MUSCLESIGHT: A BENCHMARK FOR ARM-MOUNTED ULTRASOUND REPRESENTATIONS OF MUSCLE STATE IN MANUAL TASKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Arm-mounted ultrasound images the muscles that move the hand, including deep muscles that surface electromyography cannot isolate. Existing models report strong gesture and pose decoding, but are evaluated mostly within participants and sessions. We introduce MUSCLESIGHT, a dataset and benchmark for testing whether ultrasound-based hand decoding generalizes: 37.7 hours from 33 participants across 113 sessions and 37 manual tasks, including isolated motions, static postures, and interactions with 72 distinct objects. Synchronized multi-view videos provide audited hand-pose and gesture labels. Our benchmark evaluates frame- and clip-based gesture classification, pose regression, and pose tracking within sessions and on held-out sessions and participants. Within sessions, a 0.28,M-parameter CNN reaches 89.5% gesture accuracy (majority baseline: 21.0%). On held-out sessions and participants, however, gesture classifiers fail to beat the majority baseline, pose regressors improve over a constant mean pose by at most $1.8°$, and only initialized pose tracking transfers reliably. Learned representations strongly encode recording session, revealing substantial session-specific structure. Few-shot calibration on each new session improves every calibrated model, while simply adding training participants does not: a calibrated pose model trained on 2 participants outperforms an uncalibrated model trained on 28. Data, labels, splits and code will be released.

1 INTRODUCTION

Decoding hand movement, pose, and force from muscle activity enables camera-free input for prosthetics, virtual and augmented reality, and robot teleoperation and learning (Zhao et al., 2026; He et al., 2026; Verma, 2025). Surface electromyography (sEMG) has become the dominant sensing modality for this purpose. Its ecosystem is relatively mature: large public datasets are available (Atzori et al., 2014; Salter et al., 2024), benchmark protocols explicitly test generalization across recording sessions and users (Yang et al., 2024; Salter et al., 2024), and recent models trained on thousands of participants can transfer to previously unseen users (Kaifosh et al., 2025).

Ultrasound offers a complementary view of the same underlying muscle activity. Rather than measuring electrical potentials at the skin surface, B-mode ultrasound directly images a cross-section of the muscle and resolves spatial structure, including deeper muscles that surface electrodes observe only through mixed signals (Sikdar et al., 2014; He et al., 2019; Lee et al., 2025). This richer spatial information has enabled strong results in forearm-based hand decoding: published models report up to 98.8% accuracy across 12 gestures (Bimbraw et al., 2024a) and 95.6% across 9 finger motions (Lee et al., 2025). However, these numbers are difficult to interpret as evidence of generalizable hand decoding. Most existing studies evaluate participants who also appear in training: ten of the twelve works we survey in the appendix report no result on an unseen participant, and the downloadable B-mode datasets contain only six participants each (Bimbraw, 2025; 2026). As a result, it remains unclear how much current models learn hand-related muscle dynamics versus session- and participant-specific appearance.

We introduce MUSCLESIGHT (Figure 1), a dataset and benchmark for decoding the hand from arm-mounted B-mode ultrasound. It contains 37.7 h of ultrasound from 33 participants in 113 sessions,

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

recorded at 13 forearm and upper-arm sites during 37 manual tasks; 32 participants recorded two or more sessions, so a model can be tested on a new session of a seen participant as well as on a new participant. It is organized as two datasets that share all participants: MUSCLESIGHT-Hand, with isolated joint motions and static postures, and MUSCLESIGHT-HOI, with hand-object interaction involving 72 everyday objects. Synchronized stereo video provides the labels: 20 finger joint angles on both datasets, in the UmeTrack representation used for sEMG pose decoding so that ultrasound and sEMG results are comparable (Han et al., 2022; Xi et al., 2026), and 17 gesture classes and movement phases on MUSCLESIGHT-Hand. Each label layer combines automatic estimation with human annotation and auditing. The benchmark defines four tasks, gesture classification and pose regression from one frame and pose regression and tracking from a one-second window, each evaluated within sessions, on held-out sessions and on held-out participants.

Evaluating current models on MUSCLESIGHT reveals a substantial generalization gap that is obscured by within-session evaluation, the protocol used in most prior work. Models that perform strongly within a session fall to near the level of a constant predictor that ignores the ultrasound when evaluated on held-out sessions or participants. Only pose tracking, which is initialized with the true pose, transfers reliably. Our contributions are:

- **Dataset.** MUSCLESIGHT, a public arm-mounted B-mode ultrasound dataset with more than five times the participants of any downloadable forearm-ultrasound dataset of hand gestures or pose and, unlike all of them, hand-object interaction. Recordings are paired with synchronized stereo video, and the hand-pose labels use the joint-angle representation of sEMG pose decoders (§3).

- **Benchmark.** Four tasks, from single-frame gesture classification to pose tracking from video, each evaluated within sessions, on held-out sessions and on held-out participants, which separates generalization from the recall of seen sessions. Baselines range from a compact CNN trained from scratch to frozen pretrained image encoders and a pretrained video network (§4).

- **Findings.** Evidence that current models fit session-specific structure: the representations of the models that fail strongly encode the recording session, neither pretraining on ultrasound nor a larger video network closes the gap, and calibration on the new session, not a larger training cohort, improves pose regression. These results identify cross-session and cross-participant generalization as the central open problem for ultrasound hand decoding (§5).

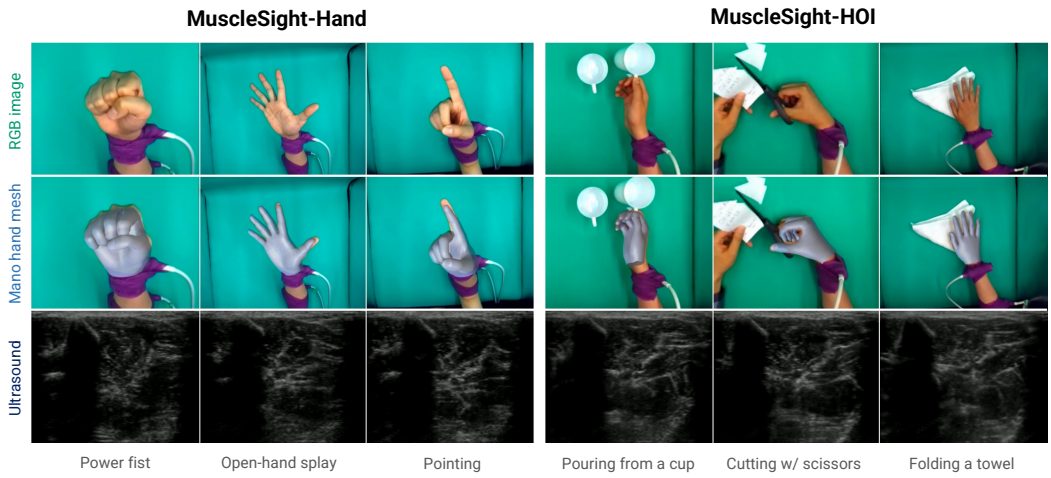


Figure 1: **MUSCLESIGHT images the muscles that drive manual action.** Each column is one instant: the overhead camera frame, the fitted MANO hand, and the ultrasound frame recorded within 20 ms. One participant and one probe site throughout, so only the task changes between columns.

Table 1: Publicly available forearm ultrasound datasets of hand gestures or hand pose. every one we found (see the appendix). Size denotes recording duration.

	Muscle sensing	Scale		
Resource	Modality	Subj	Size	HOI
Ultra-Pro (Yang et al., 2022)	A-mode	4	1.0 h	✗
Bimbraw (acoustic reflector) (Bimbraw, 2025)	B-mode	6	4.0 h	✗
Bimbraw + Vicon (Bimbraw, 2026)	B-mode	6	3.0 h	✗
MUSCLESIGHT-Hand (ours)	**B-mode**	**33**	**6.0 h**	✗
MUSCLESIGHT-HOI (ours)	**B-mode**	**33**	**11.5 h**	✓

2 RELATED WORK

Datasets of forearm muscle activity. Public sEMG datasets of hand gestures and pose are numerous and large, up to 370 hours from 193 users with motion-capture pose (Salter et al., 2024; Atzori et al., 2014; Cognolato et al., 2020; Di Domenico et al., 2025; Lomele et al., 2026; Xi et al., 2026) (five of them compared in the appendix). Public forearm ultrasound data are far smaller (Table 1): an A-mode dataset, whose transducers record one-dimensional depth traces rather than images, of four participants (Yang et al., 2022), one released participant of a ten-person A-mode study (Grandi Sgambato et al., 2025), and two B-mode datasets of six participants each (Bimbraw, 2025; 2026). The largest B-mode study, with 30 participants, shares its data only on request (Lee et al., 2025).

Hand gesture and pose from forearm ultrasound. Prior work classifies a few gestures from one frame with small convolutional networks trained from scratch, often one per participant (Bimbraw et al., 2026a; 2024b), or from short clips with spatiotemporal convolutions (Bimbraw et al., 2024a; Lee et al., 2025), and regresses joint angles from a single frame or A-mode sample (Bimbraw et al., 2023; Grandi Sgambato et al., 2025; Spacone et al., 2026). Only two of the twelve works we survey report results on held-out participants (Zadok et al., 2026; Grandi Sgambato et al., 2025), and none evaluates frozen pretrained image encoders with learned probes.

Generalization and adaptation in myoelectric interfaces. The sEMG literature treats cross-session and cross-user evaluation as standard: EMGBench benchmarks leave-one-subject-out generalization and few-shot adaptation (Yang et al., 2024), emg2pose holds out users, sensor placements and movement types (Salter et al., 2024), and decoders trained on thousands of users generalize to new ones (Kaifosh et al., 2025). For ultrasound the evidence is fragmentary: accuracy falls when the probe is replaced and per-session fine-tuning recovers part of it (Rothenberg et al., 2024), unsupervised domain adaptation on unlabeled data from a new session recovers part, but not all, of the accuracy of an A-mode classifier (Lykourinas et al., 2024), and SonoRank conditions on a rest recording taken at the start of use (Zadok et al., 2026).

Hand pose datasets from vision. Hand-object datasets provide 3D hand pose from multi-view or motion capture (Hampali et al., 2020; Chao et al., 2021; Fan et al., 2023; Liu et al., 2024) but do not measure the muscles. We use vision only for labels, in the UmeTrack joint angles used by EgoEMG (Han et al., 2022; Xi et al., 2026), so that ultrasound and sEMG results are comparable.

