From Pose to Muscle: Multimodal Learning for Piano Hand Muscle Electromyography

Ruofan Liu[1,2] **Yichen Peng**[1] **Takanori Oku**[2] **Chen-Chieh Liao**[1]
Erwin Wu[1] **Shinichi Furuya**[2] **Hideki Koike**[1]
{liu.r.ac, peng.y.ag, liao.c.aa, wu.e.aa}@m.titech.ac.jp
t-oku@shibaura-it.ac.jp, shinichi.furuya@sony.com, koike@c.titech.ac.jp
[1]Institute of Science Tokyo, [2]Sony Computer Science Laboratories

Abstract

Muscle coordination is fundamental when humans interact with the world. Reliable estimation of hand muscle engagement can serve as a source of internal feedback, supporting the development of embodied intelligence and the acquisition of dexterous skills. However, contemporary electromyography (EMG) sensing techniques either require prohibitively expensive devices or are constrained to gross motor movements, which inherently involve large muscles. On the other hand, EMGs exhibit dependency on individual anatomical variability and task-specific contexts, resulting in limited generalization. In this work, we preliminarily investigate the latent pose-EMG correspondence using a general EMG gesture dataset. We further introduce a multimodal dataset, PianoKPM Dataset, and a hand muscle estimation framework, PianoKPM Net, to facilitate high-fidelity EMG inference. Subsequently, our approach is compared against reproducible competitive baselines. The generalization and adaptation across unseen users and tasks are evaluated by quantifying the training set scale and the included data amount.

1 Introduction

Human-world interaction is an interdependent loop involving explicit visual, auditory, haptic cues, and implicit physiological signals. Among these, internal information, such as hand muscle electromyography (EMG), is critical for embodied interaction [1, 2], assistive technology [3, 4], and skill acquisition [5, 6], yet remains difficult to access in real-world ubiquitous scenarios due to the cost, obtrusiveness, and expertise required by conventional sensing systems [7].

On the other hand, the emergence of large-scale datasets and deep learning networks increasingly facilitates cross-modal synthesis, which enables transformation between different modalities. For instance, text to image [8], audio to motion [9], or text to music [10]. These developments offer new insights into inferring internal physiological signals like EMG from readily accessible modalities. Also, the invisible yet informative muscle activations are highly correlated with body postures, which is validated in several recently released large-scale datasets containing EMG and motion data. The public datasets, like EMGBench [11] and emg2pose [12], demonstrate the feasibility of classifying and estimating gestures from EMG. However, the inverse task to infer EMG from postures or other modalities, remains substantially challenging and underexplored. The difficulty stems from the richer semantic information embedded in EMG compared to posture alone, as there is no strict one-to-one mapping and the same postures may correspond to different muscle recruitment patterns.

To pioneer this promising yet insufficiently studied field, this work begins by leveraging emg2pose [12], currently the largest and most diverse paired dataset of EMG and hand poses. We pretrain a foundational model capable to learn latent relationships between muscular activations and hand postures, which serves as a preliminary step toward assessing the feasibility of pose-to-EMG inference.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

We further focus on a specific domain of piano performance, a representative scenario that involves intricate and dexterous hand motions. Expert-level pianists rely heavily on precise internal muscle synergies for advanced and technical performances. However, prior research highlights that muscle recruitment patterns can vary significantly across different domains [13], which is also confirmed when we evaluate the pretrained model on PianoMotion10M [14], a hand motion dataset specifically built for piano but without muscle dynamics.

To enhance the model's performance and generalization during piano playing, we introduce the Piano Keystroke-Pose-Muscle Dataset (PianoKPM Dataset) that contains synchronized, high-fidelity data of hand muscle EMG sampled at 2000 Hz, keystroke motions at 1000 Hz, audio, and multi-view RGB videos at 60 FPS. The dataset also includes 3D hand motion data estimated using the MANO model [15] and state-of-the-art (SOTA) method HaMeR [16]. To the best of our knowledge, this is the largest publicly available EMG dataset of professional piano performance, comprising data from 20 expert pianists performing 7 distinct musical tasks, with 12.64 hours of high-quality recordings.

Leveraging this dataset, we propose the Piano Keystroke-Pose-Muscle Network (PianoKPM Net) to infer high-frequency EMG from pose data. Additionally, we explore the benefits of incorporating other modalities, such as key-pressing events, to improve model performance. A comparative evaluation of several baselines and our method is provided to demonstrate the applicability. Moreover, various electrode placement [17, 18], subject anatomy [19, 20], and kinematics may introduce substantial discrepancies in EMG data distribution, posing challenges for generalization. We subsequently investigate the model's performance on held-out users and tasks based on the training set scale and quantify the amount of data required from an unseen distribution to recover high accuracy.

In summary, this paper makes three contributions:

- **A hand muscle EMG estimation framework**, PianoKPM Net, leverages human-centric hand motion data and tool-centric piano keystroke data.
- **A multimodal piano performance dataset**, PianoKPM Dataset, contains simultaneous hand postures, keystroke motions, audio, and miniature hand muscle EMG.
- **A comprehensive evaluation** compares our EMG inference approach with several baselines in piano playing and analyzes its generalization and adaptation across users and tasks.

Instructions regarding accessing and using our PianoKPM dataset and network are provided at `https://github.com/ruofanliu0129/PianoKPMNet.git`.

2 Related Work

2.1 Muscular Dynamics Estimation

Muscular dynamics have shown great potential in augmenting various domains like control interfaces [21, 22], sign recognition [23, 24], motor optimization [25, 26], and recently mixed reality [27, 28]. However, conventional sensor-based muscle measurements demand scarce and resource-intensive instrumentation [29]. Prior research explores computer-assistive algorithms to estimate myoelectrical signals to avoid the high cost, encompassing simulation-based and learning-based methods [30]. Musculoskeletal simulations emerge as promising approaches, while most rely on inverse dynamics and hence are constrained to pre-selected distributions and entail substantial computational overhead [31–38]. On the other hand, data-driven models enhance generalization and lower required resource [39–43], albeit at the inevitable cost of accuracy [44]. All the aforementioned principally focus on the gross motors of large muscles. Several prior studies [45, 46] target muscles involved in non-posed facial expressions and hand movements, yet these early efforts suffer from limited accuracy or small participant validation. Consequently, the intricate and critical small hand muscles in dexterous activities remain largely underexplored [47].

