STAMP: TASK-STRUCTURED TRANSFORMERS FOR STREAMING SEMG-TO-POSE

Anonymous authors
Paper under double-blind review

ABSTRACT

Streaming hand-pose estimation from surface electromyography (sEMG) requires causal reconstruction from noisy, non-stationary signals. We propose STAMP, a task-structured Transformer that organizes electrode interactions, temporal evidence, pose evolution, and motion-dependent computation. SpatioTemporal Electrode Modeling (STEM) preserves the electrode axis and combines spatial softmax attention with temporally decayed causal linear attention, informed by sensing geometry. Affine Scan provides compact, parallelizable pose updates through gated propagation and correction, while history-conditioned Motion MoE adapts computation across timescales. On emg2pose, STAMP achieves the lowest cross-user mean angular and whole-hand landmark errors in every Regression and Tracking split among matched retrained implementations. Controlled replacements, including near-parameter-matched alternatives, support structured electrode modeling, historical expert selection, and complementary pose propagation and correction. At nearly matched capacity, STAMP has lower errors and steeper fitted scaling than vEMG2Pose over 32–158 training users, outperforming the 158-user baseline with only 95 users. Increasing capacity generally yields further gains within the measured grid. Dataset-specific retraining with interface-only adaptation also yields the lowest pooled errors on EgoEMG and NinaPro, supporting architectural portability across electrode layouts.

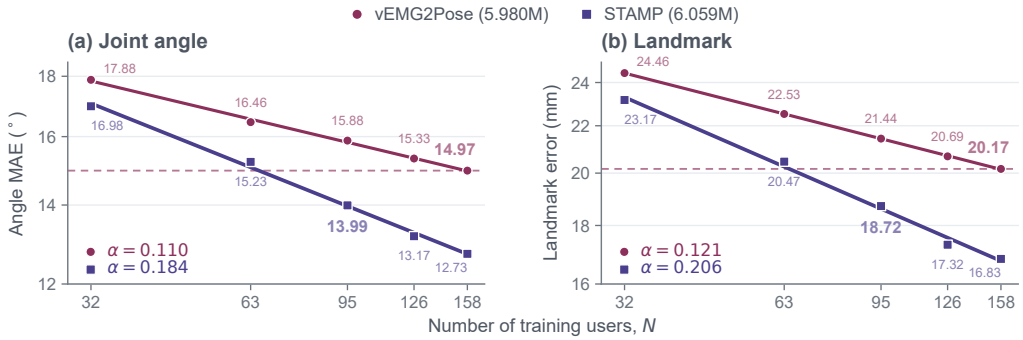


Figure 1: **Lower errors and steeper user scaling at nearly matched capacity.** Lines fitted under $E(N) = AN^{-\alpha}$; protocols and full results are provided in Appendix C.

1 INTRODUCTION

Streaming wrist sEMG-to-pose estimation imposes stricter requirements than offline sequence regression. At every step, the model must infer an absolute hand pose from an indirect, noisy muscle signal, update that estimate without future observations, and remain stable when anatomy, electrode contact, placement, or motion changes the input distribution. Population-trained sEMG models suggest that a shared decoder can generalize across people and recording conditions (Kaifosh et al., 2025). This setting motivates task structure alongside larger training sets and model capacity.

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.

The prevailing causal recipe combines local feature extraction with recurrent state, as in vEMG2Pose (a TDS encoder with an LSTM decoder) (Hannun et al., 2019; Hochreiter & Schmidhuber, 1997; Salter et al., 2024), or replaces recurrence with attention for flexible temporal interaction (Vaswani et al., 2017). These effective sequence mechanisms do not by themselves specify how to preserve electrode structure while aggregating history, which quantity should persist as the task state, or which history should control specialization. Consequently, the model must recover sensing relations, pose dynamics, and motion-regime structure implicitly from finite and shifted training data.

For sEMG-to-pose, these omissions create three concrete mismatches. First, the electrode and time axes impose different requirements: a small sensor array calls for cross-electrode interaction, whereas a growing signal history requires causal temporal aggregation. Wrist electrodes also form a ring, whose first and last channels are neighbors; placement shifts alter the observed channel pattern (Ameri et al., 2020). The representation should retain these axes and expose sensing geometry. Second, muscle activity is strongly informative about change, whereas the output is an absolute pose. Independent position proposals can ignore persistence; integrating increments preserves motion but accumulates drift. Third, anatomy, contact, session, action, and movement speed alter the sEMG–pose relationship (Farina et al., 2014; Côté-Allard et al., 2019; Lehmler et al., 2022; Yang et al., 2024). A fixed computation path must compromise across static, transitional, slow, and fast regimes. The design therefore structures perception along the electrode and time axes, makes pose propagation and correction the recurrent update, and conditions computation on causal motion history.

STAMP implements this design in a streaming Transformer (Figure 2). Its SpatioTemporal Electrode Modeling (STEM) encoder supplies causal features to a two-stage Affine Scan decoder, with Motion MoE providing history-conditioned expert computation. On emg2pose, STAMP leads every five-second Regression and Tracking split among matched retrained implementations. Near-parameter-matched encoder and non-MoE alternatives, together with geometry, routing-history, and decoding-path replacements, test the design choices. User scaling tests whether the complete architecture's advantage persists as training data grow: at nearly matched capacity, STAMP has lower errors at every tested user count and improves faster over the observed interval (Figure 1). Dataset-specific retraining with interface adaptation tests portability on EgoEMG and NinaPro.

Our contributions are:

- **STEM for causal spatiotemporal perception.** An electrode-preserving encoder stacks STEM blocks combining spatial softmax attention with temporally decayed causal linear attention, conditioned on input sensing-layout codes.
- **Parallel pose-space propagation and correction.** Affine Scan combines pose increments and absolute proposals in an associative update, enabling parallel-prefix training and constant-state streaming.
- **History-conditioned motion specialization.** A router conditions residual expert selection on motion summaries at multiple timescales.

2 Related Work

sEMG variability and generalization. sEMG reflects neural activation, but anatomy and the measurement interface also shape the observed signal (Farina et al., 2014). NinaPro established a multi-subject resource spanning sEMG, hand movements, and glove kinematics (Atzori et al., 2014); transfer, benchmark, and electrode-displacement studies further document cross-user, temporal, and placement shifts (Côté-Allard et al., 2019; Lehmler et al., 2022; Yang et al., 2024; Ameri et al., 2020). This literature motivates evaluation beyond the training users and conditions. Much of it studies gesture recognition or adaptation as the primary problem. In continuous causal reconstruction, distribution shift also interacts with trajectory state and error accumulation.

Continuous sEMG-to-pose estimation. Continuous kinematics models reconstruct a fine-grained trajectory rather than assign one gesture label. CNN–attention and high-density sEMG systems estimate finger kinematics (Geng et al., 2022; Guo et al., 2023; Sîmpetru et al., 2024); NeuroPose studies wrist-based tracking (Liu et al., 2021); the emg2pose benchmark formalizes user and gesture-stage shifts for wrist sEMG and introduces the velocity-integrating vEMG2Pose baseline (Salter et al.,

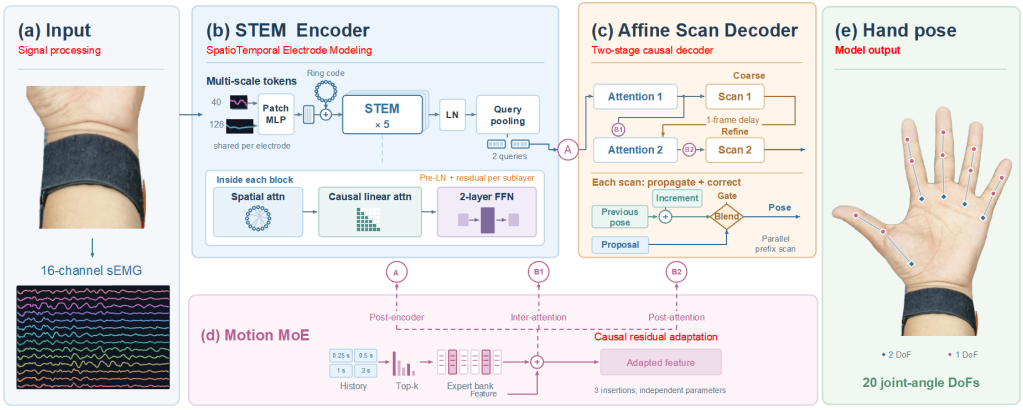


Figure 2: **STAMP introduces task structure at three levels.** The encoder embeds multi-scale per-electrode tokens with a ring code, repeats five STEM blocks of spatial attention, causal linear attention, and feed-forward transformation, then pools with two learned queries. The two-stage causal decoder delays the coarse path by one frame before refinement; each Affine Scan propagates and corrects the pose state. Independently parameterized Motion MoE banks provide causal residual adaptation at three insertion sites. Every deployed dependency is present- or past-only.

2024); and EgoEMG introduces the windowed EMGFormer family (Xi et al., 2026). Hadidi et al. revisit position–velocity decoding and report the previous best published result among streaming-compatible methods on emg2pose (Hadidi et al., 2026). STAMP explicitly represents sensing topology, propagated output state, and motion-dependent computation within one causal estimator.

Calibration-dependent REACT (Xie & Cheung, 2026), tendon-control and direct-pose CLDM (Verma, 2025; Micropilot, 2026), and bidirectional discrete-token VQ-MyoPose (Lovecchio et al., 2025) address closely related settings under different inputs or evaluation protocols. Appendix B.1 summarizes these protocols.

Structured causal sequence models. Transformers provide flexible interaction but full attention is quadratic (Vaswani et al., 2017). TimeSformer separates spatial and temporal attention (Bertasius et al., 2021), while linear attention provides a recurrent formulation for causal sequence processing (Katharopoulos et al., 2020). STEM builds on these primitives: it retains the electrode axis, uses spatial softmax attention over the small sensor array, and aggregates growing histories with decayed causal linear attention. Input geometry codes provide layout information. State-space models and retention offer other recurrent representations (Gu et al., 2022; Gu & Dao, 2024; Sun et al., 2023); parallel-prefix algorithms and simplified recurrent gates support parallel training (Blelloch, 1990; Martin & Cundy, 2018; Feng et al., 2024). STAMP's contribution is the task-specific organization and evaluation of these mechanisms, with Affine Scan maintaining predicted pose as its task state.

3 STAMP

3.1 Problem Setting and Overview

The task maps multichannel sEMG to a continuous joint-angle trajectory. emg2pose provides 16-channel wrist sEMG at 2 kHz and 20-angle targets; STAMP emits native predictions at 50 Hz.

Regression and Tracking are evaluated separately. Regression receives no true initial pose, whereas Tracking receives the true pose once at sequence start and never refreshes it. At every later step, the model may use only the declared initialization and the sEMG samples already acquired. Under the emg2pose interface, the model consumes a 1,790-sample start-up context and then emits one pose after each new 40-sample block, corresponding to a 20 ms native update interval. Appendix A gives the formal causal definition.

Encoder tokens pass through two causal decoder stages, with the first stage's one-frame-delayed pose trajectory conditioning the second (Figure 2). Three independently parameterized Motion MoE banks are placed after the encoder, between decoder attention stages, and after the final attention stage. Appendix A specifies the operators and streaming states.

3.2 Spatiotemporal electrode modeling with STEM

Shared multi-scale patch encoders produce $\mathbf{E} \in \mathbb{R}^{T\times C\times d}$, retaining time and electrode axes. Geometry codes $\mathbf{R}$ are broadcast across time and added once before the block stack; circular Fourier codes specify wrist-electrode positions. Each STEM block applies spatial attention, causal temporal attention, and a feedforward transformation with separate pre-normalization and residual connections:

$$\begin{aligned}\mathbf{H}^{(0)} &= \mathbf{E} + \mathbf{R},\\ \mathbf{S}^{(\ell)} &= \mathbf{H}^{(\ell)} + \mathrm{SA}_C\big(\mathrm{LN}_s(\mathbf{H}^{(\ell)})\big),\\ \mathbf{U}^{(\ell)} &= \mathbf{S}^{(\ell)} + \mathrm{TLA}_T\big(\mathrm{LN}_t(\mathbf{S}^{(\ell)})\big),\\ \mathbf{H}^{(\ell+1)} &= \mathbf{U}^{(\ell)} + \mathrm{FFN}\big(\mathrm{LN}_f(\mathbf{U}^{(\ell)})\big).\end{aligned} \tag{1}$$

Dropout is omitted from Equation 1. SA_C uses softmax attention over the small electrode array at each time. TLA_T applies decayed causal linear attention along each spatially mixed electrode stream, with per-head decay learned during training and fixed at inference. Prefix states summarize growing histories; because streams have already exchanged spatial information, temporal aggregation also carries cross-electrode evidence.

Five blocks and a final layer normalization form the STEM encoder; downstream learned queries pool electrode features into pose-decoder tokens. Circular input codes give the ring-aware configuration, while NinaPro uses mixed-layout codes. The codes supply layout context while spatial interactions remain learned. Appendix A.2 specifies the geometry codes and temporal recurrence.

3.3 Pose-space propagation and correction

At step t, the decoder predicts an absolute pose proposal $\mathbf{p}_t$, an angular increment $\boldsymbol{\Delta}_t$, and a per-joint sigmoid gate $\mathbf{a}_t$. It updates the pose directly as

$$\widehat{\mathbf{y}}_t = \mathbf{a}_t \odot (\widehat{\mathbf{y}}_{t-1} + \boldsymbol{\Delta}_t) + (1 - \mathbf{a}_t) \odot \mathbf{p}_t. \tag{2}$$

The gate balances persistence against observation-driven correction. Writing $\mathbf{b}_t = \mathbf{a}_t \odot \boldsymbol{\Delta}_t + (1 - \mathbf{a}_t) \odot \mathbf{p}_t$ turns Equation 2 into the affine map $\widehat{\mathbf{y}}_t = \mathbf{a}_t \odot \widehat{\mathbf{y}}_{t-1} + \mathbf{b}_t$. Two consecutive maps compose as

$$(\mathbf{a}_2, \mathbf{b}_2) \circ (\mathbf{a}_1, \mathbf{b}_1) = \big(\mathbf{a}_2 \odot \mathbf{a}_1,\ \mathbf{b}_2 + \mathbf{a}_2 \odot \mathbf{b}_1\big). \tag{3}$$

Associativity enables parallel prefix evaluation within each decoder stage, with update coefficients computed before scanning. The first stage produces a coarse pose trajectory; the second conditions on its one-step-delayed poses and pose differences to refine the prediction causally. Online inference applies the same affine recurrence once per stage and time step, maintaining pose states alongside decoder caches. Appendix A.4 provides the complete algorithm and initialization details.

