FOCUS: A FOUNDATION MODEL FOR WEARABLE A-MODE ULTRASOUND

Anonymous authors
Paper under double-blind review

ABSTRACT

Wearable A-mode ultrasound (AUS) enables continuous, non-invasive monitoring of deep tissue, and advances in conformable transducers and low-power electronics have enabled increasingly large and diverse wearable datasets. Because wearable probes communicate over bandwidth-limited wireless links that cannot stream raw echo data continuously, processing must largely run on the device, calling for compact models that can be adapted to each user with limited labeled data. However, learning approaches for wearable AUS are developed for individual tasks, subjects, or acquisition systems and trained from scratch on small datasets, and no foundation model (FM) exists for this setting. We introduce FOCUS, a compact FM for wearable AUS (6.3 M encoder parameters), pre-trained on 24.7 million unlabeled A-mode scans from five wearable and benchtop acquisition systems and multiple body locations, including the neck, upper and lower limbs. A frequency-aware multi-tokenizer maps raw signals acquired at different sampling frequencies to patches spanning an approximately constant physical depth. We evaluate FOCUS on eight subject-specific downstream tasks from five wearable datasets, spanning hand-gesture, forearm-position, and gait-phase classification, as well as upper- and lower-limb kinematic regression. Pre-training improves mean performance over the same architecture trained from scratch on all eight tasks, and fine-tuning with only 40% of the labeled training data outperforms training from scratch with the complete labeled dataset on all eight. FOCUS also outperforms the previously reported results on seven of the eight benchmarks. Across three model capacities (0.8 M, 6.3 M, and 85 M parameters), downstream performance generally improves with model size, and larger models benefit more consistently from additional pre-training data. FOCUS requires 669–856 MMACs per channel update under sequential inference, a compute budget compatible with future on-device deployment for wearable AUS.

1 INTRODUCTION

A-mode ultrasound (AUS) is emerging as a promising modality for low-cost, continuous wearable monitoring in both clinical and everyday settings (Zhou et al., 2025a; Weik et al., 2026). Unlike surface-only sensing modalities, AUS can probe structures from superficial to deeper tissues, with applications spanning musculoskeletal assessment (Brausch et al., 2022; Qu et al., 2024; Spacone et al., 2025a; Li et al., 2026), hand-gesture and motion recognition (Yang et al., 2024; Grandi Sgambato et al., 2025; Spacone et al., 2025b), cardiovascular and blood-pressure monitoring (Frey et al., 2022; Giordano et al., 2025; Zhou et al., 2025b), fetal monitoring (Park et al., 2026), and bladder and internal-organ assessment (Zhang et al., 2024; Zhou et al., 2025a). Wearable probes typically communicate over bandwidth-limited, low-power wireless links that cannot sustain continuous streaming of raw echo data (Frey et al., 2022); processing must therefore largely run on the device, which calls for compact models.

Unlike B-mode Ultrasound (US), which combines signals from multiple array elements through beamforming to reconstruct two-dimensional images, AUS records one-dimensional sequences of echo amplitudes along the beam axis, with echo time-of-flight encoding propagation depth (Kremkau, 2015). Models developed for images therefore cannot be directly applied to raw AUS scans. Existing learning approaches for wearable AUS have instead been developed for specific acquisition platforms, anatomical locations, sensor configurations, and downstream tasks, and

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.

are typically trained from scratch on small, task-specific datasets. This has limited the development of general-purpose representations that can be shared across datasets and applications.

A central challenge is the heterogeneity of acquisition hardware. Existing systems differ in sampling frequency, number of samples per scan, transducer characteristics such as center frequency and bandwidth, and the number and spatial arrangement of sensing units. The sampling frequency determines the resolution along the physical depth axis, while the number of samples together with the sampling frequency determines the total sampled depth. Transducer center frequency and bandwidth further influence the spectral content and morphology of the acquired echoes. Consequently, different systems can produce different representations of the same underlying physical domain. Despite this heterogeneity, AUS acquisitions share a common physical basis: independently of anatomical location, acquisition platform, or downstream application, the measured signals arise from acoustic propagation through tissue and reflections generated by variations in acoustic impedance. Heterogeneous AUS datasets may therefore contain common information that a general-purpose representation could capture, provided the physical interpretation of the signals is preserved. A second challenge is the limited availability of labels: wearable models are typically calibrated for each user, so the labeled data available for a given subject and task are limited.

Foundation Models (FMs) provide a natural framework for both challenges. Within B-mode US, recent FMs have demonstrated transfer across anatomical regions and downstream tasks (Jiao et al., 2024; Zhang et al., 2025; Yan et al., 2026; Ma et al., 2026), but they operate on reconstructed images. Masked autoencoding has been applied directly to AUS signals (RoßTeutscher et al., 2025), but for a fixed acquisition setting and a single downstream application. To the best of our knowledge, no FM has yet been developed for wearable AUS that learns jointly across heterogeneous datasets, acquisition systems, and downstream tasks. In this work, we address this research gap with the following contributions:

- We introduce **FOCUS**, to the best of our knowledge the first open-source FM for wearable AUS, jointly pre-trained across heterogeneous datasets and acquisition systems. A frequency-aware multi-tokenizer maps raw signals acquired at different sampling frequencies to patches spanning a similar physical depth. We pre-train FOCUS using masked self-supervision on 24.7 million A-mode scans collected with five wearable and benchtop acquisition systems at several anatomical locations (wrist, forearm, upper arm, lower limbs, neck).
- We evaluate FOCUS on eight subject-specific classification and regression tasks from five wearable datasets, spanning hand-gesture, forearm-position, and gait-phase classification, as well as upper- and lower-limb kinematic regression. Pre-training improves mean performance over the same architecture trained from scratch on all eight tasks, while FOCUS outperforms the previously reported results on seven of the eight benchmarks. Moreover, fine-tuning with only 40% of the available labeled training data outperforms the full-label from-scratch baseline across all eight tasks, demonstrating substantially improved label efficiency.
- We investigate model capacity and pre-training data scaling across three configurations (0.8 M, 6.3 M, and 85 M parameters): downstream performance generally improves with model size, and larger models benefit more consistently from additional pre-training data. With 6.3 M encoder parameters and 669–856 MMACs per channel update under sequential inference, FOCUS-Small shows promise for future on-device execution in wearable AUS systems.

2 Related Work

Learning approaches for AUS. Early AUS pipelines predominantly relied on conventional signal processing, including filtering, envelope extraction, and handcrafted feature extraction, followed by classical machine-learning methods (Yang et al., 2022; Guo et al., 2022; Grandi Sgambato et al., 2024). More recent approaches learn representations directly from raw or minimally processed AUS signals, predominantly with convolutional architectures that either process each A-mode channel independently before feature concatenation (Lian et al., 2024; Spacone et al., 2026b) or arrange multiple channels into a joint 2D representation (Grandi Sgambato et al., 2025). Transformer-based architectures have also been explored. Lykourinas et al. (2026) showed on the UltraPro dataset

of Yang et al. (2022) that a ViT performs substantially worse than convolutional architectures on raw RF signals while benefiting from envelope-based preprocessing. In multimodal systems, Li et al. (2026) and Du et al. (2026) encode multichannel AUS frames with 1D Convolutional Neural Network (CNN) tokenizers before Modeling temporal and cross-modal interactions with surface electromyography (sEMG) using Transformer encoders.

Only a few studies have explored unsupervised representation learning directly for AUS. Vostrikov et al. (2024) trained a 1D convolutional autoencoder on raw AUS traces, transferring the learned encoder across unseen subjects for gesture recognition. NVIDIA's NV-Raw2Insights-US (NVIDIA, 2026) employed self-supervised reconstruction to pre-train a 1D CNN encoder on raw, pre-beamformed data, transferring the frozen representation to speed-of-sound estimation. Roß-Teutscher et al. (2025) employed masked autoencoder pre-training with a Transformer encoder on synthetic AUS signals, demonstrating transfer to real measurements and improved label efficiency for time-of-flight estimation. However, these approaches focus on single acquisition settings and downstream applications; general-purpose, joint representation learning across heterogeneous raw AUS datasets remains unexplored.

Foundation models for B-mode ultrasound. Jiao et al. (2024) introduced *USFM*, pre-trained on more than three million US images using spatial-frequency dual masked image modeling, which Ma et al. (2026) distilled into the 5.5 M-parameter *TinyUSFM*. Similarly, Zhang et al. (2025) introduced *EchoCare*, pre-trained with masked autoencoding on 4.5 million images and evaluated across ten downstream tasks. These works show that large-scale heterogeneous pre-training with masked reconstruction yields transferable US representations. However, they operate on reconstructed two-dimensional images and cannot be directly applied to the one-dimensional raw signals acquired by A-mode systems.

Learning from heterogeneous signal representations. Related domains incorporate differences in sampling, channel configuration, and physical scale into shared models. *MOIRAI* (Woo et al., 2024) uses multiple patch sizes to process time series of different frequencies within a shared Transformer, *LUNA* (Döner et al., 2025) handles heterogeneous channel configurations in biosignals, and *M3FM* (Niu et al., 2025) incorporates voxel size during masked pre-training to relate discrete inputs to physical space. FOCUS applies this principle to AUS, where the sampling frequency determines the physical depth spanned by each sample.

3 METHODS

Figure 1 gives an overview of FOCUS. Throughout this work, we refer to each one-dimensional AUS acquisition as a *scan*. A sensing system may acquire multiple scans from different transducers, which we refer to as *channels*, either sequentially or concurrently.

3.1 PROBLEM FORMULATION

We represent each AUS scan as a one-dimensional signal $\mathbf{us} = [us_1, \ldots, us_N]^\top \in \mathbb{R}^N$, where N denotes the number of acquired samples and us_n the measured echo amplitude at the n-th sample. Given a sampling frequency f_s, the n-th sample corresponds to an approximate depth $d_n = cn/(2f_s)$, where $c \approx 1540$ m/s is the assumed speed of sound in tissue and the factor $1/2$ accounts for the round-trip propagation of the US pulse. Depth is expressed relative to the first acquired sample; any platform-specific acquisition delay only adds a constant offset. A measurement may comprise C channels. As in prior learning-based AUS approaches (Yang et al., 2022; Lian et al., 2024; Spacone et al., 2025b), scans acquired within the same acquisition cycle are treated as a single multichannel observation $\mathbf{U} = [\mathbf{us}_1, \ldots, \mathbf{us}_C] \in \mathbb{R}^{C \times N}$, without explicitly modelling the temporal offsets of sequentially acquired channels.