2.2 Relevant Datasets

EMG Datasets. Existing hand-EMG datasets can be categorized into those only involving coarse gesture labels for classification [11, 48–50] and those involving precise hand postural information for regression. Representative datasets of the latter are summarized in Row 1-3 of Table 1. Ninapro [51–54], one of the most established EMG datasets, comprises multiple subsets collected using 10

Table 1: Dataset comparisons. The first three rows list publicly available EMG-hand datasets, the following three correspond to piano performance datasets, and ours is presented in the final row.

Dataset	Year	Public	Size	Subject	Pose	Audio	MIDI	EMG
Ninapro	2014	✓	-	67	✓	-	-	✓
emg2pose	2024	✓	80M	193	✓	-	-	✓
PiMForce	2024	✓	83.2M	21	✓	-	-	✓
PianoHand2.5M	2023	✗	2.5M	21	✓	✓	✓	-
FürElise	2024	✓	2.2M	15	✓	✓	✓	-
PianoMotion10M	2024	✓	10M	14	✓	✓	✓	-
PianoKPM (Ours)	2025	✓	5M	20	✓	✓	✓	✓

Otto Bock electrodes, or 12 Delsys Trigno sensors, or dual Myo armbands, and a 22-sensor dataglove to record hand postures. emg2pose [12] employs a 16-channel sEMG-RD wristband and a 26-camera motion capture system, while PiMForce [55] utilizes an 8-channel sEMG armband with a magnetic sensing-based tracking module, to acquire hand kinematics and muscle dynamics. Some additional EMG datasets on specific skills, such as alpine skiing [56], capture only large muscle activations.

Piano Datasets. Various piano datasets are curated to support different research objectives, such as music generation [57], fingering extraction [58], and gesture recognition [59–61]. Rows 4-6 of Table 1 display datasets providing more accurate hand movement annotations. PianoHand2.5M [62] offers precise ground-truth (GT) labels with a large-scale 3D hand pose dataset from professional pianists. FürElise [63] captures synchronized MIDI key-press data and multi-view videos using a five-camera setup. PianoMotion10M [14] aggregates in-the-wild piano performance videos from the internet, including reconstructed motion, audio, and MIDI data. However, to date, no piano performance dataset simultaneously provides high-quality pose, audio, and MIDI data, along with hand muscle EMG, which is crucial for underlying fine-grained dynamics.

3 Preliminary Exploration: Contrastive Pretraining

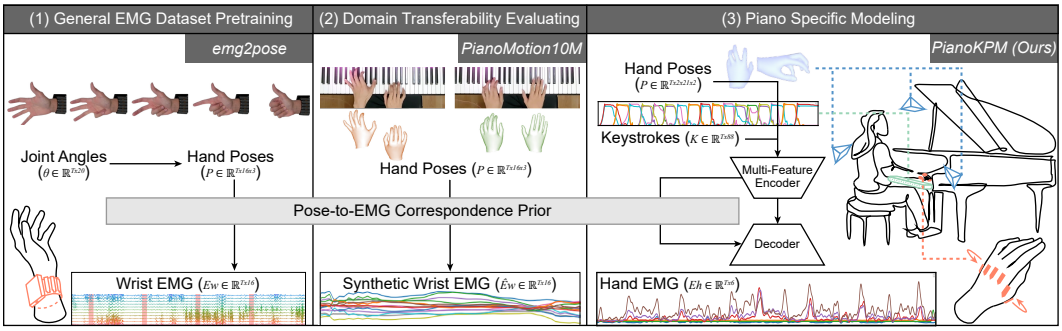

Figure 1: Overall research pipeline. We (1) use a public general EMG dataset to pre-learn latent relationships between hand kinematics and EMG, (2) evaluate this prior on a piano motion dataset but the synthetic EMG fails to yield satisfactory performance, (3) construct our specialized pose-EMG paired dataset which is subsequently used to estimate the precise EMG.

To implement cross-modal EMG estimation and mitigate the cost of large-scale paired EMG-pose collection, we exploit knowledge from public datasets. This section elaborates on the methods we use and questions we encounter in the preliminary study, including pretraining on a general EMG dataset, evaluating in a task-specific domain, and articulating the research motivations. As Figure 1 reveals, the whole pipeline from general pretraining to piano-specific modeling is the cornerstone of our EMG inference framework.

General EMG Dataset Pretraining. In Figure 1 (1), to preliminarily investigate the feasibility of estimating EMG from other modalities, we employ the public emg2pose dataset [12] to pretrain a deep learning framework capable of capturing semantic correlation between hand poses and EMG.

The dataset contains 2000 Hz, 16-channel EMG from a bipolar sEMG-RD wrist band [13], and pose annotations from a 26-camera motion capture system, spanning 193 participants, 370 hours, and 29 stages. A competitive EMG-to-pose baseline is also provided. We then modify its architecture by inverting and adapting the input-output channels to encode inductive biases, which can understand pose-to-EMG correspondence prior and facilitate downstream adaptation through fine-tuning.

Domain Transferability Evaluating. Given that EMG may exhibit distributional shifts across unseen domains, we focus on advanced piano performance with diverse dexterous finger movements to assess the generalization of the pretrained framework. To this end, we utilize the PianoMotion10M dataset [14], which comprises 116 hours of bird's-eye view piano performance videos with 10 million annotated hand poses. The pose-to-EMG correspondence prior is evaluated on this dataset, yet it fails to yield satisfactory results. Although the absence of paired EMG precludes quantitative evaluation, qualitative visualizations of the synthetic wrist EMG in Figure 1 (2) suggest that the predictions are suboptimal.

Research Questions Articulating. The above observations motivate and sharpen the focus of our research. We underscore three reasons for the degraded performance: (1) The piano hand postures differ substantially from generic datasets, undermining the generalization of learned pose-to-EMG correspondence in the target domain. (2) Surface EMG acquired from the wrist aggregates multiple underlying muscle activities, offering limited localization to isolate individual hand muscles. (3) Pose representations are inherently sparse, lacking sufficient contextual and dynamic cues essential for reliable EMG inference.

4 Methodology

To address the three research questions mentioned in Section 3, we respectively (1) construct a piano-centric dataset encompassing diverse hand postures; (2) employ EMG sensors to individually monitor six major hand muscles; (3) incorporate high-frequency keystrokes as an auxiliary modality. Leveraging this multimodal dataset, we further propose a neural network for EMG inference that achieves reliable and accurate results across general benchmarks and our piano-specific dataset.

