SCALING WRIST SEMG FOR HAND POSE ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Surface electromyography (sEMG) offers a promising alternative to vision for hand pose estimation, providing information that is complementary to visual observations. However, unlike vision, sEMG-based hand pose estimation has received limited attention. In this work, we present **CycloFormer**, a rotation-invariant Transformer that encodes the cyclic symmetry of the wrist-worn sEMG electrode ring into its architecture. It achieves state-of-the-art landmark-distance performance on `emg2pose`, and exhibits systematic scaling with model size and training data. Through comparison with the leading vision baseline, we show that an sEMG-based pose-estimation pipeline becomes more accurate when visual evidence is limited. We also derive a scaling law for sEMG hand-pose regression in the data-scarce regime. It identifies the loss-minimising model size at the current data scale and suggests that large-scale data collection is a key direction for advancing sEMG-based hand pose estimation.

1 INTRODUCTION

Vision-based approaches have improved hand pose estimation substantially over the past several years. Egocentric cameras, multi-view systems, and in-the-wild images provide dense supervision of hand pose and have demonstrated strong empirical performance across datasets and benchmarks (Zimmermann et al., 2019; Moon et al., 2020; Banerjee et al., 2025; Fan et al., 2023; Ohkawa et al., 2023; Cheng et al., 2023; Romero et al., 2017; Kanazawa et al., 2018; Rong et al., 2021; Pavlakos et al., 2024; Potamias et al., 2025; Park et al., 2022; Jiang et al., 2023; Wang et al., 2024; Grauman et al., 2024; Banerjee et al., 2025; Chao et al., 2021; Hampali et al., 2020).

Vision nonetheless has structural limits. It is constrained by occlusion, poor lighting conditions, and limited field of view. During object interactions or tool use, fingers curl around objects, fingertips vanish behind surfaces, and the forces that distinguish a secure grasp from a failed one leave no visual trace. Despite occlusion-robust architectures explicitly designed to address this failure mode (Park et al., 2022; Jiang et al., 2023), contact-time hand pose remains a persistent challenge for vision-based approaches.

Wrist surface electromyography (sEMG) measures the electrical activity of the muscles that drive the fingers, rather than reflected light. Because the signal originates upstream of visible motion, it remains observable during occlusion and line-of-sight failures that limit vision. Wrist sEMG also correlates with pinch force (Sîmpetru et al., 2022), offering a promising foundation for various applications.

sEMG sensors are now available as consumer wristband hardware, and there is growing evidence that the modality generalises across users (Kaifosh et al., 2025; Salter et al., 2024). Despite these advantages, sEMG remains far less explored as a scalable modality for hand pose estimation.

Early sEMG-based pose models used convolutional or recurrent networks and were trained on comparatively small laboratory collections (Liu et al., 2021; Sîmpetru et al., 2022). `emg2pose` (Salter et al., 2024) later released recordings from 193 users and 370 hand-hours, and an LSTM baseline (*vemg2pose*). This sEMG-labelled pose set is still small in comparison to vision datasets: EgoScale (Zheng et al., 2026), for example, uses 20,854 hours of egocentric recordings. Published pose models on `emg2pose` do not use donning rotation as an exact invariance. They have also never been compared against vision baselines such as *HaMeR* (Pavlakos et al., 2024), and there has been little work on scaling law analysis of the kind Kaplan et al. (2020) and Hoffmann et al. (2022) fit on language.

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.

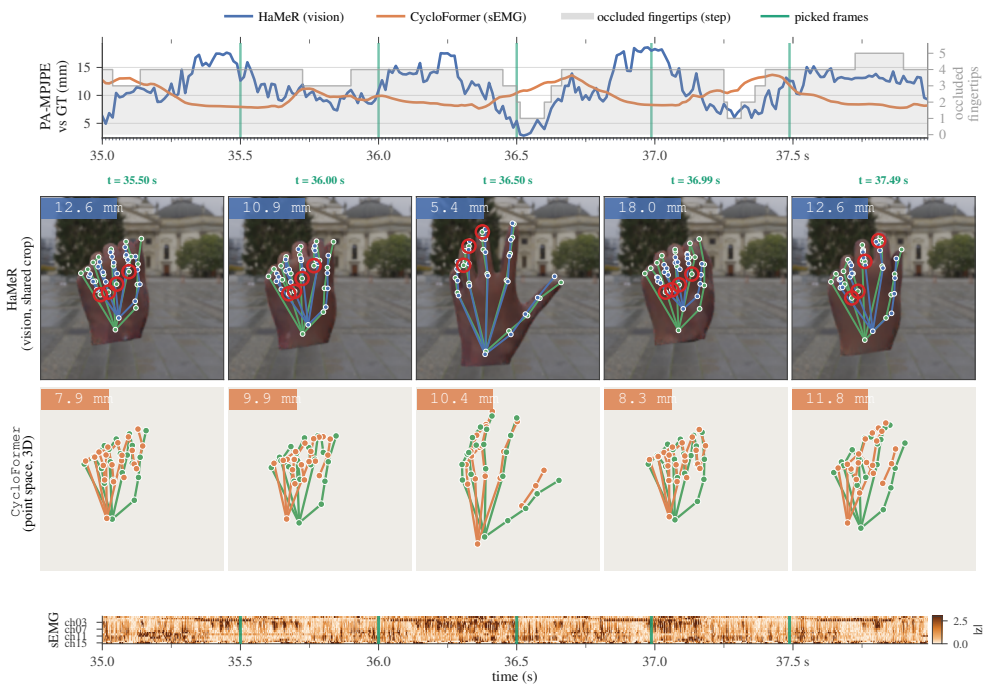

Figure 1: *HaMeR* (blue) and **CycloFormer** (orange) on a 3 s segment of HandClawGraspFlicks taken from the `emg2pose` validation split (240 frames). `emg2pose` has no camera; the *HaMeR* row is the frozen model on a render of the labelled pose. The grey step is the occluded-fingertip count; teal marks the five frames below, drawn with the ground-truth skeleton in green. Red on the *HaMeR* row marks a hidden fingertip, and the strip at the bottom is the sixteen-channel recording on the same interval.

We address these shortcomings systematically. We first develop a state-of-the-art sEMG decoder. We then use this model to compare representative sEMG and vision pose-estimation pipelines under varying levels of fingertip occlusion, thereby characterising how their relative accuracy changes across hand configurations. To our knowledge, this is the first such comparison in the literature; it is motivated by Section 3.5 of *HaMeR* (Pavlakos et al., 2024) and Section 4.2 of `emg2pose` (Salter et al., 2024). Finally, we characterise how sEMG-based pose estimation scales with model size and dataset size. We find that current sEMG benchmarks operate in a data-limited regime: increasing model width at fixed data eventually produces an error upturn that the standard additive scaling form does not capture. Our augmented scaling law improves forecasting on the 75% and large-quadrant holdouts

Our contributions are three-fold:

1. CycloFormer, a wrist decoder that is invariant to sensor rotation by construction (Section 3): The 3.53M variant achieves state-of-the-art landmark-distance accuracy on `emg2pose` across all three generalisation regimes, while also attaining the best PA-MPJPE (Table 1).

2. Comparison between sEMG and egocentric vision under varying levels of fingertip occlusion (Section 4): Using the paired dataset `EgoEMG` (Xi et al., 2026), we compare **CycloFormer** against *HaMeR* (Pavlakos et al., 2024), a state-of-the-art monocular hand-pose estimator, on synchronised sEMG and real egocentric RGB. Vision is more accurate when the hand is visible, but the relative advantage shifts toward sEMG as self-occlusion increases, with **CycloFormer** surpassing *HaMeR* when all five fingertips are occluded (Figure 3). A controlled rendered `emg2pose` analysis reproduces the same high-occlusion crossover and shows that it persists across appearance, crop, and blur (appendix D), and an occlusion-aware vision baseline (Park et al., 2022). To our knowledge, this is the first quantitative comparison of the two sensing pipelines stratified by degree of self-occlusion.

3. Scaling law for the data-scarce sEMG regime (Section 5): We observe the scaling-law form of Hoffmann et al. (2022), written for single-epoch regime, underestimates loss in the data-limited regime. We augment that form with an empirical interaction term $c(N/D)^{\gamma}$ which fits eight widths $\times$ five data fractions with R^2=0.988 against 0.906 of Hoffmann et al. (2022). The fitted data exponent from Equation (1) exceeds the parameter exponent ($\beta > \alpha$); together with the interaction penalty for increasing N/D, this indicates that improving performance in this regime is more sensitive to additional data than to additional model scale alone. To ground our scaling law, we evaluate two held-out test protocols against Chinchilla (Hoffmann et al., 2022), and beat them on two of three (Table 4).

Together, these contributions establish state-of-the-art performance on `emg2pose`, demonstrate complementary behaviour between wrist sEMG and vision on paired `EgoEMG` data (Xi et al., 2026), and show that sEMG-based hand pose estimation scales predictably with data and model size.

2 Related Work

Hand pose from images. Vision datasets for 3D hand pose span controlled multi-view capture (FreiHAND (Zimmermann et al., 2019), InterHand2.6M (Moon et al., 2020)), egocentric multi-view hand–object tracking (HOT3D (Banerjee et al., 2025)), mixed egocentric and allocentric bimanual capture (ARCTIC (Fan et al., 2023)), egocentric hand pose during object assembly (Assembly-Hands (Ohkawa et al., 2023)), and in-the-wild 2D hand understanding (Hands23 (Cheng et al., 2023)). Most current hand-pose estimation models follow HMR (Kanazawa et al., 2018) and FrankMocap (Rong et al., 2021) and regress MANO (Romero et al., 2017) pose and shape from an RGB image. For example, WiLoR (Potamias et al., 2025) localises the hand with a fully convolutional detector and then reconstructs and refines MANO with a Transformer. Our vision baseline, *HaMeR* (Pavlakos et al., 2024), takes a cropped monocular image and regresses MANO and camera parameters with a ViT-H encoder and a Transformer decoder. *HandOccNet* (Park et al., 2022) estimates the mesh when part of the hand is occluded, and Jiang et al. (2023) reconstruct the hand from one RGB image with an occlusion-aware texture. The egocentric recordings these models draw on have grown in the same way. Ego4D (Grauman et al., 2022), Ego-Exo4D (Grauman et al., 2024), and EgoDex (Hoque et al., 2026), which pairs 829 hours of video with 3D hand trajectories, are examples, and EgoScale (Zheng et al., 2026), on the other hand, uses 20,854 hours of egocentric video.

Hand pose from sEMG. Several works have studied pose regression from sEMG using datasets with relatively few participants (Liu et al., 2021; Sîmpetru et al., 2022). Liu et al. (2021) used an eight-channel Myo armband to estimate 3D finger joint angles in a 12-user study, and analyzed generalisation across users and band-remounting sessions. Sîmpetru et al. (2022) records 13 participants with a 320-electrode HD-sEMG system and predicts joint positions, joint angles, and pinch force with a 3D-convolutional network. `emg2pose` (Salter et al., 2024) later introduced a large-scale sEMG dataset containing 193 users, 751 sessions, and 370 hand-hours. Kaifosh et al. (2025) decode handwriting, gestures, and wrist velocity from wrist sEMG on held-out participants, and Mehlman et al. (2025) scale a convolutional Transformer on emg2qwerty to 109M parameters (30.5% held-out character error). emg2tendon (Verma, 2025) map sEMG to tendon control of a musculoskeletal hand to predict force. Across sEMG-to-pose models, the network is fit to the channel order in its training data, and none treats a rotation of the band as an exact invariance.

Electrode shift and ring symmetry. In sEMG, the electrode shift has mostly been studied for gesture classification, and the remedies either train over several placements or correct the shift before classification. Hargrove et al. (2006) train on patterns recorded over a range of electrode positions, Boschmann & Platzner (2014) test simulated shifts of ± 1 cm on HD-sEMG images, and Pereira et al. (2024) train on shifted eight-channel subsets of a 128-channel grid, whereas Kim et al. (2018) and Li et al. (2021) estimate the rotation of the band from a short calibration movement and correct the input before recognition. EEG and ECG face a related problem of channel identity: Saeed et al. (2021) learn to reorder shuffled or missing EEG channels into a canonical montage, DIVER-0 (Han et al., 2025) is equivariant to any permutation of EEG channels, and an interchanged ECG cable is treated as a recording error to be detected from the waveform (Batchvarov et al., 2007). Architectures such as SE(3)-Transformer (Fuchs et al., 2020) and Equiformer (Liao & Smidt, 2023) on 3D graphs, and Cyclic-Symmetry CNNs (Dieleman et al., 2016), Circular CNNs (Schubert et al., 2019), and

PanoSwin (Ling et al., 2023) on rotated or wrap-around images, show that a symmetry can be built into the network instead of learned from data. None of these applies such a constraint to a ring of sEMG electrodes, where rotating the band keeps every channel's neighbours and changes only where the channel order starts.

Topology-aware and equivariant architectures for sensor arrays. End-to-end group equivariance has been demonstrated for Transformers on point clouds (SE(3)-Transformer (Fuchs et al., 2020), Equiformer (Liao & Smidt, 2023)) and for $E(2)$ images (GE-ViT (Xu et al., 2023)), with LieTransformer (Hutchinson et al., 2021) establishing that cyclic-group (C_n) self-attention is achievable in principle. Cyclic Z_n priors at the convolutional level are Cyclic-Symmetry CNNs (Dieleman et al., 2016) and Circular CNNs (Schubert et al., 2019), with cylindrical / panoramic Transformers (PanoFormer (Shen et al., 2022), PanoSwin (Ling et al., 2023), Cylinder3D (Zhu et al., 2021)) handling wrap-around in attention windows or voxelisation; mixed cyclic-spatial $\times$ linear-temporal designs (DUN-SRE (Zhu et al., 2025), ESTAG (Wu et al., 2023)) realise the $S^1 \times \mathbb{R}$ separable structure only in CNN/GNN form. SSTAF (Muna et al., 2025) respects electrode and time axes architecturally. DIVER-0 (Han et al., 2025) achieves end-to-end *full-permutation* ($S_n \times \mathbb{R}$) equivariance for an EEG Transformer via temporal RoPE plus binary channel-attention biases.

Scaling laws. Kaplan et al. (2020) fit power laws relating language-model loss to model size, training data, and compute in the single-epoch regime, where each token is used once. Hoffmann et al. (2022) keep that regime and use a similar form to decide how model size and data should grow under a fixed compute budget. The same form has been fit for vision (Zhai et al., 2022; Alabdulmohsin et al., 2023), multimodal models (Henighan et al., 2020; Cherti et al., 2023; Aghajanyan et al., 2023), and speech (Droppo & Elibol, 2021; Cuervo & Marxer, 2024). When tokens are repeated over several epochs, Muennighoff et al. (2023) replace the parameter and data counts by saturating effective quantities N' and D'. Their Appendix F reports models above 2B parameters having higher validation loss than models between 200M and 1B. A rise at large width also appears in interpolation: Hastie et al. (2022) compute the ridgeless linear test risk as $p/n \to \gamma$ and show that it diverges at $\gamma = 1$, Advani et al. (2020) find a linear teacher–student training spike when samples and parameters are equal, and Mei & Montanari (2022); Maloney et al. (2022) find a ridgeless random-feature peak when features and samples are equal. Bahri et al. (2024) distinguish these behaviours as resolution-limited and variance-limited regimes. The laws fitted to trained networks above fall or level off as the model grows at a fixed amount of data, whereas the rise at large width is either reported without a term for it or derived for linear and random-feature models.

3 CycloFormer

We consider the problem of decoding a sequence of 16-channel sEMG into a time series of 20-DoF hand joint angles. A key artefact of this setting is that the electrode array is worn as a forearm band, and its orientation varies between sessions (Salter et al., 2024). As a result, the same underlying muscle activity can appear as a cyclic permutation of input channels. A robust decoder should therefore produce identical predictions under such permutations. We build this inductive bias directly into the model.

3.1 Methodology

Architecture. We formulate sEMG-to-pose regression as a continuous mapping $f_\theta : \boldsymbol{x} \mapsto \boldsymbol{y}$, from a raw 16-channel sEMG window $\boldsymbol{x} \in \mathbb{R}^{16\times T}$ to 20-DoF UmeTrack joint angles $\boldsymbol{y} \in \mathbb{R}^{20\times T'}$ (Han et al., 2022; Salter et al., 2024). A rotation corresponds to a cyclic permutation $\sigma_k \in \mathbb{Z}_{16}$. We seek a model that satisfies $f_\theta(\sigma_k \boldsymbol{x}) = f_\theta(\boldsymbol{x})$ for every k. **CycloFormer** achieves this $\mathbb{Z}_{16}$ invariance by construction using three stages:

1. Inputs are 5-second windows ($T = 10{,}000$ at 2 kHz). A channel-shared time-depth-separable CNN (TDS-CNN; Hannun et al., 2019; Salter et al., 2024; Chollet, 2017) downsamples this signal $40\times$ to $T' = T/40 = 250$ frames. Since identical convolutional weights process each channel independently, this commutes with σ_k.

2. A factored space–time Transformer (Vaswani et al., 2017; Bertasius et al., 2021; Arnab et al., 2021; Liu et al., 2022) mixes information across the $C = 16$ channels and the $T' = 250$ time frames

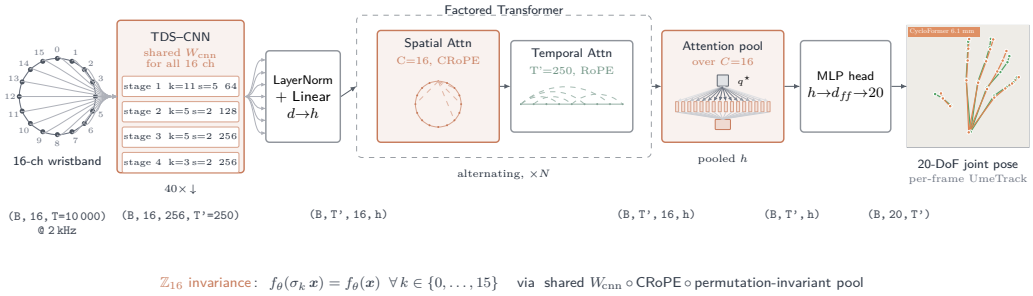

Figure 2: **Architecture** ($40\times$ downsampling, T'=250). The sEMG input is processed by a single shared TDS-CNN, projected to Transformer width, and passed through N alternating spatial/temporal attention blocks with period-16 RoPE on the channel index / linear RoPE on time. Attention pooling collapses the channel axis; an MLP head predicts joint angles.

post-CNN time steps. To maintain equivariance along the spatial axis, we employ a circular rotary position embedding (CRoPE) (Su et al., 2021); bounding attention scores strictly to the relative cyclic distance $j - i$ preserving σ_k-equivariance through the Transformer layers.

3. A learnable query $q^\star$ pools the channel axis via softmax-weighted attention (Lee et al., 2019). A weighted sum over the channel tokens, this output is invariant to permutation satisfying $\mathbb{Z}_{16}$.