3.2 FREQUENCY-AWARE TOKENIZATION

Each AUS scan is treated independently and divided into non-overlapping patches. Since the pre-training datasets are acquired at different f_s, a fixed patch size in samples would span different

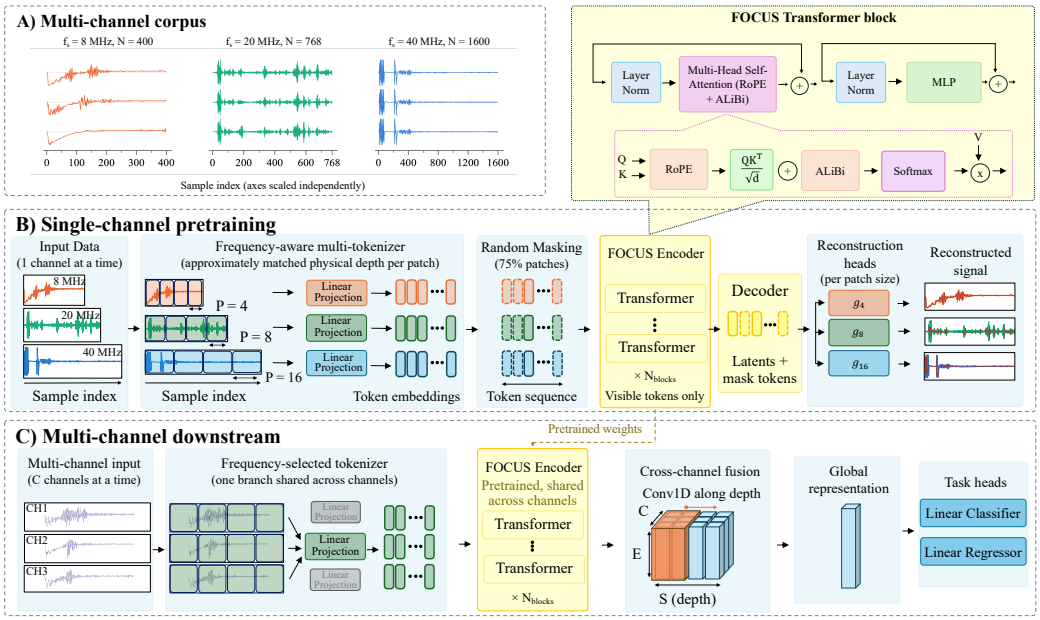


Figure 1: Overview of FOCUS. (A) Heterogeneous pre-training corpus. (B) Frequency-aware tokenization and single-channel masked autoencoding. (C) Downstream adaptation with cross-channel fusion.

physical depth ranges across datasets. Inspired by MOIRAI (Woo et al., 2024), we therefore adapt the patch size to f_s.

A patch of P samples spans $\Delta d(P, f_s) = cP/(2f_s)$. Given the available patch sizes $\mathcal{P} = \{4, 8, 16\}$, we select $P^*(f_s) = \arg\min_{P \in \mathcal{P}} |\Delta d(P, f_s) - \Delta d_{\text{patch}}|$, with $\Delta d_{\text{patch}} = 0.3$ mm, a trade-off between physical depth resolution and sequence length. For the acquisition systems considered, the resulting patch depths range from 0.308 to 0.385 mm. This design allows FOCUS to use the raw, unfiltered signals on their native sample grids, without resampling; at deployment, only the tokenizer branch corresponding to the target sampling frequency is required. Signals are routed through the corresponding tokenizer branch $\mathcal{B}^{P^*}$ (Figure 1), producing $S = \lfloor N/P^*(f_s) \rfloor$ tokens ($62 \leq S \leq 100$ across the pre-training datasets). Sequences are padded within each batch, and a boolean *validity mask* prevents padded positions from contributing to subsequent operations, including attention. Each patch is linearly projected into an E-dimensional embedding ($E = 256$, unless otherwise stated), yielding $\mathbf{X} \in \mathbb{R}^{S \times E}$.

3.3 Masked autoencoder pre-training

We pre-train an asymmetric encoder–decoder Masked Autoencoder (MAE) (He et al., 2022) (as in RoßTeutscher et al. (2025)) with 8 and 4 pre-layer-normalization Transformer blocks (Appendix A.3) and a masking ratio of $r = 0.75$ (He et al., 2022; RoßTeutscher et al., 2025); padded tokens are never masked. Token positions are encoded with Rotary positional encoding (RoPE) (Su et al., 2024), and Attention with Linear Biases (ALiBi) (Press et al., 2021) adds a locality prior, $\text{score}_{ij}^{(h)} \leftarrow \text{score}_{ij}^{(h)} - m_h|p_i - p_j|$, computed from the original patch indices p_i, p_j so that relative positions are preserved under random masking (Appendix A.7). Decoded tokens are mapped back to raw samples by a linear head $g_{P^*(f_s)} : \mathbb{R}^{E_{\text{dec}}} \to \mathbb{R}^{P^*(f_s)}$, selected by the same routing as the tokenizer. Instead of the squared error of the original MAE, we use the Huber loss ℓ_δ, averaged over the samples of each patch into e_i, and add a weak term over the visible tokens $\mathcal{V}$ to the loss over the masked tokens $\mathcal{M}$: $\mathcal{L}_{\text{rec}} = |\mathcal{M}|^{-1} \sum_{i \in \mathcal{M}} e_i + \alpha |\mathcal{V}|^{-1} \sum_{i \in \mathcal{V}} e_i$, with $\alpha = 0.2$, computed in the normalized signal space (Section 4.1).

Table 1: Pre-training corpus. P is the patch size chosen by the frequency-aware tokenizer (all patches span 0.31–0.39 mm). *Share*: fraction of scans in the full corpus and in one epoch after capping the two largest sources at 200k scans per epoch. Acquisition details in Appendix A.1.

			f_s	f_c			Depth		**Share** [%]	
Source	**Site**	**Platform**	[MHz]	[MHz]	N	P	[mm]	**Scans**	corpus	epoch
Multi-site (new)	upper, lower limbs neck, jaw	WULPUS	8	2.25	400	4	38.5	0.54 M	2.2	28.9
Pulse	wrist	Custom	20[a]	10	500[a]	8	19.3	0.45 M	1.8	23.9
Biceps, calf	upper arm, calf	Custom	40	3.5	1600[b]	16	30.8	0.49 M	2.0	26.0
Gastrocnemius	lower leg	Verasonics	20	2–8	768	8	29.6	15.90 M	64.2	10.6
K2MUSE	lower limb	ELONXI	20	5	800[b]	8	30.8	7.37 M	29.8	10.6
Total		5 platforms						24.75 M	100	100

[a] Decimated from 2 GHz, yielding 500 samples from the original 50000. **b** Long scans are truncated to N samples (Further details in Appendix A.1).

3.4 Downstream adaptation and cross-channel fusion

During pre-training, each AUS channel is processed independently. Downstream tasks, however, may rely on complementary information across channels. For downstream adaptation, we therefore retain the pre-trained tokenizer and encoder, discard the decoder and reconstruction heads, and provide all valid tokens to the encoder without masking. Each channel is encoded independently, and the channel-wise representations are stacked into $\mathbf{Z} \in \mathbb{R}^{C\times S\times E}$. A CNN-based cross-channel fusion block combines the channel and embedding dimensions of $\mathbf{Z}$ and applies a one-dimensional convolution along the token sequence, so that each filter jointly mixes all channels and embedding features over consecutive tokens, providing a local depth-wise inductive bias. The fused features are flattened and passed to a task-specific fully connected prediction head (details in Appendix A.3). Depending on the downstream setting, the pre-trained components are either frozen or jointly fine-tuned with these modules (Section 4.2). Because pre-training is single-channel, relationships between channels are learned only during downstream adaptation, from the labeled data of each task.

4 Experiments

4.1 Pre-training

Data and preprocessing. FOCUS is pre-trained on $24{,}753{,}374$ scans from five sources acquired with five platforms, covering the forearm, wrist, upper arm, neck, and lower limb. The corpus combines recordings originally collected for hand-gesture recognition, cardiovascular monitoring, muscle contraction and fatigue analysis, and lower-limb motion monitoring, together with a dedicated WULPUS (Frey et al., 2022)-based unlabeled dataset (Table 1; details in Appendix A.1). The two largest sources account for 94% of the scans. To prevent them from dominating the pre-training objective, we randomly sample $200\,000$ scans from each of them at every epoch, while using all scans from the remaining sources. Each source is randomly split into 80/10/10% training, validation, and test sets.

Each scan is normalized to $[-1, 1]$ using per-scan min–max normalization, making amplitudes comparable across acquisition systems; downstream data are normalized per channel using the minimum and maximum of the training data. We pre-train directly on raw scans, which require no system-specific assumptions about the transducer center frequency; band-pass filtered and envelope inputs are compared in Appendix A.6.

Setup and reconstruction. Pre-training runs for 400 epochs on four NVIDIA A100 GPUs (26.9 hours in total, 6.7 h wall-clock) in PyTorch 2.2.0, with a batch size of 512 per GPU (global batch size 2048), a learning rate of 10^{-3} with 10 warm-up epochs and cosine decay to 10^{-4}, and a weight decay of 0.05. FOCUS reconstructs the dominant echo structures of all acquisition systems within the masked regions, with smoother reconstructions in low-amplitude regions, and preserves the spectral content of each system (Appendix A.2).

4.2 DOWNSTREAM EVALUATION

Benchmarks. We evaluate FOCUS on eight tasks from five datasets: finger-gesture classification and wrist-angle regression on UltraPro (Yang et al., 2022); hand-pose and forearm-position classification on PosePos-USVR (Spacone et al., 2026b); 3- and 23-Degrees-of-Freedom (DoF) hand and wrist kinematic regression on HWR-3DoF (Spacone et al., 2025b) and HWR-23DoF (Spacone et al., 2026a); and gait-phase classification and lower-limb joint-angle regression on K2MUSE (Li et al., 2026). Table 2 lists the subjects and evaluation protocol of each benchmark alongside the results; acquisition parameters and complete protocols are given in Table 3 and Appendix A.4. UltraPro was acquired with a custom platform that is not present in the pre-training corpus. The WULPUS and K2MUSE downstream datasets share their platform with pre-training data but are disjoint from it: the WULPUS datasets originate from separate data collections and acquisition protocols, and pre-training uses only K2MUSE recordings collected under different experimental conditions (fatigue, sensor-shift, and inter-day protocols).

Downstream setup We evaluate FOCUS under three training regimes: *from scratch (FS)*, where the architecture described in Section 3.4 is trained from randomly initialized weights; *frozen-encoder (FE)*, where the pre-trained encoder is frozen, and its representations are directly provided to the trainable cross-channel fusion CNN followed by the lightweight task-specific prediction head; and *full fine-tuning (FT)*, where the pre-trained encoder is jointly optimized with the cross-channel convolutional fusion module and prediction head.

In addition, a *Baseline CNN (BaseCNN)* adapted from Spacone et al. (2026b) is trained from scratch, with its parameter count approximately matched to the corresponding FOCUS architecture (details in Appendix A.3). For all regimes and baselines, hyperparameters are optimized independently for each task (Appendix A.5), and experiments are repeated with 3 seeds on a single GPU. The final checkpoint is evaluated. Metrics are first averaged over the subjects of each benchmark. To enable comparison with previously reported results, Table 2 reports the standard deviation across subjects; all other results report the standard deviation across the three downstream seeds.

All downstream evaluations are *subject-specific*, following the protocol of each reference work to enable direct comparison with previously reported results: a separate model is trained and evaluated for each subject. In the *intra-session* setting (UltraPro, K2MUSE), training, validation, and test data originate from the same recording sessions, with no sensor repositioning; (split details in A.4). In the *inter-session* setting (PosePos-USVR, HWR-3DoF, HWR-23DoF), entire recording sessions, including sensor repositioning, are held out for testing. We adopt subject-specific evaluations following prior work, therefore enabling a comparison; we do not assess generalization to unseen subjects.

