

A Overall Appendix Structure

This appendix offers extended details to support and complement the main manuscript. Section B elaborates on the specifics of the PianoKPM Dataset, including a datasheet, dataset insights, and limitations. Section C describes the proposed PianoKPM Net, completing the architecture, training and inference details, and limitations. Section D outlines the experimental implementation for architectural and held-out evaluations and includes additional visualization results that further substantiate the main findings. Finally, we discuss the broader implications in Section E, offering perspectives on potential impacts and future applications.

B PianoKPM Dataset Details

B.1 Datasheet

A standardized datasheet is provided following the methodology proposed by Gebru et al. [63].

B.1.1 Motivation

PianoKPM Dataset aims to construct the first multimodal dataset capturing professional pianists' muscle activities (EMG), hand postures, audio, and keystroke motions during piano performances. Since EMG can provide non-invasive access to internal neuromuscular signals, this dataset enables research on the pose-EMG correspondence understanding and data-driven modeling of dexterous motor control. Our primary objective is EMG inference, a task that offers significant potential to enhance embodied interaction, skill acquisition, healthcare, and rehabilitation.

B.1.2 Composition

The dataset consists of 35,000 PKL files, including 7,000 sample sets, and each contains EMG, keystrokes, and hand pose data captured from three camera views. All data have been temporally synchronized and undergone standard preprocessing procedures, including filtering, normalizing, downsampling, and cleaning. In total, the dataset comprises performance data collected from 20 professional pianists across 7 designed tasks, each repeated 50 times per participant. The release also includes two dataset configuration JSON files, specifying the training, validation, and test splits used in architectural and held-out evaluations in Section 5. All data have been fully anonymized to remove personally identifiable information. The dataset can be found in: <https://github.com/Nips25PianoEMGResearcher/PianoKPMNet.git>. We plan to release the raw dataset in the future, additionally including 3-view videos (720p 60FPS), raw EMG signals (2000 Hz), audio, and raw keystroke signals (1000 Hz).

B.1.3 Collection Process

Data Collection for Muscle Activity Estimation Using Deep Learning


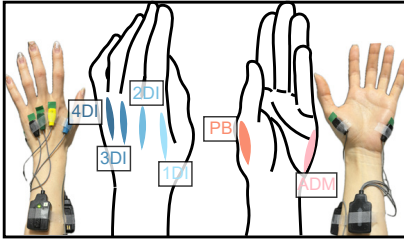
Aim: Estimate hand muscle activity in a piano performance using deep learning techniques.

Main Task: Participants will wear 6 sensors on their left hand and repeatedly perform 7 musical piece. The following four types of data will be collected:

Collected Data: Electromyography (EMG) data, hand motion videos, keystroke data, performance audio.

Duration: Approximately 2 hours

Compensation: 6,000 JPY



After each performance, please relax your hand and keep it horizontal, making sure not to touch the keyboard.

We will give a signal before each performance, so please begin only after the signal is given. *like this* →

If you would like to take a break, feel free to let us know at any time.

Figure 6: The introduction slide presented to participants before the study. The upper-right corner illustrates the six specified hand muscles and the two types of EMG sensor setups.

Apparatus Setup. Figure 2 (a) demonstrates that all modality data is collected in a standardized piano studio equipped with a Shigeru Kawai SK2L grand piano (L: 180 cm, W: 152 cm, H: 102 cm, 324 kg). As Figure 6 upper-right corner shows, two types of wireless EMG sensors (Delsys, Natick, MT, USA) are employed to record GT EMG at 2000 Hz with wireless transmission latency ≤ 40 ms. A Trigno Quattro sensor (4-channel, 25 g) records signals from Muscle 1DI to 4DI on the dorsal side of the hand, while two Trigno Mini sensors (1-channel, 19 g) captured Muscle ADM and PB on the palmar side. All sensor heads (25 x 12 x 7 mm) are placed on target muscles, with sensor bodies (27 x 46 x 13 mm) affixed to the forearm in non-obstructive positions. The skin is prepped with alcohol to reduce the impedance. To minimize interference with performing, we employ a markerless multiview motion capture system comprising three synchronized RGB cameras (1280 x 720, 60 FPS) with audio. Cameras are mounted above the keyboard center, far left, and far right around the piano. Keystroke motions are recorded using a contactless optical sensor system from prior work [62], which measures the vertical displacement of all 88 keys at 1 ms temporal and 0.01 mm spatial resolution.