4.1 PianoKPM Dataset

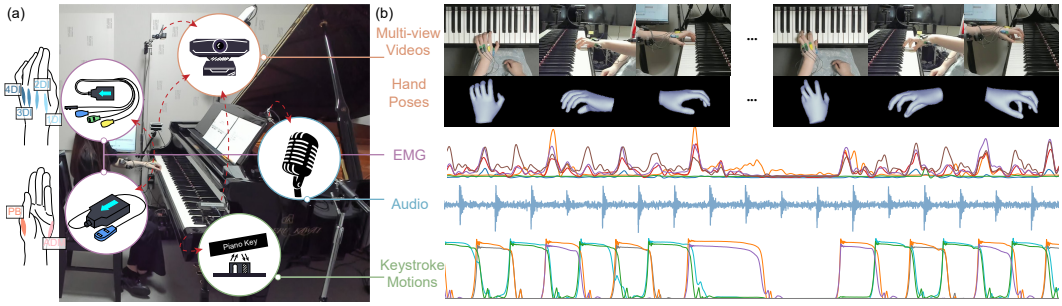

Figure 2: Dataset composition. (a) The studio, apparatus, and six hand muscles. (b) PianoKPM Dataset consists of 60 FPS videos captured from three different views, hand pose annotations, and 2000 Hz 6-channel EMG, 1000 Hz 88-channel keystrokes, and audio collected by specific equipment.

As shown in Figure 2 (a), all data are collected in a studio equipped with a grand piano, three RGB cameras, two EMG sensors, a fixed microphone, and a keystroke sensing system [64]. From a professional pianist and neuromuscular science researcher's advice, EMG signals are recorded from six hand muscles: Abductor Digiti Minimi (ADM), Pollicis Brevis (PB), First to Fourth Dorsal Interosseous (1DI - 4DI). The annotated hand postures are extracted using the SOTA pose estimation method HaMeR [16]. Appendix B provides more dataset details based on standardized datasheet [65].

4.1.1 Data Preprocessing

Inter-Frame Discrepancy-based Motion Refinement. Vision-based pose reconstruction achieves promising results, while inferior outcomes may still arise due to inevitable occlusion and rapid

movement. To clean and refine hand motions, we detect and remove the abnormal frames, then interpolate and smooth the remaining valid data. Given that outlier frames always exhibit variations from their adjacency, we compute the inter-frame differences $d_i = \|\mathbf{x}_i - \mathbf{x}_{i+1}\|_2$ to find the error frames. We then clean the data using time-informed linear interpolation, given the smooth and consistent nature of the hand motion over short durations. For an anomalous frame f, an interpolation coefficient α is computed with its temporally adjacent valid frames, then the current pose values are interpolated via $p = (1-\alpha) \cdot p_{prev} + \alpha \cdot p_{next}$. Once all frames contain valid poses, a moving average filter [66, 67] is applied to enhance temporal smoothness and reject motion artifacts. We further commission a professional data annotation service to verify the final dataset quality.

EMG Preprocessing. To ensure quality, computability, and alignment with other modalities, EMG is preprocessed through filtering, normalizing, downsampling, and task-based cropping. To reduce the noise of raw EMG signals sampled at 2000 Hz, we apply a low-pass Butterworth filter (the 5th-order, cutoff frequency: 12 Hz) [68]. Considering inter-muscle and inter-subject variability in EMG amplitude, we normalize each muscle's EMG by its maximum voluntary contraction (MVC) [69, 70], which is empirically recorded through an MVC test session before data collection trials. This procedure transforms EMG into relative values between -1 and 1, enabling more informative within- and across-subject comparisons of muscle engagement. Subsequently, the 2000 Hz EMG is downsampled to 1000 Hz to synchronize with other modalities. All data are cropped based on the onset and offset of keystroke events, and only those periods when dedicated muscle activities relate to piano playing are included. See Appendix B.1.4 for more EMG preprocessing details.

4.1.2 Data Analysis

Data Statistics. PianoKPM Dataset consists of 20 highly-skilled pianists, 7 tasks, 7,000 performances, and 21,000 videos with a total length of approximately 12.64 hours. Figure 2 (b) illustrates an example performance that includes synchronized 720p 60 FPS RGB videos from three views, corresponding 3D hand pose annotations, 1000 Hz EMG, 1000 Hz keystroke motions, and audio. After preprocessing, our dataset has valid data for 5,026,566 frames of videos and hand postures, along with 28,706,987 frames of EMGs and keystrokes. The scale and variability can ensure diversity and consistency across modalities, serving as the basis for downstream multimodal modeling to accurately predict EMG. Appendix B.1.2 includes more details about the dataset composition.

Data Quality. Following prior work [71], we adopt precision, recall, and F1 metrics, to evaluate the quality of the reconstructed hand motions. Our evaluation protocol defines TP as a frame where the model's prediction validly matches the ground truth. FP refers to a frame that is flagged as invalid by the outlier detection algorithm in Section 4.1.1, while FN denotes a missing frame with the ground truth presence of a hand. Across all reconstructed hand sequences from the three camera views, we obtain 4.83M TPs, 0.14M FPs, and 0.06M FNs. Based on these counts, our model achieves 97.22 precision, 98.74 recall, and 97.97 F1 over the full dataset. See Appendix B.2 for dataset insights [72].

4.2 PianoKPM Net

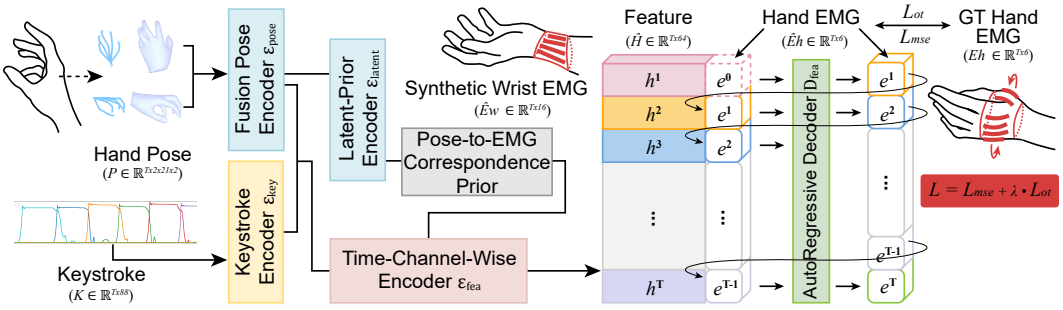

Figure 3: Network architecture. A multimodal hand muscle EMG estimation network enhances human-centric hand pose data by integrating tool-centric keystroke motions.