4.3 DOWNSTREAM RESULTS

Pre-training improves performance over the same architecture trained from scratch. The comparison between FS and FT isolates the effect of pre-training: both use the same architecture, data, and protocol, and differ only in their initialization. FT improves the mean performance on all eight tasks (Table 2, column Δ). For classification, the gains range from 0.8 percentage points on K2MUSE-gait to 24.1 on PosePos-USVR-Pose; for regression, the largest gain is on HWR-23DoF, where the root mean squared error (RMSE) decreases from 19.37° to 13.27°. Gains are smaller where FS already performs well, such as on UltraPro-Wrist and K2MUSE-joint. Variability across seeds is reported in Appendix A.9.

Comparison with prior work and a parameter-matched CNN. FT exceeds the results reported by the reference works on seven of the eight benchmarks, and FE obtains the highest accuracy on PosePos-USVR-Pose. FT also outperforms the parameter-matched BaseCNN on all eight tasks, although BaseCNN outperforms FS on four of them, indicating that the gains are not explained by the Transformer architecture alone.

Frozen encoder. With the pre-trained encoder frozen, FE outperforms FS on five tasks (both UltraPro tasks, both PosePos-USVR tasks, and HWR-23DoF). Results are comparable on the two K2MUSE tasks, and worse on HWR-3DoF (RMSE of 11.27° versus 10.13°). Fine-tuning the encoder improves over FE on seven of the eight tasks, with the largest improvements on UltraPro-

Table 2: Downstream results. One model is trained per subject; entries are mean$_{\pm\text{std}}$ across subjects (seeds averaged first; sample standard deviation), following each reference work. Accuracy in %, RMSE in degrees. Grey rows give the dataset, number of subjects, evaluation protocol, and whether its acquisition platform is *seen* in pre-training (downstream recordings are always disjoint from the corpus). Δ: FT minus FS, the gain attributable to pre-training. *Labels needed*: smallest fraction of the training labels with which FT exceeds full-label FS (Section 4.4). Best in **bold**, second underlined. Secondary metrics and seed variability in Appendix A.9.

		Prior	From scratch		Pre-trained FOCUS			
Task	**Metric**	**work**	**BaseCNN**	**FS**	**FE**	**FT**	**Δ**	**Labels needed**
UltraPro (Yang et al., 2022); 4 subjects; intra-session; platform **unseen ✗**								
Gesture	Acc. ↑	$79.7_{\pm 7.9}$ ‡	$78.1_{\pm 6.6}$	$77.3_{\pm 7.4}$	$\underline{80.5}_{\pm 6.5}$	$\mathbf{84.1}_{\pm 5.4}$	$+6.8$	40%
Wrist angle	R^2 ↑	$\underline{0.924}_{\pm 0.021}$ ‡	$0.718_{\pm 0.077}$	$0.919_{\pm 0.034}$	$\underline{0.924}_{\pm 0.032}$	$\mathbf{0.932}_{\pm 0.028}$	$+0.013$	40%
PosePos-USVR (Spacone et al., 2026b); 5 subjects; inter-session; platform seen ✓								
Hand pose	Bal. Acc. ↑	$69_{\pm 7}$	$68.0_{\pm 7.9}$	$51.4_{\pm 9.8}$	$\mathbf{77.4}_{\pm 6.2}$	$\underline{75.5}_{\pm 5.0}$	$+24.1$	20%
Forearm position	Bal. Acc. ↑	$76_{\pm 6}$	$79.1_{\pm 4.8}$	$62.9_{\pm 4.9}$	$\underline{80.8}_{\pm 4.1}$	$\mathbf{83.0}_{\pm 2.8}$	$+20.1$	20%
HWR-3DoF (Spacone et al., 2025b); 4 subjects; inter-session; platform seen ✓								
3-DoF kinematics	RMSE ↓	$11.26_{\pm 1.11}$ †	$10.26_{\pm 1.59}$	$\underline{10.13}_{\pm 2.06}$	$11.27_{\pm 1.60}$	$\mathbf{9.34}_{\pm 1.69}$	-0.79	20%
HWR-23DoF (Spacone et al., 2026a); 2 subjects; inter-session; platform seen ✓								
23-DoF kinematics	RMSE ↓	$14.60_{\pm 1.98}$	$\underline{14.43}_{\pm 2.71}$	$19.37_{\pm 5.42}$	$15.76_{\pm 1.21}$	$\mathbf{13.27}_{\pm 0.26}$	-6.10	20%
K2MUSE (Li et al., 2026); 19 / 40 subjects; intra-session (trial-wise); platform seen ✓								
Gait phase	Acc. ↑	$\mathbf{71.9}_{\pm 16.9}$	$69.8_{\pm 15.9}$	$70.0_{\pm 14.7}$	$70.0_{\pm 15.6}$	$\underline{70.8}_{\pm 15.9}$	$+0.8$	40%
Joint angles	RMSE ↓	6.70	$5.28_{\pm 4.73}$	$5.18_{\pm 4.40}$	$\underline{5.15}_{\pm 4.29}$	$\mathbf{4.76}_{\pm 4.53}$	-0.42	20%

† Recomputed from Spacone et al. (2025b) (Table III) on the four subjects whose data were available (subjects 2–5 of the original work). ‡ Best results of Yang et al. (2022) under the split used here (their Tables III–IV; gesture: MTCNN, wrist: MOGP), recomputed from the per-session values. With a 40 s/10 s split without validation, they report $80.35_{\pm 7.54}\%$ (MOGP) and $R^2 > 0.935$ (MOGP).

Gesture ($+3.6$ percentage points) and HWR-23DoF (RMSE $-2.49°$). The frozen representation therefore already transfers to most tasks, while FT further specializes it.

Larger gains on inter-session benchmarks. The gains from pre-training are largest on PosePos-USVR ($+24.1$ and $+20.1$ percentage points) and HWR-23DoF (coefficient of determination (R^2) from 0.132 to 0.488), two of the three benchmarks evaluated across sessions with sensor repositioning, and are smaller on the intra-session UltraPro and K2MUSE benchmarks. This pattern is consistent with pre-training being particularly beneficial under acquisition-induced distribution shift. Supporting this interpretation, t-SNE projections of the fused features show greater separation between training and validation data for PosePos-USVR than for UltraPro (Appendix A.8). However, these benchmarks also differ in task, acquisition platform, and baseline FS performance, with substantially more room for improvement on PosePos-USVR and HWR-23DoF; moreover, the third inter-session benchmark, HWR-3DoF, shows only a limited gain. The comparison therefore does not isolate acquisition-induced shift as the sole cause of the observed improvements, and FOCUS does not explicitly enforce session-invariant representations. Additionally, because the largest gains occur on benchmarks whose platform and anatomical sites are also represented in the pre-training corpus, our experiments do not distinguish the contribution of corpus heterogeneity from that of in-domain pre-training data.

Exception: K2MUSE-gait. On K2MUSE-gait, FT improves over FS only from 70.0% to 70.8%, a small change relative to the inter-subject variability, and does not reach the 71.9% reported by the original work, which also found limited gait-phase information in AUS signals (Li et al., 2026). Our experiments do not distinguish whether this limitation stems from the signals or from the pre-training strategy.

4.4 Label efficiency

We train FT and FS on 20%, 40%, 60%, 80%, and 100% of each downstream training set. Each subset is sampled within every acquisition session and, for classification tasks, within every class, so that it preserves the session and class proportions of the complete training set. All hyperparameters, including the number of epochs and the batch size, are kept identical to those used with the complete

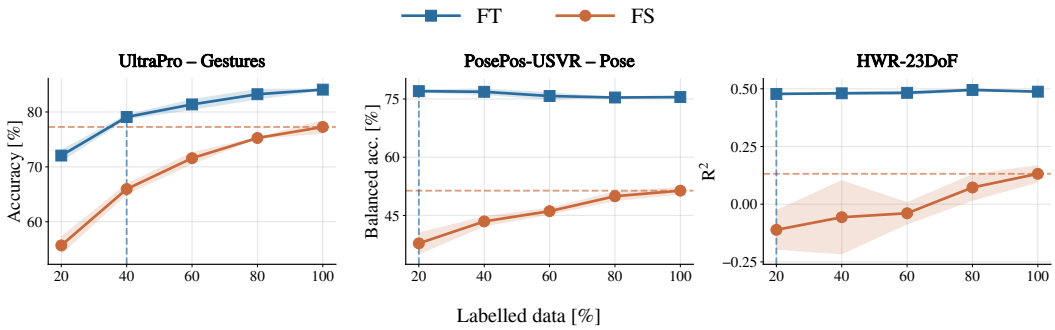


Figure 2: Label efficiency on three representative benchmarks: FT (pre-trained) versus FS (same architecture from scratch) as a function of the fraction of labeled training data. The horizontal dashed line marks FS trained on the complete training set, and the vertical dashed line the smallest fraction of labels at which FT already exceeds it. All eight benchmarks are shown in Appendix A.10. Mean and ± 1 standard deviation over three seeds.

training set (Table 4); smaller fractions therefore also receive proportionally fewer optimization steps, which makes the comparison with full-data FS conservative for FT. Because the subsets are drawn from the same recording sessions, these experiments measure the dependence on labeled examples, not on the number of recorded sessions or on the annotation effort.

With only 20% of the labeled data, FT already outperforms FS trained on the complete training set on five of the eight tasks, and with 40% on all eight (Figure 2; all tasks in Appendix A.10). At 40%, FT exceeds full-data FS by 25.5 percentage points on PosePos-USVR-Pose and raises the R^2 on HWR-23DoF from 0.132 to 0.481. FS improves steadily as labeled data are added on all eight tasks, whereas FT remains comparatively stable across fractions, except on the two UltraPro tasks. The UltraPro tasks are also where FT with 20% of the labels remains below full-label FS (Table 2, last column), and K2MUSE-gait is essentially tied at 20%. UltraPro is the only benchmark acquired with a platform absent from the pre-training corpus, which may contribute to its stronger dependence on labeled data, although our experiments do not isolate this factor.

4.5 Model capacity and pre-training data scale

Following SensorFM (Narayanswamy et al., 2026), we scale the embedding dimension and encoder depth while keeping the architecture and pre-training objective fixed to assess the model capacity. We compare *FOCUS-XSmall* (0.8 M parameters, $E = 128$, 4 encoder blocks), *FOCUS-Small* (6.3 M, 256, 8), and *FOCUS-Big* (85.1 M, 768, 12); full configurations are given in Appendix A.10. Each model is pre-trained on 20%, 40%, 60%, 80%, and 100% of the corpus for 400 epochs with a global batch size of 2048, and the final checkpoint is evaluated. Since the number of epochs is fixed and the per-epoch cap on the two largest sources (Section 4.1) is scaled with the fraction, the number of optimization steps grows proportionally with the subset size, from 61.6k at 20% to 309.2k at 100%, while the contribution of each source to an epoch is unchanged. The data-scaling results therefore measure the combined effect of more pre-training data and more training compute. FOCUS-XSmall underperforms the Small and Big configurations (Figure 3) on all eight tasks, often by a large margin, suggesting that its capacity is insufficient to exploit the pre-training data. With the complete corpus, FOCUS-Big outperforms FOCUS-Small on seven of the eight tasks, by smaller margins. The effect of additional pre-training data depends on capacity: FOCUS-Big shows a generally positive trend with task-specific fluctuations, FOCUS-Small is largely flat, and FOCUS-XSmall can degrade as the corpus grows, for example on UltraPro-Gesture from 83.1% with 20% of the corpus to 81.3% with the complete corpus. Larger models thus appear better able to exploit additional pre-training data.