Composing these equivariant blocks with the invariant pool yields rotational invariance. Under all sixteen cyclic channel shifts of the test set, the 48M **CycloFormer-**$80\times$ keeps the same error ($0°$ / 0 mm peak to peak), while vemg2pose varies by $12.2°$ / 13.9 mm peak to peak (Figure 5). Formal commutation proofs, computational complexity derivations, and the empirical rotation sweep are deferred to appendix B.1.

To ensure the gains are not artefacts of random initialisation, we evaluate seed variance for the 1.49M and 4.18M configurations; inter-seed spread is negligible relative to the performance gaps between variants. Full training details and hyperparameters are provided in appendix B.2.

3.2 Results

Table 1: Hand pose regression on the `emg2pose` test split. Mean ± standard deviation. Ang.: joint-angle MAE; LD: landmark distance; PA-MPJPE: Procrustes-aligned mean per-joint position error.

	Ang. (°)			LD (mm)			PA-MPJPE (mm)‡		
Model (params)	User	Stage	User+Stage	User	Stage	User+Stage	User	Stage	User+Stage
SensingDynamics	15.5 ± 1.4	18.8 ± 1.6	18.7 ± 1.6	21.8 ± 2.1	26.6 ± 2.0	27.2 ± 2.0	—	—	—
NeuroPose	13.2 ± 1.1	17.2 ± 1.7	17.5 ± 1.5	17.5 ± 1.3	24.0 ± 2.1	24.9 ± 1.7	—	—	—
VQ-MyoPose†	10.2 ± 1.1	21.9 ± 3.7	14.6 ± 0.7	14.7 ± 2.4	27.2 ± 2.6	18.8 ± 0.9	—	—	—
vemg2pose (6M)	12.2 ± 1.3	15.2 ± 1.6	15.8 ± 1.4	15.8 ± 1.9	20.4 ± 2.2	21.6 ± 2.0	11.4 ± 1.3	16.3 ± 1.7	17.2 ± 1.5
vemg2pose-tran (6M)	12.5 ± 1.3	16.1 ± 1.6	16.2 ± 1.3	16.3 ± 1.8	21.7 ± 2.1	22.3 ± 1.8	—	—	—
EMGFormer-S ** (3.5M)	12.5 ± 1.1	**11.1 ± 1.2**	**12.3 ± 1.1**	15.8 ± 1.6	20.6 ± 2.1	21.4 ± 1.6	11.4 ± 1.0*	16.4 ± 1.6*	17.2 ± 1.3*
CycloFormer (3.53M)	**11.5 ± 1.0**	13.7 ± 1.6	14.6 ± 1.6	**14.7 ± 1.3**	**18.5 ± 2.0**	**20.0 ± 2.2**	**10.7 ± 1.0**	**14.7 ± 1.6**	**16.1 ± 1.7**
CycloFormer (6.80M)	11.3 ± 1.1	13.1 ± 1.5	14.3 ± 1.4	14.3 ± 1.4	17.5 ± 2.0	19.5 ± 2.0	10.5 ± 1.0	14.0 ± 1.6	15.7 ± 1.5
CycloFormer (56.90M)	10.9 ± 1.3	11.7 ± 1.5	13.7 ± 1.4	13.6 ± 1.6	15.6 ± 2.0	18.4 ± 2.0	9.9 ± 1.2	12.3 ± 1.6	14.8 ± 1.5

‡ PA-MPJPE is reported so that Section 4 can share a metric with *HaMeR* (Pavlakos et al., 2024).
† VQ-MyoPose (Lovecchio et al., 2025) is trained and evaluated on a 12-subject, 27-stage subset and is not comparable on the full splits.
* PA-MPJPE from the released *EMGFormer-S* checkpoint, evaluated by us.
** Concurrent work under review; angle numbers are from the arXiv preprint.

Regression. On `emg2pose` test split (Table 1), each variant of our model outperforms *vemg2pose* (6.0M) (Salter et al., 2024) on all splits and metrics. *EMGFormer-S* (3.5M) has the lowest Stage and User+Stage joint-angle MAE, whereas the 3.53M **CycloFormer** has the lower PA-MPJPE at comparable size on all splits.

Gesture classification. We freeze **CycloFormer** after pose-regression training and train a linear head on four `emg2pose` stage groups as per CPEP (Cui et al., 2025): finger pinches, thumb rotation, thumb swipes, and claw/grasp/flick. Macro accuracy is 87.3% at 3.53M and 89.4% at 56.90M (micro 88.2% / 90.0%). CPEP reports 65.0% for *vemg2pose* (velocity prediction), 73.4% for `emg2pose` (joint angles prediction), and 78.2% on their contrastive EMG–pose model, which they particularly train for this classification (appendix B.4.1).

Table 2: **Component ablation of CycloFormer-**$80\times$. Each variant of the 4.18M model is trained from scratch with identical hyperparameters. Rows are sorted by *Stage* angular error (worst → best).

		User			*Stage*			*User+Stage*		
Variant	Params	Ang. (°)	LD (mm)	PA (mm)	Ang. (°)	LD (mm)	PA (mm)	Ang. (°)	LD (mm)	PA (mm)
w/o Spatial Attn (temporal only)	4.18M	12.9 ± 1.2	16.8 ± 1.6	12.2 ± 1.2	16.7 ± 1.6	23.0 ± 2.0	18.3 ± 1.6	17.4 ± 1.4	24.4 ± 1.7	19.7 ± 1.4
Unshared CNN (full size)	∼11M	12.0 ± 1.2	15.5 ± 1.5	11.2 ± 1.1	15.4 ± 1.6	21.0 ± 2.0	16.7 ± 1.6	15.9 ± 1.4	22.0 ± 1.8	17.7 ± 1.5
Unshared CNN (param-matched)	4.18M	12.4 ± 1.3	15.8 ± 1.6	11.4 ± 1.2	14.5 ± 1.6	19.6 ± 2.2	15.5 ± 1.7	15.8 ± 1.4	21.4 ± 1.9	17.3 ± 1.5
w/o CRoPE (no spatial PE)	4.18M	12.0 ± 1.3	15.4 ± 1.7	11.2 ± 1.2	14.3 ± 1.6	19.4 ± 2.1	15.6 ± 1.7	16.0 ± 1.5	21.9 ± 2.0	17.8 ± 1.8
Mean Pool (no Attn Pool)	4.18M	$\mathbf{11.9 \pm 1.2}$	$\mathbf{15.2 \pm 1.6}$	$\mathbf{11.0 \pm 1.2}$	13.8 ± 1.6	18.7 ± 2.1	15.0 ± 1.7	15.4 ± 1.4	21.1 ± 1.9	17.0 ± 1.7
CycloFormer (ours)	4.18M	12.0 ± 1.4	$\mathbf{15.2 \pm 1.7}$	11.1 ± 1.2	$\mathbf{13.7 \pm 1.6}$	$\mathbf{18.3 \pm 2.0}$	$\mathbf{14.6 \pm 1.6}$	$\mathbf{15.2 \pm 1.5}$	$\mathbf{20.6 \pm 2.1}$	$\mathbf{16.6 \pm 1.7}$

Ablations. We present two versions of our **CycloFormer** model with $80\times$ and $40\times$ downsampling. We have performed ablations, seed-variance analysis and scaling law analysis with $80\times$ downsampling model to save compute whereas we report the **CycloFormer**-$40\times$ model in the sota table.

We evaluate the contribution of each architectural component on the `emg2pose` benchmark. We find that spatial attention provides the largest gain, $+3.0°$ on *Stage* (Table 2). This highlights the importance of modelling cross-channel interactions. Channel-shared CNN weights provide $+1.7°$ at matched parameters, indicating that weight sharing acts as a useful inductive bias. CRoPE ($+0.6°$) and the attention pool ($+0.1°$) provide relatively smaller improvements. Notably, these components matter most in the challenging generalisation regimes (*Stage*/*User+Stage*), where the model must extrapolate to unseen gestures and configurations. In the easier *User* split, each component contributes less than $0.2°$.

We also compare **CycloFormer** against a matched TDS+Vanilla Transformer baseline trained with and without cyclic channel-rotation augmentation. We find that simple rotation augmentation does not improve performance in our setting and in some cases degrades it (Table 10). This supports our choice of encoding rotational structure in the architecture itself. For complete details on ablations, see appendix B.3.

4 sEMG AND VISION COMPARISON

Our goal is to compare strong practical pose-estimation pipelines for the two sensing modalities. For vision, we use *HaMeR*, a state-of-the-art monocular hand-pose estimator, together with *HandOccNet*, a model explicitly designed for hand occlusion; for sEMG, we use **CycloFormer**.

We compare wrist sEMG with vision and slice by occlusion. In addition, we recompute the vision baselines when information from adjacent frames is added as well as when noise is added. The datasets are `EgoEMG` (Xi et al., 2026) where both sEMG and vision modalities pre-exist, and `emg2pose` where we synthetically render the image (appendix D). All reported errors are PA-MPJPE (Zimmermann et al., 2019; Pavlakos et al., 2024; Ionescu et al., 2014) to facilitate cross-modal comparison.

Datasets. `EgoEMG` (Xi et al., 2026) records wrist sEMG (eight channels), egocentric RGB video, and motion capture for both hands at the same time. `emg2pose` (Salter et al., 2024) records wrist sEMG (sixteen channels) with UmeTrack (Han et al., 2022) pose labels but does not capture the corresponding video; we fit MANO (Romero et al., 2017) to each labelled frame and render the hand without any held object.

`EgoEMG` provides the crucial paired real RGB–sEMG setting, but its sEMG training corpus is substantially smaller than `emg2pose` (18+ versus 370 recording hours, roughly 20× less data). We therefore use EgoEMG primarily to test whether the cross-modal occlusion trend holds on real paired data, while using rendered `emg2pose` to study the same comparison in the larger-data regime where **CycloFormer** achieves substantially lower absolute pose error.

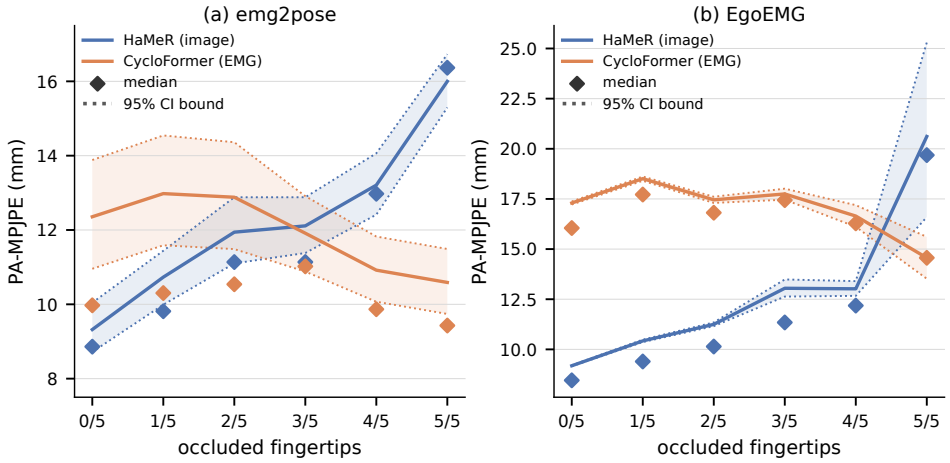

Figure 3: PA-MPJPE vs. occluded-fingertip count. (a) `emg2pose`: *HaMeR* on synthetically rendered image and **CycloFormer** on the matching sEMG. (b) `EgoEMG`: *HaMeR* on real image and **CycloFormer** on sEMG. Diamonds are medians and the bands show 95% confidence interval.

Vision baselines. We choose *HaMeR* because it provides a state-of-the-art monocular vision baseline, while *HandOccNet* provides a complementary baseline explicitly developed for occluded-hand reconstruction. Thus, the observed trend is evaluated both for a leading general hand-pose model and for an occlusion-aware model.

HaMeR's published checkpoint is applied to one frame at a time, taken from the official protocol its authors use for video (Pavlakos et al., 2024). On `EgoEMG`, *HaMeR* reads the hand crops, the left crop is mirrored as *HaMeR* only applies to right-hands, and we score MANO vertices at the 21 motion-capture marker positions; on `emg2pose` it ingests images from a rendering pipeline and is scored against the fitted MANO joints. *HandOccNet* (Park et al., 2022) designed for occluded hands is reported on both datasets. To test robustness, baselines are recomputed after the input is blurred (9 px) and after jitter is added to the crop. *HaMeR* is a per-frame estimator; as a control we freeze those predictions and smooth neighbours sampled at 2 Hz.

Results on `EgoEMG`. Given `EgoEMG` 's substantially smaller sEMG training set, we focus here on the relative behaviour of the two sensing pipelines across occlusion levels rather than absolute sEMG error. *HaMeR*'s error rises under self-occlusion, from 9.18 mm at zero occluded fingertips to 20.61 mm at five. **CycloFormer** is comparatively stable and reaches 14.56 mm in the five-fingertip bucket, where it outperforms *HaMeR* with non-overlapping confidence intervals (Figure 3). Blur and crop jitter added to the 256 pixel square images, predictably worsens across the board except for crop jitter at maximum occlusion, where it improves. Bilateral smoothing leaves the *HaMeR* mean at 9.91 mm; averaging over ± 0.5 s and ± 2.5 s raises it to 10.41 mm and 12.46 mm. Providing HaMeR with temporal context through cross-validated filtering also has negligible effect on aggregate error. *HandOccNet* performs worse than *HaMeR* across all EgoEMG occlusion buckets. **CycloFormer** training and *HaMeR* eval details provided at appendix C.4.1.

Table 3: *HaMeR* and *HandOccNet* on right-hand frames of `EgoEMG` 5 Hz frames, by occluded-fingertip count. PA-MPJPE (mm) at the 21 marker positions.

Condition	Mean	0	1	2	3	4	5
HaMeR	9.91	**9.18**	10.42	11.25	13.05	13.02	20.61
+ motion blur (9 px)	10.10	9.35	10.66	11.41	13.18	13.00	23.81
+ crop jitter (85%, 10% crop)	10.15	9.40	10.68	11.52	13.32	13.11	19.20
+ bilateral (σ_t=1.5 s, σ_p=5 mm)	9.91	9.20	**10.40**	**11.22**	**12.99**	**12.96**	20.25
+ average ($\pm$0.5 s)	10.41	9.88	10.72	11.42	13.14	13.22	17.97
+ average ($\pm$2.5 s)	12.46	12.16	12.58	13.00	14.44	14.90	17.68
HandOccNet	18.11	16.95	19.60	19.65	20.16	18.65	16.46
CycloFormer	17.67	17.30	18.51	17.45	17.75	16.65	**14.56**

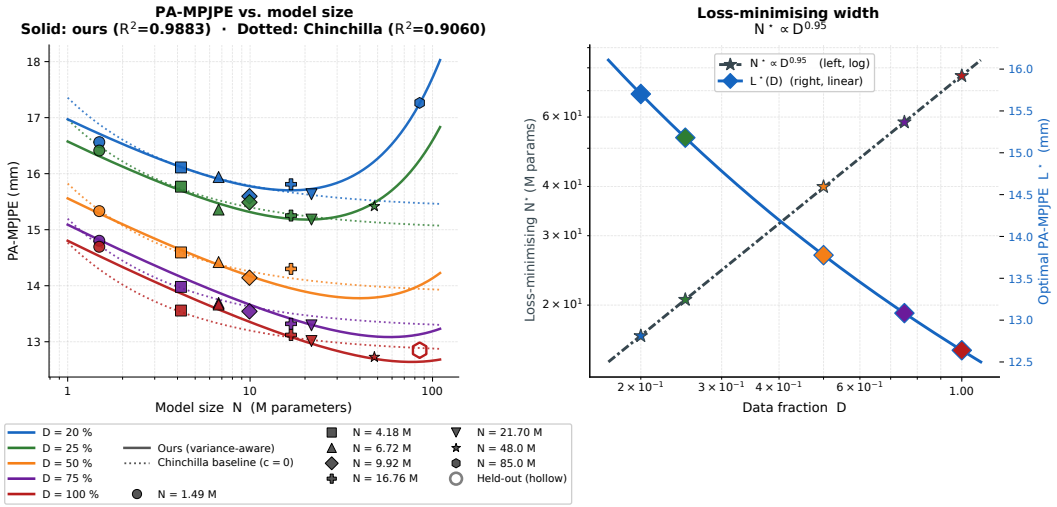

Figure 4: **Left:** test PA-MPJPE vs. model size by `emg2pose` training-data fraction. Markers are runs; solid lines are Equation (1); dotted lines are Chinchilla (c=0). The hollow marker (N=84.77M, D=100%) is held out from the fit. **Right:** loss-minimising $N^\star(D) \propto D^{0.95}$ at fixed D and the loss along that curve.

Results on rendered `emg2pose`. The controlled rendered analysis on `emg2pose` reproduces the same qualitative pattern: *HaMeR* error increases with self-occlusion, while **CycloFormer** remains comparatively robust. Session-bootstrap confidence intervals overlap for one to three occluded fingertips and separate at four and five. The high-occlusion ordering persists under changes in appearance, crop, motion blur, temporal filtering, and when *HaMeR* is replaced by *HandOccNet* (appendix D).

5 Scaling laws for emg2pose pose regression

Scaling laws tell us the amount of data and parameters needed to reach a given error/accuracy. Language and vision already have such laws (Kaplan et al., 2020; Hoffmann et al., 2022). To our knowledge, scaling behaviour has not been systematically characterised for wrist-worn sEMG hand-pose regression.

We therefore study model/data scaling on `emg2pose` and read from it the most promising scaling axis among data and parameters. The grid is eight model widths from 1.49M to 84.77M and five training fractions from 20% to 100% of the training set; we use 31 of 40 cells to fit (Figure 4; appendix E.1).

Form. Hoffmann et al. (2022) fit $L = \varepsilon_\infty + aN^{-\alpha} + bD^{-\beta}$. On this grid, the test PA-MPJPE versus the width at fixed D is U-shaped (Figure 4). We keep these three terms and add $c(N/D)^\gamma$:

$$L(N, D) \;=\; \varepsilon_\infty \;+\; a\,N^{-\alpha} \;+\; b\,D^{-\beta} \;+\; c\left(\frac{N}{D}\right)^{\gamma} \tag{1}$$

Here N is the number of parameters of **CycloFormer** (millions) and D is the fraction of the training split of `emg2pose`. Rescaling the abscissa to $u = N/D^{\gamma/(\gamma+\alpha)}$ plots data-fraction curves on a shared axis (appendix E.1).

Fit. We fit Equation (1) with a Huber-style soft-ℓ_1 loss (τ=0.05) on test PA-MPJPE, using three spatial layers and three temporal layers at varying width of **CycloFormer**. The fitted values are α=0.056, β=0.419, c=6.7×10^{-3}, γ=1.01, with R^2=0.988 and RMS residual 0.127 mm on the 31 in-sample cells; Chinchilla (c=0) as R^2=0.906 on the same cells.

Holdout validation. We evaluate whether Equation (1) predicts held-out training configurations using three multi-cell protocols (Table 4). First, we fit on all available cells at the 20% and 25% data fractions and forecast the five trained cells at 50%. Forecast RMSE is 0.291 mm for Equation (1) and

Table 4: Scaling holdouts (PA-MPJPE, mm).