3 THE MUSCLESIGHT DATASET

MUSCLESIGHT records B-mode ultrasound of the forearm and upper arm while seated participants perform cued manual tasks in view of two stereo cameras (Figure 2): 37.7 h from 33 participants in 113 sessions, 17.5 h of it inside 4,225 trials. 32 participants recorded two or more sessions, so a model can be tested on a new session of a seen participant as well as on a new participant. MUSCLESIGHT-Hand carries pose, phase and posture labels, and MUSCLESIGHT-HOI pose labels only; both include all 33 participants (Figure 1); the appendix shows example recordings over time and gives per-task counts.



Figure 2: **Overview of the MUSCLESIGHT dataset and acquisition protocol.** (a) Two stereo cameras, arm-mounted ultrasound probes and the task display. (b) Representative objects. (c) Labeled probe sites on the forearm and upper arm. (d) Modalities, postures and manipulation tasks.

3.1 ACQUISITION SETUP

An ArtUs EXT-2H beamformer (Telemed) with LF9-5N60-A3 linear-array probes records 800 × 600 B-mode frames at a transmit frequency of 7 MHz and an imaging depth of 5 cm, at 24–67 fps depending on the session. The probes are worn in holders strapped to the arm for the whole session, in transverse orientation, and 82 sessions use two probes at two sites. Two ZED 2i stereo cameras record the hand at 30 fps, an oblique side camera in every session and an overhead camera in 77. All streams are timestamped in software on one Lab Streaming Layer clock (Kothe et al., 2025).

3.2 DATA COLLECTION PROTOCOL

A session is one continuous recording of cued trials; a trial's *hold window*, in which the instructed movement or posture is performed, supplies all benchmark items. The 37 tasks, listed in the appendix, form 5 families: isolated joint motions (A, 830 trials), static hand postures (B, 774), object interactions (C, 1,547), controls comprising rest, passive movement and maximum voluntary contraction (D, 616), and manipulation primitives with varied objects (E, 458). The object tasks use 72 object variants in 10 functional categories (Figure 2b). Families A and B form MUSCLESIGHT-Hand, and C, D and E form MUSCLESIGHT-HOI (§**??**). The first 22 participants performed families A to D and the last 11 A, B, D and E; 90.8% of trials are at moderate effort, and 4,170 of 4,225 pass quality control. Probes were placed at 13 sites, 7 on the forearm and 6 on the upper arm (Figure 2c); six forearm sites account for 96.9% of the probe recordings.

3.3 ANNOTATION

Hand pose. A hand-reconstruction network (Potamias et al., 2025) estimates a MANO hand (Romero et al., 2017) in every camera view at 10 Hz, and one view per recording, the overhead camera where it sees the hand, supplies the pose. Inverse kinematics converts each estimate into the 20 finger joint angles of the UmeTrack hand model (Han et al., 2022), the representation used by EgoEMG (Xi et al., 2026); the two wrist angles cannot be recovered.

Movement phases and postures. MUSCLESIGHT-Hand trials are segmented into task-specific movement phases, and the held phases define the posture classes.

Table 2: **Image-to-gesture and image-to-pose benchmarks on MuscleSight-Hand.** Image2gesture (accuracy, higher is better) and Image2pose (MAE in degrees, lower is better) under the three evaluation splits. Probes read frozen DINOv3 or EchoCLIP features. Mean over three random seeds (five for the pose probes), with the standard deviation in small type; best in bold, second best underlined.

	Gesture classification accuracy ↑			Hand-pose regression MAE (degrees) ↓		
Model	Random split	Held-out session	Held-out participant	Random split	Held-out session	Held-out participant
No input						
Constant predictor	0.210	0.248	0.205	18.82	18.81	20.06
Trained end to end on the frame						
Lite-CNN	**0.895**$_{\pm 0.014}$	0.152$_{\pm 0.014}$	0.174$_{\pm 0.024}$	**6.40**$_{\pm 0.04}$	18.37$_{\pm 0.04}$	19.91$_{\pm 0.22}$
Lite-CNN + augmentation	0.682$_{\pm 0.038}$	0.189$_{\pm 0.011}$	**0.187**$_{\pm 0.003}$	10.08$_{\pm 0.12}$	**17.03**$_{\pm 0.10}$	**18.72**$_{\pm 0.42}$
Frozen pretrained encoder + probe						
DINOv3 ViT-B/16, linear	0.247$_{\pm 0.001}$	0.133$_{\pm 0.002}$	0.117$_{\pm 0.001}$	16.37$_{\pm 0.00}$	17.67$_{\pm 0.01}$	19.39$_{\pm 0.01}$
DINOv3 ViT-B/16, MLP	0.541$_{\pm 0.010}$	**0.243**$_{\pm 0.009}$	**0.187**$_{\pm 0.007}$	11.64$_{\pm 0.04}$	18.13$_{\pm 0.10}$	20.40$_{\pm 0.12}$
EchoCLIP, linear	0.283$_{\pm 0.003}$	0.123$_{\pm 0.005}$	0.132$_{\pm 0.003}$	16.18$_{\pm 0.00}$	17.83$_{\pm 0.01}$	19.41$_{\pm 0.01}$
EchoCLIP, MLP	0.699$_{\pm 0.006}$	0.213$_{\pm 0.004}$	0.183$_{\pm 0.009}$	9.99$_{\pm 0.02}$	18.87$_{\pm 0.15}$	20.36$_{\pm 0.09}$

Table 3: **Video-to-pose benchmarks.** Pose regression and pose tracking (MAE in degrees, lower is better) on MuscleSight-Hand and MuscleSight-HOI. *Seen*: windows withheld from the training sessions, scored by the same network as *Sess.* (held-out session); *Part.*: held-out participant. Right: the temporal models on MuscleSight-Hand with participants held out. Mean over three random seeds, with the standard deviation below; best in bold, second best underlined.

	MuscleSight-Hand						MuscleSight-HOI					
	Pose regression MAE (degrees) ↓			Pose tracking MAE (degrees) ↓			Pose regression MAE (degrees) ↓			Pose tracking MAE (degrees) ↓		
Model	Seen	Sess.	Part.	Seen	Sess.	Part.	Seen	Sess.	Part.	Seen	Sess.	Part.
Constant predictor	18.91	18.86	19.99	9.25	8.85	9.82	13.75	14.05	14.08	7.13	6.77	7.21
Lite-CNN	15.46 ±1.78	17.61 ±0.42	**19.06** ±0.20	9.20 ±0.30	9.00 ±0.02	9.95 ±0.03	12.37 ±0.18	**14.05** ±0.07	**14.55** ±0.23	7.20 ±0.07	6.94 ±0.02	7.34 ±0.04
+ temporal conv.	**11.76** ±2.45	**17.30** ±0.25	19.32 ±0.24	**7.45** ±0.25	**7.98** ±0.04	**8.69** ±0.16	**11.10** ±0.94	14.58 ±0.07	14.70 ±0.24	**6.39** ±0.03	**6.45** ±0.03	**6.83** ±0.02

	MuscleSight-Hand		
		Held-out participant	
Model		Reg.	Track.
Temporal conv.		19.32 ±0.24	**8.69** ±0.16
Transformer		20.23 ±0.79	9.04 ±0.05
R(2+1)D-18		**18.44** ±0.48	9.09 ±0.09

Each layer combines automatic proposals with human annotation and auditing: on MUSCLESIGHT-Hand a person resolved ambiguous choices of the participant's hand, corrected phase boundaries and inspected every posture frame, and on both datasets a person audited recordings of the fitted hand. The appendix details each stage.

4 BENCHMARK

Every task reads frames from a single ultrasound probe at a time, and both probes contribute items in dual-probe sessions; the cameras supply the labels only, each paired with the nearest ultrasound frame within 50 ms.

4.1 TASKS

Two tasks read one frame (Table 2) and two read a one-second window (Table 3); the latter use the pose regression and tracking formulations of emg2pose (Salter et al., 2024). **Image2gesture** (gesture classification) assigns a frame to one of 17 held postures, over 27,312 MUSCLESIGHT-Hand frames sampled at 4 Hz. **Image2pose** (hand-pose regression) predicts the 20 joint angles of a frame, over 91,025 MUSCLESIGHT-Hand frames at 5 Hz. **Video2pose regression** predicts the angles at each of the ten frames of a one-second window, over 31,118 windows on MUSCLESIGHT-Hand and 69,957

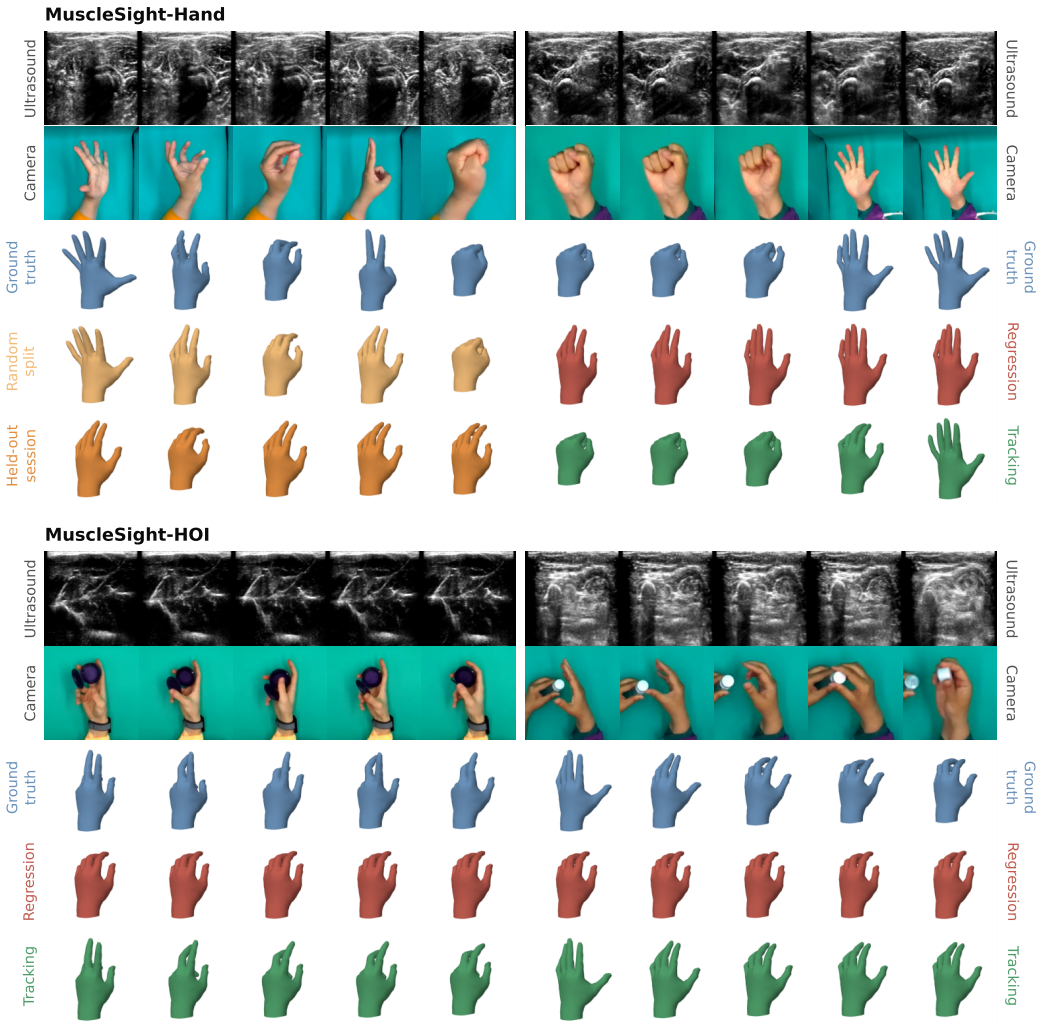


Figure 3: **Qualitative hand-pose predictions.** Rows: ultrasound input, camera frame, ground-truth pose and the predictions of two models, rendered with the UmeTrack hand model. Top left: Image2pose under the random and held-out-session splits, on five MUSCLESIGHT-Hand frames held out by every split. Top right and bottom: Video2pose regression and tracking with participants held out; the MUSCLESIGHT-Hand window is the best case among that participant's large movements.