3.4 History-conditioned specialization and streaming

The same temporal context is not equally useful for a rapid finger transition and a slowly held posture. For each decay scale m, the router maintains the causal summary

$$\begin{aligned}\mathbf{q}_t^{(m)} &= \lambda_m \mathbf{q}_{t-1}^{(m)} + (1 - \lambda_m)\psi(\mathbf{u}_t),\\ \widetilde{\mathbf{u}}_t^{\ell} &= \mathbf{u}_t^{\ell} + \sum_{e\in \mathrm{TopK}(\boldsymbol{\pi}_t^{\ell},8)} \pi_{t,e}^{\ell} E_e^{\ell}(\mathbf{u}_t^{\ell}).\end{aligned} \tag{4}$$

The concatenated summaries, spanning 0.25–2 seconds, determine the routing weights $\boldsymbol{\pi}_t^{\ell}$; each bank activates eight of 32 residual experts. Short summaries respond to transitions, while longer summaries retain slower context. During streaming inference, the encoder prefix state, decoder caches, pose states, and router summaries advance once per signal block, with no recomputation of the preceding sequence. Appendix A specifies the expert banks and streaming states; Appendix H details the causality checks.

Table 1: Five-second results on emg2pose and EgoEMG. Each cell reports mean ± standard deviation for joint-angle MAE (°) / whole-hand landmark distance (mm); lower is better. EgoEMG follows the released protocol; an additional evaluation with audited splits is reported in Appendix B.4.

Model	Params (M)	User (°/mm) ↓	Stage (°/mm) ↓	User+Stage (°/mm) ↓	Overall test (°/mm) ↓
(a) *emg2pose — Regression*					
EMGFormer-M†	6.639	12.02±1.35/15.47±1.82	14.16±1.72/19.07±2.28	15.23±1.39/20.81±1.99	13.99±1.74/18.79±2.37
NeuroPose†	6.355	13.48±1.04/17.85±1.41	18.17±1.67/24.85±2.04	18.29±1.24/25.45±1.37	17.75±1.99/24.24±2.62
vEMG2Pose†	5.980	12.35±1.26/16.02±1.84	15.23±1.57/20.59±2.10	15.69±1.32/21.65±1.75	14.98±1.69/20.20±2.33
vEMG2Pose† (Multi-task)	5.980	11.94±1.19/15.39±1.68	14.83±1.59/19.86±2.15	15.19±1.32/20.83±1.89	14.58±1.70/19.48±2.35
STAMP	6.059	**11.65±1.33/14.65±1.64**	**12.52±1.55/16.57±2.02**	**14.28±1.39/19.27±2.01**	**12.48±1.53/16.46±2.00**
(b) *emg2pose — Tracking*					
EMGFormer-M†	6.639	8.37±1.15/11.22±1.88	11.95±1.58/16.05±2.00	11.71±1.31/16.12±1.54	11.62±1.79/15.62±2.31
NeuroPose†	6.355	9.67±1.44/13.65±2.20	14.01±1.77/19.69±2.21	13.44±1.56/19.34±1.61	13.61±2.07/19.14±2.67
vEMG2Pose†	5.980	7.73±0.95/10.29±1.53	11.17±1.43/15.10±1.81	11.06±1.07/15.44±1.52	10.86±1.64/14.68±2.14
STAMP	6.059	**7.30±0.82/9.47±1.29**	**10.04±1.29/13.44±1.64**	**10.16±0.87/13.94±1.18**	**9.80±1.42/13.10±1.87**
(c) *EgoEMG — Regression*					
EMGFormer-M†	6.617	17.80±1.04/29.52±1.72	13.03±1.59/21.21±2.59	18.06±1.51/29.59±2.18	13.67±2.16/22.29±3.63
NeuroPose†	6.355	**16.36±0.61/27.57±2.00**	16.18±1.17/26.83±2.23	**16.60±0.86/28.05±1.82**	16.25±1.09/26.94±2.22
vEMG2Pose†	5.964	17.51±0.74/29.76±1.84	14.53±1.39/23.44±2.42	17.69±1.19/29.07±1.57	14.95±1.59/24.26±3.03
STAMP	6.060	17.91±1.16/29.27±2.57	**12.22±1.78/19.47±3.13**	18.25±1.24/29.80±1.93	**12.96±2.52/20.70±4.41**
(d) *EgoEMG — Tracking*					
EMGFormer-M†	6.617	10.22±0.54/16.51±1.14	10.11±2.05/16.59±3.41	11.04±0.42/17.82±0.84	10.18±1.92/16.66±3.22
NeuroPose†	6.355	10.19±0.61/16.44±1.15	10.22±2.06/16.81±3.42	11.04±0.41/17.90±1.18	10.28±1.94/16.84±3.25
vEMG2Pose†	5.964	10.89±0.35/17.45±0.75	11.49±1.62/18.47±2.65	12.45±0.77/19.93±1.29	11.49±1.52/18.45±2.52
STAMP	6.060	**9.68±0.45/15.77±0.81**	**9.74±1.76/15.98±2.87**	**10.40±0.37/16.88±0.94**	**9.79±1.65/16.03±2.70**

Sources: EMGFormer-M: Xi et al. (2026); NeuroPose: Liu et al. (2021); vEMG2Pose: Salter et al. (2024); multi-task training: Hadidi et al. (2026).
† Retrained baselines; implementations and dataset-specific protocols are detailed in Appendix B.

4 EXPERIMENTS

4.1 EVALUATION PROTOCOL

We evaluate on emg2pose, EgoEMG, and NinaPro under dataset-specific protocols, using validation-selected checkpoints. Table 1 reports five-second Regression and Tracking on the wristband datasets emg2pose and EgoEMG, with joint-angle MAE and whole-hand landmark distance for User, Stage, User+Stage, and Overall. Within each split, errors are aggregated per user before computing the cross-user mean ± population SD. Overall first pools each user's eligible windows across conditions, then equally weights unique users: 178 for emg2pose and 40 for released-protocol EgoEMG.

Tracking receives the true pose only at initialization; Tracking on EgoEMG is an author-defined extension. NinaPro is reported separately in Table 6 (Appendix B.2) as a mixed-layout portability evaluation, using Regression-only angle MAE with 1.095-second inputs and 0.2-second scored intervals. Baseline sources and retraining are identified in the result tables.

The Multi-task vEMG2Pose variant is jointly trained on Regression and Tracking and evaluated on Regression; the remaining emg2pose vEMG2Pose rows use task-specific single-task checkpoints. Benchmark-grid accuracy and native-stream causality are assessed separately; Appendix B details data splits, training, adaptation, and output alignment.

4.2 ACCURACY ACROSS TASKS AND SENSING INTERFACES

Under the released emg2pose protocol, STAMP has the lowest cross-user mean angle and landmark error in every five-second Regression and Tracking split among matched retrained implementations, including single-task and Multi-task vEMG2Pose. Regression margins over the strongest comparator are larger on Stage and smaller on User, suggesting greater gains under movement-stage shift. On EgoEMG, STAMP lowers Overall Regression angle/landmark error by $5.23\%/7.14\%$ relative to EMGFormer-M and leads every aligned Tracking split, reaching $9.79°/16.03$ mm Overall error.

To test sensitivity to future-neighbor interpolation in benchmark scoring, we re-evaluate nine affected model–task configurations with frozen checkpoints, replacing feature interpolation in vEMG2Pose with past-only selection and output interpolation in both vEMG2Pose and STAMP

with delayed zero-order hold. Across 36 conditions, angle MAE and landmark distance change by at most $0.122°$ and 0.193 mm, respectively; STAMP remains more accurate than single-task vEMG2Pose on both metrics in all four emg2pose Tracking conditions (Appendix B.5).

The NinaPro evaluation demonstrates architectural portability beyond a pure wristband array. With dataset-specific retraining, we adapt the input geometry encoding and output head to eight forearm-ring electrodes, four anatomically placed channels, and 22 target angles while retaining the STEM blocks, Affine Scan, and Motion MoE. In Table 6, STAMP attains the lowest pooled MAE at $16.83°$, 4.13% below NeuroPose, and the lowest Stage error at $16.20°$.

4.3 Ablation studies

Table 2 evaluates cumulative replacements starting from vEMG2Pose: replacing the LSTM decoder with Affine Scan, replacing the encoder with STEM, and adding Motion MoE. Affine Scan lowers seven of eight reported errors while reducing parameters by 53.9%; the STEM replacement and Motion MoE each lower all eight. The encoder comparison covers the complete patch representation, input geometry, and STEM block stack.

Table 2: Module-level cumulative ablation on five-second emg2pose Regression. Values are mean $\pm$ population standard deviation across users for Angle (°) / Landmark (mm), both minimized; rows introduce Affine Scan, the complete STEM encoder, and Motion MoE in sequence.

Configuration	Params (M) ↓	User (°/mm) ↓	Stage (°/mm) ↓	User+Stage (°/mm) ↓	Overall test (°/mm) ↓
vEMG2Pose†	5.980	$12.35_{\pm 1.26}/16.02_{\pm 1.84}$	$15.23_{\pm 1.57}/20.59_{\pm 2.10}$	$15.69_{\pm 1.32}/21.65_{\pm 1.75}$	$14.98_{\pm 1.69}/20.20_{\pm 2.33}$
TDS–Affine	**2.754**	$12.30_{\pm 1.31}/16.09_{\pm 1.90}$	$15.09_{\pm 1.60}/20.44_{\pm 2.14}$	$15.50_{\pm 1.29}/21.41_{\pm 1.77}$	$14.85_{\pm 1.70}/20.07_{\pm 2.34}$
STEM–Affine	4.433	$11.79_{\pm 1.08}/15.09_{\pm 1.40}$	$13.98_{\pm 1.49}/18.64_{\pm 1.96}$	$14.83_{\pm 1.33}/20.22_{\pm 1.91}$	$13.80_{\pm 1.53}/18.36_{\pm 2.07}$
STAMP	6.059	$\mathbf{11.65_{\pm 1.33}/14.65_{\pm 1.64}}$	$\mathbf{12.52_{\pm 1.55}/16.57_{\pm 2.02}}$	$\mathbf{14.28_{\pm 1.39}/19.27_{\pm 2.01}}$	$\mathbf{12.48_{\pm 1.53}/16.46_{\pm 2.00}}$

† Baselines retrained using the dataset-specific implementations and evaluation protocols in Appendix B.

Table 3 tests six replacements within full STAMP under the same training, validation-selection, and evaluation protocol. A temporal Transformer replaces the STEM encoder while retaining Affine Scan and Motion MoE; linear sinusoidal positions replace circular encoding; and the wider-backbone control removes Motion MoE and widens the encoder. These three controls are nearly parameter-matched to full STAMP. Current-only routing removes historical summaries but retains the temporal encoder. Position-only decoding outputs absolute pose proposals; increment-only decoding integrates learned pose increments after initialization, without ongoing absolute-pose correction.

Table 3: Replacements relative to full STAMP on five-second emg2pose Regression. Overall pools windows within each user before reporting the cross-user mean $\pm$ population SD. Split-wise results for the first three replacements appear in Table 22.

Replacement	Params (M)	Angle MAE (°) ↓	Landmark distance (mm) ↓
STAMP (full model)	6.059	$\mathbf{12.48_{\pm 1.53}}$	$\mathbf{16.46_{\pm 2.00}}$
STEM encoder → Temporal Transformer	6.060	$14.06_{\pm 1.71}$	$18.89_{\pm 2.33}$
Circular encoding → Linear sinusoidal	6.059	$12.91_{\pm 1.53}$	$17.00_{\pm 2.03}$
Motion MoE → Wider backbone	6.056	$13.54_{\pm 1.51}$	$17.95_{\pm 2.03}$
History-conditioned routing → Current-only routing	6.059	$13.05_{\pm 1.52}$	$17.18_{\pm 2.04}$
Affine Scan → Position-only	6.059	$13.21_{\pm 1.54}$	$17.45_{\pm 2.04}$
Affine Scan → Increment-only	6.059	$13.48_{\pm 1.48}$	$17.93_{\pm 1.97}$

STEM outperforms the nearly parameter-matched temporal Transformer, and Motion MoE outperforms the wider encoder, supporting their organization beyond parameter count. Circular encoding also improves on linear sinusoidal positions at equal capacity. For all three controls, both errors increase most on Stage, linking the differences primarily to movement-stage generalization (Appendix F.2, Table 22). Current-only routing raises angle/landmark error by $4.50\%/4.39\%$, supporting historical expert selection. Position-only and increment-only decoding raise errors by

5.78%/6.04% and 7.96%/8.96%, respectively; the latter degrades more among the two decoder replacements, supporting propagation with ongoing correction.

4.4 MECHANISM ANALYSIS

Figure 3 shows spatial maps progressing from circular neighborhoods to cross-electrode mixing and global integration with two opposing concentrations. Temporal attention broadens in middle blocks before refocusing on recent evidence (statistics in Appendix G.2). The attention maps characterize STEM's learned organization, while the replacement studies provide the corresponding accuracy evidence.

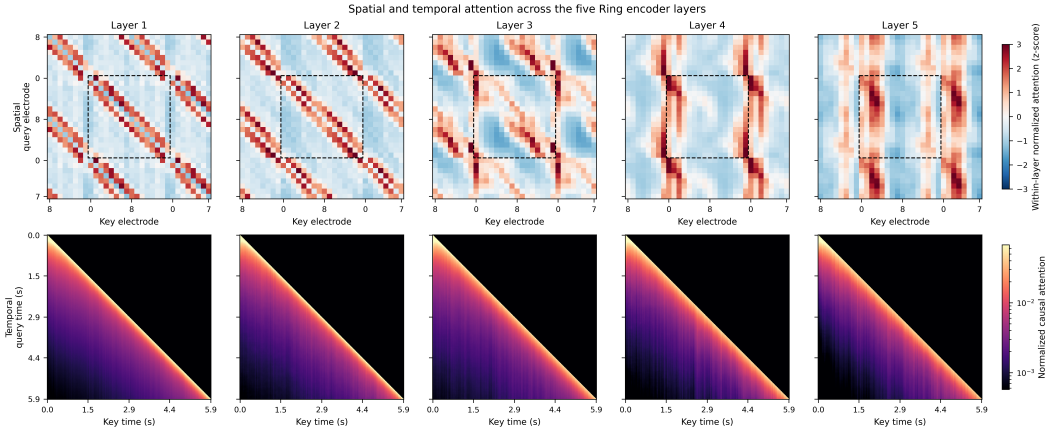


Figure 3: **Spatial and temporal attention across the five STEM blocks.** Dashed boxes mark the original 16-electrode matrices, exterior bands are circular display copies, and black temporal triangles denote excluded future positions.