Participants Recruitment. We recruit twenty highly skilled professional pianists (16 identified as females, 4 as males), aged 20-42 (M : 26.1, SD : 5.1), with 10-38 years (M : 20.9, SD : 6.0) of formal piano training. All participants have previously received top awards in piano competitions and are actively engaged in piano-related education and research, indicating their professional-level expertise. Twelve participants (60%) have prior experience performing with EMG sensors, suggesting a high degree of familiarity and comfort with the equipment. This indicates that over half of the participants can minimize potential interference from the sensors, supporting the reliability of the recorded EMG data. Ethical review processes are done by a local Institutional Review Board (IRB). Before the study, all participants are informed about the research using an introduction slide in Figure 6 and are asked to review and sign an IRB-reviewed consent form. They can ask questions and are free to withdraw from the study at any time. A retraction of consent form is also provided in advance to allow them to revoke consent if desired. All collected data are fully anonymized to remove any personally identifiable information. Each pianist is compensated at a rate of 3,000 JPY per hour (in total 120,000 JPY across all participants).

ID	Name	Sheet Music
T1	Ascending scale	
T2	Descending scale	
T3	Third Dyad 1	
T4	Arpeggio	
T5	Third Dyad 2	
T6	Third Dyad Trill	
T7	All Finger Trill	

Figure 7: Task descriptions. The name, sheet music, and fingering patterns are provided.

Task Design. The performance tasks are designed to capture a wide range of kinematics. In consultation with a professional pianist and neuromuscular researcher, we design seven left-hand tasks of comparable difficulty: two *Scales* in opposite directions, one *Arpeggio*, two *Third Dyads*, one *Third Dyad Trill*, and one *All-finger Trill*. All tasks are exclusively restricted to the left hand, as pianists typically exhibit less accurate strength control over the left hand compared to the right. Descriptions of the detailed sheet music and fingering patterns in each task can be found in Figure 7. In Appendix D.2, we follow the recommendation of an expert piano instructor to select the T4, *Arpeggio*, as the held-out task, which typically involves wider finger spans and faster positional transitions, encompassing distinct and challenging muscle activation patterns compared to other tasks.

Performance Styles. Each participant repeats each task fifty times, covering several distinct performance styles to induce varied muscle activations. For each task, trials 1 to 20 involve a progressive increase in intensity and volume, gradually from pianississimo (*ppp*) to fortississimo (*fff*). Trials 21 to 30 are performed in a *Legato* style, emphasizing smooth transitions with relaxed arm movement and equal finger descent and ascent. Trials 31 to 40 follow a *Staccato* style, characterized by short, crisp sounds, impulsive breaths, and rapid finger release. Finally, trials 41 to 50 are executed as fast as possible to capture high-speed muscular dynamics. These diverse performing styles provide a rich range of muscle recruitment patterns, laying a solid foundation for the collected data diversity and subsequent estimation algorithm’s robustness.

77 B.1.4 Preprocessing/cleaning/labeling

EMG and Keystroke Preprocessing. As detailed in Section 4.1.1, EMG signals undergo a series of preprocessing steps, including filtering, normalization, downsampling, and temporal alignment. Keystroke motions are first normalized to the range -1 to 1 based on the individual key’s specific minimum and maximum heights recorded before the collection, followed by synchronization with the corresponding EMG and pose data.

