EgoEMG: A Multimodal Egocentric Dataset with Bilateral EMG and Vision for Hand Pose Estimation*

Ziheng Xi, Jiayi Yu, Yitao Wang, Yanbo Duan, Jianjiang Feng, Jie Zhou
Department of Automation, Tsinghua University
{xizh21, jy-yu23, wangyita23, duanyb24}@mails.tsinghua.edu.cn
{jfeng, jzhou}@tsinghua.edu.cn

Abstract

Surface electromyography (sEMG) records muscle activity during hand movement and can be decoded to recover detailed hand articulation. EMG and egocentric vision are complementary for hand sensing: EMG captures fine-grained finger articulation even under occlusion and poor lighting, while vision provides global hand configuration. However, no existing dataset synchronizes both modalities. We present *EgoEMG*, a multimodal egocentric dataset for bimanual hand pose estimation. *EgoEMG* includes bilateral wristband EMG with 16 total channels (8 per wrist) sampled at 2 kHz, 120 Hz IMU, egocentric wide-angle RGB video, external RGB-D video, and mocap-derived hand motion with wrist articulation angles. The dataset covers 41 participants performing 60 gesture classes, including 30 single-hand gestures and 30 bimanual gestures, totaling more than 10 hours of recording. We also introduce a benchmark with three tasks—EMG-to-pose, vision-to-pose, and EMG+vision fusion—under a shared joint-angle prediction target and common generalization split axes (cross-gesture, cross-user, and combined). As baselines, we evaluate EMGFormer for EMG-to-pose and generic ResNet/ViT backbones for vision-to-pose. We further study a residual fusion architecture that improves over matched lightweight vision-only baselines. Together, *EgoEMG* and its benchmark establish a foundation for future research on multimodal hand pose estimation with EMG and vision.

1 Introduction

Hand pose estimation from wearable sensing is a building block for augmented reality, prosthetic control, and physical intelligence [Salter et al., 2024, Atzori and Müller, 2014, Zhao et al., 2026]. Surface EMG is attractive because it measures muscle activation directly [Farina et al., 2014] and remains usable under self-occlusion, motion blur, and challenging lighting. Recent work has shown that EMG can recover detailed hand articulation [Salter et al., 2024, Verma, 2025, Lovecchio et al., 2025]. At the same time, egocentric vision has emerged as a powerful modality for hand tracking in first-person views [Pavlakos et al., 2024, Kwon et al., 2021], but relies on line-of-sight and struggles with fast motion blur and self-occlusion during manipulation. These properties suggest a potential complementarity between the two modalities: EMG provides fine-grained finger articulation even when the hand is occluded, while vision recovers global hand configuration and spatial context. Unlocking this complementarity requires a dataset that synchronizes both modalities with pose labels—yet current EMG pose datasets provide no synchronized egocentric video.

We address these gaps with *EgoEMG*, a controlled multimodal egocentric dataset for bimanual hand pose estimation. *EgoEMG* contains synchronized bilateral wristband EMG (16 total channels, 8 per wrist at 2 kHz), per-hand IMU, egocentric wide-angle RGB, external RGB-D, and hand motion

*Code and data: https://github.com/zhenqis123/EgoEMG

Preprint.

reconstructed into MANO parameters [Romero et al., 2017] with wrist articulation angles, covering 41 participants and 60 gesture classes (30 single-hand, 30 bimanual) over 10 hours of recording.

We provide a benchmark with three tasks—EMG-to-pose, vision-to-pose, and EMG+vision fusion—sharing the same joint-angle prediction target and generalization splits (cross-gesture, cross-user, and combined). On EMG-to-pose, EMGFormer variants outperform the prior TDS+LSTM baseline by up to 22%, with EMGFormer-S achieving the best efficiency trade-off (42% fewer parameters) and EMGFormer-L the best absolute accuracy. On vision-to-pose, ResNet backbones outperform generic ViT regressors at lower parameter counts, while the hand-specialized WiLoR baseline achieves the strongest standalone visual accuracy. A residual EMG+vision fusion architecture consistently improves over matched lightweight generic vision-only baselines, suggesting EMG provides complementary pose information beyond single-frame visual features; whether this complementarity extends to stronger specialized vision models remains an open question for future work.

Our contributions are:

- *EgoEMG*, to our knowledge the first dataset jointly providing bilateral wristband EMG and IMU, egocentric RGB, external RGB-D, bimanual MANO annotations with wrist joint angles, and per-frame gesture class labels—released publicly upon acceptance.
- A three-task hand pose estimation benchmark (EMG-to-pose, vision-to-pose, EMG+vision fusion) with cross-gesture, cross-user, and combined generalization splits under a unified joint-angle prediction target.
- Baseline results showing that EMGFormer provides a strong EMG-to-pose reference on both EMG2Pose and *EgoEMG*, that hand-specialized vision modeling further improves over generic ResNet/ViT baselines in this egocentric setting, and that EMG+vision fusion consistently improves over matched lightweight vision-only baselines.

2 Related work

Hand pose datasets and benchmarks. Public EMG datasets cover gesture recognition [Atzori and Müller, 2014, Pradhan et al., 2022, Kaczmarek et al., 2019, Du et al., 2017, Amma et al., 2015], high-density recordings [Jiang et al., 2021], force estimation [Shen et al., 2024], typing [Sivakumar et al., 2024], and demographically diverse cohorts [Gowda et al., 2025]. For continuous pose estimation, EMG2Pose [Salter et al., 2024] provides synchronized bilateral EMG with mocap-derived joint angles, while MyoKi [Andreas et al., 2025] and Jarque-Bou et al. [Jarque-Bou et al., 2019] pair forearm EMG with instrumented-glove kinematics. Cross-user generalization has been studied through benchmarks such as EMGBench [Yang et al., 2024]. In contrast, visual hand-pose benchmarks provide rich RGB annotations for monocular, egocentric, and hand-object settings [Zimmermann et al., 2019, Yu et al., 2022, Moon et al., 2023, Banerjee et al., 2024, Kwon et al., 2021, Huang et al., 2024], but do not include synchronized wrist EMG. Existing resources therefore lack a unified benchmark that combines egocentric video, bilateral wristband EMG, wrist articulation, and continuous hand-pose labels, limiting the study of multimodal pose regression from both visual observations and muscle activity.

EMG-based pose estimation. Decoding hand kinematics from surface EMG has progressed from single-DoF [Ngeo et al., 2012, 2014] and multi-DoF finger-angle regression to full-hand pose estimation. NeuroPose [Liu et al., 2021] demonstrated 3D hand-pose tracking from wearable EMG, while EMG2Pose [Salter et al., 2024] later established a larger-scale benchmark using a TDS+LSTM architecture. Recent work further improves reconstruction fidelity through tendon-driven motion modeling [Verma, 2025], vector-quantized pose tokenization [Lovecchio et al., 2025], and generalization-focused pretraining [Zhao et al., 2026]. In parallel, EMG foundation models and self-supervised approaches [Fasulo et al., 2025, Mehlman et al., 2025, Raghu et al., 2024, Laamerad et al., 2024] explore large-scale pretraining, but primarily target gesture classification rather than continuous pose. Despite these advances, EMG-only pose estimation remains sensitive to inter-user variability, and datasets based on instrumented gloves [Andreas et al., 2025, Jarque-Bou et al., 2019] provide limited degrees of freedom and can suffer from sensor drift and hysteresis.

Vision-based hand pose estimation and multimodal fusion. Monocular RGB hand-pose estimation has advanced rapidly with the MANO hand model [Romero et al., 2017], large-scale annotated datasets [Zimmermann et al., 2019, Yu et al., 2022, Moon et al., 2023, Huang et al., 2024], and

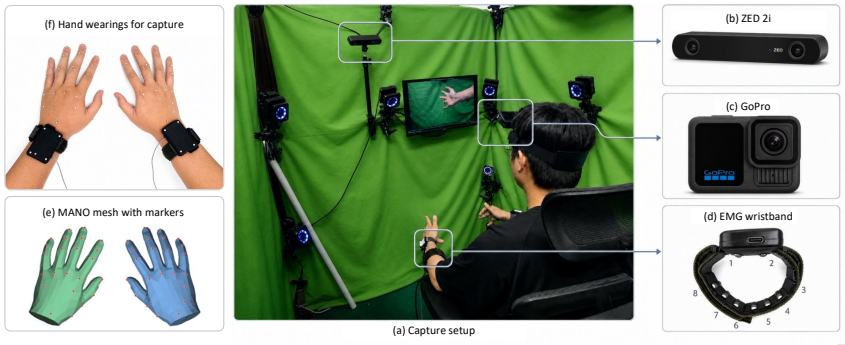


Figure 1: Data collection setup for *EgoEMG*. Bilateral EMG wristbands, head-mounted egocentric RGB, external ZED 2i RGB-D, and optical motion capture with hand markers for pose labels.

Transformer-based reconstructors [Pavlakos et al., 2024, Potamias et al., 2024]. Egocentric systems and datasets such as UmeTrack [Han et al., 2022], HOT3D [Banerjee et al., 2024], H2O [Kwon et al., 2021], and WristPP [Xi et al., 2026] further support first-person hand tracking and interaction understanding in wearable and egocentric settings. However, vision-based methods remain fundamentally constrained by line of sight: self-occlusion, inter-hand occlusion, object manipulation, and motion blur can substantially degrade pose estimates. EMG provides complementary muscle-activation cues that are independent of image visibility. Recent work has begun exploring EMG+vision multimodal learning for gesture understanding [Hao et al., 2024, Zandigohar et al., 2024, Cui et al., 2026] and cross-modal EMG-pose representation [Cui et al., 2025, Liu et al., 2025], but these efforts focus on gesture classification or offline representation mapping rather than continuous hand-pose regression. The lack of synchronized egocentric video, bilateral EMG, and continuous pose labels has therefore prevented systematic study of EMG+vision fusion for hand-pose regression.

3 *EgoEMG* Dataset

3.1 Overview and Capture Setup

EgoEMG is collected with a synchronized multimodal capture setup, including bilateral EMG, IMU, egocentric RGB, external RGB-D, and optical motion capture. Each participant wears two WAVELETECH EMG wristbands, one on each wrist, with each wristband recording 8 channels of surface EMG at 2 kHz. The visual setup contains a head-mounted wide-angle RGB camera for egocentric hand observation and an external ZED 2i RGB-D camera mounted on the motion-capture rack. Hand motion is tracked using 21 reflective markers per hand at 120 Hz with an FZMotion optical motion capture system. All modalities are temporally synchronized via host-timestamp soft-synchronization followed by linear interpolation to a common timeline (Figure 1).

We recruited 41 participants (23 male, 18 female; mean age 24; $\sim$15 min recording each). *EgoEMG* contains 60 gesture classes (30 single-hand, 30 bimanual) with per-frame gesture labels for gesture-conditioned evaluation. Gesture prompts are displayed in randomized order; single-hand gestures are performed mirror-symmetrically with both hands while varying wrist orientation; bimanual gestures follow demonstrated two-hand interactions, capturing natural coordination patterns. Figure 2 shows representative synchronized samples from all modalities.

3.2 Pose Label Reconstruction

Hand pose labels are reconstructed from optical motion-capture markers into MANO parameters [Romero et al., 2017]. We use a learning-based *markers2mano* pipeline that maps 21 marker positions to MANO pose, shape, and translation parameters. The pipeline uses fixed MANO surface vertices as virtual marker correspondences, allowing captured 3D marker positions to be aligned with the canonical MANO mesh. Full details of the *markers2mano* architecture, training data, and augmentation pipeline are provided in Appendix B.1.

Compared with the per-frame inverse-kinematics solver used by EMG2Pose, which has a reported 12.7% invalid-frame rate, our learning-based reconstruction reduces the invalid-frame rate to 3.6%, a

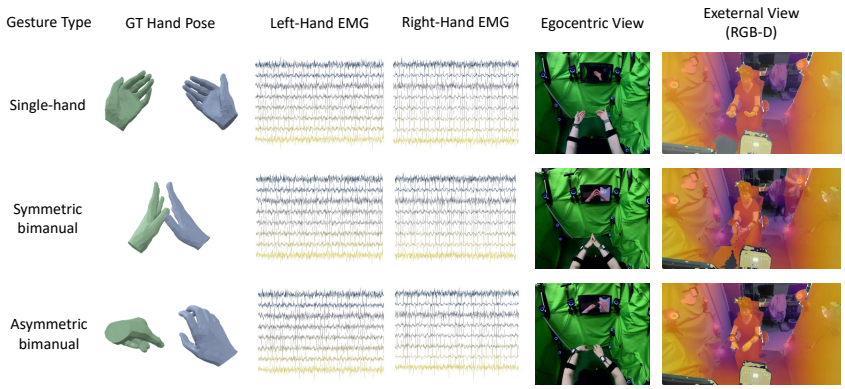


Figure 2: Representative synchronized samples from *EgoEMG*. Each row (3.9 s window) shows GT hand pose, bilateral EMG, egocentric RGB at window center, and external RGB-D.