With subsequent commands fixed, the learned update halves an injected 10° state disturbance within 0.12 s and removes 90% within 0.44 s (Figure 4; protocol in Appendix G.3). Forced retention preserves the error; forced reset corrects it immediately but discards persistence.



Figure 4: **Affine Scan under controlled state perturbations.** (A) Open-loop recovery with fixed subsequent commands. (B,C) Closed-loop changes in the gate and final output relative to unperturbed trajectories. Shading denotes bootstrap uncertainty across 24 windows.

The routing analysis tests association with finger-motion direction (Figure 5). Circularly shifting direction labels within each window preserves temporal autocorrelation while breaking direction–routing alignment. All three routers exceed the structured-null 95th percentile ($p = 0.002$ each), with conditional normalized mutual information of 0.184, 0.196, and 0.170 (Appendix G.4).

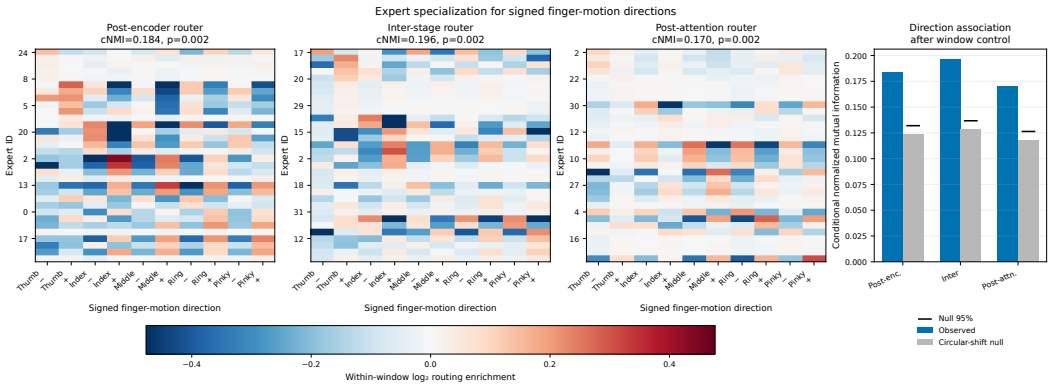


Figure 5: **Motion-expert specialization for ten signed finger directions.** Heatmaps show within-window routing enrichment. The final panel compares observed conditional normalized mutual information with circular-shift nulls; black ticks mark the null 95th percentiles.

4.5 Scaling with data and model capacity

The user-scaling experiment tests how the complete architecture benefits from additional training data (Figure 1). Fixed nested subsets of 32–158 users retain all eligible recordings with validation and test sets unchanged, increasing both user diversity and data volume. These runs use independently trained checkpoints, including at full data, rather than the main-table checkpoints. STAMP's relative angle/landmark error reductions over vEMG2Pose grow from $5.03\%/5.27\%$ at 32 users to $14.99\%/16.57\%$ at 158 users. Within this interval, fitted reductions per doubling are $11.98\%/13.32\%$, versus $7.34\%/8.06\%$ for vEMG2Pose.

The data–capacity grid tests how data and capacity interact within STAMP (Figure 6). With all training users, the 1.37M-parameter XS model reaches $14.24°/18.98$ mm, below the 15.90M-parameter Large model's $15.78°/21.41$ mm at 20% data, identifying a data-limited regime in which additional capacity does not compensate for reduced data. At 20% data, Small-to-Middle scaling raises angular error from $16.88°$ to $16.98°$, showing that gains vary across the grid. These single-seed comparisons summarize complete-model scaling over the evaluated data–capacity range. Appendix C gives measurements and fits; Appendix C.1.2 reports the exploratory capacity scan.

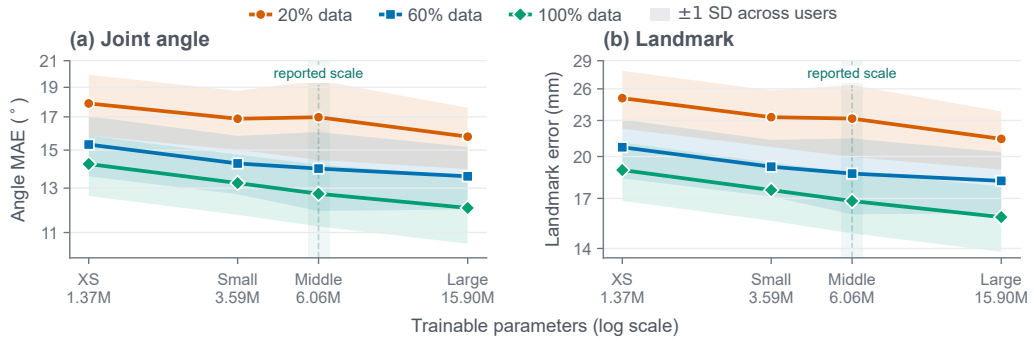


Figure 6: **STAMP improves with data and model capacity over the measured grid.**

4.6 Native streaming and long-horizon accuracy

STAMP passes all 23 causality tests (Appendix H). The MOD-295 deployment configuration reaches 1.216 ms P99 engine latency at 50 Hz; this is steady-state execution time, excluding start-up context acquisition and surrounding-system delays (Appendix I).

On thirty-second Tracking, STAMP with absolute positions reaches $14.936°/23.064$ mm under user-first aggregation, versus $13.055°/17.385$ mm for the single-task vEMG2Pose baseline. Adapting STAMP from its converged absolute-position checkpoint to RoPE lowers error to $11.618°/15.685$ mm; a matched ten-epoch continuation with Absolute positions does not improve the thirty-second result (Appendix E). Thus five-second accuracy, native causality, and long-horizon accuracy remain distinct tests.

5 Discussion and limitations

STEM's learned per-head decay controls feature memory and is fixed at inference; MoE routing varies with the input history. Near-parameter-matched replacements separate architectural organization from parameter count, while the circular-versus-linear control isolates geometry encoding. These controls evaluate the complete STEM representation and its geometry encoding; factorization, temporal decay, multiscale patching, and explicit rotation equivariance remain finer-grained questions. Scaling complements these fixed-data controls: the complete model's advantage persists and widens over the tested user counts, characterizing complete-model rather than module-specific scaling.

The cross-dataset study evaluates architectural portability through dataset-specific retraining and interface adaptation: STAMP changes its sensor-geometry and target adapters, and every model is optimized within its benchmark. STAMP leads on Stage across the three datasets, while rankings on splits involving held-out users vary across datasets.

6 Conclusion

STAMP organizes streaming sEMG decoding through STEM spatiotemporal perception, Affine Scan pose updates, and Motion MoE historical conditioning. Its leading five-second emg2pose cross-user means among matched retrained models are complemented by lower errors and steeper fitted user scaling at nearly matched capacity; additional capacity generally helps within the measured grid. Dataset-specific retraining extends the architecture to EgoEMG and NinaPro, including mixed electrode layouts. Native execution is causal, while long-horizon accuracy also depends on position treatment.

AI use statement

Generative AI assisted with interpreting author-provided reports, drafting and editing, literature discovery, figure refinement, and limited reproducibility and build-script edits. It did not generate synthetic data, prove mathematical claims, or autonomously run the reported experiments. The authors verified assisted text, equations, citations, figures, code changes, and numerical claims against source materials and take responsibility for all final content.

Ethics statement

This work studies wearable biosignals, which may contain identifying or health-relevant information. All datasets evaluated in this manuscript are existing public benchmarks, and reported results comprise aggregate metrics and illustrative pose trajectories; no personally identifying participant information is included. This paper does not introduce or evaluate a new participant dataset.

Reproducibility statement

Section 3 explains the model's task-structured design, while Appendix A specifies its operators and causal recurrences. Section 4 defines the dataset splits, alignment, metrics, and comparison protocol. The remaining appendix separates main-comparison metrics from ablation diagnostics and additionally reports anatomical breakdowns, scaling protocols, the causality audit, and deployment details. The manuscript identifies checkpoint-selection rules, hardware boundaries, and benchmark split conventions used for each reported result.

REFERENCES

Ali Ameri, Mohammad Ali Akhaee, Erik Scheme, and Kevin Englehart. A deep transfer learning approach to reducing the effect of electrode shift in EMG pattern recognition-based control. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(2):370–379, 2020. doi: 10.1109/TNSRE.2019.2962189.

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

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 813–824, 2021.

Guy E. Blelloch. Prefix sums and their applications. Technical Report CMU-CS-90-190, Carnegie Mellon University, 1990.

Ulysse Côté-Allard, Cheikh Latyr Fall, Alexandre Drouin, Alexandre Campeau-Lecours, Clément Gosselin, Kyrre Glette, François Laviolette, and Benoit Gosselin. Deep learning for electromyographic hand gesture signal classification using transfer learning. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(4):760–771, 2019. doi: 10.1109/TNSRE.2019.2896269.

EgoEMG Authors. EgoEMG official code repository. `https://github.com/zhenqis123/EgoEMG`, 2026. Configuration version `f7cbd048`.

Dario Farina, Ning Jiang, Hubertus Rehbaum, Aleš Holobar, Bernhard Graimann, Hans Dietl, and Oskar C. Aszmann. The extraction of neural information from the surface EMG for the control of upper-limb prostheses: Emerging avenues and challenges. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 22(4):797–809, 2014. doi: 10.1109/TNSRE.2014.2305111.

Leo Feng, Frederick Tung, Mohamed Osama Ahmed, Yoshua Bengio, and Hossein Hajimirsadeghi. Were RNNs all we needed? *arXiv preprint arXiv:2410.01201*, 2024. URL `https://arxiv.org/abs/2410.01201`.

Yanjuan Geng, Zhebin Yu, Yucheng Long, Liuni Qin, Ziyin Chen, Yongcheng Li, Xin Guo, and Guanglin Li. A CNN-attention network for continuous estimation of finger kinematics from surface electromyography. *IEEE Robotics and Automation Letters*, 7(3):6297–6304, 2022. doi: 10.1109/LRA.2022.3169448.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *Conference on Language Modeling*, 2024.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022.

Weiyu Guo, Ning Jiang, Dario Farina, Jingyong Su, Zheng Wang, Chuang Lin, and Hui Xiong. Multi-attention feature fusion network for accurate estimation of finger kinematics from surface electromyographic signals. *IEEE Transactions on Human-Machine Systems*, 53(3):512–519, 2023. doi: 10.1109/THMS.2023.3252817.

Nima Hadidi, Johannes Lee, Ebrahim Feghhi, Michael Yuan, and Jonathan C. Kao. Re-evaluating position and velocity decoding for hand pose estimation with surface electromyography. *arXiv preprint arXiv:2603.08212*, 2026. doi: 10.48550/arXiv.2603.08212. URL `https://arxiv.org/abs/2603.08212`.

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

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

Néstor J. Jarque-Bou, Manfredo Atzori, and Henning Müller. A large calibrated database of hand movements and grasps kinematics. *Scientific Data*, 7:12, 2020. doi: 10.1038/s41597-019-0349-2.

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

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 5156–5165, 2020.

Stephan Johann Lehmler, Muhammad Saif-ur Rehman, Tobias Glasmachers, and Ioannis Iossifidis. Deep transfer learning compared to subject-specific models for sEMG decoders. *Journal of Neural Engineering*, 19(5):056039, 2022. doi: 10.1088/1741-2552/ac9860.

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

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*, 2025. URL https://hands-workshop.org/files/2025/VQ-MyoPose.pdf.

Eric Martin and Chris Cundy. Parallelizing linear recurrent neural nets over sequence length. In *International Conference on Learning Representations*, 2018.

Meta Reality Labs. pose_modules.py in the official emg2pose repository. https://github.com/facebookresearch/emg2pose/blob/main/emg2pose/pose_modules.py, 2024. Accessed 2026-09-22.

Micropilot. emg2tendon pretrained models and model card. https://huggingface.co/Micropilot/emg2tendon, 2026. Accessed 2026-09-22.

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, Datasets and Benchmarks Track*, volume 37, 2024. doi: 10.52202/079017-1770. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/64f8884f6ba3d9ace5bb647c4f917896-Abstract-Datasets_and_Benchmarks_Track.html.

Raul C. Sîmpetru, Andreas Arkudas, Dominik I. Braun, Marius Oßwald, Daniela Souza de Oliveira, Bjoern Eskofier, Thomas M. Kinfe, and Alessandro Del Vecchio. Learning a hand model from dynamic movements using high-density EMG and convolutional neural networks. *IEEE Transactions on Biomedical Engineering*, 71(12):3556–3568, 2024. doi: 10.1109/TBME.2024.3432800.

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

Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. Retentive network: A successor to transformer for large language models. *arXiv preprint arXiv:2307.08621*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Sagar Verma. emg2tendon: From sEMG signals to tendon control in musculoskeletal hands. In *Proceedings of Robotics: Science and Systems*, Los Angeles, CA, USA, 2025. doi: 10.15607/RSS.2025.XXI.135. URL https://www.roboticsproceedings.org/rss21/p135.html.

Ziheng Xi, Jiayi Yu, Yitao Wang, Yanbo Duan, Jianjiang Feng, and Jie Zhou. EgoEMG: A multimodal egocentric dataset with bilateral EMG and vision for hand pose estimation. *arXiv preprint arXiv:2605.05712*, 2026. doi: 10.48550/arXiv.2605.05712. URL https://arxiv.org/abs/2605.05712.

Eric Xie and Hei Shing Cheung. REACT: A conditioning framework for user-adaptive sEMG hand pose estimation. *arXiv preprint arXiv:2605.30127*, 2026. doi: 10.48550/arXiv.2605.30127. URL https://arxiv.org/abs/2605.30127.

Jehan Yang, Maxwell Soh, Vivianna Lieu, Douglas J. Weber, and Zackory Erickson. EMGBench: Benchmarking out-of-distribution generalization and adaptation for electromyography. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/59fe60482e2e5faf557c37d121994663-Paper-Datasets_and_Benchmarks_Track.pdf.