FOCUS-Small represents a practical operating point for future edge deployment. Its 6.3 M-parameter encoder is comparable in size to compact biosignal foundation models designed for embedded execution (Fasulo et al., 2026) and to Transformer architectures already deployed at the edge (Liang et al., 2023; Ruciński et al., 2026). With a sequential channel inference scheme,

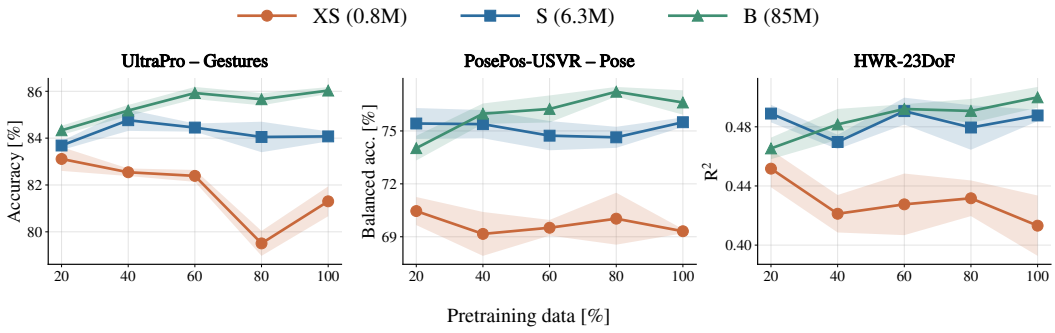


Figure 3: Downstream performance as a function of model capacity and pre-training data fraction on three representative benchmarks (all eight in Appendix A.10). Mean and ± 1 standard deviation over three seeds.

FOCUS-Small requires 669–856 MMACs per channel update and about 6.5–6.7 MB of weights under typical INT8 quantization schemes (Jacob et al., 2018) (memory and computational cost details are given in Appendix A.11).

5 Conclusion

In this work, we introduced FOCUS, to the best of our knowledge, the first FM for wearable AUS designed for joint pre-training across heterogeneous acquisition systems. A frequency-aware multi-tokenizer maps signals acquired at different sampling frequencies to patches spanning approximately consistent physical depths, allowing a shared model to be pre-trained on approximately 25 million AUS scans collected using five acquisition systems.

Across eight subject-specific downstream tasks, pre-training improves the mean performance over the same architecture trained from scratch on all eight tasks, and FOCUS outperforms the previously reported results on seven of the eight benchmarks. Fine-tuning with only 40% of the labeled training data outperforms the same architecture trained from scratch on the complete training set on all eight tasks; because these subsets are drawn from the same recording sessions, this reflects a reduced dependence on labeled examples rather than on acquisition effort. Downstream performance generally improves with model size, and larger models appear better able to exploit additional heterogeneous pre-training data.

Our experiments show that masked pre-training on the heterogeneous corpus improves over training from scratch, but they do not isolate the contribution of corpus heterogeneity. Quantifying the contribution of individual acquisition systems (e.g., single-source or leave-one-source-out pre-training) is a priority for future work. Future work also includes incorporating multichannel information directly during pre-training (Döner et al., 2025), assessing adaptation to unseen subjects (Grandi Sgambato et al., 2025) and encouraging session-invariant representations (Shangguan et al., 2024). It will also explore scaling the pre-training corpus to additional subjects, anatomical locations, and acquisition systems, including A-mode domains beyond wearable sensing such as non-destructive testing. Finally, we will evaluate on-device deployment, reporting measured latency, memory, and energy on a wearable-class processor.

To support benchmarking across heterogeneous AUS acquisition systems, we will release the FOCUS models, training code, and evaluation framework.[1]

AI use statement

In this work, we used generative AI tools, including Anthropic Claude and OpenAI Codex, to assist with performance engineering of the experimental code, including refactoring and improving execution efficiency, and to improve the grammar, clarity, and concision of author-written manuscript text.

[1]Code and models will be released after the review process to preserve anonymity.

The tools were not used to formulate research questions or hypotheses, design experiments, generate data or experimental results, or draw scientific conclusions. The authors reviewed all retained AI-assisted code changes and validated the final implementation by executing the relevant experiments and inspecting their recorded outputs. All AI-assisted language edits were reviewed and approved by the authors. The authors take responsibility for the final text, code, claims, and artifacts.

Ethics statement

All datasets used in this work contain data acquired from human participants. For previously collected datasets, we exclusively use data that were either publicly released by the authors of the corresponding studies or provided directly by the original authors upon request. The collection and ethical oversight of these data were conducted within the respective original studies, and we refer to the corresponding publications for their specific ethical approval.

For the data newly collected as part of this work, all experimental procedures were conducted in accordance with the Declaration of Helsinki and were approved by the responsible institutional ethics committee. The corresponding protocol identifier is withheld during review to preserve author anonymity and will be included in the final version of the manuscript. All participants provided written informed consent prior to participation, including consent for the use of the collected data for research purposes.

Reproducibility Statement

To support reproducibility, the code required to reproduce the experiments and results presented in this work will be made publicly available on GitHub. The corresponding pre-trained FOCUS model checkpoints and the datasets collected specifically for this work will be released through Hugging Face. Links to these resources will be provided after the review process to preserve author anonymity.

References

Lukas Brausch, Holger Hewener, and Paul Lukowicz. Muscle contraction a-scan data annotated by volunteers. OpenML Dataset 41971, 2019. URL https://www.openml.org/d/41971.

Lukas Brausch, Holger Hewener, and Paul Lukowicz. Muscle fatigue a-scan data of a single volunteer (study 2/2). OpenML Dataset 43076, 2021a. URL https://www.openml.org/d/43076.

Lukas Brausch, Holger Hewener, and Paul Lukowicz. Muscle fatigue a-scan data of 21 volunteers (study 1/2). OpenML Dataset 43075, 2021b. URL https://www.openml.org/d/43075.

Lukas Brausch, Holger Hewener, and Paul Lukowicz. Classifying muscle states with one-dimensional radio-frequency signals from single element ultrasound transducers. *Sensors*, 22 (7):2789, 2022.

Berkay Döner, Thorir Mar Ingolfsson, Luca Benini, and Yawei Li. Luna: efficient and topology-agnostic foundation model for eeg signal analysis. *Advances in Neural Information Processing Systems*, 38:70682–70708, 2025.

Zhao Du, Junyi Yang, Xiangxin Li, Wujing Cao, Meng Yin, and Xinyu Wu. Stable online hand gesture decoding with hybrid semg-ultrasound sensing and multimodal transformers. In *2026 IEEE 20th International Conference on Control and Automation (ICCA)*, pp. 1808–1813. IEEE, 2026.

Matteo Fasulo, Giusy Spacone, Thorir Mar Ingolfsson, Yawei Li, Luca Benini, and Andrea Cossettini. Tinymyo: a tiny foundation model for flexible emg signal processing at the edge. *IEEE Transactions on Biomedical Engineering*, pp. 1–12, 2026. doi: 10.1109/TBME.2026.3736157.

Sebastian Frey, Sergei Vostrikov, Luca Benini, and Andrea Cossettini. Wulpus: a wearable ultra low-power ultrasound probe for multi-day monitoring of carotid artery and muscle activity. In

2022 IEEE International Ultrasonics Symposium (IUS), pp. 1–4, 2022. doi: 10.1109/IUS54386.2022.9958156.

Marco Giordano, Christoph Leitner, Christian Vogt, Luca Benini, and Michele Magno. Pulse: Accurate and robust ultrasound-based continuous heart-rate monitoring on a wrist-worn iot device. *IEEE Internet of Things Journal*, 2025.

Bruno Grandi Sgambato, Milia H Hasbani, Deren Y Barsakcioglu, Jaime Ibáñez, Anette Jakob, Marc Fournelle, Meng-Xing Tang, and Dario Farina. High performance wearable ultrasound as a human-machine interface for wrist and hand kinematic tracking. *IEEE Transactions on Biomedical Engineering*, 71(2):484–493, 2024.

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

Lin Guo, Zongxing Lu, Ligang Yao, and Shaoxiong Cai. A gesture recognition strategy based on a-mode ultrasound for identifying known and unknown gestures. *IEEE Sensors Journal*, 22(11): 10730–10739, 2022.

K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009, 2022.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2704–2713, 2018.

Jing Jiao, Jin Zhou, Xiaokang Li, Menghua Xia, Yi Huang, Lihong Huang, Na Wang, Xiaofan Zhang, Shichong Zhou, Yuanyuan Wang, et al. Usfm: A universal ultrasound foundation model generalized to tasks and organs towards label efficient image analysis. *Medical image analysis*, 96:103202, 2024.

Frederick W Kremkau. *Sonography principles and instruments*. Elsevier Health Sciences, 2015.

Jiwei Li, Bi Zhang, Xiaowei Tan, Wanxin Chen, Zhaoyuan Liu, Juanjuan Zhang, Weiguang Huo, Jian Huang, Lianqing Liu, and Xingang Zhao. K2muse: A human lower-limb multimodal walking dataset spanning task and acquisition variability for rehabilitation robotics. *The International Journal of Robotics Research*, 0(0):02783649261444046, 2026. doi: 10.1177/02783649261444046. URL https://doi.org/10.1177/02783649261444046.

Yue Lian, Zongxing Lu, Xin Huang, Qican Shangguan, Ligang Yao, Jie Huang, and Zhoujie Liu. A transfer learning strategy for cross-subject and cross-time hand gesture recognition based on a-mode ultrasound. *IEEE Sensors Journal*, 24(10):17183–17192, 2024.

Yinan Liang, Ziwei Wang, Xiuwei Xu, Yansong Tang, Jie Zhou, and Jiwen Lu. Mcuformer: Deploying vision tranformers on microcontrollers with limited memory. *Advances in Neural Information Processing Systems*, 36:8501–8512, 2023.

Yi-Wen Liu. Hilbert transform and applications. *Fourier transform applications*, pp. 291–300, 2012.

Antonios Lykourinas, Chinmay Pendse, Francky Catthoor, Veronique Rochus, Xavier Rottenberg, and Athanassios Skodras. Ultrasound-based human-machine interfaces: Raw RF data for robust and parameter-efficient CNNs. In *European Signal Processing Conference (EUSIPCO)*, pp. 1417–1421, 2026.

Chen Ma, Jing Jiao, Shuyu Liang, Junhu Fu, Qin Wang, Zeju Li, Yuanyuan Wang, and Yi Guo. Tinyusfm: Towards compact and efficient ultrasound foundation models. *IEEE Journal of Biomedical and Health Informatics*, 2026.

Girish Narayanswamy, Maxwell A Xu, A Ali Heydari, Samy Abdel-Ghaffar, Marius Guerard, Kara Vaillancourt, Zhihan Zhang, Jake Garrison, Levi Albuquerque, Dimitris Spathis, et al. Towards a general intelligence and interface for wearable health data. *arXiv preprint arXiv:2605.22759*, 2026.

Chuang Niu, Qing Lyu, Christopher D Carothers, Parisa Kaviani, Josh Tan, Pingkun Yan, Mannudeep K Kalra, Christopher T Whitlow, and Ge Wang. Medical multimodal multitask foundation model for lung cancer screening. *Nature Communications*, 16(1):1523, 2025.

