arXiv:2412.02725v1 [cs.CV] 2 Dec 2024

emg2pose: A Large and Diverse Benchmark for Surface Electromyographic Hand Pose Estimation

Sasha Salter*, Richard Warren*, Collin Schlager*, Adrian Spurr, Shangchen Han, Rohin Bhasin†, Yujun Cai, Peter Walkington, Anuoluwapo Bolarinwa, Robert Wang, Nathan Danielson, Josh Merel†, Eftychios Pnevmatikakis, and Jesse Marshall

Reality Labs, Meta

Figure 1: We introduce the *emg2pose* dataset and benchmark to facilitate the development of pose estimation models from sEMG. Our *vemg2pose* model is capable of estimating in real-time hand pose (lower) from held-out users wearing an sEMG wristband (top). See text for further details.

Abstract

Hands are the primary means through which humans interact with the world. Reliable and always-available hand pose inference could yield new and intuitive control schemes for human-computer interactions, particularly in virtual and augmented reality. Computer vision is effective but requires one or multiple cameras and can struggle with occlusions, limited field of view, and poor lighting. Wearable wrist-based surface electromyography (sEMG) presents a promising alternative as an always-available modality sensing muscle activities that drive hand motion. However, sEMG signals are strongly dependent on user anatomy and sensor placement, and existing sEMG models have required hundreds of users and device placements to effectively generalize. To facilitate progress on sEMG pose inference, we introduce the *emg2pose benchmark*, the largest publicly available dataset of high-quality hand pose labels and wrist sEMG recordings. *emg2pose* contains 2kHz, 16 channel sEMG and pose labels from a 26-camera motion capture rig for 193 users, 370 hours, and 29 *stages* with diverse gestures - a scale comparable to vision-based hand pose datasets. We provide competitive baselines and challenging tasks evaluating real-world generalization scenarios: *held-out users, sensor placements*, and *stages*. *emg2pose* provides the machine learning community a platform for exploring complex generalization problems, holding potential to significantly enhance the development of sEMG-based human-computer interactions.

*Equal contribution.
†Work done while at Meta.

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.

1 Introduction

Despite rapid progress in computing hardware and software, current input devices can be inefficient and non-intuitive for new and emerging computing platforms. This is particularly evident for spatial interactions, such as those encountered in virtual and augmented reality, where conventional input devices like controllers, keyboards, and mice do not always offer intuitive control schemes nor sufficient degrees of freedom to enable precise control (e.g., object manipulation). Interactions based on hand movements offer a high-dimensional continuous input that is instinctive, universal, and particularly well suited to spatial interactions [Han et al., 2020]. Furthermore, existing input schemes can be viewed as low dimensional summaries of hand movements, for instance a mouse click tells you that a finger has pressed a button. As such, hand kinematics is a potentially holistic and encompassing modality, covering existing inputs and extending them in a natural manner. High fidelity hand tracking enables various AR/VR applications including gaming [Han et al., 2020], virtual teaching [Shrestha et al., 2022], teleoperations [Santos Carreras, 2012, Darvish et al., 2023], haptics [Scheggi et al., 2015], embodied realism [Wang et al., 2020], sports analytics [Gatt et al., 2020], and healthcare and rehabilitation [Krasoulis et al., 2017].

Given the high utility and broad appeal of effective hand pose estimation, there have been diverse approaches developed across many sensing modalities: optical approaches (e.g. monocular, multi-view, depth-based, motion capture, infrared) using fixed [Cai et al., 2018, Mueller et al., 2018, Ge et al., 2016, Supančič et al., 2018, Park et al., 2020] or head-mounted cameras [Han et al., 2018]; wearable data gloves using magnetic [Parizi et al., 2019], inertial [Yang et al., 2021], capacitative [Truong et al., 2018], and stretch sensors [Shen et al., 2016, Tashakori et al., 2024, Luo et al., 2021]; smart rings [Parizi et al., 2019]; wrist and forearm wearables that use impedance tomography [Zhang and Harrison, 2015], inertial measurement units [Laput and Harrison, 2019], acoustics [Laput et al., 2016] or ultrasound [McIntosh et al., 2017]. Each modality comes with its own hardware constraints and limitations. Optical approaches can struggle with occlusions, poor lighting conditions, and limited field of view, and often require multiple cameras for effective inference, which places constraints on the overall size of the device. Alternatively, glove wearables can hinder dexterous manipulation [Roda-Sales et al., 2020] and forearm wearables typically only support discrete gesture classification.

Surface electromyography (sEMG) sensing on the wrist or forearm provides an appealing alternative that does not struggle with occlusion, field of view, poor lighting, or physical encumberance. sEMG uses electrodes on the skin to measure electrical potentials generated by muscles during movement [Stashuk, 2001]. Specifically, sEMG detects the electrical activity that occurs when spinal motor neurons activate the muscle fibers that drive motion [Merletti and Farina, 2016]. As such, sEMG is particularly well suited for kinematic inference and numerous approaches have been developed [Liu et al., 2021, Quivira et al., 2018, Sosin et al., 2018, Sîmpetru et al., 2022b]. Nevertheless, learning a universal sEMG-to-pose model that *generalizes* to new participants and kinematics is particularly challenging. This is due to sEMG sensing containing many axes of variation, primarily: *user anatomy*, *sensor placement*, and *hand kinematics* [CTRL-labs at Reality Labs et al., 2024, Liu et al., 2021]. User anatomy and sensor placement both influence the locations of the sensors relative to the muscles. Hand kinematics influence what combination of muscle activities are sensed. Given the number of generative dimensions, sEMG models are particularly data-hungry [CTRL-labs at Reality Labs et al., 2024], necessitating many samples across these axes to effectively learn universal models that generalize (see Section 4.4 experiments). Existing datasets lack scale across each of these generative dimensions, thus hindering the development of generic models [Atzori et al., 2014].

Another complication of sEMG is that it encodes muscle activity, which relates more closely to motion than the pose that we would like to recover. As such, direct pose inference from sEMG is particularly challenging (see Section 4), potentially requiring reasoning over long historical sEMG sequences to disambiguate pose from sequences of indirect motion measurements. Extracting relevant information from long sequences, or contexts, in the presence of ambiguity has been extensively explored in fields such as CV [Brunetti et al., 2018, Kirillov et al., 2023, Pan et al., 2018], natural language [Achiam et al., 2023, Kojima et al., 2022, Gu et al., 2021], and robotics [Lauri et al., 2022, Dunion et al., 2024, Jang et al., 2022]. Despite this, prior sEMG works have shown promising results for personalized or single-user pose inference settings [Liu et al., 2021, Sîmpetru et al., 2022b].

To facilitate progress toward developing universal sEMG-to-pose models, we introduce the *emg2pose benchmark* dataset, a large-scale dataset of simultaneously recorded high-fidelity wrist sEMG recordings and hand pose labels. High resolution sEMG recordings are obtained with the sEMG-RD

wrist band [CTRL-labs at Reality Labs et al., 2024](see Section 3.1) and high precision pose labels are obtained from a 26-camera motion capture rig that offers benefits compared to multi-view computer vision [Liu et al., 2021, Sosin et al., 2018]. To our knowledge, this is the largest publicly-available sEMG hand pose dataset, spanning 193 users, 370 hours, and 29 diverse kinematic categories, called *stages*, each containing diverse low-level behaviors, called *gestures*. In addition, the 80M labelled frames that our dataset contains compares favourably with even the newest and largest CV equivalents [Sener et al., 2022, Yu et al., 2020] in both number of frames as well as subjects (see Table 2). We additionally provide three competitive baselines and challenging hand pose inference benchmarks, investigating generalization to unseen *users*, *stages*, and *user-stage* combinations. Instructions regarding accessing and using the *emg2pose benchmark* is provided in `https://github.com/facebookresearch/emg2pose`. Given the high potential impact of sEMG input devices, and the similar research challenges to existing fields, we believe this benchmark will be of great value to the machine learning community.

2 Related Work

sEMG Datasets: There are several publicly available sEMG datasets for tasks other than pose regression, specifically pose (sequence) classification. Data have been collected with either clinical-grade high-density electrode arrays and amplifiers [Amma et al., 2015, Du et al., 2017, Jiang et al., 2021] or consumer-grade hardware that has fewer channels and lower temporal resolution [Palermo et al., 2017]. Clinical-grade hardware offers hundreds of recording channels and acquisition rates >1 kHz but are impractical due to lengthy donning procedures that include shaving the skin before applying conductive gel and the electrode arrays. In contrast, existing consumer-grade hardware is easier to deploy, but is limited by lower bandwidth and channel counts and thus may not provide the level of fidelity required for pose estimation. In contrast, our dataset uses the sEMG-RD band [CTRL-labs at Reality Labs et al., 2024], that can be quickly donned, records 16 channels at >2 kHz and has proven performant for generalized pose classification modelling.

Table 1: The largest publicly available sEMG datasets. We report the maximum number of gestures performed for Du et al. [2017], Atzori et al. [2014] and non amputee subjects for Atzori et al. [2014]. *# Gest.* represents the number of poses participants were instructed to perform. Note that pose category definitions vary significantly across datasets.

Dataset	# Sess.	# Subj.	# Sess. / subj	# Gest.	Inc. Pose
Palermo et al. [2017]	100	10	10	7	No
Amma et al. [2015]	25	5	5	27	No
Du et al. [2017]	69	23	3	22	No
Jiang et al. [2021]	40	20	2	34	No
Atzori et al. [2014]	67	67	1	53	Yes
Ours	751	193	4	50	Yes

Atzori et al. [2014] released a pose regression dataset on consumer-grade emg technologies. This dataset is composed of 37 hours of simultaneously recorded kinematics and sEMG, over a total of 67 sensor placements. In contrast, our dataset includes 193 users, 370 hours, and 751 sessions, which should allow us to train models that generalize favourably across these axes (see Table 1 for scale comparison with existing datasets). Our dataset contains *gesture* categories as well as joint angles.

Pose Regression from sEMG: Several papers have studied pose regression from sEMG, although without open sourcing datasets. Liu et al. [2021] use the MyoBand to estimate hand pose across diverse movements in an 11 participant dataset. They test sEMG decoding models of hand pose across users and sessions with both convolutional (NeuroPose; see Section 3.5) and LSTM architectures. Sîmpetru et al. [2022a] (SensingDynamics; see Section 3.5) use a clinic-grade system to collect several dozen minute datasets in a set of 13 participants. They use a custom 3D convolutional architecture to predict hand joint angles, landmark positions, and grip force, reporting tracking with low error in a held-out test set within each participant. These datasets are limited in scale, with only 11 or 13 participants, and 15 or 20 minutes of data per participant for Liu et al. [2021], Sîmpetru et al. [2022a], respectively, likely limiting generalization across users. In contrast, our dataset includes 193 users and 370 hours, aiding the development of generic models that generalize across users (see Section 4.4).

Pose from Computer Vision: Computer vision (CV) based hand pose estimation has received considerable attention in recent years, usually taking depth, RGB, or both as input, and leveraging large open-sourced datasets [Mueller et al., 2017, 2018, Spurr et al., 2018, 2020, 2021, Wan et al., 2019, Boukhayma et al., 2019]. Labels are either obtained using marker-based motion capture [Fan et al., 2023] - whose markers create an input distributional shift due to lack of markers during

deployment - or using alternate approaches with lower quality labels or inputs, such as multi-view cameras [Zimmermann et al., 2019, Moon et al., 2024], synthetic data [Zimmermann and Brox, 2017], and magnetic sensors [Yuan et al., 2017]. In contrast, motion capture markers afford high quality labels for sEMG, but they do not affect the data from which predictions are generated.

