positions remain active and aggregate information from the remaining visible channels for channel-level reconstruction. Pre-STFT masking exposes the encoder to the spectral consequences of missing sensor observations and encourages robustness to incomplete channel configurations. Additional signal augmentations are described in the Appendix.

Gesture classification. A classifier operates on the contextualized CLS representation. Cross-entropy loss $\mathcal{L}_{\mathrm{CE}}$ preserves gesture-discriminative information in the shared representation.

Cross-dataset gesture alignment. Shared gesture labels provide semantic correspondence across heterogeneous datasets. A projection head maps the CLS representation to a normalized contrastive embedding. Samples with the same gesture label from different datasets form cross-dataset positive pairs. The cross-dataset objective $\mathcal{L}_{\mathrm{Cross}}$ aligns these representations across different sensor configurations.

We also use the standard supervised contrastive objective $\mathcal{L}_{\mathrm{SupCon}}$ (Khosla et al., 2020). The objective uses all same-gesture positives in a batch. $\mathcal{L}_{\mathrm{SupCon}}$ promotes overall gesture-level class structure. In contrast, $\mathcal{L}_{\mathrm{Cross}}$ explicitly promotes alignment across datasets. The full contrastive formulation is provided in Appendix A.

Masked channel-representation reconstruction. The reconstruction objective recovers clean latent channel representations from channel-masked observations. The shared STFT encoder and the corresponding dataset-specific adapter produce clean channel-token targets with stop-gradient. The contextualized output retained at each masked channel position uses information from the visible channels. A reconstruction head predicts the corresponding clean channel-token representation from this output.

The reconstruction loss $\mathcal{L}_{\mathrm{Rec}}$ combines cosine distance and mean-squared error. The loss assigns greater weight to explicitly masked channels. The objective encourages the model to infer missing channel-level information from the remaining sensor observations. The complete formulation is provided in the Appendix.

Source pre-training objective. The complete source pre-training objective is

$$\mathcal{L}_{\mathrm{source}} = \mathcal{L}_{\mathrm{CE}} + \lambda_{\mathrm{sup}}\mathcal{L}_{\mathrm{SupCon}} + \lambda_{\mathrm{cross}}\mathcal{L}_{\mathrm{Cross}} + \lambda_{\mathrm{rec}}\mathcal{L}_{\mathrm{Rec}}. \tag{2}$$

The loss weights and implementation details are provided in Appendix B.

3.4 Transfer to an Unseen Target

The target dataset is completely excluded from heterogeneous source pre-training. We transfer the pretrained STFT encoder, channel-index embeddings, Transformer, gesture classifier, projection head, and reconstruction head to the target configuration. A new residual adapter is initialized for the unseen target configuration. All transferred components remain trainable during target adaptation. The source datasets are not revisited during transfer.

Masked target adaptation. Masked target adaptation uses only the target training split. The same raw-waveform channel masking used during source pre-training is applied to the target signals. The pretrained shared components and the new target adapter are jointly updated using gesture classification, supervised contrastive learning, and channel-representation reconstruction.

Only one target dataset is available during this stage. Cross-dataset positive pairs therefore cannot be constructed, and $\mathcal{L}_{\mathrm{Cross}}$ is omitted. The adaptation objective is

$$\mathcal{L}_{\mathrm{adapt}} = \mathcal{L}_{\mathrm{CE}} + \lambda_{\mathrm{sup}}\mathcal{L}_{\mathrm{SupCon}} + \lambda_{\mathrm{rec}}\mathcal{L}_{\mathrm{Rec}}. \tag{3}$$

Supervised fine-tuning. After masked target adaptation, supervised fine-tuning uses clean target training signals without input corruption. The STFT encoder, target adapter, Transformer, and gesture classifier are updated end-to-end. The fine-tuning objective is

$$\mathcal{L}_{\mathrm{FT}} = \mathcal{L}_{\mathrm{CE}}. \tag{4}$$

Model selection uses the target validation split. The held-out test split is used only for final evaluation.

4 Experiments

4.1 Datasets and Protocol

We evaluate our framework on five sEMG datasets with native channel counts ranging from 2 to 16 (Table 1).

We compare the gesture sets across all five datasets and identify five gestures that represent the same hand or wrist movement in every dataset. These five gestures form the shared label space: *Hand Open*, *Hand Close*, *Rest*, *Wrist Flexion*, and *Wrist Extension*. All recordings are resampled to 2,000 Hz and segmented into 0.5-s windows with a 500-ms stride. Each native channel is represented by a log-magnitude STFT over 20–450 Hz. We preserve the native channel count of

Table 1: sEMG datasets used in our experiments. Gestures reports the number of original/shared gesture classes after harmonization.

Data source	Placement	Ch.	Subjects	Gestures	Sampling rate
Wrist-2ch (Jeong et al., 2025)	Wrist	2	9	8/5	1,259 Hz
Mendeley-4ch (Ozdemir et al., 2022)	Forearm	4	40	10/5	2,000 Hz
Forearm-8ch (Jeong et al., 2025)	Forearm	8	9	8/5	1,924 Hz
NinaPro-12ch (Atzori et al., 2014)	Forearm	12	40	18/5	2,000 Hz
GrabMyo-16ch (Jiang et al., 2022)	Forearm	16	43	17/5	2,048 Hz

Table 2: Balanced accuracy (%) across five target datasets. All heterogeneous pre-training methods exclude the target dataset. Best and second-best results are shown in bold and underlined.

Method	Wrist 2ch	Mendeley 4ch	Forearm 8ch	NinaPro 12ch	GrabMyo 16ch	Avg.
Single-Dataset	50.40	51.15	78.44	52.38	80.47	62.57
SAFT-style Supervised	53.53	63.02	83.54	<u>68.35</u>	**85.16**	70.72
Backbone + Supervised	55.13	62.74	85.95	63.35	80.94	69.62
Source-only (Ours)	<u>61.00</u>	<u>63.95</u>	<u>87.76</u>	68.24	81.64	<u>72.52</u>
Source-only + Adapt (Ours)	**62.96**	**65.40**	**90.46**	**69.16**	<u>82.78</u>	**74.15**

each dataset. Full preprocessing details are provided in Appendix B. All evaluations are subject-independent. Wrist-2ch and Forearm-8ch use leave-one-subject-out (LOSO) evaluation. The remaining datasets use subject-wise 5-fold cross-validation. Training, validation, and test participants are disjoint. For source-only transfer, the target dataset is completely excluded from heterogeneous source pre-training. The remaining four datasets form the source collection. Masked adaptation and supervised fine-tuning use only the target training split. Wrist-2ch and Forearm-8ch contain the same nine participants. To prevent cross-configuration subject leakage, we also remove the held-out target participant from the paired source dataset during source pre-training. Source-pretraining checkpoints for these two targets are therefore constructed independently for each LOSO fold.