Table 1: Comparison of *EgoEMG* with representative public EMG datasets. "EMG Ch." counts total channels across both wrists (per-wrist×2 for bilateral datasets); "Bilateral" indicates dual-wrist EMG capture; "Wrist labels" indicates explicit wrist articulation labels; "Pose labels" indicates continuous hand pose annotations.

Dataset	Subj.	EMG Ch.	Bilateral	Ego RGB	RGB-D	Wrist labels	Pose labels	Gestures
Jarque-Bou 2019	22	8	–	–	–	–	18 (g)	26
NinaPro	27–78	10–16	–	–	–	✓	22 (g)*	53
MyoKi	36	16	–	–	–	–	18 (g)	74
PiMForce	21	8	–	–	–	–	20 (g)	63
EMG2Pose	193	32 (16×2)	✓	–	–	–	20 (m)	50
***EgoEMG* (ours)**	**41**	**16 (8×2)**	✓	✓	✓	✓	**22 (m)**	**60**

* NinaPro DB1, 2 and 5 provide continuous joint angles (22-DoF); remaining sub-databases provide gesture labels only.
(g) = instrumented glove; (m) = optical motion capture.

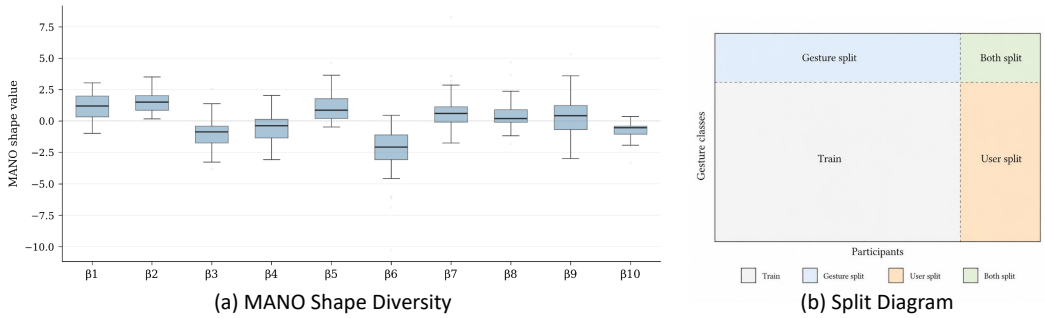
3.5× reduction. The resulting MANO meshes achieve a mean marker-to-mesh alignment error of 4.3 mm. In addition to MANO parameters, *EgoEMG* provides per-frame joint angles in a 22-DoF scalar format (20 finger angles plus 2 wrist articulation angles) obtained by optimizing UmeTrack joint angles to match MANO landmark positions; the conversion procedure is described in Appendix B.3. As an independent cross-modal check, we reproject reconstructed MANO meshes onto egocentric RGB frames using the calibrated camera model: the mean per-marker alignment error is 4.3 mm on *EgoEMG* (3.2 mm on the public InterHand2.6M subset), and visual inspection of randomly sampled reprojections confirms consistent mesh-to-image alignment across gesture families (Figure 11). Per-gesture error distributions are provided in Appendix B.

Wrist articulation angles are computed directly from motion-capture markers, independent of the MANO reconstruction pipeline. Markers attached to each EMG armband define a forearm coordinate frame, from which we compute two wrist degrees of freedom: flexion/extension and radial/ulnar deviation; the full derivation is provided in Appendix B.2.

3.3 Dataset Statistics and Positioning

Table 1 compares *EgoEMG* with representative public EMG datasets. Among existing datasets, only EMG2Pose and *EgoEMG* provide bilateral EMG with motion-capture-derived continuous hand-pose labels. *EgoEMG* further contributes synchronized egocentric RGB, external RGB-D, IMU, and explicit wrist articulation labels, enabling direct study of cross-modal hand pose estimation.

Hand shape and gesture diversity. *EgoEMG* estimates per-subject MANO shape parameters ($\boldsymbol{\beta} \in \mathbb{R}^{10}$) by averaging shape coefficients for each participant. Figure 3(a) shows the distribution across 41 participants, demonstrating substantial inter-subject hand-shape variation. The 60 gesture classes include 30 single-hand gestures and 30 bimanual interaction gestures, covering both symmetric coordination and asymmetric manipulation. The full gesture vocabulary is provided in Appendix A.1; per-dimension MANO shape statistics are reported in Appendix A.2.



(a) MANO Shape Diversity (b) Split Diagram

Figure 3: **(a)** MANO shape diversity across *EgoEMG* participants (82 hands). **(b)** Evaluation splits on the participant–gesture matrix (Train, Gesture/User/Both).

3.4 Evaluation Splits and Release Format

Following the protocol of EMG2Pose [Salter et al., 2024], we define two primary generalization axes: *gesture* and *user*. Based on these axes, we evaluate three splits: a *gesture split*, where held-out gestures are evaluated on seen participants; a *user split*, where held-out participants are evaluated on all gestures; and a combined *both split*, where both gestures and participants are held out. All modalities follow the same split assignment, preserving synchronized multimodal examples and preventing leakage along the held-out axis. We hold out 10 gestures for the gesture split and 6 participants for the user split, with an overall train-to-validation/test ratio of approximately 7:3. Figure 3(b) illustrates the split design.

The released dataset includes synchronized EMG and IMU signals, egocentric RGB frames, external RGB-D frames, MANO-based and UmeTrack-based hand-pose annotations, wrist articulation angles, timestamps, and calibration metadata. The prediction target is a 22-DoF joint-angle vector per hand: 20 finger angles and 2 wrist articulation angles. The same target definition and semantic ordering are used for both left and right hands.

4 Benchmark and Baselines

4.1 Benchmark Tasks

As an initial benchmark, *EgoEMG* defines three prediction tasks under the same 22-DoF target definition and evaluation split axes. The target consists of 20 finger joint angles and 2 wrist articulation angles per hand. The three tasks are: *EMG-to-pose*, which predicts joint angles from EMG windows; *vision-to-pose*, which predicts joint angles from single egocentric RGB hand crops; and *EMG+vision fusion*, which uses both modalities to predict the same output. The external RGB-D and IMU streams, as well as temporal multi-frame vision models, are included in the release to support future extensions but are not benchmarked in this paper.

4.2 Evaluation Protocol

We evaluate all tasks on the gesture, user, and both splits defined in §3. The primary metric is mean absolute error (MAE) over joint angles:

$$\text{MAE} = \frac{1}{JT} \sum_{j=1}^{J} \sum_{t=1}^{T} |\hat{\theta}_{j,t} - \theta_{j,t}| \quad [^\circ], \tag{1}$$

where J is the number of predicted joint angles and T is the number of evaluated time steps. For wrist analysis, we additionally report MAE on flexion/extension and radial/ulnar deviation. Unless otherwise specified, all errors are reported on the held-out test split. EMG evaluation uses the full predicted trajectory over each EMG window, whereas the vision-only and fusion models are evaluated on the window center frame, where visual evidence is available; cross-modality comparisons should therefore be interpreted with this difference in supervision granularity in mind.

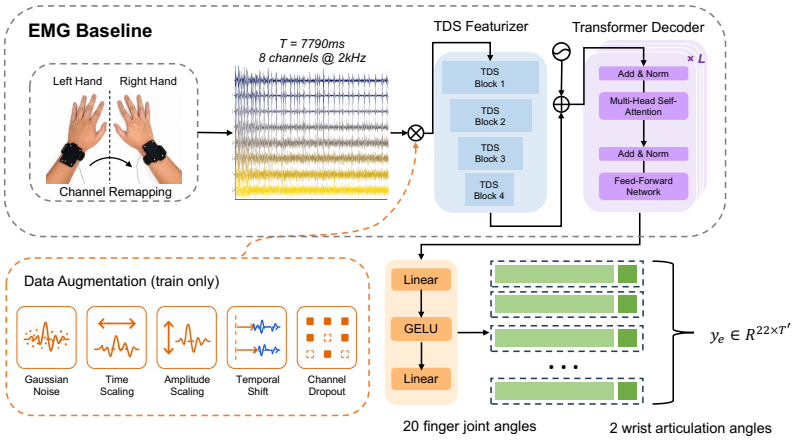


Figure 4: EMGFormer architecture for EMG-to-pose. Bilateral EMG is encoded by a TDS-style temporal convolutional frontend, decoded by a Transformer with RoPE positional encoding, and projected to joint angles including wrist articulation.

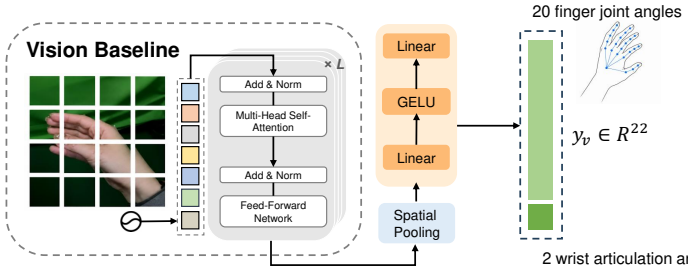


Figure 5: Vision-to-pose baseline. A 256×256 hand crop from the center frame is processed by a ResNet or ViT backbone, followed by an MLP head that regresses 22 joint angles.

4.3 EMG Baseline

As the EMG-to-pose baseline, we introduce EMGFormer (Figure 4), a temporal model that maps raw EMG windows to per-frame joint angles. A TDS-style temporal convolutional frontend [Hannun et al., 2019] downsamples 2 kHz EMG signals to latent features at approximately 37 Hz, followed by a Transformer decoder with rotary positional encoding (RoPE) [Su et al., 2024] to model long-range temporal dependencies. Left-hand EMG is mirrored to the right-hand convention during training. Each forward pass predicts one hand's joint-angle trajectory, including wrist articulation for *EgoEMG*. We provide three model sizes; implementation details, architecture specifications, and ablations are provided in Appendix C.1.

4.4 Vision Baseline

For vision-to-pose, we evaluate seven single-frame reference baselines. Six are generic image backbones across two families: ResNet [He et al., 2016] ({18, 50, 152}) and ViT [Dosovitskiy et al., 2021] ({Small, Base, Large}). ResNet backbones are initialized from ImageNet [Deng et al., 2009]-pretrained weights, while ViT backbones are initialized from DINOv2 [Oquab et al., 2024]-pretrained weights. Each generic model takes a 256×256 egocentric hand crop from the window center frame, extracts a feature vector with the backbone, and regresses the 22-DoF target through an MLP prediction head. We additionally evaluate WiLoR [Potamias et al., 2024] as a hand-specialized vision baseline initialized from a pretrained MANO-based hand-pose checkpoint and fine-tuned on the same center-frame protocol (Figure 5). Model sizes and training configurations are provided in Appendix C.2.

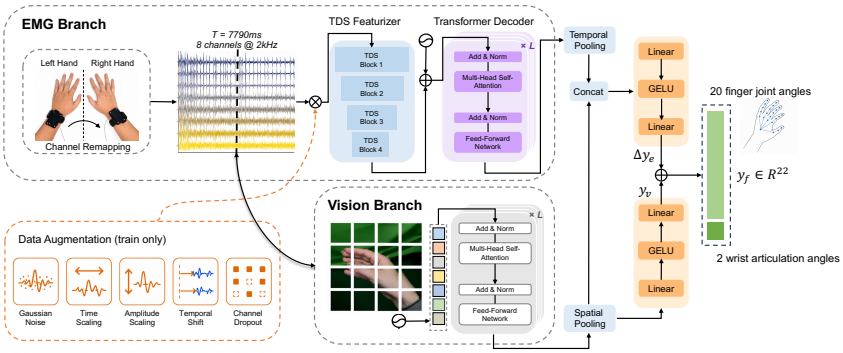


Figure 6: EMG+vision fusion baseline. Vision branch predicts $\mathbf{y}_v$; EMG branch predicts residual $\Delta\mathbf{y}_{\text{emg}}$ (zero-initialized head).

4.5 Fusion Baseline

The EMG+vision baseline is designed as a simple residual fusion model (Figure 6). The vision branch first predicts a vision-only pose $\mathbf{y}_v$, while the EMG branch predicts an additive correction $\Delta\mathbf{y}_{\text{emg}}$. The final prediction is $\mathbf{y} = \mathbf{y}_v + \Delta\mathbf{y}_{\text{emg}}$. This design makes the fusion model start from a strong visual baseline and learn only the pose information contributed by EMG.