The potential of estimating EMG from human-centric videos is grounded in the observation that muscle activities are inherently linked to the executed motor actions [54, 73]. Different movement

patterns lead to distinct muscle coordination dynamics, hence producing discriminative EMG signals [48, 74]. However, postural data alone cannot fully decode muscle activity complexity, due to its sparsity and the absence of external force information during human-object interactions. For instance, the same poses performed with or without extrinsic resistance may elicit different muscle engagement, which highlights the need to incorporate tool-centric information for context-aware EMG estimation. Consequently, we augment hand movements with precise keystroke motions. As Figure 3 shows, we devise a multi-branch encoding architecture to utilize the complementary inputs: (1) Input-specific encoders, $\mathcal{E}_{pose}$ and $\mathcal{E}_{key}$, to extract shallow features from postures and keystrokes, (2) Latent-prior encoder, $\mathcal{E}_{latent}$, to integrate prior correlations between pose dynamics and EMG latent space, (3) Time-channel-wise encoder, $\mathcal{E}_{fea}$, to disentangle temporal and feature dependencies. An auto-regressive decoder, $\mathcal{D}_{fea}$, is subsequently introduced to reconstruct EMG aligned with the ground-truth channels from the encoded feature representations. Appendix C.1 provides further details on the structure and parameters of each layer and the design of the loss function, and Appendix C.2 describes the implementation, including hyperparameter settings, computational resources, and possible latency.

4.2.1 Multi-Branch Feature Encoder

Input-Specific Encoder. We first extract shallow features independently from the hand and keystroke motions using input-specific encoders. Due to the inaccuracy of 3D pose depth information mentioned in Appendix B.1.4, the pose encoder uses concatenated projected 2D keypoints $\mathbf{P} \in \mathbb{R}^{T\times V\times J\times 2}$ (T: time length, V: views, J: joints) of each frame to perform cross-view fusion, which reshapes, aligns and aggregates the input vector to produce 3D postural features through: $\mathbf{F}_p = \mathcal{E}_{pose}(\mathbf{P}) \in \mathbb{R}^{T\times D_p}$. In parallel, the keystroke encoder takes the 1D keystroke sequence $\mathbf{K} \in \mathbb{R}^{T\times N}$ (N: keys) as input and outputs keystroke features by: $\mathbf{F}_k = \mathcal{E}_{key}(\mathbf{K}) \in \mathbb{R}^{T\times D_k}$.

Latent-Prior Encoder. To exploit pre-trained priors capturing the statistical heuristics between hand motion and EMG, a latent encoder is introduced to align pose representations with the latent prior module's inputs, which is useful for injecting inductive bias into the model [75]. Formally, the latent encoder maps the input pose features into a physiologically meaningful space by: $\mathbf{Z}_l = \mathcal{E}_{latent}(\mathbf{F}_p) \in \mathbb{R}^{T\times D_l}$, capturing compositional rules of muscular activations in response to hand motion. It is then integrated into the main time-channel encoder, guiding its feature construction with relational biases to facilitate learning about entities, relations, and rules for composing them.

Time-Channel-Wise Encoder. We revise the TDS block in previous research [76] as the core component of our time-channel-wise encoder, which is designed to decouple temporal dynamics from channel-wise feature interactions. All intermediate features are first combined by addition along the feature dimension: $\mathbf{F} \in \mathbb{R}^{T\times C} = [\mathbf{F}_p + \mathbf{F}_k + \mathbf{Z}_l]$, and then mapped to a higher-dimensional space by: $C = channels \times feature\_width$. Subsequently, the reshaped pseudo-image tensor $\mathbf{F}' \in \mathbb{R}^{T\times c\times w}$ allows us to treat the feature space as a 2D spatial domain and align the input shape with the following blocks. This vector is further passed through a module consisting of two sequential blocks, each comprising a 2D convolutional layer. The final embedded feature is computed as: $\hat{\mathbf{H}} = \mathcal{E}_{fea}(\mathbf{F}') \in \mathbb{R}^{T\times D_f}$. Compared to other convolutional encoders [77, 78], this decomposition facilitates temporal-channel separability, allowing the network to learn across time and channels independently.

D_p, D_k, D_l, and D_f refer to the output dimensions of the pose, keystroke, latent prior, and fused embedded features after their respective encoders. More details are introduced in Appendix C.1.

4.2.2 Auto-Regressive Decoder

Given that the target EMG is sequential, our decoder is designed as an auto-regressive structure. At each timestep t, the model predicts the hand muscle EMG $\hat{\mathbf{e}_\mathbf{h}}^t \in \mathbb{R}^M$ (M: muscles) based on the encoded features $\hat{\mathbf{h}}^t$ and the prediction $\hat{\mathbf{e}_\mathbf{h}}^{t-1}$ from the last timestep, which is thereby computed as: $\hat{\mathbf{E}_\mathbf{h}} = \mathcal{D}_{fea}(\{[\hat{\mathbf{h}}^t \| \hat{\mathbf{e}_\mathbf{h}}^{t-1}]\}_{t=1}^{T})$. Teacher forcing is not employed during training and the model relies solely on its previous predictions. A multilayer perceptron (MLP) is opted as the decoder backbone. Compared to long short-term memory (LSTM) architectures chosen in previous work [79], the MLP achieves competitive accuracy while offering faster convergence and lower computational overhead, making it preferable in the training and inference phases.

4.2.3 Precision-Structure Hybrid Loss

In addition to the standard Mean Squared Error (MSE) loss, we incorporate Optimal Transport (OT) loss to balance local accuracy with global structure preservation. MSE is widely used due to its effectiveness in penalizing numerical precision. However, training with it alone leads the model to converge to an over-smoothed solution that suppresses peaks, valleys, and muscle activation transitions. On the other hand, previous studies have demonstrated the effectiveness of OT distance in assessing muscle synergy similarity [80]. Therefore, Sinkhorn-based OT loss [81, 82] is specifically tailored for our training, which measures the transport cost required to morph the predicted sequence into the ground truth as a distributional matching problem. OT loss can jointly consider amplitude, temporal progression, and inter-muscle coordination as: $\mathcal{L}_{ot} = \mathcal{W}_\epsilon(\hat{\mathbf{E}}_\mathbf{h}, \mathbf{E}_\mathbf{h})$, where $\mathcal{W}_\epsilon$ denotes the entropic-regularized Wasserstein distance between two point clouds in $\mathbb{R}^M$ over T time steps. Finally, we combine them as: $\mathcal{L} = \lambda_{mse} \cdot \mathcal{L}_{mse} + \lambda_{ot} \cdot \mathcal{L}_{ot}$, to encourage low-level numerical precision and preserve higher-level physiological structures of EMG sequences.

