# CPEP: Contrastive Pose-EMG Pre-training Enhances Gesture Generalization on EMG Signals

**Wenhui Cui**[1,2*] **Christopher Sandino**[1], **Hadi Pouransar**[1], **Ran Liu**[1], **Juri Minxha**[1],

**Ellen L. Zippi**[1], **Aman Verma**[1], **Anna Sedlackova**[1], **Erdrin Azemi**[1], **Behrooz Mahasseni**[1]

[1]Apple
{csandino, mpouransari, ran_liu3, j_minxha, ezippi,
amanverma, asedlackova, erdrin, bmahasseni}@apple.com
[2]Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California
wenhuicu@usc.edu

## Abstract

Hand gesture classification using high-quality structured data such as videos, images, and hand skeletons is a well-explored problem in computer vision. Leveraging low-power, cost-effective biosignals, e.g. surface electromyography (sEMG), allows for continuous gesture prediction on wearables. In this paper, we demonstrate that learning representations from weak-modality data that are aligned with those from structured, high-quality data can improve representation quality and enables zero-shot classification. Specifically, we propose a **C**ontrastive **P**ose-**E**MG **P**re-training (**CPEP**) framework to align EMG and pose representations, where we learn an EMG encoder that produces high-quality and pose-informative representations. We assess the gesture classification performance of our model through linear probing and zero-shot setups. Our model outperforms *emg2pose* benchmark models by up to **21%** on in-distribution gesture classification and **72%** on unseen (out-of-distribution) gesture classification.

## 1 Introduction

Biosignals, such as inertial measurement unit (IMU) and surface electromyography (sEMG), have been widely applied in rehabilitation, health monitoring, human activity recognition, and robotic control due to their low power requirements and ease of integration into wearable devices [7, 6, 3, 4]. Human gesture recognition on wearables has recently attracted significant interest and demonstrated potential across various scenarios [15, 22, 13]. However, predicting hand gestures from biosignals like sEMG, especially generalizing to unseen gestures, remains challenging [8], primarily because of the high variability and fine dexterity of human hand movements, sensor noise, and/or the limited scale of publicly available data. Thanks to recent advances in deep learning and the availability of large-scale visual data, such as videos and motion capture data [1], vision-based models have achieved remarkable success [14]. Despite these advances, cameras' high power demands and privacy concerns hinder deployment on wearables, and pose occlusions can destabilize vision-based gesture classification. This has driven a recent rise in the use of low-power sensors, such as IMUs and EMGs, for real-time gesture recognition [2, 19]. However, collecting large-scale datasets using biosignals in unconstrained environments remains challenging and expensive [18]. To overcome this, we introduce

[*] Work completed during internship at Apple.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Foundation Models for the Brain and Body.

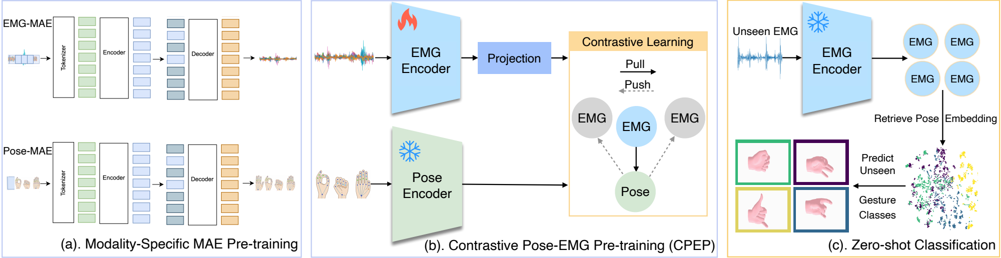

Figure 1: Overview of Contrastive Pose–EMG Pre-training (CPEP) framework. (a). MAE pre-trains EMG and pose encoders; (b). CPEP freezes the pose encoder and aligns EMG to pose via contrastive learning, enabling (c). zero-shot gesture classification by retrieval over frozen pose embeddings.

a representation alignment approach that allows the model to generalize to unseen gestures and new users at test time without the need for additional training or extensive data collection.