4.2 Compared Methods

We compare five settings using identical target splits. **Single-Dataset** trains the target model from scratch using only the target training data. **SAFT-style Supervised** uses supervised heterogeneous pre-training followed by target fine-tuning, following the general SAFT paradigm (Olsson et al., 2026). **Backbone + Supervised** uses the proposed backbone with heterogeneous source pre-training based only on gesture classification. The setting controls for the architecture and access to source data. **Source-only (Ours)** uses the complete source pre-training objectives followed directly by supervised target fine-tuning. **Source-only + Adapt (Ours)** additionally performs masked target adaptation before supervised fine-tuning. This setting is our primary transfer method.

4.3 Implementation and Evaluation

Our default model uses 128-dimensional channel tokens and a three-layer Transformer with four attention heads. The same model configuration is used for all target datasets unless otherwise stated. Balanced accuracy is the primary evaluation metric. We additionally report accuracy and macro-F1. The ablation study removes individual components from our Source-only + Adapt method while keeping all other settings fixed. Complete architecture, optimization, augmentation, masking, sampling, and training details are provided in Appendix B.

5 Results

5.1 Main Results

Table 2 reports balanced accuracy across the five target datasets. Our Source-only + Adapt method outperforms Single-Dataset training on all five target datasets. Average balanced accuracy increases from 62.57% to 74.15%, an improvement of 11.58 %p. The gains range from 2.31 %p on GrabMyo-

Table 3: Ablation of our Source-only + Adapt method across five target datasets. Values are balanced accuracy (%). Δ denotes the average change relative to the full model.

Method	Wrist 2ch	Mendeley 4ch	Forearm 8ch	NinaPro 12ch	GrabMyo 16ch	Avg.	Δ
Source-only + Adapt (Ours)	62.96	65.40	90.46	69.16	82.78	74.15	–
w/o Channel Masking	59.24	63.71	87.69	66.19	82.82	71.93	−2.22
w/o Reconstruction	60.29	64.71	90.30	68.38	82.59	73.25	−0.90
w/o Cross-Dataset SupCon	62.11	64.54	89.83	69.69	83.24	73.88	−0.27
w/o Target Adaptation	61.00	63.95	87.76	68.24	81.64	72.52	−1.63

16ch to 16.78 %p on NinaPro-12ch. The target dataset is completely excluded from heterogeneous source pre-training.

The intermediate settings separate the effects of heterogeneous source pre-training and the proposed representation learning objectives. Backbone + Supervised reaches 69.62% average balanced accuracy. Our Source-only method increases the average to 72.52%. The proposed source objectives therefore provide a 2.90 %p gain over Backbone + Supervised.

Masked target adaptation provides a further improvement. Our Source-only + Adapt method reaches 74.15% average balanced accuracy, compared with 72.52% without adaptation. Target adaptation improves performance on all five datasets. The adaptation stage uses only the target training split and does not revisit the source datasets.

SAFT-style Supervised reaches 70.72% average balanced accuracy. Our Source-only + Adapt method achieves a higher average of 74.15%. Our method achieves higher balanced accuracy on Wrist-2ch, Mendeley-4ch, Forearm-8ch, and NinaPro-12ch. SAFT-style Supervised achieves the highest performance on GrabMyo-16ch.

5.2 Ablation Study

Table 3 evaluates the contribution of each component across all five target datasets. Removing channel masking produces the largest average decrease of 2.22 %p. Performance decreases on four of the five target datasets.

Target adaptation also provides a consistent benefit. Removing target adaptation decreases average balanced accuracy by 1.63 %p. Performance decreases on all five target datasets. Removing channel-representation reconstruction decreases the average by 0.90 %p and also reduces performance on every target dataset.

Cross-dataset supervised contrastive learning has a smaller and more target-dependent effect. Removing the objective decreases average balanced accuracy by 0.27 %p. The full model performs better on Wrist-2ch, Mendeley-4ch, and Forearm-8ch. The ablated model performs slightly better on NinaPro-12ch and GrabMyo-16ch.

5.3 Class-wise Analysis

Figure 2 provides a class-wise comparison on Wrist-2ch, Mendeley-4ch, and Forearm-8ch. The three target datasets represent increasing channel counts from 2 to 8. The improvements of our Source-only + Adapt method are distributed across multiple gesture classes rather than a single dominant class. Additional results for NinaPro-12ch and GrabMyo-16ch are provided in the Appendix.

5.4 Target Representation Analysis

Figure 3 visualizes held-out $\mathbf{h}_{\mathrm{CLS}}$ representations for Wrist-2ch, Mendeley-4ch, and Forearm-8ch. The UMAP projections illustrate qualitative differences in gesture-level organization across the compared methods. We further evaluate the representations in the original embedding space. Our Source-only + Adapt method achieves a higher gesture silhouette score than Single-Dataset training on all five target datasets. The cosine-margin results are more target-dependent. The full quantitative