5 Experiments

To evaluate the methodology rigorously, we conduct experiments along two axes. (1) Architectural evaluation: We benchmark our method against several baselines on the public emg2pose dataset [12] and the proposed PianoKPM Dataset using quantitative and qualitative metrics. (2) Held-out evaluation: We investigate the model's generalization to held-out-distribution scenarios, including exploring the effect of training dataset scale and quantifying the amount of unseen data required to sustain satisfying performance. To ensure fair comparisons and align with our hybrid loss design, we report results using Root Mean Squared Error (RMSE) and Optimal Transport Distance (OTD), which can evaluate the similarity between synergistic muscle activation patterns.

5.1 Architectural Evaluation

On the emg2pose and PianoKPM datasets, in-distribution experiments across all users and tasks are conducted to compare the proposed network architecture against several baselines. To further validate the usability of multiple modalities, we perform ablation studies, isolating the effect of each input on overall estimation accuracy. Given the absence of existing work on estimating small hand muscle EMG from poses in piano performance, we revise NeuroPose [83], originally designed for EMG-to-pose regression, to support multiple input modalities. Moreover, we modify CodeTalker [84], a widely-used cross-modal synthesis model, to better suit our task and serve as a reference to our method. See Appendix D.1 for baseline implementation details.

Table 2: Results for architectural evaluations. We report the mean and standard deviation across performances. Bold indicates the best method and lowest loss with different input settings.

Dataset	Input	Methods	RMSE	OTD
emg2pose [12]	Pose	NeuroPose [83]	.067 ± .032	.039 ± .036
		CodeTalker [84]	.096 ± .027	.075± .043
		PianoKPM (Ours)	**.055 ± .034**	**.030 ± .037**
PianoKPM	Pose	NeuroPose	.166 ± .066	.070 ± .081
		CodeTalker	.171 ± .063	.064 ± .071
		PianoKPM (Ours)	**.136 ± .062**	**.033 ± .060**
	Keystroke	NeuroPose	.196 ± .096	.118 ± .172
		CodeTalker	**.152 ± .075**	**.055 ± .092**
		PianoKPM (Ours)	.191 ± .090	.106 ± .144
	Pose, Keystroke	NeuroPose (multi-input)	.155 ± .068	.061 ± .077
		CodeTalker (cross-attention)	.143 ± .071	.040 ± .082
		PianoKPM (Ours)	**.134 ± .061**	**.031 ± .058**

Quantitative Results. The method comparison and modality ablation results are exhibited in Table 2. In general, PianoKPM Net achieves competitive accuracy when using human-centric video input alone (RMSE: 0.136, OTD: 0.033), and improves more when incorporated with tool-centric keystroke input (RMSE: 0.134, OTD: 0.031). While PianoKPM Net consistently outperforms other baselines

with either video alone or the fused two inputs, CodeTalker obtains the best performance with only keystroke input. This is because the transformer-based structure and self-attention mechanism capture long-sequence temporal dependencies. Our model integrates spatial postural semantics with temporal keypressing features to improve the accuracy of EMG prediction.

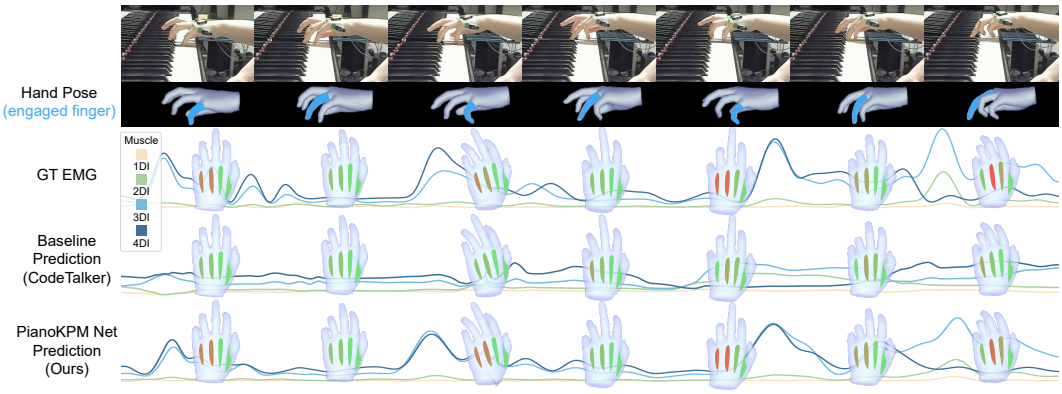

Figure 4: Qualitative results. Color gradients ranging from green (low activation) to red (high) are used to intuitively convey EMG magnitudes. See Figure 2 for the muscle localization diagram.

Qualitative Results. Since the EMG is normalized to a small range, the quantitative improvements in RMSE appear marginal. Qualitative visualizations are provided in Figure 4 to compare the intuitive results of the best-performing baseline and ours. We choose a task involving alternating little, ring, and middle finger movements. The line plots of 1DI-4DI muscle activities are displayed to compare the GT EMG, CodeTalker's, and PianoKPM Net's predictions. Our method yields amplitude and dynamics more closely aligned with GT EMG, highlighting its superior capability in modeling dexterous muscle activation patterns. See Appendix D.3 for more qualitative visualizations.

5.2 Held-Out Evaluation

Table 3: Dataset split and results for held-out evaluations. We separate two test sets to measure generalization to new users (*Cross-User*) and different performance tasks (*Cross-Task*).

	Train, Val			Test			**RMSE**	**OTD**
	Users	Tasks	Hours	Users	Tasks	Hours		
Cross-User	18	6	9.54	2	6	1.18	.209 ± .071	.095 ± .097
Cross-Task				18	1	1.72	.264 ± .094	.152 ± .188

Configurations and Overall Results. Robust and effective muscle EMG inference requires models that generalize across different individuals and tasks. Inspired by prior work [12], we similarly construct two held-out test sets intended to evaluate across these axes independently. Table 3 reports detailed statistics and results for each held-out setting. Specifically, *Cross-User* denotes unseen users during training but in-distribution performing tasks, while *Cross-Task* is the vice versa, involving unseen tasks but in-distribution users. More detailed held-out settings can be found in Appendix D.2. In the held-out evaluation, *Cross-User* (RMSE: 0.209, OTD: 0.095) and *Cross-Task* (RMSE: 0.264, OTD: 0.152) perform worse compared to the architectural evaluation (RMSE: 0.134, OTD: 0.031). Notably, the model exhibits lower generalization performance on unseen tasks compared to unseen users. It stems from the lower diversity of tasks (6) relative to users (18) in the training set, and the higher semantic complexity in the task space, particularly in structured domains like musical performance. Novel tasks often engage different muscles, activation timings, and fine-grained motor control variations that the model has not learned during training. We analyze the specific reasons in Appendix B.2 and visualize some held-out results in Appendix D.3.