The vision branch extracts a 256-d feature from the center-frame hand crop and uses a prediction head to produce $\mathbf{y}_v \in \mathbb{R}^{22}$. The EMG branch uses EMGFormer to encode the full EMG window into a temporal feature sequence, followed by learned temporal attention pooling to obtain a 256-d EMG feature. The two features are concatenated and passed to a fusion head that predicts $\Delta\mathbf{y}_{\text{emg}} \in \mathbb{R}^{22}$. The last layer of the residual head is zero-initialized so that the model begins from the vision-only prediction. We train the fusion model under a center-supervised protocol: only the window center frame is supervised, matching the vision-only target and enabling direct comparison.

5 Experiments

5.1 EMG Baseline Results

EMGFormer also establishes a strong baseline on the EMG2Pose dataset. On the hardest user+stage split, EMGFormer-S achieves $12.34° \pm 1.07°$ MAE, a 22% improvement over the vemg2pose baseline [Salter et al., 2024] ($15.8° \pm 1.4°$) with only 58% of the model size (full results in Appendix D.3).

Table 2 reports EMG-to-pose results on *EgoEMG* for EMGFormer variants and two traditional architectures across three test conditions (gesture/user/both). The baselines are: vemg2pose [Salter et al., 2024], a recurrent EMG-to-pose model with an LSTM temporal core; and NeuroPose, a convolutional encoder-decoder baseline [Liu et al., 2021]. We do not include CLDM [Verma, 2025] or VQ-MyoPose [Lovecchio et al., 2025] on *EgoEMG* because neither method has released public code, precluding reproduction on a new dataset; we compare against their published numbers on the EMG2Pose benchmark in Appendix D.3.

EMGFormer variants achieve the best overall accuracy, with EMGFormer-M and EMGFormer-L both outperforming the strongest traditional baseline. The gains are especially pronounced on the gesture split, indicating that EMGFormer better captures gesture-level articulation under the seen-user, held-out-gesture setting. Cross-user generalization remains the central challenge: larger temporal models improve gesture-split performance but do not by themselves close the gap on unseen participants. EMGFormer-M offers the best accuracy–capacity trade-off, matching EMGFormer-L in average MAE with fewer than half the parameters.

5.2 Vision-to-Pose Results

Table 3 reports vision-only and EMG+vision results on *EgoEMG*. We first analyze the vision-only rows, which include six generic single-frame baselines spanning two backbone families, ResNet and ViT, together with one hand-specialized baseline, WiLoR. Among the generic backbones, ResNet-152 achieves the best vision-only performance across all splits. ResNet-50 also outperforms ViT-Large

Table 2: EMG-to-pose results on *EgoEMG*. MAE is reported in degrees; lower is better. Gesture/User/Both columns report per-user mean ± standard deviation. Avg is the per-sample weighted mean MAE across all test splits (pooled gesture, user, and both). Shaded rows highlight EMGFormer variants; bold indicates best per column and underline indicates second best.

Method	Params	Gesture	User	Both	Avg.
EMGFormer variants					
EMGFormer-S	3.5M	$12.8_{\pm 1.4}$	$\mathbf{15.6_{\pm 2.5}}$	$17.4_{\pm 1.2}$	$\underline{14.7}$
EMGFormer-M	6.6M	$\underline{11.8_{\pm 1.6}}$	$\mathbf{15.6_{\pm 1.4}}$	$17.4_{\pm 0.9}$	**14.2**
EMGFormer-L	16.3M	$\mathbf{11.7_{\pm 1.5}}$	$15.7_{\pm 3.0}$	$17.7_{\pm 1.1}$	**14.2**
Traditional architectures					
vemg2pose	6.0M	$15.0_{\pm 1.4}$	$16.3_{\pm 1.7}$	$\underline{17.3_{\pm 1.3}}$	15.9
NeuroPose	6.4M	$15.8_{\pm 1.2}$	$\underline{15.7_{\pm 1.3}}$	$\mathbf{16.3_{\pm 0.7}}$	16.1

Table 3: Vision-only and EMG+vision fusion results on *EgoEMG*. MAE is reported in degrees; lower is better. Subscripts denote standard deviation across users. Avg is the per-sample weighted mean MAE across all test splits. Shaded rows highlight EMG+vision fusion baselines. Bold indicates the better result within each matched vision-only/fusion pair. Underlines mark the strongest generic vision-only backbone.

Input	Method	Params	Gesture	User	Both	Avg
ResNet-based comparison						
Vision	V-RN18	11.5M	$5.1_{\pm 1.3}$	$6.8_{\pm 1.7}$	$6.8_{\pm 0.9}$	5.9
EMG+Vision	F-RN18+S	15.5M	$\mathbf{4.6}_{\pm 1.3}$	$\mathbf{6.4}_{\pm 1.4}$	$\mathbf{6.4}_{\pm 0.8}$	**5.4**
ViT-based comparison						
Vision	V-ViT-S	21.9M	$5.0_{\pm 1.3}$	$7.3_{\pm 2.4}$	$7.2_{\pm 0.5}$	6.0
EMG+Vision	F-ViT-S+S	25.9M	$\mathbf{4.5}_{\pm 1.3}$	$\mathbf{6.8}_{\pm 1.2}$	$\mathbf{6.8}_{\pm 0.6}$	**5.5**
Additional vision-only backbones						
Vision	V-RN50	23.5M	$4.5_{\pm 1.3}$	$6.2_{\pm 2.2}$	$6.2_{\pm 0.7}$	5.3
Vision	V-RN152	58.2M	$\underline{4.3}_{\pm 1.2}$	$6.1_{\pm 2.3}$	$6.1_{\pm 0.7}$	5.1
Vision	V-ViT-B	86.2M	$4.8_{\pm 1.3}$	$7.0_{\pm 2.3}$	$6.9_{\pm 0.6}$	5.8
Vision	V-ViT-L	303.8M	$4.4_{\pm 1.3}$	$6.6_{\pm 2.2}$	$6.6_{\pm 0.7}$	5.4
Hand-specialized vision baseline						
Vision	V-WiLoR	631.6M	$\mathbf{3.9}_{\pm 1.3}$	$\mathbf{5.6}_{\pm 2.4}$	$\mathbf{5.7}_{\pm 0.7}$	**4.7**

despite using substantially fewer parameters, suggesting that convolutional backbones remain highly competitive as generic egocentric references in this data regime. Scaling reduces error within both generic backbone families, while cross-user and both splits remain harder than the gesture split across all vision-only models. The relative underperformance of generic ViT regressors may reflect the scale of the available supervised training data, as transformer-based vision models often benefit from larger datasets. In contrast, the hand-specialized WiLoR baseline achieves the strongest standalone visual accuracy, reducing the average MAE to 4.7°. This indicates that task-specific hand-pose pretraining and architecture remain important even in the controlled *EgoEMG* setting.

5.3 EMG+Vision Fusion Results

We evaluate EMG+vision fusion using the residual architecture described in §4.5, with EMGFormer-S as the EMG backbone. The fusion model predicts the window center frame, where the egocentric RGB observation is available, making its MAE directly comparable to the vision-only baseline under the same supervision target. Table 3 reports matched vision-only and EMG+vision comparisons.

Fusion consistently improves over matched lightweight generic vision-only baselines across all backbone families and splits. The improvement is largest on the gesture split, while user-split gains are smaller, suggesting that EMG is particularly informative for fine articulation but remains affected by inter-subject variability. At the same time, V-WiLoR remains stronger than the current lightweight fusion baselines, indicating that the next benchmark milestone is to test whether EMG also improves

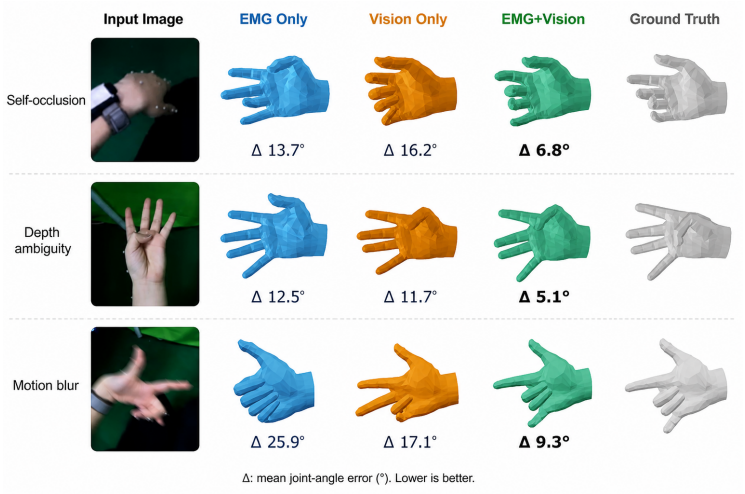


Figure 7: Qualitative comparison on *EgoEMG*. The fused model (green) recovers articulation closer to GT under motion blur, self-occlusion, and depth ambiguity.

a strong hand-specialized visual model rather than only generic backbones. A per-gesture analysis across all 60 gesture classes (Appendix D.5) shows that fusion improves over vision-only on nearly all gestures. An occlusion-stratified analysis (Appendix D.6) shows a positive correlation between hand self-occlusion and fusion gain. Figure 7 illustrates this complementarity: when visual evidence is ambiguous, EMG retains pose cues independent of image appearance, and the fused model recovers articulation closer to the ground truth. For reference under the same center-frame supervision target used by fusion, we also provide EMG-only center-frame results in Appendix D.2.

6 Discussion and Conclusion

EgoEMG provides a multimodal benchmark for hand pose estimation with synchronized bilateral EMG, IMU, egocentric RGB, external RGB-D, and hand-pose annotations including wrist articulation. By using a shared 22-DoF prediction target and unified split axes, the benchmark aligns EMG-to-pose, vision-to-pose, and EMG+vision fusion under a common label space while retaining modality-specific supervision protocols.

Our experiments highlight three findings. First, cross-user generalization remains the central challenge for EMG-based pose estimation: errors increase more on unseen participants than on unseen gestures. Second, EMG and vision offer complementary information. Vision provides the stronger standalone signal in the current benchmark, and the hand-specialized WiLoR baseline is the strongest vision-only model, whereas EMG captures pose-relevant cues for fine articulation and improves performance when fused with matched lightweight vision baselines. Third, scaling behavior depends on the data regime. The proposed EMGFormer architecture improves over vemg2pose on the original EMG2Pose benchmark, but larger variants do not consistently improve on *EgoEMG*. This suggests that increasing model capacity alone is insufficient under limited subject coverage and motivates large-scale EMG pretraining, cross-dataset transfer, and subject-robust adaptation.

Limitations and future work. The present study has several limitations. First, *EgoEMG* includes 41 participants, which is substantial for synchronized multimodal capture but still limited for modeling the full extent of inter-subject EMG variability. Expanding the dataset to more diverse participants would enable a more systematic study of cross-user robustness. Second, the current vision benchmark is limited to single-frame models. Since *EgoEMG* provides synchronized multi-view video and RGB-D streams, it can support future work on temporal, multi-view, and RGB-D-based methods. The benchmark also does not yet include rich hand-object interaction with instrumented or densely annotated objects; future extensions should test whether the same multimodal signals remain helpful when object contact introduces additional occlusion and pose ambiguity. Third, our fusion baselines are intentionally simple and are only evaluated with lightweight generic visual branches. A stronger next step is to test whether EMG also improves specialized vision models such as WiLoR. More

expressive architectures, including cross-attention, uncertainty-aware fusion, and explicit temporal alignment, may better exploit the complementary strengths of EMG and vision.

More broadly, *EgoEMG* provides a testbed for studying when physiological signals add information beyond visual observations, especially under unreliable visual evidence such as self-occlusion and fast motion. We hope it will support future work on subject-robust, multimodal hand-pose estimation under a unified benchmark.

Ethics Statement

All participants provided informed consent under an IRB-approved protocol and were compensated at a standard hourly rate commensurate with local guidelines. Participants were explicitly informed of their right to withdraw from the study at any time and to request deletion of their data; no participant exercised this right. The consent form explained the purpose of data collection, the modalities recorded, and the intended research use, and is included in the supplementary materials.

The dataset contains no personally identifiable information: facial regions in egocentric and external camera frames are automatically blurred, and participant identifiers are replaced with anonymous numeric IDs. While EMG and motion capture data do not reveal sensitive health information beyond gross motor function, we note that MANO shape parameters and EMG signals could in principle serve as soft biometric identifiers under certain conditions. The dataset will be released under a CC-BY-NC 4.0 research-use license that prohibits re-identification attempts and commercial use without explicit permission. We commit to hosting the dataset on a stable platform (e.g., Zenodo) with a DOI and versioned releases, and will provide a contact email for data-related inquiries and error reporting.