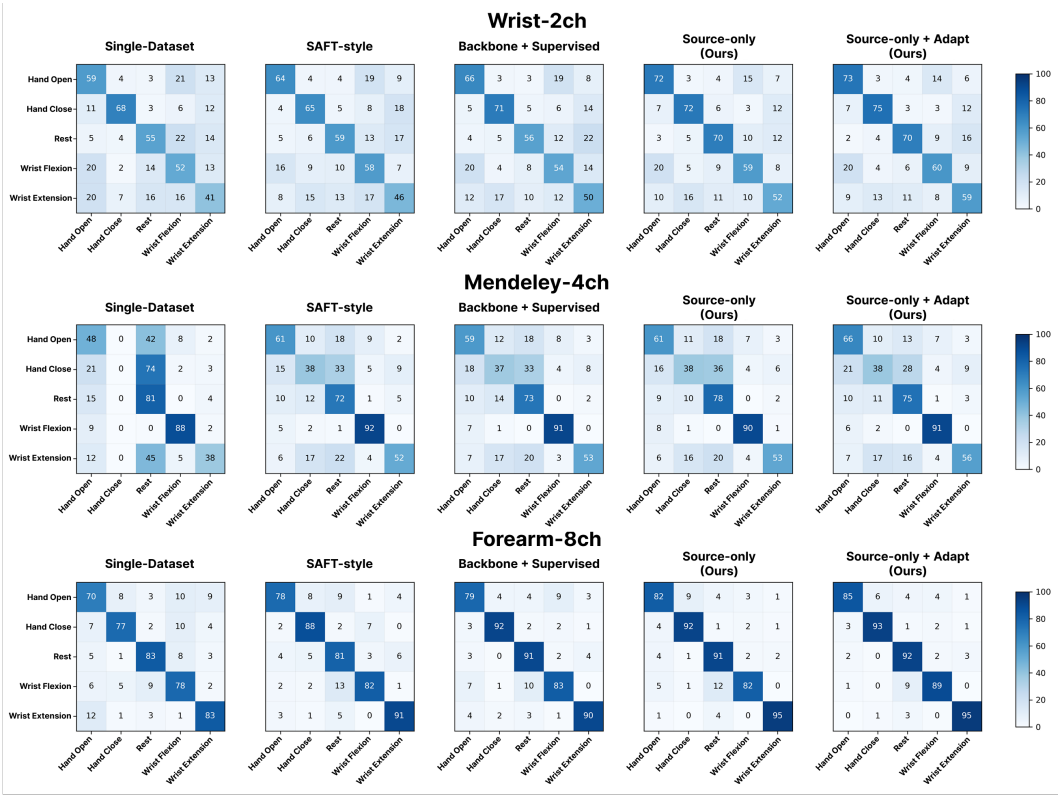

Figure 2: Row-normalized confusion matrices (%) for Wrist-2ch, Mendeley-4ch, and Forearm-8ch across the five compared methods. Rows denote true gesture classes and columns denote predicted classes. Results for NinaPro-12ch and GrabMyo-16ch are provided in the Appendix.

analysis and additional UMAP visualizations for NinaPro-12ch and GrabMyo-16ch are provided in Appendix C.3.

6 Discussion

Our results show that heterogeneous sEMG datasets can provide reusable knowledge for previously unseen sensor configurations. Our Source-only + Adapt method improves over Single-Dataset training across all five target datasets, although the target dataset is completely excluded from heterogeneous source pre-training. The results support the use of shared gesture identity as a bridge between heterogeneous datasets with different sensor configurations.

The ablation results clarify how the proposed components contribute to transfer. Channel masking produces the largest average contribution and reduces the dependence on individual sensor observations. Channel-representation reconstruction provides additional gains across all five target datasets. Masked target adaptation also improves performance on every target dataset by adjusting the source-pretrained representation to the new sensing configuration. Cross-dataset supervised contrastive learning provides a smaller and more target-dependent contribution. Its average effect remains positive, but the benefit is not uniform across datasets.

The representation analysis provides additional evidence beyond classification accuracy. Our Source-only + Adapt method achieves higher silhouette scores than Single-Dataset training across all five target datasets. The cosine-margin results are more target-dependent. These results indicate that improved classification is accompanied by changes in gesture-level representation structure, although no single representation metric fully explains the downstream performance.

The framework does not require explicit anatomical correspondence between electrodes across datasets. Each dataset preserves its native sensor configuration, while dataset-specific adapters ac-

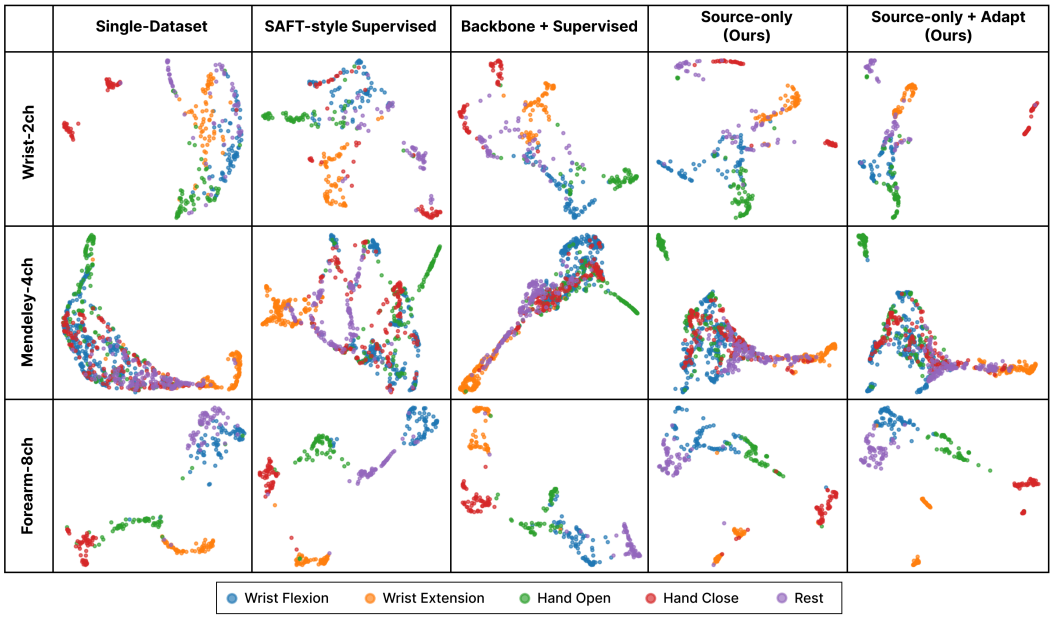

Figure 3: UMAP visualization of held-out target representations for Wrist-2ch, Mendeley-4ch, and Forearm-8ch across the five compared methods. Colors denote gesture classes. UMAP is used only for visualization; quantitative metrics are computed in the original representation space.

count for configuration-dependent variation. The shared encoder and Transformer learn gesture-relevant structure from heterogeneous observations. Shared gesture labels provide semantic correspondence across these otherwise different sensing configurations. The transferable representation therefore operates at the gesture level rather than through a common EMG channel layout.

Several limitations remain. Our experiments use five gestures that can be consistently matched across all datasets. The current evaluation therefore does not address partially overlapping or disjoint gesture vocabularies. The five datasets span channel counts from 2 to 16, but broader hardware, electrode layouts, and acquisition conditions remain to be evaluated. Target adaptation and supervised fine-tuning also require labeled target training data. Future work could extend the framework to partially shared gesture spaces and few-shot, semi-supervised, or unlabeled target adaptation.

7 Conclusion

We presented a cross-dataset representation learning framework for heterogeneous sEMG gesture recognition. The framework uses shared gesture identity as semantic correspondence across datasets with different sensor configurations. Variable-channel encoding, masked channel-representation learning, and cross-dataset gesture alignment enable heterogeneous source pre-training without constructing a common channel layout. Across five datasets spanning 2–16 channels, our Source-only + Adapt method consistently outperforms Single-Dataset training. The target dataset is completely excluded from heterogeneous source pre-training and is used only during target adaptation and supervised fine-tuning. The results demonstrate that shared gesture identity can support representation transfer across heterogeneous sEMG sensor configurations.

AI Use Statement

We used AI in a limited manner for grammar and style polishing only. It did not contribute to research ideation, experimental design, or analysis. The authors take full responsibility for all content.