on MUSCLESIGHT-HOI. **Video2pose tracking** is also given the pose at the first frame and is scored on the other nine.

4.2 EVALUATION SPLITS

Within a task the item pool is fixed and only the split changes. **Within-session**: the test items come from training sessions, as frames assigned at random (*random split*) or as 15% of the training sessions' windows scored by the held-out-session network (*seen*). Adjacent frames and overlapping windows make this largely a test of recall, yet it matches most published forearm-ultrasound protocols. **Held-out session**: 10 participants appear in training, at a probe site already seen for them. **Held-out participant**: 5 test, 6 validation and 22 training participants, stratified by protocol group.

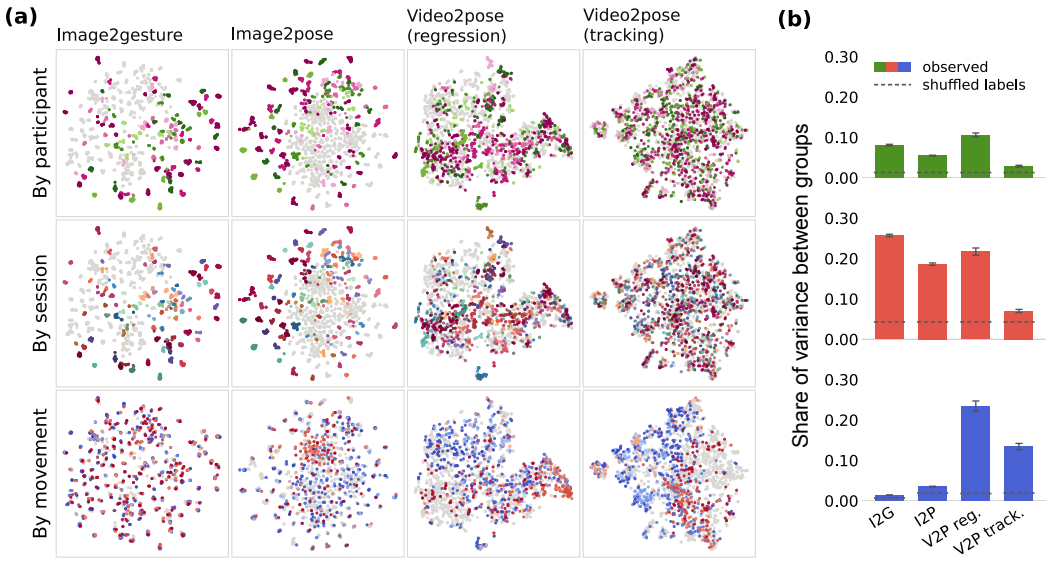


Figure 4: **Learned representations are organized by session. (a)** t-SNE of 2,500 items per model, colored by participant, session and movement phase; in each coloring the eight commonest groups are colored and everything else is gray. **(b)** Fraction of variance between groups (± 1 standard deviation over bootstrap resamples), with the bars in the order of the columns in (a), abbreviated I2G, I2P, V2P reg. and V2P track., and a key swatch carrying all three row colors; dashed lines: the same fraction with shuffled labels. Models are retrained on the held-out-participant split with one seed.

4.3 Evaluation Protocol

Gesture classification is scored by accuracy, or balanced accuracy (chance 0.059) in §5.3, and pose by the mean absolute error (MAE) in degrees over the 20 angles. Each score is compared with a *constant predictor* that ignores the ultrasound: it outputs the training fold's most frequent posture or mean pose, or holds the initial pose for tracking. Gesture and window models are read at their best validation epoch, and single-frame pose models after a fixed schedule.

4.4 Baselines

Lite-CNN is a five-block convolutional network on 128×128 frames with 0.28 M parameters, the scale of prior networks for this modality (0.11 M in the code released with Bimbraw et al., 2026a), trained from scratch with and without augmentation. **Frozen encoders**, DINOv3 ViT-B/16 (Siméoni et al., 2025) and the echocardiography model EchoCLIP (Christensen et al., 2024), are read by linear or MLP probes. **Temporal models** follow the per-frame Lite-CNN with no temporal layer, two temporal convolutions, the second dilated (2.0 M parameters) or two Transformer layers (2.4 M), and R(2+1)D-18 (Tran et al., 2018) (31.4 M) is fine-tuned from Kinetics-400. Tracking models predict the displacement from the given pose. The appendix gives the training details.

5 Experiments

5.1 Benchmark Results

Within-session evaluation overstates performance. When the test items come from sessions seen in training, every task appears nearly solved by the smallest model (Tables 2 and 3): Lite-CNN reaches 0.895 accuracy (constant predictor: 0.210) and 6.40° MAE (18.82°), and the Video2pose models 11.76° for regression and 7.45° for tracking (18.91° and 9.25°).

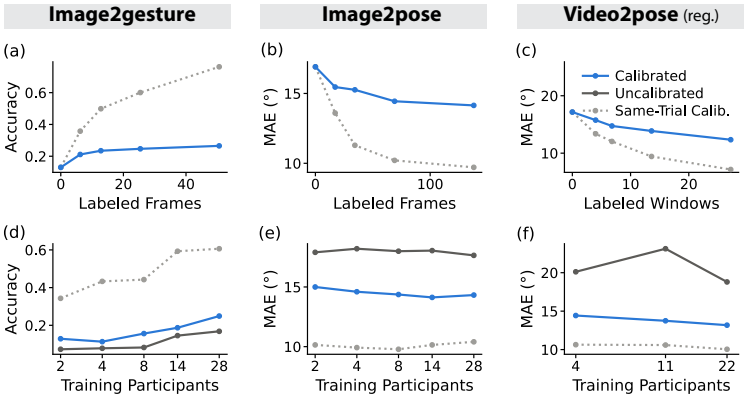


Figure 5: **Few-shot session calibration (a–c) and participant scaling (d–f). (a–c)** Performance after fine-tuning on k labeled items of the test session. **(d–f)** Models trained on n participants, before (gray) and after (blue) calibration. Dotted: calibration and evaluation drawn from the same trials as an upper bound control.

Performance does not transfer across sessions or participants. No gesture model exceeds the constant predictor on held-out sessions (0.243, against 0.248) or participants (0.187, against 0.205), and the best pose-regression models improve on it by at most $1.8°$. With the network held fixed, withholding the session raises the Video2pose regression error of the temporal convolution model by $5.5°$ on MUSCLESIGHT-Hand and $3.5°$ on MUSCLESIGHT-HOI (Table 3, *Seen* against *Sess.*). Figure 3 shows the gap: the random-split Image2pose model reproduces each posture, whereas the held-out models predict nearly the same flexed hand for every posture.

Ultrasound pretraining and model scale do not close the gap. On held-out participants, EchoCLIP, pretrained on cardiac ultrasound, is no better than DINOv3 on Image2pose ($19.41°$ against $19.39°$ with linear probes), and the 31.4 M-parameter R(2+1)D-18 reaches $18.44°$ on Video2pose regression, against $19.99°$ for the constant predictor.

Pose tracking transfers to new participants. Given the initial pose, tracking reaches $8.69 \pm 0.16°$ on held-out participants of MUSCLESIGHT-Hand, against $9.82°$ for holding the initial pose, and $6.83°$ against $7.21°$ on MUSCLESIGHT-HOI; withholding the session costs it only $0.53°$ and $0.06°$.

5.2 REPRESENTATION ANALYSIS

Figure 4 measures the fraction of each representation's variance lying between sessions, participants and movement phases, against shuffled labels. The between-session fraction is 0.257, 0.186 and 0.217 for Image2gesture, Image2pose and Video2pose regression, 4.3 to 6.0 times the shuffled 0.043, whereas the single-frame models carry little movement-phase variance (0.014 and 0.035). The tracking model is the exception, with 0.070 between sessions (1.6 times the shuffled level) and 0.134 between movement phases (7.1 times): the only representation not dominated by session belongs to the only model that transfers. The items are drawn from all folds, so part of this structure reflects the fit to training sessions.

5.3 FEW-SHOT SESSION CALIBRATION

As in the adaptation setting of EMGBench (Yang et al., 2024), we fine-tune all parameters of a model trained on 28 other participants (22 for Video2pose) for 300 steps on k labeled items from one session of each held-out participant, drawn from trials that are never scored (Figure 5a–c). Image2gesture balanced accuracy rises from 0.130 to 0.265 with about 51 frames, Image2pose error falls from $16.9°$ to $14.1°$, and Video2pose regression error from $17.2°$ to $12.3°$. Drawing the items from the scored trials instead (0.763, $9.7°$, $7.2°$) measures the fit to seen trials.

5.4 Scaling the Number of Training Participants

Training from scratch on 2 to 28 other participants (4 to 22 for Video2pose) with a fixed number of steps (Figure 5d–f) improves Image2gesture, from 0.073 to 0.168 balanced accuracy, or from 0.129 to 0.249 with calibration, whose gain does not shrink. It does not improve the pose models: Image2pose error is 17.9° with 2 participants and 17.6° with 28, and Video2pose regression varies non-monotonically (20.1°, 23.1° and 18.8°). With calibration, 2 participants outperform 28 uncalibrated ones by 2.6° on Image2pose, and 4 outperform 22 by 4.4° on Video2pose. Every point is a separate model trained from scratch, so the largest cohort does not reproduce the uncalibrated scores of §5.3.