Reproducibility Statement

The repository contains training code, evaluation scripts, and EMGFormer baseline experiments with exact configuration files for all reported results. Appendix C.1 provides per-variant hyperparameters, optimizer settings, and compute requirements. The public release will include dataset documentation, split definitions, preprocessing scripts, benchmark configurations, and an anonymized download link. All EMG2Pose baseline experiments were run on 6 NVIDIA RTX 4090 GPUs with bf16 mixed precision; each EMGFormer variant requires approximately 2–8 GPU-hours for 200 epochs of training. Total compute for all reported experiments is approximately 128 GPU-hours (120 for main baselines + 8 for additional analyses).

Acknowledgments and Disclosure of Funding

We thank the anonymous reviewers.

References

Christoph Amma, Thomas Krings, Jonas Böer, and Tanja Schultz. Advancing muscle-computer interfaces with high-density electromyography. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI)*, pages 929–938, 2015.

Benedikt Andreas, Konstantin Werner, Yiming Hou, Kanishk Dwivedi, Claudio Castellini, and Petra Beckerle. A database of multimodal myography and hand kinematics during realistic daily life activities. *Scientific Data*, 12:1507, 2025.

Manfredo Atzori and Henning Müller. Ninapro database: A resource for semg-based hand movement research. *Scientific Data*, 1:140053, 2014.

Prithvijit Banerjee, Shikhar Tuli, Rohan Chacko, Taein Kwon, Abhinav Gupta, and Martial Hebert. Hot3d: A large-scale egocentric dataset for 3d hand and object tracking. In *European Conference on Computer Vision (ECCV)*, 2024.

Jason Cui, Christian Sandino, Hossein Pouransari, Leo Liu, Erion Minxha, Francisco Zippi, Sagar Verma, Lenka Sedlackova, Arber Azemi, and Hooman Mahasseni. Cross-modal pretraining for electromyography and pose learning. *arXiv preprint arXiv:2509.04699*, 2025. URL https://arxiv.org/abs/2509.04699.

Jason Cui, Christian Sandino, Hossein Pouransari, Leo Liu, Erion Minxha, Francisco Zippi, Arber Azemi, and Hooman Mahasseni. Embridge: Bridging emg and vision-language models for gesture understanding. In *International Conference on Learning Representations (ICLR)*, 2026.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations (ICLR)*, 2021.

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.

Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. Arctic: A dataset for dexterous bimanual hand-object interaction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Dario Farina, Aleš Holobar, Roberto Merletti, and Roger M. Enoka. The extraction of neural strategies from the surface emg: An update. *Journal of Applied Physiology*, 117(11):1215–1230, 2014.

Matteo Fasulo, Giusy Spacone, Thorir Mar Ingolfsson, Yawei Li, Luca Benini, and Andrea Cossettini. TinyMyo: A tiny foundation model for flexible EMG signal processing at the edge. *arXiv preprint arXiv:2512.15729*, 2025. URL https://arxiv.org/abs/2512.15729.

Harshavardhana T. Gowda, Neha Kaul, Carlos Carrasco, Marcus A. Battraw, Safa Amer, Saniya Kotwal, Selena Lam, Zachary McNaughton, Ferdous Rahimi, Sana Shehabi, Jonathon S. Schofield, and Lee M. Miller. A database of upper limb surface electromyogram signals from demographically diverse individuals. *Scientific Data*, 12(1):517, 2025. doi: 10.1038/s41597-025-04825-z.

Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Ho-3d: A benchmark for hand-object 3d pose estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Shangchen Han, Po-chen Wu, Yubo Zhang, Beibei Liu, Linguang Zhang, and Zheng Wang. UmeTrack: Unified multi-view end-to-end hand tracking for VR. In *SIGGRAPH Asia Conference Papers*, 2022. URL https://arxiv.org/abs/2211.00099.

Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Y. Ng. Sequence-to-sequence speech recognition with time-depth separable convolutions. In *Interspeech*, 2019.

Shunran Hao, Dongxu Gao, Zhaojie Ju, and Qing Gao. Multimodal hand gesture recognition based on the fusion of surface electromyography and vision. In *International Conference on Mechatronics and Machine Vision in Practice (M2VIP)*, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-Excitation Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Weipeng Huang, Hang Guo, Lixin Yang, Angela Yao, and Yulan Guo. Gigahand: A large-scale annotated dataset for 3d hand reconstruction from egocentric and exocentric views. *arXiv preprint arXiv:2412.04244*, 2024. URL https://arxiv.org/abs/2412.04244.

Néstor J. Jarque-Bou, Margarita Vergara, Joaquín L. Sancho-Bru, Verónica Gracia-Ibáñez, and Alba Roda-Sales. A calibrated database of kinematics and EMG of the forearm and hand during activities of daily living. *Scientific Data*, 6(1), 2019.

X. Jiang et al. Open access dataset and toolbox of high-density surface electromyogram recordings. *PhysioNet*, 2021.

Paweł Kaczmarek, Tomasz Mańkowski, and Jakub Tomczyński. Putemg: Surface emg dataset for repetitive hand activities. *Sensors*, 19(16):3548, 2019.

Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2O: Two hands manipulating objects for first person interaction recognition. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Abdelhak Laamerad, Alireza Shabanpour, Md Islam, and Sina Mohammadi. Masked autoencoding for high-density emg modeling. *arXiv preprint arXiv:2410.17261*, 2024. URL `https://arxiv.org/abs/2410.17261`.

Ruofan Liu, Yichen Peng, Takanori Oku, Chen-Chieh Liao, Erwin Wu, Shinichi Furuya, and Hideki Koike. From pose to muscle: Multimodal learning for piano hand muscle electromyography. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Yilin Liu, Shijia Zhang, and Mahanth Gowda. Neuropose: 3d hand pose tracking using emg wearables. In *Proceedings of the Web Conference 2021*, pages 1471–1482, 2021.

Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Francesco Lovecchio, Aravind Mamidanna, Randolph Jansen, Dario Farina, and Tomas Verstraten. Vq-myopose: Vector-quantized modeling for myoelectric hand pose estimation. In *Proceedings of the HANDS Workshop*, 2025.

Nicholas Mehlman, Jean-Christophe Gagnon-Audet, Michael Shvartsman, Kelvin Niu, Alexander H. Miller, and Shagun Sodhani. Scaling and distilling transformer models for sEMG. *Transactions on Machine Learning Research (TMLR)*, 2025.

Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2.6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(9):11065–11080, 2023.

Jimson Ngeo, Tomoya Tamei, and Tomohiro Shibata. Continuous estimation of finger joint angles using muscle activation inputs from surface EMG signals. In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2756–2759, 2012.

Jimson G Ngeo, Tomoya Tamei, and Tomohiro Shibata. Continuous and simultaneous estimation of finger kinematics using inputs from an EMG-to-muscle activation model. *Journal of NeuroEngineering and Rehabilitation*, 11(1):122, 2014.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research*, 2024.

Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. In *European Conference on Computer Vision (ECCV)*, 2024.

Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild. In *European Conference on Computer Vision (ECCV)*, 2024.

Gopal Pradhan, Jing He, and Ning Jiang. Grabmyo: A benchmark for wearable emg gesture recognition. *Scientific Data*, 9:733, 2022.

Shyam Raghu, Sarah MacIsaac, and Erik Scheme. Vicreg for electromyography representation learning. *arXiv preprint arXiv:2409.11632*, 2024. URL `https://arxiv.org/abs/2409.11632`. Submitted to Computers in Biology and Medicine.

Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. 36(6), 2017.

Sasha Salter, Richard Warren, Collin Schlager, Adrian Spurr, Wonjun Han, Anup Bhasin, Lili Cai, Matt Walkington, Modupe Bolarinwa, Weikai Wang, Chase Danielson, Joshua Merel, Eftychios Pnevmatikakis, and Luke Marshall. emg2pose: A large and diverse benchmark for surface electromyographic hand pose estimation. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2024.

Yilin Shen, Jiankai Hao, Hao-Shu Fang, and Cewu Lu. Pimforce: Large-scale dataset and benchmark for emg-based force estimation during grasp. In *PiMForce: Large-Scale Dataset and Benchmark for EMG-Based Force Estimation during Grasp*, 2024.

R. C. Sîmpetru, A. Arkudas, D. I. Braun, M. Osswald, D. S. de Oliveira, B. Eskofier, T. M. Kinfe, and A. Del Vecchio. Sensing the full dynamics of the human hand with a neural interface and deep learning. *Science Robotics*, 9(93), 2024.

Aditya Sivakumar, Josh Seely, Yu Du, Robert Bittner, Benjamin Berenzweig, Modupe Bolarinwa, Alexandre Gramfort, and Michael Mandel. Emg2qwerty: A large-scale dataset for wrist-worn emg typing. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2024.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063.

Sagar Verma. emg2tendon: From semg signals to tendon control in musculoskeletal hands. In *Robotics: Science and Systems (RSS)*, 2025.

Ziheng Xi, Zihang Ao, Yitao Wang, Mingze Gao, Wanmei Zhang, Jianjiang Feng, and Jie Zhou. Wristpp: A wrist-worn system for hand pose and pressure estimation. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pages 1–30, 2026.

Jehan Yang, Maxwell Soh, Vivianna Lieu, Douglas J. Weber, and Zackory Erickson. EMGBench: Benchmarking out-of-distribution generalization and adaptation for electromyography. *arXiv preprint arXiv:2410.23625*, 2024. URL `https://arxiv.org/abs/2410.23625`.

Bowen Yu, Subhash Suresh, Mustafa Mukadam, Daniel Maturana, Chin Lih, and Tat-Jen Cham. Dexycb: A benchmark for enabling grasp perception. In *International Conference on Robotics and Automation (ICRA)*, 2022.

Mehrshad Zandigohar, Mo Han, Mohammadreza Sharif, Sezen Yağmur Günay, Mariusz P. Furmanek, Mathew Yarossi, Paolo Bonato, Cagdas Onal, Taşkın Padır, Deniz Erdoğmuş, and Gunar Schirner. Multimodal fusion of EMG and vision for human grasp intent inference in prosthetic hand control. *Frontiers in Robotics and AI*, 11:1312554, 2024.

Qianyou Zhao, Wenqiao Li, Chiyu Wang, and Kaifeng Zhang. DexEMG: Towards dexterous teleoperation system via EMG2Pose generalization. *arXiv preprint arXiv:2603.05861*, 2026. URL `https://arxiv.org/abs/2603.05861`.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5745–5753, 2019.

Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless hand reconstruction from single rgb images. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10832–10841, 2019.

Christian Zimmermann, Max Argus, and Thomas Brox. Contrastive representation learning for hand shape estimation. In *DAGM German Conference on Pattern Recognition*, 2021.

A Dataset details

A.1 Gesture Vocabulary

Table 4 lists the complete gesture vocabulary of *EgoEMG* with descriptions.

A.2 MANO Shape Diversity Statistics

Table 5 reports per-dimension statistics of the MANO shape parameters ($\boldsymbol{\beta} \in \mathbb{R}^{10}$) across *EgoEMG*'s 41 participants (82 hands).

A.3 Dataset Card and Release Format

License. The dataset will be released under CC-BY-NC 4.0 for research use, and the baseline code will be distributed under the MIT License. Third-party assets, including the MANO model [Romero et al., 2017] and the training data for *markers2mano*, remain subject to their original licenses.

File schema. Data is organized by episode. Each episode is stored as a parquet file containing synchronized multimodal time series: bilateral EMG (16 channels at 2 kHz, raw and filtered), per-hand IMU, an egocentric RGB frame from a head-mounted wide-angle GoPro (1280×720 at 60 fps), an external RGB-D frame from a ZED 2i sensor (RGB + depth, 1280×720 at 30 fps), MANO pose parameters $\boldsymbol{\theta} \in \mathbb{R}^{48}$ and shape parameters $\boldsymbol{\beta} \in \mathbb{R}^{10}$, wrist flexion/extension and radial/ulnar deviation angles, timestamps, calibration metadata, and split assignments. Video frames are stored as MP4 files and referenced by frame index; each frame includes synchronization diagnostics (`stale` flag and `delta_ms`). Pre-cropped 256×256 left- and right-hand patches, generated by reprojecting mocap keypoints onto the egocentric view and extracting per-hand bounding boxes, are provided as per-episode LMDB archives to support vision-based training without requiring users to re-run the projection pipeline. Windows of 7790 EMG samples are extracted on-the-fly during training.

Access. The dataset will be hosted with a download script and checksum verification. During review, an anonymized access link will be provided.