Through our experiments, we found that due to the inherently noisy nature of EMG signals, conventional self-supervised pre-training methods are suboptimal and fail to achieve high-quality gesture classification. We also found that supervised training approaches, which directly predict poses from EMG signals [17], have limited generalization capacity to unseen gestures or users. Therefore, we explore a contrastive pre-training strategy in this work. Aligning image–text representations through contrastive pre-training has been shown to enhance the quality of both text and image embeddings [16, 9]. Unlike CLIP [16] model, which is trained on billions of image–text pairs, we align pre-trained unimodal encoders to reduce the amount of paired data required. We introduce **C**ontrastive **P**ose-**E**MG **P**re-training (**CPEP**), a method that enables zero-shot classification of new gestures without the need for task- or data-specific fine-tuning. Specifically, we use two encoders: (a) an EMG encoder, pre-trained on unlabeled EMG signals via self-supervised learning, and (b) a pose encoder, pre-trained on unlabeled pose data to produce high-quality pose representations. A contrastive learning objective then aligns the EMG representation with its corresponding pose representation in the embedding space, while separating mismatched pairs. CPEP encourages the EMG encoder to capture pose-relevant information and develop a rich understanding of diverse hand kinematics, enabling generalization to unseen gestures by extrapolating within the structured latent space. In CPEP, pose encoder provides richer supervisory signals by capturing structural and semantic relationships that are difficult to learn from the weak modality (EMG) alone, thus guiding the weak-modality encoder toward more discriminative and generalizable representations.

We utilize the large-scale public EMG dataset *emg2pose* [17], which provides simultaneous paired EMG and pose recordings to pre-train our CPEP model and perform downstream evaluations. Leveraging the metadata available in the emg2pose dataset, we design a gesture classification task to evaluate EMG representation quality. Following CLIP evaluation protocol [16], we validate the learned EMG representations through zero-shot classification and linear probing, demonstrating superior performance on both in-distribution and unseen gestures compared to *emg2pose* benchmark models. In summary, our contributions are two-fold: (i) We propose a contrastive pre-training strategy to enhance EMG representation learning from noisy EMG signals by aligning them with high-quality pose representations. (ii) To the best of our knowledge, CPEP is the first framework enabling zero-shot classification of unseen gestures from EMG signals.

## 2 Methods

In this section, we formalize the problem, describe the pre-training of unimodal EMG and pose encoders, and present our contrastive pre-training strategy CPEP for aligning Pose-EMG representations. An overview of the full pipeline is shown in Fig. 1.

### 2.1 Problem Definition

**Definition 1** ***EMG, Pose data, and Gesture classes.*** *Let $\mathcal{X} = \{\mathbf{x}_i \in \mathbb{R}^{C\times T}\}_{i=1}^{N}$ denote multi-channel EMG sequences with $C$ channels and time length $T$, and let $\mathcal{P} = \{\mathbf{p}_i \in \mathbb{R}^{J\times T}\}_{i=1}^{N}$*

*denote the paired pose sequences of joint angles, where $J$ is the number of joints in a pre-defined hand-skeleton [17]. Here, $N$ is the total number of paired samples.*

*Let $\mathcal{Y} = \{1, \ldots, K\}$ denote the gesture classes, and let $\{y_i\}_{i=1}^N$ be the labels with $y_i \in \mathcal{Y}$. Each pose $p_i$ has a unique label $y_i$, although multiple samples may share the same label. The paired dataset is*