Hand Pose Extracting and Preprocessing. To accurately capture the hand postures, we adopt the differentiable parametric MANO model [15] as the basis representation for hand annotation. The 3D joint positions $\mathbf{P}_{3D} \in \mathbb{R}^{21 \times 3}$ and mesh vertices $\mathbf{V}_{3D} \in \mathbb{R}^{778 \times 3}$ are computed with pose θ and shape β , through functions $\mathbf{V}_{3D} = M(\theta, \beta)$ and $\mathbf{P}_{3D} = P_{\text{reg}}(M(\theta, \beta))$. To reduce setup complexity and better align with real-world general scenarios, we deliberately avoid multi-camera calibration. Instead, we first apply ViTPose [64] to detect 2D pose keypoints and bounding boxes of the hand region, based on which each frame is cropped to solve inhomogeneity across hand spatial locations. Subsequently, HaMeR [16], a SOTA transformer-based hand pose estimation model, is employed to reconstruct 3D hand postures with improved accuracy and robustness. The PianoKPM Dataset comprises 5.0 million images from three-viewpoint cameras, and each frame is annotated with 3D hand postures generated by HaMeR. However, in preliminary experiments, we observe the depth (Z-axis) estimation is not sufficiently accurate to meet the requirements for precise pose inference. As such, in the subsequent network training stage, we instead use the projected 2D keypoints from the right-view to replace the depth cues from the top-view as $\mathbf{P}_{2D} \in \mathbb{R}^{V \times J \times 2}$ (V : views, J : joints) and customize a fusion pose encoder to extract semantically meaningful 3D pose representations. Due to the limitations of single-frame posture estimation algorithms based solely on visual input, we apply refinement and post-processing to the extracted hand pose data in Section 4.1.1.

100 B.1.5 Uses

The dataset, associated code, and dataset split configurations are released to advance academic research in EMG-based learning and modeling for purely non-commercial purposes. The provided codebase, implemented on the popular framework PyTorch, is modular and adaptable to alternative application scenarios. We welcome and encourage its use as a reference for related research on biosignal inference and multimodal learning.

106 B.1.6 Distribution and Maintenance

An example subset of the PianoKPM Dataset and the code for reproducing experimental results are available at <https://github.com/Nips25PianoEMGResearcher/PianoKPMNet.git>. The full dataset and raw data will be also hosted on a public cloud storage platform in the future. Contributions are welcome from the broader research community, and ongoing maintenance, updates, and issue tracking will be managed and distributed through the GitHub repository.

112 B.2 Dataset Insights

The collection of the PianoKPM dataset enables the analysis of correlations between EMG and other modalities. Figure 8 illustrates the visual distribution of EMG features across different tasks and users. We align the EMG of each performance to a fixed length then reduce the feature dimensionality into a two-dimensional map using t-SNE [71], and the results reveal clear and distinct clustering by task. For instance, T7 (brown) is predominantly located in the lower-left region, while T5 (light green) is in the lower-right, indicating EMGs in different tasks have different characteristics. On the

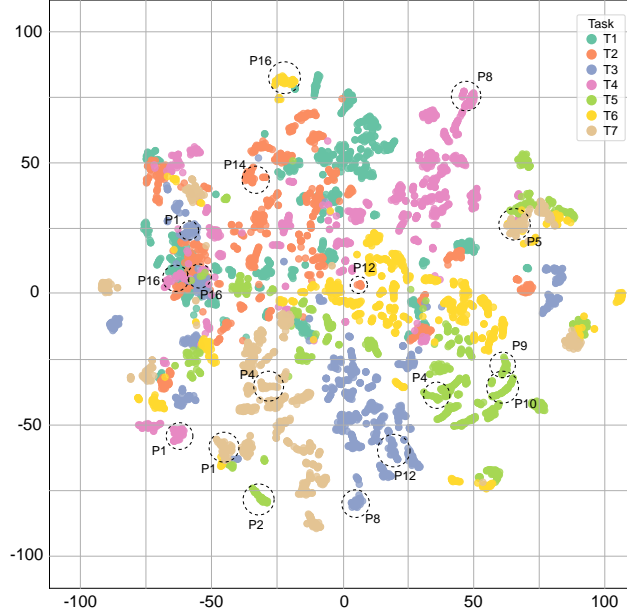


Figure 8: Visualization for EMG features after t-SNE dimensionality reduction. Different colors distinguish different tasks, and dashed circles highlight some example users’ EMG data.

other hand, different users’ EMG features of the same task may appear nearby. For example, P9 and P10 performing T5 (light green, lower-right) exhibit similar EMG patterns. This suggests that the model has the potential to learn cross-user invariant features. Such capacity may partly explain the observation in Section 5.2 that the model generalizes more effectively across users than across tasks. However, the left region of the plot shows overlap among different tasks, indicating that the EMGs for certain tasks are less distinguishable. To address this issue, we design the PianoKPM Net, with an advanced feature extraction module, novel network architecture, and specialized loss function.