A.4 EMG Signal Processing

All EMG signals are recorded at 2 kHz. The release includes both raw and filtered EMG for each channel. Filtering is applied per channel in the frequency domain via FFT mask with the following pipeline:

1. Per-channel mean subtraction (DC removal).
2. Narrow-band notch filters at 50 Hz (power line frequency) and 100 Hz (first harmonic).
3. Broadband bandpass filter 20–850 Hz.
4. Soft roll-off above 900 Hz to suppress high-frequency noise.

No amplitude normalization is applied; filtered EMG values are retained in mV. Both raw (`observation.emg.left/right`) and filtered (`observation.emg.left_filtered/right_filtered`) streams are preserved in the release. Training uses the filtered EMG by default.

Table 4: Complete gesture vocabulary of *EgoEMG* (60 classes) with per-gesture descriptions.

Gesture	Description
Single-hand (30 gestures)	
ASL1–ASL9	American Sign Language digits 1 through 9
Claw3	Three-finger claw pose (index + middle + ring flexed)
Claw5	Five-finger claw pose (all fingers flexed)
FreeAction	Free-form unconstrained hand action
ILY	"I Love You" hand sign (thumb, index, pinky extended)
IndexBow	Index finger flexion/extension bowing
IndexMiddleClaw	Index and middle fingers clawed, others extended
JoystickCircle	Circular joystick manipulation on the palm with thumb
JoystickSlide	Linear joystick sliding on the palm with thumb
MiddleBow	Middle finger flexion/extension bowing
Nine	Hand forming number 9 (four fingers curled, index up)
PalmYaw	Palm facing up/down rotation about the yaw axis
PinchMiddle	Thumb to middle fingertip pinch
PinkyBow	Pinky finger flexion/extension bowing
Rest	Relaxed hand in neutral resting posture
RingAndThumb	Ring finger touching thumb tip
RingBow	Ring finger flexion/extension bowing
Rock	"Rock on" hand sign (index and pinky extended)
Thumb	Thumb up
nocontact_disperse_palm	Fingers spread apart with palm open, no contact
nocontact_free	Free-form finger motion without contact
nocontact_grab	Simulated grasping motion without contact
Symmetric bimanual (18 gestures)	
Clap	Both hands clapping together symmetrically
CrossHand	Both hands crossed fingers
CrossStretch	Hands opposite fingers and stretched
FingerTipTouch	Matching fingertips of both hands touching
FistBump	Two fists bumping together
Gaming	Both hands holding a game controller
HandClasp	Hands clasped together
HandRub	Rubbing palms together
IndexTapping	Both index fingers tapping in the air
Kiss	Fingertips of both hands touching (Thumb and Middle), then separating
PalmStack	One palm stacked on top of the other
Prayer	Palms pressed together in prayer position
MiddleOppo	Middle and Ring fingers of one hand oppose the other hand
SymOpen	Both hands opening symmetrically from closed to open
SymSwing	Both hands swinging symmetrically side to side
ThumbWrestle	Thumbs wrestling each other
Typing	Both hands typing on a virtual keyboard
raw	Unconstrained bimanual hand movement
Asymmetric bimanual (12 gestures)	
FingerPullLeft	Left hand pulls right-hand fingers
FingerPullRight	Right hand pulls left-hand fingers
PalmRoll	Palms facing, rolling around each other
PinkyHook	Pinky fingers hooked together
Squeeze	Both hands squeezing an imaginary object
Beijing	Two hands crossing to make a "Bei" Chinese character
Checky	Two hands with thumbs forming a "Checky" sign
PairClaw	Both hands clawing with asymmetric finger configurations
PairOK	Both hands forming OK signs
Picture	Hands forming a picture frame rectangle
PinchWring	Two hands pinching and wringing
ThumbOppo	Thumbs of both hands in opposition at different orientations

Table 5: Per-dimension MANO shape coefficient statistics across *EgoEMG* participants. MANO shape parameters are PCA-derived and have no fixed semantic interpretation.

Dimension	Mean	Std	Min	Max	P95–P5 range
β_1	1.087	1.051	-0.966	3.034	4.000
β_2	1.505	0.793	0.183	3.506	3.323
β_3	-0.981	1.076	-3.804	2.547	6.351
β_4	-0.537	1.069	-3.077	2.050	5.128
β_5	1.054	0.996	-0.461	4.638	5.099
β_6	-2.269	1.711	-10.253	0.450	10.703
β_7	0.727	1.437	-1.735	8.254	9.989
β_8	0.371	1.000	-1.802	4.683	6.485
β_9	0.365	1.464	-2.978	5.344	8.322
β_{10}	-0.738	0.553	-3.308	0.366	3.674

B Label reconstruction and quality

B.1 Markers2MANO Pipeline

This section details the pipeline used to recover MANO parameters from optical motion capture marker positions during *EgoEMG* data collection.

B.1.1 Marker Selection

We select 21 vertices from the MANO mesh surface as virtual markers, distributed across the hand topology to capture wrist, finger, and palm geometry: indices 191, 88, 253, 708, 729 (wrist/palm), 144, 87, 295, 319 (thumb), 220, 365, 407, 445 (index/middle/ring), and 183, 477, 518, 556, 83, 589, 635, 673 (pinky/palm). These indices are consistent across all MANO shape instances $\boldsymbol{\beta}$, enabling direct correspondence between captured 3D marker positions and the canonical MANO mesh.

B.1.2 Graph Transformer Architecture

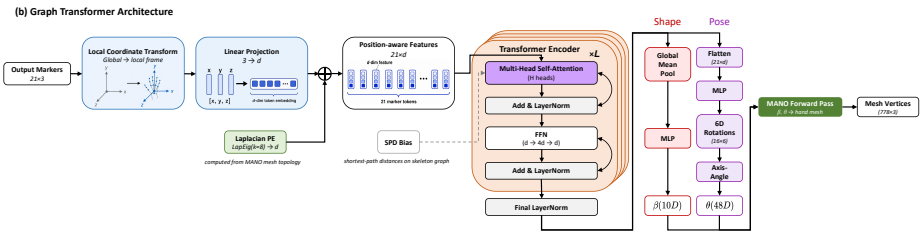


Figure 8: Architecture of *markers2mano* pipeline.

The primary inference model is a Graph Transformer that maps the 21 marker positions $\mathbf{M} \in \mathbb{R}^{21\times 3}$ to MANO pose $\boldsymbol{\theta} \in \mathbb{R}^{48}$ and shape $\boldsymbol{\beta} \in \mathbb{R}^{10}$ parameters. The architecture is shown in Figure 8. The model consists of:

- **Input embedding**: Each marker position is projected to a latent dimension d via a linear layer ($\mathbb{R}^3 \to \mathbb{R}^d$).
- **Graph positional encoding**: Graph Laplacian eigenvectors ($k = 8$) of the MANO hand-skeleton connectivity graph are projected through a linear layer ($\mathbb{R}^8 \to \mathbb{R}^d$) and *added* to the per-marker input embeddings as positional encoding. This provides each marker with a geometric signature of its location in the hand topology.
- **Structural attention bias**: The shortest-path distances between marker pairs on the hand skeleton graph are embedded and reshaped into a per-head attention bias matrix. This additive bias encourages anatomically adjacent markers (e.g., successive joints along a finger) to attend more strongly to each other.
- **Transformer encoder**: L standard Transformer layers with multi-head self-attention (with the SPD bias added to the pre-softmax attention scores), LayerNorm, and a feed-forward network with $4\times$ expansion ratio ($d \to 4d \to d$). Residual connections wrap each sub-layer.
- **Output heads**: After a final LayerNorm, two separate MLP heads produce the predictions. The *shape head* applies global mean pooling over the 21 marker features followed by a 3-layer MLP to regress $\boldsymbol{\beta} \in \mathbb{R}^{10}$. The *pose head* flattens all per-marker features ($21 \cdot d$) and passes them through a 3-layer MLP to regress 6D rotation representations [Zhou et al., 2019] for the 16 hand joints, which are then converted to axis-angle parameters $\boldsymbol{\theta} \in \mathbb{R}^{48}$.
- **MANO forward pass**: The predicted $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ are passed through the MANO layer to obtain the final mesh vertices $\mathbf{V} \in \mathbb{R}^{778\times 3}$.

B.1.3 Training and Data Augmentation

The model is trained on hand meshes derived from 11 public datasets—GigaHand [Huang et al., 2024], HOT3D [Banerjee et al., 2024], ARCTIC [Fan et al., 2023], HOI4D [Liu et al., 2022], HanCo [Zimmermann et al., 2021], DexYCB [Yu et al., 2022], H2O [Kwon et al., 2021], InterHand2.6M [Moon et al., 2023], FreiHand [Zimmermann et al., 2019], HO3Dv3 [Hampali et al., 2020], and HO3Dv2 [Hampali et al., 2020]—whose combined raw footage exceeds 195 million

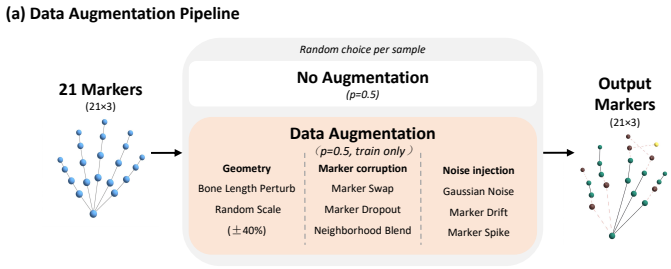


Figure 9: Data augmentation pipeline for *markers2mano* training.

frames. To balance the training distribution, we assign each dataset a sampling ratio based on its scale tier: large-scale datasets (100M+ frames, e.g., GigaHand) at 0.14%; medium-scale datasets (1M+, e.g., HOT3D, InterHand2.6M) at 1.5%; and small-scale datasets (100K+, e.g., FreiHand, HO3D) at 15–20%. This yields approximately 0.7M training meshes (Figure 10). This multi-source setup covers a diverse range of hand shapes, poses, and object interactions, ensuring generalization to the gesture space of *EgoEMG*.

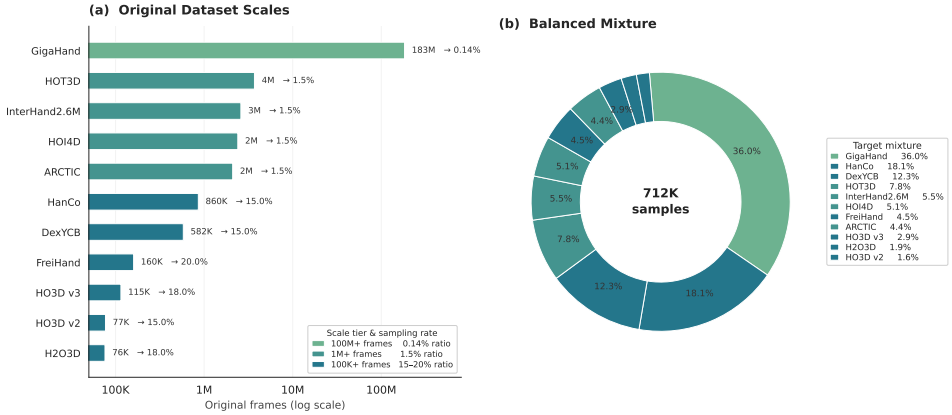


Figure 10: Composition of the *markers2mano* training data. Left: original frame counts of 11 public datasets (log scale), spanning 76K to 183M frames. Each dataset is assigned a sampling ratio by scale tier—0.14% (GigaHand, 100M+), 1.5% (1M+), or 15–20% (100K+)—shown as annotations. Right: target distribution of the training mixture after applying the tiered sampling ratios, yielding approximately 0.7M training meshes.

The loss function combines two terms evaluated on the MANO mesh output:

$$\mathcal{L} = \lambda_{\text{verts}} \frac{1}{V} \sum_{v=1}^{V} \|\mathbf{v}_v^{\text{pred}} - \mathbf{v}_v^{\text{gt}}\|_2^2 + \lambda_{\text{markers}} \frac{1}{N} \sum_{n=1}^{N} \|\mathbf{m}_n^{\text{pred}} - \mathbf{m}_n^{\text{aug}}\|_2^2, \tag{2}$$

where $\mathbf{v}^{\text{pred}}$, $\mathbf{v}^{\text{gt}}$ are the predicted and ground-truth MANO mesh vertices ($V = 778$), and $\mathbf{m}^{\text{pred}}$, $\mathbf{m}^{\text{aug}}$ are the subset of 21 predicted mesh vertices corresponding to the virtual markers and the augmented input markers ($N = 21$), respectively. We set $\lambda_{\text{verts}} = 1.0$ and $\lambda_{\text{markers}} = 5.0$, placing stronger emphasis on accurate marker-level reconstruction.

Data augmentation. To ensure robustness against realistic motion-capture noise, we apply a multi-step augmentation pipeline to the input marker positions prior to model ingestion. The pipeline comprises eight perturbation types organized into three categories; 50% of the samples in each batch bypass augmentation entirely, thereby retaining an unperturbed signal.