Reproducibility Statement

We provide the model architecture and training objectives in Section 3, and the datasets and evaluation protocols in Section 4. Appendix A and Appendix B provide additional methodological and implementation details. Additional evaluation and analysis details are provided in Appendix C. Code will be made publicly available upon publication.

References

Ali Ameri, Mohammad Ali Akhaee, Erik Scheme, and Kevin Englehart. A deep transfer learning approach to reducing the effect of electrode shift in emg pattern recognition-based control. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(2):370–379, 2019.

Manfredo Atzori, Arjan Gijsberts, Claudio Castellini, Barbara Caputo, Anne-Gabrielle Mittaz Hager, Simone Elsig, Giorgio Giatsidis, Franco Bassetto, and Henning Müller. Electromyography data for non-invasive naturally-controlled robotic hand prostheses. *Scientific data*, 1(1): 140053, 2014.

Manfredo Atzori, Matteo Cognolato, and Henning Müller. Deep learning with convolutional neural networks applied to electromyography data: A resource for the classification of movements for prosthetic hands. *Frontiers in neurorobotics*, 10:9, 2016.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Martin Colot, Cédric Simar, Ana Maria Cebolla Alvarez, and Gianluca Bontempi. Linear non-conservative unsupervised domain adaptation for cross-subject emg gesture recognition. *Biomedical Signal Processing and Control*, 111:108283, 2026.

Ulysse Côté-Allard, Cheikh Latyr Fall, Alexandre Drouin, Alexandre Campeau-Lecours, Clément Gosselin, Kyrre Glette, François Laviolette, and Benoit Gosselin. Deep learning for electromyographic hand gesture signal classification using transfer learning. *IEEE transactions on neural systems and rehabilitation engineering*, 27(4):760–771, 2019.

Wenhui Cui, Christopher M Sandino, Hadi Pouransari, Ran Liu, Juri Minxha, Ellen Zippi, Erdrin Azemi, and Behrooz Mahasseni. Embridge: Enhancing gesture generalization from emg signals through cross-modal representation learning. In *International Conference on Learning Representations*, volume 2026, pp. 149668–149684, 2026.

Qingfeng Dai, Yongkang Wong, Mohan Kankanhali, Xiangdong Li, and Weidong Geng. Improved network and training scheme for cross-trial surface electromyography (semg)-based gesture recognition. *Bioengineering*, 10(9):1101, 2023.

Yu Du, Wenguang Jin, Wentao Wei, Yu Hu, and Weidong Geng. Surface emg-based inter-session gesture recognition enhanced by deep domain adaptation. *sensors*, 17(3):458, 2017.

Weidong Geng, Yu Du, Wenguang Jin, Wentao Wei, Yu Hu, and Jiajun Li. Gesture recognition by instantaneous surface emg images. *Scientific reports*, 6(1):36571, 2016.

Yao Guo, Jiayan Liu, Yonglin Wu, Xinyu Jiang, Yalin Wang, Long Meng, Xiangyu Liu, Feng Shu, Chenyun Dai, and Wei Chen. semg-based inter-session hand gesture recognition via domain adaptation with locality preserving and maximum margin. *International Journal of Neural Systems*, 34(03):2450010, 2024.

Jiayuan He, Xinyue Niu, Penghui Zhao, Chuang Lin, and Ning Jiang. From forearm to wrist: Deep learning for surface electromyography-based gesture recognition. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 32:102–111, 2023.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 15979–15988. IEEE, 2022.

Takayuki Hoshino, Suguru Kanoga, Masashi Tsubaki, and Atsushi Aoyama. Comparing subject-to-subject transfer learning methods in surface electromyogram-based motion recognition with shallow and deep classifiers. *Neurocomputing*, 489:599–612, 2022.

Zhenghao Huang, Huilin Yao, Kaikai Wang, and Lin Shu. Reading your actions: Learning generalizable action representations via pre-training aemg. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20108–20117, 2026.

Andrés Jaramillo-Yánez, Marco E Benalcázar, and Elisa Mena-Maldonado. Real-time hand gesture recognition using surface electromyography and machine learning: A systematic literature review. *Sensors*, 20(9):2467, 2020.

Hanseok Jeong, Fangqing Li, and Sang Ho Yoon. Two-channel emg and imu-based hand gesture classification method using wearable wristband. *KIISE Transactions on Computing Practices*, 31: 135–140, 03 2025. doi: 10.5626/KTCP.2025.31.3.135.

Ning Jiang, A Pradhan, and J He. Gesture recognition and biometrics electromyogram (grabmyo). *PhysioNet. doi: https://doi. org/10.13026/dzhr-nj10*, 2022.

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.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Jianfeng Li, Xinyu Jiang, Jiahao Fan, Yanjuan Geng, Fumin Jia, and Chenyun Dai. Deep end-to-end transfer learning for robust inter-subject and inter-day hand gesture recognition using surface emg. *Biomedical Signal Processing and Control*, 100:106892, 2025.

Wei Li, Ping Shi, and Hongliu Yu. Gesture recognition using surface electromyography and deep learning for prostheses hand: state-of-the-art, challenges, and future. *Frontiers in neuroscience*, 15:621885, 2021.

Mansooreh Montazerin, Elahe Rahimian, Farnoosh Naderkhani, S Farokh Atashzar, Svetlana Yanushkevich, and Arash Mohammadi. Transformer-based hand gesture recognition from instantaneous to fused neural decomposition of high-density emg signals. *Scientific reports*, 13(1): 11000, 2023.

Sike Ni, Mohammed AA Al-qaness, Ammar Hawbani, Dalal Al-Alimi, Mohamed Abd Elaziz, and Ahmed A Ewees. A survey on hand gesture recognition based on surface electromyography: Fundamentals, methods, applications, challenges and future trends. *Applied Soft Computing*, 166:112235, 2024.

Alexander E Olsson, Nebojša Maleševic, Anders Björkman, and Christian Antfolk. Calibration-free electromyography motor intent decoding using large-scale supervised pretraining. *Advanced Intelligent Systems*, pp. e202500791, 2026.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Mehmet Akif Ozdemir, Deniz Hande Kisa, Onan Guren, and Aydin Akan. Dataset for multi-channel surface electromyography (semg) signals of hand gestures. *Data in brief*, 41:107921, 2022.

Lizhi Pan, Dingguo Zhang, Ning Jiang, Xinjun Sheng, and Xiangyang Zhu. Improving robustness against electrode shift of high density emg for myoelectric control through common spatial patterns. *Journal of neuroengineering and rehabilitation*, 12(1):110, 2015.