A Formal specification of STAMP

A.1 Causal task definition

Let $\mathbf{X} \in \mathbb{R}^{T_s \times C}$ be multichannel sEMG and $\mathbf{Y} \in \mathbb{R}^{T_p \times J}$ the target joint-angle trajectory. emg2pose uses $C = 16$, $J = 20$, 2 kHz input, and 50 Hz targets; our EgoEMG-aligned setting uses $C = 8$ and $J = 22$ (20 finger and two wrist angles); the NinaPro adaptation uses $C = 12$ and $J = 22$. Let $\mathbf{s}_0$ denote task-specific initialization: $\mathbf{s}_0 = \varnothing$ for Regression and $\mathbf{s}_0 = \mathbf{y}_0$ for Tracking, where the true initial pose is supplied once and never refreshed. At native output step k, a causal model may use only the signal prefix available when that output is emitted and this declared initialization. For the emg2pose streaming interface, 1,790 context samples are consumed once and each subsequent 40-sample block emits one pose:

$$\widehat{\mathbf{y}}_k = f_\theta(\mathbf{X}_{1:\tau_k}, \mathbf{s}_0), \qquad \tau_k = 1790 + 40(k+1), \quad k \geq 0. \tag{5}$$

The native interface has a 20 ms algorithmic block delay.

A.2 STEM encoder and geometry conditioning

Per-electrode temporal patches. For electrode c, the same patch encoder is applied at two temporal scales:

$$\mathbf{e}_{t,c}^{(s)} = \phi_s(\mathbf{X}_{40t-L_s+1:40t,c}), \qquad L_s \in \{40, 128\}. \tag{6}$$

The 40-sample scale matches the 20 ms output interval, while the 128-sample scale captures a 64 ms local context. Patch-encoder weights are shared across electrodes. Both patch streams are aligned to stride 40. Learned branch logits are normalized by softmax to obtain shared scalar weights; their weighted sum is followed by layer normalization to form $\mathbf{E}$. The weights do not depend on the current input.

Circular position encoding. For wrist arrays, electrode positions lie on a ring. For index $c \in \{0, \ldots, C-1\}$, we use Fourier features

$$\mathbf{r}_c = [\sin(2\pi kc/C), \cos(2\pi kc/C)]_{k=1}^{K_f}, \tag{7}$$

The resulting fixed basis is repeated and truncated deterministically to the token width; it introduces no trainable parameters. The coordinates are periodic across electrodes $C-1$ and 0. The codes are added once before the block stack, $\mathbf{H}^{(0)} = \mathbf{E} + \mathbf{R}$. NinaPro uses the mixed-layout codes in Appendix B.2.

STEM block. STEM corresponds to `FactorizedLinearBlock` in the implementation. Equation 1 specifies its order: spatial softmax attention across the C electrodes at each time, decayed causal linear attention along each spatially mixed electrode stream, and a two-layer GELU feedforward network. Each sublayer has its own pre-layer normalization and residual addition. The

implementation includes attention dropout, dropout on attention outputs, and dropout after the feed-forward activation and output projection; these are omitted from the main equation. The block preserves the $T \times C \times d$ layout.

The encoder contains five independently parameterized blocks of width 224, with eight attention heads and feedforward width 896, followed by a final layer normalization. Query pooling follows the encoder.

Decayed causal temporal attention. Within each block, query, key, value, and output projections are shared across electrode streams, whose prefix states are maintained separately. The positive feature map is $\phi(x) = \mathrm{ELU}(x) + 1$. For head h, $\gamma_h = \sigma(\alpha_h)$ is learned and initialized from half-lives logarithmically spaced from 4 to 256 tokens; it is fixed at inference. Suppressing electrode and block indices, the streaming update is

$$\mathbf{S}_{t,h} = \gamma_h \mathbf{S}_{t-1,h} + \phi(\mathbf{k}_{t,h})\mathbf{v}_{t,h}^\top, \quad \mathbf{r}_{t,h} = \gamma_h \mathbf{r}_{t-1,h} + \phi(\mathbf{k}_{t,h}), \quad \mathbf{o}_{t,h} = \frac{\phi(\mathbf{q}_{t,h})^\top \mathbf{S}_{t,h}}{\max(\phi(\mathbf{q}_{t,h})^\top \mathbf{r}_{t,h}, 10^{-6})}. \tag{8}$$

Training evaluates this recurrence in 32-token chunks with analytically decayed carry states.

Learned electrode pooling. After the encoder's final normalization, $K = 2$ learned queries attend to its C electrode states $\mathbf{h}_{t,1:C}$ and yield two downstream pose-decoder tokens per time step:

$$\mathbf{z}_{t,k} = \mathrm{Attn}(\mathbf{q}_k, \mathbf{h}_{t,1:C}), \qquad k \in \{1, 2\}. \tag{9}$$

A.3 Pose–motion Affine Scan

At each step, the decoder predicts a pose proposal $\mathbf{p}_t$, a learned per-frame motion increment $\mathbf{v}_t$, and gate logits $\boldsymbol{\ell}_t$:

$$(\mathbf{p}_t, \mathbf{v}_t, \boldsymbol{\ell}_t) = g(\mathbf{h}_t), \qquad \mathbf{a}_t = \sigma(\boldsymbol{\ell}_t). \tag{10}$$

The pose state is updated elementwise as

$$\mathbf{x}_t = \mathbf{a}_t \odot \mathbf{x}_{t-1} + \mathbf{a}_t \odot \mathbf{v}_t + (1 - \mathbf{a}_t) \odot \mathbf{p}_t. \tag{11}$$

The increment $\mathbf{v}_t$ is expressed in angular units per output frame. Defining

$$\mathbf{b}_t = \mathbf{a}_t \odot \mathbf{v}_t + (1 - \mathbf{a}_t) \odot \mathbf{p}_t, \tag{12}$$

Eq. 11 becomes the affine map $F_t(\mathbf{x}) = \mathbf{a}_t \odot \mathbf{x} + \mathbf{b}_t$. Two such maps compose as

$$(\mathbf{a}_2, \mathbf{b}_2) \circ (\mathbf{a}_1, \mathbf{b}_1) = (\mathbf{a}_2 \odot \mathbf{a}_1,\ \mathbf{b}_2 + \mathbf{a}_2 \odot \mathbf{b}_1), \tag{13}$$

which is associative. Training can therefore use a parallel prefix scan with $O(\log T)$ dependency depth, while online inference stores only the previous pose and performs one $O(J)$ update.

A.4 Causal two-stage refinement

The first decoder produces causal features $\mathbf{h}_t^{(1)}$ and a coarse pose trajectory $\mathbf{x}_t^{(1)}$ through its Affine Scan. The refinement decoder conditions these features on the one-step-delayed coarse pose and its frame-to-frame difference:

$$\mathbf{u}_t^{(2)} = \mathrm{Condition}\Big(\mathbf{h}_t^{(1)},\ \Big[\mathbf{x}_{t-1}^{(1)},\ \mathbf{x}_{t-1}^{(1)} - \mathbf{x}_{t-2}^{(1)}\Big]\Big). \tag{14}$$

The pose difference is computed from the coarse trajectory; it is distinct from the first stage's learned increment $\mathbf{v}_{t-1}^{(1)}$. The conditioned features then pass through causal attention and a second Affine Scan to obtain $\mathbf{x}_t^{(2)}$. The final prediction is $\widehat{\mathbf{y}}_t = \mathbf{x}_t^{(2)}$. Algorithm 1 summarizes the two-stage computation, with $\mathbf{V}$ collecting the learned per-frame increments $\mathbf{v}_t$ defined above.

Algorithm 1 Two-stage Affine Scan

Inputs: causal perceptual tokens $\mathbf{Z}$, initial pose $\mathbf{x}_0$, and task mask $\mathbf{M}$ (zero during position-only initialization, one otherwise).

1 **function** $\mathrm{Rollout}(\mathbf{H}, \mathbf{x}_0, \mathbf{M})$
2 $\quad(\mathbf{P}, \mathbf{V}, \mathbf{L}) \leftarrow \mathrm{CommandHead}(\mathbf{H})$
3 $\quad\mathbf{A} \leftarrow \mathbf{M} \odot \sigma(\mathbf{L})$
4 $\quad\mathbf{B} \leftarrow \mathbf{A} \odot \mathbf{V} + (1 - \mathbf{A}) \odot \mathbf{P}$
5 $\quad(\overline{\mathbf{A}}, \overline{\mathbf{B}}) \leftarrow \mathrm{InclusivePrefixScan}(\mathbf{A}, \mathbf{B}; \circ)$
6 $\quad$**return** $\overline{\mathbf{A}} \odot \mathbf{x}_0 + \overline{\mathbf{B}}$
7 $\mathbf{H}^{(1)} \leftarrow \mathrm{CausalDecoder}_1(\mathbf{Z})$
8 $\mathbf{X}^{(1)} \leftarrow \mathrm{Rollout}(\mathbf{H}^{(1)}, \mathbf{x}_0, \mathbf{M})$
9 $\mathbf{s}_t \leftarrow [\mathbf{x}^{(1)}_{t-1},\ \mathbf{x}^{(1)}_{t-1} - \mathbf{x}^{(1)}_{t-2}], \quad t = 1, \ldots, T$
10 $\mathbf{H}^{(2)} \leftarrow \mathrm{CausalDecoder}_2(\mathbf{H}^{(1)}, \mathbf{S})$
11 $\widehat{\mathbf{Y}} \leftarrow \mathrm{Rollout}(\mathbf{H}^{(2)}, \mathbf{x}_0, \mathbf{M})$
12 **return** $\widehat{\mathbf{Y}}$

The inclusive scan composes maps in chronological order using Equation 13, so its tth output represents $F_t \circ \cdots \circ F_1$. In line 9, $\mathbf{x}^{(1)}_0 = \mathbf{x}^{(1)}_{-1} = \mathbf{x}_0$ initializes the delayed pose and makes its first difference zero. The mask $\mathbf{M}$ forces absolute-pose prediction during position-only initialization. Within each stage, all update coefficients are generated before the scan; the second stage uses the completed coarse trajectory during parallel training and the available coarse history during streaming inference.

Decoder temporal position handling. Position encoding is applied in the causal softmax-attention decoder, not in the encoder's decayed linear attention. The base decoder adds an absolute sinusoidal code before attention. The long-horizon variant instead applies rotary position embedding (RoPE) to query and key after their learned projections and before scaled dot-product attention; the 250-token causal cache is unchanged (Su et al., 2024). Both variants use only current and past decoder tokens; Appendix E specifies their checkpoints and inference-only wrapping controls.

A.5 Causal multi-timescale Motion MoE

We place residual expert banks (A) after the encoder, (B1) between the two decoder attention stages, and (B2) after the final attention stage. Each bank contains $N = 32$ residual multilayer-perceptron (MLP) experts. The three banks and their router projections have independent parameters (96 experts total); only the architectural template and four decay scales are shared. For feature $\mathbf{u}^\ell_t$, the routed update is

$$\widetilde{\mathbf{u}}^\ell_t = \mathbf{u}^\ell_t + \sum_{e \in \mathrm{TopK}(\boldsymbol{\pi}^\ell_t, 8)} \pi^\ell_{t,e} E^\ell_e(\mathbf{u}^\ell_t). \tag{15}$$

Expert output projections are initialized to zero, so optimization begins from the shared path.

The router summarizes motion at four causal decay scales, corresponding to 0.25, 0.5, 1, and 2 seconds. For decay λ_m, it maintains

$$\mathbf{q}^{(m)}_t = \lambda_m \mathbf{q}^{(m)}_{t-1} + (1 - \lambda_m)\psi(\mathbf{u}_t), \tag{16}$$

and computes $\boldsymbol{\pi}_t$ from the concatenated summaries.

The training/evaluation implementation evaluates all 32 experts and then masks the top-8 mixture weights. Deployment packs the frozen expert weights and executes only selected rows using batched matrix multiplications. Routing runs on the GPU. Appendix I gives the numerical checks and timing protocol.

A.6 Streaming state and causality

At inference, STAMP stores (i) a patch buffer, (ii) temporal linear-attention prefix states, (iii) two causal-attention key–value caches, (iv) the previous pose for each Affine Scan, and (v) four router summaries per insertion point. After warm-up, one 20 ms signal block produces one pose frame

without revisiting the full history. The largest encoder context spans 1,790 input samples, or approximately 895 ms at 2 kHz. This context is acquired at start-up.

B EVALUATION PROTOCOLS

The retrained comparison family is STAMP, EMGFormer-M, NeuroPose, and task-matched vEMG2Pose; the emg2pose evaluation additionally includes our retrained vEMG2Pose (Multi-task), following the joint Regression–Tracking training strategy of Hadidi et al. (2026). For Table 1, eligible windows are first pooled within each test user; user means are then summarized with equal weight and population SD. User/Stage/User+Stage/Overall contain 20/158/20/178 users for emg2pose and 5/36/5/40 for EgoEMG. Overall pools each user's windows across conditions before summarizing unique users. The released EgoEMG task is Regression; our Tracking extension supplies the true pose once at initialization (Xi et al., 2026). For consistent reporting, every result table uses the axes *User*, *Stage*, and *User+Stage*. Stage denotes the held-out motion condition: movement stage on emg2pose and held-out gesture or action on EgoEMG and NinaPro. All models within each benchmark share training data, split manifests, window construction, checkpoint selection, and evaluation. Unless stated otherwise, training uses Adam, batch size 64, FP32, validation-selected checkpoints, and $\mathcal{L}_{\text{angle}} + 0.01\mathcal{L}_{\text{fingertip}}$; NinaPro uses masked angular loss. The backbone learning rate is 10^{-4}; final Motion MoE experts and routers use 10^{-3}. The emg2pose main-model and cumulative-ablation configurations each use seed 42. The emg2pose vEMG2Pose Tracking evaluations at five and thirty seconds share the seed-42 single-task checkpoint selected by minimum validation `val_loss` at epoch 75 (SHA-256 prefix: `1d63d3ae4002`). These comparisons and the long-horizon experiments run on an NVIDIA H200 system; EgoEMG uses a separate H200 system; NinaPro and scaling use RTX 5090 GPUs; latency uses an RTX 5060 Laptop GPU.

Table 4: Published Hadidi et al. (Hadidi et al., 2026) results under their training and alignment protocols: Angle (°) / Landmark (mm), mean ± SD across users.