B.3 Dataset Limitations

Data Fidelity. While we leverage the SOTA algorithm for 3D hand annotations, the single-frame-based inference still suffers from temporal jitter and error estimations. Since vision-based approaches rely on camera inputs, these limitations are often difficult to overcome due to occlusion and low lighting. Moreover, our system only captures the 3D positions of 21 hand joints, without estimating the hand mesh, elbow, shoulder, or other upper-body parts, which are nevertheless closely related to muscle activation patterns. On the other hand, during data collection, we occasionally encounter sensor detachment issues caused by hand perspiration during piano performance. Although immediate measures are taken, for example, cooling the hands, re-sanitizing the skin, and repositioning the sensors, minor variations in sensor placement may inevitably occur and affect the fidelity of the recorded EMG. While this variability, rather than being a limitation, may contribute positively to the model’s robustness by exposing it to intra-subject domain shifts. In real-world applications, even EMG signals from the same user may vary across sessions due to subtle changes in sensor placement or physiological conditions, making this type of noise a meaningful aspect of the learning process.

Ethics and Privacy. The combination of hand muscle EMG, hand motions, and keystrokes in the PianoKPM Dataset can serve as a unique biometric signature of a pianist’s neuromotor behavior. Such data may potentially be used to identify individuals or infer sensitive information, physical conditions, or even health status. Appropriate anonymization, encrypted storage, and controlled safeguards are essential to protect participant privacy. Moreover, EMG and motion data recorded during the piano performance may unintentionally capture a pianist’s style interpretation. Clear consent procedures should be established to ensure that participants understand how their data may be analyzed, used, or shared. Despite these concerns, there are societal benefits from the development of the dexterous

skill EMG-motion dataset, supporting internal muscle state feedback, motor supervision, fatigue monitoring, and even embodied interaction applications. See Appendix E for more potential impacts.

C PianoKPM Net Details

C.1 Network Architecture

Multi-Branch Feature Encoder. To extract meaningful representations from input sequences, we utilize a stack of two 1D convolutional blocks. Each block consists of a 1D convolution layer followed by a ReLU activation and dropout with a rate of 0.1. Layer normalization is subsequently applied to stabilize training and improve generalization. Specifically, the two blocks expand the input dimension (postures for 84, keystrokes for 88) to 256 channels with kernel sizes of 11 and 5, strides of 1 and 1, and padding of 5 and 2, to preserve the input size. For layer normalization, we apply it across the channel dimension after transposing the temporal axis, ensuring consistency with the input-output shape requirements. This design allows the network to jointly model global and local temporal patterns hierarchically. Subsequently, all features are combined via element-wise addition and fed into the Time-Channel-Wise Encoder in Section 4.2.1. To capture temporal-channel dependencies in the feature sequences, we adopt a hierarchical time-channel-wise encoder, inspired by a time-depth separable (TDS) structure [75]. The encoder consists of sequential modules, each comprising three main components, a 1D convolution, a stack of TDS blocks, and a linear projection. Conv1DBlocks apply over the time dimension with kernel sizes of 17 and 9, strides of 1 and 1, and padding of 8 and 4, mapping the input channels to a higher-dimensional space determined by $C = channels \times feature_width$. This projection increases representational capacity and aligns the input shape with the expected configuration of the following TDS blocks. TDS blocks include a special module to perform a 2D convolution with a fixed kernel size along the time axis and a channel-wise depth separation along the feature axis. The reshaped vectors enable grouped convolution over temporal windows. The output is then fused back to the original shape and combined via residual connection. A layer normalization follows to stabilize training. Specifically, the kernel size of the Conv2DBlocks layers is (1, 9) and (1, 5), with a stride of (1, 1). The input and output channels are 256 and 64 respectively.

Auto-Regressive Decoder. In Section 4.2.2, for the backbone of the decoder, we utilize a lightweight fully connected Multi-Layer Perceptron (MLP) module to map the high-dimensional fused feature vector into a low-dimensional target space. The MLP is designed to support optional layer normalization and output scaling to enhance training stability and numerical conditioning. The overall architecture of the decoder is straightforward. At each time step, the model concatenates the 64-dimensional embedded encoder features of the current frame with the 6-dimensional predicted EMG from the previous frame. This combined vector is fed into two fully connected layers of size 512 for each, followed by LeakyReLU activation applied. Layer Normalization is applied after hidden layers to mitigate internal covariate shifts and facilitate faster convergence. The decoder finally outputs a 6-dimensional predicted EMG vector. Additionally, a final multiplicative scaling factor of 0.01 is applied to the output to regularize the prediction, which preserves sufficient capacity to capture the nonlinear mapping between high-level features and target signals.