Generalization Across Dataset Scale. Several experiments are conducted to investigate the effect of the training set scale on generalization Figure 5 (a) exhibits that adding more training users or tasks contributes to a reduction in out-of-distribution estimation error since the model is exposed to a broader range of individual anatomies and hand kinematics, mitigating overfitting. But *Cross-Task*

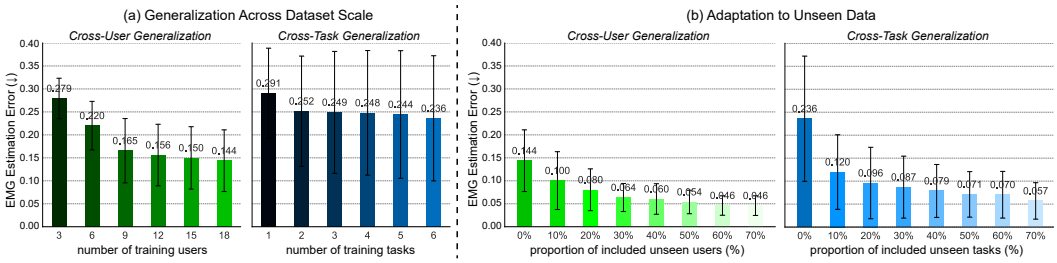

Figure 5: Held-out evaluation results. (a) Model generalization ability concerning the number of users or tasks included in training. (b) Model adaptation ability with different amounts of data from a new user or task. Distributions are across performance segments. The average EMG estimation errors (the sum of MSE and OTD) are indicated for reference, and error bars are standard deviations.

generalization is less pronounced than *Cross-User*, perhaps due to the lower feature overlap between the held-out task and in-distribution tasks. We include an additional held-out test in Appendix D.2.

Adaptation to Unseen Data. We further explore whether folding a small amount of unseen data into training can facilitate few-shot learning. Figure 5 (b) proves that including 30% of the new user data can reduce EMG estimation error by 55.6%, while 30% of new task data yields a greater reduction of 63.1%. Notably, including just 10% data from the unseen task domain already improves accuracy by approximately 50%, underscoring the critical role of kinematic diversity in enhancing EMG prediction generalization.

6 Limitations and Future Work

PianoKPM Dataset: The PianoKPM Dataset is the first large-scale, high-fidelity piano EMG dataset consisting of expert-level pianists' hand muscle activities. However, the lack of novices' and intermediates' performances restricts the dataset's generalization to a broader population, and synergistic wrist-forearm muscles remain unexplored in hand control. In addition, the current seven tasks are insufficient to capture the complexity of bimanual coordination required in advanced repertoires such as sonatas and fugues. Future investigations include more demographic diversity, muscular variations, and musical tasks. See Appendix B.3 for other dataset limitations.

PianoKPM Net: We propose PianoKPM Net, a promising approach for pose-to-EMG inference that outperforms several comparative baselines. However, Table 2 shows that using only keystrokes as input, CodeTalker achieves the best predictions, which may be attributed to the transformer's effectiveness at encoding long temporal dependencies, while PianoKPM is superior for multimodal learning. Future work may incorporate a transformer-based encoder or other sequence modeling frameworks to extract features from temporal keystroke sequences, for example, advanced diffusion-based techniques [85, 86], perception models [87, 88], and other multimodal networks [89]. On the other hand, our findings also highlight the persistent challenge of generalization across users and tasks. Future work could explore few-shot finetuning and transfer learning techniques, such as correlation alignment (CORAL) [90] for unsupervised domain adaptation and invariant risk minimization (IRM) [91] for domain generalization, to facilitate universally robust models capable of handling distributional variation. Other aspects for network limitations are introduced in Appendix C.3.

7 Conclusions

This paper presents PianoKPM Net, a framework for inferring dexterous hand muscle EMG during piano performance from accessible modalities. We conduct preliminary studies on a general EMG-to-pose dataset. Due to its unsatisfactory transfer results in piano-playing contexts, we introduce PianoKPM Dataset, a large-scale, high-fidelity, open-source multimodal EMG dataset, collected from expert pianists. Trained on this dataset, PianoKPM Net achieves promising estimation from pose input alone and further improves with keystroke integration. Comprehensive evaluations assess the model architectures and generalization capabilities. Together, PianoKPM Net and PianoKPM Dataset establish a foundation for low-cost access to internal physiological and myoelectric signals, advancing human augmentation and high-dimensional human-machine interaction.

Acknowledgments

This study is supported by JST CRONOS under Grant No. JPMJCS24N8, and JST ASPIRE under Grant No. JPMJAP2404.

References

[1] Ethan Eddy, Erik J Scheme, and Scott Bateman. A framework and call to action for the future development of emg-based input in hci. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, pages 1–23, New York, NY, USA, 2023. Association for Computing Machinery.

[2] T. Scott Saponas, Desney S. Tan, Dan Morris, Ravin Balakrishnan, Jim Turner, and James A. Landay. Enabling always-available input with muscle-computer interfaces. In *Proceedings of the 22nd Annual ACM Symposium on User Interface Software and Technology*, UIST '09, page 167–176, New York, NY, USA, 2009. Association for Computing Machinery.

[3] Momona Yamagami, Alexandra A Portnova-Fahreeva, Junhan Kong, Jacob O. Wobbrock, and Jennifer Mankoff. How do people with limited movement personalize upper-body gestures? considerations for the design of personalized and accessible gesture interfaces. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '23, pages 1–15, New York, NY, USA, 2023. Association for Computing Machinery.

[4] Ray Antonius and Hendra Tjahyadi. Electromyography gesture identification using cnn-rnn neural network for controlling quadcopters. *Journal of Physics: Conference Series*, 1858(1):012075, April 2021.

[5] Jakob Karolus, Annika Kilian, Thomas Kosch, Albrecht Schmidt, and Paweł W. Wozniak. Hit the thumb jack! using electromyography to augment the piano keyboard. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*, DIS '20, page 429–440, New York, NY, USA, 2020. Association for Computing Machinery.

[6] Ruofan Liu, Yichen Peng, Takanori Oku, Erwin Wu, Shinichi Furuya, and Hideki Koike. Pianokeystroke-emg: Piano hand muscle electromyography estimation from easily accessible piano keystroke. In *SIGGRAPH Asia 2024 Posters*, SA '24, pages 1–2, New York, NY, USA, 2024. Association for Computing Machinery.

[7] Andrea Manca, Andrea Cereatti, Lynn Bar-On, Alberto Botter, Ugo Della Croce, Marco Knaflitz, Nicola Maffiuletti, Davide Mazzoli, Andrea Merlo, Silvestro Roatta, Andrea Turolla, and Franca Deriu. A survey on the use and barriers of surface electromyography in neurorehabilitation. *Frontiers in Neurology*, 11:573616, October 2020.