The gestural diversity of CV-based datasets mostly focuses on exploring the full static pose space of the hand [Yuan et al., 2017, Zimmermann and Brox, 2017, Zimmermann et al., 2019], interaction with objects [Fan et al., 2023, Samarth et al., 2020, Hampali et al., 2020] or hand-hand interactions [Moon et al., 2020, 2024]. Conversely, our dataset focuses on movements of the hand because sEMG, unlike CV, is more closely related to motion than pose. Furthermore, our dataset has 80M frames and 193 subjects, comparing favorably to CV datasets (see Table 2, reporting million frames, subjects and fps).

Table 2: Largest CV hand datasets. The *per hand* row counts the data from the left and right hands independently, whereas the *across hands* row pools those data.

Dataset	# Frames	# Subjects	# FPS
Yuan et al. [2017]	2.2M	10	60
Moon et al. [2020]	2.6M	27	5-30
Moon et al. [2024]	1.5M	10	5-30
Samarth et al. [2020]	2.9M	50	n/a
Fan et al. [2023]	2.1M	10	30
Liu et al. [2022]	2.4M	4	n/a
Sener et al. [2022]	111M	53	n/a
Yu et al. [2020]	24M	453	60
Ours (per hand)	80M	193	60
Ours (across hands)	40M	193	60

Pose from Other Modalities: In addition to vision and sEMG, there exists a diverse range of additional wearable approaches to pose inference (see Section 1), which typically focus on pose (sequence) classification. For example, Achenbach et al. [2023] released a dataset for pose classification using commercially available sensor gloves. Other datasets typically use bespoke hardware and are small in scale, with the exception of a large (50 participant, 25 class) dataset available for classification using commercially available smartwatches [Laput and Harrison, 2019].

3 emg2pose Benchmark

3.1 sEMG Device

Data are collected using the 16 channel bipolar sEMG-RD wrist band from CTRL-labs at Reality Labs et al. [2024]. They demonstrate the effectiveness of this device for generalized pose sequence classification across 6400 participants, the largest study to date. This high performance is achieved without the need for high-density sEMG platforms [Amma et al., 2015], with a similar form factor and ease of use to other low-density platforms [Rawat et al., 2016] (see Figs. 1 and 2 for a visual depiction of the device). In contrast to the previously used low-density Thalmic Labs Myo band [Liu et al., 2021] that streams data at 200Hz, across 8 channels and with 8-bits, sEMG-RD senses at 2kHz, across 16 channels and with 12-bits. For more details see Appendix B.1.

3.2 Dataset

Table 3: *emg2pose* dataset statistics, reporting mean and standard deviation. Three separate test sets measure generalization to new users, types of behaviors (stages), and user-behavior combinations (user, stage). Note that the overall hours is the sum of the hours across all splits. The number of hours counts the right-handed and left-handed data separately for each participant.

	Train	Val		Test			Overall
		User	*User, Stage*	*User*	*Stage*	*User, Stage*	
Subjects	158	15	15	20	158	20	193
Unique stages	23	23	6	23	6	6	29
Hours	250.9	21.7	4.6	31.9	54.2	7.0	370.3
Hours / subject	1.6 ± 0.4	1.4 ± 0.5	0.3 ± 0.1	1.6 ± 0.3	0.3 ± 0.1	0.3 ± 0.0	1.9 ± 0.5
Sessions / subject	3.9 ± 0.6	3.8 ± 0.6	3.7 ± 0.6	3.9 ± 0.3	3.8 ± 0.7	3.8 ± 0.5	3.9 ± 0.6

Consenting participants (see Appendix A) stood in a 26 camera motion capture array (Appendix B.2). A research assistant placed 19 motion capture markers on each of the participants' hands (Han et al. [2018]) and an sEMG-RD band on each wrist [CTRL-labs at Reality Labs et al., 2024]. All sEMG and motion capture data were streamed to a real-time data acquisition system at 2kHz and 60 Hz, respectively. We time-aligned device streams using software timestamps, which we found to show less than 10ms relative latency between devices. Motion capture data were post-processed using an offline inverse kinematics (IK) solver to reconstruct the joint angles of the hand (Appendices A

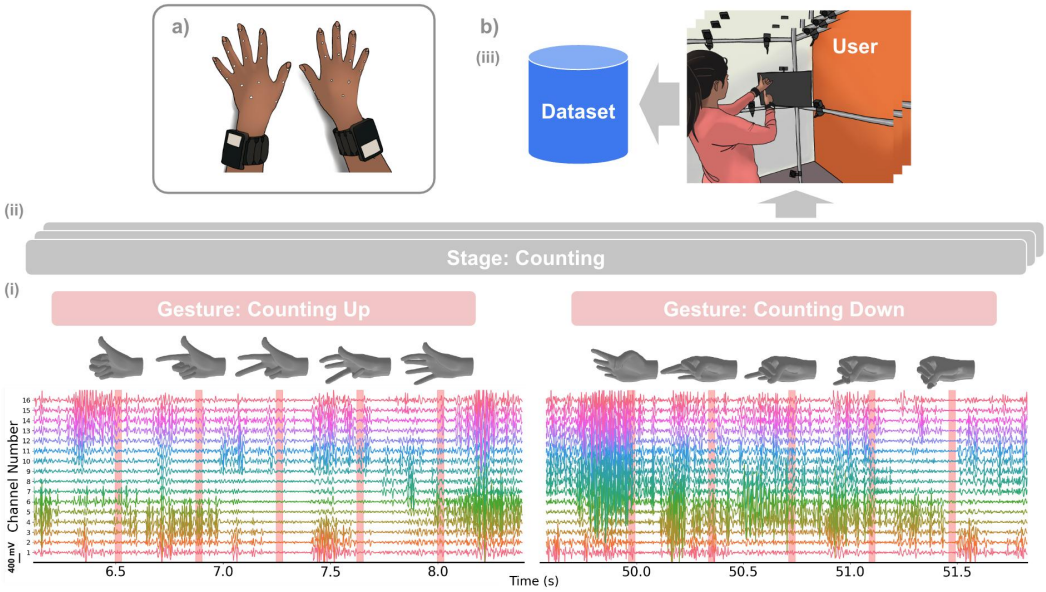

Figure 2: *Dataset composition*: a) sEMG-RD wrist-band and motion capture marker (white dots) setup. b) Dataset breakdown. i) Users are prompted to perform a sequence of movement types (*gestures*), such as counting up and down. sEMG and poses are recorded simultaneously. ii) Groups of specific gesture types comprise a *stage*, such as counting. Stages are partitioned into train/val/test splits (see Section 3.4). Our dataset consists of 29 diverse stages. iii) Each of the 193 users perform various stages, donning on-and-off the wrist band. In total we record 370 hours of data.

and B.2). The IK solver failed for 12.7% of frames, typically due to simultaneously occluded markers. Finally, joints angles were linearly interpolated to 2 kHz to match the sample rate of sEMG.

Participants followed a standardized data collection protocol across a diverse set of 45-120 s *stages* in which participants were prompted to perform either a mix of 3-5 similar *gestures* in random orderings (e.g. specific *finger counting* orderings such as *ascending* or *descending*) or unconstrained freeform movements (see Appendices A and B.3 for further details). *Stages* can be viewed as a categorization of *gestures*. For example, the *Counting* stage categorizes *Counting Up* and *Counting Down* gestures (see Fig. 2). During data collection, the majority of users donned on-and-off the device 4 times, with a small fraction only thrice. Each group of stages with a single band placement is referred to as a *session*. We report the prompted movements for each stage in detail in Appendix B.3. During each stage, we prompted participants using videos and verbal instructions by the research assistant. Participants were instructed to move their hands across their body and between their waist and shoulders to ensure a range of different postures were sampled. See Fig. 2 for a visualization of the data collection.

The full dataset is organized hierarchically by participant, session, and stage. In total, we collected data from 193 participants, spanning 370 hours, 751 sessions, 29 diverse stages (see Appendix B.3 for further details and Table 3 for statistics). Note that the number of hours counts the right-handed and left-handed data separately for each participant, although they were collected simultaneously. To our knowledge, this is the only open-sourced sEMG and motion capture dataset and is of similar scale to those in the CV literature [Yuan et al., 2017, Brahmbhatt et al., 2020, Moon et al., 2020, 2024]. The entire dataset consists of 25, 253 HDF5 files, each consisting of time-aligned sEMG and joint angles for a single hand in a single stage.

3.3 Tasks

The *emg2pose benchmark* includes two benchmark tasks: *pose regression* and *pose tracking*.

Regression: For this task, previously explored in Liu et al. [2021], Sîmpetru et al. [2022b], one must regress from sEMG to hand joint angle sequences. Without knowledge of the initial hand pose and velocity, this is a partially observable task [Spaan, 2012], and thus particularly challenging for the

reasons mentioned in Section 1. Pose regression is the most challenging task and is meant to promote continued research with applications including unimodal pose prediction in settings where computer vision is infeasible or unreliable.

Tracking: For this simpler task, one must regress from sEMG to hand joint angle sequences whilst being provided with the initial hand pose in the sequence. Providing the initial pose addresses the partial observability dilemma. Nevertheless, this task still poses the generalization challenges discussed in Section 1. The tracking task is meant to promote initial research and progress, and has several real-world applications. An effective tracker would provide great value in settings where: the user is prompted to match a given pose before tracking commences; visual pose prediction feedback is provided, allowing the user to adjust their pose to correct for erroneous initial predictions; and when ground truth pose estimates are intermittently available, such as from computer vision settings whenever partial or full occlusions occur.

Evaluation: We evaluate on 5 second trajectories and report test set mean absolute *joint angular error* ($^\circ$) and mean (Euclidean) *landmark distance* (mm). Landmarks correspond to joint and fingertip Cartesians. We do not regress to wrist angles, which were not recorded for this dataset. Landmarks corresponding to the most proximal joint for fingers other than the thumb always have zero error because the wrist does not move. These landmarks are therefore excluded from our metrics. We obtain landmark locations by passing joint angles through a default hand model. This introduces bias, as it will not perfectly align with each user's anatomy. We leave addressing this limitation for future work. In real world applications, it will be important to not only improve mean performance for these metrics, but also lower percentile scores across the population.

3.4 Held-Out Settings

Effective pose inference requires models that generalize across *device placements*, *users*, and *hand kinematics*. Prior works have only investigated generalization across a subset of these axes, such as user [Liu et al., 2021, CTRL-labs at Reality Labs et al., 2024] or device placement [Liu et al., 2021, Palermo et al., 2017], but generalization to new types of kinematics has not been explicitly explored. In contrast, we provide three separate test sets intended to measure these axes independently. The statistics of each held-out scenario are reported in Table 3. In short, *users* corresponds to unseen users, but in-distribution kinematics (*stages*). *Stages* represents unseen kinematic categories, but in-distribution users. Finally, *users, stages* constitute held-out users and stages, and is of greatest value as the most encompassing real-world deployment setting. Both held out user scenarios all constitute new device placements, which vary across all sessions. We break down train, validation and test splits roughly using $0.7 : 0.1 : 0.2$ ratio with exact splits shown in Table 3. Held-out users are randomly sampled and held-out stages are chosen to be visually out-of-distribution with respect to the training stages. See Fig. 3 for a breakdown of which stages are in the training and held-out sets, Table 6 for details regarding each stage, and Appendix B.2.1 for further dataset details.