6 Limitations and Future Work

The leave-one-participant-out evaluation is the most important extension. The pose labels come from a vision pipeline, with conversion error, no wrist angles. The recordings are transverse and mostly at moderate effort, site coverage is uneven, and the data come from one laboratory. Two-probe fusion, longer windows, the single-frame tasks on MuscleSight-HOI and adaptation from unlabeled test-session data are not yet evaluated.

7 Conclusion

MuscleSight pairs arm-mounted B-mode ultrasound from 33 participants with camera-derived hand-pose labels in two datasets, MuscleSight-Hand and MuscleSight-HOI, and its benchmark evaluates hand decoding on held-out sessions and participants as well as within sessions. Current models perform well within a session but not generalize across sessions and participants. The representations of most models strongly encode the recording session, and only pose tracking, given the initial pose, transfers to new sessions and subjects. Few-shot calibration on the new session narrows the gap, whereas more training participants do not improve pose regression, which points to session-invariant representations and efficient calibration as the directions for progress. We expect held-out-session and held-out-participant evaluation to become the standard for ultrasound interfaces, as it is for sEMG.

AI use statement

We used a video-understanding model (Gemini video understanding) for an initial segmentation of the recorded trials into movement phases, followed by human review and correction in an annotation tool. Generative AI tools were also used to suggest improvements to writing clarity and wording. The authors evaluated these suggestions and reviewed the final manuscript.

Ethics statement

The study was approved by the institutional review board of the authors' institution; the protocol number is withheld for anonymous review. All participants gave written informed consent. Faces will be masked in the released camera videos. Ultrasound was acquired within the acoustic output limits of the imaging system, and the probes were operated by trained research staff. Participants received no monetary compensation. The terms of use will be specified in the license released with the data.

Reproducibility statement

The raw recordings, timestamp tables, protocol metadata, annotation layers, frozen item tables, split definitions and evaluation code will be released upon acceptance, together with their license. The exact commit and configuration files will be listed with the release, every model was trained on a single NVIDIA L40S or RTX 4090 GPU, and the appendix gives the architectures and hyperparameters.

REFERENCES

Manfredo Atzori, Arjan Gijsberts, Claudio Castellini, Barbara Caputo, Anne-Gabrielle Mittaz Hager, Simone Elsig, Giorgio Giatsidis, Franco Bassetto, and Henning Müller. Electromyography data for non-invasive naturally-controlled robotic hand prostheses. *Scientific Data*, 1:140053, 2014. doi: 10.1038/sdata.2014.53.

Keshav Bimbraw. Forearm ultrasound gestures dataset with and without acoustic reflector. Zenodo, 2025. URL https://doi.org/10.5281/zenodo.17386583. Data record for the article in *IEEE Sensors Journal*, doi:10.1109/JSEN.2025.3621577.

Keshav Bimbraw. Forearm ultrasound dataset for hand gesture classification and finger joint angle estimation. Zenodo, 2026. URL https://doi.org/10.5281/zenodo.21825018. Data record for the article in *IEEE Transactions on Medical Robotics and Bionics* 5(1), 2023.

Keshav Bimbraw, Christopher J. Nycz, Matthew Schueler, Ziming Zhang, and Haichong K. Zhang. Simultaneous estimation of hand configurations and finger joint angles using forearm ultrasound. *IEEE Transactions on Medical Robotics and Bionics*, 5(1):120–132, 2023. doi: 10.1109/TMRB.2023.3237774. Data record: https://doi.org/10.5281/zenodo.21825018.

Keshav Bimbraw, Ankit Talele, and Haichong K. Zhang. Hand gesture classification based on forearm ultrasound video snippets using 3D convolutional neural networks. In *2024 IEEE Ultrasonics, Ferroelectrics, and Frequency Control Joint Symposium (UFFC-JS)*, pp. 1–4, 2024a. doi: 10.1109/UFFC-JS60046.2024.10794148. arXiv:2409.16431.

Keshav Bimbraw, Haichong K. Zhang, and Bashima Islam. Forearm ultrasound based gesture recognition on edge. In *2024 IEEE 20th International Conference on Body Sensor Networks (BSN)*, pp. 1–4, 2024b. doi: 10.1109/BSN63547.2024.10780669. arXiv:2409.09915.

Keshav Bimbraw, Yichuan Tang, and Haichong K. Zhang. Acoustic reflector-based forearm ultrasound for hand gesture classification. *IEEE Sensors Journal*, 26(10):14896–14905, 2026a. doi: 10.1109/JSEN.2025.3621577.

Keshav Bimbraw, Ye Wang, Jing Liu, and Toshiaki Koike-Akino. GPT sonography: Hand gesture decoding from forearm ultrasound images via a large vision-language model. *IEEE Access*, 14: 70724–70736, 2026b. doi: 10.1109/ACCESS.2026.3687477. arXiv:2407.10870.

Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox. DexYCB: A benchmark for capturing hand grasping of objects. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Matthew Christensen, Milos Vukadinovic, Neal Yuan, and David Ouyang. Vision–language foundation model for echocardiogram interpretation. *Nature Medicine*, 30:1481–1488, 2024. doi: 10.1038/s41591-024-02959-y.

Matteo Cognolato, Arjan Gijsberts, Valentina Gregori, Gianluca Saetta, Katia Giacomino, Anne-Gabrielle Mittaz Hager, Andrea Gigli, Diego Faccio, Cesare Tiengo, Franco Bassetto, Barbara Caputo, Peter Brugger, Manfredo Atzori, and Henning Müller. Gaze, visual, myoelectric, and inertial data of grasps for intelligent prosthetics. *Scientific Data*, 7:43, 2020. doi: 10.1038/s41597-020-0380-3.

Dario Di Domenico, Inna Forsiuk, Simon Müller-Cleve, Simone Tanzarella, Florencia Garro, Andrea Marinelli, Michele Canepa, Matteo Laffranchi, Michela Chiappalone, Chiara Bartolozzi, Lorenzo De Michieli, Nicolò Boccardo, and Marianna Semprini. Reach&Grasp: a multimodal dataset of the whole upper-limb during simple and complex movements. *Scientific Data*, 12:233, 2025. doi: 10.1038/s41597-025-04552-5.

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

Bruno Grandi Sgambato, Bálint K. Hodossy, Deren Yusuf Barsakcioglu, Xingchen Yang, Anette Jakob, Marc Fournelle, Meng-Xing Tang, and Dario Farina. Virtual reality interactions via a user-generic ultrasound human-machine interface for wrist and hand tracking. *Nature Communications*, 16(1):11062, 2025. doi: 10.1038/s41467-025-66001-6.

Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent 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 (CVPR)*, pp. 3196–3206, 2020. doi: 10.1109/CVPR42600.2020.00326.

Shangchen Han, Po-chen Wu, Yubo Zhang, Beibei Liu, Linguang Zhang, Zheng Wang, Weiguang Si, Peizhao Zhang, Yujun Cai, Tomas Hodan, Randi Cabezas, Luan Tran, Muzaffer Akbay, Tsz-Ho Yu, Cem Keskin, and Robert Wang. UmeTrack: Unified multi-view end-to-end hand tracking for VR. In *SIGGRAPH Asia 2022 Conference Papers*, 2022. doi: 10.1145/3550469.3555378.

Botao He, Zhi Wang, Linna Kuang, Ishaan Ghosh, Jitendra Malik, Cornelia Fermüller, Tingfan Wu, Jiayuan Mao, Ruoshi Liu, Haozhi Qi, and Yiannis Aloimonos. ForceBand: Learning forceful manipulation with sEMG. *arXiv preprint arXiv:2606.26093*, 2026.

Jiayuan He, Henry Luo, Jie Jia, John T. W. Yeow, and Ning Jiang. Wrist and finger gesture recognition with single-element ultrasound signals: A comparison with single-channel surface electromyogram. *IEEE Transactions on Biomedical Engineering*, 66(5):1277–1284, 2019. doi: 10.1109/TBME.2018.2872593.

Patrick Kaifosh, Thomas R. Reardon, and CTRL-labs at Reality Labs. A generic non-invasive neuromotor interface for human-computer interaction. *Nature*, 645(8081):702–711, 2025. doi: 10.1038/s41586-025-09255-w.

Christian Kothe, Seyed Yahya Shirazi, Tristan Stenner, David Medine, Chadwick Boulay, Matthew I. Grivich, Fiorenzo Artoni, Tim Mullen, Arnaud Delorme, and Scott Makeig. The Lab Streaming Layer for synchronized multimodal recording. *Imaging Neuroscience*, 3:IMAG.a.136, 2025. doi: 10.1162/IMAG.a.136.

Yohan Lee, Keunsoo Ko, Jaeyeop Jang, Jung-Min Lee, and Changhan Yoon. A deep learning framework for finger motion recognition using forearm ultrasound imaging. *Scientific Reports*, 15: 39719, 2025. doi: 10.1038/s41598-025-23348-6. Data available only on request to the corresponding author.

Jiwei Li, Bi Zhang, Xiaowei Tan, Wanxin Chen, Zhaoyuan Liu, Juanjuan Zhang, Weiguang Huo, Jian Huang, Lianqing Liu, and Xingang Zhao. K2MUSE: A human lower-limb multimodal walking dataset spanning task and acquisition variability for rehabilitation robotics. *The International Journal of Robotics Research*, 2026. doi: 10.1177/02783649261444046.

Yun Liu, Haolin Yang, Xu Si, Ling Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, and Li Yi. TACO: Benchmarking generalizable bimanual tool-ACtion-object understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

G. Lomele, T. Lencioni, S. D'Ambrosio, A. Comanducci, F. Lucchetti, A. Marzegan, C. Derchi, S. Garzonio, T. Atzori, M. Rabuffetti, P. Castiglioni, M. Ferrarin, and L. Fornia. High-density EEG and multi-muscle EMG dataset during object prehension with a sensorized grasping box in humans. *Scientific Data*, 13:932, 2026. doi: 10.1038/s41597-026-07242-y.

Antonios Lykourinas, Xavier Rottenberg, Francky Catthoor, and Athanassios Skodras. Unsupervised domain adaptation for inter-session re-calibration of ultrasound-based HMIs. *Sensors*, 24 (15):5043, 2024. doi: 10.3390/s24155043.

Roger Pallarès-López, Duarte Folgado, Uriel Magana-Salgado, Jessica Rosendorf, Enya Ryu, Micha Feigin-Almon, Hugo Gamboa, Luca Daniel, Brian W. Anthony, and Praneeth Namburi. A multimodal biomechanics dataset with synchronized kinematics and internal tissue motions during reaching. *Scientific Data*, 2026. doi: 10.1038/s41597-026-07019-3.

Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9826–9836, 2024.

Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. WiLoR: End-to-end 3D hand localization and reconstruction in-the-wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. arXiv:2409.12259.

Manuel Sam Ribeiro, Jennifer Sanger, Jing-Xuan Zhang, Aciel Eshky, Alan Wrench, Korin Richmond, and Steve Renals. TaL: A synchronised multi-speaker corpus of ultrasound tongue imaging, audio, and lip videos. In *IEEE Spoken Language Technology Workshop (SLT)*, 2021. arXiv:2011.09804.

Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 36(6): 245:1–245:17, 2017. doi: 10.1145/3130800.3130883.

Jack Rothenberg, Keshav Bimbraw, and Haichong K. Zhang. Improving intersession reproducibility for forearm ultrasound based hand gesture classification through an incremental learning approach. In *IEEE Ultrasonics, Ferroelectrics, and Frequency Control Joint Symposium (UFFC-JS)*, pp. 1–4, 2024. doi: 10.1109/UFFC-JS60046.2024.10793746. arXiv:2409.16415.

Sasha Salter, Richard Warren, Collin Schlager, Adrian Spurr, Shangchen Han, Rohin Bhasin, Yujun Cai, Peter Walkington, Anuoluwapo Bolarinwa, Robert Wang, Nathan Danielson, Josh Merel, Eftychios Pnevmatikakis, and Jesse Marshall. emg2pose: A large and diverse benchmark for surface electromyographic hand pose estimation. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, volume 37, pp. 55703–55728, 2024. doi: 10.52202/079017-1770.

Siddhartha Sikdar, Huzefa Rangwala, Emily B. Eastlake, Ira A. Hunt, Andrew J. Nelson, Jayanth Devanathan, Andrew Shin, and Joseph J. Pancrazio. Novel method for predicting dexterous individual finger movements by imaging muscle activity using a wearable ultrasonic system. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 22(1):69–76, 2014. doi: 10.1109/TNSRE.2013.2274657.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Giusy Spacone, Luca Benini, and Andrea Cossettini. Toward whole hand and wrist kinematic tracking with a wearable A-Mode ultrasound probe. *IEEE Open Journal of Ultrasonics, Ferroelectrics, and Frequency Control*, 6:232–235, 2026. doi: 10.1109/OJUFFC.2026.3727192. arXiv:2606.22333.

Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6450–6459, 2018.

Sagar Verma. emg2tendon: From sEMG signals to tendon control in musculoskeletal hands. In *Proceedings of Robotics: Science and Systems (RSS)*, 2025. doi: 10.15607/RSS.2025.XXI.135.

Ziheng Xi, Jiayi Yu, Yitao Wang, Yanbo Duan, Jianjiang Feng, and Jie Zhou. EgoEMG: A multimodal egocentric dataset with bilateral EMG and vision for hand pose estimation. *arXiv preprint arXiv:2605.05712*, 2026.

Jehan Yang, Maxwell Soh, Vivianna Lieu, Douglas J. Weber, and Zackory Erickson. EMGBench: Benchmarking out-of-distribution generalization and adaptation for electromyography. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, volume 37, pp. 50313–50342, 2024. doi: 10.52202/079017-1593.

Xingchen Yang, Yifan Liu, Zongtian Yin, Pu Wang, Peilin Deng, Zhengen Zhao, and Honghai Liu. Simultaneous prediction of wrist and hand motions via wearable ultrasound sensing for natural control of hand prostheses. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 30:2517–2527, 2022. doi: 10.1109/TNSRE.2022.3197875. Ultra-Pro data record: `https://doi.org/10.6084/m9.figshare.20448489`.

Dean Zadok, Alon Wolf, Alex M. Bronstein, and Oren Salzman. SonoRank: Towards calibration-free real-time finger flexion detection from forearm ultrasound sequences. *arXiv preprint arXiv:2607.07542*, 2026.

Qiang Zhang, Noor Hakam, Oluwasegun Akinniyi, Ashwin Iyer, Xuefeng Bao, and Nitin Sharma. AnkleImage — an ultrafast ultrasound image dataset to understand the ankle joint muscle contractility. *Scientific Data*, 2024. doi: 10.1038/s41597-024-04285-x.

Qianyou Zhao, Wenqiao Li, Chiyu Wang, and Kaifeng Zhang. DexEMG: Towards dexterous teleoperation system via EMG2Pose generalization. *arXiv preprint arXiv:2603.05861*, 2026.

A Dataset Details

A.1 Acquisition and Synchronization

The ultrasound system records B-mode frames of 800×600 pixels at 24–67 fps, depending on the acquisition settings of the session. The frame rate stored in the video container does not match the acquisition rate, so all timing uses per-frame timestamps. In the 82 dual-probe sessions the two probes are recorded as separate streams, and probe A is the stream the capture software binds to the first of its two ports, a binding fixed in the software rather than stored in the recordings. The ZED 2i cameras record side-by-side stereo pairs of 1280×720 images at 30 fps.

All sensor streams are timestamped by Lab Streaming Layer (Kothe et al., 2025) on one free-running clock, and streams are aligned by these timestamps, never by frame index. The ultrasound stream also carries wall-clock timestamps, and the protocol software records its trial markers in wall-clock time, so a linear fit between the two clocks, made per session because their offset drifts by about 10 ms over twenty minutes, maps every marker onto the sensor clock. Row N of a timestamp table corresponds to frame N of its video, and 2,733 of the 18,031,722 recorded frames were dropped. The benchmark pairs each fitted hand pose with the ultrasound frame of the same probe that is nearest in time and discards a pose that has none within 50 ms. Synchronization is therefore by software and looser than the hardware triggers of several related datasets (Table 9). Table 4 summarizes the corpus.

A.2 Tasks and Protocol

Table 5 lists the 37 tasks. Protocol markers time each trial (start, cue, beginning and end of the hold window, end), and the hold window is the interval in which the instructed movement or posture is performed. At recording time the operator rates each trial as good, acceptable, repeat or unusable, and 4,170 of the 4,225 trials are good or acceptable. Trials are cued at light, moderate or maximal effort, and 90.8% are moderate. The protocol changed during collection: the first 22 participants performed families A to D and the last 11 families A, B, D and E, so the object tasks of MuscleSight-HOI come from two disjoint groups of participants, and the splits are stratified by group (§C.2).

The probes are worn on one arm per session. The arm fields of the protocol form proved unreliable, so a person read the instrumented arm from the camera for each of the 109 sessions with bare-hand trials: the left arm in 58 sessions and the right in 51, with no session placing one probe on each arm.

A.3 Objects

The object tasks use 72 distinct object variants in 10 functional categories (Table 6), spanning rigid household objects, compliant objects, articulated mechanisms, tools, electronics, stationery and personal-care items. The collection comprises about 80 fixed physical instances, and some tasks use sets of small objects: the push-pin tasks, for example, involve the simultaneous manipulation of 5 to more than 20 visually similar pins. The objects were selected to vary in geometry, size, material, compliance and the interactions they afford.

A.4 Composition

Figure 6 crosses task with imaging site in each dataset. Site codes give the segment (FA forearm, UA upper arm), the aspect (anterior, posterior or lateral) and the level (proximal, mid or distal). Coverage is crossed but uneven: six forearm sites carry 96.9% of the probe recordings and the upper-arm sites 2.7%.

A.5 Example Recordings

Figure 7 shows one recording from each dataset over its hold window, with both cameras, both probes and, for MuscleSight-Hand, the reviewed phase track.

Table 4: MUSCLESIGHT at a glance. Acquisition facts, stated in full, including the ones on which the resource does not compare favorably. Synchronization is reported here; Table 1 compares modality and scale only.

	Value	
Corpus		
Participants	33	31 with demographics; trials unevenly distributed
Sessions	113	
Trials recorded	4,225	
Trials passing QC	4,170	49 unusable, 6 repeat
Frames	18,031,722	2,733 dropped
Duration	37.7 h	wall clock across both streams
Protocol		
Tasks	37	A: 830 / B: 774 / C: 1,547 / D: 616 / E: 458
Imaging sites, primary probe	10	top six carry 96.9% of trials
Imaging sites, both probes	13	upper arm only 2.7%
Probe orientations	1	transverse in every trial
Effort levels	3	light: 187 / maximal: 202 / moderate: 3,836
Dual-probe trials	3,015	two sites imaged at once
Synchronization and pairing		
Alignment	software	LSL timestamps, linear clock fit; no hardware trigger
Hand-pose rate	10 Hz	every third camera frame
Pairing tolerance	50 ms	pose to nearest ultrasound frame, else dropped
Annotation		
Hand pose, MUSCLESIGHT-Hand	1,492 recordings	Audited by a human annotator
Hand pose, MUSCLESIGHT-HOI	2,231 recordings	Audited by a human annotator
Movement phases	1,440 trials	MUSCLESIGHT-Hand only; 583 reviewed by a person
Posture labels	27,990 frames	all inspected by a person; 678 flagged

B ANNOTATION

The benchmark labels come from the two cameras in three layers: hand pose for every annotated trial, and movement phases and posture classes for the trials of MUSCLESIGHT-Hand. Each layer is proposed automatically and then reviewed in part by a person (Table 4).

B.1 HAND POSE

Detection. Every trial that passes quality control and has at least the side camera is annotated over its hold window. In each view, that is, each image of each stereo camera, the detector of WiLoR (Potamias et al., 2025) proposes hand boxes on every third frame (10 Hz), and WiLoR estimates a MANO hand (Romero et al., 2017) in each box; up to three candidates are kept per frame.

Choice of the participant's hand. A view can also contain the participant's other hand, the operator's hand, a reflection in the metal enclosure or the probe bundle strapped to the forearm. The candidates are linked over time into tracks, and one track per view is proposed as the participant's hand from the colored strap at the wrist, the detector's estimate of handedness and the image brightness. On MUSCLESIGHT-Hand, for every recording the proposal marked as ambiguous, a person inspected the uncropped scene with every track drawn and overruled the proposal where it was wrong. On MUSCLESIGHT-HOI, where a second hand is in view far more often, the proposal, calibrated on these human decisions, is used without review.

Fitting and choice of camera. The two images of a stereo camera are fused by averaging their 45 finger-joint parameters, which are expressed in the hand's own frame, and frames on which the two images disagree by more than 10 mm are discarded. One camera supplies the pose of each recording: the overhead camera where it is present and sees the hand, which holds for 75.7% of

Table 5: The task battery. *DOF* is the joint the task moves, which the protocol varies systematically and crosses against the imaging site. *Trials* counts every recorded trial; *QC* counts those passing quality control.