Shriram Tallam Puranam Raghu, Dawn T MacIsaac, and Erik J Scheme. Self-supervised learning via vicreg enables training of emg pattern recognition using continuous data with unclear labels. *Computers in Biology and Medicine*, 185:109479, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Kang Wang, Yiqiang Chen, Yingwei Zhang, Xiaodong Yang, and Chunyu Hu. Iterative self-training based domain adaptation for cross-user semg gesture recognition. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:2974–2987, 2023.

Runsheng Wang, Xinyue Zhu, Ava Chen, Jingxi Xu, Lauren Winterbottom, Dawn M Nilsen, Joel Stein, and Matei Ciocarlie. Reactemg: Stable, low-latency intent detection from semg via masked modeling. *arXiv preprint arXiv:2506.19815*, 2025.

Zihao Wang, Huiying Wan, Long Meng, Zheng Zeng, Metin Akay, Chen Chen, and Wei Chen. Optimization of inter-subject semg-based hand gesture recognition tasks using unsupervised domain adaptation techniques. *Biomedical Signal Processing and Control*, 92:106086, 2024.

Dezhen Xiong, Daohui Zhang, Xingang Zhao, and Yiwen Zhao. Deep learning for emg-based human-machine interaction: A review. *IEEE/CAA Journal of Automatica Sinica*, 8(3):512–533, 2021.

Hao Xu and Anbin Xiong. Advances and disturbances in semg-based intentions and movements recognition: A review. *IEEE Sensors Journal*, 21(12):13019–13028, 2021.

Jaemo Yang, Doheun Cha, Dong-Gyu Lee, and Sangtae Ahn. Stcnet: Spatio-temporal cross network with subject-aware contrastive learning for hand gesture recognition in surface emg. *Computers in Biology and Medicine*, 185:109525, 2025.

Aaron J Young, Levi J Hargrove, and Todd A Kuiken. The effects of electrode size and orientation on the sensitivity of myoelectric pattern recognition systems to electrode shift. *IEEE transactions on biomedical engineering*, 58(9):2537–2544, 2011a.

Aaron J Young, Levi J Hargrove, and Todd A Kuiken. Improving myoelectric pattern recognition robustness to electrode shift by changing interelectrode distance and electrode configuration. *IEEE Transactions on Biomedical Engineering*, 59(3):645–652, 2011b.

Wenli Zhang, Tingsong Zhao, Jianyi Zhang, and Yufei Wang. Lst-emg-net: Long short-term transformer feature fusion network for semg gesture recognition. *Frontiers in Neurorobotics*, 17: 1127338, 2023.

Zhen Zhang, Shilong Liu, Yanyu Wang, Wei Song, and Yuhui Zhang. Online cross session electromyographic hand gesture recognition using deep learning and transfer learning. *Engineering Applications of Artificial Intelligence*, 127:107251, 2024.

A METHODOLOGICAL DETAILS

A.1 FORMAL PROBLEM DEFINITION

We consider a collection of K heterogeneous source datasets $\{\mathcal{D}_k\}_{k=1}^{K}$, where

$$\mathcal{D}_k = \{(\mathbf{X}_i^{(k)}, y_i^{(k)})\}_{i=1}^{N_k}, \qquad \mathbf{X}_i^{(k)} \in \mathbb{R}^{C_k \times T}. \tag{5}$$

Here, C_k denotes the number of available sEMG channels in dataset k, T denotes the temporal window length, and $y_i^{(k)}$ denotes the gesture label. The datasets may differ in channel count, electrode configuration, hardware, participants, and acquisition protocol.

We define a shared gesture label space $\mathcal{Y}_{\text{shared}}$ containing gestures that can be consistently matched across the participating datasets. Samples assigned to the same shared label are treated as observations of the same underlying gesture despite differences in sensing configuration.

For source-only transfer, a target dataset $\mathcal{D}_{\text{tgt}}$ is completely excluded from heterogeneous source pre-training. The remaining datasets form the source collection. After source pre-training, the learned representation is transferred to the target using only its training split.

A.2 VARIABLE-CHANNEL REPRESENTATION

For each available channel, the shared spectral encoder produces a D-dimensional embedding

$$\mathbf{e}_{i,c} = E_{\text{STFT}}(\mathbf{x}_{i,c}), \qquad \mathbf{e}_{i,c} \in \mathbb{R}^D. \tag{6}$$

The same encoder E_{STFT} is applied independently to every available channel across all datasets. A dataset-specific residual adapter then transforms each channel embedding:

$$\tilde{\mathbf{e}}_{i,c} = A_{d_i}(\mathbf{e}_{i,c}), \tag{7}$$

where d_i denotes the dataset identity.

Channel-index embeddings are added to the adapted channel embeddings. A learnable CLS token is prepended to form the Transformer input. The shared Pre-LayerNorm Transformer produces

$$\mathbf{H}_i = [\mathbf{h}_{i,\text{CLS}}; \mathbf{h}_{i,1}; \ldots; \mathbf{h}_{i,C_i}]. \tag{8}$$

The CLS representation $\mathbf{h}_{i,\text{CLS}}$ provides the sample-level representation used for gesture classification and contrastive learning. The channel-level outputs $\mathbf{h}_{i,c}$ are used for channel-representation reconstruction.

Samples are dynamically padded along the channel dimension within each mini-batch. The maximum channel count in the current mini-batch determines the padded sequence length. A validity mask identifies real and padded channel positions. Padded positions are excluded from Transformer self-attention through the key-padding mask. This design preserves the native channel count of each dataset without constructing a common fixed-channel signal layout.

A.3 CHANNEL MASKING

Channel masking is applied to the normalized raw sEMG waveform before STFT encoding. For each training sample, a subset of available channels is selected for masking. The selected waveforms are replaced by zeros, while at least one physical channel remains visible.

The Transformer key-padding mask prevents masked channel positions from contributing as key/value observations in self-attention. Their query positions remain active. The Transformer can therefore produce contextualized outputs at masked channel positions using information from the CLS token and the remaining visible channels. These outputs are used to reconstruct the corresponding clean channel-token representations.

Channel masking is applied during source pre-training and masked target adaptation. No temporal masking is used in the final model.

A.4 Gesture Classification

The contextualized CLS representation is passed to a gesture classifier. The classification objective is standard cross-entropy:

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{N}\sum_{i=1}^{N} \log p(y_i \mid \mathbf{h}_{i,\mathrm{CLS}}). \tag{9}$$

A.5 Cross-Dataset Gesture Alignment

The CLS representation is mapped through a projection head and ℓ_2-normalized:

$$\mathbf{z}_i = \frac{g_{\mathrm{proj}}(\mathbf{h}_{i,\mathrm{CLS}})}{\|g_{\mathrm{proj}}(\mathbf{h}_{i,\mathrm{CLS}})\|_2}. \tag{10}$$

We use standard supervised contrastive learning $\mathcal{L}_{\mathrm{SupCon}}$, where all samples with the same gesture label form positive pairs.