$$\mathcal{D} = \{(x_i, p_i, y_i)\}_{i=1}^N,$$

*where $x_i$ and $p_i$ are a paired EMG–pose sample and $y_i$ is its gesture label. We split $\mathcal{D}$ into $\mathcal{D}_{\mathrm{tr}}$, $\mathcal{D}_{\mathrm{val}}$, and $\mathcal{D}_{\mathrm{test}}$. Let $\mathcal{Y}_{\mathrm{in}} \subset \mathcal{Y}$ be the in-distribution gesture classes and $\mathcal{Y}_{\mathrm{unseen}} \subset \mathcal{Y}$ be the unseen gesture classes, where $\mathcal{Y}_{\mathrm{in}} \cap \mathcal{Y}_{\mathrm{unseen}} = \emptyset$. For $S \in \{\mathrm{tr}, \mathrm{val}, \mathrm{test}\}$,*

$$\mathcal{D}_S^{\mathrm{in}} = \{(x, p, y) \in \mathcal{D}_S : y \in \mathcal{Y}_{\mathrm{in}}\}, \qquad \mathcal{D}_S^{\mathrm{unseen}} = \{(x, p, y) \in \mathcal{D}_S : y \in \mathcal{Y}_{\mathrm{unseen}}\}.$$

**Definition 2** ***Encoders.*** *Let $\mathcal{E}_x$ and $\mathcal{E}_p$ denote encoder functions mapping EMG and pose data respectively into latent embeddings in $\mathbb{R}^d$.*

### 2.2 Unimodal Encoder Pre-training

We adopt the standard Transformer encoder [21] backbone with a linear tokenizer to map raw signals into $d$-dimensional token embeddings. A patch length of $S$ along the temporal dimension results in $K = \left\lfloor \frac{T}{S} \right\rfloor$ non-overlapping tokens. We flatten the channels at each time-point within the patch and use linear tokenizer to project each patch into an embedding $\mathbf{z_i} \in \mathbb{R}^d$. Following Masked autoencoder (MAE) [5], given a mask ratio $r \in (0, 1)$, we sample a random mask set $\mathcal{M} \subset \{1, \ldots, K\}, |\mathcal{M}| = K \times r$. The encoder $\phi$ processes only the unmasked tokens $\{\mathbf{z}_i : i \notin \mathcal{M}\}$, and the decoder $\psi$ reconstructs the full sequence. We optimize the masked-token reconstruction via mean squared error: $\mathcal{L}_{\mathrm{MAE}} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \left\| \psi\big(\phi(\{\mathbf{z}_j\}_{j \notin \mathcal{M}})\big)_i - \mathbf{z}_i \right\|_2^2$. We apply this pre-training independently to the EMG and pose data, yielding two pre-trained encoders $\mathcal{E}_x = \phi_{\mathrm{EMG}}$ and $\mathcal{E}_p = \phi_{\mathrm{pose}}$, which produce meaningful representations for contrastive alignment.

### 2.3 Contrastive Pose-EMG Pre-training (CPEP)

In our approach, the hand pose data, offering high-quality, low-noise gesture information, serves as the anchor modality. In contrast, EMG signals exhibit lower signal-to-noise ratio (SNR). To align EMG and pose representations, we append a lightweight projection head $h_\phi$ to the pre-trained EMG encoder $\mathcal{E}_x$, mapping its [CLS] token into the same $d$-dimensional space as the pose encoder's [CLS] token. The [CLS] token is chosen for its ability to capture global context and provide a compact summary of the entire sequence [5]. During contrastive training, we keep the pose encoder $\mathcal{E}_p$ frozen and only update the EMG encoder $\mathcal{E}_x$ (together with the projection head $h$) to avoid the pose representations being altered when enforcing alignment with EMG representations. For each paired sample $(x_i, p_i)$ we compute $u_i = h(\mathcal{E}_x(x_i))_{[\mathrm{CLS}]}$, $v_i = (\mathcal{E}_p(p_i))_{[\mathrm{CLS}]}$, where the projection head $h$ is randomly initialized and trained, and the pose encoder $\mathcal{E}_p$ is frozen. We $\ell_2$-normalize embeddings and use cosine similarity. Let $\tilde{u}_i = \frac{u_i}{\|u_i\|}, \quad \tilde{v}_j = \frac{v_j}{\|v_j\|}, \quad s_{ij} = \frac{\tilde{u}_i^\top \tilde{v}_j}{\tau}$, with temperature $\tau$. Following CLIP [16], the symmetric Information Noise-Contrastive Estimation (InfoNCE) objective [20] is:

$$\mathcal{L}_{\mathrm{CPEP}} = \frac{1}{2N} \sum_{i=1}^N \left[ -\log \frac{\exp(s_{ii})}{\sum_{j=1}^N \exp(s_{ij})} - \log \frac{\exp(s_{ii})}{\sum_{j=1}^N \exp(s_{ji})} \right].$$

This loss pulls EMG embedding $\tilde{u}_i$ toward its matching pose $\tilde{v}_i$ while pushing away all non-matching poses, yielding pose-informed EMG embeddings that support generalizable gesture classification. The embedding dimension is 256 and CPEP was trained with a batch size of 256.

### 2.4 Evaluation Protocols

We employ two evaluation protocols to examine the learned representation quality. **Linear probing** (LP): with labeled EMG data $\{(x_i, y_i)\}$, we freeze $\mathcal{E}_x^*$ and train a randomly-initialized linear classifier $\mathcal{C}$ on top of its embeddings, reporting accuracy on a held-out split. **Zero-shot classification** (ZS) is performed as k-nearest-neighbor voting in the embedding space, following standard practice in representation learning [12, 16]. For each EMG sample, we retrieve its TopK nearest poses in the

| Split (totals) | Subset | Gesture Counts | | User Counts | |
|---|---|---|---|---|---|
| | | In-dist. | Unseen | In-dist. | Unseen |
| $\mathcal{D}_{tr}$ (23 gestures / 158 users) | $\mathcal{D}^{\text{in}}_{\text{probe-tr}}$ | 4 | 0 | 158 | 0 |
| $\mathcal{D}_{val}$ (29 gestures / 15 users) | $\mathcal{D}_{\text{tune}}$ | 4 | 4 | 0 | 3 |
| | $\mathcal{D}^{\text{unseen}}_{\text{probe-tr}}$ | 0 | 4 | 0 | 12 |
| $\mathcal{D}_{test}$ (29 gestures / 158 users) | $\mathcal{D}_{\text{eval}}$ | 4 | 4 | 0 | 20 |

Table 1: Dataset splits with gesture and user counts. Four unseen gestures evaluated out of six total.

embedding space, then vote the corresponding gesture labels to determine the predicted gesture. Given a test EMG sample $x_j$, let $\mathcal{R}_j = \mathrm{TopK}_{p \in \mathcal{P}_{\text{test}}}\big(\mathcal{E}_x(x_j)^\top \mathcal{E}_p(p)\big)\big)$ be the set of $k$ pose samples with highest cosine similarity to $x_j$. We then predict $\hat{y}_j = \mathrm{mode}\big\{\, y(p) \mid p \in \mathcal{R}_j \big\}$, where $y(p)$ denotes the gesture class associated with pose $p$.

# 3 Experiments and Results

## 3.1 Data

**Dataset:** We use *emg2pose* [17], a large-scale open-source EMG dataset containing 370 hours of sEMG and synchronized hand pose data across 193 consenting users, 29 different behavioral groups that include a diverse range of discrete and continuous hand motions such as making a fist or counting to five. The hand pose labels are generated using a high-resolution motion capture system. The full dataset contains over 80 million pose labels and is of similar scale to the largest computer vision equivalents. Each user completed four recording sessions per gesture category, each with different EMG band placement. Each session lasted 45–120 s, during which users repeatedly performed a mix of 3–5 similar gestures or unconstrained freeform movements. We use non-overlapping 2-second windows as input sequences. EMG is instance-normalized, band-pass filtered (2–250 Hz), and notch-filtered at 60 Hz. For more details, please refer to [17].