Task	Description	DOF	Trials	QC
Isolated joint motions — 6 tasks, 830 trials				
A1	Elbow flexion / extension	elbow	149	140
A2	Wrist flexion / extension	wrist	132	130
A3	Wrist pronation / supination	forearm	134	133
A4	Finger flexion (mass)	finger	131	131
A5	Thumb opposition	finger	144	143
A6	Wrist radial / ulnar deviation	wrist	140	138
Static hand postures — 6 tasks, 774 trials				
B1	Power fist (sustained hold)	finger	139	139
B2	Open hand splay	finger	130	127
B3	Pointing (index extend)	finger	119	119
B4	Peace sign / V-sign	finger	122	122
B5	Pinch posture (thumb-index)	finger	132	129
B6	Sequential finger taps	finger	132	131
Object interactions — 12 tasks, 1547 trials				
C1	Lift and hold a cup	multi-joint (object)	239	237
C2	Pour from a bottle	multi-joint (object)	143	142
C3	Squeeze a stress ball	multi-joint (object)	300	298
C4	Turn a key / doorknob	multi-joint (object)	116	115
C5	Use scissors (cut paper)	multi-joint (object)	106	103
C6	Write with a pen	multi-joint (object)	105	104
C7	Pull open a drawer	multi-joint (object)	105	102
C8	Unscrew a jar lid	multi-joint (object)	178	177
C9	Wipe the desk	multi-joint (object)	48	48
C10	Click / press the button	multi-joint (object)	88	87
C11	Open the box / lid	multi-joint (object)	72	71
C12	Fold the towel	multi-joint (object)	47	47
Controls — 3 tasks, 616 trials				
D1	Rest (arm supported)	control	118	114
D2	Passive movement	control	156	149
D3	Maximum voluntary contraction (MVC)	control	342	339
Manipulation primitives — 10 tasks, 458 trials				
EG1	Acquire / hold	multi-joint (object)	46	45
EG2	Relocate / place	multi-joint (object)	65	65
EG3	Translate by force	multi-joint (object)	55	54
EG4	Reorient	multi-joint (object)	52	52
EG5	Open / close / articulate	multi-joint (object)	49	49
EG6	Transfer contents	multi-joint (object)	26	26
EG7	Activate / trigger	multi-joint (object)	46	46
EG8	Deform / reshape	multi-joint (object)	45	45
EG9	Tool-mediated effect	multi-joint (object)	36	36
EG10	Attach / detach / fasten	multi-joint (object)	38	37

the MUSCLESIGHT-Hand items and 79.8% of the MUSCLESIGHT-HOI items, and the side camera otherwise. Fusing the two cameras did not improve on a single camera in our measurements.

Conversion to joint angles. MANO and the UmeTrack hand model (Han et al., 2022) differ in skeleton, bone lengths and rest posture, so the conversion is inverse kinematics rather than a closed-form mapping. One similarity transform maps MANO's canonical frame onto UmeTrack's. It is estimated once from the two rest poses, refined jointly with per-frame angles on a sample of postures, and shared by both datasets, so that their angles follow one convention. For every frame, L-BFGS fits the 20 finger angles, each passed through a sigmoid scaled to the model's joint limits, so that the UmeTrack landmarks reach the MANO joints; the hand shape is the model's mean shape, as in the released code of EgoEMG (Xi et al., 2026). Part of the landmark distance is a fixed offset

Table 6: Object inventory by functional category. Shares are of the 72 variants and are rounded.

Category	Variants	Share (%)
Personal care and medical	12	16.7
Office and stationery	12	16.7
Mechanisms, articulated and elastic objects	11	15.3
Soft and compliant objects, textiles	9	12.5
Electronics and accessories	8	11.1
Hand tools and measurement	7	9.7
Rigid household objects, storage and surfaces	5	6.9
Drinkware	4	5.6
Bottles and dispensers	3	4.2
Food and consumables	1	1.4
Total	72	100

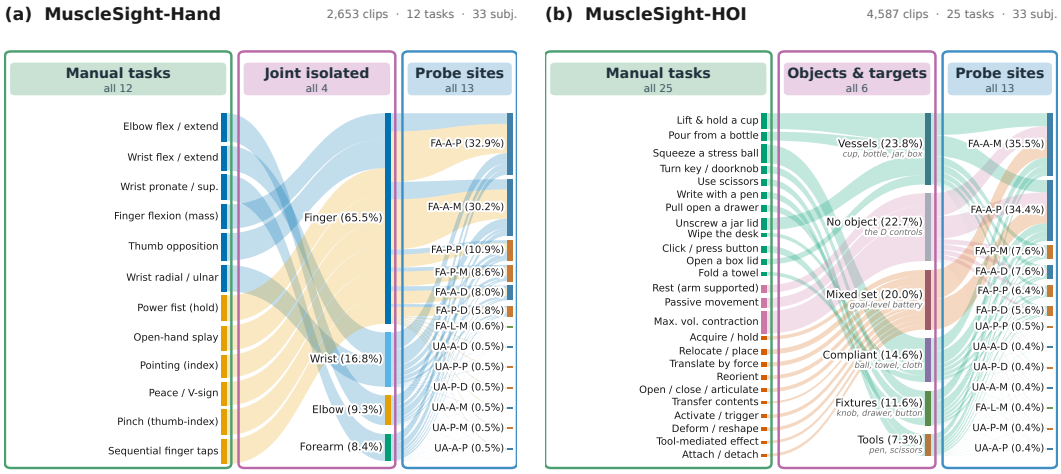


Figure 6: **Dataset composition, one taxonomy per dataset.** Each panel is a three-stage Sankey over the probe-clips of one half of the corpus, a probe-clip being one (trial, probe placement) pair, so 4,225 trials recorded with one or two simultaneous probes yield more clips than trials. Ribbon width is clip count and every ribbon keeps the color of its task family. (a) MUSCLESIGHT-Hand, families A and B, which the protocol makes free-limb or isometric by design: every clip in this half would land on a single *no object* node, so the middle column draws the joint the task isolates instead. That is the axis that varies here, and it is the one crossed against imaging site in §3.2. (b) MUSCLESIGHT-HOI, families C, D and E, where the middle column is the object or contact target the protocol puts in the hand. It is prescribed per task rather than annotated per trial, and the E battery deliberately varies it, which is why those tasks share one node; the D controls (rest, passive movement and maximum voluntary contraction) hold nothing and form the *no object* node. Coverage is crossed but far from uniform in both halves: the upper-arm sites carry 2.7% of trials between them, and a reader following any single task to the right sees it land on two sites and not thirteen. Within the task column, tasks are ordered and colored by family, so a family reads as a block. Percentages are within each panel's own dataset and so do not sum across the two. Site codes read segment, aspect and level: `FA` or `UA` for fore- or upper arm, `A`, `P` or `L` for anterior, posterior or lateral, which is the bar color, and `P`, `M` or `D` for proximal, mid or distal. Every participant recorded A, B and D, and each recorded exactly one of C and E, so the object tasks of MUSCLESIGHT-HOI come from two disjoint groups of people.

between the two skeletons at the wrist and knuckles, which no finger angle can change. The part the angles can move, which measures the posture error of the conversion, has a median of 2.9 mm on MUSCLESIGHT-Hand and 2.8 mm on MUSCLESIGHT-HOI, against 2.8 mm reported for the same conversion by Xi et al. (2026).

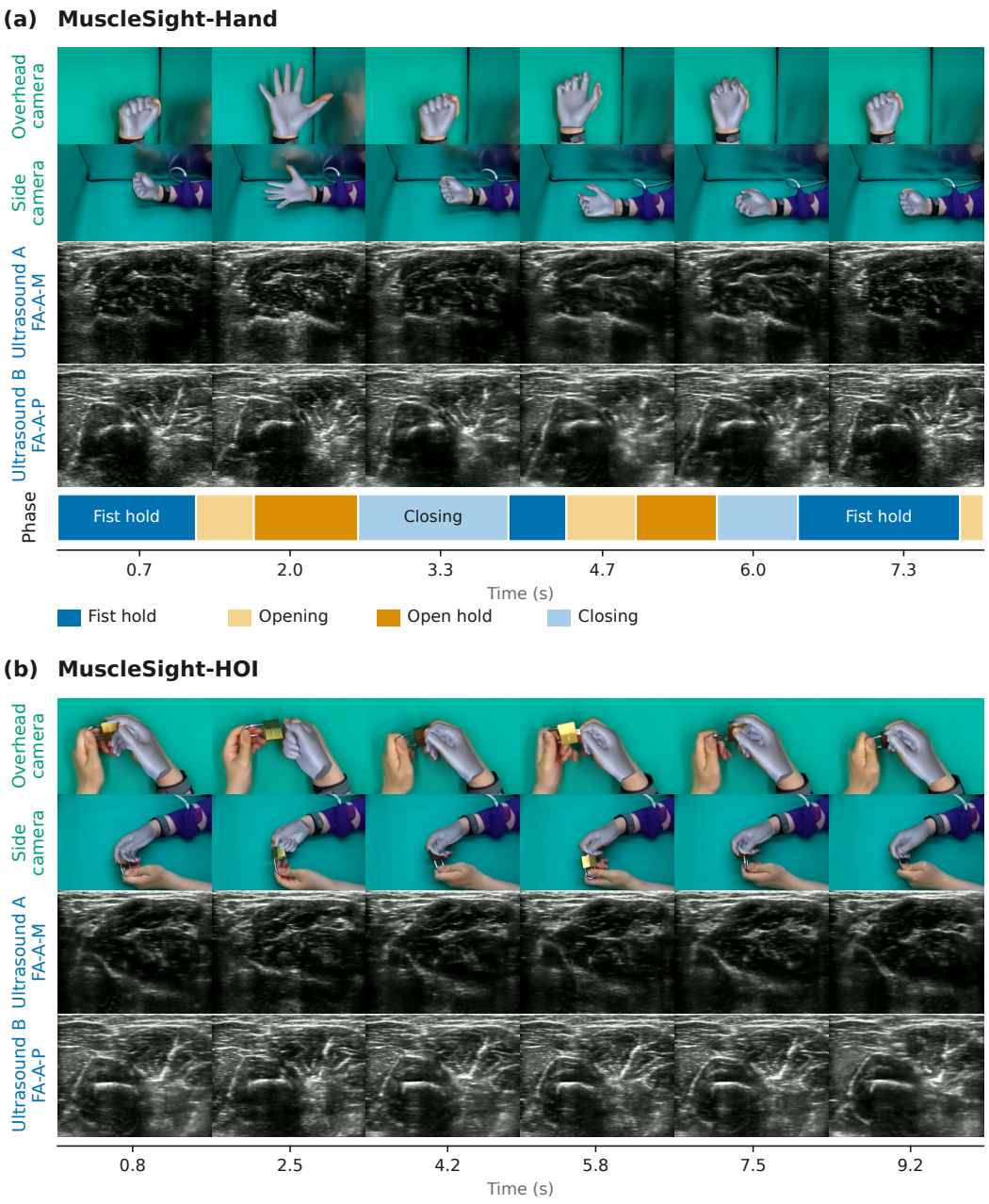


Figure 7: **Two recordings over time. (a)** MUSCLESIGHT-Hand: repeated mass finger flexion (A4), with the reviewed phase track. **(b)** MUSCLESIGHT-HOI: turning a key (C4). One participant, with probe A at `FA-A-M` and probe B at `FA-A-P`; columns are six instants of the hold window, and every stream is within 12 ms of the overhead frame.