- **Structural**: (1) Bone-length perturbation—individual bone segments are randomly scaled by up to $\pm 5\%$ of their length, simulating soft-tissue artifacts. (2) Random global scaling—after root-centering, all coordinates are multiplied by a factor in $[0.6, 1.4]$, varying the apparent hand scale.

- **Identity**: (3) Marker swap—spatially proximate markers (within 15 mm) are randomly exchanged with probability 0.3 (capped at three swaps per frame). (4) Marker dropout—up to three markers are dropped and replaced with positions interpolated from their neighbors. (5) Neighborhood blending—each marker is blended with a random convex combination of its neighboring mesh vertices, retaining a self-weight of 0.5.
- **Noise**: (6) Gaussian noise—i.i.d. $\mathcal{N}(0, 1\,\mathrm{mm}^2)$ noise is added to each coordinate, and per-marker dropout is applied with 10% probability. (7) Marker drift—up to three markers receive a systematic spatial offset of up to 5 mm in a random direction. (8) Marker spike—a single marker is randomly displaced to an anomalously distant position (2–5$\times$ the hand scale) with 10% probability, simulating gross tracking failures.

To assess the fidelity of markers2mano reconstruction, we report the reconstruction error of the original markers and virtual vertices of the reconstructed meshes. *markers2mano* shows 3.2 mm of per-vertex error on the public dataset, and 4.3 mm of per-marker error on the *EgoEMG*. We also manually reproject meshes onto egocentric RGB frames sampled randomly to validate the label quality, as shown in Figure 11.

B.2 Wrist Articulation Angle Computation

Wrist articulation angles are computed directly from motion-capture markers, independent of the MANO reconstruction pipeline. Markers attached to each EMG armband (4 markers on the right wrist, 5 on the left) define a forearm coordinate frame, from which we compute two wrist degrees of freedom: flexion/extension θ_{fe} and radial/ulnar deviation θ_{ru}.

Three non-collinear armband markers define a forearm frame with normalized basis vectors $\hat{\mathbf{F}} = \mathrm{normalize}(\mathbf{F})$ (forearm direction) and $\hat{\mathbf{L}} = \mathrm{normalize}(\mathbf{L})$ (leftward across dorsal surface), with normal $\hat{\mathbf{N}} = \mathrm{normalize}(\hat{\mathbf{F}} \times \hat{\mathbf{L}})$. Let $\hat{\mathbf{H}}$ be the unit vector from wrist to middle finger MCP, and $\hat{\mathbf{H}}_{\text{proj}} = \hat{\mathbf{H}} - (\hat{\mathbf{H}} \cdot \hat{\mathbf{N}})\hat{\mathbf{N}}$ its projection onto the $(\hat{\mathbf{F}}, \hat{\mathbf{L}})$ plane. The wrist angles are:

$$\theta_{\text{fe}} = \arcsin(\hat{\mathbf{H}} \cdot \hat{\mathbf{N}}), \qquad \theta_{\text{ru}} = \mathrm{atan2}\big(\hat{\mathbf{H}}_{\text{proj}} \cdot \hat{\mathbf{L}},\ \hat{\mathbf{H}}_{\text{proj}} \cdot \hat{\mathbf{F}}\big),$$

with extension and radial deviation defined as positive. For the left hand, $\mathbf{N} = \mathbf{L} \times \mathbf{F}$ to maintain consistent sign conventions.

B.3 Joint Angle Conversion from MANO to UmeTrack / EMG2Pose Format

In addition to MANO parameters $(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{t})$, *EgoEMG* provides per-frame joint angles in the 20-DoF scalar format used by EMG2Pose and UmeTrack [Salter et al., 2024, Han et al., 2022], enabling direct comparison with prior EMG-to-pose benchmarks that use scalar joint angle representations. The conversion is formulated as landmark-level inverse kinematics.

Conversion pipeline. Given MANO pose and shape parameters for a frame, we first run MANO forward kinematics to obtain 21 joint positions. We align these into the UmeTrack coordinate frame via a fixed transformation (scale, global orientation, and translation) precomputed by jointly optimizing alignment between the two models' rest-pose meshes. We then optimize 20 scalar joint angles $\mathbf{a} \in \mathbb{R}^{20}$ such that the UmeTrack forward kinematics produces landmark positions matching the aligned MANO targets:

$$\mathcal{L}(\mathbf{a}) = \frac{1}{20} \sum_{i=1}^{20} \big\| \mathbf{p}_i^{\text{umetrack}}(\mathbf{a}) - \mathbf{p}_i^{\text{mano}} \big\|_2^2, \tag{3}$$

where $\mathbf{p}_i^{\text{mano}}$ are the aligned MANO joint positions (after applying the fixed scale, rotation, and translation) and $\mathbf{p}_i^{\text{umetrack}}(\mathbf{a})$ are the corresponding UmeTrack landmarks computed via the UmeTrack forward kinematics (scalar joint angles $\rightarrow$ axis-angle rotations via Rodrigues $\rightarrow$ skinning). The 21 MANO joints are mapped to 20 UmeTrack landmarks via a predefined correspondence; the MANO wrist joint is excluded because it coincides with the UmeTrack coordinate origin and carries no pose information.

Optimization. We minimize $\mathcal{L}(\mathbf{a})$ with L-BFGS (strong Wolfe line search, learning rate 0.1, 100 outer steps with up to 50 inner iterations each). Joint range constraints from the UmeTrack hand model are enforced via sigmoid reparameterization: $\mathbf{a} = \mathbf{a}_{\min} + (\mathbf{a}_{\max} - \mathbf{a}_{\min}) \cdot \sigma(\mathbf{z})$, where $\mathbf{z}$ is the unconstrained optimizable variable. The optimization runs per-episode on GPU with batch size 256; episodes exceeding GPU memory are processed in smaller chunks.

Output. The resulting 22-DoF prediction target combines 20 finger joint angles (indices 0–19) and 2 wrist articulation angles (indices 20–21). The mean per-landmark fitting error of the optimized UmeTrack joints relative to the MANO targets is 2.8 mm, confirming that the conversion introduces negligible additional error. Table 6 lists the complete semantic ordering per dimension. The conversion script is included in the dataset release.

Table 6: Complete 22-DoF joint angle target semantics. Each dimension is reported in degrees. FE = flexion/extension, AA = abduction/adduction. Indices 0–3 cover the thumb, 4–7 the index finger, 8–11 the middle finger, 12–15 the ring finger, and 16–19 the pinky.

Index	Finger	Joint	Motion
0	Thumb	CMC	FE
1	Thumb	CMC	AA
2	Thumb	MCP	FE
3	Thumb	IP	FE
4	Index	MCP	AA
5	Index	MCP	FE
6	Index	PIP	FE
7	Index	DIP	FE
8	Middle	MCP	AA
9	Middle	MCP	FE
10	Middle	PIP	FE
11	Middle	DIP	FE
12	Ring	MCP	AA
13	Ring	MCP	FE
14	Ring	PIP	FE
15	Ring	DIP	FE
16	Pinky	MCP	AA
17	Pinky	MCP	FE
18	Pinky	PIP	FE
19	Pinky	DIP	FE
20	–	Wrist	Flexion/Extension
21	–	Wrist	Radial/Ulnar deviation

We show reconstruction samples in Figure 11.

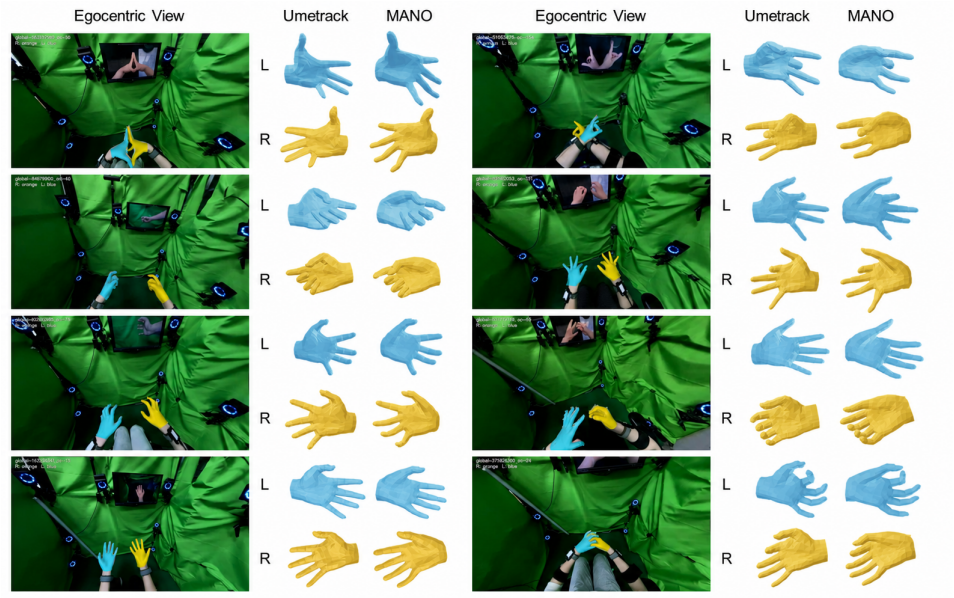


Figure 11: Mesh reprojection and visualization of UmeTrack and MANO mesh reconstruction on *EgoEMG*. The reconstructed MANO meshes (from *markers2mano*) are overlaid on egocentric RGB frames, demonstrating accurate hand pose reconstruction and realistic mesh-to-image alignment.

C Training and implementation details

C.1 EMGFormer Training Details

Optimization. All EMGFormer variants are trained for 200 epochs with AdamW [Loshchilov and Hutter, 2019] ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay 10^{-4}), initial learning rate 10^{-4} with cosine annealing to 5×10^{-6}, batch size 800, and bf16 mixed precision, minimizing an L1 (MAE) loss with an additional fingertip-distance loss term (weight 0.01). We use a linear warmup for the first 500 steps. All experiments were conducted on six NVIDIA RTX 4090 GPUs (24 GB each).

Data augmentation. Applied on-the-fly during training: channel dropout (25% probability per channel), frequency masking (3 masks of up to 128 bins), additive Gaussian noise (SNR 25–35 dB with 50% probability), and temporal jitter ($\pm$40 ms).

Architecture and per-size configuration. Table 7 summarizes the architecture and training configuration per EMGFormer variant.

Table 7: Architecture and training configuration per EMGFormer variant. "Params" includes the TDS encoder, Transformer decoder, and regression head.

Variant	Layers	Hidden dim	Heads	FFN dim	Params	GPU-hours
EMGFormer-S	3	256	4	512	3.5M	2
EMGFormer-M	6	256	8	1024	6.6M	4
EMGFormer-L	8	384	12	1536	16.3M	8

TDS featurizer. The TDS-style temporal convolutional frontend [Hannun et al., 2019] maps the raw 2 kHz EMG window ($T = 7790$ samples, 16 channels) to a 256-dimensional feature sequence. The architecture consists of two initial strided Conv1d blocks followed by two TDS stages. Each TDS stage begins with a strided Conv1d for temporal subsampling, followed by a TDSConvEncoder composed of alternating depth-separable temporal convolutions (reshaping the channel axis into a 2D feature map) and channel-wise fully-connected blocks. Squeeze-and-excitation [Hu et al., 2018] is applied globally after each TDS stage. The layer-wise configuration is:

1. Conv1d: $16 \to 256$, kernel 11, stride 5

2. Conv1d: $256 \to 256$, kernel 5, stride 2
3. TDS Stage 1: in-conv kernel 9, stride 5; 1 block of TDSConv (kernel width 5, channels 8, feature width 32)
4. TDS Stage 2: in-conv kernel 3, stride 1; 1 block of TDSConv (kernel width 3, channels 8, feature width 32)

The final output is a $(256, 146)$ feature map, corresponding to an effective frame rate of approximately 37 Hz with a receptive field spanning the full input window.

C.2 Vision and Fusion Training Details

Vision baseline. We provide six fully fine-tuned generic vision backbones: ResNet-{18, 50, 152} and ViT-{Small, Base, Large}. ResNet backbones are initialized from ImageNet-pretrained weights, while ViT backbones are initialized from DINOv2-pretrained weights. Each generic backbone extracts a feature vector from the window center frame crop (256×256), which is fed through a prediction head to regress 22 joint angles. The prediction head is a 2-layer MLP: Linear($d_{\text{backbone}} \to 512$) $\to$ ReLU $\to$ Dropout(0.1) $\to$ Linear($512 \to 22$). All generic vision models are trained with the same L1 (MAE) loss plus fingertip-distance loss (weight 0.01) as the EMG baseline. We additionally fine-tune WiLoR [Potamias et al., 2024] as a hand-specialized baseline using its pretrained MANO-based checkpoint, with the same center-frame supervision target and output dimensionality. Generic backbones use AdamW [Loshchilov and Hutter, 2019] with cosine annealing and bf16 mixed precision on 6 NVIDIA RTX 4090 GPUs; WiLoR uses AdamW with learning rate 10^{-5} and cosine annealing over 30 epochs. Table 8 lists the per-backbone configuration.