3.5 Baselines

We provide three baselines: open-source re-implementations of the *NeuroPose* and *SensingDynamics* network architectures [Liu et al., 2021, Sîmpetru et al., 2022a], and a new *vemg2pose* model. Algorithm details can be found in Appendix C.

vemg2pose: sEMG meaures underlying muscle activity, and therefore relates more strongly to hand movements than the static pose of the hand. Therefore, *vemg2pose* ("Velocity-based emg2pose") predicts joint angular velocities, which are then integrated to produce joint angle predictions. sEMG is first embedded via a causal strided convolutional *featurizer*, which temporally down-samples sEMG from 2 kHz to 50 Hz. A *Time-Depth Separable Convolution* (TDS) network is used for the featurizer, as it has been shown to be effective and parameter-efficient in the automatic speech recognition literature [Hannun et al., 2019] (see Appendix C for implementation details). The features at each time-step are then concatenated to the joint angle predictions at the previous time step and fed to an LSTM *decoder*, which produces the next velocity prediction. Those velocities are added to the previous joint angles to produce the next prediction. vemg2pose is therefore auto-regressive with respect to its own predictions. Finally, predictions are linearly up-sampled to match the sample rate of the joint angles targets. For the tracking task, the initial joint angles are set to the ground truth,

according to the motion capture labels. For the regression task, the initial state is also predicted by the decoder (see Appendix C for further details).

NeuroPose: NeuroPose and vemg2pose differ in their prediction spaces and network architectures. Whereas vemg2pose predicts angular velocities, NeuroPose predicts joint angles directly. NeuroPose uses a U-Net architecture with residual bottleneck layers. Briefly, a convolutional encoder spatially and temporally down-samples sEMG while extracting features which are then refined via a stack of residual blocks. Finally, a decoder generates pose predictions at the original sample rate via convolutions and up-sampling layers. Because our sEMG device measures at 10x the temporal frequency and 2x the spatial frequency of the MyoBand used in Liu et al. [2021], we increase the temporal and spatial down and up-sampling of NeuroPose's featurizer and decoder (by 8x and 2x, respectively), such that the receptive field remains comparable to the original model. See Liu et al. [2021] for full model details and Appendix C for further details.

SensingDynamics: SensingDynamics and NeuroPose primarily differ in their architectures. Instead of a U-Net, SensingDynamics' featurizer comprises of 2d convolutions over sEMG channels and time, with learnable SMU activations [Biswas et al., 2021], batch normalisation, circular padding across channels, and dropout layers. The decoder comprises of a 3-layered MLP. Uniquely, SensingDynamics additionally passes 20Hz low-passed filtered sEMG as input to the featurizer. See Sîmpetru et al. [2022a] for full model details and Appendix C for further details.

Training Setup: All algorithms are trained to minimize the L1 error between predicted and ground truth joint angles as well as the Euclidean error between between predicted and ground truth fingertip locations. The joint angle loss term has a weight of 1 and the fingertip loss term has a weight of $.01$. We train on 1-6 seconds of non-overlapping trajectories. The training trajectory length - in addition to other hyperparameters - is optimized independently for each algorithm (see Table 7). We train for 500 epochs with a 50 epoch early stopping criterion. Time-points for which motion capture data are not available are skipped during training and evaluation. We use a batch size of 64 per GPU. We train on Amazon EC2 `g5.48xlarge` instances which have 8x NVIDIA T4 GPUs for less than a day.

4 Experiments

4.1 Benchmark Results

Table 4: *Regression* test set results. Mean and standard deviation are reported across users. Bold indicates the significance of a Wilcoxon signed-rank test comparing vemg2pose to NeuroPose and Sensing Dynamics for each metric and condition (.01 threshold adjusted to .0008 via Bonferroni correction, see Appendix C.7 for details).

Test Set	Baseline	Angular Error ($^\circ$)	Landmark Distance (mm)
User	SensingDynamics	15.5 $\pm$ 1.4	21.8 $\pm$ 2.1
	NeuroPose	13.2 $\pm$ 1.1	17.5 $\pm$ 1.3
	vemg2pose	**12.2** $\pm$ **1.3**	**15.8** $\pm$ **1.9**
Stage	SensingDynamics	18.8 $\pm$ 1.6	26.6 $\pm$ 2.0
	NeuroPose	17.2 $\pm$ 1.7	24.0 $\pm$ 2.1
	vemg2pose	**15.2** $\pm$ **1.6**	**20.4** $\pm$ **2.2**
User, Stage	SensingDynamics	18.7 $\pm$ 1.6	27.2 $\pm$ 2.0
	NeuroPose	17.5 $\pm$ 1.5	24.9 $\pm$ 1.7
	vemg2pose	**15.8** $\pm$ **1.4**	**21.6** $\pm$ **2.0**

We report *regression* results in Table 4 and *tracking* results in Table 5. We do not report standard deviation across model seeds, as we observed these to be negligible. Results are further broken down by stage, finger, and joint in Figs. 3, 10 and 11, respectively. For the regression task, vemg2pose outperforms both NeuroPose and SensingDynamics with respect to both angular errors and landmark distances. In general, accuracy degrades most for the held-out *user, stage* combination, which is the hardest of all transfer scenarios. For the tracking task - in which the initial ground truth pose is provided - errors are lower overall, as expected (see Section 3.3). For this task, we do not report

scores for NeuroPose and SensingDynamics, as these models were not originally designed to leverage knowledge of the initial ground truth pose during inference.

Table 5: *Tracking* test set results. Mean and standard deviation are reported across users.

Test Set	Baseline	Angular Error (°)	Landmark Distance (mm)
User	vemg2pose	7.7 ± 1.0	10.3 ± 1.5
Stage	vemg2pose	11.2 ± 1.4	15.2 ± 1.9
User, Stage	vemg2pose	11.0 ± 1.0	15.4 ± 1.4

Performance varies considerably across users for all models and tasks, potentially due to anatomical differences (Tables 4 and 5). Performance varies significantly across stages (Fig. 3), which is likely a result of the amount and type of movements in each stage. Stages with limited movement (StaticHands, WristFlex) may be easier for the model track because they involve very limited postural transitions. Stages with complex hand poses and dynamic articulation of individual fingers (Gesture2, Pointing) are more challenging and have higher errors. Moreover, Fig. 10 shows that performance varies significantly across fingers and finger joints, with the thumb the most reliably predicted, followed by the index, middle, ring, and pinky fingers. We also find that proximal joint angles of the fingers are easier to track than distal joint angles (Fig. 11). Together, this suggests that stages with high amounts of thumb movements (e.g. ThumbRotations) may be easier to track than those with more general finger movements (e.g. Freestyle1).

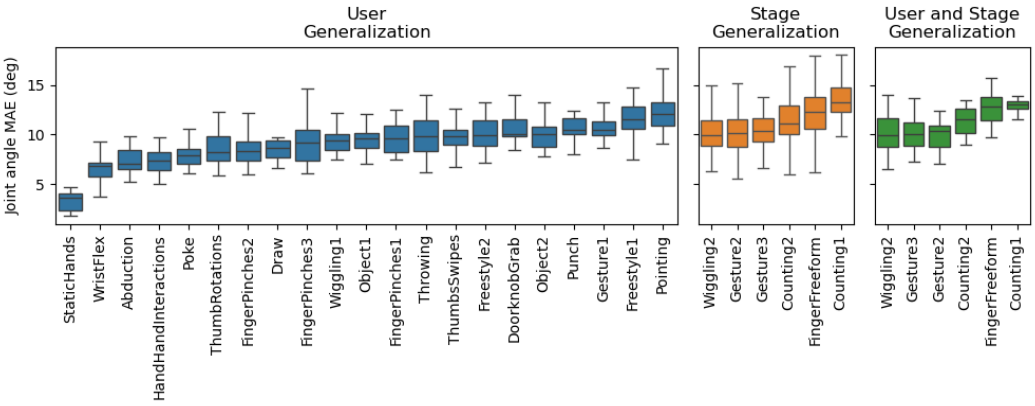

Figure 3: *vemg2pose tracking performance break down* by stage and generalization condition. Distributions are over users. Note the variability in performance across stages. Each box shows the median and interquartile range (IQR), and whiskers show the minimum and maximum values that are within 1.5 times the IQR of the lower and upper quartiles.

4.2 Analysis on Challenging Stages for Vision-Based Systems

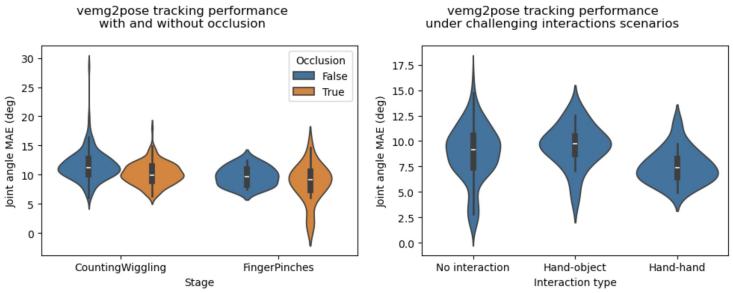

Figure 4: vemg2pose tracking results *with/without occlusion (left) and physical interactions (right)*. Distributions are over users. See Appendix D.1 for more details.

Some stages were specifically designed to test behaviors that are known to be challenging for vision-based hand pose estimation (see Appendix D.1 for details). We found that stages with hand-hand

interactions or hand-object interactions have similar model performance compared to stages without such interactions (Fig. 4, right), although differences in behavioral distribution across these stages makes direct comparison challenging. Furthermore, we find that visual occlusion does not impact sEMG based pose reconstruction, as expected. Stages in which the hand is occluded from a CV based headset tracking system have similar accuracy compared to stages without occlusion in which the same behaviors are performed (Fig. 4, left).

4.3 Qualitative Analysis

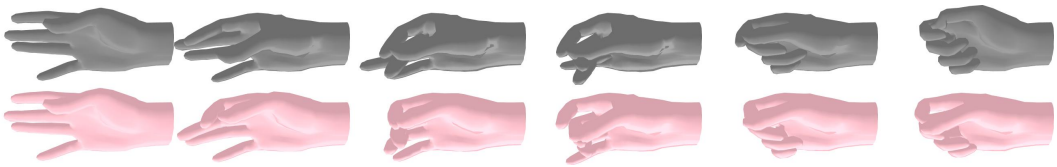

Figure 5: *Median percentile held-out user and stage* (Counting2). Top: motion capture; bottom: vemg2pose, tracking predictions. Clips unroll evenly left-to-right over a 2 second segment.

We plot *vemg2pose, tracking* real-time online and offline kinematic predictions for *held-out users and stages* in Figs. 1 and 5 (see Appendix C.5 for online setup details). This is the most challenging scenario, representing generalization to held-out kinematics, user anatomy, and device placement. For Fig. 5, we plot a median-performance representative held-out stage (Counting2, see Table 6) and user. As seen, individual finger movements are mostly tracked, but not always. We visualize top and bottom percentile (15% and 85%) offline kinematics for the held-out users and stages generalization setting in Figs. 12 to 15. In general, we observed three challenges specific to sEMG pose inference: angular drift due to sensing that strongly relates to pose derivatives (see the ring finger in Fig. 15); movements related to harder-to-sense intrinsic hand muscles, such as the finger adduction/abduction present in the "vulcan" gesture (Table 6); movements related to smaller and fewer muscles, such as pinky (see Fig. 10) and distal joint motion (see Fig. 1).