**Experiment Design:** We evaluate on two disjoint gesture sets drawn from the public *emg2pose* corpus. First, we select four representative single-hand motions covering various finger movements as our **in-distribution (in-dist.) gestures**. Second, from the six held-out classes that are not seen during training, we exclude the two-handed gesture and the highly variable "finger freeform" class, yielding four **unseen gestures**. Note that we include gestures like "*CountingWiggling*" and "*FingerPinches*" that are known to be challenging for vision-based hand pose estimation due to visual occlusion [17]. Details of gesture classes are in Appendix A. For data splits, we follow the public train $\mathcal{D}_{tr}$, val $\mathcal{D}_{val}$ , test $\mathcal{D}_{test}$ splits in [17] and define our data splits for downstream gesture classification tasks as shown in Table 1. The model is pre-trained on the full $\mathcal{D}_{tr}$. A linear head is trained on $\mathcal{D}^{\text{in}}_{\text{probe-tr}}$, and report final accuracy on the evaluation set $\mathcal{D}^{\text{in}}_{\text{eval}}$. For unseen gestures (which appear only in the original val and test splits), we train on $\mathcal{D}^{\text{unseen}}_{\text{probe-tr}}$, and report accuracy on $\mathcal{D}^{\text{unseen}}_{\text{eval}}$. Zero-shot classification is evaluated only on $\mathcal{D}_{\text{eval}}$. All users in $\mathcal{D}_{\text{eval}}$ are unseen, so the results also assess user-level generalization. A held-out dataset $\mathcal{D}_{\text{tune}}$, strictly disjoint from all other sets, is reserved for hyper-parameter tuning.

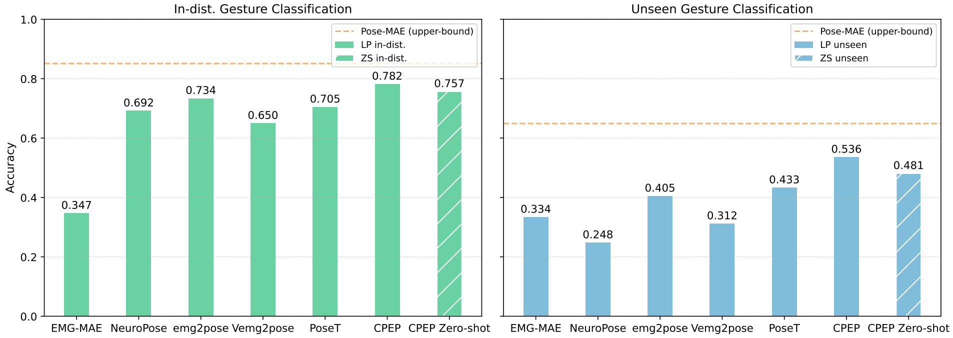

Figure 2: Classification accuracy on in-distribution and unseen gestures. Linear probing is reported for all methods; zero-shot is reported only for CPEP (other methods do not support zero-shot).

### 3.2 Gesture Classification Results

We assess the quality of EMG representations on both **in-dist.** and **unseen** gestures using the two evaluation protocols described earlier: zero-shot classification (ZS) and linear probing (LP). For comparative analysis, we compare our CPEP model against: (i) a supervised encoder–decoder Transformer called **PoseT** trained to regress poses directly from EMG; (ii) the supervised pose regression models (**emg2pose**, **Vemg2pose**, and **NeuroPose**) released with the *emg2pose* benchmark [17]; and (iii) a self-supervised MAE model pre-trained using EMG data only (**EMG-MAE**). For a fair comparison under linear probing, we freeze each baseline's encoder and train a softmax linear head on top of its embeddings. For context, we include an **upper-bound** by linear probing Pose-MAE features, since the pose encoder serves as the alignment anchor.