NVIDIA. Adaptive ultrasound imaging with physics-informed NV-Raw2Insights-US AI. `https://huggingface.co/blog/nvidia/raw2insights-adaptive-ultrasound-imaging`, 2026. Hugging Face Blog, accessed August 31, 2026.

Geonho Park, Yizhou Bian, Hao Huang, Sai Zhou, Siyu Qin, Muyang Lin, Xinyi Yang, Aaron Lee, Anand Ramkumar, Mariana Tome, et al. Fetal monitoring for high-risk pregnancies using a wearable ultrasound patch. *Nature Biotechnology*, pp. 1–10, 2026.

Ofir Press, Noah A Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*, 2021.

Mengjiao Qu, Jie Zhou, Dongze Lv, Zexin Wang, Yinfei Zheng, Guojun Zhang, and Jin Xie. Continuously monitoring of muscle fatigue based on a wearable micromachined ultrasonic transducer probe. *Sensors and Actuators A: Physical*, 365:114892, 2024.

Immanuel RoßTeutscher, Klaus S Drese, and Thorsten Uphues. Masked autoencoders for ultrasound signals: Robust representation learning for downstream applications. *IEEE Access*, 13:212620–212634, 2025.

Szymon Ruciński, Pietro Bonazzi, Engin Türetken, Simon Narduzzi, Michele Magno, and Nadim Maamari. Ftervit: Fully ternary vision transformer. *arXiv preprint arXiv:2605.21171*, 2026.

Qican Shangguan, Yue Lian, Shaoxiong Cai, Jun Wu, Ligang Yao, and Zongxing Lu. Dann-repositing strategy for zero retraining long-term hand gesture recognition using wearable a-mode ultrasound. *IEEE Transactions on Instrumentation and Measurement*, 73:1–11, 2024. doi: 10.1109/TIM.2024.3460939.

Giusy Spacone, Sebastian Frey, Christoph Leitner, Luca Benini, and Andrea Cossettini. Towards fully wearable muscle fatigue assessment with a-mode ultrasound. In *2025 IEEE International Ultrasonics Symposium (IUS)*, pp. 1–4, 2025a. doi: 10.1109/IUS62464.2025.11201415.

Giusy Spacone, Sergei Vostrikov, Victor Kartsch, Simone Benatti, Luca Benini, and Andrea Cossettini. Tracking of wrist and hand kinematics with ultra low power wearable a-mode ultrasound. *IEEE Transactions on Biomedical Circuits and Systems*, 19(3):536–548, 2025b.

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

Giusy Spacone, Sebastian Frey, Enzo Baraldi, Mattia Orlandi, Luca Benini, and Andrea Cossettini. A wearable multimodal ultrasound+ inertial system for real-time virtual reality interaction. *arXiv preprint arXiv:2606.17741*, 2026b. Accepted at IEEE SMC 2026.

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

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

Sergei Vostrikov, Andrea Cossettini, Christoph Leitner, Christian Baumgartner, and Luca Benini. Aepus: a tool for the automated extraction of pennation angles in ultrasound images with low signal-to-noise ratio for plane-wave imaging. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 1520–1526. IEEE, 2022.

Sergei Vostrikov, Matteo Anderegg, Luca Benini, and Andrea Cossettini. Unsupervised feature extraction from raw data for gesture recognition with wearable ultralow-power ultrasound. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 71(7):831–841, 2024.

David Weik, Richard Nauber, Erik Kaiser, Nele Kirsch, Robert Kunz, Lennart Schierling, Christoph Leitner, Luca Benini, Hsiao-Chuan Liu, Qifa Zhou, Jochen Hampe, Gerhard Fettweis, Moritz Herzog, and Christian Kupsch. Current trends in ultrasound wearables: Spotlight on system architecture. *IEEE Reviews in Biomedical Engineering*, pp. 1–21, 2026. doi: 10.1109/RBME.2026.3664011.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. On layer normalization in the transformer architecture. In *International conference on machine learning*, pp. 10524–10533. PMLR, 2020.

Bingcong Yan, Chunlei Li, Jingliang Hu, Yilei Shi, Xiao Xiang Zhu, and Lichao Mou. Towards real-world ultrasound understanding: Large vision-language models from multi-image examinations with long-form reports. *arXiv preprint arXiv:2607.01908*, 2026.

Xingchen Yang, Zhenfeng Chen, Nalinda Hettiarachchi, Jipeng Yan, and Honghai Liu. A wearable ultrasound system for sensing muscular morphological deformations. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 51(6):3370–3379, 2021.

Xingchen Yang, Yifan Liu, Zongtian Yin, Pu Wang, Peilin Deng, Zhengen Zhao, and Honghai Liu. Simultaneous prediction of wrist and hand motions via wearable ultrasound sensing for natural control of hand prostheses. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 30:2517–2527, 2022.

Xingchen Yang, Claudio Castellini, Dario Farina, and Honghai Liu. Ultrasound as a neurorobotic interface: A review. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 54(6):3534–3546, 2024.

Hongyuan Zhang, Yuheng Wu, Mingyang Zhao, Zhiwei Chen, Rebecca Li, Fei Zhu, Haohan Zhao, Xiaohua Yuan, Meng Yang, Chunli Qiu, et al. A fully open and generalizable foundation model for ultrasound clinical applications. *arXiv preprint arXiv:2509.11752*, 2025.

Lin Zhang, Colin Marcus, Dabin Lin, David Mejorado, Scott Joseph Schoen Jr, Theodore T Pierce, Viksit Kumar, Sara V Fernandez, David Hunt, Qian Li, et al. A conformable phased-array ultrasound patch for bladder volume monitoring. *Nature Electronics*, 7(1):77–90, 2024.

Sai Zhou, Geonho Park, Muyang Lin, Xinyi Yang, and Sheng Xu. Wearable ultrasound technology. *Nature Reviews Bioengineering*, 3(10):835–854, 2025a.

Sai Zhou, Geonho Park, Katherine Longardner, Muyang Lin, Baiyan Qi, Xinyi Yang, Xiaoxiang Gao, Hao Huang, Xiangjun Chen, Yizhou Bian, et al. Clinical validation of a wearable ultrasound sensor of blood pressure. *Nature biomedical engineering*, 9(6):865–881, 2025b.

A Appendix

A.1 Pre-training Datasets Description

Table 1 summarizes the datasets used to construct the pre-training corpus, collected from different body locations and targeting different tasks. The acquisition platforms employed differ in sampling frequency f_s, number of samples per US channel N, and consequently the sampled depth range. They also employ transducers with different center frequencies f_c. Following the frequency-aware multi-tokenizer strategy described in Section 3.2, we additionally report the patch size P selected for each dataset.

The following datasets from prior works are employed to construct the pre-training corpus:

- **Wrist dataset for heart-rate monitoring** This portion comprises data (Giordano et al., 2025) collected from the wrist while resting and after doing light exercise[2]. A custom acquisition system was used, with data sampled at 2 GHz; transducers with a center frequency of 10 MHz were employed. Each US scan consists of 50000 samples, which maps to 500 after resampling.
- **Upper-arm and calf muscle datasets** This portion comprises A-mode US recordings acquired from the calf during muscle contraction experiments and from the biceps brachii during fatigue experiments (Brausch et al., 2022; 2019; 2021a;b) [3]. Data were acquired using a custom platform with single-element transducers (Panametrics, 3.5 MHz center frequency) and contribute a total of 488 175 A-mode scans to the pre-training corpus. The muscle-contraction recordings contain 3000 samples per frame, whereas the fatigue recordings contain 2048 samples. To harmonize the sampled depth range across datasets and remove portions of the traces that do not contain relevant information, both datasets are truncated to the first 1600 samples before pre-training.
- **Lower-limb dataset for lateral gastrocnemius monitoring**[4] This portion comprises data (Vostrikov et al., 2022) collected from the lateral gastrocnemius muscle. Data were collected using a linear array US probe (Ge-9LD, General Electric, Healthcare, Chicago, IL, USA), connected to a benchtop system (Verasonics Vantage 256). This portion accounts for 15 899 520 scans.
- **Lower-limb monitoring dataset.** This portion of the corpus is derived from the K2MUSE dataset (Li et al., 2026), which comprises multimodal lower-limb recordings acquired under diverse locomotion and acquisition conditions. A-mode US signals were acquired using a commercially available four-channel system (ELONXI, China) equipped with single-element transducers with a center frequency of 5 MHz. Each A-mode frame contains 1000 samples acquired at an axial sampling frequency of 20 MHz. The dataset includes recordings acquired across different walking conditions, as well as under non-ideal acquisition conditions comprising muscle fatigue, sensor shifts, and inter-day variability. Although A-mode US signals were synchronously recorded under these non-ideal conditions, the corresponding analyses and motion-intention recognition experiments in the original work were conducted using sEMG features rather than US. We therefore use the A-mode US recordings from the fatigue, sensor-shift, and inter-day experiments exclusively for FOCUS pre-training, contributing a total of 7 374 232 A-mode scans. The reference walking recordings used for our downstream evaluation are excluded from pre-training. Scans used for pre-training are truncated to 800 samples by removing the tail, thereby limiting the maximum sequence length across the pre-training corpus.

In addition to existing datasets, we collected a new corpus using the WULPUS platform (Frey et al., 2022), referred to as the **Multi-site WULPUS dataset**. Data were collected at a sampling frequency of 8 MHz using transducers with a center frequency of 2.25 MHz. Each scan comprises 400 samples. Recordings were acquired from multiple sensing locations, including the wrist, forearm, upper arm, lower leg, neck, and mandibular region. The acquisition protocol spans both controlled and unconstrained conditions, including static and free hand and wrist gestures, finger movements, grip and shaking movements, walking and foot movements, and free upper-limb movements. The corpus further includes neck and mandibular recordings acquired during vocalized and silent speech, chewing, and head movements. Transducers were removed and repositioned between selected recordings to increase the acquisition variability represented in the pre-training corpus. Overall, this dataset contributes 542 946 *unlabelled* A-mode scans to the pre-training corpus.

A.2 Depth and Frequency Reconstruction results

Figure 4 reports representative reconstructions in both time and frequency domains for datasets acquired with different hardware configurations. Specifically, the left plot shows data acquired with WULPUS at 8 MHz using 2.25 MHz transducers, the middle plot shows the lower-limb dataset acquired at 20 MHz with 5 MHz transducers, and the right plot shows the calf and biceps datasets acquired with 3.5 MHz transducers at 40 MHz. Across these heterogeneous settings, FOCUS re-

[2]Dataset was accessed via direct request to the authors.

[3]Additional data not directly available through the open-source release were obtained upon request from the authors.

[4]Dataset was accessed via direct request to the authors.