Published configuration	User (°/mm) ↓	Stage (°/mm) ↓	User+Stage (°/mm) ↓
Position, multi-task (Regression)	$\mathbf{11.54 \pm 1.19 / 14.74 \pm 1.64}$	$14.02 \pm 1.62 / 18.61 \pm 2.16$	$\mathbf{14.58 \pm 1.29}/19.78 \pm 1.83$
Velocity, multi-task (Regression)	$11.64 \pm 1.25 / \mathbf{14.74 \pm 1.74}$	$\mathbf{13.85 \pm 1.62 / 18.33 \pm 2.16}$	$14.63 \pm 1.34 / \mathbf{19.77 \pm 1.95}$
Position, single-task (Tracking)	$7.35 \pm 0.92 / 9.67 \pm 1.48$	$10.21 \pm 1.32 / 13.73 \pm 1.69$	$10.25 \pm 0.93 / 14.18 \pm 1.35$

Table 8 records the retrained implementations and alignment paths. The vEMG2Pose (Multi-task) implementation follows the training specification of Hadidi et al. (2026); its configuration is given below.

B.1 REFERENCE PROTOCOLS

REACT encodes 15 recordings from other sessions of the test user using a bidirectional GRU and FiLM-conditions a pretrained vEMG2Pose backbone (Xie & Cheung, 2026). Published User/Stage/User+Stage angular errors are $12.0/14.6/15.3°$ for Regression and $7.4/10.9/10.8°$ for Tracking.

emg2tendon evaluates CLDM for direct pose and 39-dimensional tendon control (Verma, 2025); its public checkpoint targets tendon control (Micropilot, 2026). VQ-MyoPose uses a VQ-VAE and bidirectional EMG-to-token GRU on a 12-subject, 27-stage subset (Lovecchio et al., 2025).

B.2 NINAPRO MIXED-LAYOUT PORTABILITY

Task construction. NinaPro DB2 provides synchronized 12-channel sEMG from 40 intact participants; DB9 provides calibrated CyberGlove-II kinematics for the corresponding recordings (Atzori et al., 2014; Jarque-Bou et al., 2020). Exercises 1 and 2 contain 40 movements with six repetitions per participant, yielding 9,600 subject–gesture–repetition units. The model receives sEMG as input and DB9 kinematics as supervision.

The 12 input channels comprise eight electrodes distributed around the forearm and four electrodes placed at the flexor digitorum superficialis, extensor digitorum, biceps, and triceps. STAMP represents this mixed topology with 18 fixed geometry features: two channel-type indicators,

eight Fourier ring coordinates, four anatomical-site indicators, two limb-level indicators, and two flexor/extensor group indicators. A learned projection maps these features to the token width, replacing the circular input codes before the STEM block stack. It adds 4,032 parameters (0.067% of the adapted model); the STEM blocks, Affine Scan, and Motion MoE retain their architecture. Every model, including STAMP, is retrained under the same data, split, window, loss, and evaluator after changing its interface to 12 input channels and 22 outputs.

Causal windows and splits. Each input contains 1,790 history samples and a 400-sample scored interval at 2 kHz, for 2,190 samples (1.095 s) in total; the stride is 100 samples (50 ms). Missing history at a sequence boundary is zero padded, and the loss is evaluated only on the scored interval. No true initial pose is provided. The fixed split uses 28/4/8 train/validation/test participants and 32/8 seen/unseen movements. Table 5 accounts for every training, validation, and test unit.

Table 5: NinaPro DB2+DB9 split accounting. User denotes held-out participants; Stage denotes held-out movements; User+Stage combines both shifts.

Partition	Units	Duration (h)	Windows
Train: seen users, seen movements	5,376	7.303	507,206
Validation: User	768	1.057	73,429
Validation: User+Stage	192	0.293	20,424
Test: User	1,536	2.141	148,872
Test: Stage	1,344	1.933	134,535
Test: User+Stage	384	0.544	37,841

Targets and aggregation. The network retains a 22-dimensional output interface. One DB9 dimension has no reliable calibrated observation and is excluded from both training and evaluation by a validity mask. Overlapping-window errors are first accumulated within each subject–gesture–repetition unit. Each split is the equal-weight mean of its unit MAEs. The pooled metric instead weights unit error e_u by the number n_u of valid angle values,

$$E_{\text{pool}} = \frac{\sum_u n_u e_u}{\sum_u n_u}. \tag{17}$$

The three test partitions contain 3,264 mutually exclusive units and 2,698,483,200 valid angle values.

Table 6: NinaPro DB2+DB9 — Regression; 1.095 s input, 0.2 s scored interval; Angle (°). Results use the mixed-layout portability protocol; lower is better. Split cells report the equal-unit mean ± population standard deviation across subject means (User/User+Stage: $n = 8$; Stage: $n = 28$). Overall retains the valid-angle-value weighting defined above and reports the corresponding weighted population standard deviation across 36 subject means.

Model	Params (M) ↓	User ↓	Stage ↓	User+Stage ↓	Overall test ↓
EMGFormer-M†	6.629	18.007±1.974	17.740±3.325	21.901±3.148	17.587±2.597
NeuroPose†	6.355	**17.100±1.447**	18.606±3.138	**21.606±2.915**	17.553±2.222
vEMG2Pose† (Regression)	**5.975**	18.282±2.148	20.120±3.405	21.999±3.157	18.556±2.769
STAMP	6.064	17.771±2.081	**16.202±3.295**	22.216±3.492	**16.829±2.978**

† Baselines retrained using the dataset-specific implementations and evaluation protocols in Appendix B.

B.3 EgoEMG aligned-protocol audit and auxiliary results

Each EgoEMG row in Table 1 evaluates 2,792 windows containing 627,139,040 valid angle elements across User, Stage, User+Stage and both hands. Training uses train-labeled frames from EgoEMG, ShowEE, and EgoEMG-Incre; evaluation uses EgoEMG held-out labels. Results use the user-first aggregation defined above.

In the released configuration, window membership uses the center label (EgoEMG Authors, 2026). Of 40,855 training-window positions per hand, 3,354 (8.21%) contain non-train frames. Raw-frame overlap affects 616 of 754 Stage windows (81.70%), 4 of 538 User windows (0.74%), and none of the 104 User+Stage windows. User/User+Stage identities also occur in train-labeled frames.

Table 7: EgoEMG auxiliary results: mean ± population SD across 40 users after pooling eligible windows per user. Velocity is in rad/s; acceleration and jerk are evaluator-scaled finite-difference proxies in a.u. Loss, fingertip and anatomical errors are minimized. STAMP Regression uses the validation-selected checkpoint.

(a) Spatial, objective, and motion diagnostics

Task	Model	Fingertip (mm) ↓	Test loss ↓	Velocity (rad/s)	Accel. proxy (a.u.)	Jerk proxy (a.u.)
Regression	EMGFormer-M†	37.690±6.297	0.6155±0.0993	0.5582±0.0963	1.1165±0.1925	2.2332±0.3851
Regression	NeuroPose†	45.522±3.765	0.7388±0.0539	0.1084±0.0266	0.2168±0.0533	0.4336±0.1065
Regression	vEMG2Pose†	40.970±5.354	0.6707±0.0795	0.3084±0.0449	0.0043±0.0006	0.0063±0.0009
Regression	**STAMP**	**34.827±7.578**	**0.5744±0.1187**	0.7637±0.1168	0.0293±0.0048	0.0440±0.0072
Tracking	EMGFormer-M†	27.815±5.560	0.4559±0.0882	0.1962±0.0678	0.3925±0.1357	0.7850±0.2714
Tracking	NeuroPose†	28.160±5.626	0.4610±0.0892	0.0257±0.0159	0.0514±0.0319	0.1027±0.0638
Tracking	vEMG2Pose†	31.071±4.461	0.5112±0.0702	0.1570±0.0235	0.0025±0.0003	0.0036±0.0004
Tracking	**STAMP**	**26.849±4.782**	**0.4394±0.0759**	0.0792±0.0182	0.0018±0.0004	0.0027±0.0006

(b) Anatomical angular MAE (°): finger groups

Task	Model	Thumb	Index	Middle	Ring	Little
Regression	EMGFormer-M†	11.352±2.164	13.092±2.060	14.904±2.286	14.049±2.412	16.139±2.668
Regression	NeuroPose†	12.857±1.332	15.213±1.608	18.088±1.503	16.839±1.204	19.300±1.753
Regression	vEMG2Pose†	12.412±1.704	14.074±1.762	16.321±1.669	15.417±1.840	17.557±2.185
Regression	STAMP	**10.756±2.307**	**12.483±2.339**	**13.815±2.659**	**13.101±2.755**	**15.120±3.163**
Tracking	EMGFormer-M†	8.366±1.527	10.293±1.981	10.406±2.042	9.947±2.251	11.495±2.857
Tracking	NeuroPose†	8.413±1.553	10.371±2.006	10.459±2.107	9.935±2.264	11.586±2.822
Tracking	vEMG2Pose†	9.145±1.207	11.282±1.611	12.198±1.747	11.610±1.862	13.059±2.350
Tracking	**STAMP**	**7.952±1.323**	**9.860±1.642**	**10.081±1.811**	**9.621±1.961**	**11.106±2.525**

(c) Anatomical angular MAE (°): joint levels and wrist

Task	Model	Proximal	Intermediate	Distal	Wrist flex.	Wrist dev.
Regression	EMGFormer-M†	12.639±1.986	20.365±3.867	9.987±1.849	**13.955±3.337**	**8.658±2.733**
Regression	NeuroPose†	15.134±1.332	24.461±2.113	11.110±1.800	16.587±2.954	11.630±2.363
Regression	vEMG2Pose†	13.783±1.628	22.348±2.930	10.710±1.624	16.002±3.026	9.832±2.193
Regression	STAMP	**11.958±2.363**	**18.876±4.358**	**9.427±1.802**	14.832±3.471	9.091±2.624
Tracking	EMGFormer-M†	9.785±1.759	13.881±3.504	6.954±1.519	14.494±3.128	7.524±1.951
Tracking	NeuroPose†	9.824±1.756	13.951±3.551	7.013±1.538	15.523±3.481	7.574±1.874
Tracking	vEMG2Pose†	10.525±1.317	16.376±2.890	8.409±1.640	15.033±2.634	8.517±1.713
Tracking	**STAMP**	**9.324±1.407**	**13.526±3.199**	**6.724±1.389**	**13.592±2.502**	**7.301±1.577**

† Baselines retrained using the dataset-specific implementations and evaluation protocols in Appendix B.

The same held-out pool supplies checkpoint selection and reporting. Appendix B.4 specifies the additional split.

For the alignment record, a native output is indexed by its emission time. Sample-wise zero-look-ahead requires that the prediction scored at target sample t use no EMG after t, including preprocessing, resampling, and output alignment.

The released pose stack linearly interpolates TDS features from 25 to 50 Hz with `align_corners=True`, then decoded poses from 50 Hz to 2 kHz for scoring (Salter et al., 2024; Meta Reality Labs, 2024; Hadidi et al., 2026). Interpolation between anchors uses the subsequent feature or pose token. Our EMGFormer-M and NeuroPose adaptations select no-later-than-current features. STAMP emits a native pose after each 40-sample block; benchmark scoring uses the released 50 Hz-to-2 kHz linear output alignment.

vEMG2Pose (Multi-task) uses the Regression model's causal TDS encoder and two-layer, 512-hidden-unit LSTM pose–velocity decoder. Each batch is evaluated as Regression with zero initialization and Tracking with a true initial pose; losses have weights 0.125 and 0.875. Both branches use position proposals for the first 12 native steps, then integrate predicted velocity. Table 1 evaluates this checkpoint as Regression.

Table 8: Method provenance and reproducible emg2pose alignment record. Panel (a) pairs cited formulations with our implementations. Panel (b) records the evaluation paths for the models in Table 1. "future neighbor" refers to interpolation between native anchors, not to the native model update. Panel (c) identifies the score grids and generated artifacts.

(a) Cited formulation and evaluated implementation

Cited formulation	Evaluated implementation(s)
EMGFormer (Xi et al., 2026)	EMGFormer-M†; emg2pose causal adaptation and EgoEMG-aligned implementation
NeuroPose (Liu et al., 2021)	NeuroPose†; causal adaptation
vEMG2Pose (Salter et al., 2024); multi-task training (Hadidi et al., 2026)	vEMG2Pose†; retrained Regression, Multi-task, and Tracking variants

(b) Recorded paths for locally evaluated emg2pose measurements

Implementation	Native output	Feature alignment	Dense score alignment	Latest-input dependency
EMGFormer-M† adaptation	token prediction	past-token selection	causal hold at 2 kHz	no later than scored sample
NeuroPose† adaptation	sample-aligned prediction	audited 319-sample shift	none after alignment	no later than scored sample
vEMG2Pose† variants	50 Hz pose	25→50 Hz linear	50 Hz→2 kHz linear	future neighbor possible at both sites
STAMP	50 Hz pose	none	50 Hz→2 kHz linear	native: chunk end; dense: next pose token

(c) Score grid and artifact record

Evidence	Scored outputs per 5 s	Evaluation source	Persisted artifacts
Locally evaluated accuracy	10,000 samples at 2 kHz	released `EMG2PoseEvaluation`	`metrics.csv`, `pooled_metrics.csv`
Native streaming audit	250 updates at 50 Hz	prefix and stateful-stream tests	test results and raw latency traces

† Baselines retrained using the dataset-specific implementations and evaluation protocols in Appendix B.

B.4 EgoEMG evaluation with audited splits

The additional EgoEMG split below is used for training, checkpoint selection, and final testing. All four methods use eight EMG channels, 22 joint-angle outputs, and seed 42. Regression receives no true initial pose; Tracking receives it once. Checkpoints and hyperparameters are selected on validation data, with a separate final test set.

Identity and gesture partitioning. We retain only EgoEMG recordings with fully auditable identities and gesture labels. ShowEE is excluded from this auditable pool, and EgoEMG-Incre is excluded because usable gesture-class labels are unavailable. Session suffixes are removed to group recordings from the same person under one normalized identity. With seed 42, the 39 identities are assigned to disjoint training, validation, and test identity groups of 27, 6, and 6; the 60 gesture IDs are independently assigned to disjoint groups of 40, 10, and 10. Table 9 specifies their combinations. User combines unseen identities with seen gestures; Stage combines seen identities with unseen gestures; User+Stage holds out both.

Table 9: Identity–gesture combinations for the additional EgoEMG evaluation. Group sizes are identity-group sizes before window filtering.