4.4 Dataset Scale Analysis

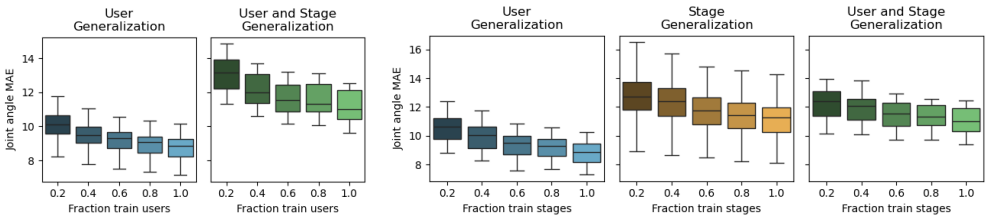

Figure 6: *Generalization vs. number of training users (left two) or stages (right three) for vemg2pose tracking*. We subsampled the training users/stages but evaluated on the same held-out users/stages. As seen, performance improves with the number of training users/stages, demonstrating the importance of our dataset scale for effective generalization. Box plots take the same format as Fig. 3.

We ran experiments to *demonstrate the importance of the scale* of our dataset for effective generalization. In Fig. 6, we show that increasing the number of training users considerably reduces the error for held-out users, perhaps because models are exposed to sEMG from users with a variety of wrist anatomies. We also show that increasing the number of stages per-user improves performance across all modes of generalization, demonstrating the importance of behavioural diversity.

4.5 Quantifying Generalization Difficulty across Users and Stages

To directly quantify the *difficulty of generalizing* across held-out *stages* and *users*, we performed experiments in which a subset of the data from the held-out users and stages were either folded into the training set or excluded entirely. Fig. 7 shows that excluding specific users and stages from the training set markedly degrades performance, demonstrating the difficulty of generalizing across these dimensions. Refer to Fig. 7 for a detailed description of the experimental setup.

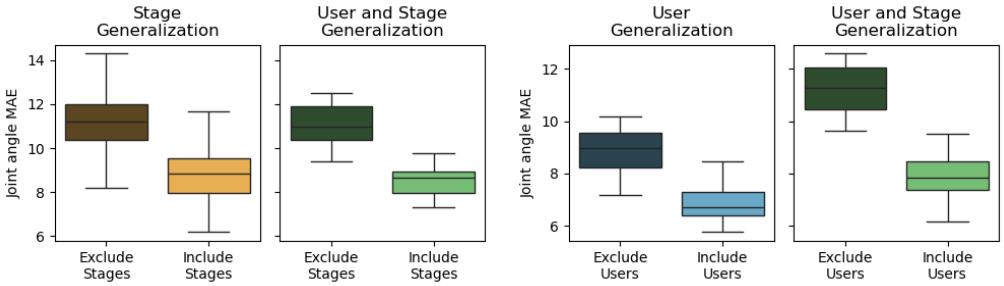

Figure 7: Excluding stages (left) or users (right) from the training set *markedly decreases performance for these stages/users*. For the *include stages/users* condition, we include 70% of the data from the held-out stages/users in the training set. For the *exclude stages/users* condition we exclude that 70% entirely. Both test sets are identical allowing us to isolate the influence of holding out stages/users. Data are from a tracking task with a vemg2pose model. Distributions are over users.

5 Limitations and Future Work

Modelling: We provide an initial investigation into generalized sEMG-to-pose modelling and open-source our baselines to the community. Nevertheless, there remains a plethora of unexplored, potentially fruitful sequence modelling directions, such as state space and diffusion-based methods. Pose estimation in the presence of uncertainty introduced by sensor noise and anatomical variability could also be addressed with probabilistic methods [Danelljan et al., 2020]. Model personalization has also been shown to be beneficial [CTRL-labs at Reality Labs et al., 2024, Liu et al., 2021], yet we do not explore this avenue here. In addition, our models obtain mean landmark distance errors that are higher than reported in the CV literature [Boukhayma et al., 2019, Mueller et al., 2017], despite having the advantage of not having to infer the wrist position or user's anatomy. Addressing this performance gap will be of great importance. Finally, the lack of broader access to the sEMG-RD wrist-band [CTRL-labs at Reality Labs et al., 2024] might be limiting, as this precludes human-in-the-loop testing of models.

Metrics: Our landmark distance metrics use a default hand model to convert joint angles to joint positions. The mismatch between the hand model and user anatomy will bias this metric. In general, our metrics do not capture the physical plausibility of model predictions. For example, we have observed that vemg2pose sometimes predicts unfeasible kinematics, such as intra-finger penetration. Providing metrics that capture these failure modes will be of value, especially for embodied applications [Yuan et al., 2023]. Simulators of the hand [Caggiano et al., 2022] could be leveraged in a manner similar to Yuan et al. [2023] to ensure physical constraints are adhered to. Finally, our held-out *user, stage* test scenario is meant to best represent real-world in the wild performance. Nevertheless, it does not cover a potential range of signal aggressors such as: electrode-skin contact artifacts; impedance changes from sweat; electrical interference from external devices; and non-stationarity due to muscle fatigue. While these aggressors likely play a minor role in sEMG variability, they may be important to include in future datasets and test sets.

Dataset: We discuss dataset limitations in Appendix B.4.

Ethical and Societal Implications: We discuss ethical and societal implications in Appendix B.5.

6 Conclusion

We introduce the *emg2pose benchmark*, the largest, diverse, and open-source dataset of high-fidelity sEMG recordings and hand pose labels. We introduce competitive benchmark models that can track or regress to hand pose for held-out users, stages and sessions, although there remains significant room to improve these models in future research. Due to the myriad sources of variability in sEMG signals, deciphering the relationship between sEMG and movement in a manner that generalizes across people and kinematics will likely require new algorithmic advances, taking inspiration from related machine learning fields. Large datasets like *emg2pose* should thus facilitate progress in both sEMG decoding and machine learning applied to biosignals more broadly. Progress will enable intuitive, high-dimensional human-computer interfaces that we perceive as extensions of ourselves.

7 Acknowledgements

We thank Patrick Kaifosh and TR Reardon for their sponsorship and vision and the entire CTRL-labs team for their collaboration and support. We thank Carl Hewitt and Migmar Tsering for help with data collection, Steve Olsen and Mark Hogan for assistance setting up motion capture recordings, John Choi and Diogo Peixoto for technical assistance and advice, and Dano Morrison and Sunaina Rajani for assistance with visualizations.

References

P. Achenbach, S. Laux, D. Purdack, P. N. Müller, and S. Göbel. Give me a sign: Using data gloves for static hand-shape recognition. Sensors, 23(24):9847, 2023.

J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023.

C. Amma, T. Krings, J. Böer, and T. Schultz. Advancing muscle-computer interfaces with high-density electromyography. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, pages 929–938, 2015.

M. Atzori, A. Gijsberts, C. Castellini, B. Caputo, A.-G. M. Hager, S. Elsig, G. Giatsidis, F. Bassetto, and H. Müller. Electromyography data for non-invasive naturally-controlled robotic hand prostheses. Scientific data, 1(1):1–13, 2014.

K. Biswas, S. Kumar, S. Banerjee, and A. K. Pandey. Smu: smooth activation function for deep networks using smoothing maximum technique. arXiv preprint arXiv:2111.04682, 2021.

A. Boukhayma, R. d. Bem, and P. H. Torr. 3d hand shape and pose from images in the wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10843–10852, 2019.

S. Brahmbhatt, C. Tang, C. D. Twigg, C. C. Kemp, and J. Hays. Contactpose: A dataset of grasps with object contact and hand pose. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16, pages 361–378. Springer, 2020.

A. Brunetti, D. Buongiorno, G. F. Trotta, and V. Bevilacqua. Computer vision and deep learning techniques for pedestrian detection and tracking: A survey. Neurocomputing, 300:17–33, 2018.

V. Caggiano, H. Wang, G. Durandau, M. Sartori, and V. Kumar. Myosuite–a contact-rich simulation suite for musculoskeletal motor control. arXiv preprint arXiv:2205.13600, 2022.

Y. Cai, L. Ge, J. Cai, and J. Yuan. Weakly-supervised 3d hand pose estimation from monocular rgb images. In Proceedings of the European conference on computer vision (ECCV), pages 666–682, 2018.

CTRL-labs at Reality Labs, D. Sussillo, P. Kaifosh, and T. Reardon. A generic noninvasive neuromotor interface for human-computer interaction. bioRxiv, 2024. doi: 10.1101/2024.02.23.581779. URL `https://www.biorxiv.org/content/early/2024/02/28/2024.02.23.581779`.

M. Danelljan, L. V. Gool, and R. Timofte. Probabilistic regression for visual tracking. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 7183–7192, 2020.

K. Darvish, L. Penco, J. Ramos, R. Cisneros, J. Pratt, E. Yoshida, S. Ivaldi, and D. Pucci. Teleoperation of humanoid robots: A survey. IEEE Transactions on Robotics, 39(3):1706–1727, 2023.

Y. Du, W. Jin, W. Wei, Y. Hu, and W. Geng. Surface emg-based inter-session gesture recognition enhanced by deep domain adaptation. Sensors, 17(3):458, 2017.

M. Dunion, T. McInroe, K. S. Luck, J. Hanna, and S. Albrecht. Conditional mutual information for disentangled representations in reinforcement learning. Advances in Neural Information Processing Systems, 36, 2024.

Z. Fan, O. Taheri, D. Tzionas, M. Kocabas, M. Kaufmann, M. J. Black, and O. Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12943–12954, 2023.

I. T. Gatt, T. Allen, and J. Wheat. Accuracy and repeatability of wrist joint angles in boxing using an electromagnetic tracking system. Sports Engineering, 23:1–10, 2020.

L. Ge, H. Liang, J. Yuan, and D. Thalmann. Robust 3d hand pose estimation in single depth images: from single-view cnn to multi-view cnns. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3593–3601, 2016.

T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, and K. Crawford. Datasheets for datasets. Communications of the ACM, 64(12):86–92, 2021.

A. Gu, K. Goel, and C. Ré. Efficiently modeling long sequences with structured state spaces. arXiv preprint arXiv:2111.00396, 2021.

S. Hampali, M. Rad, M. Oberweger, and V. Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 3196–3206, 2020.

S. Han, B. Liu, R. Wang, Y. Ye, C. D. Twigg, and K. Kin. Online optical marker-based hand tracking with deep labels. Acm transactions on graphics (tog), 37(4):1–10, 2018.

S. Han, B. Liu, R. Cabezas, C. D. Twigg, P. Zhang, J. Petkau, T.-H. Yu, C.-J. Tai, M. Akbay, Z. Wang, et al. Megatrack: monochrome egocentric articulated hand-tracking for virtual reality. ACM Transactions on Graphics (ToG), 39(4):87–1, 2020.

S. Han, P. Wu, Y. Zhang, B. Liu, L. Zhang, Z. Wang, W. Si, P. Zhang, Y. Cai, T. Hodan, R. Cabezas, L. Tran, M. Akbay, T. Yu, C. Keskin, and R. Wang. Umetrack: Unified multi-view end-to-end hand tracking for VR. In SIGGRAPH Asia 2022 Conference Papers, SA 2022, Daegu, Republic of Korea, December 6-9, 2022, 2022.

A. Hannun, A. Lee, Q. Xu, and R. Collobert. Sequence-to-sequence speech recognition with time-depth separable convolutions. arXiv preprint arXiv:1904.02619, 2019.

J. N. Ingram, K. P. Körding, I. S. Howard, and D. M. Wolpert. The statistics of natural hand movements. Experimental brain research, 188:223–236, 2008.