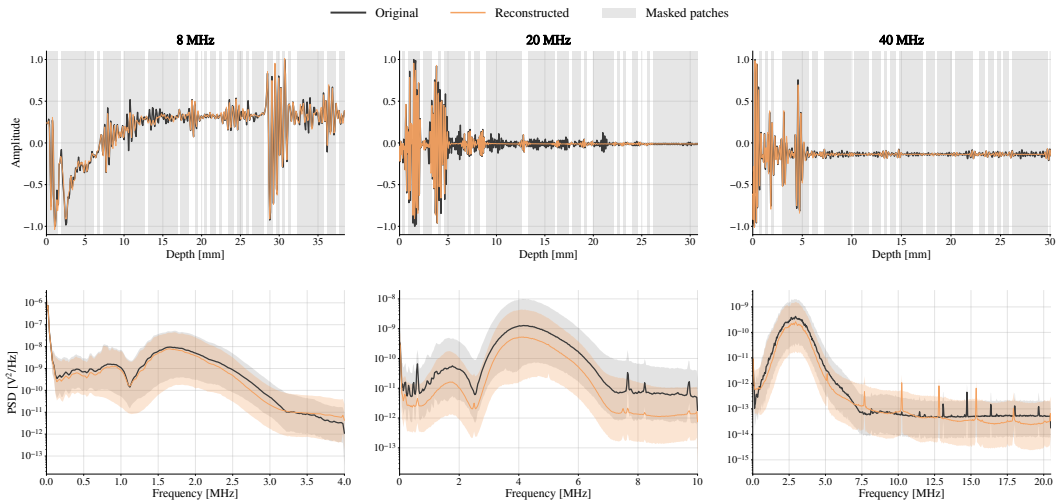


Figure 4: Masked reconstruction of pre-training test signals from sources acquired at three different sampling frequencies: forearm for hand-gesture recognition ($f_s = 8$ MHz, left), lower limb monitoring ($f_s = 20$ MHz, center), and calf muscle contraction ($f_s = 40$ MHz, right). Top: one test signal per source as a function of depth. The original signal is shown in dark gray, the reconstruction in orange, and the patches hidden from the encoder are shaded. Bottom: power spectral density of the original and reconstructed signals over the test-set of each source. Lines show the median and shaded bands the 10th–90th percentile range. Left plot: data acquired with WULPUS at 8 MHz using 2.25 MHz transducers; Middle plot: lower-limb (Li et al., 2026) dataset acquired at 20 MHz with 5 MHz transducers; Right plot: calf and biceps datasets (Brausch et al., 2022) with 3.5 MHz transducers at 40 MHz.

constructs the main temporal echo structures while also preserving the characteristic spectral content of each dataset, including dominant components around the corresponding transducer center frequency. These examples therefore provide qualitative evidence that the model can handle differences not only in sampling frequency and signal discretization, but also in transducer-specific spectral characteristics.

A.3 Architecture Details

Transformer blocks. Encoder and decoder blocks follow a pre-layer normalization design (Xiong et al., 2020) and consist of multi-head self-attention (Vaswani et al., 2017), followed by a feed-forward network with a $4\times$ hidden-dimension expansion and GELU activation (Hendrycks & Gimpel, 2016). Unless otherwise stated, FOCUS uses 8 encoder blocks, 4 decoder blocks, and 8 attention heads (FOCUS-Small in Table 9).

Cross-channel fusion block. The stacked channel-wise representations $\mathbf{Z} \in \mathbb{R}^{B\times C\times S\times E}$ are reshaped by combining the channel and embedding dimensions, and a one-dimensional convolution with $F = 32$ filters, kernel size $k = 4$, and stride $s = 4$ is applied along the token sequence, followed by group normalization and a GELU activation. The block produces $\mathbf{Z}_{\text{fused}} \in \mathbb{R}^{B\times F\times D}$, where $D = \lfloor (S-k)/s \rfloor + 1$ ($D = S/4$ for S divisible by 4), which is flattened into an FD-dimensional representation and provided to the task-specific fully connected prediction head.

Baseline CNN. The BaseCNN is adapted from Spacone et al. (2026b). It consists of two convolutional blocks, each comprising a 7×1 Conv2D layer followed by batch normalization, ReLU, dropout, and 4×1 max pooling along the depth dimension. The resulting features are flattened and processed by two fully connected layers followed by the output layer. The number of convolutional filters is adjusted for each downstream task such that the total parameter count is approximately matched to the corresponding FOCUS architecture (Table 5).

A.4 DOWNSTREAM DATASETS

We evaluate FOCUS across a wide range of downstream tasks, including classification and regression. Table 3 provides an overview of the evaluated datasets included, together with the acquisition specifications (acquisition platform employed, number of acquisition channels, transducer central frequency, sampling rate), the task description, and the corresponding evaluation metrics. Specifically, performance is evaluated on the following datasets:

Table 3: Downstream datasets used to evaluate FOCUS and their acquisition, task, and evaluation settings. f_s denotes the sampling frequency and f_c the transducer center frequency.

Dataset	**Platform**	f_s **[MHz]**	**Depth [mm]**	f_c **[MHz]**	**# Ch**	**Task**	**Evaluation**		
							Metric	**Setting**	**Data split**
UltraPro	Custom	20	38.5	5	8			Intra-session	80%, 5%, 15% train-validation-test
UltraPro-Gesture						Classification (6 gestures)	Accuracy		
UltraPro-Wrist						Regression (1 DoF)	R^2		
PosePos-USVR	WULPUS	8	38.2	2.25	6			Inter-session	5-fold CV. 4 sessions for training-validation, 1 session for test.
PosePos-USVR-Pose						Classification (6 poses)	Accuracy		
PosePos-USVR-Position						Classification (3 positions)	Accuracy		
HWR-3DoF	WULPUS	8	38.5	2.25	4	Regression (3 DoF)	RMSE, R^2	Inter-session	9-fold CV. 8 sessions for training-validation, 1 session for test.
HWR-23DoF	WULPUS	8	38.5	2.25	4	Regression (23 DoF)	RMSE, R^2	Inter-session	3-fold CV. 2 sessions for training-validation, 1 session for test.
K2MUSE	ELONXI	20	38.5	5	4			Intra-session	5-fold trial-wise CV[1]
K2MUSE-gait						Classification (5 gait phases)	Accuracy		
K2MUSE-joint						Regression (3 DoF)	RMSE, R^2		

[1] We categorize this as an *intra-session* protocol, since no sensor repositioning is performed between training and test data. However, following the methodology of the original work, the CV split is performed at the trial level, with one complete trial held out as the test fold in each iteration.

- **UltraPro:** This dataset (Yang et al., 2022) comprises recordings from four transradial amputees. Eight A-mode US transducers with a center frequency of 5 MHz were distributed around the forearm of the amputated side. Each A-mode scan contains 1000 samples acquired at a sampling frequency of 20 MHz, while the transducers were sequentially driven at a frame rate of 10 Hz. A custom acquisition system was used (Yang et al., 2021).
  Participants performed six finger gestures while concurrently rotating the wrist between pronation and supination. Each gesture was performed for 50 s, and three repeated sessions were recorded for each participant. Two downstream tasks are considered: six-class finger-gesture classification and continuous wrist pronation–supination angle regression (1 DoF). Classification accuracy is used to evaluate gesture recognition, whereas the coefficient of determination (R^2) is used for wrist-angle regression.
  We adopt an *intra-session* evaluation protocol. Following Yang et al. (2022) (deep learning model evaluations) each recording session is treated independently, resulting in subject- and session-specific models. Each 50 s recording is split chronologically, with the first 40 s (80%) used for training. Of the remaining 10 s, the first 2.5 s (5%) are used for validation and the final 7.5 s (15%) for testing, resulting in an 80%/5%/15% train/validation/test split. This preserves the temporal ordering of the acquisition and ensures that the training, validation, and test partitions are temporally disjoint.
- **PosePos-USVR:** This dataset (Spacone et al., 2026b) comprises recordings from five subjects performing hand gestures and forearm movements for real-time virtual-reality interaction. A-mode US signals were acquired from the forearm and upper arm using the WULPUS platform (Frey et al., 2022), using six transducers with a center frequency of 2.25 MHz. Each A-mode frame is sampled at 8 MHz, and transducers were sequentially excited at a rate of 30 Hz. Five acquisition sessions were collected across multiple days, explicitly introducing inter-session variability. Two downstream classification tasks are considered: hand-pose recognition (6 classes) and forearm-position estimation (3 classes).

Following the original evaluation protocol, subject-specific models are evaluated using the *inter-session* setting. For each fold, four acquisition sessions are used for model training, and the remaining session is held out for testing. The data from the four training-validation sessions are further split into 70% for training and 30% for validation. Classification accuracy is used as the evaluation metric.

- **HWR-3DoF:** This dataset (Spacone et al., 2025b) [5] comprises recordings from five subjects (four considered in this work) performing hand and wrist movements. A-mode US signals were acquired using the WULPUS platform (Frey et al., 2022) and four transducers with a center frequency of 2.25 MHz. Each A-mode frame contains 400 samples acquired at a sampling frequency of 8 MHz. The four transducers are sequentially acquired at 50 Hz, resulting in a complete four-channel arm scan at 12.5 Hz. Nine acquisition sessions were collected for each subject, with the armband removed and repositioned between consecutive sessions to introduce inter-session variability.
  The downstream task consists of simultaneously estimating three hand and wrist DoFs: wrist flexion–extension, wrist radial–ulnar deviation, and metacarpophalangeal flexion–extension. Following the original work, subject-specific models are evaluated using an *inter-session* setting, where eight sessions are used for training and the remaining session for testing. 90% of data belonging to the train-validation fold are used for training, 10% for validation. Performance is evaluated using RMSE and R^2.
- **HWR-23DoF:** This dataset (Spacone et al., 2026a)[6] comprises recordings from two subjects performing hand and wrist movements. A-mode US signals were acquired using the WULPUS platform (Frey et al., 2022) and four transducers with a center frequency of 2.25 MHz, equally spaced around the forearm. Each A-mode frame contains 400 samples, with transducers excited in a round-robin fashion at a frame rate of 30 Hz. Three acquisition sessions were collected for each subject, with armband repositioning between sessions to introduce inter-session variability.
  The downstream task consists of simultaneously regressing 23 hand and wrist DoFs, comprising 20 finger joint angles and three wrist angles. Following the original work, subject-specific models are evaluated using the *inter-session* setting (3-fold CV). 90% of data belonging to the train-validation fold are used for training, and 10% for validation.
  Performance is evaluated using RMSE and R^2.
- **K2MUSE-gait:** This downstream task is derived from the K2MUSE corpus (Li et al., 2026). The acquisition platform is the same as for the corresponding portion of the corpus used for pre-training and described in Section A.1. Four A-mode US channels were synchronously acquired from the left leg at a frame rate of 20 Hz. The acquisition protocol comprises treadmill walking under different locomotion conditions, including level-ground walking and ascending and descending ramps at 5° and 10°, with walking speeds of 0.5, 1.0, and 1.5 m/s. Each walking condition comprises five repeated trials.
  The downstream task consists of classifying the gait cycle into 5 discrete gait phases. We evaluate the 19 participants included in the publicly available release `https://k2muse.github.io/usage/index#gait-phase-classification` using subject-specific models and 5-fold trial-wise CV, with one trial held out for testing in each fold. We categorize this as an *intra-session* protocol, since no sensor repositioning is performed between training and test trials. Classification accuracy is used as the evaluation metric.
- **K2MUSE-joint:** This downstream task is derived from the same K2MUSE corpus (Li et al., 2026) and uses the four A-mode US channels acquired at a frame rate of 20 Hz. The task consists of predicting the left-leg hip, knee, and ankle joint angles during level-ground walking at 1.0 m/s, resulting in a three-output regression task. Following the original evaluation protocol, subject-specific models are evaluated for 40 participants using a 5-fold trial-wise CV, where each trial is held out in turn as the test fold (`https://k2muse.github.io/usage/index#joint-angle-regression`). We categorize this as an *intra-session* protocol, since no sensor repositioning is performed between training and test trials. The root mean square error (RMSE), expressed in degrees ($^\circ$), is used as the evaluation metric.