We additionally define a cross-dataset positive set for anchor i:

$$\mathcal{P}_{\mathrm{cross}}(i) = \{j \neq i \mid y_j = y_i,\ d_j \neq d_i\}. \tag{11}$$

The cross-dataset supervised contrastive objective is

$$\mathcal{L}_{\mathrm{Cross}} = -\frac{1}{|\mathcal{I}_{\mathrm{cross}}|}\sum_{i\in\mathcal{I}_{\mathrm{cross}}} \frac{1}{|\mathcal{P}_{\mathrm{cross}}(i)|}\sum_{p\in\mathcal{P}_{\mathrm{cross}}(i)} \log \frac{\exp(\mathbf{z}_i^\top \mathbf{z}_p/\tau)}{\sum_{a\neq i}\exp(\mathbf{z}_i^\top \mathbf{z}_a/\tau)}, \tag{12}$$

where τ is the contrastive temperature. $\mathcal{I}_{\mathrm{cross}}$ contains anchors with at least one valid cross-dataset positive. This objective explicitly aligns samples that share gesture identity but originate from different datasets.

A.6 Channel-Representation Reconstruction

For each physical channel, the reconstruction target is computed from the corresponding clean waveform. The shared STFT encoder and dataset-specific adapter produce the clean channel-token target:

$$\mathbf{r}_{i,c}^{\mathrm{clean}} = \mathrm{sg}\left[A_{d_i}\left(E_{\mathrm{STFT}}(\mathbf{x}_{i,c})\right)\right], \tag{13}$$

where $\mathrm{sg}[\cdot]$ denotes stop-gradient. The reconstruction target is therefore a clean latent channel representation rather than the raw waveform.

Given the channel-masked input, the Transformer retains an output at each physical channel position. The reconstruction head predicts the clean channel-token target from the corresponding contextualized output:

$$\hat{\mathbf{r}}_{i,c} = g_{\mathrm{rec}}(\mathbf{h}_{i,c}). \tag{14}$$

The channel-level reconstruction error is

$$\ell_{\mathrm{rec}}(i,c) = 1 - \cos(\hat{\mathbf{r}}_{i,c}, \mathbf{r}_{i,c}^{\mathrm{clean}}) + \alpha\left\|\hat{\mathbf{r}}_{i,c} - \mathbf{r}_{i,c}^{\mathrm{clean}}\right\|_2^2. \tag{15}$$

Let $\mathcal{V}_i$ denote the set of physical channels in sample i. The reconstruction objective is

$$\mathcal{L}_{\mathrm{Rec}} = \frac{\sum_i \sum_{c \in \mathcal{V}_i} w_{i,c}\, \ell_{\mathrm{rec}}(i,c)}{\sum_i \sum_{c \in \mathcal{V}_i} w_{i,c}}, \tag{16}$$

where $w_{i,c} = 3$ for explicitly masked channels and $w_{i,c} = 1$ for visible physical channels. Reconstruction is therefore evaluated over all real channels, with greater emphasis on recovering representations at masked positions.

A.7 Source Pre-training and Target Transfer

Source pre-training jointly optimizes

$$\mathcal{L}_{\mathrm{source}} = \mathcal{L}_{\mathrm{CE}} + \lambda_{\mathrm{rec}}\mathcal{L}_{\mathrm{Rec}} + \lambda_{\mathrm{sup}}\mathcal{L}_{\mathrm{SupCon}} + \lambda_{\mathrm{cross}}\mathcal{L}_{\mathrm{Cross}}. \tag{17}$$

For an unseen target configuration, we transfer the shared STFT encoder, channel-index embeddings, Transformer, classifier, projection head, and reconstruction head. A new residual adapter is initialized for the target dataset.

Masked target adaptation uses only the target training split. The adaptation objective is

$$\mathcal{L}_{\mathrm{adapt}} = \mathcal{L}_{\mathrm{CE}} + \lambda_{\mathrm{rec}}\mathcal{L}_{\mathrm{Rec}} + \lambda_{\mathrm{sup}}\mathcal{L}_{\mathrm{SupCon}}. \tag{18}$$

The cross-dataset objective is not used during target adaptation because only one target dataset is present. All transferred components remain trainable during adaptation.

Final supervised fine-tuning uses clean target signals and optimizes $\mathcal{L}_{\mathrm{CE}}$. The STFT encoder, target adapter, Transformer, and classifier are updated end-to-end. The original source datasets are not revisited during either target adaptation or supervised fine-tuning.

B Experimental Details

B.1 Datasets and Gesture Harmonization

Wrist-2ch and Forearm-8ch contain sEMG recordings from the same nine participants (Jeong et al., 2025). Wrist-2ch contains two channels placed around the wrist and was originally sampled at 1,259 Hz. Forearm-8ch contains eight channels distributed around the forearm and was originally sampled at 1,924 Hz. Both datasets contain eight original gesture classes.

Mendeley-4ch (Ozdemir et al., 2022) contains 4-channel sEMG recordings from 40 participants sampled at 2,000 Hz and contains 10 original gesture classes. NinaPro DB2-12ch (Atzori et al., 2014) contains 12-channel recordings from 40 participants sampled at 2,000 Hz. GrabMyo-16ch (Jiang et al., 2022) contains 16-channel recordings from 43 participants sampled at 2,048 Hz.

To construct a common gesture label space, we compare the original gesture sets of the five datasets and identify movements with the same hand or wrist posture across all datasets. Five gestures satisfy this criterion: *Hand Open*, *Hand Close*, *Rest*, *Wrist Flexion*, and *Wrist Extension*. We map the corresponding dataset-specific gesture labels to these five shared classes. All other gesture classes are excluded because they do not have a corresponding movement in every dataset. The same mapping is used for every compared method.

B.2 Signal Preprocessing

All recordings are resampled to 2,000 Hz. Each recording is segmented into non-overlapping 0.5-s windows with a 500-ms stride. Each window therefore contains 1,000 samples per channel.

Signal normalization. Each 500-ms window is independently normalized per channel using a robust z-score based on the median and median absolute deviation (MAD):

$$\tilde{\mathbf{x}}_c = \frac{\mathbf{x}_c - \mathrm{median}(\mathbf{x}_c)}{1.4826\ \mathrm{MAD}(\mathbf{x}_c) + \epsilon}. \tag{19}$$

The normalized waveform is clipped to $[-5, 5]$. Channel masking and signal augmentations are applied after this waveform normalization and before spectral encoding.