Partition	Identity group	Gesture group
Training	Training (27)	Training (40)
Validation User	Validation (6)	Training (40)
Validation Stage	Training (27)	Validation (10)
Validation User+Stage	Validation (6)	Validation (10)
Test User	Test (6)	Training (40)
Test Stage	Training (27)	Test (10)
Test User+Stage	Test (6)	Test (10)

Windows and aggregation. Evaluation uses five-second windows. Training stride is 1,200 input samples; validation/test stride is 12,000 samples. A 6,000-sample exclusion zone is applied on each side of every semantic split boundary. The audit verifies disjoint identity and gesture groups and zero retained boundary-crossing windows. Each hand contributes 504 User, 296 Stage, and 76 User+Stage test windows, totaling 1,752 across both hands. Subject-mean results pool windows within each normalized identity, then compute an equal-user mean and population SD. Pooled first combines eligible conditions within each identity.

Table 10: EgoEMG with audited splits: cross-user mean ± population standard deviation for joint-angle MAE (°) / whole-hand landmark distance (mm). Five-second Regression and Tracking use separate, validation-selected checkpoints with seed 42. Pooled combines eligible windows within each normalized identity before summarizing users. Lower is better; bold marks the lowest mean within each task and metric before rounding.

Model	Params (M)	User (°/mm) ↓	Stage (°/mm) ↓	User+Stage (°/mm) ↓	Pooled (°/mm) ↓
(a) *Regression*					
EMGFormer-M†	6.617	15.72±1.35/25.84±2.99	15.21±1.40/26.63±2.41	15.66±0.83/26.74±1.74	15.30±1.37/26.49±2.47
NeuroPose†	6.355	15.58±1.25/25.87±2.82	15.29±1.50/26.65±2.38	15.70±0.69/26.91±1.51	15.34±1.43/26.51±2.40
vEMG2Pose†	5.964	15.89±1.12/26.49±2.28	15.50±1.40/26.89±2.30	15.92±0.67/27.28±1.86	15.57±1.33/26.82±2.24
STAMP	6.060	**15.34±1.43/25.56±2.86**	**15.16±1.42/26.38±2.35**	**15.39±0.58/26.42±1.35**	**15.19±1.39/26.23±2.38**
(b) *Tracking*					
EMGFormer-M†	6.617	9.14±1.66/**14.76±2.84**	9.20±1.83/16.12±3.48	**8.02±1.50/12.92±3.84**	9.17±1.77/15.83±3.38
NeuroPose†	6.355	9.15±1.61/14.80±2.71	9.19±1.81/16.06±3.42	8.05±1.46/13.01±3.73	9.16±1.74/15.79±3.31
vEMG2Pose†	5.964	11.66±1.13/18.57±2.08	11.58±1.68/19.62±2.90	11.37±1.08/18.00±3.17	11.59±1.57/19.42±2.77
STAMP	6.060	**9.00±1.18**/14.76±1.85	**9.05±1.63/15.84±3.01**	8.08±1.12/13.34±2.93	**9.02±1.53/15.61±2.85**

† Baselines retrained under the audited split protocol in this section.
For Tracking User landmark error, STAMP and EMGFormer-M both round to 14.76 mm; their unrounded means are 14.75864 and 14.75726 mm, respectively.

Subject-mean results.

Table 11: EgoEMG with audited splits: element-weighted joint-angle MAE (°) / whole-hand landmark distance (mm). All valid angle or landmark elements across hands and windows contribute directly, without a user-level macro-average. These are the same predictions as Table 10; only aggregation changes. Lower is better; bold marks the lowest error within each task and metric.

Model	User (°/mm) ↓	Stage (°/mm) ↓	User+Stage (°/mm) ↓	Pooled (°/mm) ↓
(a) *Regression*				
EMGFormer-M†	16.007/26.625	**15.357**/26.830	15.774/27.168	15.767/26.741
NeuroPose†	15.871/26.598	15.619/27.115	15.686/26.901	15.770/26.799
vEMG2Pose†	16.154/27.076	15.696/27.189	15.820/27.124	15.970/27.118
STAMP	**15.657/26.277**	15.448/**26.814**	**15.336/26.333**	**15.559/26.463**
(b) *Tracking*				
EMGFormer-M†	9.295/15.185	9.186/15.823	8.513/**14.276**	9.190/15.322
NeuroPose†	9.301/15.192	9.173/15.762	8.539/14.340	9.192/15.311
vEMG2Pose†	11.837/18.962	11.635/19.496	11.692/19.204	11.756/19.164
STAMP	**9.111/15.037**	**9.023/15.569**	**8.458**/14.407	**9.024/15.162**

† Baselines retrained under the audited split protocol in this section.

Element-weighted results.

B.5 Temporal-alignment sensitivity

This evaluation uses the five-second emg2pose and released-protocol EgoEMG comparisons in Table 1; Appendix B.3 describes the released alignment. Across nine model–task configurations, checkpoints are frozen, and test windows, splits, task initialization, masks, kinematic reconstruction, metrics, and user-first aggregation remain unchanged. vEMG2Pose replaces feature interpolation with selection of the most recent available feature and output interpolation with delayed zero-order hold. STAMP changes only output alignment. EMGFormer-M and NeuroPose retain their existing past-only alignment and are not recomputed.

Scoring starts after the configured left context has been acquired. On the 2 kHz grid, output holding uses a 40-sample stride: samples 0–39 use the initial pose, samples 40–79 use the first native prediction, and subsequent intervals hold the latest available prediction. Regression initializes with the

zero pose; Tracking uses the one true pose supplied at the context boundary. EgoEMG's final partial block holds the last available prediction without stretching the native sequence. Tables 12 and 13 report the resulting errors.

Table 12: Five-second temporal-alignment sensitivity. Error cells report user-first Overall angle MAE (°) / whole-hand landmark distance (mm). Overall pools eligible windows within each user across conditions before equally weighting unique users. Maximum absolute changes are taken over User, Stage, User+Stage, and Overall; bold marks the largest change for each metric across configurations.

Configuration	Task	Released Overall (°/mm)	Past-only Overall (°/mm)	Max. $\lvert\Delta\rvert$ (°/mm)
emg2pose				
vEMG2Pose†	Regression	14.984/20.204	15.073/20.265	0.090/0.083
vEMG2Pose† (Multi-task)	Regression	14.575/19.483	14.667/19.571	0.092/0.090
vEMG2Pose†	Tracking	10.861/14.680	10.927/14.715	0.068/0.040
STAMP	Regression	12.484/16.456	12.603/16.584	**0.122**/0.129
STAMP	Tracking	9.797/13.098	9.865/13.147	0.071/0.051
EgoEMG				
vEMG2Pose†	Regression	14.953/24.263	14.998/24.326	0.086/0.132
STAMP	Regression	12.956/20.697	13.011/20.874	0.061/**0.193**
vEMG2Pose†	Tracking	11.487/18.445	11.481/18.334	0.013/0.150
STAMP	Tracking	9.790/16.025	9.790/16.026	0.002/0.007

† Retrained vEMG2Pose checkpoints from Table 1; Tracking uses the single-task checkpoint.

Table 13: Five-second emg2pose Tracking under released and past-only alignment. Each cell stacks angle MAE (°) above whole-hand landmark distance (mm), as user-first mean ± population SD. Overall pools windows per user across conditions before the cross-user summary. Bold marks the lower mean, with its SD, for each metric and condition within each alignment.

	Released		Past-only	
Condition	vEMG2Pose†	STAMP	vEMG2Pose†	STAMP
	7.733±0.948	**7.298±0.819**	7.786±0.942	**7.338±0.815**
User	10.287±1.534	**9.470±1.287**	10.325±1.523	**9.495±1.269**
	11.172±1.434	**10.040±1.290**	11.239±1.425	**10.111±1.283**
Stage	15.104±1.810	**13.442±1.643**	15.139±1.800	**13.493±1.633**
	11.063±1.069	**10.162±0.868**	11.131±1.078	**10.228±0.883**
User+Stage	15.435±1.522	**13.937±1.175**	15.475±1.521	**13.985±1.176**
	10.861±1.640	**9.797±1.420**	10.927±1.634	**9.865±1.418**
Overall	14.680±2.139	**13.098±1.872**	14.715±2.130	**13.147±1.865**

† Retrained single-task vEMG2Pose Tracking checkpoint from Table 1.

C Scaling experiments

The user-data study uses fixed nested subsets of 32, 63, 95, 126, and 158 training users (20%–100%), retaining all eligible recordings per user without duration caps. Validation and test data remain fixed. The models are vEMG2Pose (5.980M) and STAMP (6.059M; 1.32% parameter difference). Each configuration uses one training seed and RTX-5090 checkpoints separate from the H200 main comparison. Metrics are summarized across users. Table 14 reports measurements and population SDs; for each mean error E and user count N, we fit

$$E(N) = AN^{-\alpha} \tag{18}$$

in log space without an irreducible-error term.

The joint data–model study uses XS (1.370M), Small (3.590M), Middle (6.059M), and Large (15.900M) at 20%, 60%, and 100% data, all with single-node eight-RTX-5090 DDP. The three

Middle measurements reuse the matching user-data runs. For data fraction D and parameter count P in millions, we fit

$$E(D, P) = E_\infty + AD^{-\alpha} + BP^{-\beta}. \tag{19}$$

Coefficients are estimated separately for angle and landmark error by nonlinear least squares with nonnegative terms, using all twelve cells. Table 15 reports the coefficients, in-grid R^2, RMSE, and maximum residual.

The fixed-20% scan evaluates 12 STAMP configurations from 1.37M to 148.44M parameters. Width, feedforward capacity, and selected depths vary; splits, training, evaluation, and validation selection remain fixed. Results use the same user-first aggregation.

C.1 EMG2POSE USER-DATA AND JOINT DATA–MODEL SCALING

Table 14: emg2pose Regression user-scaling measurements and log-space fits. Cells are cross-user mean ± population SD; ↓ is better. Doubling is the fitted reduction in mean error for a 2× increase in training users.

Train	Users	Angle (°) ↓		Landmark (mm) ↓		Fingertip (mm) ↓	
fraction	N	vEMG2Pose†	STAMP	vEMG2Pose†	STAMP	vEMG2Pose†	STAMP
20%	32	17.876±2.491	**16.977±2.524**	24.461±3.074	**23.170±3.211**	41.667±5.469	**39.331±5.732**
40%	63	16.458±2.203	**15.230±2.364**	22.528±2.901	**20.467±3.074**	38.310±5.127	**34.612±5.454**
60%	95	15.877±2.069	**13.992±2.065**	21.438±2.721	**18.717±2.737**	36.450±4.849	**31.431±4.801**
80%	126	15.330±1.909	**13.173±1.820**	20.686±2.555	**17.315±2.286**	35.041±4.570	**28.907±3.958**
100%	158	14.973±1.719	**12.728±1.475**	20.170±2.346	**16.829±1.985**	34.080±4.190	**28.148±3.479**

Metric	Model	α ↑	log-space R^2	Doubling reduction ↑
Angle	vEMG2Pose†	0.110	**0.997**	7.34%
Angle	**STAMP**	**0.184**	0.996	11.98%
Landmark	vEMG2Pose†	0.121	**1.000**	8.06%
Landmark	**STAMP**	**0.206**	0.994	13.32%
Fingertip	vEMG2Pose†	0.126	**1.000**	8.36%
Fingertip	**STAMP**	**0.217**	0.993	13.97%

† Baselines retrained using the dataset-specific implementations and evaluation protocols in Appendix B.

C.1.1 JOINT DATA–MODEL GRID

Table 15: emg2pose Regression data–model scaling study (single-node 8×RTX 5090 DDP). Panel (a) reports mean ± population standard deviation across test users for every cell; Panel (b) fits the cross-user means with $E(D, P) = E_\infty + AD^{-\alpha} + BP^{-\beta}$, where D is a fraction and P is in millions. Lower error, RMSE, and maximum absolute residual are better.

		20% data		60% data		100% data	
STAMP scale	Params (M)	Angle (°) ↓	Landmark (mm) ↓	Angle (°) ↓	Landmark (mm) ↓	Angle (°) ↓	Landmark (mm) ↓
XS	1.370	17.881±2.041	25.082±2.801	15.317±1.730	20.732±2.359	14.235±1.610	18.977±2.138
Small	3.590	16.879±1.871	23.294±2.544	14.265±1.563	19.225±2.102	13.251±1.496	17.558±1.973
Middle (STAMP)	6.059	16.977±2.524	23.170±3.211	13.992±2.065	18.717±2.737	12.728±1.475	16.829±1.985
Large	15.900	**15.778±1.815**	**21.414±2.408**	**13.591±1.593**	**18.196±2.152**	**12.067±1.520**	**15.810±1.994**

Metric	E_∞	A	α	B	β	R^2 ↑	RMSE ↓	Max. residual ↓
Angle (°)	0 (bound)	10.276	0.194	4.279	0.282	0.9905	0.170	0.360
Landmark (mm)	0 (bound)	13.499	0.224	6.197	0.324	0.9914	0.253	0.532

C.1.2 BROAD CAPACITY SCAN AT FIXED DATA

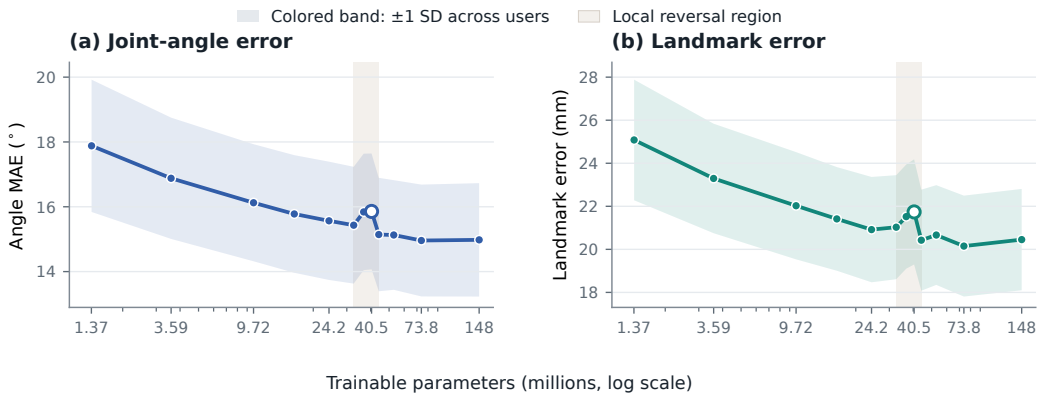


Figure 7: Fixed-20% capacity scan. Markers show cross-user mean angle and landmark errors; bands show ± 1 population SD across 178 users. The shaded vertical region marks the configurations around 40.46M parameters.

Table 16: Complete fixed-20% capacity scan. Test cells are mean $\pm$ population standard deviation across users for one prespecified seed. Validation loss and cross-user mean errors are minimized. Bold marks the shared local mean-error peak and the best post-peak mean for each test metric.