[5] Dataset was accessed via direct request to the authors.

[6] Dataset was accessed via direct request to the authors.

A.5 Additional Details on donwstream set-up

Table 4 reports the hyperparameters employed for each downstream task under the architectures evaluated in this work. The reader can refer to the Downstream setup paragraph of Section 4.2 for additional details.

Table 4: Hyperparameters selected for downstream adaptation. Hyperparameters are optimized independently for each downstream task. The FE setting uses the same hyperparameters as FT, while FS is optimized independently. Dropout is used in the CNN cross-channel fusion module.

Dataset / Task	Batch size	Dropout	LR	Epochs	Weight decay	Layer-wise decay
From scratch (FS)						
UltraPro-Gesture	32	0.0	0.001	60	0.1	-
UltraPro-Wrist	32	0.0	0.001	60	0.1	-
PosePos-USVR-Pose	256	0.0	0.001	24	0.1	-
PosePos-USVR-Position	256	0.0	0.001	24	0.1	-
HWR-3DoF	256	0.0	0.001	60	0.01	-
HWR-23DoF	256	0.0	0.001	80	0.1	-
K2MUSE-gait	16	0.0	0.001	8	0.1	-
K2MUSE-joint	16	0.0	0.001	80	0.1	-
Fine-tuning (FT) / Frozen encoder (FE)						
UltraPro-Gesture	32	0.0	0.001	30	0.05	0.9
UltraPro-Wrist	32	0.0	0.001	30	0.05	0.9
PosePos-USVR-Pose	256	0.2	0.001	12	0.2	0.9
PosePos-USVR-Position	256	0.2	0.001	12	0.2	0.9
HWR-3DoF	256	0.0	0.001	15	0.05	0.9
HWR-23DoF	256	0.1	0.001	40	0.2	0.9
K2MUSE-gait	16	0.0	0.001	8	0.05	0.75
K2MUSE-joint	16	0.0	0.001	40	0.05	0.75

Table 5: Hyperparameters selected for the CNN baseline on each downstream benchmark. The CNN baseline is trained from scratch for all tasks.

Benchmark	Batch size	Dropout	LR	Epochs	Weight decay	# Filters
UltraPro-Gesture	32	0.25	0.0001	120	0.0	[3, 6]
UltraPro-Wrist	32	0.25	0.0001	120	0.0	[3, 6]
PosePos-USVR-Pose	256	0.25	0.001	12	0.0	[10, 21]
PosePos-USVR-Position	256	0.25	0.001	12	0.0	[10, 21]
HWR-3DoF	256	0.05	0.001	15	0.0	[15, 31]
HWR-23DoF	256	0.05	0.0001	80	0.0	[15, 31]
K2MUSE-gait	16	0.05	0.0001	32	0.0	[6, 12]
K2MUSE-joint	16	0.05	0.001	80	0.0	[6, 12]

A.6 Ablation on different pre-processing strategy

We evaluate different pre-processing strategies for model pre-training: raw, band-pass filtered, and envelope-based extraction. Raw US signals are first pre-processed to suppress out-of-band components while accounting for the different transducer characteristics across datasets. Specifically, each signal is filtered using a 4th-order Butterworth band-pass filter centered at the transducer center frequency f_c, with a fractional bandwidth of 80%. Accordingly, the lower and upper cutoff frequencies are set to $0.6f_c$ and $1.4f_c$, respectively. The signal envelope is subsequently extracted using the Hilbert transform (Liu, 2012). Table 6 shows the results. Raw is the best input on three of the four regression benchmarks, while the envelope is the best on all four classification benchmarks. No representation is therefore uniformly superior: the optimal pre-processing depends on the downstream task, with the phase-preserving raw signal favoring continuous kinematic regression and the demodulated envelope favoring discrete classification.

Table 6: Ablation of the input representation the encoder is pre-trained on. Entries are the mean ± standard deviation over three downstream seeds; accuracies are percentages and every RMSE is expressed in degrees. Best results are highlighted in bold.

Benchmark	Metric	Raw	Band-pass	Envelope
UltraPro-Gesture	Accuracy ↑	84.1 ± 0.2	83.8 ± 0.4	$\mathbf{85.0 \pm 0.3}$
UltraPro-Wrist	R^2 ↑	0.932 ± 0.000	0.922 ± 0.001	$\mathbf{0.938 \pm 0.001}$
PosePos-USVR-Pose	Balanced acc. ↑	75.5 ± 0.3	76.1 ± 0.8	$\mathbf{78.2 \pm 1.7}$
PosePos-USVR-Position	Balanced acc. ↑	83.0 ± 0.4	82.9 ± 1.2	$\mathbf{84.6 \pm 0.5}$
HWR-3DoF	RMSE ↓	$\mathbf{9.34 \pm 0.03}$	9.67 ± 0.03	9.78 ± 0.02
	R^2 ↑	$\mathbf{0.750 \pm 0.002}$	0.735 ± 0.002	0.728 ± 0.001
HWR-23DoF	RMSE ↓	$\mathbf{13.27 \pm 0.03}$	13.80 ± 0.07	13.34 ± 0.11
	R^2 ↑	$\mathbf{0.488 \pm 0.004}$	0.448 ± 0.003	0.466 ± 0.004
K2MUSE-gait	Accuracy ↑	70.8 ± 0.0	70.5 ± 0.1	$\mathbf{71.0 \pm 0.0}$
K2MUSE-joint	RMSE ↓	$\mathbf{4.76 \pm 0.00}$	4.89 ± 0.00	4.80 ± 0.00
	R^2 ↑	$\mathbf{0.775 \pm 0.001}$	0.770 ± 0.000	0.774 ± 0.001

A.7 Ablation: advantages of using ALiBi

We evaluate the same pre-training and downstream strategy with the multi-tokenizer, just comparing whether the additive, symmetric distance bias is applied to the attention scores or not. Table 7 shows the results. ALiBi is consistently better on two of the four regression benchmarks, HWR-3DoF and HWR-23DoF on both metrics, and on the two PosePos-USVR classification benchmarks. Removing it is never better on any benchmark, and on the remaining tasks the two encoders are comparable. The bias is therefore never harmful; it does not add any parameters to learn, and we therefore decided to keep it.

Table 7: Ablation of the ALiBi positional bias. Entries are the mean ± standard deviation over three downstream seeds; accuracies are percentages and every RMSE is expressed in degrees. Best results are highlighted in bold.

Benchmark	Metric	With ALiBi	Without ALiBi
UltraPro-Gesture	Accuracy ↑	$\mathbf{84.1 \pm 0.2}$	83.9 ± 0.4
UltraPro-Wrist	R^2 ↑	$\mathbf{0.932 \pm 0.000}$	0.931 ± 0.001
PosePos-USVR-Pose	Balanced acc. ↑	$\mathbf{75.5 \pm 0.3}$	74.0 ± 0.9
PosePos-USVR-Position	Balanced acc. ↑	$\mathbf{83.0 \pm 0.4}$	80.4 ± 0.5
HWR-3DoF	RMSE ↓	$\mathbf{9.34 \pm 0.03}$	9.61 ± 0.03
	R^2 ↑	$\mathbf{0.750 \pm 0.002}$	0.742 ± 0.002
HWR-23DoF	RMSE ↓	$\mathbf{13.27 \pm 0.03}$	13.72 ± 0.17
	R^2 ↑	$\mathbf{0.488 \pm 0.004}$	0.449 ± 0.018
K2MUSE-gait	Accuracy ↑	$\mathbf{70.8 \pm 0.0}$	$\mathbf{70.8 \pm 0.0}$
K2MUSE-joint	RMSE ↓	4.758 ± 0.003	$\mathbf{4.756 \pm 0.001}$
	R^2 ↑	$\mathbf{0.775 \pm 0.001}$	$\mathbf{0.775 \pm 0.000}$

A.8 Qualitative Analysis of Inter-Session Distribution Shift

To further investigate the effect of the evaluation protocol, we visualize the learned feature distributions for an *intra-session* and an *inter-session* dataset. Figure 5 shows t-SNE projections of the features extracted after the cross-channel fusion module for training and validation samples. On UltraPro-Gesture, where no sensor repositioning is applied, the two distributions largely overlap. In contrast, PosePos-USVR-Pose exhibits a more pronounced separation between training and validation samples, consistent with the additional variability introduced by acquisitions performed across different sessions and sensor placements. While t-SNE provides only a qualitative visualization and should not be interpreted as a quantitative measure of distribution shift, this observation is consistent with the greater challenge associated with the inter-session evaluation setting.

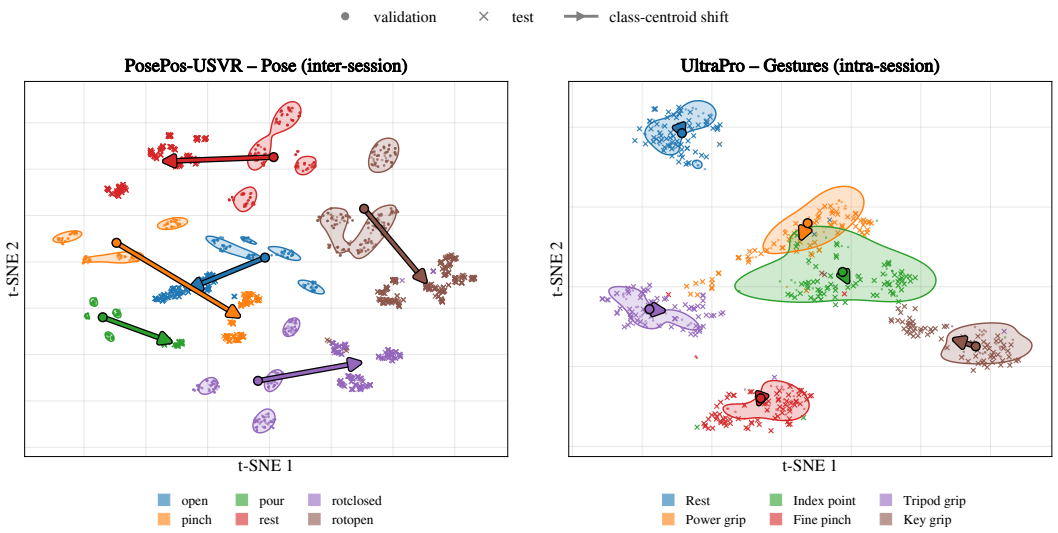


Figure 5: t-SNE visualization of features extracted after the cross-channel fusion module for UltraPro-Gesture and PosePos-USVR-Pose. Points correspond to training and validation samples, as indicated by the color coding. Training and validation representations largely overlap for the intra-session UltraPro-Gesture dataset, whereas a more pronounced separation is observed for the inter-session PosePos-USVR-Pose. The visualization provides qualitative evidence of the increased acquisition-induced distribution shift associated with the inter-session evaluation protocol.

A.9 Robustness across runs

Table 8: Downstream performance across all evaluation benchmarks, reporting variability across repeated runs. Results are reported as mean±std, where performance is first averaged across subjects for each run and subsequently aggregated across runs. In the *Original* column, the standard deviation is instead across subjects, as in Table 2.