Spectral representation. For each available channel, we compute an STFT using a 128-sample Hann window, an FFT size of 128, and a hop length of 32 samples. Frequency bins between 20 and 450 Hz are retained. We apply a log-magnitude transformation using $\log(1 + x)$. Each resulting channel spectrogram is then standardized using its own mean and standard deviation over the time–frequency bins before convolutional encoding.

Each dataset retains its native number of channels. No artificial channels are introduced, and recordings are not converted to a common fixed-channel signal layout.

B.3 Subject-Independent Evaluation

All evaluation protocols are subject-independent. Wrist-2ch and Forearm-8ch use leave-one-subject-out evaluation over their nine participants. Mendeley-4ch, NinaPro DB2-12ch, and GrabMyo-16ch use subject-wise 5-fold cross-validation. Training, validation, and test participants are disjoint.

Wrist-2ch and Forearm-8ch were collected from the same participants. For these two target datasets, the held-out target participant is also removed from the paired source dataset during heterogeneous source pre-training. This procedure prevents participant leakage across the two sensing configurations. A separate source-pretraining checkpoint is therefore constructed for each Wrist-2ch and Forearm-8ch LOSO fold.

B.4 Architecture and Optimization

The shared channel encoder produces 128-dimensional channel tokens. Each dataset-specific residual adapter uses a bottleneck dimension of 64. The shared Transformer contains three Pre-LayerNorm Transformer blocks with four attention heads. The feed-forward dimension is 512, and dropout is 0.1. The contrastive projection head produces a 64-dimensional normalized embedding.

The channel masking ratio is 0.35. Amplitude jitter of 0.15 and signal-relative Gaussian noise with scale 0.02 are applied before STFT encoding. The reconstruction objective uses $\alpha = 0.25$. Explicitly masked channels receive three times the reconstruction weight of visible channels. The contrastive temperature is $\tau = 0.1$.

We use $\lambda_{\mathrm{rec}} = 1.0$, $\lambda_{\mathrm{sup}} = 0.1$, and $\lambda_{\mathrm{cross}} = 0.1$.

Source pre-training is performed for 35 epochs with a learning rate of 5×10^{-4}. During source pre-training, source datasets are sampled uniformly, while gesture classes are balanced within each source dataset using a weighted random sampler.

Masked target adaptation is performed for 15 epochs with a learning rate of 1×10^{-4}. Final target fine-tuning is performed for up to 50 epochs with a learning rate of 1×10^{-4}. We use AdamW with weight decay 10^{-4}, gradient clipping at 5.0, and a mini-batch size of 128. Early stopping and model selection use validation balanced accuracy. Unless otherwise stated, experiments use random seed 42.

C Additional Analyses

C.1 Sensitivity to the Number of Source Datasets

We examine how transfer performance changes as the number of heterogeneous source datasets increases. For four target datasets, we vary the number of source datasets used during source pre-training from one to four. For each source count, we report the mean and standard deviation across the corresponding source-dataset combinations when multiple combinations are available.

Table 4: Sensitivity to the number of source datasets. Values are balanced accuracy (%). Results are averaged across the corresponding source-dataset combinations when multiple combinations are available.

# Sources	Wrist-2ch	Mendeley-4ch	Forearm-8ch	NinaPro-12ch	4-target Avg.
1	57.10 ± 3.74	62.40 ± 2.38	86.99 ± 2.03	66.91 ± 2.05	68.35
2	59.38 ± 2.34	63.95 ± 0.49	89.05 ± 1.09	67.99 ± 1.63	70.09
3	60.48 ± 1.64	64.57 ± 0.62	89.62 ± 0.55	68.87 ± 0.46	70.89
4	61.38	65.40	89.95	69.16	71.47

Table 5: Sensitivity to window stride for our Source-only + Adapt method. The window length is fixed at 500 ms. Values are balanced accuracy (%).

Stride	Overlap	Wrist	Mendeley	Forearm	NinaPro	GrabMyo	Avg.
100 ms	80%	60.26	64.29	**92.63**	**74.71**	**85.50**	**75.48**
250 ms	50%	58.96	**65.83**	91.95	72.46	84.10	74.66
500 ms	0%	**62.96**	65.40	90.46	69.16	82.78	74.15

Performance improves consistently as more heterogeneous source datasets are included. The four-target average balanced accuracy increases from 68.35% with one source dataset to 71.47% with four source datasets. The same increasing trend is observed for all four evaluated target datasets. These results show that increasing the number of heterogeneous source datasets provides additional benefit for cross-dataset transfer.

C.2 Sensitivity to Window Stride

We evaluate the sensitivity of our Source-only + Adapt method to the temporal stride used for window segmentation. The window length remains fixed at 500 ms. Table 5 compares strides of 100, 250, and 500 ms, corresponding to overlaps of 80%, 50%, and 0%.

Performance remains strong across all three settings. The average balanced accuracy ranges from 74.15% to 75.48%. The best stride varies across target datasets. These results show that the proposed transfer framework does not depend on highly overlapping input windows.

C.3 Full Target Representation Analysis

We evaluate the held-out target representations in the original embedding space using the cosine margin and gesture silhouette score. Table 6 reports mean and standard deviation across evaluation folds.

Our Source-only + Adapt method achieves a higher silhouette score than Single-Dataset training on all five target datasets. The cosine-margin results are more target-dependent. Our method achieves the highest cosine margin on Wrist-2ch, Mendeley-4ch, and Forearm-8ch. Single-Dataset produces the highest margin on NinaPro-12ch, while SAFT-style Supervised produces the highest margin on GrabMyo-16ch. The results indicate that no single representation metric completely explains downstream classification performance.

C.4 Representation Metrics

Let $\mathbf{q}_i$ denote the normalized representation of sample i and $\boldsymbol{\mu}_c$ the normalized centroid of gesture class c. The cosine margin is

$$m_i = \mathrm{sim}(\mathbf{q}_i, \boldsymbol{\mu}_{y_i}) - \max_{c \neq y_i} \mathrm{sim}(\mathbf{q}_i, \boldsymbol{\mu}_c), \tag{20}$$

where $\mathrm{sim}(\cdot, \cdot)$ denotes cosine similarity. We compute the mean margin within each evaluation fold and report the mean and standard deviation across folds.

Table 6: Representation quality across five target datasets. Values are mean ± standard deviation across held-out evaluation folds. Higher values indicate stronger gesture-level separation.