E. Jang, A. Irpan, M. Khansari, D. Kappler, F. Ebert, C. Lynch, S. Levine, and C. Finn. Bc-z: Zero-shot task generalization with robotic imitation learning. In Conference on Robot Learning, pages 991–1002. PMLR, 2022.

X. Jiang, X. Liu, J. Fan, X. Ye, C. Dai, E. A. Clancy, M. Akay, and W. Chen. Open access dataset, toolbox and benchmark processing results of high-density surface electromyogram recordings. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 29:1035–1046, 2021.

J. D. M.-W. C. Kenton and L. K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of naacL-HLT, volume 1, page 2. Minneapolis, Minnesota, 2019.

A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 4015–4026, 2023.

T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. Large language models are zero-shot reasoners. Advances in neural information processing systems, 35:22199–22213, 2022.

A. Krasoulis, I. Kyranou, M. S. Erden, K. Nazarpour, and S. Vijayakumar. Improved prosthetic hand control with concurrent use of myoelectric and inertial measurements. Journal of neuroengineering and rehabilitation, 14:1–14, 2017.

G. Laput and C. Harrison. Sensing fine-grained hand activity with smartwatches. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pages 1–13, 2019.

G. Laput, R. Xiao, and C. Harrison. Viband: High-fidelity bio-acoustic sensing using commodity smartwatch accelerometers. In Proceedings of the 29th Annual Symposium on User Interface Software and Technology, UIST '16, page 321–333, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450341899. doi: 10.1145/2984511.2984582. URL https://doi.org/10.1145/2984511.2984582.

M. Lauri, D. Hsu, and J. Pajarinen. Partially observable markov decision processes in robotics: A survey. IEEE Transactions on Robotics, 39(1):21–40, 2022.

Y. Liu, S. Zhang, and M. Gowda. Neuropose: 3d hand pose tracking using emg wearables. In Proceedings of the Web Conference 2021, pages 1471–1482, 2021.

Y. Liu, Y. Liu, C. Jiang, K. Lyu, W. Wan, H. Shen, B. Liang, Z. Fu, H. Wang, and L. Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 21013–21022, 2022.

Y. Luo, Y. Li, P. Sharma, W. Shou, K. Wu, M. Foshey, B. Li, T. Palacios, A. Torralba, and W. Matusik. Learning human–environment interactions using conformal tactile textiles. Nature Electronics, 4 (3):193–201, 2021.

J. McIntosh, A. Marzo, M. Fraser, and C. Phillips. Echoflex: Hand gesture recognition using ultrasound imaging. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, pages 1923–1934, 2017.

R. Merletti and D. Farina. Surface electromyography: physiology, engineering, and applications. John Wiley & Sons, 2016.

G. Moon, S.-I. Yu, H. Wen, T. Shiratori, and K. M. Lee. Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16, pages 548–564. Springer, 2020.

G. Moon, S. Saito, W. Xu, R. Joshi, J. Buffalini, H. Bellan, N. Rosen, J. Richardson, M. Mize, P. De Bree, et al. A dataset of relighted 3d interacting hands. Advances in Neural Information Processing Systems, 36, 2024.

F. Mueller, D. Mehta, O. Sotnychenko, S. Sridhar, D. Casas, and C. Theobalt. Real-time hand tracking under occlusion from an egocentric rgb-d sensor. In Proceedings of the IEEE International Conference on Computer Vision, pages 1154–1163, 2017.

F. Mueller, F. Bernard, O. Sotnychenko, D. Mehta, S. Sridhar, D. Casas, and C. Theobalt. Ganerated hands for real-time 3d hand tracking from monocular rgb. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 49–59, 2018.

F. Palermo, M. Cognolato, A. Gijsberts, H. Müller, B. Caputo, and M. Atzori. Repeatability of grasp recognition for robotic hand prosthesis control based on semg data. In 2017 International Conference on Rehabilitation Robotics (ICORR), pages 1154–1159. IEEE, 2017.

X. Pan, P. Luo, J. Shi, and X. Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In Proceedings of the european conference on computer vision (ECCV), pages 464–479, 2018.

F. S. Parizi, E. Whitmire, and S. Patel. Auraring: Precise electromagnetic finger tracking. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 3(4):1–28, 2019.

G. Park, T.-K. Kim, and W. Woo. 3d hand pose estimation with a single infrared camera via domain transfer learning. In 2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), pages 588–599. IEEE, 2020.

Plotly Technologies Inc. Collaborative data science, 2015. URL https://plot.ly.

F. Quivira, T. Koike-Akino, Y. Wang, and D. Erdogmus. Translating semg signals to continuous hand poses using recurrent neural networks. In 2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), pages 166–169. IEEE, 2018.

S. Rawat, S. Vats, and P. Kumar. Evaluating and exploring the myo armband. In 2016 International Conference System Modeling & Advancement in Research Trends (SMART), pages 115–120. IEEE, 2016.

A. Roda-Sales, J. L. Sancho-Bru, M. Vergara, V. Gracia-Ibáñez, and N. J. Jarque-Bou. Effect on manual skills of wearing instrumented gloves during manipulation. Journal of biomechanics, 98: 109512, 2020.

B. Samarth, T. Chengcheng, D. T. Christopher, C. K. Charles, and H. James. Contactpose: A dataset of grasps with object contact and hand pose. In European Conference on Computer Vision (ECCV), 2020.

L. Santos Carreras. Increasing haptic fidelity and ergonomics in teleoperated surgery. Technical report, EPFL, 2012.

S. Scheggi, L. Meli, C. Pacchierotti, and D. Prattichizzo. Touch the virtual reality: using the leap motion controller for hand tracking and wearable tactile devices for immersive haptic rendering. In ACM SIGGRAPH 2015 Posters, pages 1–1. 2015.

F. Sener, D. Chatterjee, D. Shelepov, K. He, D. Singhania, R. Wang, and A. Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 21096–21106, 2022.

Z. Shen, J. Yi, X. Li, M. H. P. Lo, M. Z. Chen, Y. Hu, and Z. Wang. A soft stretchable bending sensor and data glove applications. Robotics and biomimetics, 3(1):22, 2016.

S. Shrestha, C. Fermüller, T. Huang, P. T. Win, A. Zukerman, C. M. Parameshwara, and Y. Aloimonos. Aimusicguru: Music assisted human pose correction. arXiv preprint arXiv:2203.12829, 2022.

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. bioRxiv, pages 2022–07, 2022a.

R. C. Sîmpetru, M. Osswald, D. I. Braun, D. S. Oliveira, A. L. Cakici, and A. Del Vecchio. Accurate continuous prediction of 14 degrees of freedom of the hand from myoelectrical signals through convolutive deep learning. In 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pages 702–706. IEEE, 2022b.

I. Sosin, D. Kudenko, and A. Shpilman. Continuous gesture recognition from semg sensor data with recurrent neural networks and adversarial domain adaptation. In 2018 15Th international conference on control, automation, robotics and vision (ICARCV), pages 1436–1441. IEEE, 2018.

M. T. Spaan. Partially observable markov decision processes. In Reinforcement learning: State-of-the-art, pages 387–414. Springer, 2012.

A. Spurr, J. Song, S. Park, and O. Hilliges. Cross-modal deep variational hand pose estimation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 89–98, 2018.

A. Spurr, U. Iqbal, P. Molchanov, O. Hilliges, and J. Kautz. Weakly supervised 3d hand pose estimation via biomechanical constraints. In European conference on computer vision, pages 211–228. Springer, 2020.

A. Spurr, A. Dahiya, X. Wang, X. Zhang, and O. Hilliges. Self-supervised 3d hand pose estimation from monocular rgb via contrastive learning. In Proceedings of the IEEE/CVF international conference on computer vision, pages 11230–11239, 2021.

D. Stashuk. Emg signal decomposition: how can it be accomplished and used? Journal of Electromyography and Kinesiology, 11(3):151–173, 2001.

J. S. Supančič, G. Rogez, Y. Yang, J. Shotton, and D. Ramanan. Depth-based hand pose estimation: methods, data, and challenges. International Journal of Computer Vision, 126:1180–1198, 2018.

A. Tashakori, Z. Jiang, A. Servati, S. Soltanian, H. Narayana, K. Le, C. Nakayama, C.-l. Yang, Z. J. Wang, J. J. Eng, et al. Capturing complex hand movements and object interactions using machine learning-powered stretchable smart textile gloves. Nature Machine Intelligence, 6(1):106–118, 2024.

H. Truong, S. Zhang, U. Muncuk, P. Nguyen, N. Bui, A. Nguyen, Q. Lv, K. Chowdhury, T. Dinh, and T. Vu. Capband: Battery-free successive capacitance sensing wristband for hand gesture recognition. In Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems, pages 54–67, 2018.

C. Wan, T. Probst, L. V. Gool, and A. Yao. Self-supervised 3d hand pose estimation through training by fitting. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 10853–10862, 2019.

J. Wang, F. Mueller, F. Bernard, S. Sorli, O. Sotnychenko, N. Qian, M. A. Otaduy, D. Casas, and C. Theobalt. Rgb2hands: real-time tracking of 3d hand interactions from monocular rgb video. ACM Transactions on Graphics (ToG), 39(6):1–16, 2020.

Z. Yang, S. Yan, B.-J. F. van Beijnum, B. Li, and P. H. Veltink. Hand-finger pose estimation using inertial sensors, magnetic sensors and a magnet. IEEE sensors journal, 21(16):18115–18122, 2021.

Z. Yu, J. S. Yoon, I. K. Lee, P. Venkatesh, J. Park, J. Yu, and H. S. Park. Humbi: A large multiview dataset of human body expressions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2990–3000, 2020.

S. Yuan, Q. Ye, B. Stenger, S. Jain, and T.-K. Kim. Bighand2. 2m benchmark: Hand pose dataset and state of the art analysis. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4866–4874, 2017.

Y. Yuan, J. Song, U. Iqbal, A. Vahdat, and J. Kautz. Physdiff: Physics-guided human motion diffusion model. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 16010–16021, 2023.

Y. Zhang and C. Harrison. Tomo: Wearable, low-cost electrical impedance tomography for hand gesture recognition. In Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology, pages 167–173, 2015.

C. Zimmermann and T. Brox. Learning to estimate 3d hand pose from single rgb images. In Proceedings of the IEEE international conference on computer vision, pages 4903–4911, 2017.

C. Zimmermann, D. Ceylan, J. Yang, B. Russell, M. Argus, and T. Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 813–822, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 5.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Appendix B.5.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] In Section 1 we link to `https://github.com/facebookresearch/emg2pose` that provides all the relevant information.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 3, Appendices A to C, and Tables 3, 6 and 7.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] For example, see Tables 4 and 5 and Fig. 3. We include standard deviation and percentile statistics calculated across users, stages. We do not report errors bars due to multiple model random seeds, as we saw very minimal differences.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 3.5.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Section 3.5 for details on existing baselines that we provide and architectures we use. See Section 3.2 and Appendix B.2 for details on the Optitrack system used to collect data. See Appendix B.1 for wrist-band details. See Appendix A for details including python packages we build on.
   (b) Did you mention the license of the assets? [Yes] Dataset and code are licensed under a CC-BY-NC-SA 4.0 license, which is listed in Appendix A and repository.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We provide videos in zip format to reviewers and further instruction regarding how to access the dataset in Appendix B.6.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] See Appendix A.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] In Section 3.2 and Appendix A we discuss how we remove personally identifiable information.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [No] We provide a detailed description of the experimental instructions and a list of the movements participants were asked to perform (see Appendices A and B).
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [Yes] We discuss the consenting process in Appendix A and the approval of all research under an external IRB
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] We recruited all participants through a third-party vendor that determined their compensation via market rates. We give details in Appendix A.