Precision-Structure Hybrid Loss. In the preliminary exploration of EMG inference networks, we found that using only MSE loss causes the network to output "safe" stable EMG values, likely since the GT EMG may exhibit subtle fluctuations. To encourage the network to better learn the EMG structural variations driven by pose and keystroke inputs, we introduce an additional loss based on optimal transport, OT loss, mentioned in Section 4.2.3. The original optimal transport problem between two discrete probability distributions $\mu = \sum_i a_i \delta_{x_i}$ and $\nu = \sum_j b_j \delta_{y_j}$ is formulated as: $\mathcal{L}_{original_ot}(a, b) = \min_{\gamma \in \Pi(a, b)} \sum_{i, j} \gamma_{ij} C_{ij}$. But it is a linear programming (LP) problem without gradient hence not possible for backprop. Thus, referring to prior work [80], we add entropy regularization $H(\gamma)$ to use the Sinkhorn-based optimal transport loss from the Python Geomloss library and compute a soft alignment between the predicted and ground-truth distributions as:

$$\mathcal{L}_{ot} = \mathcal{W}_\epsilon(a, b) = \min_{\gamma \in \Pi(a, b)} \sum_{i, j} \gamma_{ij} C_{ij} - \epsilon H(\gamma) \quad (1)$$

Here, $H(\gamma) = -\sum_{i,j} \gamma_{ij} \log \gamma_{ij}$ is the entropy of the transport plan and ϵ controls the regularization strength. Our subsequent experiments validate that the combination of MSE and OT losses facilitates EMG inference in high local accuracy and faithful global pattern preservation.

C.2 Implementation Details

Hyperparameters. We train the model for 200 epochs to ensure sufficient convergence. The batch size is set to 64 to balance computational efficiency and training stability. EMG and keystroke signals sampled at 1000 Hz use a window length of 1024 for both training and inference. During training, a sliding window with an overlap of 256 is applied for data augmentation. For the 60 FPS pose data, the window length is 60 with an overlap of 15. The number of workers and threads is specified as 0 to ensure reproducibility. The model employs an AdamW optimizer with a learning rate of 0.0001, and a StepLR scheduler with a step size of 20 and a decay factor (gamma) of 0.5. The loss function is a weighted combination of MSE and OT losses, where the weights λ_{mse} and λ_{ot} are both set to 1.

Compute Resources. Model training is performed on a high-performance computing system with an AMD EPYC 9654 96-core/192-thread processor, 768 GiB DDR5-4800 RAM, and NVIDIA H100 SXM5 GPUs, and the entire process takes approximately 14 hours. Notably, neither multi-GPU parallelism nor mixed-precision training is employed. Model inferring is conducted on a more accessible setup with an Intel Core i9-10900X CPU, 128 GB RAM, and an NVIDIA GeForce RTX 4090 with 24 GB GPU memory. The model contains 3.7 million parameters and achieves batch inference within 170 ms (*latency* $\approx 170ms$). Therefore, the inference is lightweight enough to run on desktop-grade hardware without delay or out-of-memory issues, which confirms its suitability for interactive applications requiring timely feedback.

C.3 Network Limitations

Generalization. Generalization remains challenging across nearly all EMG-based research, and our work is no exception. Prior studies have primarily approached this issue through two strategies: (1) expanding dataset scale to capture a wider range of intra- and inter-subject variability [12, 13], or (2) adopting domain generalization methods to enable robust cross-domain transfer [11]. Our future work will extend in both directions. (1) Addressing dataset and network bias: The current dataset may overrepresent certain demographics (e.g., professional pianists), potentially introducing bias and limiting the model’s applicability to broader populations. Future models should be trained on more demographically and behaviorally diverse datasets and incorporate explicit evaluations of generalization across ages, expertise levels, and physical conditions. (2) Improving distributional generalization: Our current experiments focus on the relationship between training set coverage and generalization to unseen users or tasks. A promising direction is to employ transfer learning or few-shot learning to better adapt models across different distributions. Alternatively, future work may leverage large-scale multimodal foundation models to encode stronger muscle-pose priors that facilitate generalization in low-data or domain-shift scenarios.