Figure 2 presents a summary of the experimental results. Since supervised baselines do not support zero-shot (k-nearest neighbor) classification in the embedding space, we report only their LP results. We report macro accuracy (mean per-class accuracy) to account for class imbalance across gestures. We omit standard deviations over random seeds because they were negligible relative to the performance gains. CPEP outperforms all baselines in terms of accuracy under both linear probing and zero-shot settings. Remarkably, zero-shot classification on unseen gestures achieves performance comparable to linear probing, without requiring any additional task-specific training. Furthermore, our zero-shot CPEP approach exceeds the linear probing results of the baselines, highlighting its superior generalizability through contrastive learning. These findings support our claim that aligning EMG signals with high-quality pose embeddings enables effective generalization to unseen gestures.

### 3.3 Ablation Study

**Impact of CPEP design choices.** Table 2 summarizes the ablation results for different CPEP design choices. Initializing the EMG encoder from the pre-trained MAE consistently outperforms random initialization (*CPEP-$\mathcal{E}_x$ RandInit*) and results in a more stable contrastive training. Random initialization slows convergence and degrades both zero-shot and linear-probing performance. One important finding is that randomly initializing both encoders ($\mathcal{E}_x$ and $\mathcal{E}_p$) fails to converge. Freezing both encoders and training only the projection head (*CPEP-$\mathcal{E}_x$ Frozen*) leads to significantly lower accuracy, highlighting the necessity of fine-tuning the EMG encoder during the feature alignment phase. Making the pose encoder trainable alongside the EMG encoder (updating both) also fails to converge, suggesting that altering the pose anchor harms alignment. Furthermore, average pooling over tokens (*CPEP-AvgPool*) underperforms the `[CLS]` token.

| | LP in-dist. | ZS in-dist. | LP unseen | ZS unseen |
|---|---|---|---|---|
| CPEP-$\mathcal{E}_x$ Frozen | 0.372 | 0.344 | 0.326 | 0.298 |
| CPEP-$\mathcal{E}_x$ RandInit | 0.748 | 0.701 | 0.479 | 0.454 |
| CPEP-AvgPool | 0.761 | 0.711 | 0.518 | 0.454 |
| CPEP (ours) | **0.782** | **0.757** | **0.536** | **0.481** |

Table 2: Ablations of CPEP training setup.

**Impact of $S_{\text{emg}}$, $S_{\text{pose}}$, and mask ratio $r$.** We study the effect of the unimodal MAE pre-training choices (patch length $S$ and mask ratio $r$) on downstream classification, with CPEP training setup fixed. Fig. 3 summarizes the results across different hyper-parameter settings. Overall, for small values of $S_{\text{pose}}$, higher $r$ values achieve better results. With $S_{\text{pose}}$=200, $r$=0.5 outperforming others. Using longer EMG patches consistently degrades performance.

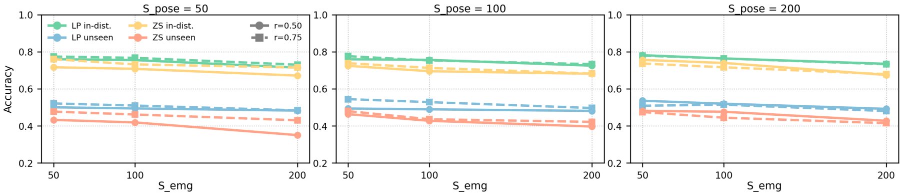

Figure 3: Ablation of MAE pre-training setup: temporal patch length $S$ and mask ratio $r$.

## 4 Discussion and Conclusion

We introduced CPEP, a contrastive pre-training framework that aligns EMG with pose representations. By leveraging pose as a rich supervisory signal, CPEP learns pose-informed EMG embeddings that capture structural and semantic relationships and are linearly separable in the latent space. Across in-distribution and unseen gesture classification tasks, CPEP demonstrated strong generalization and zero-shot capabilities. Future work includes extending CPEP to additional EMG datasets and other biosignals (e.g. IMU), as well as exploring tasks such as human activity recognition. Overall, CPEP provides an effective approach to zero-shot gesture classification from EMG, and can serve as a foundation for broader multi-modal biosignal applications.