A Datasheet

We provide a datasheet in accordance with Gebru et al. [2021].

Motivation: The motivation for *emg2pose* is to address the lack of wide-spread, sufficiently large, non-invasive surface electromyographic (sEMG) datasets with high-quality ground-truth annotations for a concrete task. sEMG as a technology has the potential to revolutionize how humans interact with computers, and this public dataset is motivated to facilitate progress in this domain without needing specialized hardware. The task we consider is hand pose inference, as a potentially holistic and encompassing modality, with many biomimetic applications. This dataset was created by the CTRL-Labs research group within Reality Labs, Meta.

Composition: The entire dataset consists of $25,253$ HDF5 files, each consisting of time-aligned sEMG and joint angles for a single hand in a single stage. In total, we collected data from 193 participants, spanning 370 hours and 29 diverse stages. The number of hours includes both the right-handed and left-handed data for each participant, which were collected simultaneously. Each HDF5 file includes sEMG data from one hand, the stage label, and the joint angles. sEMG is recorded at 2kHz, high pass filtered at 40 Hz, and rescaled such that the noise floor has a standard deviation of 1. We also flip the sign of the left-handed EMG data to account for the reversal of polarity caused by wearing the band on the left vs. right hand. Additionally, the dataset includes a metadata file in CSV format containing dataset split information (train, val, and test). All metadata have been de-identified to remove any personally identifiable information and does not identify any sub-population. See Section 3 for additional details on the dataset and Table 3 for statistics about the dataset such as the number of participants, total duration, number of sessions and stages. See Table 6 for further details with regards to the stage composition. The configuration for the precise data splits used in our experiments can be found in the following link: `https://github.com/facebookresearch/emg2pose`.

Collection Process: We recruited participants through a third-party vendor, who compensated participants at market rates. All recruitment and on-boarding followed an external IRB-approved protocol. We provided participants with information about the study, and before study initiation asked them to review and sign an IRB-reviewed consent form. We gave all participants the opportunity to ask questions before the study and were able to discontinue participation at any point. To ensure participant well-being, on-site research administrators monitored participants during the study protocol. All data have been de-identified to remove any personally identifiable metadata. Participants stood in a 26 camera motion capture array (Appendix B.2). A research assistant placed 19 motion capture markers on each of the participant's hands (Han et al. [2018]) and an sEMG-RD band on each wrist [CTRL-labs at Reality Labs et al., 2024](Appendix B.2). All sEMG and motion capture data were streamed to a real-time data acquisition system at 2kHz and 60 Hz, respectively (Appendices B.1 and B.2). Participants followed a standardized data collection protocol across a diverse set of 30-120 s *stages* in which participants were prompted to perform a mix of 3-5 gestures. We organized the data collection into two repetitions of two different groups of 15 and 26 stages with a different band placement for each. Each group of stages with a single band placement is referred as a *session*. For further stage and data collection details see Appendices B.2.1 and B.3.

Preprocessing/Cleaning/Labeling: sEMG recordings in the dataset are sampled at 2 kHz with a bit depth of 12 bits, with a maximum signal amplitude of 6.6 mV, and are bandpass filtered with -3 dB cutoffs at 20 Hz and 850 Hz before digitization (see Appendix B.1).

Joint angles were estimated from motion capture recordings using a custom inverse kinematics pipeline using a personalized hand model according to Han et al. [2018]. Briefly, 19 reflective markers were attached to each hand, and their 3D coordinates were tracked via a commercial Optitrack system with 26 cameras around the participant. A ConvNet then assigned labels to each marker. The labeled markers were registered to positions on a calibrated hand mesh to determine landmark positions. An inverse kinematics solver produced the final joint angles. We applied a conservative 15 Hz low pass filter (Ingram et al. [2008]) to the final joint angles to ensure there is no residual jitter. The mean absolute difference between the filtered and unfiltered signal was only 0.32 degrees across 500 recordings.

This model produced an estimate of joint angles for the MCP, PIP and DIP joints for each finger as well as the IP, MCP and CMC joints of the thumb. Each joint had a degree of freedom for flexion and extension, while each MCP joint had an additional degree of freedom for abduction and adduction.

Following joint angle estimation, we used a forward kinematic algorithm using a generic hand model to produce estimates of landmark positions [Han et al., 2022]. We used the center of each joint, as well as the fingertips, as landmark positions for evaluation. Finally, joint angles were low-pass filtered at 15 Hz to remove tracking noise, and temporally upsampled to match to 2 kHz sample rate of sEMG.

Uses: The dataset and the associated tooling are meant to be used only to advance sEMG-based research topics of interest within the academic community for purely non-commercial purposes and applications. Our code for baseline models, built on top of frameworks such as PyTorch, PyTorch Lightning and Hydra, is designed such that it can be easily extended to the exploration of different models and novel techniques for this task. The dataset and the associated code are not intended to be used in conjunction with any other data types.

Distribution and Maintenance: The dataset and the code to reproduce the baselines are accessible via `https://github.com/facebookresearch/emg2pose`. The dataset is hosted on Amazon S3 and the code to reproduce the baseline experiments on GitHub under the CC-BY-NC-SA 4.0 license. We welcome contributions from the research community. Any future update, as well as ongoing maintenance such as tracking and resolving issues identified by the broader community, will be performed and distributed through the GitHub repository.

B Dataset Details

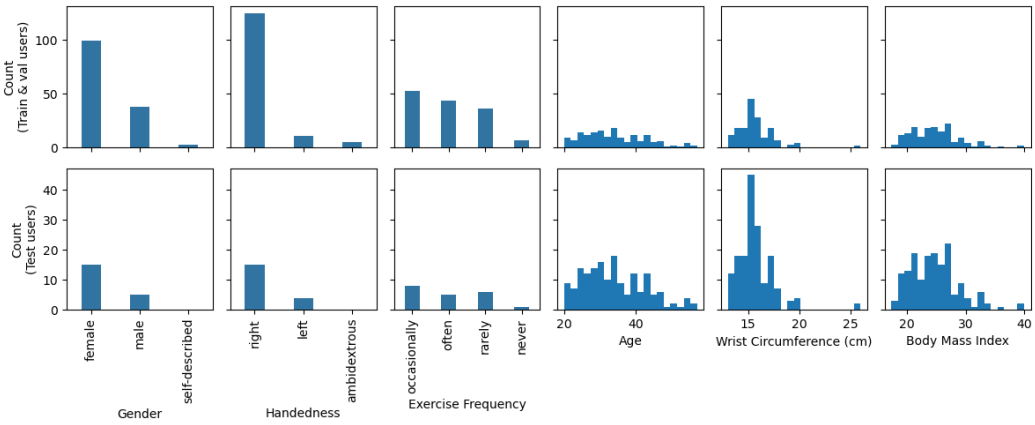

Figure 8: *Participant demographic information.* Train and val users are shown in the top row, and test users on the bottom. Notice that test users are representative of the population of train/val users.

B.1 sEMG Sensing

sEMG data were collected using the sEMG-RD [CTRL-labs at Reality Labs et al., 2024] consisting of 16 differential electrode pairs utilizing dry gold-plated electrodes. The 16 electrodes are arranged on a rigid ribbon, leaving a gap between electrodes 0 and 15 on the ulnar side of the wrist close to the ulnar styloid. Identical bands are worn on the left and right hands, with the same electrode indices aligning with the same anatomical features, but the polarity of the differential sensing being reversed. The band is tightened with an elastic strap and the size of the gap depends on the subject's wrist size and tightness. The band is manufactured in three different sizes to account for large changes in wrist size and the electrodes themselves are spring-loaded to further adjust across small variations in wrist sizes. In contrast, the previously used low-density Thalmic Labs Myo band [Rawat et al., 2016] only streams data at 200Hz, across 8 channels and at 8-bits.

B.2 Motion Capture

All motion capture data were collected using a 26 camera motion camera array at 60 Hz (Prime13W Optitrack) in an external data collection facility. Before data collection participants donned an sEMG

band on each wrist and 19 3mm facial motion capture markers in order. We placed markers at the base of each fingernail and between the DIP and PIP and PIP and MCP joints of each finger. For the thumb we placed markers between the IP and MCP joint, on the MCP joint, and between the MCP and CMC joints. We additionally placed markers in a triangular pattern on the dorsal side of the hand [Han et al., 2018]. Participants additionally wore a 3D printed frame of an XR headset, tethered to a PC, that was not relevant to the present data collection. Before collection, participants were asked to perform a series of 17 calibration gestures with each hand. These gestures were used as input to a custom optimization software that estimated the size of the hand and the position of motion capture markers relative to joints [Han et al., 2018]. We saved personalized hand model information to a separate file to be used offline to estimate joint angles from the collected motion capture data.

During data collection, sEMG data were streamed over Bluetooth to a real-time data collection application. Motion capture data were recorded over ethernet using Motive2 (Naturalpoint) and then streamed to the same data collection pipeline. sEMG and motion capture datastreams were assigned software timestamps based on their arrival at the data pipeline. Internal testing bounded the relative latency between the two recording pathways to below 10 ms, approximately the Nyquist limit of the 60 Hz Optitrack recording.

B.2.1 Data Collection Protocol

Data collection was divided into 4 different sessions (band placements). Participants performed two repetitions of two different groups of prompted stages. In each stage participants were asked to follow along a video of a set of example movements, either a mix of discrete gestures or freeform unprompted movements. Stages lasted 45 to 60s, while freeform stages lasted 60 to 120s. During data collection, users donned on-and-off the device on average 3.9 times in total, see Table 3. We call these *sessions* and are clearly annotated in our dataset. Participants performed all movements while standing or sitting on a tall stool. During each stage participants were asked to move their hand from right to left and up and down to ensure a broad range of postures.

B.3 Stage Descriptions

The data collection protocol was designed to capture a wide range of kinematics. Each stage consistent of a particular set of instructed kinematics. Descriptions of the kinematics performed in each stage can be found in Table 6.

B.4 Dataset Limitations

While our dataset is the largest and highest fidelity open-sourced to date, it is smaller than those used in CTRL-labs at Reality Labs et al. [2024], which may hinder generalization. While we provide high quality pose labels from motion capture using the inverse kinematics approach from Han et al. [2018], as a camera-based method it still suffers from occlusion, hindering label quality for gestures such as fist clenching. We additionally do not track wrist movements, which are important for how we interact with the world. Alternate labelling methods, such as stretch-sensing gloves, could address these limitations, at the potential expense of lower quality labels and impaired dexterity. Finally, future datasets could include both camera and sEMG sensors, which could be combined to improve pose inference in contexts where camera-based tracking fails such as occlusion.

B.5 Ethical and Societal Implications

The broader usage of sEMG and the specific development of sEMG pose estimation models may pose novel ethical and societal considerations. A highly performant emg-to-pose model running on a device placed on the wrist or forearm could store or transmit information about a person's actions, and appropriate safeguards to encrypt and limit access to this information may be warranted. There are numerous societal benefits for the development of sEMG models for pose estimation. sEMG allows one to directly interface a person's neuromotor intent with a computing device. This can be used to create novel device controls for the general population and can also be used to develop adaptive controllers for those who struggle to use existing computer interfaces.

Table 6: *Stage descriptions.* For video examples, visit `https://github.com/facebookresearch/emg2pose`.