Protocol	Fit	Forecast	Equation (1)	Chinchilla
D=50% column	$D \leq 25\%, N \leq 84.77\text{M}$	$D = 50\%, N \leq 84.77\text{M}$	0.291	**0.136**
D=75% column	$D \leq 50\%, N \leq 84.77\text{M}$	$D = 75\%, N \leq 84.77\text{M}$	**0.120**	0.271
Large (model, data)	$D \leq 50\%$ or $N \leq 16.7\text{M}$	$D > 50\%, N > 16.7\text{M}$	**0.110**	0.249

0.136 mm for Chinchilla (Hoffmann et al., 2022). Chinchilla has fewer fitted coefficients and forecasts more accurately in this setting, where only two data fractions are available for fitting. Second, we fit on all available cells at 20%, 25%, and 50% and forecast the six trained cells at 75%; RMSE is 0.120 mm for Equation (1) and 0.271 mm for Chinchilla. Third, we fit on cells with $D \leq 50\%$ or $N \leq 16.77\text{M}$ and forecast the four cells in the large-model, large-data quadrant: $(21.70\text{M}, 75\%)$, $(21.70\text{M}, 100\%)$, $(48.00\text{M}, 100\%)$, and $(84.77\text{M}, 100\%)$. Forecast RMSE is 0.110 mm for Equation (1) and 0.249 mm for Chinchilla. Across the three protocols, Equation (1) achieves lower forecast RMSE than Chinchilla on the 75% and large-quadrant holdouts, while Chinchilla achieves lower RMSE on the 50% holdout.

What the fit implies. On this grid the data term falls faster than the width term: doubling D shrinks $bD^{-\beta}$ by about 25% (β=0.42), and doubling N shrinks $aN^{-\alpha}$ by about 4% (α=0.06). Increasing N without more data raises $c(N/D)^{\gamma}$; so the loss-minimising width grows with the data, $N^{\star} \propto D^{0.95}$ (Figure 4, right). At the full training split, $N^{\star}$ is about 77M, the scale of the largest models we train.

6 Limitations

Modelling. We provide a $\mathbb{Z}_{16}$-invariant encoder for a closed ring of 16 identical bipolar electrodes and evaluate it on `emg2pose` pose regression. The consumer wrist sEMG is typically a circular band (Kaifosh et al., 2025), but there are other topologies such as HD-EMG grids (Rahimi et al., 2024). Our architecture will at the minimum need a matched graph or grid positional encoding. Measurements below the observed error range in Equation (1) are planning estimates.

Dataset. Force regression is untested on this pose-only, screened corpus (Salter et al., 2024). We keep this work as a potential future work plan. Privacy implications of sEMG as a biometric are in appendix F.

Scaling law. The constants of scaling laws should be considered to the *exact* values only for this family of model architectures.

7 Conclusion

We present a systematic study of wrist sEMG as a sensing modality for hand pose estimation. Using **CycloFormer**, a rotation-invariant decoder, we achieve state-of-the art performance on prior benchmarks. We compare sEMG and strong egocentric-vision pose-estimation pipelines under stratified by levels of self-occlusion. Across both rendered `emg2pose` and paired real `EgoEMG` data, vision is more accurate when the hand is clearly visible, whereas the relative advantage shifts toward sEMG under severe self-occlusion. Our observed complementarity is a conservative lower bound for real-world occlusion scenarios: the synthetic vision pipeline captures self-occlusion but not object occlusion. We further characterise the scaling behaviour of sEMG and find that performance is currently data-limited. Extrapolating these trends indicate that additional data is likely to yield larger gains.

AI use statement

In this work, we used generative AI tools for assisting with translation. We have not used generative AI tools for helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing, analysing existing literature or providing feedback on research methodology or experiments, supporting qualitative and thematic data analysis, discovering research topics or identifying gaps and brainstorming, or interpreting results, and generating synthetic datasets and cleaning and reformatting datasets are not applicable to this work.

We have used generative AI tools for writing code for creating scientific figures, debugging software code. We have reviewed all AI-assisted work. We mainly used AI for writing code for visualisation generation and for writing scripts that monitor training runs and notify us if a run died, together with the error. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Pose labels, sEMG, and (on `EgoEMG`) egocentric video used in this work come from the public `emg2pose` (Salter et al., 2024) and `EgoEMG` (Xi et al., 2026) corpora. Both dataset papers report informed consent and an IRB-approved collection protocol. We do not collect new human-subjects recordings. Wrist sEMG is a biometric signal; we discuss privacy and dual-use considerations in appendix F.

REPRODUCIBILITY STATEMENT

Architecture, training recipe, and invariance proofs are in Section 3 and appendices B.1 to B.3. The cross-modal protocol, renderer validation, and retargeter residuals are in Section 4 and appendices C.1 to C.3. Scaling-law fits, holdouts, and the master-curve derivation are in Section 5 and appendix E.1. Code and configuration files will be released with the camera-ready.

REFERENCES

Madhu S Advani, Andrew M Saxe, and Haim Sompolinsky. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446, 2020.

Armen Aghajanyan, Lili Yu, Alexis Conneau, Wei-Ning Hsu, Karen Hambardzumyan, Susan Zhang, Stephen Roller, Naman Goyal, Omer Levy, and Luke Zettlemoyer. Scaling laws for generative mixed-modal language models. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2023.

Ibrahim Alabdulmohsin, Xiaohua Zhai, Alexander Kolesnikov, and Lucas Beyer. Getting ViT in shape: Scaling laws for compute-optimal model design. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. ViViT: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences (PNAS)*, 121(27):e2311878121, 2024.

Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, Jakob Julian Engel, and Tomas Hodan. HOT3D: Hand and object tracking in 3D from egocentric multi-view videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7061–7071, 2025.

Velislav N. Batchvarov, Marek Malik, and A. John Camm. Incorrect electrode cable connection during electrocardiographic recording. *Europace*, 9(11):1081–1090, 2007. doi: 10.1093/europace/eum198.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences (PNAS)*, 116(32):15849–15854, 2019.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

Alexander Boschmann and Marco Platzner. Towards robust HD EMG pattern recognition: Reducing electrode displacement effect using structural similarity. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 4547–4550, 2014. doi: 10.1109/EMBC.2014.6944635.

Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Tianyi Cheng, Dandan Shan, Ayda Sultan Hassen, Richard E. L. Higgins, and David F. Fouhey. Towards a richer 2D understanding of hands at scale. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2829, 2023.

François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Santiago Cuervo and Ricard Marxer. Scaling properties of speech language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

Wenhui Cui, Christopher Sandino, Hadi Pouransari, Ran Liu, Juri Minxha, Ellen Zippi, Aman Verma, Anna Sedlackova, Erdrin Azemi, and Behrooz Mahasseni. CPEP: Contrastive pose-EMG pre-training enhances gesture generalization on EMG signals. In *NeurIPS 2025 Workshop on Foundation Models for the Brain and Body*, 2025. URL https://arxiv.org/abs/2509.04699.

Sander Dieleman, Jeffrey De Fauw, and Koray Kavukcuoglu. Exploiting cyclic symmetry in convolutional neural networks. In *ICML*, 2016.

Jasha Droppo and Oguz Elibol. Scaling laws for acoustic models. In *Proc. Interspeech*, pp. 2576–2580, 2021.

Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, et al. Project Aria: A new tool for egocentric multi-modal AI research. *arXiv preprint arXiv:2308.13561*, 2023.

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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Fabian B. Fuchs, Daniel E. Worrall, Volker Fischer, and Max Welling. SE(3)-transformers: 3d roto-translation equivariant attention networks. In *NeurIPS*, 2020.

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*, 16:11062, 2025. doi: 10.1038/s41467-025-66001-6.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, et al. Ego4D: Around the world in 3,000 hours of egocentric video. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, et al. Ego-Exo4D: Understanding skilled human activity from first- and third-person perspectives. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. Oral presentation.

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)*, 2020.

Danny Dongyeop Han, Ahhyun Lucy Lee, Taeyang Lee, Yonghyeon Gwon, Sebin Lee, Seongjin Lee, David Keetae Park, Shinjae Yoo, Jiook Cha, and Chun Kee Chung. DIVER-0: A fully channel equivariant EEG foundation model. *arXiv preprint arXiv:2507.14141*, 2025. ICML 2025 Workshop on GenBio.

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 *ACM SIGGRAPH Asia 2022 Conference Papers*, 2022.

Awni Hannun, Ann Lee, Qiantong Xu, and Ronan Collobert. Sequence-to-sequence speech recognition with time-depth separable convolutions. In *Proc. Interspeech*, pp. 3785–3789, 2019. doi: 10.21437/Interspeech.2019-2460.

Levi J. Hargrove, Kevin Englehart, and Bernard Hudgins. The effect of electrode displacements on pattern recognition based myoelectric control. In *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*, 2006. doi: 10.1109/IEMBS.2006.260681.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of Statistics*, 50(2):949–986, 2022.

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B. Brown, Prafulla Dhariwal, Scott Gray, Chris Hallacy, Benjamin Mann, Alec Radford, Aditya Ramesh, Nick Ryder, Daniel M. Ziegler, John Schulman, Dario Amodei, and Sam McCandlish. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020.

Jinuk Heo, Hyelim Choi, Yongseok Lee, Hyunsu Kim, Harim Ji, Hyunreal Park, Youngseon Lee, Cheongkee Jung, Hai-Nguyen Nguyen, and Dongjun Lee. Hand tracking: Survey. *International Journal of Control, Automation and Systems*, 22(6):1761–1778, 2024. doi: 10.1007/s12555-024-0298-1.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Ryan Hoque, Peide Huang, David J. Yoon, Mouli Sivapurapu, and Jian Zhang. Egodex: Learning dexterous manipulation from large-scale egocentric video. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/07fcc6e2b89439d3ee5ab60939aaa6a0-Paper-Conference.pdf.

Michael J. Hutchinson, Charline Le Lan, Sheheryar Zaidi, Emilien Dupont, Yee Whye Teh, and Hyunjik Kim. LieTransformer: Equivariant self-attention for Lie groups. In *ICML*, 2021.

Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014. doi: 10.1109/TPAMI.2013.248.

Zheheng Jiang, Hossein Rahmani, Sue Black, and Bryan M. Williams. A probabilistic attention model with occlusion-aware texture regression for 3D hand reconstruction from a single RGB image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

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, September 2025. doi: 10.1038/s41586-025-09255-w.

Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Minjae Kim, Keehoon Kim, and Wan Kyun Chung. Simple and fast compensation of sEMG interface rotation for robust hand motion recognition. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(12):2397–2406, 2018. doi: 10.1109/TNSRE.2018.2878439.

Mike Lambeta, Tingfan Wu, Ali Sengül, Victoria Rose Most, Nolan Black, Kevin Sawyer, Romeo Mercado, Haozhi Qi, Alexander Sohn, Byron Taylor, Norb Tydingco, Gregg Kammerer, Dave Stroud, Jake Khatha, Kurt Jenkins, Kyle Most, Neal Stein, Ricardo Chavira, Thomas Craven-Bartle, Eric Sanchez, Yitian Ding, Jitendra Malik, and Roberto Calandra. Digitizing touch with an artificial multimodal fingertip. *arXiv preprint arXiv:2411.02479*, 2024.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2019.

Ziyou Li, Xingang Zhao, Guangjun Liu, Bi Zhang, Daohui Zhang, and Jianda Han. Electrode shifts estimation and adaptive correction for improving robustness of sEMG-based recognition. *IEEE Journal of Biomedical and Health Informatics*, 25(4):1101–1110, 2021. doi: 10.1109/JBHI.2020.3012698.

Yi-Lun Liao and Tess Smidt. Equiformer: Equivariant graph attention transformer for 3d atomistic graphs. In *ICLR*, 2023.

Zhixin Ling, Zhen Xing, Xiangdong Zhou, Manliang Cao, and Guichun Zhou. PanoSwin: A pano-style swin transformer for panorama understanding. In *CVPR*, 2023.

Jiyao Liu, Hao Wu, Li Zhang, and Yanxi Zhao. Spatial-temporal transformers for EEG emotion recognition. In *Proceedings of the 6th International Conference on Advances in Artificial Intelligence (AAI)*, pp. 116–120. ACM, 2022. doi: 10.1145/3571560.3571577.

Yilin Liu, Shijia Zhang, and Mahanth K. Gowda. NeuroPose: 3D hand pose tracking using EMG wearables. In *Proceedings of the Web Conference 2021 (WWW '21)*, pp. 1471–1482. ACM, 2021. doi: 10.1145/3442381.3449890.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Rossana Lovecchio, Pranav Mamidanna, Bart Jansen, Dario Farina, and Tom Verstraten. VQ-MyoPose: Movement tokenization improves decoding of hand kinematics from surface EMG wristbands. In *HANDS Workshop at the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. URL `https://hands-workshop.org/files/2025/VQ-MyoPose.pdf`. Non-proceedings track.

Alexander Maloney, Daniel A Roberts, and James Sully. A solvable model of neural scaling laws. *arXiv preprint arXiv:2210.16859*, 2022.

Nicholas Mehlman, Jean-Christophe Gagnon-Audet, Michael Shvartsman, Kelvin Niu, Alexander H. Miller, and Shagun Sodhani. Scaling and distilling transformer models for sEMG. *Transactions on Machine Learning Research (TMLR)*, 2025. URL `https://openreview.net/forum?id=hFPWThwUiZ`. arXiv:2507.22094.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75 (4):667–766, 2022.

Meta. Meta ray-ban display: AI glasses with an EMG wristband. Meta Newsroom, September 2025. URL https://about.fb.com/news/2025/09/meta-ray-ban-display-ai-glasses-emg-wristband/. Announced 17 September 2025.

Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. InterHand2.6M: A dataset and baseline for 3D interacting hand pose estimation from a single RGB image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. Scaling data-constrained language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Ummay Maria Muna, Md. Mehedi Hasan Shawon, Md Jobayer, Sumaiya Akter, and Saifur Rahman Sabuj. SSTAF: Spatial-spectral-temporal attention fusion transformer for motor imagery classification. *IEEE Access*, 13:158855–158869, 2025. doi: 10.1109/ACCESS.2025.3605265.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. In *International Conference on Learning Representations (ICLR)*, 2020.

Takehiko Ohkawa, Kun He, Fadime Sener, Tomas Hodan, Luan Tran, and Cem Keskin. AssemblyHands: Towards egocentric activity understanding via 3d hand pose estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

JoonKyu Park, Yeonguk Oh, Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. HandOccNet: Occlusion-robust 3D hand mesh estimation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

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

João Pereira, Dimitrios Halatsis, Bálint K. Hodossy, and Dario Farina. Tackling electrode shift in gesture recognition with HD-EMG electrode subsets. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1786–1790, 2024. doi: 10.1109/ICASSP48485.2024.10448329.

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.

Fernando Quivira, Toshiaki Koike-Akino, Ye Wang, and Deniz Erdogmus. Translating sEMG signals to continuous hand poses using recurrent neural networks. In *IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pp. 166–169, 2018. doi: 10.1109/BHI.2018.8333395.

Farnaz Rahimi, Mohammad Ali Badamchizadeh, Raul C. Sîmpetru, Sehraneh Ghaemi, Bjoern M. Eskofier, and Alessandro Del Vecchio. Simultaneous control of human hand joint positions and grip force via HD-EMG and deep learning. *arXiv preprint arXiv:2410.23986*, 2024.

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), 2017.

Yu Rong, Takaaki Shiratori, and Hanbyul Joo. FrankMocap: A monocular 3D whole-body pose estimation system via regression and integration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2021.

Aaqib Saeed, David Grangier, Olivier Pietquin, and Neil Zeghidour. Learning from heterogeneous EEG signals with differentiable channel reordering. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021. doi: 10.1109/ICASSP39728.2021.9413712.

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 37 (NeurIPS 2024) Datasets and Benchmarks Track*, 2024.

Stefan Schubert, Peer Neubert, Johannes Pöschmann, and Peter Protzel. Circular convolutional neural networks for panoramic images and laser data. In *IEEE Intelligent Vehicles Symposium*, 2019.

Zhijie Shen, Chunyu Lin, Kang Liao, Lang Nie, Zishuo Zheng, and Yao Zhao. PanoFormer: Panorama transformer for indoor 360° depth estimation. In *ECCV*, 2022.

Raul C. Sîmpetru, Andreas Arkudas, Dominik I. Braun, Marius Osswald, Daniela Souza de Oliveira, Bjoern Eskofier, Thomas M. Kinfe, and Alessandro Del Vecchio. Sensing the full dynamics of the human hand with a neural interface and deep learning. *bioRxiv*, 2022. doi: 10.1101/2022.07.29.502064.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021.

Arvin Tashakori, Zenan Jiang, Amir Servati, Saeid Soltanian, Harishkumar Narayana, Katherine Le, Caroline Nakayama, Chieh-ling Yang, Z. Jane Wang, Janice J. Eng, and Peyman Servati. Capturing complex hand movements and object interactions using machine learning-powered stretchable smart textile gloves. *Nature Machine Intelligence*, 6:106–118, 2024. doi: 10.1038/s42256-023-00780-9.

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

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

Chen Wang, Haochen Shi, Weizhuo Wang, Ruohan Zhang, Li Fei-Fei, and C. Karen Liu. DexCap: Scalable and portable mocap data collection system for dexterous manipulation. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024. arXiv:2403.07788.

Liming Wu, Zhichao Hou, Jirui Yuan, Yu Rong, and Wenbing Huang. Equivariant spatio-temporal attentive graph networks to simulate physical dynamics. In *NeurIPS*, 2023.

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, 2026. URL `https://arxiv.org/abs/2605.05712`.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. On layer normalization in the transformer architecture. In *ICML*, 2020.

Renjun Xu, Kaifan Yang, Ke Liu, and Fengxiang He. $E(2)$-equivariant vision transformer. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 216 of *Proceedings of Machine Learning Research*, pp. 2356–2366, 2023.

Peyton R. Young, Kihun Hong, Eden J. Winslow, Giancarlo K. Sagastume, Marcus A. Battraw, Richard S. Whittle, and Jonathon S. Schofield. The effects of limb position and grasped load on hand gesture classification using electromyography, force myography, and their combination. *PLOS ONE*, 20(4):e0321319, 2025. doi: 10.1371/journal.pone.0321319.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12104–12113, 2022.

Ruijie Zheng, Dantong Niu, Yuqi Xie, Jing Wang, Mengda Xu, Yunfan Jiang, Fernando Castañeda, Fengyuan Hu, You Liang Tan, Letian Fu, Trevor Darrell, Furong Huang, Yuke Zhu, Danfei Xu, and Linxi Fan. EgoScale: Scaling dexterous manipulation with diverse egocentric human data, 2026. URL https://arxiv.org/abs/2602.16710.

Xinge Zhu, Hui Zhou, Tai Wang, Fangzhou Hong, Yuexin Ma, Wei Li, Hongsheng Li, and Dahua Lin. Cylindrical and asymmetrical 3d convolution networks for LiDAR segmentation. In *CVPR*, 2021.

Yuliang Zhu, Jing Cheng, Qi Xie, Zhuo-Xu Cui, Qingyong Zhu, Yuanyuan Liu, Xin Liu, Jianfeng Ren, Chengbo Wang, and Dong Liang. Dun-sre: Deep unrolling network with spatiotemporal rotation equivariance for dynamic mri reconstruction. *IEEE Journal of Selected Topics in Signal Processing*, 2025.

Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas 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 (ICCV)*, 2019.

A Other Related Works

Other sensors for continuous hand kinematics. Wrist-worn cameras and IMUs are recorded together with hand pose in HOT3D (Banerjee et al., 2025) and on the Project Aria glasses (Engel et al., 2023). The Meta Neural Band places sEMG and an IMU on the same wrist (Meta, 2025). Heo et al. (2024) survey sensors worn off the wrist. Textile-plus-IMU gloves report 1.2–1.5° joint-angle RMSE against optical motion capture (Tashakori et al., 2024). Sonomyography recovers wrist and hand kinematics from ultrasound (Grandi Sgambato et al., 2025). Force myography classifies gestures under changes of limb pose and load (Young et al., 2025). Wrist-worn radar and electrical-impedance sensors have been used for discrete gesture classification, and fingertip sensors such as Digit 360 record local contact geometry (Lambeta et al., 2024).

Continuous joint-angle regression from sEMG. Quivira et al. (2018) record an eight-channel Myo armband and Leap Motion kinematics from seven users and regress continuous hand pose with an LSTM. `emg2pose` defines two tasks on the same UmeTrack labels (Salter et al., 2024). Regression maps a window of sEMG to the joint-angle sequence, with no initial pose provided. Tracking predicts angular velocities and integrates them, and is given the first pose. NeuroPose predicts the angles with a convolutional encoder–decoder (Liu et al., 2021). *vemg2pose* predicts velocities with a TDS featurizer and an LSTM, then integrates (Salter et al., 2024). Table 1 reports the regression task.

B CycloFormer

B.1 Architecture Details

Channel-shared TDS-CNN backbone. The 16 channels are processed independently by an identical Time-Depth-Separable (TDS) convolutional stack (Hannun et al., 2019; Chollet, 2017). Each of the four stages applies a 1-D convolution of stride $s \in \{5, 2, 4, 2\}$ and kernel size $k \in \{11, 5, 5, 3\}$, expanding the channel dimension through $[32, 64, 128, 128]$, followed by two TDS residual blocks (a depth-wise temporal convolution and a 1×1 point-wise convolution with GELU). Cumulative downsampling is $5 \times 2 \times 4 \times 2 = 80\times$, producing $T' = 125$ tokens per channel from the input of 10,000 frames, each of dimension $d_{\text{cnn}} = 128$. Critically, the exact same parameters W_{cnn} are applied to every channel. This is implemented efficiently as a depth-wise group convolution where the group size equals the 16 input channels. This is the original stem used for component ablations, the 48M rotation sweep, and the scaling grid. The 3.53M, 6.80M, and 56.90M models in Table 1 use $40\times$ downsampling (T'=250), kernels $[11, 5, 5, 3]$, strides $[5, 2, 2, 2]$, and $(N_{\text{spat}}, N_{\text{temp}}) = (3, 3)$. CNN widths and Transformer size are in Table 6.

Linear projection. Following a LayerNorm (Xiong et al., 2020) over the feature axis, a single shared linear map $\mathbb{R}^{d_{\text{cnn}}} \to \mathbb{R}^{d_{\text{model}}}$ lifts every (t, c) token to the Transformer width. Because this projection is identical across all channels and time steps, no per-channel embedding exists anywhere in the network.

Factored spatial–temporal Transformer. The model alternates spatial blocks (attending across the 16 channels at each time step) and temporal blocks (attending across T' time steps within each channel) (Liu et al., 2022), repeated N_{spat} and N_{temp} times. Each block is a pre-norm residual unit with explicit projections W_q, W_k, W_v, W_o, scaled-dot-product attention, and a 2-layer GELU FFN of width d_{ff}. Residuals and dropout ($p = 0.15$) follow standard pre-norm Transformer practice. We deliberately do *not* use PyTorch's native `nn.MultiheadAttention`; its fused in-place projection wrapper obscures the order of operations and silently breaks the CRoPE relative-position guarantee.

Memory efficiency. A standard full attention mechanism over the flattened sequence would require computing an attention matrix of size $(CT') \times (CT')$, yielding $\mathcal{O}((CT')^2)$ complexity. By strictly factoring the attention into sequential spatial and temporal blocks (Bertasius et al., 2021), the complexity is drastically reduced to $\mathcal{O}(C^2T' + T'^2C)$. Given $T' = 125$ and $C = 16$ on this stem, this factorisation is critical for maintaining a low memory footprint suitable for wearable constraints.

Circular Rotary Position Embedding (CRoPE). Spatial blocks use a circular variant of RoPE (Su et al., 2021) parameterised by 16 cyclic positions. For head dimension d_h, half the channels carry sine–cosine pairs at frequencies $\omega_p = 2\pi p/16$, $p = 1, \ldots, d_h/2$. The rotation matrix R_i at channel position i is applied to the projected query and key:

$$q_i = R_i(W_q x_i), \qquad k_j = R_j(W_k x_j), \qquad \text{score}(i, j) = (W_q x_i)^\top R_{j-i}(W_k x_j).$$

Application order is mathematically rigid: $R_i^\top R_j = R_{j-i}$ holds *only* when no linear transformation sits between the two rotations. Rotating before W_q, W_k would inject an absolute-position-dependent term $W_q^\top W_k$ into the score, destroying the symmetry. Temporal blocks utilise standard linear RoPE (Su et al., 2021) with the identical apply-after-projection convention.

Permutation-invariant attention pooling. After the final block, the $(B, T', 16, d_{\text{model}})$ feature tensor is collapsed along the channel axis via a learned-query attention pool (Lee et al., 2019). A single learnable token $\boldsymbol{q} \in \mathbb{R}^{d_{\text{model}}}$ attends over the 16 channel tokens using multi-head attention, returning one d_{model}-dimensional vector per time step. Because this pool is a sum weighted by softmax attention, it is inherently invariant to any permutation of the channel axis, and therefore explicitly invariant to the $\mathbb{Z}_{16}$ cyclic shifts. A LayerNorm is applied to the final pooled output.

Output head. A 2-layer MLP (LayerNorm $\to$ Linear$(d_{\text{model}}, d_{\text{ff}})$ $\to$ GELU $\to$ Dropout $\to$ Linear$(d_{\text{ff}}, 20)$) maps each pooled time-step embedding to 20 joint-angle predictions, matching the UmeTrack DoF.

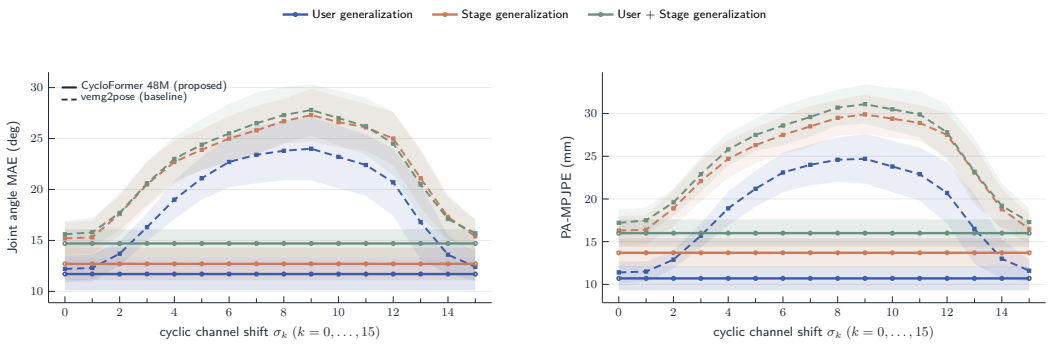

Figure 5: $\mathbb{Z}_{16}$ **invariance.** Test-set error as a function of cyclic channel shift k for **CycloFormer** (48M) and *vemg2pose* (6M). **CycloFormer**, by geometric construction, remains perfectly flat across all 16 offsets with variance $< 10^{-5}$. Conversely, *vemg2pose* degrades by $\geq 10°$ at $k = 8$, the offset maximally distant from the training distribution.

B.1.1 Formal $\mathbb{Z}_{16}$ Commutation

Let σ_k denote a cyclic shift by k positions along the channel axis, where $k \in \mathbb{Z}_{16}$. The shared CNN, linear projections, feed-forward networks, and temporal RoPE operate identically across all channels. Because they contain no channel-specific parameters, they are strictly *equivariant* to σ_k (i.e., they commute with the shift operator). Similarly, the CRoPE attention mechanism depends exclusively on the cyclic offset $(j - i) \bmod 16$, which is preserved under cyclic shifts, making the spatial attention layers equivariant as well.

Finally, the global attention pool computes a permutation-invariant weighted sum. Composing these equivariant feature extractors with a strictly invariant pooling layer mathematically guarantees that the entire network f_θ is invariant to cyclic shifts: $f_\theta(\sigma_k \boldsymbol{x}) = f_\theta(\boldsymbol{x})$. Figure 5 is that map on $k \in \{0, \ldots, 15\}$.

On `emg2pose`, donning is take-off and put-on of the band between sessions (Salter et al., 2024). Their placement augmentation is a channel shift in $\{-1, 0, +1\}$. The corpus does not label longitudinal slip of the band along the forearm. ThumbRotations in the stage list is a thumb gesture.

B.1.2 Model Configurations

We sweep **CycloFormer** across eight parameter scales to establish the scaling laws presented in Section 5. All configurations share an identical structural recipe: CNN channels $[32, 64, 128, 128]$ with strides $[5, 2, 4, 2]$, $(N_{\text{spat}}, N_{\text{temp}}) = (3, 3)$, dropout 0.15, and attention pooling. Only the Transformer width, FFN dimension, and attention head count are scaled (Table 5).

Table 5: **CycloFormer** architectural configurations for the scaling-law analysis (original $80\times$ stem). † 4.18M and ‡ 48M are the ablation and rotation-sweep widths. The SoTA models in Table 1 (3.53M / 6.80M / 56.90M) use a $40\times$ stem and a wider CNN.

Variant	Params	d_{model}	d_{ff}	Heads	$d_{\text{ff}}/d_{\text{model}}$
smallest	1.49M	128	256	4	2.00
smaller†	4.18M	256	512	8	2.00
small	6.72M	320	768	10	2.40
large	9.92M	384	1024	12	2.67
larger	16.77M	448	1792	14	4.00
largest	21.70M	512	2048	16	4.00
anchor‡	47.99M	768	3072	24	4.00
huge	84.77M	1024	4096	32	4.00

Table 6: **SoTA** $40\times$ **widths** (Table 1). All three use $(N_{\text{spat}}, N_{\text{temp}}) = (3, 3)$, kernels $[11, 5, 5, 3]$, strides $[5, 2, 2, 2]$, dropout 0.15, and attention pooling. CNN stage widths and Transformer size vary.

Params	CNN channels	d_{model}	d_{ff}	Heads
3.53M	$[32, 64, 128, 128]$	224	512	8
6.80M	$[64, 128, 256, 256]$	256	768	8
56.90M	$[128, 256, 512, 512]$	768	3072	24

B.2 Training Details

This subsection is the original $80\times$ family: component ablations, the 48M rotation sweep, and the scaling grid. The 3.53M, 6.80M, and 56.90M models in Table 1 use the $40\times$ recipe in Appendix B.2.1.

Loss. We train using L1 supervision on joint angles, supplemented by a forward-kinematic 3D-landmark auxiliary term to enforce physical consistency:

$$\mathcal{L} = \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_1 + 0.1 \cdot \|\mathrm{FK}(\hat{\boldsymbol{\theta}}) - \mathrm{FK}(\boldsymbol{\theta})\|_1,$$

where $\boldsymbol{\theta} \in \mathbb{R}^{20\times T'}$ represents the ground-truth joint angles and $\mathrm{FK}(\cdot)$ denotes the 3D landmark positions derived via UmeTrack forward kinematics. This configuration exactly matches the *vemg2pose* training protocol (Salter et al., 2024). All models process an input window of 10,000 samples (5 s at 2 kHz).

Optimisation. We utilise AdamW (Loshchilov & Hutter, 2019) with the following hyperparameters: $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, and weight decay 10^{-4}. We apply gradient clipping at $\|g\|_2 \leq 1.0$. The learning rate is cosine annealing ($T_{\max} = 100$) with peak 3×10^{-4} and $\eta_{\min} = 10^{-6}$, with no warmup. Stochastic Weight Averaging (SWA) starts at epoch 85 at 10^{-7}. Automatic mixed precision (`bf16`) is used throughout.

Schedule and hardware. We train eight model variants spanning 1.49M to 84.77M parameters for 100 epochs each. The effective batch size is scaled with model capacity (Table 7) to maintain a stable gradient signal-to-noise ratio during the scaling sweep. All experiments are conducted on NVIDIA A100 (40 GB) and V100 (32 GB) GPUs. Total compute for the scaling sweep on the 100% data split is $\sim$590 A100-hours. The best checkpoint for each model is selected by minimising the validation loss (`val_loss`). The test split is strictly isolated from all training and model-selection procedures.

Table 7: **Batch size scaling.** Per-GPU batch sizes across the **CycloFormer** scaling sweep.

Params (M)	1.49	4.2	6.7	9.9	16.77	21.7	48.0	84.77
Per-GPU batch	64	64	64	128	192	256	512	768

Table 8: **Learning-rate warmup.** Cosine over 100 epochs in every case. Only the 56.90M SoTA run uses a warmup (five epochs from 1% of peak) and a 15-epoch augmentation ramp.

Family	Widths	Warmup
$80\times$ scaling and ablations	1.49M–84.77M	none
$40\times$ Table 1	3.53M, 6.80M	none
$40\times$ Table 1	56.90M	5 epochs from 1% peak

Reproducibility. Our software stack includes PyTorch 2.3, PyTorch Lightning 2.2, and Hydra 1.3. We use a fixed global seed of 42 for all reported metrics; see Appendix B.2.2 for the corresponding seed-variance characterisation. The full inference command, checkpoint, Hydra overrides, and network configuration files are provided in the supplementary material.

B.2.1 SoTA 40× models

The 3.53M, 6.80M, and 56.90M models in Table 1 share 40× downsampling (T=10,000 → T'=250), kernels $[11, 5, 5, 3]$, strides $[5, 2, 2, 2]$, two residual blocks per stage, $(N_{\text{spat}}, N_{\text{temp}}) = (3, 3)$, dropout 0.15, Circular RoPE on the channel index, and a learned-query attention pool to 20 UmeTrack angles. CNN channels and Transformer width are in Table 6.

Training reads native `emg2pose` HDF5 through `WindowedEmgDataModule` (window 10,000, padding $[0, 0]$, stride unset, `skip_ik_failures` true, 10 windows per session). The optimiser is AdamW at 3×10^{-4} with weight decay 10^{-4} and cosine annealing over 100 epochs, $\eta_{\min} = 5\times10^{-6}$. The 3.53M and 6.80M models have no warmup. The 56.90M model warms up for five epochs from 1% of the peak LR, then cosine; input augmentations ramp from zero over the first 15 epochs (Table 8). The effective batch is scaled from 192 to 768 with width. Training is 100 epochs, gradient clip 1.0, full precision (no `bf16`), and SWA from epoch 85 at 10^{-7}. The loss is angle ℓ_1 with weight 1.0 plus a fingertip-distance term with weight 0.01. In-network augmentations are channel dropout 0.2, mix 0.1, amplitude jitter 0.08, and temporal-chunk mask 0.05; the datamodule applies `ExtractToTensor`. There is no rotation augmentation.

B.2.2 Seed-Variance Robustness

To check that the reported gains are not a random-init artefact, we re-train two **CycloFormer** configurations (smallest, 1.49M; smaller, 4.18M) at the 20% data split with four seeds each. Inter-seed coefficient of variation stays below 1% on both validation loss and validation MAE for both configurations (Table 9), a one-way ANOVA on epoch-wise val loss across the four seeds returns $F = 0.52$, $p = 0.67$, and inter-seed spread (Figure 6) is small relative to the gaps between architectural variants reported in the ablation table.

Table 9: Seed variance for **CycloFormer** at the 20% data split, across two model sizes. CoV $< 1\%$ on val loss and val MAE for both models; one-way ANOVA on epoch-wise val loss across the four seeds gives $F = 0.52$, $p = 0.67$, so the reported gains are not an artefact of random initialisation. Bold marks the best seed in each block.

Seed	Best val loss	Best val MAE (°)	Best epoch	Δ loss vs. mean	Δ MAE vs. mean (°)
Smallest (1.49M params)					
42	0.549441	13.8603	55	+0.001329	+0.0348
52	**0.546681**	**13.8526**	66	−0.001431	+0.0270
62	0.547992	13.7937	55	−0.000120	−0.0318
72	0.548335	13.7956	36	+0.000223	−0.0300
Mean	0.548113	13.8256	—	—	—
Std dev	0.001137	0.0358	—	—	—
CoV	0.207%	0.259%	—	—	—
Smaller (4.18M params)					
42	**0.534385**	**13.5792**	43	−0.003821	−0.0728
52	0.537069	13.6055	64	−0.001138	−0.0464
62	0.541534	13.7104	53	+0.003327	+0.0584
72	0.539838	13.7127	63	+0.001631	+0.0608
Mean	0.538206	13.6519	—	—	—
Std dev	0.003143	0.0697	—	—	—
CoV	0.584%	0.510%	—	—	—

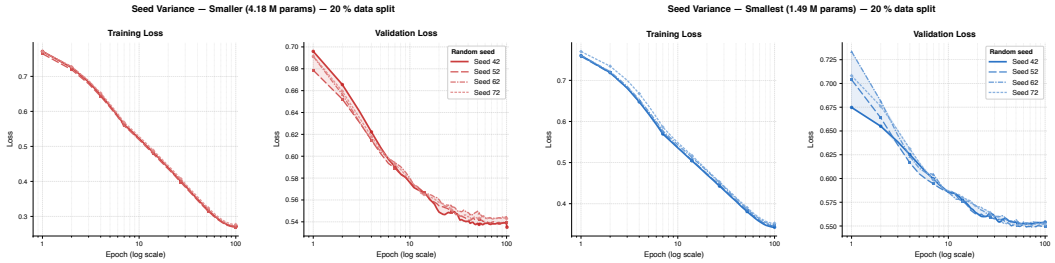

Figure 6: Seed variance for the smaller (left) and smallest (right) model configurations. Inter-seed spread is small relative to the differences between architectural variants.

B.3 Ablations

Spatial attention contributes the most to performance ($+3.0^\circ$ on *Stage*; Table 2), followed by channel-shared CNN weights ($+1.7^\circ$), CRoPE ($+0.6^\circ$), and the attention pool ($+0.1^\circ$). By testing an unshared CNN variant matched to the exact parameter count of our model, we rule out a pure capacity explanation. Weight sharing acts as a fundamental inductive prior rather than a simple parameter restrictor. Furthermore, these architectural priors matter most on the harder *Stage* and *User+Stage* splits, proving they are critical for generalising to gesture kinematics unseen during training.