Other Modalities. Recently, multimodal learning frameworks have been a new rising research hotspot, demonstrating the benefits of incorporating diverse inputs into model training [88]. While the present study focuses on hand motions and keystrokes containing direct physiological and kinematic associations with EMG signals, we retain audio, a post-execution modality, to enable future investigations into multimodal fusion strategies. This choice lays the groundwork for extensible research on richer sensory integration. Looking forward, we plan to incorporate additional modalities such as audio, touch pressure, and visual sheet music, to improve the accuracy, robustness, and semantic interpretability of EMG inference. These efforts are expected to support the development of more comprehensive models for high-level EMG reasoning and nuanced performance understanding.

D Experiment Details

D.1 Baselines Implementation in Architectural Evaluation

NeuroPose [82]: While NeuroPose infers 3D hand poses from EMG, our objective reverses to estimate EMG from motions, optionally enhanced with auxiliary modalities like keystrokes. To this

end, we modify the original NeuroPose U-Net architecture to accommodate differences in input structures and prediction goals. Specifically, NeuroPose utilizes a single-modality input, but our framework incorporates multimodal inputs. Inspired by diffusion-step embeddings in DiffWave [84], we encode keystroke information as constraints and add it to the input of residual layers. The encoder is composed of three sequential Conv-BN-ReLU-MaxPool layers, progressively downsampling the feature sizes by (4×4), (2×4), and (2×2) over temporal and spatial domains. This is followed by five residual blocks with a consistent kernel size of 3×2, while between the first and second residual blocks, we inject the encoded keystroke features into the pose representation. As well, the decoder consists of three similar Conv-BN-ReLU-UpSample modules, upsampling the feature sizes by (4×4), (8×4), and (8×4). A final linear projection maps the output to a 6-channel EMG signal, aligning with our target muscles.

CodeTalker [83]: While the original CodeTalker targets speech-driven 3D facial animation, we build upon its Transformer-based architecture and extend it to our pose-to-EMG inference task. Concretely, pose data and keystroke sequences are first processed by their modality-specific encoders, to temporally align with distinct frequency. The encoded pose representations are subsequently fused via addition, yielding a unified pose embedding. The core component is a transformer block that takes the fused pose embedding as the query (Q) and the encoded keystroke embedding as the key (K/V). This block consists of three sub-modules, each containing an LN-XATTN-ResNet layer and an LN-MLP-ResNet feedforward layer, which enable the model to inject keystroke-informed dynamics into the pose features, facilitating more physiologically plausible EMG prediction. A final FC layer maps the 512-dimension hidden features to the 6-channel EMG output.

D.2 Held-Out Evaluation

Detailed Configurations. Here, we provide a more detailed elaboration of the held-out protocols described in Section 5.2 as well as an additional held-out test set. For *Cross-User*, the held-out users are randomly sampled and for *Cross-Task*, the held-out task (T4 in Figure 7) is chosen by a professional piano teacher, which should be visually out-of-distribution concerning the training stages. In Figure 9 (a), besides *Cross-User* and *Cross-Task*, a most challenging but practically significant scenario, *Cross-User-Task*, is conducted to involve both unseen users and tasks. We partition the dataset into training, validation, and test sets with an approximate ratio of 70% : 10% : 20%. Validation sets are sampled from the same distribution as training sets, while each test set corresponds to a specific condition as described above.

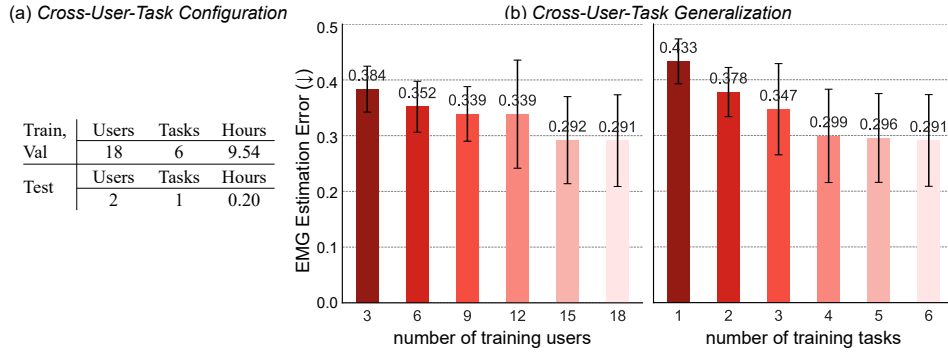


Figure 9: *Cross-User-Task* configuration and results. (a) The test set is split to include both unseen users and task. (b) Model generalization ability across training dataset scale. The bar charts take the same format as in Figure 5.