Stage	Movements
FingerPinches1	Finger pinches (4) D-pad style thumb swipes (4) Thumb rotations
Object1	Drink from and rotate a cup Squeeze a soft toy
Counting1	Fingers counting up and down (5)
Counting2	Fingers counting up and down (5) Wiggling the fingers Abducting and adducting the fingers
DoorknobGrab	Mimic opening a doorknob pull fingers into a loose fist use index fingers as a trigger
Throwing	Mimic swinging ping pong paddle Mimic throwing a ball
Abduction	Series of finger abduction movements
FingerFreeform	Freeform movements of fingers
FingerPinches2	Single and multiple finger pinches (multiple fingers touching the thumb simultaneously) (7)
HandHandInteractions	Slide fingers along the palm Clap hands Co-wiggle fingers
Wiggling1	Wiggling and spreading fingers
Punch	Pull hand torwards body while grasping fingers Punching motion
Gesture1	Form and unform a claw Bring fingers together in loose fist Flick fingers individually (4)
StaticHands	Move hand from waist level to chest Press hands together and move slowly
FingerPinches3	Like FingerPinches1, but with the hands occluding eachother
Wiggling2	Like Counting2, but with the hands occluding eachother
Unconstrained	The hand that is not prompted to move during a particular stage
Gesture2	Extend index and pinky while curling middle fingers Make scissor cutting motion
FingerPinches2	Index finger pinches Middle finger pinches D-pad style thumb swipes (4)
Pointing	Point individual fingers (5) Snap middle finger and thumb
Freestyle1	Free style movements with one hand
Object2	Play with blocks Move chess pieces
Draw	Poking Mimic drawing Pinching, rotating with hands both close and far
Poke	Poking Pinch and rotate wrist
Gesture3	Extend thumb and pinky, curl middle fingers Vulcan salute Peace sign
ThumbsSwipes	D-pad style thumb swipes (4) Slowly fold and unfold all fingers
ThumbRotations	Thumb rotations
Freestyle2	Free style movement with both hands
WristFlex	Wrist flexion Wrist abduction

B.6 Dataset Instructions

Data is hosted on Amazon S3, and code with readme instructions as to how to reproduce experimental results are found on: `https://github.com/facebookresearch/emg2pose`.

C Algorithm details

C.1 vemg2pose

vemg2pose consists of a convolutional *featurizer* and an LSTM *decoder*. First, the featurizer converts sEMG to features:

$$\mathbf{z}_t = f(\mathbf{e}_{t-R:t}) \tag{1}$$

where f is the featurizer, R is the featurizer receptive field, and $\mathbf{e}_t$ and $\mathbf{z}_t$ are vectors of sEMG and features at time t, respectively. Next, the decoder produces angular velocity predictions as a function of the features and the previous joint angles. Velocities are integrated to produce angular predictions:

$$\mathbf{s}_t^v = \pi(\mathbf{z}_t, \mathbf{s}_{t-1}) \tag{2}$$

$$\mathbf{s}_t = \mathbf{s}_{t-1} + \mathbf{s}_t^v \tag{3}$$

where π is the decoder, $\mathbf{s}_t$ is the angular prediction at time t, and $\mathbf{s}_t^v$ is the angular velocity prediction at time t. For the *tracking* task, the ground truth first state is provided: $\mathbf{s}_0 := \mathbf{s}_0^*$. For *regression*, the ground truth state is unknown. Therefore, the decoder produces angle and angular velocity predictions ($\mathbf{s}^p$ and $\mathbf{s}^v$, respectively). The angular predictions are used for the first P time steps (250 ms in our case), and velocities are integrated thereafter:

$$\mathbf{s}_t^p, \mathbf{s}_t^v = \pi(\mathbf{z}_t, \mathbf{s}_{t-1}) \tag{4}$$

$$\mathbf{s}_t = \begin{cases} \mathbf{s}_t^p & \text{if } t < P, \\ \mathbf{s}_{t-1} + \mathbf{s}_t^v & \text{if } t \geq P \end{cases} \tag{5}$$

The LSTM has two hidden layers of size 512. We scale its output by .01, as we find that this improves training. A *Time-Depth Separable Convolution* (TDS) network is used for the featurizer, as it has been shown to be effective in the automatic speech recognition literature [Hannun et al., 2019]. The featurizer first applies three 1D convolutions over time with 256 features, kernel widths of 11, 5, and 17, and strides of 5, 2, and 4. There are then 4 TDS blocks with channel and feature widths of 16 and kernel widths of 9, 9, 5, and 5. Overall, the featurizer reduces the sample rate to 25 Hz, and a final linear up-sampling brings them to 50 Hz. We use layer norms as described in [Hannun et al., 2019].

C.2 NeuroPose

We implement the NeuroPose U-Net architecture as described in Liu et al. [2021], with minor modifications to account for differences in recording device and joint angle targets. The encoder of the original NeuroPose has 40x temporal down-sampling achieved via a series of strides. To account for the 10x greater sample rate of our device, we double each of 3 temporal strides to yield 360x down-sampling. Similarly, we double the spatial stride of the final encoder convolution to account for the 2x spatial resolution of our device. We similarly modify the up-sampling in the decoder by the same factors and add a final linear project to achieve 20 dimensional angular predictions.

Note that the original NeuroPose uses a velocity regularization term, which we do not explore here. We find that predicting velocities rather than joint angles is sufficient to achieve smooth predictions, and precludes having to tune the weight on the velocity regularization term.

C.3 SensingDynamics

The original SensingDynamics was designed for a high-density sEMG device with 320 electrodes spread over 5 separate patches on the forearm and wrist [Sîmpetru et al., 2022a]. The architecture uses 3d convolutions over channels, patches, and time. In contrast, the sEMG-RD wrist band from CTRL-labs at Reality Labs et al. [2024] does not have separate patches and has distinct channel densities and temporal resolutions. To account for these discrepancies, we use 2d convolutions over

sEMG channels and time, and modified the convolutional kernel, strides and dilations to match the effective receptive fields and strides of the original setup.

Note that the original SensingDynamics smooths the output predictions with a moving average filter of 150 ms. We find that predicting velocities rather than joint angles is sufficient to achieve similarly smooth predictions.

C.4 Training Setup

For each algorithm, we performed a hyperparamter sweep over the following parameters, with each explored independently: training window length (2000-12000 samples at 2kHz, 3 different values), learning rate (.001 or .0001), gradient norm clipping (none or 1), and whether the decoder used an MLP or LSTM (for (v)emg2pose). The most performant setting was used for each algorithm, as reported in Table 7 (for emg2pose explanation see Appendix D.2). A learning rate of 0.001 was universally optimal. To improve generalization across device placements, we use *rotation augmentation*, wherein we spatially rotate the sEMG channels by 1, 0, or -1 (uniformly sampled). Augmentation is only applied during training.

Table 7: *Algorithm comparison.* We extend the window lengths to account for receptive fields.

	Baseline	Predictions	Network	Grad clip	Window
Tracking	emg2pose	Angles	TDS + MLP	1	5790
	vemg2pose	Velocities	TDS + LSTM	1	11790
Regression	emg2pose	Angles	TDS + MLP	1	5790
	vemg2pose	Velocities	TDS + LSTM	0	11790
	NeuroPose	Angles	U-Net	0	4000
	SensingDynamics	Angles	2d Conv + MLP	0	10167

C.5 Online vemg2pose

To enable online deployment of vemg2pose it must be setup to handle sEMG data being received sequentially in discrete packets of variable temporal lengths. As such, we created a variant of vemg2pose that uses buffers to append the current packet of data to the previous ones. In addition to storing all received data in a buffer, we additionally keep track of which data have been processed already and which have not (this is a function of the network receptive field and the stride), so as not to make duplicate predictions. For Fig. 1, we trained a *vemg2pose, tracking* model with this internal variant. This internal variant achieved joint angular errors almost identical to those reported in Table 5, as expected. We do not open-source this setup, as it is only useful with access to the sEMG-RD band for online testing.

C.6 Hand mesh visualizations of prediction trajectories

We generated the articulated hand meshes representing prediction trajectories (as depicted in Fig. 5 and Appendix D.5) from sequences of joint angles using the forward kinematic and a mesh-skinning algorithms provided by UmeTrack [Han et al., 2022]. We use the generic, default hand model provided by UmeTrack. To generate the figures, we render the meshes using the Plotly and Plotly-Kaleido visualization packages [Plotly Technologies Inc., 2015].

C.7 Statistical Analysis

The Wilcoxon statistical analyses reported in Table 4 were performed on data aggregated across time for each user. That is, metrics were computed at each temporal sample, then averaged across time for each user within each experimental condition. Statistics aggregated within each user and condition are similarly used to construct distributions for all other plots and tables.

D Detailed Analyses

D.1 Analysis of Stages that are Challenging for Vision-Based Systems

We compared the same stages with and without occlusion, and found that occlusion did not negatively impact model performance, as expected (Fig. 4, left). Each subject performed the CountingWiggling and FingerPinches stages under two conditions: with the hands in front them - such that they would be visible to a headset based CV tracking system - and with the hands very close to or very far away from the body - such that they would be occluded.

We also compared stages with hand-object interactions, hand-hand interactions, and no interactions (Fig. 4, right). Hand-object interactions consisted of the Object1 and Object2 stages, in which participants interacted with a cup, a soft toy, blocks, and chess pieces. Hand-hand interactions (HandHandInteractions stage) consisted of sliding the fingers across the opposite palm, clapping the hands together, and wiggling the fingers such that the fingertips of opposite hands tap against one another. These interaction types are known to be challenging for vision-based systems. Nonetheless, performance for these stages was comparable or superior to performance in stages without any interactions. Note, however, that the behavioral distributions are different across these stages, which makes direct comparison of metrics challenging.

D.2 Velocity vs. Positional Predictions

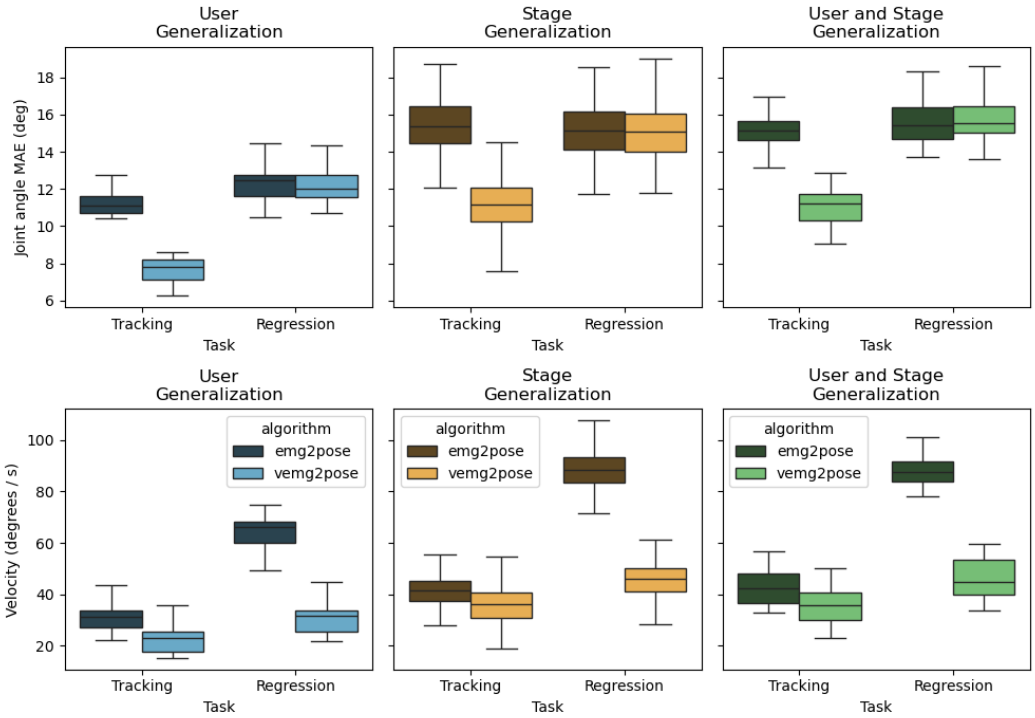

Figure 9: *vemg2pose* vs. *emg2pose* for tracking and regression tasks. Distributions are over users. Box plots take the same format as Fig. 3.