What each row tests. Each row of Table 2 replaces one architectural piece of the 4M **CycloFormer**. All models are trained from scratch with identical optimisers, schedules, window lengths, batch sizes, and seeds.

w/o Spatial Attention (temporal only). Spatial attention blocks are removed, leaving only the temporal Transformer over per-channel features and the channel pool. Without cross-channel mixing to explicitly model the spatial relationships of the sensor array (Liu et al., 2022), the model loses 3.0° on *Stage* and 2.2° on *User+Stage*. This is the largest single contribution in the table and the only ablation that approaches the performance gap to *vemg2pose* (Salter et al., 2024).

Unshared CNN (full size). The channel-shared TDS-CNN is replaced with 16 independent per-channel CNNs at the same per-channel width. This raises the CNN parameter count by roughly $16\times$ to nearly 11M. The model still loses 1.7° on *Stage* despite carrying $2.5\times$ more total parameters, proving that per-channel specialisation cannot recover what the shared inductive prior provides.

Unshared CNN (param-matched). The unshared variant is scaled down so the 16 per-channel CNNs sum to the exact footprint of the shared baseline (4.18M). The shared model outperforms this variant by 0.8° on *Stage* at equal compute. This isolates the weight-sharing inductive bias entirely from any capacity effects.

w/o CRoPE (no spatial position encoding). Replacing the Circular Rotary Position Embedding (Su et al., 2021) with no positional encoding forces the spatial attention to view an unordered set of 16 channel tokens. This drops performance by 0.6° on *Stage*. Because the attention pool remains permutation-invariant, the whole network retains $\mathbb{Z}_{16}$-invariance. Therefore, CRoPE's contribution is not the symmetry itself, but the relative-distance bias that allows spatial attention to exploit the physical ring topology of the band.

Mean pool (no attention pool). The learned-query softmax pool (Lee et al., 2019) is replaced with a uniform mean over the 16 channel tokens. Both operations are permutation-invariant. However, the attention pool provides a 0.1° gain on *Stage* by allowing the model to dynamically up-weight channels carrying the strongest signal at any given time step.

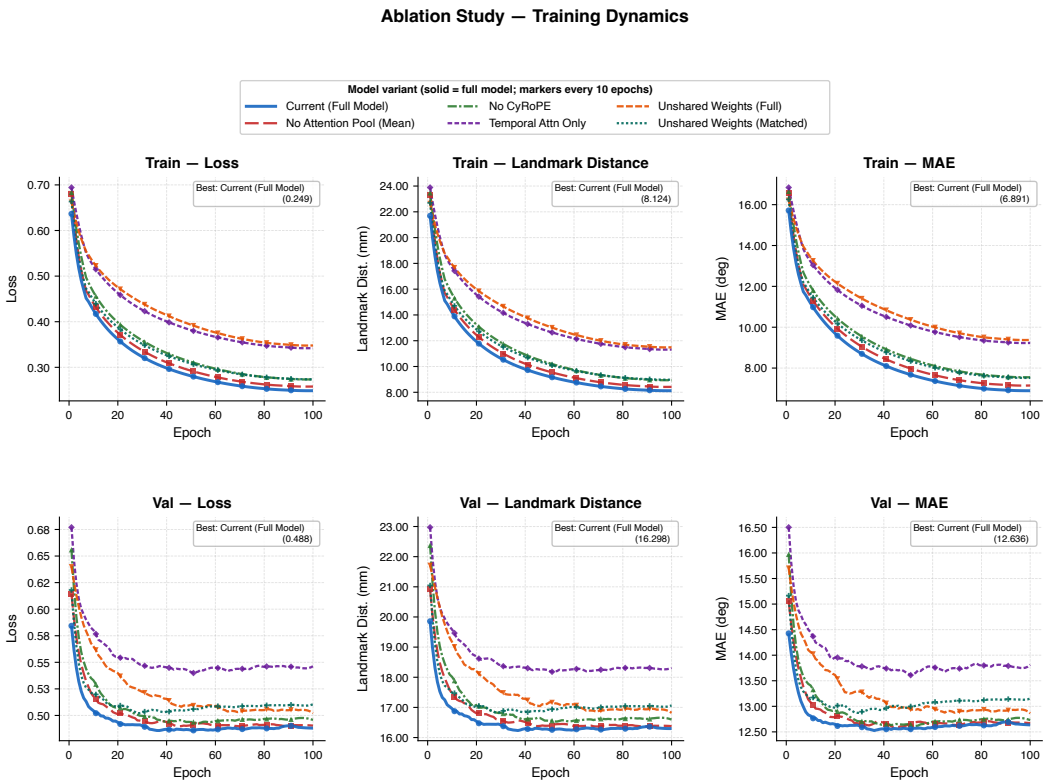

Figure 7: **Training dynamics of ablation variants.** The convergence curves corroborate the final metrics in Table 2. Removing spatial attention entirely yields the largest performance degradation. Unsharing the CNN weights causes the second largest drop. CRoPE provides a smaller but consistent refinement over an unordered spatial set.

Geometric prior vs. rotation augmentation. A non-equivariant network can approximate $\mathbb{Z}_{16}$ invariance through cyclic channel-shift data augmentation. However, this forces the model to expend parameter capacity learning an auxiliary symmetry rather than the core task (Bronstein et al., 2021). In contrast, **CycloFormer** encodes this invariance architecturally and exactly. To isolate these two approaches, we train the TDS+Vanilla Transformer baseline (from Table 1) with a cyclic channel-shift augmentation of ± 3. We hold all training hyperparameters strictly fixed against the unaugmented run. We then evaluate both models at $k = 0$ (no rotation) on the test split (Table 10).

The ± 3 channel-shift aug trails the unaugmented transformer at $k = 0$ by $2.4°$ User and $4.3°$ Stage. At that same $k = 0$, the 4M **CycloFormer** is $13.7°$ Stage against $14.5°$ unaugmented and $18.8°$ with ± 3; User is $12.0°$ against $11.6°$ unaugmented. The 48M model is $12.7°$ Stage (Table 10). This is the $k = 0$ number next to the commutation argument in Appendix B.1.

Table 10: **Geometric prior vs. rotation augmentation.** Test-set angular MAE and landmark distance evaluated at $k = 0$ (no rotation at inference). The baseline TDS+Vanilla Transformer is trained both without (top row) and with (second row) cyclic channel-rotation augmentation of ± 3. **CycloFormer** is exactly $\mathbb{Z}_{16}$-invariant by construction and trains without rotation augmentation. Displayed as mean $\pm$ standard deviation across users.

			User		*Stage*		*User+Stage*	
Model	Params	Rot. aug.	Ang. (°)	LD (mm)	Ang. (°)	LD (mm)	Ang. (°)	LD (mm)
TDS+Vanilla Transformer	5M	—	11.6 ± 1.2	14.8 ± 1.7	14.5 ± 1.6	19.5 ± 2.1	15.0 ± 1.3	20.5 ± 1.9
TDS+Vanilla Transformer	5M	± 3	14.0 ± 1.4	19.3 ± 2.2	18.8 ± 1.6	27.1 ± 2.1	18.3 ± 1.5	26.8 ± 1.8
CycloFormer	4M	—	12.0 ± 1.4	15.2 ± 1.7	13.7 ± 1.6	18.3 ± 2.0	15.2 ± 1.5	20.6 ± 2.1
CycloFormer	48M	—	$\mathbf{11.7 \pm 1.6}$	$\mathbf{14.7 \pm 1.9}$	$\mathbf{12.7 \pm 1.6}$	$\mathbf{17.1 \pm 2.1}$	$\mathbf{14.7 \pm 1.4}$	$\mathbf{19.8 \pm 1.9}$

Input window length. We evaluate the 48M **CycloFormer** at $4\times$ the default temporal context (10,000 to 40,000 samples; 5 s to 20 s at 2 kHz). Performance remains virtually identical. Every metric shifts by less than 0.3 units. The factored spatial-temporal architecture successfully extracts all relevant kinematic structure within the standard 5 s window.

Table 11: **Effect of input window length.** Evaluation of the 48M **CycloFormer**. Each cell: Ang. (°) / LD (mm) / PA (mm), averaged over users in each generalisation split. Ang.: joint-angle MAE; LD: landmark distance; PA: PA-MPJPE.

Window	*Stage*	*User*	*User+Stage*
10,000 samples (5 s)	12.7 / 17.1 / 13.7	11.7 / 14.7 / 10.7	14.7 / 19.8 / 16.0
40,000 samples (20 s)	12.6 / 17.3 / 13.7	11.5 / 14.6 / 10.7	14.6 / 20.1 / 16.1

Spatial vs. temporal depth at fixed compute. We investigate the optimal allocation of Transformer layers across the spatial and temporal dimensions under a fixed compute budget. Table 12 compares our balanced 3+3 configuration against asymmetric 2+4 and 4+2 splits, holding the total depth constant at six layers.

Table 12: **Spatial vs. temporal depth allocation.** Performance of the 4M **CycloFormer** (50% data split) under different spatial and temporal layer distributions ($N_{\text{spat}} + N_{\text{temp}} = 6$). The balanced 3+3 configuration strictly outperforms both asymmetric allocations across all generalisation splits. While the absolute gaps are small ($\leq 0.4°$, ≤ 0.7 mm), the uniform advantage of the 3+3 split indicates that spatial (cross-channel) and temporal contexts provide comparable signal, making symmetric allocation optimal. Ang.: joint-angle MAE; LD: landmark distance; PA: PA-MPJPE.

Setting	**Split**	**Users**	**Ang. (°)**	**LD (mm)**	**PA (mm)**
2+4	*Stage*	158	15.25 ± 1.69	20.78 ± 2.17	16.65 ± 1.77
	User	20	12.43 ± 1.25	16.13 ± 1.59	11.67 ± 1.18
	User+Stage	20	16.10 ± 1.11	22.17 ± 1.66	18.07 ± 1.42
4+2	*Stage*	158	$\mathbf{14.96 \pm 1.66}$	20.35 ± 2.16	16.31 ± 1.75
	User	20	12.29 ± 1.24	15.91 ± 1.50	11.62 ± 1.17
	User+Stage	20	15.88 ± 1.43	21.96 ± 1.94	17.80 ± 1.65
3+3	*Stage*	158	14.97 ± 1.65	$\mathbf{20.25 \pm 2.17}$	$\mathbf{16.15 \pm 1.72}$
	User	20	$\mathbf{12.21 \pm 1.15}$	$\mathbf{15.65 \pm 1.37}$	$\mathbf{11.44 \pm 1.13}$
	User+Stage	20	$\mathbf{15.72 \pm 1.24}$	$\mathbf{21.50 \pm 1.57}$	$\mathbf{17.45 \pm 1.39}$

Augmentation mix. We evaluate the impact of input-domain noise by toggling the jitter ($\sigma = 0.15$) and channel dropout ($p = 0.1$) augmentations. Table 13 reports the best-validation-loss epoch. Architectural regularisers, such as temporal-chunk dropout, remain active across both runs. Removing input noise reduces training MAE by 26%. However, validation MAE improves by only 5%. The input augmentations successfully narrow the train-val gap by a factor of 2.3 while slightly raising the absolute validation error.

Table 13: **Input-noise augmentation ablation.** Evaluation of the best-val-loss epoch. The input-domain pair (jitter $\sigma = 0.15$, channel-dropout $p = 0.1$) is toggled. Removing the augmentations sharply drops training error but provides minimal validation gains, widening the generalisation gap.

Input-noise aug.	Epoch	Train MAE (°)	Train LD (mm)	Val MAE (°)	Val LD (mm)
Jitter + ChDrop	67	11.21	14.39	12.95	17.11
None	59	8.26	9.93	12.31	15.79

Double descent and implicit regularisation. Adding jitter and channel dropout shrinks the generalisation gap from $4.05°$ to $1.74°$. However, this elevates both curves simultaneously. Training MAE rises by $2.95°$ and validation MAE rises by $0.64°$. A classical bias-variance perspective views this

as a standard trade-off. The double-descent framework (Belkin et al., 2019; Nakkiran et al., 2020) offers a more precise reading. At 4M parameters, our model sits near the interpolation threshold. Here, the error curve is non-monotone with respect to regularisation strength. Explicit augmentations stack redundantly with the implicit regularisation of SGD. The gap closes because both errors are artificially elevated, not because variance is genuinely absorbed. This perfectly matches the failure mode predicted by double descent at the interpolation boundary.

Robustness under inference-time input perturbations. We evaluate the robustness of **CycloFormer** and *vemg2pose* against hardware-induced noise by applying two input-domain perturbations strictly at inference time. Neither model is exposed to these perturbations during training.

First, we introduce per-channel log-normal amplitude jitter, $x_c \leftarrow x_c \cdot \exp(\varepsilon_c)$ where $\varepsilon_c \sim \mathcal{N}(0, 0.15^2)$. This simulates the natural variance in contact impedance across different donning sessions. Second, we apply Bernoulli channel dropout, $x_c \leftarrow x_c \cdot m_c$ where $m_c \sim \mathrm{Bern}(0.9)$, to simulate intermittent electrode failure. Both perturbations are applied directly to the raw input windows prior to the first network layer. We report test-set angular MAE, landmark distance, and PA-MPJPE averaged across three random seeds for the perturbation sampling. Because the underlying model checkpoints remain fixed, the reported variability strictly reflects the architectures' vulnerability to the input noise distribution.

Table 14: **Robustness under inference-time input perturbations.** Test-set error with per-channel log-normal jitter ($\sigma_{\log} = 0.15$) and channel dropout ($p = 0.1$) applied simultaneously at inference. Displayed as mean $\pm$ standard deviation across users, with perturbation noise averaged over 3 seeds. **CycloFormer** improves over *vemg2pose* on every split and metric, widening the gap by up to $2.0°$ angular MAE and 2.9 mm landmark distance on the hardest generalisation regime. Ang.: joint-angle MAE; LD: landmark distance; PA: PA-MPJPE.

Model	Split	Ang. (°)	LD (mm)	PA (mm)
vemg2pose	*Stage*	16.9 ± 1.6	23.2 ± 2.0	18.3 ± 1.7
	User	13.5 ± 1.2	17.9 ± 1.7	12.8 ± 1.2
	User+Stage	17.1 ± 1.4	23.7 ± 1.8	19.0 ± 1.5
CycloFormer (4M)	*Stage*	$\mathbf{14.9 \pm 1.7}$	$\mathbf{20.3 \pm 2.2}$	$\mathbf{16.3 \pm 1.8}$
	User	$\mathbf{12.6 \pm 1.4}$	$\mathbf{16.3 \pm 1.6}$	$\mathbf{12.0 \pm 1.2}$
	User+Stage	$\mathbf{16.1 \pm 1.4}$	$\mathbf{22.2 \pm 2.0}$	$\mathbf{18.1 \pm 1.6}$

Under these challenging conditions, **CycloFormer** consistently outperforms *vemg2pose* across every generalisation split and metric (Table 14). This demonstrates that our geometric priors do not merely overfit to pristine training conditions; they yield representations that are fundamentally more resilient to physical, out-of-distribution sensor noise. Consistent with the component ablations in Table 2, this performance gap is most pronounced on the *Stage* split. Structural inductive biases are most critical when the network must generalise to unseen gesture kinematics and degraded signal quality simultaneously.

B.4 Extended Performance Analysis

We expand on the main-text results by probing the quality and distribution of our model's predictions. First, a transfer evaluation tests if our learned features generalise to discrete downstream tasks. Second, a granular head-to-head comparison with *vemg2pose* decomposes kinematic errors by user, gesture, and frame. Finally, an internal failure analysis isolates where **CycloFormer**'s residual errors concentrate.

B.4.1 Linear probing

Following Cui et al. (2025), we freeze the encoder after pose-regression training and train a linear classifier on its embeddings. Stage groups, as per CPEP: finger pinches, thumb rotation, thumb swipes, and claw/grasp/flick. Macro accuracy is 87.3% at 3.53M and 89.4% at 56.90M (micro 88.2% / 90.0%).

CPEP reports 65.0% for *vemg2pose* (velocity prediction), 73.4% for a `emg2pose` (joint angles prediction), and 78.2% on their contrastive EMG–pose model, which they particularly train for this classification. Zero-shot on that contrastive model is 75.7%. The 56.90M head is 89.4%, against 65.0% / 73.4% / 78.2%.

On `emg2pose`, donning is take-off and put-on of the band between sessions (Salter et al., 2024). The linear head is trained on frozen features from that encoder (Appendix B.1.1).

Table 15: **Linear-probe gesture classification.** Macro accuracy (%) on the four CPEP classes.

Method	Pretraining objective	Macro Acc. (%)
EMG-MAE (Cui et al., 2025)	masked autoencoder	34.7
vemg2pose (Salter et al., 2024)	pose regression (variational)	65.0
NeuroPose (Liu et al., 2021)	pose regression	69.2
PoseT (Cui et al., 2025)	pose regression (Transformer)	70.5
emg2pose sup. (Salter et al., 2024)	pose regression	73.4
CPEP zero-shot (Cui et al., 2025)	contrastive (cross-modal)	75.7
CPEP linear probe (Cui et al., 2025)	contrastive (cross-modal)	78.2
CycloFormer (ours, 3.53M)	pose regression	87.3
CycloFormer (ours, 56.90M)	**pose regression**	**89.4**

B.4.2 Head-to-Head with *vemg2pose*

We compare **CycloFormer** (4M) to *vemg2pose* along three axes: per-user (Figure 8), per-stage (Figure 9), and frame-level paired (Figure 10). The two models stay closely matched user-by-user with **CycloFormer**'s IQRs slightly tighter than *vemg2pose*'s, and **CycloFormer** matches or beats *vemg2pose* on most stages. Per-frame Pearson $r \approx +0.55$ across the validation sample means the two models are confused by the same frames, but the cloud sits above $y = x$ for most sessions (mean $\Delta \approx +1.4^\circ$ dataset-wide), so the gain is uniformly distributed across frames rather than concentrated on a few easy stages. The regressions cluster on contact-rich and bimanual gestures (PlayBlocksChess, HandDeskSeparateClaspedChest, GraspPunch, FingerTouchPalm in *User*; HandOverHandCountingUpDownFingerWigglingSpreading in *User+Stage*, the largest single regression at $+1.9^\circ$), where the EMG signal is most confounded by external load or contralateral activity.

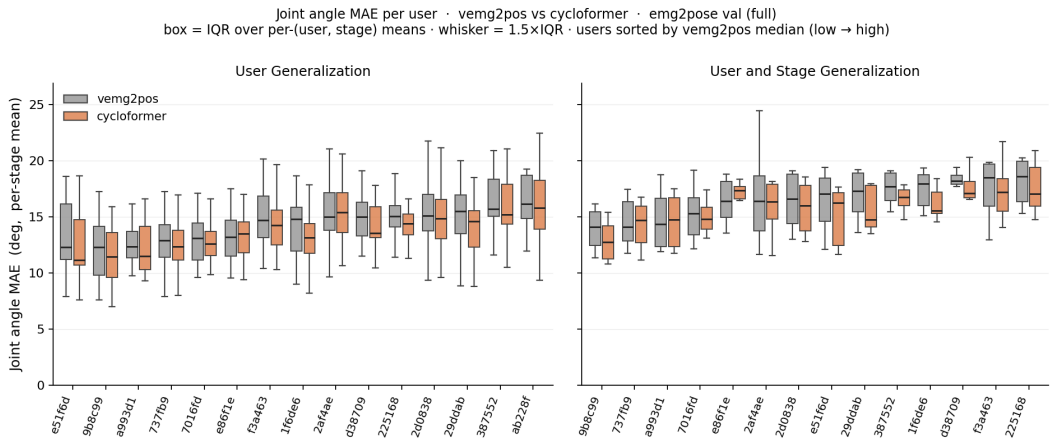

Figure 8: **Per-user MAE on the `emg2pose` validation split, *vemg2pose* vs. CycloFormer (4M).** Each box is the inter-quartile range over per-(user, stage) mean MAE, the dash is the median across those (user, stage) means and whiskers extend to $1.5\times$IQR. Users are sorted left-to-right by ascending *vemg2pose* median. Per-user variability is the dominant axis of disagreement: the easiest-user vs. hardest-user spread ($\sim 12^\circ \rightarrow \sim 17^\circ$) is comparable to the gap between gestures. The two models stay closely matched across users; **CycloFormer**'s IQRs are usually a touch tighter than *vemg2pose*'s.