Params (M)	Validation loss ↓	Angle (°) ↓	Landmark (mm) ↓
1.37	0.566831	17.881±2.041	25.082±2.801
3.59	0.544181	16.879±1.871	23.294±2.544
9.72	0.532242	16.125±1.806	22.027±2.492
15.90	0.523752	15.778±1.815	21.414±2.408
24.19	0.524319	15.567±1.825	20.917±2.447
32.60	0.520001	15.431±1.802	21.029±2.417
36.78	0.528782	15.841±1.798	21.520±2.419
40.46	**0.531487**	**15.856±1.788**	**21.739±2.444**
44.14	0.514715	15.143±1.750	20.426±2.348
52.89	0.516613	15.130±1.692	20.665±2.317
73.84	0.515603	**14.959±1.726**	**20.152±2.345**
148.44	0.514246	14.979±1.750	20.453±2.354

D Complete auxiliary metrics

Velocity is the mean prediction velocity magnitude in rad/s. Acceleration and jerk are evaluator-scaled finite differences of predictions, reported in arbitrary units (a.u.).

D.1 Regression diagnostics

Table 17: Complete Regression diagnostics: cross-user mean ± population SD. Panel (a) gives the main comparison; Panel (b) gives the cumulative ablation chain. Velocity, acceleration, and jerk summarize predictions.

Window	Model	Loss ↓	Angle (°) ↓	Landmark (mm) ↓	Velocity (rad/s)	Acc. proxy (a.u.)	Jerk proxy (a.u.)
(a) Main comparison							
5 s	EMGFormer-M†	0.561±0.071	13.987±1.743	18.789±2.369	1.031±0.142	2.0624±0.2839	4.1252±0.5679
5 s	NeuroPose†	0.725±0.080	17.754±1.988	24.238±2.616	0.571±0.100	1.1413±0.2005	2.2828±0.4011
5 s	vEMG2Pose† (Regression)	0.603±0.069	14.984±1.688	20.204±2.334	0.727±0.125	0.0086±0.0013	0.0088±0.0013
5 s	vEMG2Pose† (Multi-task)	0.582±0.069	14.575±1.698	19.483±2.351	0.783±0.140	0.0093±0.0014	0.0095±0.0014
5 s	**STAMP**	**0.492±0.060**	**12.484±1.528**	**16.456±2.003**	0.963±0.155	0.0193±0.0027	0.0196±0.0027
30 s	EMGFormer-M†	0.608±0.103	14.744±2.569	20.621±3.219	0.841±0.146	1.6828±0.2928	3.3657±0.5856
30 s	NeuroPose†	0.700±0.100	17.152±2.510	23.361±3.166	0.529±0.143	1.0578±0.2859	2.1156±0.5719
30 s	vEMG2Pose† (Regression)	0.593±0.082	14.977±2.087	19.706±2.667	0.650±0.189	0.0076±0.0021	0.0085±0.0023
30 s	vEMG2Pose† (Multi-task)	0.558±0.091	14.163±2.271	18.579±2.983	0.704±0.208	0.0085±0.0023	0.0094±0.0026
30 s	**STAMP**	**0.513±0.084**	**12.842±2.046**	**17.298±2.792**	0.896±0.211	0.0178±0.0035	0.0190±0.0037

Window	Configuration	Loss ↓	Angle (°) ↓	Landmark (mm) ↓	Velocity (rad/s)	Acc. proxy (a.u.)	Jerk proxy (a.u.)
(b) Matched architectural ablations							
5 s	vEMG2Pose† (Regression)	0.603±0.069	14.984±1.688	20.204±2.334	0.727±0.125	0.0086±0.0013	0.0088±0.0013
5 s	TDS–Affine	0.598±0.070	14.845±1.704	20.070±2.344	0.875±0.129	0.0138±0.0019	0.0140±0.0019
5 s	STEM–Affine	0.549±0.062	13.802±1.534	18.361±2.068	1.058±0.148	0.0257±0.0029	0.0260±0.0029
5 s	**STAMP**	**0.492±0.060**	**12.484±1.528**	**16.456±2.003**	0.963±0.155	0.0193±0.0027	0.0196±0.0027
30 s	vEMG2Pose† (Regression)	0.593±0.082	14.977±2.087	19.706±2.667	0.650±0.189	0.0076±0.0021	0.0085±0.0023
30 s	TDS–Affine	0.614±0.091	15.357±2.355	20.495±2.998	0.955±0.211	0.0158±0.0034	0.0171±0.0036
30 s	STEM–Affine	0.552±0.075	13.696±1.958	18.736±2.429	1.062±0.226	0.0251±0.0041	0.0265±0.0044
30 s	**STAMP**	**0.513±0.084**	**12.842±2.046**	**17.298±2.792**	0.896±0.211	0.0178±0.0035	0.0190±0.0037

† Baselines retrained using the dataset-specific implementations and evaluation protocols in Appendix B.

E TRACKING AND THIRTY-SECOND LONG-HORIZON DIAGNOSTICS

For thirty-second evaluations, each window contains 60,000 sEMG samples and 1,500 target frames. Windows do not overlap, and model state is not reset within a window. Regression receives no true initial pose; Tracking receives it once at window start. All configurations use the same User (890 windows; $n = 20$ users), Stage (1,706; $n = 156$), and User+Stage (222; $n = 19$) splits and skip the same inverse-kinematics failures. We first aggregate eligible windows within each user and then report the mean and population standard deviation across users; the Overall column combines each of 176 unique users' eligible split windows before the cross-user summary.

E.1 FIVE-SECOND TRACKING DIAGNOSTICS

The five- and thirty-second vEMG2Pose Tracking results use the same single-task checkpoint as Table 1.

Table 18: Five-second Tracking diagnostics: mean ± population SD across 178 users. Panel (a) reports accuracy and prediction-motion statistics; Panel (b) reports anatomical angular MAE.

Model	Loss ↓	Angle (°) ↓	Landmark (mm) ↓	Velocity (rad/s)	Acc. proxy (a.u.)	Jerk proxy (a.u.)
(a) User-first diagnostics						
EMGFormer-M†	0.463±0.070	11.624±1.792	15.621±2.311	0.782±0.156	1.5649±0.3114	3.1301±0.6228
NeuroPose†	0.559±0.081	13.607±2.065	19.144±2.674	0.303±0.087	0.6056±0.1738	1.2113±0.3477
vEMG2Pose†	0.435±0.065	10.861±1.640	14.680±2.139	0.623±0.135	0.0055±0.0012	0.0057±0.0013
STAMP	**0.388±0.056**	**9.797±1.420**	**13.098±1.872**	0.694±0.141	0.0103±0.0020	0.0105±0.0020

Model	Thumb (°) ↓	Index (°) ↓	Middle (°) ↓	Ring (°) ↓	Pinky (°) ↓	Proximal (°) ↓	Mid (°) ↓	Distal (°) ↓
(b) Anatomical angular MAE								
EMGFormer-M†	9.43±1.65	11.65±2.01	11.77±2.05	11.92±2.00	13.35±2.12	9.60±1.39	14.73±2.66	12.56±2.43
NeuroPose†	10.50±1.81	14.01±2.34	14.16±2.45	14.04±2.27	15.33±2.46	11.22±1.60	17.95±3.21	14.03±2.80
vEMG2Pose†	9.29±1.54	10.83±1.86	10.84±1.86	10.86±1.76	12.48±1.96	8.86±1.21	13.90±2.53	11.84±2.16
STAMP	**8.58±1.41**	**9.74±1.59**	**9.71±1.63**	**9.78±1.57**	**11.18±1.67**	**8.12±1.09**	**12.15±2.11**	**10.80±1.93**

† Baselines retrained using the dataset-specific implementations and evaluation protocols in Appendix B.

Table 19: Thirty-second emg2pose accuracy: mean ± population SD across users (User: $n = 20$; Stage: $n = 156$; User+Stage: $n = 19$; Overall: 176 unique users). Overall pools eligible windows per user before summarizing users. Bold marks the lowest mean within each task and metric.

	User		Stage		User+Stage		Overall	
Model	Angle (°) ↓	Landmark (mm) ↓	Angle (°) ↓	Landmark (mm) ↓	Angle (°) ↓	Landmark (mm) ↓	Angle (°) ↓	Landmark (mm) ↓
Regression								
EMGFormer-M†	11.920±1.458	16.033±1.727	15.026±2.539	21.067±3.071	15.268±1.696	21.730±2.190	14.744±2.569	20.621±3.219
NeuroPose†	12.377±1.048	16.634±1.573	17.644±2.184	24.032±2.629	17.394±1.472	24.258±1.641	17.152±2.510	23.361±3.166
vEMG2Pose† (Regression)	12.260±1.023	15.423±1.562	15.237±2.031	20.105±2.492	15.672±1.795	21.237±2.105	14.977±2.087	19.706±2.667
vEMG2Pose† (Multi-task)	11.239±1.262	14.300±1.547	14.451±2.191	18.984±2.858	14.726±1.769	20.016±2.255	14.163±2.271	18.579±2.983
STAMP (absolute)	11.607±1.309	14.892±1.570	12.935±2.096	17.484±2.849	14.350±1.625	19.876±1.799	12.842±2.046	17.298±2.792
STAMP (long-horizon RoPE)	**11.112±1.340**	**14.059±1.689**	**11.939±1.830**	**16.091±2.473**	**13.729±1.756**	**18.752±2.305**	**11.901±1.790**	**15.962±2.434**
Tracking; one true initial pose								
EMGFormer-M†	10.086±2.386	14.651±3.554	17.834±3.697	24.734±4.051	17.452±2.809	25.022±3.282	17.109±4.109	23.808±4.747
NeuroPose†	10.623±2.575	15.661±3.872	19.055±4.094	26.641±4.539	18.708±3.068	26.995±3.545	18.269±4.521	25.635±5.261
vEMG2Pose† (Tracking)	**8.376±1.554**	**11.066±2.279**	13.531±2.115	18.002±2.577	13.458±1.835	18.802±2.565	13.055±2.462	17.385±3.087
STAMP (absolute)	14.047±1.106	22.792±2.004	15.028±1.919	23.083±2.927	15.061±1.310	23.597±2.077	14.936±1.857	23.064±2.821
STAMP (long-horizon RoPE)	9.050±1.190	11.804±1.693	**11.862±1.579**	**16.045±2.161**	**12.497±1.518**	**17.293±2.255**	**11.618±1.696**	**15.685±2.351**

† Baselines retrained using the dataset-specific implementations and evaluation protocols in Appendix B.

E.2 Thirty-second cross-model accuracy

E.3 Long-horizon position treatment

The absolute sinusoidal model was trained over positions up to 294, corresponding to 5.9 seconds at 50 Hz, but a thirty-second window reaches 1,500 output tokens. We evaluate four reported treatments in Table 20: the original Absolute baseline uses index t; MOD-295 applies $t \bmod 295$ to that frozen model; RoPE replaces the absolute code and adapts the converged checkpoint to rotary positions at index t; and RoPE+MOD-295 applies the wrapped index to the same RoPE checkpoint without further optimization. As a matched optimization control, we also continue the Absolute model for ten epochs with the same data and learning-rate schedule used for RoPE; this control does not improve the thirty-second evaluation.

Table 20: Thirty-second position treatments. Accuracy cells give cross-user mean ± population SD for Angle (°) / Landmark (mm); lower is better. Overall pools eligible windows per user before summarizing 176 unique users. Prediction-motion diagnostics use the same user-first aggregation.

Configuration	User	Stage	User+Stage	Overall
(a) Regression: Angle / Landmark				
Absolute baseline	11.607±1.309/14.892±1.570	12.935±2.096/17.484±2.849	14.350±1.625/19.876±1.799	12.842±2.046/17.298±2.792
MOD-295	**11.019±1.295/13.958±1.609**	12.007±1.784/16.103±2.302	13.737±1.725/**18.723±2.174**	11.952±1.751/15.962±2.280
RoPE	11.112±1.340/14.059±1.689	**11.939±1.830/16.091±2.473**	**13.729±1.756**/18.752±2.305	**11.901±1.790/15.962±2.434**
RoPE+MOD-295	11.124±1.332/14.064±1.675	11.976±1.833/16.102±2.451	13.800±1.764/18.814±2.301	11.936±1.794/15.973±2.413
(b) Tracking: Angle / Landmark				
vEMG2Pose† (Tracking)	**8.376±1.554/11.066±2.279**	13.531±2.115/18.002±2.577	13.458±1.835/18.802±2.565	13.055±2.462/17.385±3.087
Absolute baseline	14.047±1.106/22.792±2.004	15.028±1.919/23.083±2.927	15.061±1.310/23.597±2.077	14.936±1.857/23.064±2.821
MOD-295	10.814±1.150/15.700±1.842	12.923±1.698/18.267±2.529	13.364±1.215/19.223±1.815	12.738±1.725/18.051±2.525
RoPE	9.050±1.190/11.804±1.693	**11.862±1.579/16.045±2.161**	**12.497±1.518/17.293±2.255**	**11.618±1.696/15.685±2.351**
RoPE+MOD-295	9.048±1.194/11.786±1.692	11.878±1.579/16.082±2.165	12.515±1.506/17.335±2.226	11.632±1.698/15.718±2.360

(c) Overall user-first prediction-motion diagnostics

Task	Configuration	Velocity (rad/s)	Acc. proxy (a.u.)	Jerk proxy (a.u.)
Regression	Absolute baseline	0.896089±0.211118	0.017765±0.003472	0.018966±0.003739
Regression	MOD-295	0.896159±0.221297	0.017517±0.003632	0.018715±0.003916
Regression	RoPE	0.859333±0.212608	0.016589±0.003304	0.017725±0.003577
Regression	RoPE+MOD-295	0.857832±0.212022	0.016571±0.003295	0.017704±0.003568
Tracking	Absolute baseline	0.714259±0.202828	0.011768±0.002878	0.012747±0.003139
Tracking	MOD-295	0.698910±0.210351	0.011339±0.003003	0.012289±0.003276
Tracking	RoPE	0.717753±0.220801	0.011844±0.003233	0.012822±0.003520
Tracking	RoPE+MOD-295	0.712146±0.219582	0.011763±0.003221	0.012733±0.003505

† Baseline retrained using the dataset-specific implementation and evaluation protocol in Appendix B.

F Ablation diagnostics

F.1 Anatomical breakdown of the ablation chain

Table 21 uses the same four-stage Regression ablation chain as Table 2.