Table 8: Vision backbone configuration. All models take 256×256 center-frame crops as input and are evaluated on the same 22-DoF center-frame target. Generic ResNet/ViT models share the same MLP prediction head; WiLoR uses its native hand-pose backbone with a 22-DoF regression head.

Backbone	Params	d_{backbone}	LR	Batch size	Epochs
ResNet-18	11.5M	512	10^{-3}	900	150
ResNet-50	23.5M	2048	10^{-3}	400	150
ResNet-152	58.2M	2048	10^{-3}	180	150
ViT-S/14	21.9M	384	10^{-4}	360	150
ViT-B/14	86.2M	768	10^{-4}	160	150
ViT-L/14	303.8M	1024	10^{-4}	50	150
WiLoR	631.6M	1576	10^{-5}	18	30

Fusion baseline. The fusion model uses the residual architecture described in §4.5. The vision pathway (backbone, projection layer, and vision head) can be initialized from a vision-only checkpoint, and the EMG pathway (EMGFormer-S, comprising a TDS featurizer and a Transformer decoder) can be initialized from an EMG-only checkpoint. The fusion projection and the residual regression head are randomly initialized, with the last layer of the head set to zero so that the model starts from the vision-only baseline ($\Delta \mathbf{y}_{\text{emg}} \approx 0$). The model predicts only the center frame of the window. Training is performed with AdamW [Loshchilov and Hutter, 2019] (learning rate 1×10^{-4}, cosine annealing schedule) under the same L1 (MAE) plus fingertip-distance loss, a batch size ranging from 64 to 600 depending on the backbone, and 200–300 epochs, using bf16 mixed precision on six NVIDIA RTX 4090 GPUs.

D Extended experimental results

D.1 Per-Hand Breakdown of EMG Baseline Results

Table 9 reports per-hand MAE for all methods on *EgoEMG*, complementing the aggregate results in Table 2.

Left and right hand MAE agree within 2.0° across all models and splits, confirming that the shared-weight bilateral design and left-to-right mirroring strategy produce symmetric performance. The per-hand consistency supports the validity of processing both hands through a single model.

Table 9: Per-hand EMG-to-pose MAE on *EgoEMG*. MAE is reported in degrees; lower is better. Left and right hands are processed independently with shared model weights, and left-hand EMG is mirrored to the right-hand convention. Rows are grouped by method family; shaded rows highlight EMGFormer variants.

Method	Hand	Gesture	User	Both
EMGFormer variants				
EMGFormer-S	Left	12.4	14.7	16.5
	Right	13.2	16.6	18.4
EMGFormer-M	Left	11.5	15.5	17.0
	Right	12.1	16.6	17.8
EMGFormer-L	Left	11.4	15.3	17.6
	Right	12.0	16.0	17.7
Traditional architectures				
vemg2pose	Left	14.7	15.7	17.2
	Right	15.2	17.0	17.5
NeuroPose	Left	15.9	15.4	15.7
	Right	15.7	16.0	16.9

D.2 Center-Frame EMG-Only Reference

To partially normalize the supervision granularity gap discussed in §4.2, we additionally evaluate EMG-only models on the window center frame rather than over the full predicted trajectory. Table 10 reports these center-frame MAE values using the same gesture, user, and both splits as the main benchmark. The center-frame results are comparable but not identical to the trajectory-level evaluation: they slightly change the ranking among model sizes, with EMGFormer-S achieving the best average MAE and the best user-split result, while EMGFormer-L remains strongest on the gesture and both splits. All three variants remain tightly clustered within 0.4° average MAE.

Table 10: EMG-only center-frame reference results on *EgoEMG*. MAE is reported in degrees; lower is better. Columns report per-user mean $\pm$ standard deviation (left+right pooled). Avg is the per-sample overall MAE across all test splits. These runs use the same center-frame target as the vision-only and fusion models, but remain EMG-only. Bold indicates the best mean per column.

Method	Params	Gesture	User	Both	Avg.
EMGFormer-S	3.5M	$12.5_{\pm 1.7}$	$\mathbf{15.5_{\pm 2.5}}$	$17.0_{\pm 1.4}$	$\mathbf{14.5}$
EMGFormer-M	6.6M	$12.9_{\pm 1.7}$	$15.7_{\pm 2.6}$	$17.7_{\pm 1.6}$	14.9
EMGFormer-L	16.3M	$\mathbf{12.6_{\pm 1.6}}$	$16.1_{\pm 1.6}$	$\mathbf{17.4_{\pm 1.2}}$	14.6

D.3 EMG2Pose Benchmark Results

Table 11 reports EMGFormer results on the EMG2Pose dataset standard splits.

EMGFormer-L achieves the best user+stage MAE of $12.25^\circ \pm 1.08^\circ$, improving over the vemg2pose baseline [Salter et al., 2024] ($15.8^\circ \pm 1.4^\circ$) by 22% while using $2.7\times$ more parameters. SensingDynamics [Sîmpetru et al., 2024] and NeuroPose [Liu et al., 2021] report user+stage MAE of 18.7° and 17.5°, respectively, on the full EMG2Pose benchmark. CLDM [Verma, 2025] reports $14.7^\circ \pm 1.4^\circ$, also on the full dataset. VQ-MyoPose [Lovecchio et al., 2025][‡] was evaluated on a 12-subject subset and is not directly comparable. EMGFormer-L reduces the stage split MAE to $9.28^\circ \pm 1.09^\circ$, a 39% improvement over vemg2pose, indicating stronger temporal pattern modeling for unseen gesture stages.

D.4 Per-Joint Error Analysis

Table 12 presents the per-finger and per-joint-group mean absolute error (MAE) for EMGFormer variants on *EgoEMG*. Across all three model sizes, distal joints and wrist deviation yield the lowest errors (best MAE 10.0° for both), whereas the pinky and mid-phalanx groups are the most challenging

Table 11: EMG-only baseline results on EMG2Pose test splits. MAE is reported in degrees; lower is better. EMGFormer values report per-user mean ± standard deviation across users (n=158 for Stage, n=20 for User and User+Stage). † Numbers reported in the original papers; all evaluated on the full dataset. ‡Evaluated on a 12-subject subset of EMG2Pose (12/27 stages); not directly comparable to full-dataset results. Shaded rows highlight EMGFormer variants; bold indicates the best EMGFormer value per column.

Method	Params	User	Stage	User+Stage
Prior work				
SensingDynamics† [Sîmpetru et al., 2024]	–	$15.5_{\pm 1.4}$	$18.8_{\pm 1.6}$	$18.7_{\pm 1.6}$
NeuroPose†	6.4M	$13.2_{\pm 1.1}$	$17.2_{\pm 1.7}$	$17.5_{\pm 1.5}$
vemg2pose†	6.0M	$12.2_{\pm 1.3}$	$15.2_{\pm 1.6}$	$15.8_{\pm 1.4}$
CLDM† [Verma, 2025]	–	$11.3_{\pm 1.0}$	$14.3_{\pm 1.5}$	$14.7_{\pm 1.4}$
EMGFormer variants				
EMGFormer-S	3.5M	$12.45_{\pm 1.06}$	$11.10_{\pm 1.19}$	$12.34_{\pm 1.07}$
EMGFormer-M	6.6M	$12.40_{\pm 1.08}$	$10.18_{\pm 1.13}$	$12.39_{\pm 1.08}$
EMGFormer-L	16.3M	$\mathbf{12.26_{\pm 1.08}}$	$\mathbf{9.28_{\pm 1.09}}$	$\mathbf{12.25_{\pm 1.08}}$

(best MAE 16.7° and 20.7°, respectively). This pattern is consistent with hand kinematics and sensing difficulty. Distal joints tend to have a more constrained range of motion, making their angles easier to regularize from EMG signals. In contrast, mid-phalanx joints capture a larger portion of finger flexion and therefore exhibit greater motion variability. The higher error on the pinky is also expected: pinky motion is often weaker, more coupled with the ring finger, and less independently represented in wrist-level EMG, making it harder to distinguish from neighboring-finger activation patterns.

Table 12: Per-joint MAE (°) on *EgoEMG*, left and right hands pooled within each split, then per-sample weighted average across gesture, user, and both splits. Lower is better. Bold indicates the best mean per row.

Joint group	EMGFormer-S	EMGFormer-M	EMGFormer-L
Distal	10.2	10.1	**10.0**
Wrist deviation	10.8	10.1	**10.0**
Thumb	12.8	**12.3**	**12.3**
Index	13.4	**13.3**	13.3
Proximal	13.7	**13.3**	13.4
Ring	15.2	14.7	**14.6**
Middle finger	15.2	14.8	**14.8**
Wrist flexion	15.7	14.9	**14.6**
Pinky	17.3	**16.7**	17.0
Mid-phalanx	21.6	**20.7**	**20.7**

For comparison, Table 13 details the corresponding per-joint MAE on the EMG2Pose dataset. Because EMG2Pose lacks wrist articulation labels, our evaluation here is restricted to finger joints.

Table 13: Per-joint MAE (°) on the EMG2Pose dataset, per-sample weighted average across all three test splits (stage, user, and user+stage). Because EMG2Pose lacks wrist labels, only finger joints are reported. Bold indicates the best mean per row.

Joint group	EMGFormer-S	EMGFormer-M	EMGFormer-L
Proximal	9.3	8.9	**8.5**
Index	10.8	10.3	**9.8**
Thumb	11.1	10.6	**9.9**
Middle finger	11.1	10.5	**10.0**
Ring	11.1	10.6	**10.1**
Distal	13.3	12.8	**12.2**
Pinky	13.5	12.9	**12.3**
Mid-phalanx	14.2	13.4	**12.6**

A cross-dataset comparison shows that mid-phalanx and pinky joints are consistently among the most challenging groups across both benchmarks. However, the easiest groups differ between datasets: proximal joints achieve the lowest MAE on EMG2Pose, whereas distal joints and wrist deviation

are easiest on *EgoEMG*. The overall error magnitudes are also higher on *EgoEMG*—ranging from 10.0° to 21.6°, compared with 8.5° to 14.2° on EMG2Pose—which likely reflects differences in participant count, gesture diversity, sensor setup, and the inclusion of wrist articulation in *EgoEMG*'s 22-DoF prediction target.

D.5 Per-Gesture Modality Complementarity Analysis

To characterize where EMG provides complementary information beyond vision, we evaluate all three modalities—vision-only, EMG-only, and F-RN18+S fusion—on a per-gesture basis across the full test set (all splits, both hands). Following the vocabulary in Appendix A.1, we group the 60 gesture classes into three semantic families: single-hand (30 gestures), symmetric bimanual (18), and asymmetric bimanual (12). One bimanual gesture (raw, unconstrained) is excluded from this analysis due to its unbounded pose distribution. Table 14 reports sample-weighted per-family averages across all three modalities.

Table 14: Per-gesture-family modality comparison on *EgoEMG* (F-RN18+S fusion). Values are sample-weighted averages across all test splits; MAE is in degrees (lower is better). $\Delta = \text{MAE}_{\text{vision}} - \text{MAE}_{\text{fusion}}$; positive Δ indicates fusion improvement. Gesture families follow the vocabulary in Appendix A.1.

Gesture family	#G	#Samp.	Vis.	EMG	Fusion	Δ
Single-hand	30	2,406	6.0	15.0	5.5	$\mathbf{+0.50}$
Symmetric bimanual	17	1,288	5.9	14.0	5.5	$+0.38$
Asymmetric bimanual	12	584	6.8	14.6	6.4	$+0.41$
Overall	59	4,278	6.1	14.6	5.6	$+0.46$

Three patterns emerge. First, the fusion gain is systematic: fusion improves over vision-only on 54 of 59 analyzed gesture classes, ruling out the possibility that the aggregate improvement is driven by a few outlier gestures. Second, the largest family-level gain occurs on single-hand gestures ($\Delta = +0.50^\circ$), followed by asymmetric bimanual ($\Delta = +0.41^\circ$). This suggests that EMG is most useful when the prediction depends on subtle within-hand articulation or asymmetric cross-hand coordination that is harder to infer from a single RGB crop. Symmetric bimanual gestures show the smallest gain ($\Delta = +0.38^\circ$), as two-hand interactions of this type are often easier to discriminate visually. Third, even for the few gestures where fusion slightly underperforms vision-only (Rock, -0.22°; Typing, -0.21°; Clap, -0.12°; FingerPullLeft, -0.04°; CrossHand, -0.01°), the degradation is negligible and involves large-scale hand silhouettes whose visual features are already highly discriminative. Figure 12 visualizes the per-gesture complementarity: nearly all points lie below the diagonal, confirming that fusion provides a systematic, if modest, correction across the gesture vocabulary.