[8] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

[9] Simon Alexanderson, Rajmund Nagy, Jonas Beskow, and Gustav Eje Henter. Listen, denoise, action! audio-driven motion synthesis with diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–20, July 2023.

[10] Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023.

[11] Jehan Yang, Maxwell Soh, Vivianna Lieu, Douglas J Weber, and Zackory Erickson. Emgbench: Benchmarking out-of-distribution generalization and adaptation for electromyography. *Advances in Neural Information Processing Systems*, 37:50313–50342, 2024.

[12] Sasha Salter, Richard Warren, Collin Schlager, Adrian Spurr, Shangchen Han, Rohin Bhasin, Yujun Cai, Peter Walkington, Anuoluwapo Bolarinwa, Robert J Wang, et al. emg2pose: A large and diverse benchmark for surface electromyographic hand pose estimation. *Advances in Neural Information Processing Systems*, 37:55703–55728, 2024.

[13] Patrick Kaifosh and Thomas R Reardon. A generic noninvasive neuromotor interface for human-computer interaction. *Nature*, pages 1–10, 2025.

[14] Qijun Gan, Song Wang, Shengtao Wu, and Jianke Zhu. Pianomotion10m: Dataset and benchmark for hand motion generation in piano performance. *arXiv preprint arXiv:2406.09326*, 2024.

[15] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, 36(6):1–17, November 2017.

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

[17] Luca Mesin, Roberto Merletti, and Alberto Rainoldi. Surface emg: The issue of electrode location. *Journal of electromyography and kinesiology : official journal of the International Society of Electrophysiological Kinesiology*, 19:719–26, October 2008.

[18] Yu Mike Chi, Tzyy-Ping Jung, and Gert Cauwenberghs. Dry-contact and noncontact biopotential electrodes: Methodological review. *IEEE Reviews in Biomedical Engineering*, 3:106–119, 2010.

[19] Evan Campbell, Angkoon Phinyomark, and Erik Scheme. Current trends and confounding factors in myoelectric control: Limb position and contraction intensity. *Sensors*, 20(6), 2020.

[20] Reza Nourbakhsh and Carl Kukulka. Relationship between muscle length and moment arm on emg activity of human triceps surae muscle. *Journal of electromyography and kinesiology : official journal of the International Society of Electrophysiological Kinesiology*, 14:263–73, May 2004.

[21] Faizan Haque, Mathieu Nancel, and Daniel Vogel. Myopoint: Pointing and clicking using forearm mounted electromyography and inertial motion sensors. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 3653–3656, New York, NY, USA, 2015. Association for Computing Machinery.

[22] T Scott Saponas, Desney S. Tan, Dan Morris, and Ravin Balakrishnan. Demonstrating the feasibility of using forearm electromyography for muscle-computer interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, page 515–524, New York, NY, USA, 2008. Association for Computing Machinery.

[23] Celal Savur and Ferat Sahin. American sign language recognition system by using surface emg signal. In *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 002872–002877, 2016.

[24] Qian Zhang, Dong Wang, Run Zhao, and Yinggang Yu. Myosign: enabling end-to-end sign language recognition with wearables. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, page 650–660, New York, NY, USA, 2019. Association for Computing Machinery.

[25] Hans Kainz, Willi Koller, Elias Wallnöfer, Till Bader, Gabriel Mindler, and Andreas Kranzl. A framework based on subject-specific musculoskeletal models and monte carlo simulations to personalize muscle coordination retraining. *Scientific Reports*, 14(1):3567, February 2024.

[26] Jan Pieter Clarys and Jan Cabri and. Electromyography and the study of sports movements: A review. *Journal of Sports Sciences*, 11(5):379–448, 1993. PMID: 8301704.

[27] Jessica Sehrt, Tim Wißmann, Jan Breitenbach, and Valentin Schwind. The effects of body location and biosignal feedback modality on performance and workload using electromyography in virtual reality. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, pages 1–16, New York, NY, USA, 2023. Association for Computing Machinery.

[28] Lisa Brown Jaloza. Inside facebook reality labs: Wrist-based interaction for the next computing platform. *Tech at Meta*, 18, 2021.

[29] Róisín Howard, Richard Conway, and Andrew Harrison. A survey of sensor devices: use in sports biomechanics. *Sports Biomechanics*, 15:450–461, May 2016.

[30] Jennifer L. Hicks, Thomas K. Uchida, Ajay Seth, Apoorva Rajagopal, and Scott L. Delp. Is my model good enough? best practices for verification and validation of musculoskeletal models and simulations of movement. *Journal of Biomechanical Engineering*, 137(2):020905, February 2015.

[31] Ahmet Erdemir, Scott McLean, Walter Herzog, and Antonie J. van den Bogert. Model-based estimation of muscle forces exerted during movements. *Clinical Biomechanics*, 22(2):131–154, 2007.

[32] Scott L. Delp, Frank C. Anderson, Allison S. Arnold, Peter Loan, Ayman Habib, Chand T. John, Eran Guendelman, and Darryl G. Thelen. Opensim: Open-source software to create and analyze dynamic simulations of movement. *IEEE Transactions on Biomedical Engineering*, 54(11):1940–1950, 2007.

[33] Ajay Seth, Jennifer L Hicks, Thomas K Uchida, Ayman Habib, Christopher L Dembia, James J Dunne, Carmichael F Ong, Matthew S DeMers, Apoorva Rajagopal, Matthew Millard, et al. Opensim: Simulating musculoskeletal dynamics and neuromuscular control to study human and animal movement. *PLoS Computational Biology*, 14:e1006223, July 2018.

[34] Thomas Geijtenbeek. Scone: Open source software for predictive simulation of biological motion. *Journal of Open Source Software*, 4(38):1421, June 2019.

[35] Steve A Maas, Benjamin J Ellis, Gerard A Ateshian, and Jeffrey A Weiss. Febio: Finite elements for biomechanics. *Journal of biomechanical engineering*, 134:011005, February 2012.

[36] Claudio Pizzolato, David G. Lloyd, Massimo Sartori, Elena Ceseracciu, Thor F. Besier, Benjamin J. Fregly, and Monica Reggiani. Ceinms: A toolbox to investigate the influence of different neural control solutions on the prediction of muscle excitation and joint moments during dynamic motor tasks. *Journal of Biomechanics*, 48(14):3929–3936, 2015.