Target. The 20 angles are four per digit: for the thumb, the bend and spread at its base and the bends at its knuckle and tip; for each finger, the spread at the knuckle and the bends at the knuckle, middle and tip joints. The bending joints have much larger ranges than the spreading joints, so a mean absolute error over the 20 angles is dominated by the bending joints. The two wrist angles of the released UmeTrack model are not recovered: the conversion works in MANO's canonical frame, which discards the orientation of the wrist, and no forearm frame is available to measure it against. They are omitted from the target, which therefore says least about tasks A2 and A6, whose movement is at the wrist.

Table 7: Movement-phase vocabulary of the bare-hand tasks. Each task has its own label set of movements and the holds at their ends; *rest* marks time outside the task. Phases are proposed from the fitted joint angles and corrected in part by a person. Inter-annotator agreement has not been measured, so a reviewed label is one reviewer's judgment rather than a consensus.

Task	Phase labels
A1	elbow flexion, elbow extension, flexion hold, extension hold, rest
A2	wrist flexion, wrist extension, flexion hold, extension hold, rest
A3	pronation, supination, pronation hold, supination hold, rest
A4	closing, hand_fist_hold, opening, hand_open_hold, rest
A5	thumb to index forming, thumb to index hold, thumb to index releasing, thumb to middle forming, thumb to middle hold, thumb to middle releasing, thumb to ring forming, thumb to ring hold, thumb to ring releasing, thumb to little forming, thumb to little hold, thumb to little releasing, open hand, rest
A6	radial deviation, ulnar deviation, radial hold, ulnar hold, rest
B1	open hand, fist forming, fist hold, fist releasing, rest
B2	fist, splay forming, splay hold, splay releasing, rest
B3	fist, index extension, index hold, index flexion, rest
B4	fist, V extension, V hold, V flexion, rest
B5	open hand, pinch forming, pinch hold, pinch releasing, rest
B6	thumb to index forming, thumb to index hold, thumb to index releasing, thumb to middle forming, thumb to middle hold, thumb to middle releasing, thumb to ring forming, thumb to ring hold, thumb to ring releasing, thumb to little forming, thumb to little hold, thumb to little releasing, open hand, rest

Human audit. For each audited recording, a person watched four panels on the same crop, namely the camera image alone, WiLoR's estimate, the estimate of an independently trained network, HaMeR (Pavlakos et al., 2024), on the same pixels, and the stored pose, and judged whether the stored pose follows the hand. On MUSCLESIGHT-Hand, 749 of the 1,492 recordings were audited; their 11,111 items (5.8%) are flagged. On MUSCLESIGHT-HOI, 722 of the 2,231 recordings were audited, mostly for a wrong pose and otherwise for the wrong hand or an unusable view; their 18,971 items (4.7%) are flagged. The two networks disagree more on MUSCLESIGHT-HOI, where the object occludes the fingers.

B.2 MOVEMENT PHASES

Each MUSCLESIGHT-Hand trial is segmented into the phases of its task (Table 7): the movements, the holds at their ends, and *rest*, which marks time outside the task, such as the countdown and the return. The proposal is computed from the fitted joint angles by task-specific rules applied to one signal per task: the position of the hand for elbow flexion (A1), its orientation for radial and ulnar deviation (A6), the contact between the thumb and each fingertip for thumb opposition and finger tapping (A5, B6), and the shape of the hand otherwise. Three rules keep the proposals consistent. A hold is named by the movement that produced it, not by the level of the signal, which drifts over a trial. On A5 and B6, the open hand at the two ends of a trial is labeled *rest*, because the hand passes through partial postures there. The threshold for thumb contact is set per finger, because the thumb comes closer to the index fingertip than to the little fingertip.

A person reviewed the proposals in an annotation tool that plays the recording against an editable phase timeline, correcting boundaries and labels or excluding the trial. 583 of the 1,440 annotated trials were reviewed. The review covers A3, A4, A5 and B6 almost completely, A1 in part and the other tasks sparsely, so the reviewed subset is not a random sample, and inter-annotator agreement has not been measured. On a task that alternates between two ends, which end is which cannot be read from the sign of the signal, so a trial can be segmented correctly and named in reverse. Direction was therefore audited by a person on A3 and A6, checked against earlier human annotations on A1 and A2, and settled for B1 to B5 by the frame audit of §B.3. 66 trials were excluded, by a reviewer or, in 4 cases, for a reversed direction. The other trials carry the automatic proposal, and every item records which of the two its label comes from. The trials of MUSCLESIGHT-HOI have no phase annotation.

B.3 Posture Labels

A held phase names a posture, and a fixed translation from task and held phase to posture class merges the same posture across tasks into 17 classes: elbow flexed and extended (A1); wrist flexed and extended (A2); forearm pronated and supinated (A3); wrist radially and ulnarly deviated (A6); fist (A4, B1 to B4); open hand (A4, A5, B1, B5, B6); splay (B2); pointing (B3); V sign (B4); thumb to the index fingertip (A5, B5, B6); and thumb to the middle, ring and little fingertips (A5, B6).

Frames are sampled at 4 Hz inside every held phase that lasts at least 0.3 s after 0.08 s is trimmed from each end, and each is paired with the nearest ultrasound frame within 50 ms, giving 27,990 frames. Of these, 7,140 show the hand resting between repetitions rather than holding the instructed posture, and a column records this role. A person then inspected every frame on a review page that shows the ultrasound frame beside the camera crop the label was read from, grouped by class and probe site, and marked 678 as wrong. The marks concentrate on the resting fist of B2 to B4, which is a different event from a deliberate fist; the role column lets a run restrict to the instructed postures. Image2gesture uses the remaining 27,312 frames.

C Benchmark Details

C.1 Items

The four tasks read frozen item tables, released with the dataset, that record for every item its clip, time, label, label source, fold under every split and review verdict. Image2gesture uses the 27,312 posture frames of §B.3. Image2pose starts from the 192,739 pairs of a fitted pose and an ultrasound frame on MuscleSight-Hand, one per fitted pose, keeping the closer pose where two claim the same ultrasound frame. The window tasks group consecutive pairs without resampling: each clip is cut into runs of poses no more than 0.15 s apart, windows of ten poses (1 s) tile each run at a stride of five, and a window whose span differs from 1 s by more than 6% is dropped. This gives 32,912 windows on MuscleSight-Hand and 73,231 on MuscleSight-HOI, of which 31,118 and 69,957 lie on recordings form the benchmark. The two clips of a dual-probe trial carry the same pose sequence with different ultrasound. The release also contains windows of 2 s, which this paper does not use. Hashing every stored image confirms that no image appears twice in any table.

C.2 Splits

Folds are assigned to whole groups, so that no group spans two folds, and the two clips of a dual-probe trial always share a fold, because they share camera frames and therefore pose labels. Groups are drawn toward 70%, 15% and 15% of the clips for training, validation and test, and stratified by the task families a participant performed, because the two protocol groups performed different families (§A.2). The held-out-session split assigns whole sessions, with 17 test sessions from 14 participants, 10 of whom also have sessions in training. The held-out-participant split assigns whole participants: 22 for training, 6 for validation and 5 for testing. The random split of the single-frame tasks assigns about 15% of the frames to the test fold at random, regardless of trial or session. The *seen* evaluation of the window tasks withholds 15% of the windows of the training sessions at random and scores them with the network trained for the held-out-session split. The release also defines splits that hold out trials and movement families, which this paper does not use.

C.3 Architectures and Training

Networks. Lite-CNN has five blocks of 3×3 convolution, batch normalization, ReLU and 2×2 max pooling, with 16, 32, 64, 128 and 128 channels, followed by dropout of 0.3 and a linear output layer; it reads 128×128 frames, and its penultimate representation has 2,048 dimensions. The frozen encoders read 224×224 frames, and a probe reads the class-token embedding of DINOv3 (768 dimensions) or the pooled image embedding of EchoCLIP (512), through either a linear layer or one hidden layer of 256 units with ReLU and dropout of 0.3. The window models apply Lite-CNN with shared weights to every frame and a linear layer to the 20 angles at every frame. Between the two sits no temporal model; two one-dimensional convolutions of width 256 and kernel size 3, the second dilated by 2, with GELU activations; or two pre-norm Transformer encoder layers of width 256 with four heads, a feed-forward width of 1,024, dropout of 0.1 and learned position

embeddings. R(2+1)D-18 is the torchvision model with Kinetics-400 weights; it reads 112×112 frames replicated to three channels and normalized with the Kinetics statistics, and one linear layer maps its pooled 512-dimensional feature to the poses of all ten frames. For pose tracking, the given pose is projected to the feature width and added to every frame's feature, and the head predicts the displacement from the given pose, so an output of zero reproduces holding the initial pose.

Optimization. All models are trained with AdamW. Lite-CNN uses a learning rate of 10^{-3}, weight decay of 10^{-4} and batches of 32; on Image2gesture its cross-entropy loss is weighted by inverse class frequency and training runs for 60 epochs without augmentation and 180 with it, and on Image2pose its L1 loss in degrees is taken on the offset from the training fold's mean pose, so that an output of zero is the constant predictor, and training runs for 60 epochs. The probes train for 200 epochs at a learning rate of 10^{-3}, with batches of 512 for Image2gesture and 256 for Image2pose; the weight decay is 10^{-4} for the pose probes and is chosen from $\{10^{-4}, 10^{-3}, 10^{-2}\}$ on validation data for the gesture probes. The window models train for 60 epochs with an L1 loss in degrees, a learning rate of 3×10^{-4}, weight decay of 10^{-4} and batches of 32, without augmentation.