We compared *vemg2pose* to *emg2pose*, an otherwise identical algorithm that directly predicts joint angles rather than joint angular velocities (Fig. 9). emg2pose has similar joint angular error in the regression task, but much worse performance on the tracking task. This is likely because vemg2pose is initialized to the ground truth initial state, whereas emg2pose is merely conditioned on the ground truth initial state. For both tracking and regression tasks, vemg2pose has lower overall velocity than emg2pose, suggesting that operating in velocity space encourages smoother predictions.

Table 8: *Regression ablation* test set results. Mean and standard deviation are reported across users.

Test Set	Ablation	Angular Error (°)	Landmark Distance (mm)
User	vemg2pose-tran	12.5 ± 1.3	16.3 ± 1.8
	vemg2pose-lstm	12.2 ± 1.3	15.8 ± 1.9
Stage	vemg2pose-tran	16.1 ± 1.6	21.7 ± 2.1
	vemg2pose-lstm	15.2 ± 1.6	20.4 ± 2.2
User, Stage	vemg2pose-tran	16.2 ± 1.3	22.3 ± 1.8
	vemg2pose-lstm	15.8 ± 1.4	21.6 ± 2.0

D.3 LSTM vs Transformer Decoders

We ablated over decoder architectures, specifically LSTMs and transformers as the two most widely adopted models for sequence modelling. For the transformer, we explored the widely adopted transformer encoder BERT setup [Kenton and Toutanova, 2019]. We swept over the *number of layers* (2, 4, 6) and *number of heads* (2, 4, 8) reporting the best for both *regression* and *tracking* tasks in Tables 8 and 9. In order to fit into memory (Amazon EC2 `g4dn.metal` instances which have 8x NVIDIA T4 GPUs) we had to halve the feature dimensionality of the transformer decoder. In general, the transformer performs similarly or slightly worse than the LSTM.

Table 9: *Tracking ablation* test set results. Mean and standard deviation are reported across users.

Test Set	Ablation	Angular Error (°)	Landmark Distance (mm)
User	vemg2pose-tran	8.0 ± 1.0	10.7 ± 1.6
	vemg2pose-lstm	7.7 ± 1.0	10.3 ± 1.5
Stage	vemg2pose-tran	11.7 ± 1.4	15.9 ± 1.9
	vemg2pose-lstm	11.2 ± 1.4	15.2 ± 1.9
User, Stage	vemg2pose-tran	11.4 ± 1.1	16.0 ± 1.5
	vemg2pose-lstm	11.0 ± 1.0	15.4 ± 1.4

D.4 Performance Decomposition across Fingers and Joints

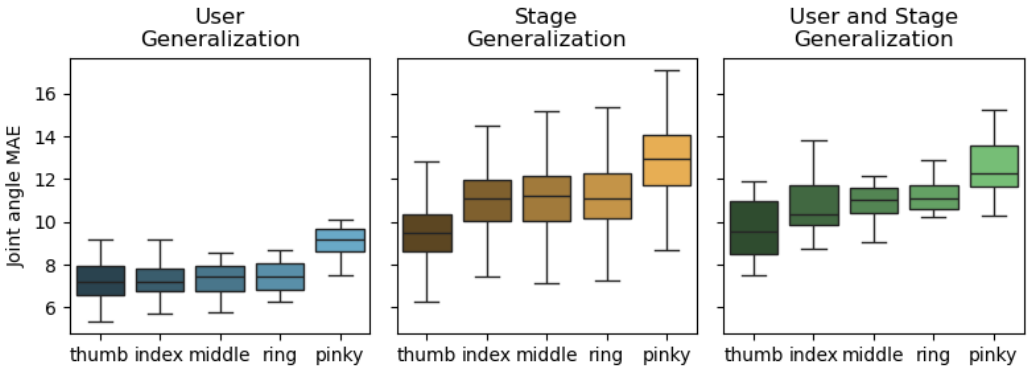

Figure 10: *Performance decomposition per finger:* for tracking task, vemg2pose. Error per finger is measured by averaging the errors of the joints associated with each finger. Distributions are over users. Box plots take the same format as Fig. 3.

We decompose *vemg2pose* tracking performance across fingers (Fig. 10), and proximal, mid, and distal joint groups (Fig. 11). For the latter, proximal, mid, and distal joints are grouped according to their distance from the palm. See Fig. 11 for further details. Reconstruction performance varies across fingers and joint groups. Thumb and pinky fingers are consistently best and worst performers, and proximal joints are more easily predicted than distal joints.

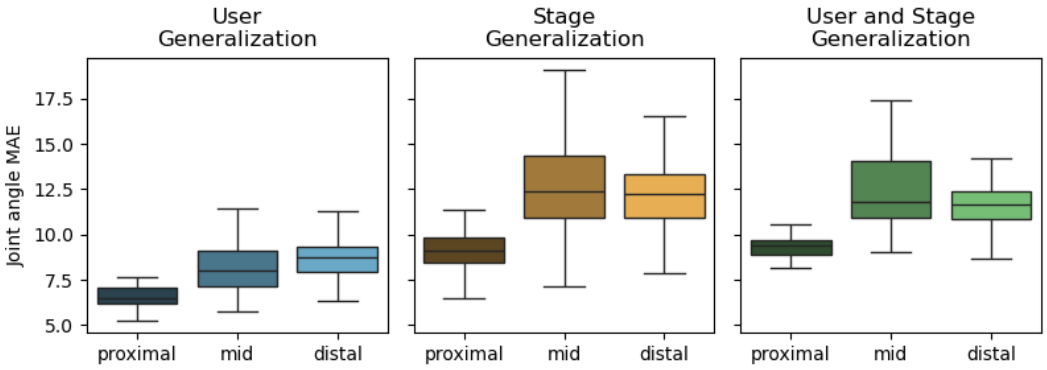

Figure 11: *Performance decomposition across joint groups:* for tracking task, vemg2pose. Performance broken down by joint according to their proximal-distal location. *Proximal* is CMC for the thumb and MCP for other fingers; *Mid* is MCP for thumb and PIP for other fingers; and *Distal* is IP for thumb and DIP for other fingers. Box plots take the same format as Fig. 3.

D.5 Tracking Trajectory Examples

We provide representative *vemg2pose, tracking* prediction trajectories for the 15%, 50% and 85% user and stage percentiles for the held-out *user, stage* scenario described in Section 3.4. Performance varies considerably across held-out users and stages, as seen in Figs. 12 to 15. We note that there may exist large variance *within* stages, for which these kinematic plots do not reflect. Minimizing these variances will be of great value. For video examples, visit `https://github.com/facebookresearch/emg2pose`.

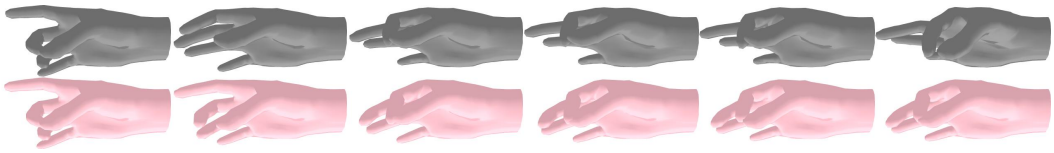

Figure 12: *Held-Out User, Stage* tracking, top 15% stage (Gesture2), median user. Top: motion capture; bottom: vemg2pose, tracking predictions. Clips unroll evenly left-to-right over a 2 seconds.

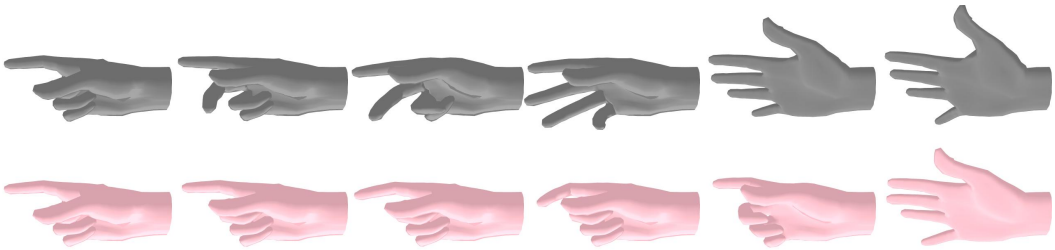

Figure 13: *Held-Out User, Stage* tracking, bottom 15% percentile stage (Counting1), median user. Top: motion capture; bottom: vemg2pose, tracking predictions. Clips unroll evenly left-to-right over a 2 seconds.

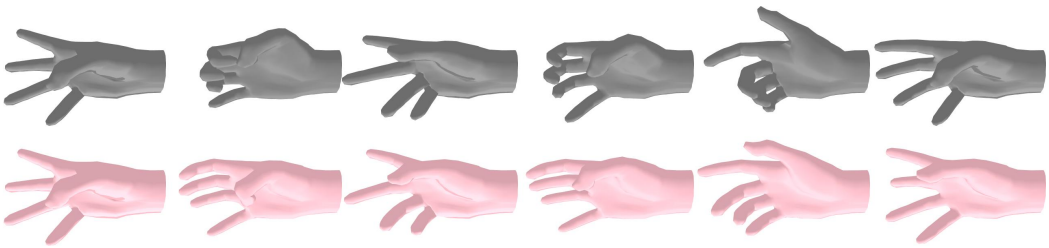

Figure 14: *Held-Out User, Stage* tracking, median stage (Counting2), top 15% percentile user. Top: motion capture; bottom: vemg2pose, tracking predictions. Clips unroll evenly left-to-right over a 2 seconds.

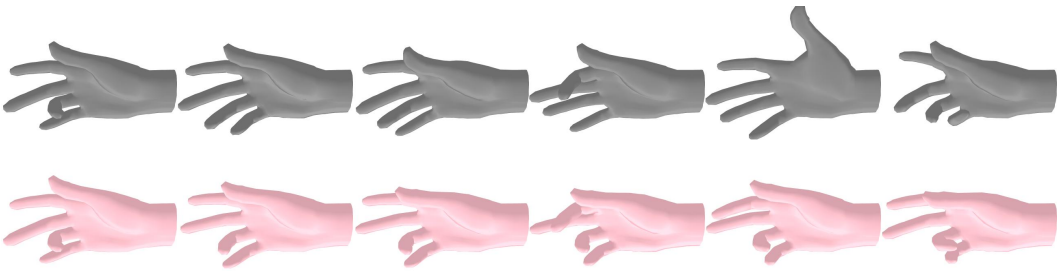

Figure 15: *Held-Out User, Stage* tracking, median stage (Wiggling2), bottom 15% percentile user. Top: motion capture; bottom: vemg2pose, tracking predictions. Clips unroll evenly left-to-right over a 2 seconds.