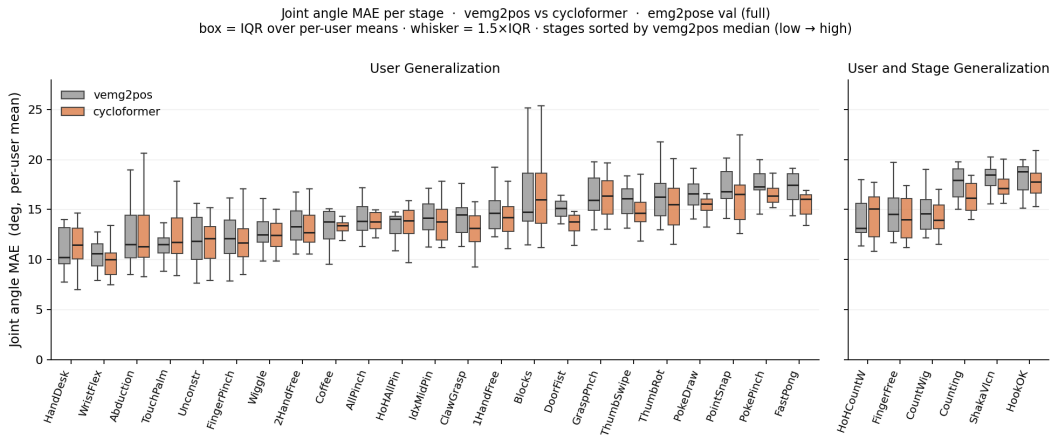

Figure 9: **Per-stage MAE on the `emg2pose` validation split, *vemg2pose* vs. CycloFormer (4M).** Each box represents the IQR of per-user mean MAE across users for that stage; the dash is the median across users and whiskers extend to $1.5\times$IQR. Stages are sorted left-to-right by ascending *vemg2pose* median, so the gestures *vemg2pose* finds easiest sit on the left. While **CycloFormer** matches or beats *vemg2pose* on most stages and the two models track each other closely, with **CycloFormer**'s median running a little lower across the harder gestures on the right, it regresses on a small but consistent set of stages dominated by hand-on-object contact and asymmetric bimanual coordination PlayBlocksChess, HandDeskSeparateClaspedChest, GraspPunch, FingerTouchPalm in *User*, and HandOverHandCountingUpDownFingerWigglingSpreading in *User+Stage* (the largest single regression at $+1.9^\circ$ median). These are exactly the stages where the EMG signal is most confounded by external load or contralateral activity, suggesting **CycloFormer**'s gains come from clean, isolated finger movements while it gives up some ground on contact-rich and two-handed gestures.

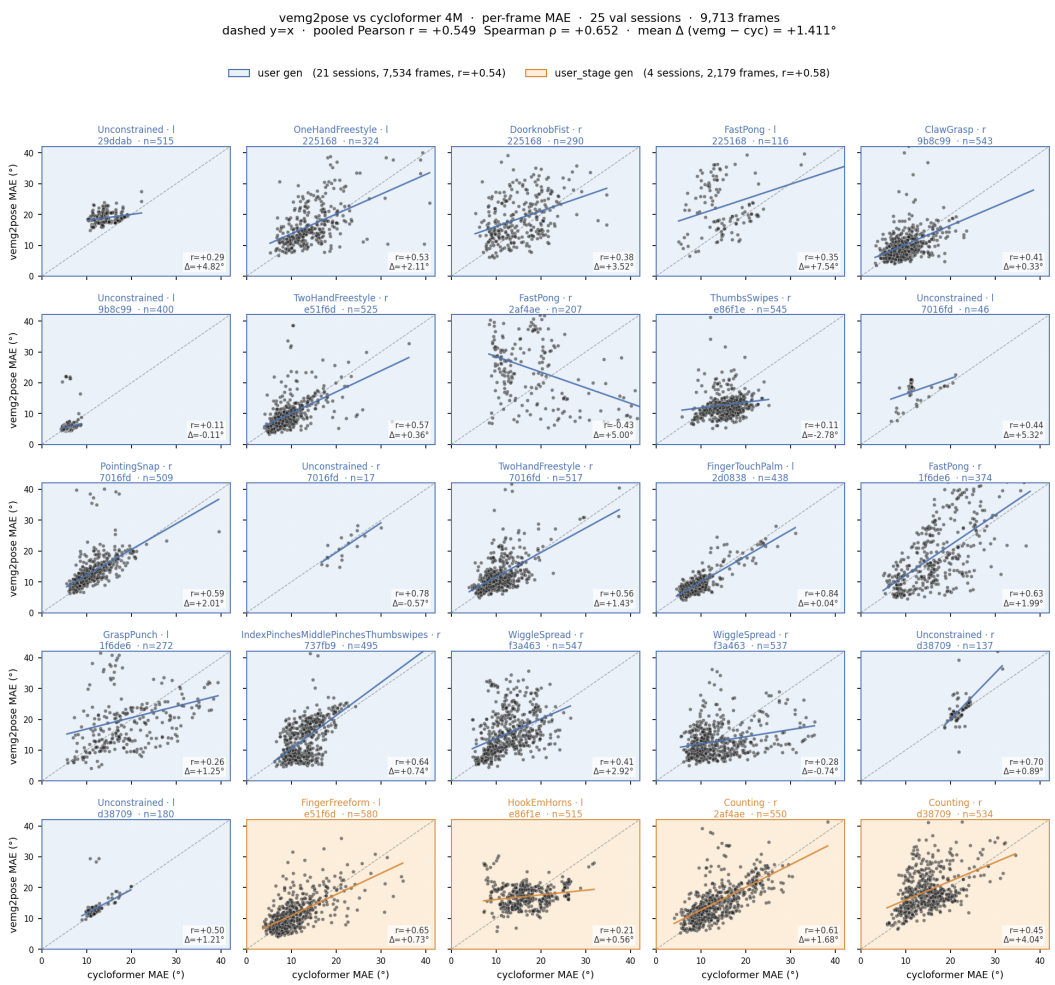

Figure 10: **Paired per-frame MAE: *vemg2pose* vs. CycloFormer (4M), 25-session validation sample.** Per-frame MAE scatter (*vemg2pose* on y-axis vs. **CycloFormer** on x-axis), 25 validation sessions, ∼9.7k frames. Dashed grey $y = x$ is parity; points above the line = **CycloFormer** is better on that frame, below = *vemg2pose* is better. Per-session correlation r and mean Δ = *vemg2pose* − **CycloFormer** are annotated; mean $\Delta \approx +1.4°$ dataset-wide (**CycloFormer** wins on average). High correlation between the two models' per-frame error (pooled $r \approx +0.55$, $\rho \approx +0.65$): they tend to be confused by the same frames. But the cloud sits above $y = x$ for most sessions, especially in *User+Stage* (orange panels), confirming **CycloFormer**'s edge isn't from a few easy stages but is broadly distributed across frames.

B.4.3 Failure-Mode Breakdown of CycloFormer

Where does residual error live within **CycloFormer**? Three lenses: gesture family (Table 16), hand-pose state (Figure 11), and joint (Figure 12 and Table 17). Family-level MAE is tight at 11.4–12.9° despite a 7× imbalance in session counts, **CycloFormer** does not overfit to the dominant free-form prompts. Per-frame error correlates negatively with hand openness (pooled $r \approx -0.32$): closed-fist and pinch frames are harder than extended-finger frames. At joint level, abduction error is strikingly bimodal: index (4.9°) and ring (5.1°) abduct over narrow anatomical ranges and are easy, whereas middle (17.4°), pinky (18.0°), and thumb CMC (16.1°) abductions span larger noisier ranges and are the hardest joints in the model, alongside ring PIP (18.1°).

Table 16: **Per-stage angular error (test, User-Stage split).** Stages are aggregated into five gesture families; n_{stages} is the number of distinct prompts in each family and n_{sessions} the total recordings pooled into the family-level mean. Two observations are worth flagging. First, the two metrics rank families *differently*: thumb-dominant gestures are the hardest in joint-angle space ($12.94°$) yet the easiest in Procrustes-aligned Cartesian space (9.84 mm), because thumb errors are concentrated in CMC abduction – a large angular swing that translates to small fingertip displacement. Conversely, finger-wiggling sits near the bottom in degrees but at the top in millimetres, since small MCP/PIP errors at long lever arms produce the largest tip displacements. Second, the spread across families is tight (11.37–$12.94°$, 9.84–11.71 mm) despite a $7\times$ imbalance in session counts, indicating that **CycloFormer** does not over-fit to the dominant free-form prompts and motivates targeted collection of pinch and wiggle data rather than more free-form play. Ang.: joint-angle MAE; PA: PA-MPJPE.

Stage family	n stages	n sessions	Ang. (°)	PA (mm)
Thumb-dominant gestures	2	156	12.94	9.84
Hand grasps / claws	3	193	12.69	11.12
Finger pinches	6	341	12.59	11.05
Finger wiggling / abduction	4	306	11.46	11.71
Free-form & dexterous play	7	1098	11.37	10.19
Macro mean (across 5 families)	22	**2094**	**12.21**	**10.78**

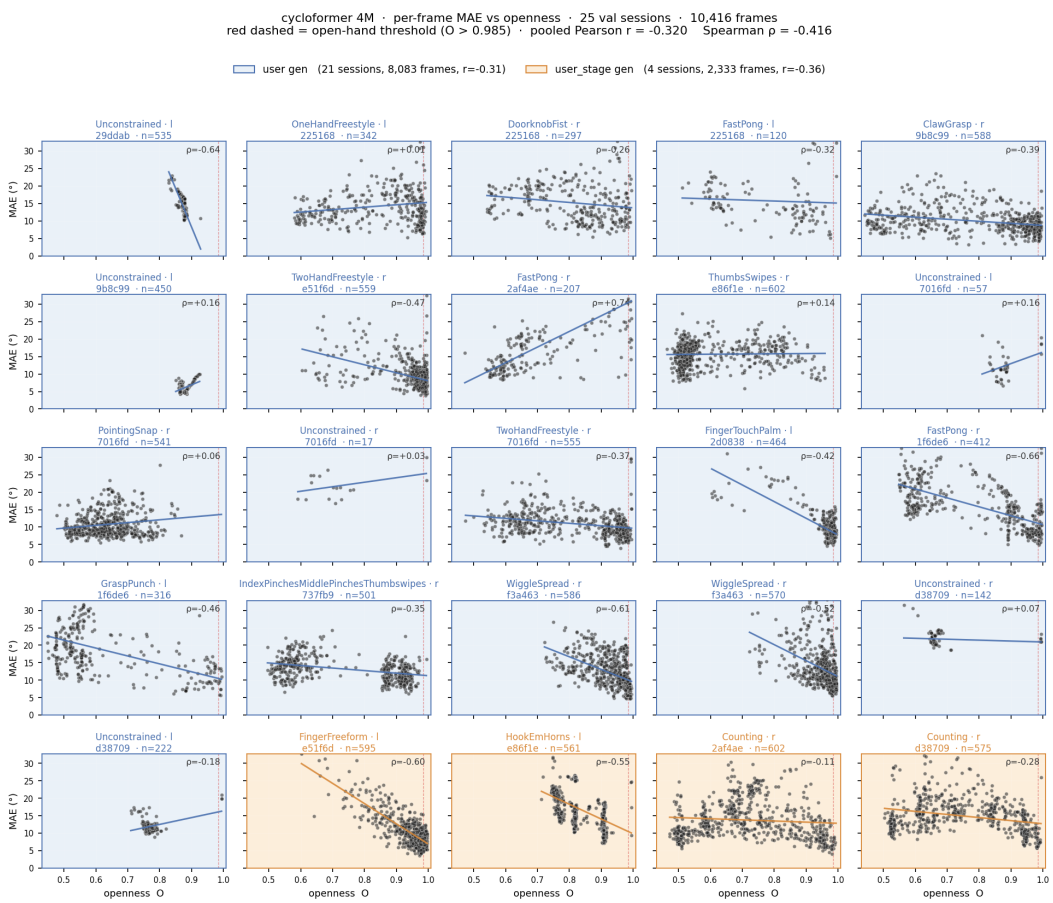

Figure 11: **CycloFormer** (4M) **per-frame MAE vs. openness, 25-session validation sample.** Per-frame MAE plotted against the geometric scale-invariant openness O^1 of the GT pose, on a 25-session sample of the validation split (10,416 frames). Each subplot is one session, panel-titled with stage and user, colour-coded by generalisation split (blue = *User*, orange = *User+Stage*). Per-session Pearson r and pooled Pearson/Spearman in the suptitle. A moderate, consistent negative correlation (pooled $r \approx -0.32$, $\rho \approx -0.42$), open hands have lower MAE than closed/curled hands. Within-session correlations vary a lot by gesture (some near zero, some $r > 0.6$), but the dataset-wide trend is clean: **CycloFormer** is more confident on extended-finger frames than on closed-fist or pinch frames.

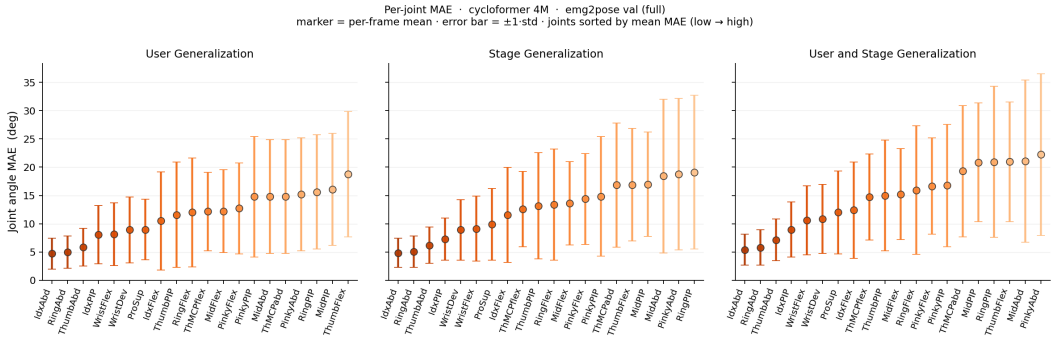

Figure 12: **Per-joint MAE on the `emg2pose` validation split (CycloFormer,** 4M**).** **CycloFormer** (4M) MAE for each of the 20 UmeTrack DoFs. A point is the mean over val frames for that joint and split; the whisker is the per-frame standard deviation. Order is increasing mean, left to right. Index and ring abduction are low. Thumb CMC abduction is 16.1° on Stage for the 48M model (Table 17).

Table 17: **Per-joint angular error (*Stage* split, 48M model).** All values in degrees (MAE). Column means use the fifteen printed cells. The blank distal row is thumb IP together with DIP of index, middle, ring, and pinky; UmeTrack ties those DIPs to PIP. Index abduction is $4.9°$ and ring $5.1°$. Thumb CMC abduction is $16.1°$. Ring PIP is $18.1°$. Index has the lowest column mean ($8.0°$).

Joint	Thumb	Index	Middle	Ring	Pinky
MCP flex/ext	12.6	11.3	13.3	13.1	14.1
PIP flex/ext	12.8	7.7	17.1	18.1	15.1
Distal (IP/DIP)	—	—	—	—	—
Abduction	16.1	4.9	17.4	5.1	18.0
Mean	**13.8**	**8.0**	**15.9**	**12.1**	**15.7**

Table 18: **Inference latency and parameter efficiency.** Wall-clock latency for a single 5-second window ($T = 10{,}000$, batch 1) on an NVIDIA A100 (40 GB, SXM4), `bf16` autocast, averaged over 2,000 timed iterations after 200 warm-up passes (cuDNN benchmark on, GPU clocks locked). Each model is loaded from its trained checkpoint, so timings include every operation that runs at deploy time — including the sequential LSTM rollout that *vemg2pose* performs at 50 Hz over the prediction horizon. Two findings stand out. *(i)* **CycloFormer** runs $\sim 8\times$ faster than *vemg2pose* at comparable parameter count because its prediction is a single parallel forward pass, whereas *vemg2pose* pays for ~ 250 sequential LSTM steps that cannot be batched along time. *(ii)* The $12\times$ jump from the 4M to the 48M **CycloFormer** in parameter count adds essentially no latency ($+0.05$ ms): both share the same depth, so the kernel-launch budget is identical and the A100 has spare FLOPs for the wider matmuls at batch 1.

Model	Params	A100 latency (ms)
vemg2pose	5.98M	89.51 ± 0.20
CycloFormer	4.18M	11.06 ± 0.04
CycloFormer	47.99M	11.11 ± 0.12

B.4.4 Inference Latency

CycloFormer regresses pose in a single parallel forward pass over the prediction window, so its wall-clock latency is dominated by depth (kernel-launch budget) rather than width. Two consequences (Table 18). First, **CycloFormer** at 4.18M parameters runs in 11.06 ± 0.04 ms per 5-second window on an A100, $\sim 8\times$ faster than *vemg2pose*'s 89.51 ± 0.20 ms at comparable parameter count, because *vemg2pose* pays for ~ 250 sequential LSTM steps that cannot be batched along time. Second, scaling **CycloFormer** from 4M to 48M parameters costs an additional 0.05 ms: depth and thus the kernel-launch budget are unchanged, and the A100 has spare FLOPs for the wider matmuls at batch 1.

B.4.5 Reconciliation with the *vemg2pose-tran* Baseline

We re-run the *vemg2pose-tran* decoder of Salter et al. (2024, Appendix D.3) on a V100. D.3 is a BERT encoder next to their LSTM; they cut transformer width in half to fit T4 memory, swept $2, 4, 6$ layers and $2, 4, 8$ heads, and the transformer did not beat the LSTM.

Architecture. The featuriser reuses the *vemg2pose* featuriser of Salter et al. (2024) unchanged: two strided `Conv1dBlocks` ($16 \to 256$, $k = 11$, $s = 5$ and $256 \to 256$, $k = 5$, $s = 2$) followed by two `TdsStages` (16 channels $\times$ 16-feature width, kernel sizes 17 and 9 with strides 4 and 2, output projected to $d_{\text{feat}} = 128$). Cumulative downsampling is $5 \cdot 2 \cdot 4 \cdot 2 = 80\times$ (2 kHz to 25 Hz); the left context is 1,790 samples, matching *vemg2pose*. The decoder head is a 6-layer pre-norm `nn.TransformerEncoderLayer` stack with $d_{\text{model}} = 256$, 8 attention heads, $d_{\text{ff}} = 256$, dropout 0.15, and bidirectional self-attention with RoPE (Su et al., 2021) on the temporal axis. A final `LayerNorm` and a linear projection map each 25 Hz token to the 20-DoF joint angles. Total parameters $\approx$5.3M; the head makes no design choices specific to sEMG beyond the use of RoPE and pre-norm.