Benchmark	Reference work	Metric	Original	BaseCNN	FS	FE	FT
UltraPro – Gestures	Yang et al. (2022)	Accuracy ↑	$79.7^{\dagger,\ddagger}_{\pm 7.9}$	$78.1_{\pm 0.8}$	$77.3_{\pm 1.2}$	$80.5_{\pm 1.0}$	$\mathbf{84.1}_{\pm 0.2}$
UltraPro – Wrist		R^2 ↑	$0.924^{\dagger,\ddagger}_{\pm 0.021}$	$0.718_{\pm 0.055}$	$0.919_{\pm 0.001}$	$0.924_{\pm 0.003}$	$\mathbf{0.932}_{\pm 0.000}$
PosePos-USVR – Pose	Spacone et al. (2026b)	Bal. Accuracy ↑	$69_{\pm 7}$	$68.0_{\pm 0.3}$	$51.4_{\pm 1.0}$	$\mathbf{77.4}_{\pm 1.2}$	$75.5_{\pm 0.3}$
PosePos-USVR – Position		Bal. Accuracy ↑	$76_{\pm 6}$	$79.1_{\pm 0.5}$	$62.9_{\pm 0.8}$	$80.8_{\pm 0.4}$	$\mathbf{83.0}_{\pm 0.4}$
HWR-3DoF	Spacone et al. (2025b)	RMSE ↓	$11.26^{\dagger}_{\pm 1.11}$	$10.26_{\pm 0.11}$	$10.13_{\pm 0.11}$	$11.27_{\pm 0.04}$	$\mathbf{9.34}_{\pm 0.03}$
		R^2 ↑	–	$0.708_{\pm 0.003}$	$0.713_{\pm 0.005}$	$0.655_{\pm 0.001}$	$\mathbf{0.750}_{\pm 0.002}$
HWR-23DoF	Spacone et al. (2026a)	RMSE ↓	$14.60_{\pm 1.98}$	$14.43_{\pm 0.33}$	$19.37_{\pm 0.58}$	$15.76_{\pm 0.18}$	$\mathbf{13.27}_{\pm 0.03}$
		R^2 ↑	$0.305_{\pm 0.05}$	$0.454_{\pm 0.017}$	$0.132_{\pm 0.038}$	$0.402_{\pm 0.012}$	$\mathbf{0.488}_{\pm 0.004}$
K2MUSE-gait	Li et al. (2026)	Accuracy ↑	$\mathbf{71.9}_{\pm 16.9}$	$69.8_{\pm 0.2}$	$70.0_{\pm 0.1}$	$70.0_{\pm 0.2}$	$70.8_{\pm 0.0}$
K2MUSE-joint		RMSE ↓	6.7	$5.28_{\pm 0.04}$	$5.18_{\pm 0.01}$	$5.15_{\pm 0.02}$	$\mathbf{4.76}_{\pm 0.00}$
		R^2 ↑	–	$0.745_{\pm 0.003}$	$0.762_{\pm 0.002}$	$0.761_{\pm 0.004}$	$\mathbf{0.775}_{\pm 0.001}$

†,‡ See the footnotes of Table 2.

Table 8 reports the results achieved by FOCUS-Small. Compared to Table 2, here we report the standard deviation computed over three downstream seeds, highlighting the stability of FOCUS across runs.

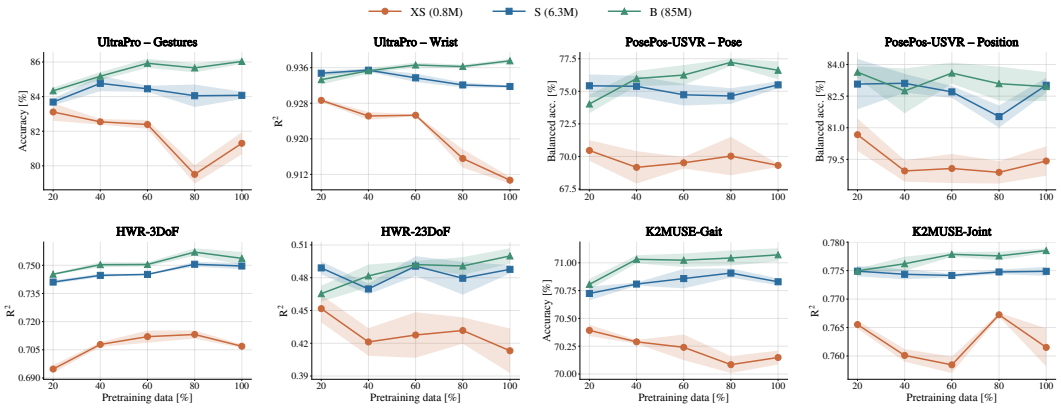
A.10 Ablation on Model Capacity and Label Efficiency

Table 9 reports the detailed configurations of FOCUS-XSmall, FOCUS-Small, and FOCUS-Big, including the corresponding model sizes used in the model-capacity ablation.

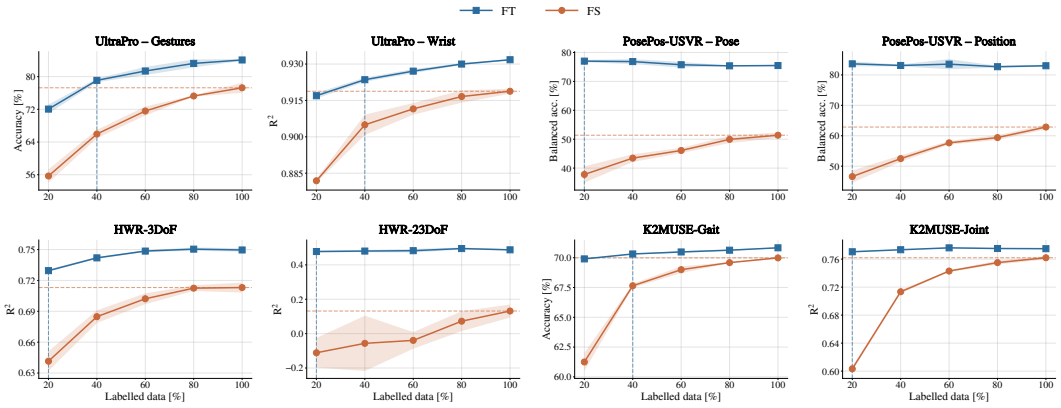
Figure 6 reports the complete results of the scaling experiments. Figure 6(a) shows downstream performance as a function of model capacity and the amount of pre-training data, while Figure 6(b) reports the effect of the amount of labelled downstream data under the FT and FS settings.

Table 9: FOCUS model configurations used to study the effect of model capacity. The embedding dimension E is shared between the encoder and decoder. Parameter counts refer to the encoder only (decoder and downstream modules excluded).

Model	Params.	E	Enc. blocks	Enc. heads	Dec. blocks	Dec. heads
FOCUS-XSmall	0.80 M	128	4	4	2	4
FOCUS-Small	6.33 M	256	8	8	4	8
FOCUS-Big	85.10 M	768	12	12	6	16



(a) Model capacity and pre-training data scale



(b) Labelled downstream data scale

Figure 6: Complete scaling results across all downstream benchmarks. **(a)** Effect of model capacity and pre-training data scale on downstream performance. **(b)** Effect of labelled downstream data availability, comparing the FT and FS settings. Lines report the mean and shaded regions indicate ± 1 standard deviation over three seeds.

A.11 Computational Complexity and Edge-Deployment Potential

Table 10 reports the parameter count and computational complexity of FOCUS-Small across all downstream tasks.

In typical wearable AUS systems (Yang et al., 2021; Frey et al., 2022), channels are acquired sequentially through time-multiplexed transmission. FOCUS is well suited to this regime: the tokenizer and the encoder process each channel independently, and channels interact only in the cross-channel readout. The encoder representation of a channel therefore depends on that channel alone and can be cached. After a start-up phase covering one full acquisition cycle, each newly acquired channel requires a single tokenizer–encoder pass, after which the readout and the prediction head are evaluated on the cached representations of all C channels. Because the encoder processes channels independently, the sequential scheme is equivalent to all-channel inference by construction.

As shown in Table 10, a channel update costs 669–856 MMACs, a 4.0–$7.9\times$ reduction with respect to all-channel inference, at the price of caching $C \times S \times E$ activations (e.g., ≈ 254 k values for UltraPro-Gesture).

Table 10: Parameter count and computational complexity of FOCUS-Small across downstream benchmarks. C denotes the number of AUS channels. Sequential inference processes one newly acquired channel while reusing cached representations of the remaining channels (Seq. = Enc./ch. + Readout); all-channel inference recomputes all C channel representations (All-ch. = C·Enc./ch. + Readout.)

		Parameters				MACs [M]			
Benchmark	C	**Enc. [M]**	**CNN [k]**	**Head [k]**	**Tot. [M]**	**Enc./ch.**	**Readout**	**Seq.**	**All-ch.**
UltraPro-Gesture	8	6.332	262.2	6.0	6.600	844.6	8.14	852.8	6765.3
UltraPro-Wrist	8	6.332	262.2	1.0	6.595	844.6	8.13	852.8	6765.3
PosePos-USVR-Pose	6	6.332	196.7	4.6	6.533	664.2	4.73	668.9	3990.0
PosePos-USVR-Position	6	6.332	196.7	2.3	6.531	664.2	4.72	668.9	3990.0
HWR-3DoF	4	6.332	131.2	2.4	6.466	671.3	3.28	674.6	2688.6
HWR-23DoF	4	6.332	131.2	18.4	6.482	671.3	3.30	674.6	2688.6
K2MUSE-gait	4	6.332	131.2	5.0	6.468	852.0	4.07	856.0	3411.9
K2MUSE-joint	4	6.332	131.2	3.0	6.466	852.0	4.06	856.0	3411.9

Table 11 estimates the parameter-storage requirements of FOCUS-Small. We assume a typical quantization scheme of embedded inference toolchains (Jacob et al., 2018): INT8 weights, INT32 biases, and FP32 normalization parameters. Only the tokenizer branch of the target sampling frequency is stored ($P = 8$ for UltraPro and K2MUSE, $P = 4$ for the WULPUS datasets). Under these assumptions, the encoder dominates the parameter storage with approximately 6.40 MB (25.3 MB in FP32), whereas the cross-channel fusion block and the prediction head together require 134–269 kB. The encoder weights can therefore be kept in external memory.

Table 11: Estimated parameter-storage footprint of FOCUS-Small, assuming a typical quantization scheme of embedded inference toolchains: INT8 weights, INT32 biases, FP32 normalization parameters (1 kB $= 10^3$ B, 1 MB $= 10^6$ B). *Encoder*: weights held in external memory and streamed once per update. Analytical parameter-storage estimates; the accuracy of the quantized model is not evaluated.

			Weights		
Benchmark	C	S	**Encoder [MB]**	**CNN + head [kB]**	**Total [MB]**
UltraPro-Gesture	8	124	6.40	268.5	6.67
UltraPro-Wrist	8	124	6.40	263.5	6.67
PosePos-USVR-Pose	6	99	6.40	201.6	6.60
PosePos-USVR-Position	6	99	6.40	199.3	6.60
HWR-3DoF	4	100	6.40	133.9	6.54
HWR-23DoF	4	100	6.40	149.9	6.55
K2MUSE-gait	4	125	6.40	136.4	6.54
K2MUSE-joint	4	125	6.40	134.4	6.54