Cross-User-Task. From Figure 9 (b), we can draw similar conclusions as those discussed in Section 5.2. In *Cross-User-Task*, increasing the number of training users and tasks both contribute to improved performance. Notably, enhancing task diversity leads to a more rapid reduction in EMG estimation error, underscoring the critical role of kinematic and postural variability in facilitating generalization. Moreover, compared to *Cross-User* and *Cross-Task*, the model exhibits inferior generalization performance under *Cross-User-Task*. This highlights the greater complexity of simultaneously

283 adapting to both unseen users and novel tasks, suggesting the need for further methodological
 284 optimization in this scenario.

285 D.3 More Qualitative Results

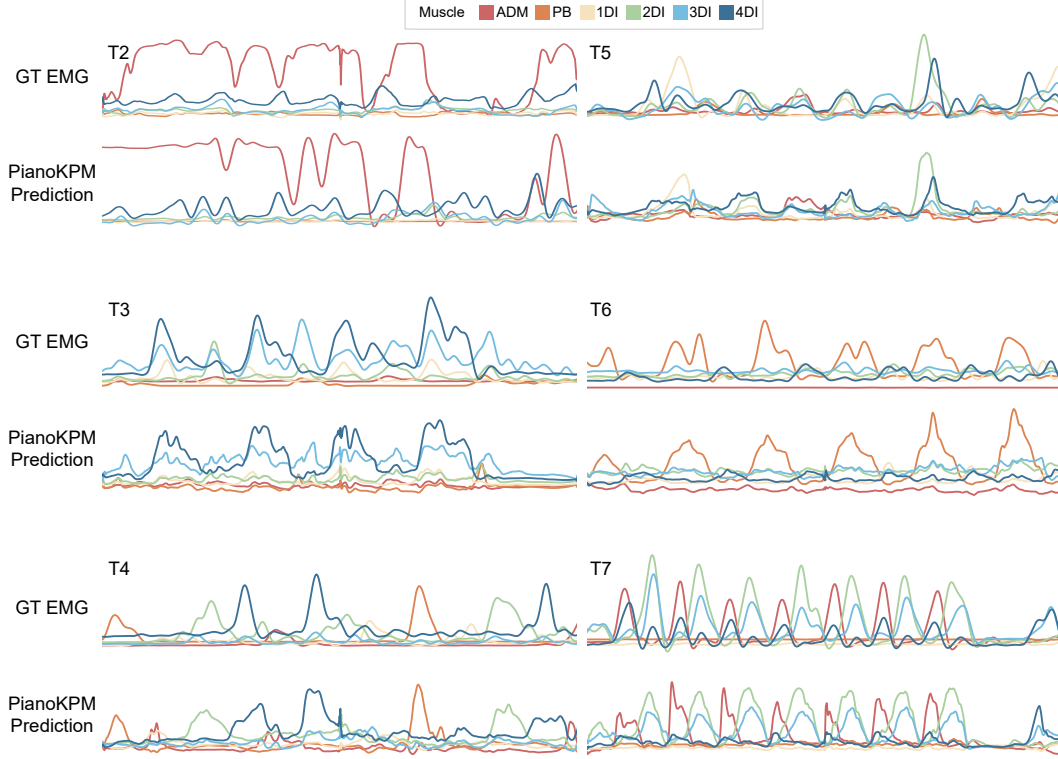


Figure 10: Visualization results for tasks T2 to T7 from several users. Six hand muscle EMGs are represented as line plots. For each task, the first row shows the GT EMG as a reference, while the second row presents the predictions given by PianoKPM Net.