Training. We use Adam at learning rate 10^{-3}, gradient clipping at $\|g\|_2 \leq 1.0$, and `ReduceLROnPlateau` on the validation loss (factor 0.5, patience 4), with patience-10 early stopping. The window length is $11{,}790$ samples ($10{,}000$ prediction $+$ $1{,}790$ left context); batch size 64, 100 epochs maximum. Mixed-precision on a single NVIDIA V100 (32 GB).

Reconciliation with Salter et al. (2024, Appendix D.3). Their T4 transformer is $12.5°$ / $16.1°$ / $16.2°$ User / Stage / User+Stage against $12.2°$ / $15.2°$ / $15.8°$ for the LSTM. Our six-layer pre-norm encoder at $d_{\text{model}} = 256$ with RoPE and a per-frame linear head is $11.6°$ / $14.5°$ / $15.0°$ (Table 19).

Table 19: **Reconciliation with Salter et al. (2024, Appendix D.3).** The top two rows of each block reproduce the original benchmark findings. The third row represents our re-run using contemporary compute and modern architectural norms. Δ is computed against *vemg2pose-lstm*, the primary baseline against which the original Transformer degradation was reported.

Split	Decoder	Ang. (°)	Δ	LD (mm)	Δ
User	*vemg2pose-lstm* (Salter et al., 2024)	12.2 ± 1.3	—	15.8 ± 1.9	—
	vemg2pose-tran (Salter et al., 2024, Appendix D.3)	12.5 ± 1.3	$+0.3$	16.3 ± 1.8	$+0.5$
	TDS+Vanilla Transformer (ours)	$\mathbf{11.6 \pm 1.2}$	-0.6	$\mathbf{14.8 \pm 1.7}$	-1.0
Stage	*vemg2pose-lstm*	15.2 ± 1.6	—	20.4 ± 2.2	—
	vemg2pose-tran	16.1 ± 1.6	$+0.9$	21.7 ± 2.1	$+1.3$
	TDS+Vanilla Transformer (ours)	$\mathbf{14.5 \pm 1.6}$	-0.7	$\mathbf{19.5 \pm 2.1}$	-0.9
User+Stage	*vemg2pose-lstm*	15.8 ± 1.4	—	21.6 ± 2.0	—
	vemg2pose-tran	16.2 ± 1.3	$+0.4$	22.3 ± 1.8	$+0.7$
	TDS+Vanilla Transformer (ours)	$\mathbf{15.0 \pm 1.3}$	-0.8	$\mathbf{20.5 \pm 1.9}$	-1.1

C Cross-Modal Comparison Details

C.1 MANO Renderer Tuning and Validation

Before the cross-modal comparison is trustworthy, the renderer must not itself be a confound. Running *HaMeR* on FreiHAND frames via our renderer yields a $+0.90$ mm gap between real and synthetic (Table 20).

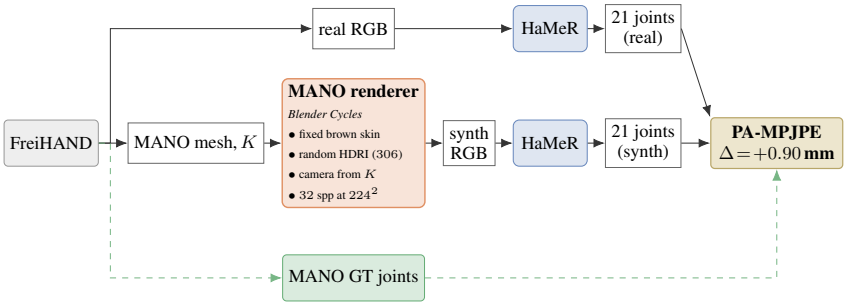

Figure 13: MANO renderer validation on FreiHAND. Real RGB is fed directly to *HaMeR*; the MANO mesh and K drive our renderer to produce a synthetic image, which is then passed through *HaMeR*. Both predictions are PA-MPJPE'd against the same MANO ground-truth joints.

Setup. We use the first $N = 200$ frames (sorted by annotation filename) from the FreiHAND (Zimmermann et al., 2019) evaluation split, which ships paired (real RGB at 224×224, MANO (Romero et al., 2017) mesh posed in OpenCV camera coordinates, per-frame intrinsics K). For each frame we (i) map vertices into Blender's world frame via $\mathrm{diag}(1, -1, -1)$, (ii) configure a Blender camera with focal length $f_x \cdot 36/W$ mm and shift derived from K, (iii) Cycles render at 32 samples with a per-frame-random HDRI from RenderIH and a single fixed skin texture (the winner of an 18-recipe optimisation sweep), and (iv) feed both the real and synthetic image to *HaMeR* (Pavlakos et al., 2024) with the same GT-derived bounding box (project the 21 GT joints with K, take their padded 2D AABB).

Each candidate frame is rendered with a fresh random rigid wrist transform: a Haar-uniform rotation on SO(3) (Shoemake's quaternion method) composed with a fixed prior orientation R_{orient} (palm-down side or first-person), plus a ± 20 mm uniform translation jitter.

Result. *HaMeR*'s PA-MPJPE is 6.44 mm on real photos and 7.34 mm on our renders of the same poses (delta $+0.90$ mm; per-frame Pearson $r = 0.591$). The full distribution is in Figure 20 and the worst-frame analysis in Figures 14 and 15.

Table 20: *HaMeR* PA-MPJPE on $N = 200$ FreiHAND eval frames, evaluated on the real photo and on our Blender render of the same MANO ground-truth mesh and intrinsics. The renderer adds $+0.90$ mm; per-frame Pearson correlation between the two paths is $r = 0.591$.

Input image	Mean (mm)	Median (mm)	Std (mm)
HaMeR (real)	6.44	5.98	3.11
HaMeR (synth)	7.34	6.57	3.35
Δ (synth $-$ real)	$+0.90$	$+0.59$	—

Per-frame correlation $r = 0.59$. A perfectly faithful renderer would give r close to 1 (per-frame error would be dominated by pose-dependent difficulty, shared between real and synth). Decomposing the variance gives $\mathrm{Var}(\text{real}) = 9.7\,\text{mm}^2$, $\mathrm{Var}(\text{synth}) = 11.2\,\text{mm}^2$, and $\mathrm{Cov} = 6.2\,\text{mm}^2$, so the synth-side image-specific noise (sleeves missing, texture seams, HDRI extremes) is roughly 44% larger than real's. Eliminating that noise alone would push r to roughly 0.80 under an additive-noise model; reaching higher would require matching real's noise sources (motion blur, scene context).

Where do the worst frames fail? To make the noise concrete, the figures below show the 5 frames where synth is hardest for *HaMeR* vs real and the 5 where synth is unexpectedly easy. Each figure is laid out as 5 columns $\times$ 2 rows: top row is the FreiHAND real photo, bottom row is the Blender synth render. GT keypoints in green, *HaMeR*'s predictions in magenta.

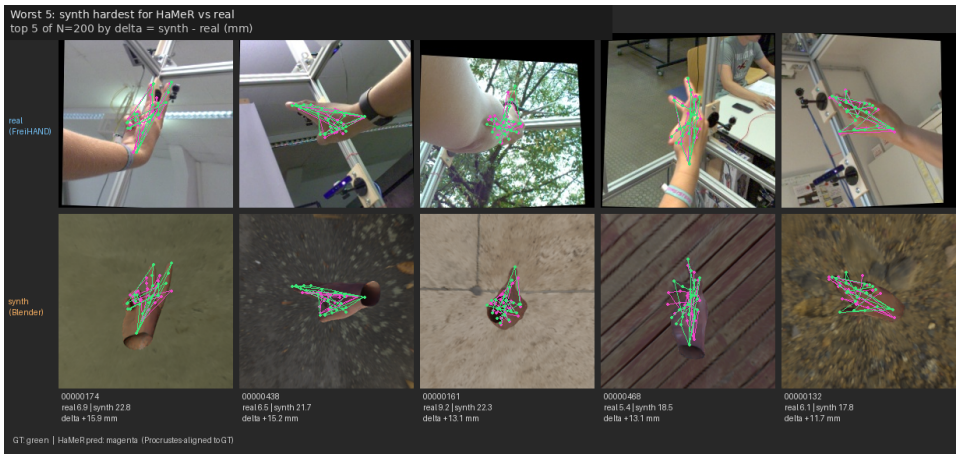

Figure 14: Worst-5 frames (synth hardest vs real).

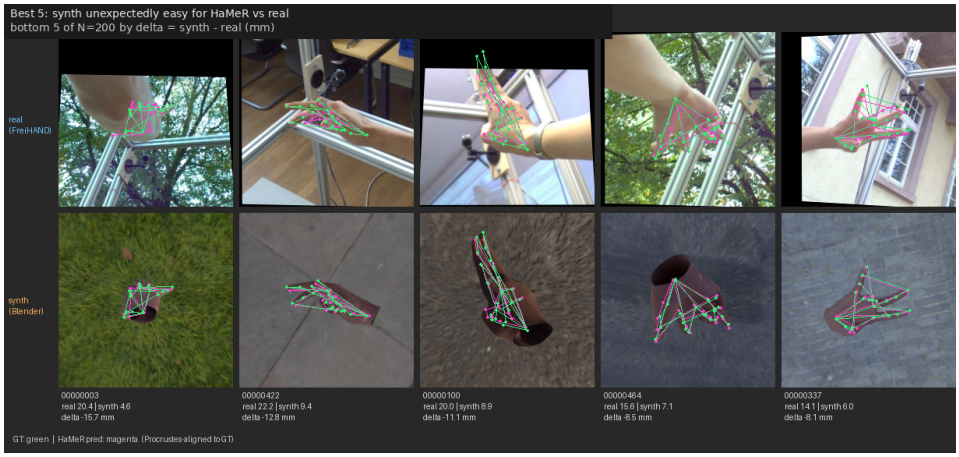

Figure 15: Best-5 frames (synth easy vs real).

C.2 MANO Retargeter Tuning and Validation

HaMeR returns 21 joints in MANO topology; `emg2pose` stores 20-DoF UmeTrack joint angles. To score *HaMeR* in its native parameterisation (no skeleton-name floor), we fit a per-frame MANO pose to the 21 UmeTrack landmarks via Adam optimisation of $\{\text{GLOBAL_ORIENT}, \text{HAND_POSE}, \text{TRANSL}\}$ with shared per-session β, fingertip-weighted joint loss, and a five-keypoint Procrustes initialisation.

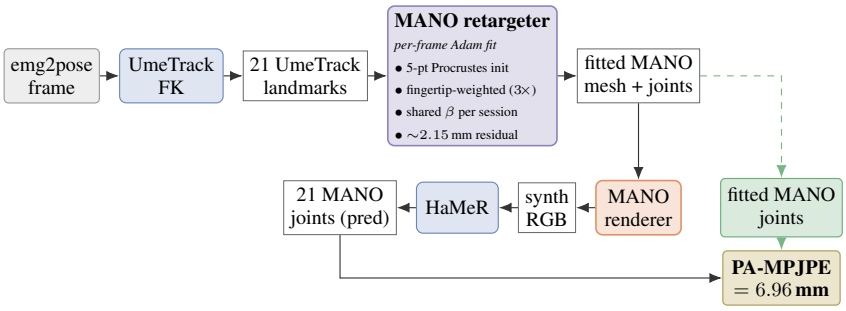

Figure 16: UmeTrack-to-MANO retargeting pipeline. Each `emg2pose` frame is forward-kinematic'd to 21 UmeTrack landmarks; a per-frame Adam fit (5-keypoint Procrustes init, fingertip-weighted joint loss, shared per-session, 2.15 mm residual) recovers a MANO mesh and joints. The fitted mesh is rendered through the same Blender Cycles pipeline as Figure 13, *HaMeR* consumes the render, and the prediction is PA-MPJPE'd against the fitted MANO joints (dashed green reference) at 6.96 mm.

`emg2pose` ships 20-DoF UmeTrack (Han et al., 2022) joint angles. For *HaMeR* scoring in MANO topology, we run per-frame Adam (600 iters, lr 0.02 cosine-annealed to 0.002) over $\{\text{GLOBAL_ORIENT}, \text{HAND_POSE}, \text{TRANSL}\}$ with shared $\beta \in \mathbb{R}^{10}$ across a session, minimising a fingertip-weighted (3×) MSE between MANO 20 joints and UmeTrack 20 joints (we drop the OpenPose thumb-CMC because UmeTrack lacks it). Initialising from a five-keypoint Procrustes (wrist + thumb tip + three MCPs) avoids two local minima that the joint loss alone does not disambiguate: an all-20-joint init lets fingertip mismatch dominate the rotation, and a four-keypoint (no thumb) init leaves the thumb twisted. A side-aware $\pm 90^\circ$ wrist roll then breaks a remaining symmetry in the palm direction. The fit converges to a 2.15 mm mean residual, comparable to FreiHAND's own MANO fits, and is replayed in `Ume-Mano fit animation.mov` (supplementary). 30 sessions × 60 frames at 6 fps give the *HaMeR*-on-`emg2pose` baseline in Table 21: 6.96 mm against fitted MANO joints, within 0.38 mm of the FreiHAND synthetic baseline above.

Table 21: *HaMeR* PA-MPJPE on `emg2pose` (30 sessions × 60 frames at 6 fps, $N = 1{,}800$ per render), as a function of which mesh family is rendered and which joint set *HaMeR* is scored against. Apples-to-apples is the (MANO, fitted MANO) cell at 6.96 mm; the ∼2.15 mm gap to the UmeTrack column is the topology-mapping floor.

	Reference (PA-MPJPE, mm)	
Rendered mesh	Fitted MANO	UmeTrack-20
MANO	**6.96**	**9.11**
UmeTrack	6.99	9.16

C.3 Cross-Modal Occlusion Comparison Protocol

This section details the cross-modal comparison pipeline upon which the results in Section 4 are reported.

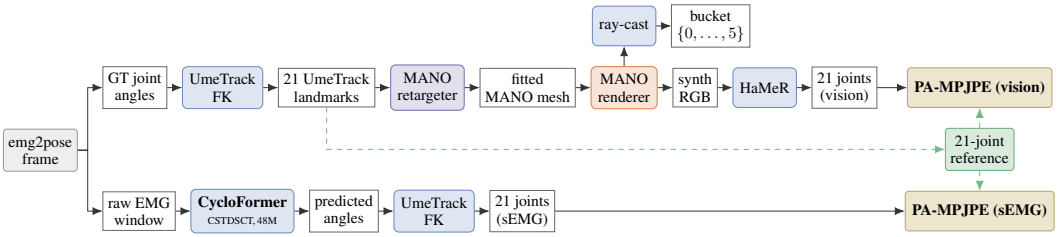

Figure 17: Cross-modal comparison by fingertip occlusion. Each held-out frame is processed by two independent predictors that never share information. A ray-cast against the rendered mesh assigns each frame to a fingertip-occlusion bucket $\{0, \ldots, 5\}$; PA-MPJPE is reported per bucket.

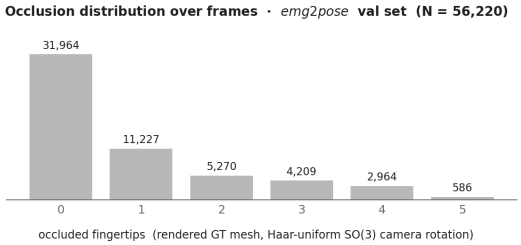

Figure 18: Distribution of frames across fingertip-occlusion buckets on the `emg2pose` validation split. Buckets are derived by ray-casting against the rendered MANO mesh and counting occluded fingertips $\{0, \ldots, 5\}$. The cross-modal comparison (Figure 3) is reported on a balanced subsample drawn from this distribution to avoid bucket-size confounds.



Figure 19: Representative rendered frames stratified by fingertip-occlusion bucket. Each row corresponds to an occlusion level $\{0, \ldots, 5\}$; columns show example frames assigned to that bucket by the ray-cast procedure described above. The visualisation confirms that the bucket assignment tracks visible-fingertip count as intended.

C.4 TEMPORAL FILTERING OF FROZEN *HaMeR*

HaMeR is run one frame at a time in the main comparison. To give its output the same amount of temporal support as the EMG window, we use the 600 centre frames from Table 25 and take neighbouring poses from the same session. Frames are sampled every 0.5 s from -2.5 to $+2.5$ s, giving at most 11 images for each centre and 6,446 images after sequence boundaries are removed.

For each neighbourhood, we copy the camera, wrist transform, handedness, lighting, and texture from its centre frame. Only the articulated hand pose changes along the recorded sequence. We then apply ***HaMeR*** independently to every image. Before filtering, we subtract the predicted wrist position from the other joints and evaluate the filtered pose against the ground truth at the centre frame.

The filter settings are listed in Table 22. The unsmoothed centre prediction is included in the same selection.

Table 22: **Temporal-filter sweep for frozen *HaMeR*.** All windows are centred on one of the 600 control frames.

Filter	Settings
Boxcar mean	$h \in \{0.5, 1.0, 1.5, 2.0, 2.5\}$ s
Coordinate median	$h \in \{0.5, 1.0, 1.5, 2.0, 2.5\}$ s
Gaussian	$\sigma_t \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.75, 1.0, 1.5, 2.5\}$ s
Local polynomial	degrees 1–3; support ± 1 to ± 2.5 s
Bilateral	$\sigma_t \in \{0.3, 0.5, 0.75, 1.0, 1.5\}$ s; $\sigma_p \in \{3, 5, 10, 20, 40\}$ mm

For the bilateral filter, a neighbouring pose at offset t receives weight

$$w_t = \exp\left(-\frac{t^2}{2\sigma_t^2}\right) \exp\left(-\frac{d(\hat{P}_t, \hat{P}_0)^2}{2\sigma_p^2}\right),$$

where d is the RMS joint distance between the wrist-relative prediction at t and the prediction at the centre. The filtered centre pose is the weighted mean of these predictions.

For each of five session-disjoint folds, we choose the filter and its settings on 24 sessions and apply that choice to the remaining six. Four folds select the bilateral filter with σ_t=1.5 s and σ_p=5 mm. The other fold selects σ_t=0.3 s and σ_p=10 mm.

On the combined held-out frames, unsmoothed ***HaMeR*** gives 12.64 mm and the selected filter gives 12.60 mm. Their paired difference is -0.04 mm, with a session-bootstrap 95% confidence interval of $[-0.11, +0.01]$ mm. The boxcar results are 14.67 mm for ± 0.5 s and 15.72 mm for ± 2.5 s. **CycloFormer** gives 11.94 mm on the same centre frames.

The temporal experiment required the neighbouring frames to be rendered again. Its unsmoothed ***HaMeR*** value is therefore 12.64 mm rather than the 13.04 mm in Table 25. All conclusions above use paired predictions from the temporal render.