Table 15 provides the complete per-gesture breakdown for all 59 analyzed gesture classes, grouped by semantic family and sorted by descending fusion gain within each family.

D.6 Fusion Benefit vs. Hand Self-Occlusion

To understand when EMG–vision fusion provides the greatest benefit, we stratify test-set predictions by the degree of hand self-occlusion.

Self-occlusion score. We quantify per-frame hand self-occlusion via z-buffer rasterization of the ground-truth MANO mesh in camera space. Concretely, for each test frame:

1. The MANO mesh vertices are transformed from world coordinates to camera coordinates using the calibrated extrinsic matrix ($\mathbf{v}_{\text{cam}} = \mathbf{R}_{C\leftarrow W}\,\mathbf{v}_W + \mathbf{t}_{C\leftarrow W}$).
2. All mesh triangles are rasterized into a depth buffer (z-buffer) at the native video resolution via pinhole projection with the calibrated intrinsic matrix $\mathbf{K}$. For each pixel covered by a triangle, the interpolated depth is stored if it is smaller than the current buffer value.
3. A vertex is deemed *visible* if, within a 5×5 pixel neighbourhood of its projected location, any z-buffer entry agrees with the vertex depth to within $\epsilon = 5\,\text{mm}$.
4. Each triangle's 3-D surface area (half the cross-product norm of two edge vectors) is distributed equally to its three vertices, yielding a per-vertex area weight w_i.

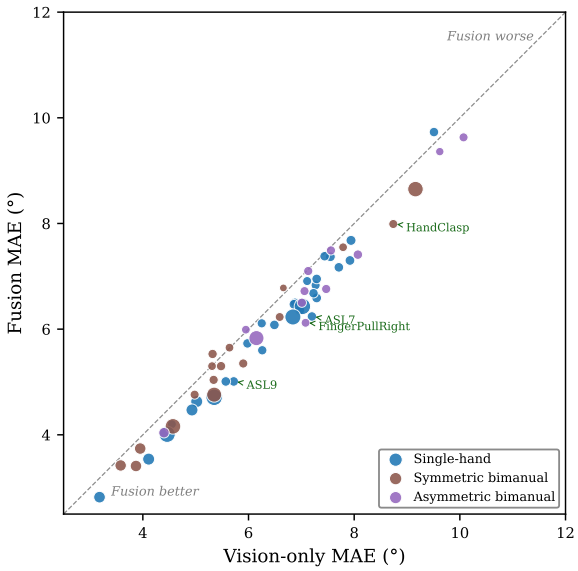


Figure 12: Per-gesture vision-only vs. fusion MAE on *EgoEMG* (F-RN18+S). Each point is one gesture class; point size reflects sample count; color indicates the semantic gesture family. The diagonal marks parity; points below the diagonal indicate fusion improvement. Top-4 most-improved gestures and the most notably worsened gestures are annotated.

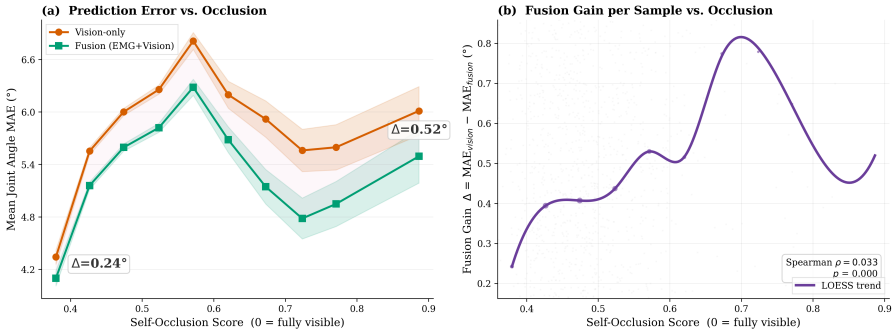


Figure 13: Fusion benefit as a function of hand self-occlusion on the *EgoEMG* test set (17,168 samples, stride reduced to $\frac{1}{4}$ of the evaluation window for denser coverage). **(a)** Mean joint-angle MAE of vision-only and EMG+Vision fusion across fixed-width (0.05) occlusion bins. The annotated Δ values indicate the fusion gain at the lowest and highest occlusion bins. **(b)** Per-bin fusion gain ($\Delta = \text{MAE}_{\text{vision}} - \text{MAE}_{\text{fusion}}$) with LOESS trend line (Spearman $\rho = 0.049$, $p = 0.001$). Error bands in (a) and the trend confidence in (b) denote ± 1 SEM.

5. The **self-occlusion score** is defined as

$$s_{\text{occ}} = 1 - \frac{\sum_{i \in \mathcal{V}_{\text{vis}}} w_i}{\sum_{i=1}^{V} w_i}, \tag{4}$$

where $\mathcal{V}_{\text{vis}}$ is the set of visible vertices. A score of 0 means the hand is fully visible; a score of 1 means it is fully occluded from the camera viewpoint.

Results. Figure 13(a) shows that both vision-only and fusion MAE vary with occlusion, while the gap between them—the fusion gain—is consistently positive across the full range, from $\Delta = 0.24°$ at the lowest occlusion bin to $\Delta = 0.52°$ at the highest. The largest fusion benefit occurs at moderate-to-high occlusion ($\Delta \approx 0.8°$ near occlusion ≈ 0.72), where visual evidence is substantially degraded. Panel (b) visualizes the same trend at the per-bin level with a LOESS smooth, showing a statistically significant positive association between occlusion and fusion gain (Spearman $\rho = 0.049$, $p = 0.001$).

Table 15: Full per-gesture modality breakdown on *EgoEMG* using F-RN18+S fusion. MAE is in degrees. $\Delta = \text{MAE}_{\text{vis}} - \text{MAE}_{\text{fus}}$. Gestures are grouped by family and sorted by descending Δ within each family. Gesture descriptions are provided in Appendix A.1.

ID	Name	N	Vis.	EMG	Fus.	Δ
Single-hand gestures						
6	ASL7	36	7.2	16.1	6.2	+0.96
8	ASL9	38	5.7	17.5	5.0	+0.71
12	ILY	40	7.3	19.2	6.6	+0.70
13	IndexBow	34	6.3	20.7	5.6	+0.66
20	PinchMiddle	324	5.3	12.2	4.7	+0.64
26	Thumb	40	7.9	24.8	7.3	+0.62
7	ASL8	42	7.0	14.8	6.4	+0.61
28	nocontact_free	292	6.8	14.2	6.2	+0.61
18	Nine	320	7.0	18.1	6.4	+0.59
17	MiddleBow	92	4.1	11.1	3.5	+0.57
16	JoystickSlide	38	5.6	19.5	5.0	+0.56
23	RingAndThumb	36	7.2	16.5	6.7	+0.55
9	Claw3	38	7.7	17.3	7.2	+0.54
27	nocontact_disperse_palm	92	4.9	12.2	4.5	+0.46
15	JoystickCircle	308	4.5	12.2	4.0	+0.45
2	ASL3	30	7.3	19.6	6.8	+0.44
11	FreeAction	38	6.5	15.8	6.1	+0.41
5	ASL6	40	6.9	18.1	6.5	+0.39
14	IndexMiddleClaw	92	5.0	12.9	4.6	+0.39
1	ASL2	34	6.9	17.7	6.5	+0.38
19	PalmYaw	82	3.2	9.8	2.8	+0.36
4	ASL5	28	4.5	14.6	4.2	+0.35
24	RingBow	38	7.3	16.9	7.0	+0.34
0	ASL1	44	7.9	17.6	7.7	+0.26
10	Claw5	36	6.0	15.6	5.7	+0.25
22	Rest	30	7.1	17.4	6.9	+0.20
21	PinkyBow	40	7.5	16.6	7.4	+0.18
3	ASL4	32	6.2	18.0	6.1	+0.14
29	nocontact_grab	36	7.4	18.6	7.4	+0.06
25	Rock	36	9.5	17.5	9.7	-0.22
Symmetric bimanual gestures						
38	HandClasp	30	8.7	16.4	8.0	+0.75
52	MiddleOppo	226	5.3	12.3	4.8	+0.59
39	HandRub	32	5.9	16.0	5.3	+0.55
36	FistBump	256	9.2	23.7	8.7	+0.51
44	Prayer	78	3.9	9.5	3.4	+0.46
42	PalmStack	262	4.6	10.2	4.2	+0.41
47	SymSwing	28	6.6	15.7	6.2	+0.36
40	IndexTapping	32	5.3	13.0	5.0	+0.30
37	Gaming	26	7.8	15.9	7.5	+0.24
46	SymOpen	30	5.0	12.6	4.8	+0.22
51	Kiss	78	4.0	10.4	3.7	+0.21
48	ThumbWrestle	34	5.5	14.6	5.3	+0.18
35	FingerTipTouch	76	3.6	7.5	3.4	+0.16
32	CrossStretch	26	5.3	11.9	5.3	+0.01
31	CrossHand	26	5.6	12.8	5.7	-0.01
30	Clap	14	6.7	14.3	6.8	-0.12
49	Typing	34	5.3	11.9	5.5	-0.21
Asymmetric bimanual gestures						
34	FingerPullRight	28	7.1	13.6	6.1	+0.96
58	PinchWring	32	7.5	17.8	6.8	+0.71
55	PairClaw	36	8.1	17.4	7.4	+0.66
43	PinkyHook	32	7.0	16.0	6.5	+0.51
53	Beijing	30	10.1	19.2	9.6	+0.44
41	PalmRoll	54	4.4	12.1	4.0	+0.36
59	ThumbOppo	30	7.1	17.5	6.7	+0.34
45	Squeeze	230	6.2	12.4	5.8	+0.32
54	Checky	22	9.6	21.1	9.4	+0.26
56	PairOK	32	7.6	15.6	7.5	+0.07
57	Picture	32	7.1	16.7	7.1	+0.03
33	FingerPullLeft	26	6.0	13.6	6.0	-0.04

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction frame the paper primarily as a new dataset and benchmark, and the strongest empirical claims are restricted to the initial EMG baselines that are actually reported; see Sections 1, 3, 4, 6, and 7.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper explicitly discusses claim scope, incomplete benchmark tracks, and missing integration details in Section 7.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not present theoretical results or formal proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper describes the benchmark protocol, split design, baseline families, per-variant hyperparameters, optimizer settings, data augmentation, compute requirements, and provides reproduction commands in Appendix C.1. The repository contains all training code, evaluation scripts, and configuration files needed to reproduce the reported baseline results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The repository contains baseline code and configuration files. The dataset will be released with an anonymized access link during review, including download script, checksum verification, split files, and example loading code (Appendix A.3).

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the benchmark split design, baseline families, per-variant hyperparameters, optimizer settings, data augmentation, and compute requirements in Appendix C.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Primary tables report mean $\pm$ standard deviation across users or user groups, as stated in the corresponding captions. We do not claim multi-seed error bars or paired significance tests in the current submission.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix C.1 reports GPU type, memory, per-variant GPU-hours, and total compute (approximately 128 GPU-hours on NVIDIA RTX 4090 GPUs with 24 GB memory).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper presents a dataset-and-benchmark contribution with explicit discussion of scope and release considerations, and nothing in the described workflow intentionally departs from the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper motivates positive use cases (assistive interfaces, AR/VR, prosthetics). Potential negative impacts include surveillance via EMG-based activity inference; the dataset's research-use license and anonymization mitigate this risk.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: The paper includes an Ethics Statement section describing informed consent, IRB approval, anonymization procedures, and a research-use license.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper cites all prior datasets and methods used. The dataset license (CC-BY-NC 4.0) and code license (MIT) are specified in Appendix A.3. Third-party assets (MANO model, training data sources for markers2mano) retain their original licenses, which are documented in the release package.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper describes the new dataset with structured documentation (Appendix A.3), including file schema, split definitions, license (CC-BY-NC 4.0 for data, MIT for code), and a datasheet covering composition, collection, preprocessing, uses, distribution, and maintenance.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: The paper includes an Ethics Statement section describing informed consent, compensation at standard hourly rates, and anonymization procedures. Participant instructions are described in Section 3.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [Yes]

Justification: The paper includes an Ethics Statement section covering informed consent, IRB-approved protocol, participant compensation, and data anonymization. Risk disclosure to participants was part of the consent process.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core methods, benchmark design, and reported experiments do not rely on an LLM as an original methodological component.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.