## References

[1] Antonino Casile, Giulia Fregna, Vittorio Boarini, Chiara Paoluzzi, Fabio Manfredini, Nicola Lamberti, Andrea Baroni, and Sofia Straudi. Quantitative comparison of hand kinematics measured with a markerless commercial head-mounted display and a marker-based motion capture system in stroke survivors. *Sensors*, 23(18):7906, 2023.

[2] Agastasya Dahiya, Dhruv Wadhwa, Rohan Katti, and Luigi G Occhipinti. Efficient hand gesture recognition using artificial intelligence and imu based wearable device. *IEEE Sensors Letters*, 2024.

[3] Chenyu Gu, Weicong Lin, Xinyi He, Lei Zhang, and Mingming Zhang. Imu-based motion capture system for rehabilitation applications: A systematic review. *Biomimetic Intelligence and Robotics*, 3(2):100097, 2023.

[4] Harish Haresamudram, Chi Ian Tang, Sungho Suh, Paul Lukowicz, and Thomas Ploetz. Past, present, and future of sensor-based human activity recognition using wearables: A surveying tutorial on a still challenging task. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 9(2):1–44, 2025.

[5] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021.

[6] Néstor J Jarque-Bou, Joaquin L Sancho-Bru, and Margarita Vergara. A systematic review of emg applications for the characterization of forearm and hand muscle activity during activities of daily living: Results, challenges, and open issues. *Sensors*, 21(9):3035, 2021.

[7] Maan Khedr and Nasser El-Sheimy. A smartphone step counter using imu and magnetometer for navigation and health monitoring applications. *Sensors*, 17(11):2573, 2017.

[8] Gierad Laput and Chris Harrison. Sensing fine-grained hand activity with smartwatches. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–13, New York, NY, USA, 2019. Association for Computing Machinery.

[9] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023.

[10] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts, 2017.

[11] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.

[12] Markus Marks, Manuel Knott, Neehar Kondapaneni, Elijah Cole, Thijs Defraeye, Fernando Perez-Cruz, and Pietro Perona. A closer look at benchmarking self-supervised pre-training with image classification, 2024.

[13] Ali Moin, Andy Zhou, Abbas Rahimi, Alisha Menon, Simone Benatti, George Alexandrov, Senam Tamakloe, Jonathan Ting, Natasha Yamamoto, Yasser Khan, et al. A wearable biosensing system with in-sensor adaptive machine learning for hand gesture recognition. *Nature Electronics*, 4(1):54–63, 2021.

[14] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers, 2023.

[15] Kyung Rok Pyun, Kangkyu Kwon, Myung Jin Yoo, Kyun Kyu Kim, Dohyeon Gong, Woon-Hong Yeo, Seungyong Han, and Seung Hwan Ko. Machine-learned wearable sensors for real-time hand-motion recognition: toward practical applications. *National science review*, 11(2):nwad298, 2024.

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[17] Sasha Salter, Richard Warren, Collin Schlager, Adrian Spurr, Shangchen Han, Rohin Bhasin, Yujun Cai, Peter Walkington, Anuoluwapo Bolarinwa, Robert Wang, Nathan Danielson, Josh Merel, Eftychios Pnevmatikakis, and Jesse Marshall. emg2pose: A large and diverse benchmark for surface electromyographic hand pose estimation, 2024.

[18] Viswanath Sivakumar, Jeffrey Seely, Alan Du, Sean R Bittner, Adam Berenzweig, Anuoluwapo Bolarinwa, Alexandre Gramfort, and Michael I Mandel. emg2qwerty: A large dataset with baselines for touch typing using surface electromyography, 2025.