C.4.1 EgoEMG TRAINING AND VISION PROTOCOL

The EgoEMG numbers use one network: the 3.53M $40\times$ **CycloFormer** from Table 1, with the architecture left as is and the data recipe taken from ***EMGFormer-S*** (Xi et al., 2026). We do not retune widths or depth for eight channels. To run the same 3.53M $40\times$ stack on eight bipolar channels, we leave the shared CNN, Transformer width, and depth unchanged, retarget circular RoPE and spatial attention from $\mathbb{Z}_{16}$ to the eight-electrode ring, pool over those eight tokens, and replace the 20-D UmeTrack head with a 22-D MANO head (twenty joint angles plus wrist pitch and yaw).

Data. EgoEMG records eight bipolar channels per wrist at 2 kHz. The head predicts 22 numbers: twenty MANO joint angles plus wrist pitch and yaw. Splits are the EgoEMG ones: train, then user, gesture, and both.

Network and loss. The stem is still $40\times$, so a 10,000-sample window becomes T'=250 tokens. CNN widths stay $[32, 64, 128, 128]$, Transformer width d=224, feed-forward 512, eight heads,

$(N_{\text{spat}}, N_{\text{temp}}) = (3, 3)$, dropout 0.15, circular RoPE on the eight-electrode ring, and attention pooling into the 22-D pose. The only training noise is *EMGFormer-S*'s channel roll of -1, 0, or $+1$ electrode, applied in the dataloader. Loss is angle ℓ_1 with weight 1.0 plus a fingertip-distance term with weight 0.01.

Optimisation. This follows *EMGFormer-S*'s recipe: AdamW (Loshchilov & Hutter, 2019) at $5{\times}10^{-5}$, weight decay 10^{-2}, cosine over 200 epochs with $\eta_{\min} = 10^{-5}$, no warmup, gradient clip 1.0, and `bf16` mixed precision. The published *EMGFormer-S* batch is 500; we used microbatch 128 with four steps of accumulation (effective 512) so the run fitted on a 40 GB A100. The checkpoint in the paper is the val-MAE snapshot at epoch 191. On the protocol of Table 3 that checkpoint scores 17.67 mm mean PA-MPJPE.

Table 23: **`EgoEMG` training** for the 3.53M **CycloFormer** used in Section 4 and Table 3. Architecture is ours; window, stride, optimiser, and the ± 1 channel roll match *EMGFormer-S*.

Channels / outputs	8 / 22 (MANO 20 + wrist pitch, yaw)
Window / train stride / eval stride	12,000 / 1,200 / 12,000
Encoder rate	$40\times$ (T'=300)
AdamW	$5{\times}10^{-5}$, wd 10^{-2}, cosine 200 epochs, $\eta_{\min} = 10^{-5}$
Batch	$128{\times}4$ (effective 512)
Precision / clip	`bf16`, $\|g\|_2 \leq 1$
Selected checkpoint	val MAE, epoch 191

Vision protocol. We do not train a vision model. *HaMeR* (Pavlakos et al., 2024) and *HandOccNet* (Park et al., 2022) are the authors' published weights, frozen. **CycloFormer** is the `EgoEMG` checkpoint above.

Frames are the released $256{\times}256$ egocentric hand crops. We keep the first valid copy of each webcam index that has a generated label, is not stale, and has at least 15 mocap keypoints, then take every 12th such frame (5 Hz). A fingertip is counted occluded only when both the tip and the inner joint are hidden, by a ray test on the motion-capture mesh in the camera.

HaMeR reads one crop. The left crop is mirrored in and the predicted x is mirrored back, because *HaMeR* is a right-hand model. We take *HaMeR*'s MANO vertices at the 21 `EgoEMG` marker indices and score PA-MPJPE against mocap: surface markers versus the capture system. Crop jitter is an 85% window whose centre is shifted by 10% of the crop, then resized back; motion blur is a 9-pixel horizontal or vertical streak. Temporal smoothing (bilateral / moving average) is applied to *HaMeR* only.

D RENDERED EMG2POSE COMPARISON

`emg2pose` supplies synchronised sEMG and UmeTrack angles (Han et al., 2022) but no camera stream. We score *HaMeR* (Pavlakos et al., 2024) on rendered images of those poses and **CycloFormer** on the corresponding sEMG, using PA-MPJPE, the metric reported in vision baselines, against the number of occluded fingertips.

D.1 METHODOLOGY

We fit MANO (Romero et al., 2017) to each UmeTrack frame by Adam retargeting (Appendix C.2) and render the hand mesh alone (Appendix C.1). Camera rotation is Haar-uniform on $\mathrm{SO}(3)$ composed with a fixed palm prior, so occluded-fingertip count is a function of pose (Appendix C.1). *HaMeR* (Pavlakos et al., 2024) is a single-frame estimator and is applied to video frames independently. It reads that image from a ground-truth joint bounding box; **CycloFormer** reads the concurrent 5 s sEMG window. *HaMeR* is scored against the fitted MANO joints that were rendered; **CycloFormer** is scored against UmeTrack. Fingertip occlusion $\{0, \ldots, 5\}$ is obtained by ray-casting against the rendered mesh. We run *HaMeR* on FreiHAND (Zimmermann et al., 2019) photos and on our renders of the same poses, so the renderer can be scored on a dataset that has both. The renderer adds $+0.90$ mm PA-MPJPE (Figure 20; Appendix C.1).

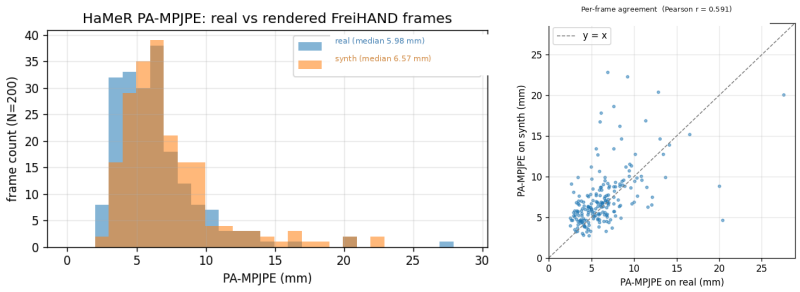

Figure 20: **Renderer check on FreiHAND.** Left: PA-MPJPE of *HaMeR* on real photos vs. our renders (N=200). Right: per-frame scatter; the diagonal is parity.

D.2 RESULTS

Occlusion profile. *HaMeR* on the official `emg2pose` test split (198 user-rows) versus **CycloFormer** is 8.30 ± 1.27 mm vs. 12.33 ± 1.87 mm PA-MPJPE Table 24 (medians 8.24 vs. 12.14). Most frames have few occluded fingertips. On the validation split, *HaMeR* PA-MPJPE increases with occluded-fingertip count (Figures 1 and 3). A 600-frame control with 100 frames per occlusion group reuses the same ids for both models. Session-cluster bootstrap intervals overlap at one, two, and three occluded fingertips and separate at four and five. *HaMeR* at zero versus five occluded fingertips is 11.10 mm vs. 16.14 mm; **CycloFormer** is 12.36 mm vs. 10.59 mm (Table 25). We take four or five occluded fingertips as the regime where this ordering is stable.

Table 24: ***HaMeR* vs. CycloFormer** on the `emg2pose` test split (198 user-rows). PA-MPJPE (mm).

Model	mean $\pm$ std	median	p_{75}	p_{95}
HaMeR	8.30 ± 1.27	8.24	9.15	10.49
CycloFormer	12.33 ± 1.87	12.14	13.36	15.73

Rendering conditions and *HandOccNet*. For each of the 600 control frames, we also render a new HDRI and texture, perturb the crop by 10% at 85% scale, and apply 9 px motion blur. These give *HaMeR* mean errors of 12.64, 12.86, and 14.17 mm, respectively (Table 25). We run *HandOccNet* (Park et al., 2022) on the baseline renders and obtain 17.44 mm mean PA-MPJPE. At five occluded fingertips, the four vision conditions range from 15.90 to 17.12 mm, whereas **CycloFormer**

is 10.59 mm. The ordering at high occlusion therefore remains the same after changing appearance, crop, blur, or the single-image vision model.

Table 25: **Bucket-balanced control** (100 frames × 6 occlusion buckets). PA-MPJPE (mm). Bold is lowest in column.

Condition	Method	0	1	2	3	4	5	Mean
Baseline	*HaMeR*	**11.10**	**12.74**	**12.68**	12.88	12.71	16.14	13.04
Baseline	**CycloFormer**	12.36	12.98	12.88	**11.91**	**10.92**	**10.59**	**11.94**
Baseline	*HandOccNet*	15.31	17.20	18.83	18.63	17.57	17.12	17.44
+ HDRI/texture	*HaMeR*	9.74	11.40	12.69	12.64	13.20	16.17	12.64
+ crop jitter	*HaMeR*	10.93	12.35	12.55	12.88	12.52	15.90	12.86
+ motion blur (9 px)	*HaMeR*	12.07	13.55	13.84	14.99	13.62	16.92	14.17

Hand configuration and contact. For each frame, we derive closed and open hand states from the labelled UmeTrack pose, thumb–fingertip contact from the distance between their labelled tips, and pinch/grasp groups from the `emg2pose` stage name. On closed hands, **CycloFormer** and *HaMeR* give 11.02 and 13.20 mm; on pinch/grasp stages they give 10.21 and 13.36 mm; and on thumb–fingertip contact they give 11.04 and 18.61 mm (Table 26). *HaMeR* is lower on open hands, at 11.23 mm against 12.71 mm by **CycloFormer**.

We also use CoffeePanicPete and PlayBlocksChess, the two `emg2pose` stages recorded while an object is held. The renderer has the hand mesh and not the object. When all occlusion groups are pooled, *HaMeR* gives 10.34 and 11.34 mm on these stages, compared with 12.44 and 12.95 mm for **CycloFormer**. After restricting the same stages to four occluded fingertips, the differences are $+2.39$ and $+1.82$ mm in favour of **CycloFormer**.

Table 26: **Contact and object subsets.** PA-MPJPE (mm). Δ = *HaMeR* − **CycloFormer**; positive favours sEMG. Pose-derived states use all frames in the subset; object stages are bucket-balanced (25 frames per bucket).

Hand state	*HaMeR*	**CycloFormer**	Δ	Δ@4	Δ@5
Closed hand	13.20	**11.02**	$+2.18$	$+1.27$	$+4.80$
Pinch/grasp-named stage	13.36	**10.21**	$+3.15$	$+4.78$	$+6.11$
Thumb–fingertip contact	18.61	**11.04**	$+7.57$	$+13.57$	$+7.73$
Open hand	11.23	12.71	-1.48	$+2.39$	—
Object1 (CoffeePanicPete)	**10.34**	12.44	-2.09	$+2.39$	$+3.31$
Object2 (PlayBlocksChess)	**11.34**	12.95	-1.61	$+1.82$	$+2.36$

Temporal context. Finally, Pavlakos et al. (2024) process video one frame at a time. We retain that exact protocol in the main comparison and separately test whether filtering their predictions over the 5 s available to **CycloFormer** changes the result. For each of the 600 control frames, we render neighbouring poses over ± 2.5 s, apply *HaMeR* independently, and select a filter by five-fold session-disjoint cross-validation (Appendix C.4). The selected filter changes PA-MPJPE from 12.64 to 12.60 mm, a difference of -0.04 mm with session-bootstrap 95% CI $[-0.11, +0.01]$ mm. Uniform averaging over ± 2.5 s instead gives 15.72 mm. On these centre frames, **CycloFormer** is 11.94 mm.

E DATA-SCARCE, MULTI-EPOCH SCALING LAWS

E.1 SCALING-LAW FIT DETAILS

Optimisation. The fit uses `scipy.optimize.differential_evolution` (seed 42, `maxiter`= 2×10^5, `popsize`= 40, polished by a local minimisation step) with per-parameter bounds $\varepsilon_\infty \in [0, L_{\min}]$, $\{a, b\} \in [0.01, 5L_{\max}]$, $\{\alpha, \beta\} \in [10^{-3}, 10]$, $c \in [10^{-6}, L_{\max}]$, $\gamma \in [0.05, 3]$. We add an L_2 regulariser $\lambda\,\varepsilon_\infty^2$ to the soft-ℓ_1 residual loss with $\lambda = 10^{-4}$. The grid spans observed losses in only the 12.4–17.3 range, so the asymptotic floor ε_∞ is statistically under-identified: the optimiser can trade ε_∞ against a shallow α in the $a\,N^{-\alpha}$ term to absorb the constant offset (different starting points produce values from $\varepsilon_\infty \approx 0$ to $\varepsilon_\infty \approx 7.6$ mm with near-identical RSS). The regulariser breaks this degeneracy by pulling $\varepsilon_\infty \to 0$, forcing the model-size and data terms to do the floor-explaining; the resulting fit costs only $\sim$4% in in-sample RMSE versus the $\lambda = 0$ optimum.

Table 27: Empirical **PA-MPJPE** (mm) on the `emg2pose` test set, on the 5×8 (N, D) grid used to fit Equation (1) (32 of 40 cells in the table; "—" = not trained). **Bold** marks cells where loss *increases* relative to the running minimum at smaller N for the same D.

Data \ Params	1.49M	4.18M	6.72M	9.92M	16.77M	21.70M	48.00M	84.77M
20%	16.58	16.13	15.95	15.61	**15.82**	**15.65**	—	**17.28**
25%	16.43	15.78	15.37	**15.50**	15.26	15.19	**15.43**	—
50%	15.34	14.61	14.43	14.15	**14.31**	—	—	—
75%	14.81	13.99	13.70	13.55	13.33	13.30	—	—
100%	14.71	13.57	**13.67**	—	13.13	13.02	12.74	**12.86**

Table 28: **Fitted scaling-law parameters on `emg2pose` (full test, 5×8 grid; 31 of 40 cells used to fit, with $(N{=}85\text{M},\ D{=}100\%)$ held out).** Equation (1) (Ours) substantially reduces in-sample RMSE over Chinchilla: PA-MPJPE $0.355 \to 0.127$ mm (-64.7%). Held-out cell $(N{=}85\text{M},\ D{=}100\%)$ prediction error: $+0.21$ mm PA-MPJPE.

Model	ε_∞	a	α	b	β	c	γ	R^2
Chinchilla ($c{=}0$)	0.08	1.97	0.69	12.72	0.12	—	—	0.906
Ours (Equation (1))	0.11	12.47	0.06	2.22	0.42	6.7×10^{-3}	1.01	0.988

Master-curve rescaling. If the interaction term governs the loss, the same combination should determine every data-fraction curve, just shifted in N. Define the rescaled axis $u = N/D^p$ with $p = \gamma/(\gamma+\alpha)$ (the same exponent that sets the U-minimum location). Substituting $N = u\,D^p$ into the two N-dependent terms of Equation (1):

$$a\,N^{-\alpha} = a\,u^{-\alpha}\,D^{-\alpha p},$$
$$c\,(N/D)^{\gamma} = c\,u^{\gamma}\,D^{-\gamma(1-p)}.$$

The two D-prefactors coincide: $\alpha p = \alpha\gamma/(\gamma+\alpha) = \gamma(1-p)$. Pulling out the common $D^{-\alpha p}$ leaves

$$L(N, D) - \varepsilon_\infty - b\,D^{-\beta} = D^{-\alpha p}\big[a\,u^{-\alpha} + c\,u^{\gamma}\big].$$

The bracket is a function of u alone. Figure 21 plots the floor-subtracted residual against u with the PA-MPJPE fit $\alpha{=}0.056$, $\gamma{=}1.01$, so $p{=}\gamma/(\gamma+\alpha){\approx}0.95$. The remaining $D^{-\alpha p}$ prefactor is omitted in that plot.

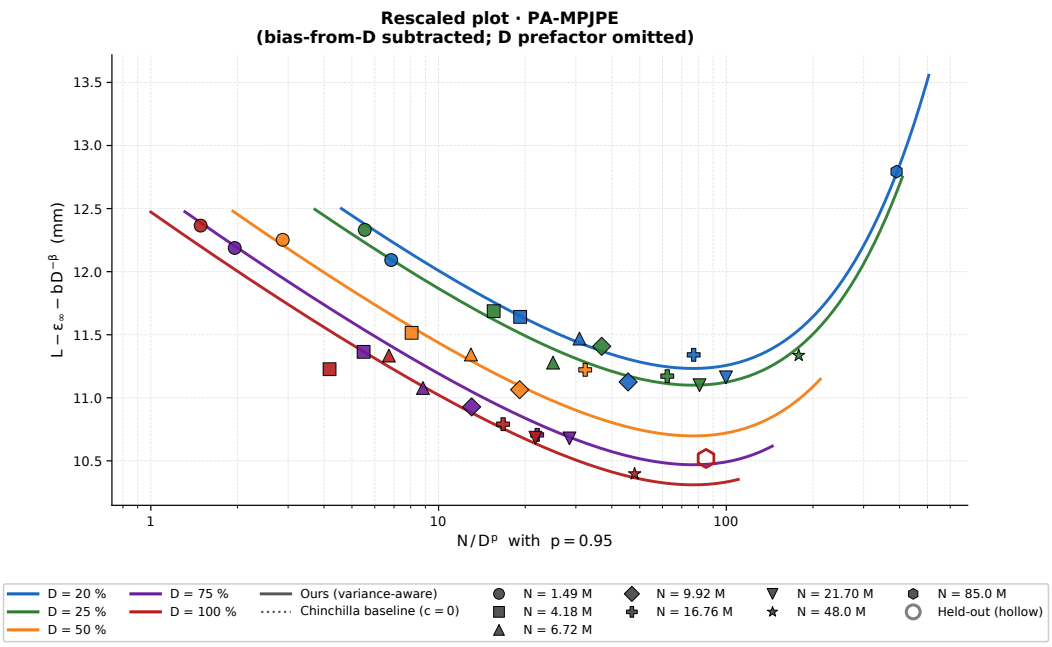

Figure 21: **Rescaled plot of Equation (1) on `emg2pose` (PA-MPJPE).** Floor-subtracted residual $L - \varepsilon_\infty - b\,D^{-\beta}$ against $u = N/D^{\hat{\gamma}/(\hat{\gamma}+\hat{\alpha})}$ with the fitted $\hat{p} = \hat{\gamma}/(\hat{\gamma}+\hat{\alpha}) \approx 0.95$ from Table 28 (α=0.056, γ=1.01). Hollow marker is the held-out (N=84.77M, D=100%) run.

F Ethical and Societal Considerations

Wrist sEMG decoding has substantial potential benefits: it enables robust hand-pose supervision in the grasping regime where vision fails, provides force-correlated signals absent from egocentric video, and can serve as a direct neuromotor interface for adaptive controllers that assist users with limited mobility or dexterity.

On the negative side, raw sEMG encodes individual biometric characteristics such as forearm anatomy, muscle activation patterns, and hand kinematics that can be used to identify users. A deployed decoder could be repurposed for covert identification or monitoring of hand activity without consent. Additionally, if training corpora are demographically narrow, performance may degrade for underrepresented users, a risk that grows in assistive and prosthetic contexts. Deployments should apply encryption, data minimisation, and access controls.