Augmentation. Where a table marks it, each training frame receives, independently, a horizontal flip with probability 0.5, a crop that keeps 80 to 100% of the side and is resized back, a horizontal shift of up to 5%, a rotation of up to 5°, a rescaling by 0.9 to 1.1, a gain of 0.85 to 1.15 and an offset of up to 0.05 of the gray range; pixels moved in from outside the frame are black. Vertical flips, elastic distortion, blur, noise and cutout are excluded, because the vertical axis is depth and the speckle pattern and tissue boundaries are the signal.

Epoch selection and seeds. The gesture models are read at the epoch that maximizes the validation accuracy smoothed over five epochs, with validation data held out from the training fold, and the window models at the epoch of lowest validation error. The single-frame pose models, both Lite-CNN and the probes, are read at the end of their fixed schedule, with the validation fold joined to training. Each score is the mean over three seeds, five for the pose probes, and the reported spread is over seeds only.

Calibration and participant scaling. Calibration starts from the trained model and fine-tunes all parameters for 300 steps at a learning rate of 10^{-4} with batches of 32. Within each task, the trials of the calibration session are dealt alternately to a calibration pool and an evaluation pool; calibration items are drawn from the first and every score is computed on the second, which no calibration budget touches, while the control draws the items from the evaluation pool instead. The scaling experiment retrains each model from scratch on 2 to 28 of the training participants, 4 to 22 for Video2pose, with the same number of optimization steps, and calibrates it on half of the largest budget. Both experiments use one seed.

Representation analysis. The benchmark runs do not store weights, so each analyzed model is retrained once, on the held-out-participant split with one seed and its own recipe. For each model, 2,500 items are embedded, each feature dimension is standardized, and the fraction of variance between groups is computed, with its spread over 80 bootstrap resamples of the items and a reference from 8 shuffles of the group labels. The t-SNE views are computed on the first 50 principal components.

Hardware. Every model was trained on a single NVIDIA L40S or RTX 4090 GPU with PyTorch 2.8.

D Prior Datasets and Methods

D.1 Datasets

Table 1 in the main text compares the downloadable forearm ultrasound datasets. Table 8 compares the downloadable electromyography datasets in the same columns, and Table 9 adds how each dataset synchronizes its streams, where MuscleSight, timestamped in software, is weaker than several hardware-triggered datasets.

Table 8: Downloadable electromyography datasets of manual tasks that also record the hand, by video or kinematics, in the columns of Table 1, with our two datasets repeated for reference. *HD-sEMG*: high-density sEMG, recorded with electrode grids. Size is in hours, or in repetitions or grasps where no duration is given; for our datasets, the time inside cued trials. Four of these five have more participants than ours.

	Muscle sensing	Scale		
Resource	Modality	Subj	Size	HOI
Ninapro DB1–3 (Atzori et al., 2014)	sEMG	78	29K reps.	✓
MeganePro MDS1 (Cognolato et al., 2020)	sEMG	45	14.4K grasps	✓
emg2pose (Salter et al., 2024)	sEMG	193	370 h	✓
Reach&Grasp (Di Domenico et al., 2025)	HD-sEMG	10	1.6K reps.	✓
EgoEMG (Xi et al., 2026)	sEMG	41	10+ h	✗
MUSCLESIGHT-Hand (ours)	**B-mode**	**33**	**6.0 h**	✗
MUSCLESIGHT-HOI (ours)	**B-mode**	**33**	**11.5 h**	✓

Table 9: Synchronization and scale of related multimodal datasets. *Sync*: hardware trigger or genlock (HW) or software timestamping (SW). Rows marked † come from the biomedical literature. Entries are as stated by each source; n.s.: not stated. TACO synchronizes its cameras in hardware and its motion capture by timestamps.

Resource	Muscle sensing	Subj.	Other streams	Sync	Public
DexYCB (Chao et al., 2021)	—	10	RGB-D, obj. pose	n.s.	yes
ARCTIC (Fan et al., 2023)	—	10	multi-view, MoCap	n.s.	yes
TACO (Liu et al., 2024)	—	14	multi-view, pose	HW+SW	yes
emg2pose (Salter et al., 2024)	sEMG, 16 ch	193	hand pose	SW	yes
K2MUSE† (Li et al., 2026)	A-mode, 4 ch	42	MoCap, GRF, sEMG	HW	yes
Bimbraw (Bimbraw, 2026)	B-mode	6	finger joint angles	HW	yes
Pallarès-López† (Pallarès-López et al., 2026)	B-mode	36	MoCap, EMG	HW	yes
AnkleImage† (Zhang et al., 2024)	B-mode	10	MoCap, EMG, GRF	HW	yes
TaL (Ribeiro et al., 2021)	B-mode, tongue	82	lip video, audio	HW	yes
MUSCLESIGHT (ours)	B-mode	33	stereo RGB, hand	SW	on acceptance

On synchronization. Our streams are timestamped by Lab Streaming Layer (Kothe et al., 2025) on one shared clock, and each fitted hand pose is paired with the nearest ultrasound frame within 50 ms; a pose without such a frame is discarded rather than interpolated. This is software timestamping, looser than the hardware triggers of several datasets above. It is adequate for labels sampled at 10 Hz, and we make no claim on this axis.

D.2 METHODS

Table 10 compares the published works that map forearm ultrasound to a hand gesture or a hand pose by what they did: how each defined its prediction task, what model it fitted and how that model was trained, what of the work can be obtained today, and whether any reported number was measured on a participant the model had not seen. Scope is every such work we could find. Our own protocol is not a row: this is a survey of prior work, and adding a row we control to a grid we also chose would not be a comparison. The *Released* column was checked against the repositories and data archives themselves rather than read from the papers, because the two disagree in one place: one work announces a public repository that exists and is empty.

Held-out participants studies are limited. Two of the twelve works report a number measured on a person absent from training: in one, macro F1 ranges from 0.46 to 0.83 across twelve held-out participants, and in the other, the median R^2 rises from 0.22 to 0.52 as the number of training participants grows from one to nine. Every accuracy above 95% in the table comes from a split in which training and test data share participants, and one work states that its clips were divided at random regardless of which participant produced them. These results are correct under their own protocols. They are the reason a number from this literature and a number from a held-out-

Table 10: **Prior work mapping forearm ultrasound to hand gestures or hand pose.** *In.*: A-mode (one-dimensional echo traces) or B-mode (images). *Task*: input → output. Accuracies carry the chance level of their label set. *Released*: what can be obtained, checked against the live record on 2026-09-20 rather than read from the paper. *Held out*: whether any reported number holds out a participant the model never trained on.

Work	In.	Task	Model and training	Released	Result as reported	Held out
Bimbraw et al. (2023)	B	frame → 11 gestures; frame → 4 knuckle angles	CNN from scratch; SVC and MLP classifiers compared	data (Zenodo, CC BY 4.0; 6 subj., Vicon); no code	82.7% of 11 (chance 9.1%); 7.35° RMSE	no
Bimbraw et al. (2026a)	B	frame → 12 gestures	5-layer CNN, ViT and SVC from scratch; one model per subject	data (Zenodo, CC BY 4.0; 6 subj.); code (MIT), no trained models	per subject, 12 gestures (chance 8.3%)	no
Bimbraw et al. (2024a)	B	clip → 12 gestures	(2+1)D residual CNN from scratch	none	98.8 ± 0.9% of 12 (chance 8.3%)	no
Bimbraw et al. (2024b)	B	frame → 4 gestures	5-layer CNN from scratch, quantized for a single-board computer	none	92% of 4 (chance 25%); 0.31 s per inference	no
Rothenberg et al. (2024)	B	frame → 5 gestures	frozen CNN features; last layer fine-tuned per new session	none	about +10 points after two fine-tuning sessions (chance 20%)	no
Bimbraw et al. (2026b)	B	frame → 5 gestures	frozen vision-language model; k labeled images in the prompt	no data; the method is a prompt	0-shot 19–21% (chance 20%); 3-shot 78.7% within, 61.3% across sessions	no
Zadok et al. (2026)	B	sequence + rest sequence → 5 finger-flexion decisions	CNN and temporal transformer (10.7 M); encoder pretrained on image reconstruction, then ranking pretraining	no data; code announced, **repository empty**	macro F1 0.63; reimplemented prior methods ≤0.21	**yes**
Lee et al. (2025)	B	clip → 9 finger motions	weighted vote over five video networks (3D CNNs, MViT, Swin3D) from scratch	data on request; no code	95.6% of 9 (chance 11.1%)	no
Spacone et al. (2026)	A	sample → 23 joint angles	11 k-parameter CNN from scratch, fine-tuned per session, on the probe's microcontroller	none	6.1–6.4° MAE pooled; 9.5–11.2° across sessions; 7.1–7.5° fine-tuned	no
Grandi Sgambato et al. (2025)	A	sample + neutral reference → 4 joint angles	ConvNeXt-T pretrained on ImageNet, fine-tuned	data for 1 of 10 subj. (CC BY-NC 4.0); code, no weights	median R^2 0.22 → 0.52 as training participants go from 1 to 9	**yes**
Yang et al. (2022)	A	frame → 6 grips and wrist rotation	multi-task network; multi-output Gaussian process	data (figshare, CC BY 4.0); no code	4 transradial amputees, 3 sessions each	no
Lykourinas et al. (2024)	A	as above, on Ultra-Pro	4-block CNN with adversarial adaptation (DANN, VADA) on unlabeled new-session data	uses Ultra-Pro; no code	DANN adds 25% accuracy on average over no re-calibration (6 grips, chance 16.7%)	no

participant protocol cannot be placed in the same column, and the reason our benchmark reports within-session, held-out-session and held-out-participant results side by side (§4.2).

No work maps a sequence to a pose sequence. The three works that regress pose map a single sample or a single frame to a single pose vector. One further work reads a sequence and emits five binary per-finger decisions rather than a trajectory. Continuous pose over time, the standard formulation in the electromyography literature, has no entry on the ultrasound side, which is why our window tasks follow the formulations of emg2pose (Salter et al., 2024).

Nine of twelve train from scratch. The exceptions are an ImageNet-pretrained convolutional backbone, an image encoder pretrained on image reconstruction, and a frozen vision-language model, which is at chance without in-context examples. Frozen large-scale image encoders, which our baselines include (§4.4), are essentially unexplored on this modality.

Data availability. Three corpora can be obtained in full, two of gestures and one that pairs B-mode with continuous knuckle angles, all under permissive licenses, and one further corpus is released for a single participant of ten. Two works release code and none releases a trained model. A prior

method can therefore be compared on our tasks only by rebuilding it from its description; our Lite-CNN baseline is a small convolutional network of comparable size (§4.4).