286 **Architectural Evaluation.** Figure 4 presents comparative results for a representative task (T1). In
 287 this section, we further show visualization results for the remaining six tasks (T2-T7) across some
 288 users. As shown in Figure 10, aside from minor fluctuations in some predicted values, PianoKPM
 289 Net successfully captures the general trends of muscle activation and the amplitude-level proximity
 290 to the GT EMG. This performance can be attributed to the precision-structure hybrid loss and the
 291 designed network architecture of the PianoKPM Net. Consequently, the network consistently delivers
 292 accurate and interpretable EMG predictions across all tasks and users.

293 **Held-Out Evaluation.** As shown in Figure 11, the predictions in the left half (*Cross-User* setting)
 294 are relatively more accurate. While the predicted EMG signals may not perfectly align with the GT in
 295 magnitude, the structures and activation trends of muscles are partially preserved, indicating that the
 296 model can capture user-invariant neuromuscular dynamics to some extent. In contrast, the right half
 297 (*Cross-Task* setting) shows more discrepancies. For example, in the top-right example, PianoKPM
 298 erroneously outputs additional activation for Muscle ADM (red); similarly, in the bottom-right
 299 example, Muscle PB (orange) activation is missing. These errors may be due to out-of-distribution
 300 tasks exhibiting kinematic patterns not included during training, which leads to incorrect model
 301 fitting. On the other hand, variations across users appear to be partially captured or characterized by
 302 our advanced network architecture.

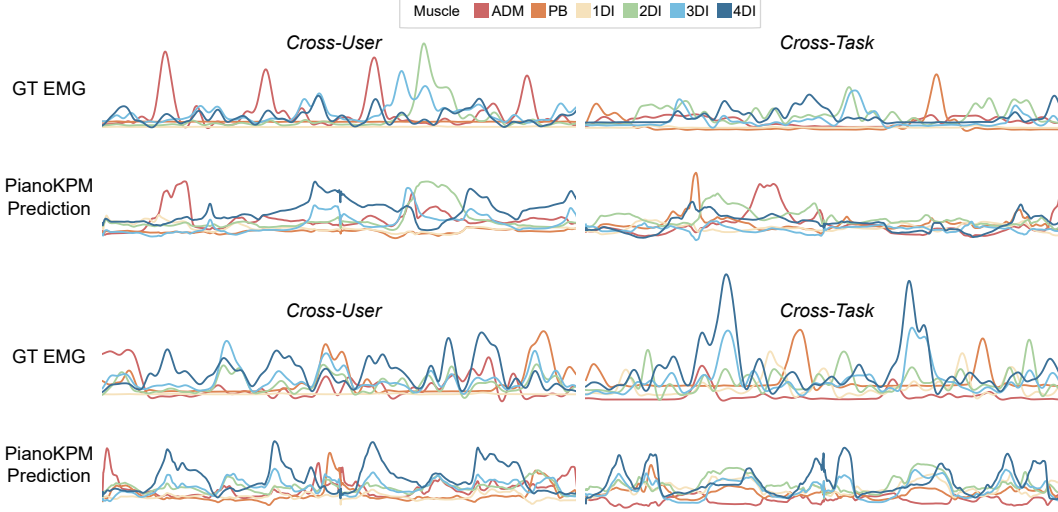


Figure 11: Visualization results under *Cross-User* and *Cross-Task* settings. The left and right respectively display *Cross-User* and *Cross-Task* results.

E Broader Impacts

The PianoKPM framework is proposed for estimating piano hand muscle EMG during dexterous motor tasks by leveraging multiple modalities, such as human-centric pose data and tool-centric keystroke data, which has broad application potential across various domains. In embodied interaction, this technology enables muscle-aware user interfaces by recognizing intent and effort to enhance adaptive feedback. In healthcare and rehabilitation, our low-cost and non-invasive EMG estimation can support remote muscle monitoring during therapy. In digital twins and biomechanical modeling, the predicted EMG can complement kinematics to construct more realistic, individualized, and physiologically grounded digital human models.

Despite its promising utility, the development of EMG estimation models introduces novel ethical and privacy concerns, such as the risk of biometric identification, unintended inference of health conditions, and non-consensual bodily monitoring. Consequently, this study prioritizes informed consent, data security, and transparency to ensure responsible use. We make sure that all participants understand the nature of data collection, provide written consent, and retain the right to withdraw at any time. All data are anonymized to protect privacy, ensuring the study’s compliance with ethical research standards.