Table 21: emg2pose five-second Regression-only anatomical angular MAE. Cells are mean ± population standard deviation across users in degrees for one prespecified seed; lower is better.

Model	Thumb (°) ↓	Index (°) ↓	Middle (°) ↓	Ring (°) ↓	Pinky (°) ↓	Proximal (°) ↓	Mid (°) ↓	Distal (°) ↓
vEMG2Pose†	13.60±2.00	13.90±1.90	15.25±2.15	15.05±2.07	17.11±2.07	11.66±1.29	20.08±2.93	16.54±2.32
TDS–Affine	13.40±2.04	13.85±1.89	15.01±2.18	14.92±2.08	17.03±2.13	11.57±1.27	19.96±3.01	16.28±2.35
STEM–Affine	12.64±1.92	12.80±1.63	13.68±1.89	13.94±1.98	15.95±2.04	10.86±1.16	18.12±2.61	15.37±2.18
STAMP	**11.39±1.79**	**11.79±1.67**	**12.24±1.92**	**12.57±1.84**	**14.44±1.99**	**9.96±1.15**	**16.05±2.56**	**13.97±2.13**

† Baselines retrained using the dataset-specific implementations and evaluation protocols in Appendix B.

F.2 Detailed architectural replacement results

Table 22: Five-second emg2pose Regression. Cells stack angle MAE (°) above landmark distance (mm), each as mean ± population SD across users. Overall first pools windows within each user across splits, then summarizes 178 unique users. Bold marks the lowest mean for each metric and column.

Replacement	Params (M)	User	Stage	User+Stage	Overall
		11.65±1.33	**12.52±1.55**	**14.28±1.39**	**12.48±1.53**
STAMP (full model)	6.059	**14.65±1.64**	**16.57±2.02**	**19.27±2.01**	**16.46±2.00**
STEM encoder →		12.01±1.30	14.24±1.68	15.13±1.44	14.06±1.71
Temporal Transformer	6.060	15.40±1.70	19.20±2.22	20.77±2.02	18.89±2.33
Circular encoding →		11.77±1.36	12.99±1.53	14.44±1.48	12.91±1.53
Linear sinusoidal	6.059	14.69±1.70	17.17±2.01	19.38±1.94	17.00±2.03
Motion MoE →		11.72±1.12	13.70±1.48	14.68±1.37	13.54±1.51
Wider backbone	6.056	14.89±1.41	18.21±1.94	19.82±1.84	17.95±2.03

G Additional mechanism details

The module probes use one fixed checkpoint and 24 five-second windows: eight each from User, Stage, and User+Stage.

G.1 Selected motion excerpts

For each split, we scan 160 distinct-recording five-second test windows at 50 Hz and discard the first 0.5 s after reset. Candidate 0.6 s excerpts are ranked by

$$s = E^{\text{tip}}_{\text{vEMG2Pose}} - 1.5E^{\text{tip}}_{\text{STAMP}}. \tag{20}$$

We select three excerpts per condition from different recordings, favoring distinct action patterns, and keep these nine excerpts fixed. FingerLand averages three-dimensional error over the 19 finger landmarks, excluding wrist and palm center. Angle and FingerLand labels average all 31 frames per excerpt. Rendered poses use a shared view and coordinate range at $t = -0.3$ s, 0 s, and $+0.3$ s; headings identify the source recording stage.

Table 23: Selected-excerpt errors: mean ± population SD across three fixed 0.6 s excerpts per condition, selected by Equation 20. FingerLand and angle errors are minimized.

	FingerLand (mm) ↓		Angle (°) ↓	
Condition	vEMG2Pose†	STAMP	vEMG2Pose†	STAMP
User	28.6±2.6	**8.2±0.5**	22.0±4.2	**10.6±2.3**
Stage	31.8±3.5	**7.9±1.9**	30.4±4.5	**6.1±0.6**
User+Stage	32.8±4.3	**9.1±1.4**	25.4±3.8	**8.7±0.7**

† Baselines retrained using the dataset-specific implementations and evaluation protocols in Appendix B.

G.2 STEM ATTENTION STATISTICS

Table 24: STEM block-wise attention summaries. Neighbor mass covers self and circular ± 1 electrodes (uniform baseline 18.75%); the uniform mean ring-distance baseline is approximately 4. Temporal mass is the share assigned to the latest 0.5 s. Half-lives summarize the eight temporal heads per block.

	Spatial attention				Temporal attention			
Block	Neighbor mass (%) ↑	Ring dist. ↓	Key CV	Query sim.	Mean lag (s)	Latest 0.5 s (%)	Median $t_{1/2}$ (s)	Max $t_{1/2}$ (s)
1	25.73	3.294	0.135	0.755	0.648	64.50	0.656	4.233
2	**31.85**	**3.060**	0.139	0.741	0.671	62.81	0.674	5.535
3	24.95	3.323	0.191	0.874	**0.711**	61.54	**0.834**	**5.545**
4	25.18	3.565	**0.486**	0.906	0.638	63.74	0.415	3.647
5	21.10	3.813	0.380	**0.974**	0.625	63.73	0.617	2.810



Figure 8: Selected User-split motion excerpts. Rows show ground truth, vEMG2Pose, and STAMP; columns show $t = -0.3$ s, 0 s, and $+0.3$ s in purple, row colors, and gold. Labels report clip-mean angular MAE (°) and FingerLand (mm) over all 31 frames; lower is better.

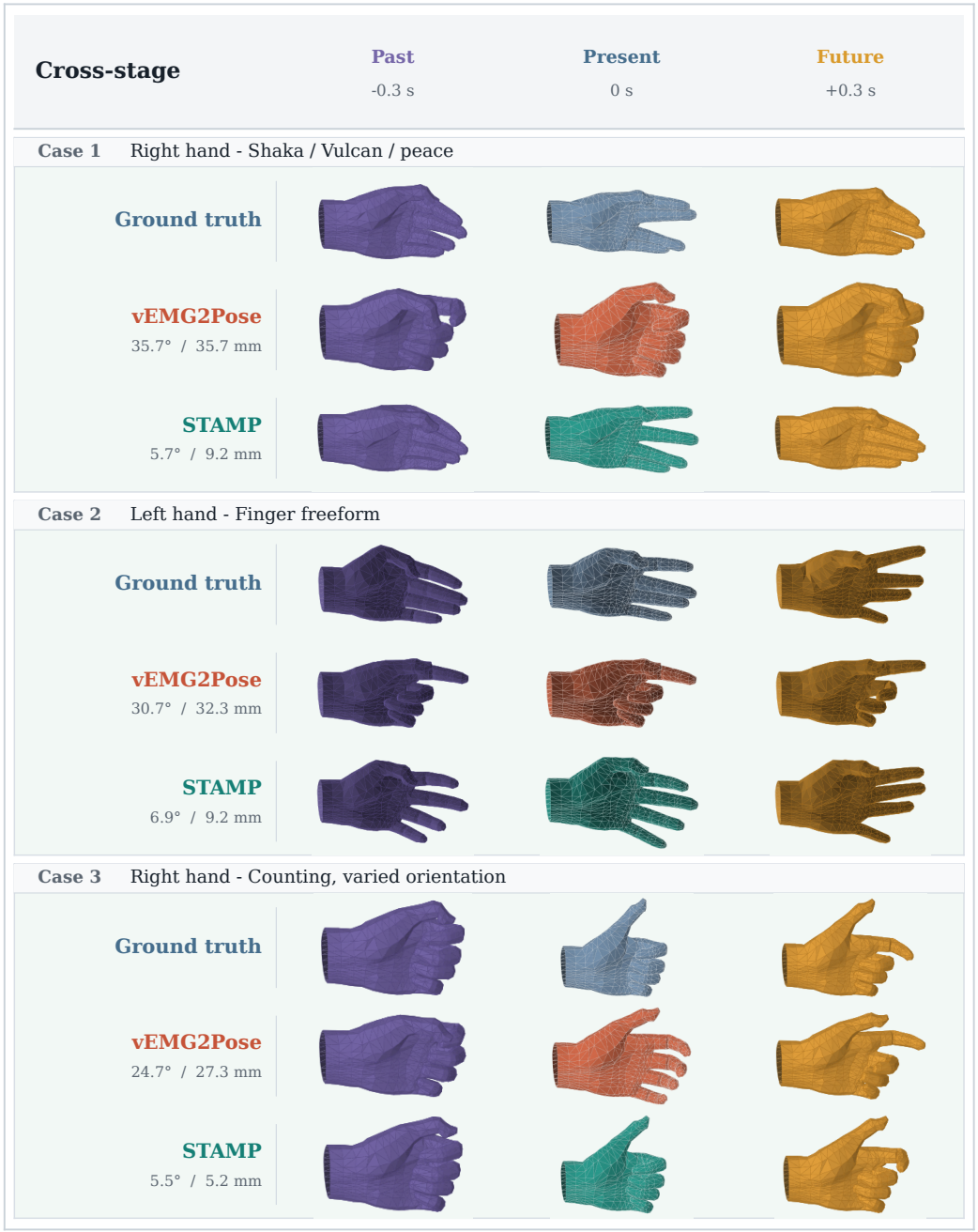


Figure 9: Selected Stage-split motion excerpts. Layout and metrics follow Figure 8.



Figure 10: Selected User+Stage-split motion excerpts. Layout and metrics follow Figure 8.

G.3 State-perturbation protocol and measurements

For two trajectories receiving identical $(\mathbf{a}_t, \mathbf{b}_t)$ commands,

$$\delta\mathbf{x}_t = \mathbf{a}_t \odot \delta\mathbf{x}_{t-1} = \left(\prod_{i=1}^{t} \mathbf{a}_i\right) \odot \delta\mathbf{x}_0. \tag{21}$$

If $0 \le a_{t,j} \le \rho < 1$ uniformly, then $\|\delta\mathbf{x}_t\|_\infty \le \rho^t \|\delta\mathbf{x}_0\|_\infty$.

We inject $+10^\circ$ into the mid-sequence `INDEX_MCP_FE` state and hold subsequent commands fixed. The measured half-life is 0.12 s; 90% decay takes 0.44 s. Forced $g = 1$ retains the offset, while $g = 0$ removes it in one step. The mean gate is 0.8763 (95% bootstrap CI $[0.8472, 0.9016]$), with a constant-gate half-life of 0.105 s.

Propagating the perturbed coarse state through the second stage gives a mean gate change of $+0.000559$ over the next 0.5 s (95% CI $[+0.000409, +0.000719]$; positive in 23/24 windows). Mean absolute final-output change is 0.148° (95% CI $[0.113^\circ, 0.185^\circ]$), or 1.48% of the injected offset.

G.4 Direction–routing statistics

The routing probe uses 5,559 tokens with target finger speed above 0.15 rad/s, assigned to ten signed flexion/extension directions. Enrichment is normalized by each window's mean routing. Direction labels are circularly shifted within windows 500 times to obtain the null distribution; permutation p-values use denominator 501.

Across the post-encoder, inter-stage, and post-attention routers, respectively, the direction cNMI is 0.1840, 0.1964, and 0.1703; the corresponding circular-shift 95th percentiles are 0.1320, 0.1367, and 0.1264. Each test yields $p = 0.002$. The routers activate 21, 23, and 13 distinct top-1 experts, while deployment uses top-$k = 8$ routing.

H Causality audit

For an input sequence $\mathbf{X}_{1:T}$ and cut point $t < T$, we compare the first t predictions from a full run with the truncated prefix using $\epsilon_{\text{prefix}}(t) = \max_{i \le t} \|f(\mathbf{X}_{1:T})_i - f(\mathbf{X}_{1:t})_i\|_\infty$. A causal implementation should satisfy $\epsilon_{\text{prefix}}(t) < 10^{-5}$ up to floating-point ordering effects. We independently compare batched evaluation with sample-by-sample streaming while carrying all states listed in Section 3. The current audit contains 23 tests, all of which pass, covering the full 50 Hz token stream, both Affine passes, future-input perturbations, batched streams, per-stream reset, long-sequence cache boundaries, causalized comparison models, offline controls, and shared-component initialization.

I Streaming latency benchmark

We benchmark the deployed STAMP engine at batch size one and 50 Hz. Each synchronous call accepts 40 new samples from 16 channels at 2 kHz and returns one 20-dimensional angle vector copied to CPU memory. The timer spans input checks and buffering, CPU-to-GPU transfer, one complete CUDA Graph replay, all recurrent-state updates, GPU-to-CPU transfer, finite-value checks, and output assembly. It excludes acquisition of the 40 new samples, input-file I/O, online filtering, process IPC, subject calibration, MANO rendering, GUI submission, and display refresh. The 1,790-sample (0.895 s) initial context, model loading, compilation, and graph capture are start-up requirements and are also excluded.

All variants use the streaming FLA recurrence, a 250-token decoder cache, top-8 execution from 32 experts at three insertion points, and modulo-295 positions. The 6.059M checkpoint matches the frozen MOD-295 accuracy configuration. Per model, four independent processes each run 300 warm-up steps and 2,000 timed steps. Model order is rotated across processes, and all 8,000 measurements are pooled.

Table 25: Steady-state synchronous streaming latency at batch size one. All latency columns are milliseconds; lower is better. The deadline is 20 ms at the 50 Hz output rate.

Configuration	Params (M)	P50 (ms) ↓	P95 (ms) ↓	P99 (ms) ↓	Mean (ms) ↓
STAMP-XS	1.373	**0.7213**	**0.8218**	**0.9023**	**0.7381**
STAMP-Small	3.588	0.8529	0.9447	1.0143	0.8681
STAMP (Middle)	6.059	1.0296	1.1348	1.2155	1.0452
STAMP-Large	15.902	1.9542	2.0769	2.1779	1.9705

I.1 Environment and numerical checks

The benchmark uses a laptop with an RTX 5060 Laptop GPU (8,151 MiB reported VRAM; driver 592.01), Intel Core Ultra 9 275HX, and 31.44 GiB memory. The native Windows 11 build 26200 environment uses Python 3.11.15, PyTorch 2.8.0+cu128, CUDA 12.8, cuDNN 9.10.2, fla-core 0.5.2, triton-windows 3.4.0.post21, and NumPy 2.2.6.

The FP32 engine disables autocast and synchronizes timing through `cpu().numpy().copy()`. Thirty deployment tests pass. Over 310 tokens, sparse FLA differs from the reference by at most 0.000492° (Small) and 0.000143° (XS); CUDA Graph and sparse eager execution agree over the first 320 tokens of every capacity.