Target	Method	Cosine Margin ↑	Silhouette Score ↑
Wrist-2ch	Single-Dataset	0.334 ± 0.156	0.112 ± 0.125
	SAFT-style Supervised	0.227 ± 0.091	0.158 ± 0.130
	Backbone + Supervised	0.241 ± 0.101	0.168 ± 0.090
	Source-only (Ours)	0.324 ± 0.128	0.187 ± 0.095
	Source-only + Adapt (Ours)	$\mathbf{0.387 \pm 0.191}$	$\mathbf{0.211 \pm 0.133}$
Mendeley-4ch	Single-Dataset	0.444 ± 0.059	0.061 ± 0.052
	SAFT-style Supervised	0.427 ± 0.079	0.149 ± 0.057
	Backbone + Supervised	0.389 ± 0.062	0.133 ± 0.060
	Source-only (Ours)	0.436 ± 0.078	0.175 ± 0.072
	Source-only + Adapt (Ours)	$\mathbf{0.446 \pm 0.077}$	$\mathbf{0.192 \pm 0.083}$
Forearm-8ch	Single-Dataset	0.618 ± 0.121	0.388 ± 0.109
	SAFT-style Supervised	0.713 ± 0.188	0.528 ± 0.196
	Backbone + Supervised	0.702 ± 0.128	0.516 ± 0.139
	Source-only (Ours)	0.756 ± 0.143	0.576 ± 0.150
	Source-only + Adapt (Ours)	$\mathbf{0.834 \pm 0.129}$	$\mathbf{0.637 \pm 0.139}$
NinaPro-12ch	Single-Dataset	$\mathbf{0.618 \pm 0.064}$	0.206 ± 0.032
	SAFT-style Supervised	0.559 ± 0.056	0.296 ± 0.014
	Backbone + Supervised	0.557 ± 0.055	0.272 ± 0.026
	Source-only (Ours)	0.583 ± 0.060	0.298 ± 0.033
	Source-only + Adapt (Ours)	0.603 ± 0.022	$\mathbf{0.326 \pm 0.019}$
GrabMyo-16ch	Single-Dataset	0.558 ± 0.029	0.357 ± 0.038
	SAFT-style Supervised	$\mathbf{0.637 \pm 0.043}$	$\mathbf{0.450 \pm 0.037}$
	Backbone + Supervised	0.497 ± 0.052	0.353 ± 0.033
	Source-only (Ours)	0.573 ± 0.043	0.393 ± 0.038
	Source-only + Adapt (Ours)	0.604 ± 0.029	0.426 ± 0.030

Gesture silhouette is computed in the original representation space using cosine distance. For large evaluation sets, we use a class-balanced subset of at most 4,000 held-out samples per fold. The same deterministic sampling procedure is applied across methods.

UMAP is used only for visualization. All quantitative representation metrics are computed in the original high-dimensional representation space.

C.5 Additional UMAP Visualizations

Figure 4 shows the held-out target representations for NinaPro-12ch and GrabMyo-16ch. UMAP is fitted independently for each method. Absolute coordinates should therefore not be compared across panels. The visualization is used only to examine qualitative gesture-level organization.

C.6 Additional Class-wise Results

Figure 5 provides the class-wise results for NinaPro-12ch and GrabMyo-16ch. These results complement the Wrist-2ch, Mendeley-4ch, and Forearm-8ch confusion matrices reported in the main paper.

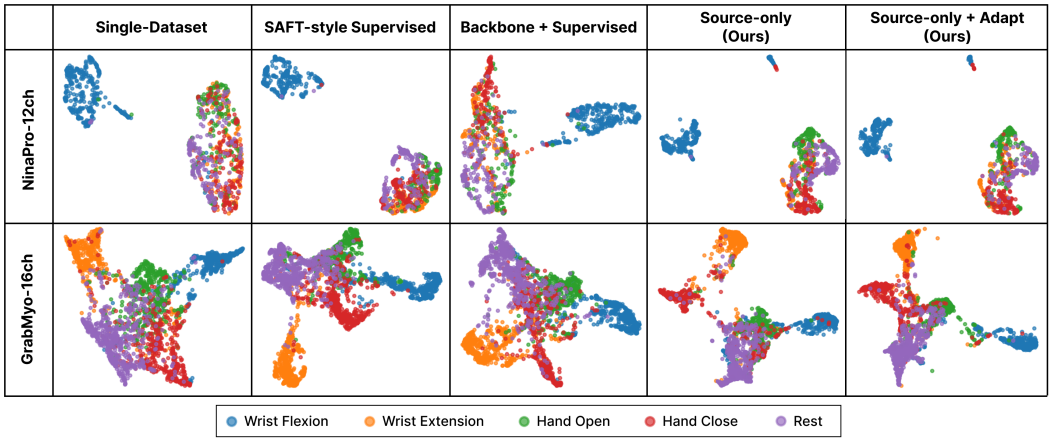

Figure 4: UMAP visualization of held-out target representations for NinaPro-12ch and GrabMyo-16ch across the five compared methods. Colors denote gesture classes. UMAP is used only for visualization; quantitative representation metrics are computed in the original embedding space.

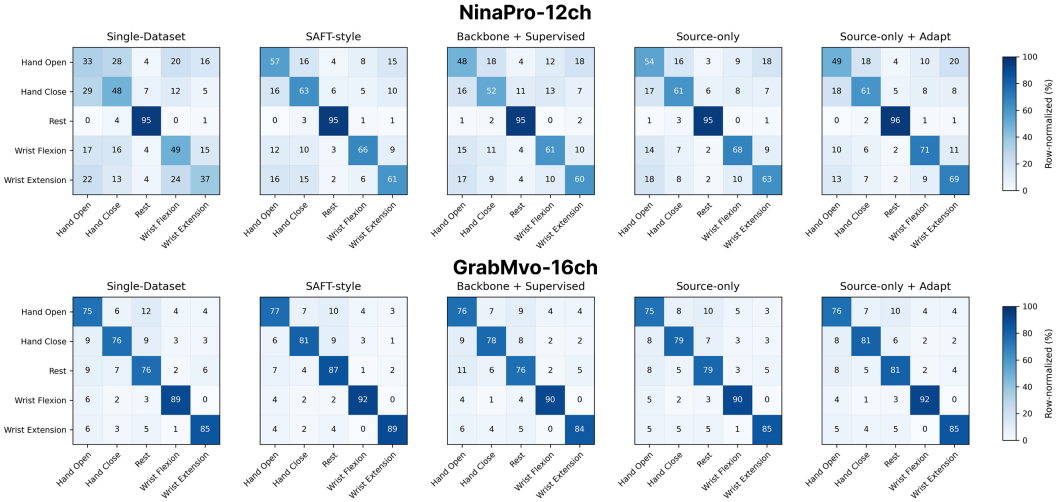

Figure 5: Row-normalized confusion matrices (%) for NinaPro-12ch and GrabMyo-16ch across the five compared methods. Rows denote true gesture classes and columns denote predicted classes.