[19] 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.

[20] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[22] Huihui Wang, Bo Ru, Xin Miao, Qin Gao, Masood Habib, Long Liu, and Sen Qiu. Mems devices-based hand gesture recognition via wearable computing. *Micromachines*, 14(5):947, 2023.

## A Details of Gestures

**In-distribution (4 classes).** (1) Thumb swipes whole hand; (2) Hand claw, grasp, and flicks; (3) Thumbs rotation, thumbs up and down; (4) finger pinches, single/multiple fingers.

**Unseen (4 classes).** (1) Hook ’em Horns, OK, and Scissors; (2) Shaka and Vulcan peace; (3) Counting up/down; (4) Counting up/down with finger wiggling and spreading.

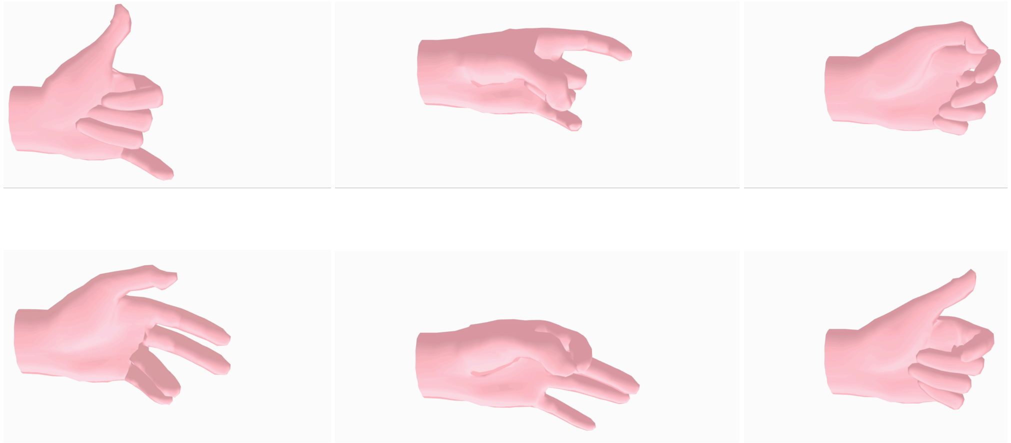

Figure 4: Example visualizations of gestures used in gesture classification tasks.

# B Implementation Details

We use 2 s windows sampled at 2 kHz for both pose and EMG. EMG is instance-normalized, band-pass filtered (2–250 Hz), and notch-filtered at 60 Hz. Following [17], we apply channel-rotation augmentation to EMG. Our MAE is an encoder–decoder Transformer model with 4 encoder layers and 2 decoder layers, and the embedding dimension is $d{=}256$. We optimize with AdamW [11] (lr $1\text{e}{-}4$, weight decay $1\text{e}{-}5$) and cosine annealing with warm restarts [10]. The masking ratio is 50%. Token length is $S{=}200$ for pose and $S{=}50$ for EMG, producing non-overlapping tokens along time. Batch size is 256, and each MAE is trained for 100 epochs. Our supervised baseline PoseT shares the same transformer architecture with MAE model, the only difference is that the outputs are pose regression predictions. The training objective is also the same as MAE, a mean squared error loss.

For CPEP, we attach a 1-layer projection head (hidden size 256) to the EMG encoder and train the EMG encoder plus projection head while keeping the pose encoder frozen. The contrastive temperature $\tau$ is learnable (initialized to 0.02). Batch size is 256 and training runs for 100 epochs. All output embeddings are $\ell_2$-normalized. All model trainings are conducted on $4\times$ NVIDIA V100 GPUs; end-to-end training of each model takes approximately 4.5 hours. In zero-shot retrieval, we precompute pose embeddings for the entire corpus and, for each EMG query, retrieve the top-$k$ neighbors by cosine similarity with $k{=}10$; the predicted label is the majority vote of the retrieved labels.