[37] Antoine Falisse, Gil Serrancolí, Christopher L Dembia, Joris Gillis, Ilse Jonkers, and Friedl De Groote. Rapid predictive simulations with complex musculoskeletal models suggest that diverse healthy and pathological human gaits can emerge from similar control strategies. *Journal of The Royal Society Interface*, 16(157):20190402, 2019.

[38] Frank C Anderson and Marcus G Pandy. Dynamic optimization of human walking. *J. Biomech. Eng.*, 123(5):381–390, 2001.

[39] Scott D. Uhlrich, Antoine Falisse, Łukasz Kidziński, Julie Muccini, Michael Ko, Akshay S. Chaudhari, Jennifer L. Hicks, and Scott L. Delp. Opencap: 3d human movement dynamics from smartphone videos. *PLoS Computational Biology*, 19(10):e1011462, 2023.

[40] Mazen Al Borno, Johanna O'Day, Vanessa Ibarra, James Dunne, Ajay Seth, Ayman Habib, Carmichael Ong, Jennifer Hicks, Scott Uhlrich, and Scott Delp. Opensense: An open-source toolbox for inertial-measurement-unit-based measurement of lower extremity kinematics over long durations. *Journal of neuroengineering and rehabilitation*, 19(1):22, 2022.

[41] Timothy Hewett, Gregory Myer, Kevin Ford, Robert Heidt, Angelo Colosimo, Scott Mclean, Antonie van den Bogert, Mark Paterno, and Paul Succop. Biomechanical measures of neuromuscular control and valgus loading of the knee predict anterior cruciate ligament injury risk in female athletes a prospective study. *The American journal of sports medicine*, 33:492–501, May 2005.

[42] Łukasz Kidziński, Bryan Yang, Apoorva Rajagopal, Scott Delp, and Michael Schwartz. Deep neural networks enable quantitative movement analysis using single-camera videos. *Nature Communications*, 11(1):4054, August 2020.

[43] M.A. Boswell, S.D. Uhlrich, Ł. Kidziński, K. Thomas, J.A. Kolesar, G.E. Gold, G.S. Beaupre, and S.L. Delp. A neural network to predict the knee adduction moment in patients with osteoarthritis using anatomical landmarks obtainable from 2d video analysis. *Osteoarthritis and Cartilage*, 29(3):346–356, 2021.

[44] Eni Halilaj, Apoorva Rajagopal, Madalina Fiterau, Jennifer L. Hicks, Trevor J. Hastie, and Scott L. Delp. Machine learning in human movement biomechanics: Best practices, common pitfalls, and new opportunities. *Journal of Biomechanics*, 81:1–11, 2018.

[45] Ruofan Liu, Yichen Peng, Takanori Oku, Chen-Chieh Liao, Erwin Wu, Shinichi Furuya, and Hideki Koike. Piamuscle: Improving piano skill acquisition by cost-effectively estimating and visualizing activities of miniature hand muscles. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, pages 1–16, New York, NY, USA, 2025. Association for Computing Machinery.

[46] Giuseppe Boccignone, Vittorio Cuculo, Giuliano Grossi, Raffaella Lanzarotti, and Raffaella Migliaccio. Virtual emg via facial video analysis. In *Image Analysis and Processing - ICIAP 2017*, pages 197–207, Cham, 2017. Springer International Publishing.

[47] Vittorio Caggiano, Huawei Wang, Guillaume Durandau, Massimo Sartori, and Vikash Kumar. Myosuite–a contact-rich simulation suite for musculoskeletal motor control. *arXiv preprint arXiv:2205.13600*, 2022.

[48] Yu Du, Wenguang Jin, Wentao Wei, Yu Hu, and Weidong Geng. Surface emg-based inter-session gesture recognition enhanced by deep domain adaptation. *Sensors*, 17(3), 2017.

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

[50] Xinyu Jiang, Xiangyu Liu, Jiahao Fan, Xinming Ye, Chenyun Dai, Edward A. Clancy, Metin Akay, and Wei Chen. Open access dataset, toolbox and benchmark processing results of high-density surface electromyogram recordings. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:1035–1046, 2021.

[51] Manfredo Atzori, Arjan Gijsberts, Claudio Castellini, Barbara Caputo, Anne-Gabrielle Mittaz Hager, Simone Elsig, Giorgio Giatsidis, Franco Bassetto, and Henning Müller. Electromyography data for non-invasive naturally-controlled robotic hand prostheses. *Scientific Data*, 1, 2014.

[52] Manfredo Atzori, Arjan Gijsberts, Simone Heynen, Anne-Gabrielle Mittaz Hager, Olivier Deriaz, Patrick van der Smagt, Claudio Castellini, Barbara Caputo, and Henning Müller. Building the ninapro database: A resource for the biorobotics community. In *2012 4th IEEE RAS & EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob)*, pages 1258–1265, 2012.

[53] Manfredo Atzori, Arjan Gijsberts, Ilja Kuzborskij, Simone Elsig, Anne-Gabrielle Mittaz Hager, Olivier Deriaz, Claudio Castellini, Henning Müller, and Barbara Caputo. Characterization of a benchmark database for myoelectric movement classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 23(1):73–83, 2015.

[54] Stefano Pizzolato, Luca Tagliapietra, Matteo Cognolato, Monica Reggiani, Henning Müller, and Manfredo Atzori. Comparison of six electromyography acquisition setups on hand movement classification tasks. *PLoS one*, 12:e0186132, October 2017.

[55] Kyungjin Seo, Junghoon Seo, Hanseok Jeong, Sangpil Kim, and Sang Ho Yoon. Posture-informed muscular force learning for robust hand pressure estimation. *Advances in Neural Information Processing Systems*, 37:87831–87873, 2024.

[56] Erwin Wu, Takashi Matsumoto, Chen-Chieh Liao, Ruofan Liu, Hidetaka Katsuyama, Yuki Inaba, Noriko Hakamada, Yusuke Yamamoto, Yusuke Ishige, and Hideki Koike. Skitech: An alpine skiing and snowboarding dataset of 3d body pose, sole pressure, and electromyography. In *Proceedings of the 6th International Workshop on Multimedia Content Analysis in Sports*, MMSports '23, page 3–8, New York, NY, USA, 2023. Association for Computing Machinery.

[57] Kun Su, Xiulong Liu, and Eli Shlizerman. Audeo: audio generation for a silent performance video. *Advances in Neural Information Processing Systems*, 33:3325–3337, 2020.

[58] Amit Moryossef, Yanai Elazar, and Yoav Goldberg. At your fingertips: Extracting piano fingering instructions from videos. *arXiv preprint arXiv:2303.03745*, 2023.

[59] Hui Liang, Jin Wang, Qian Sun, Yong-Jin Liu, Junsong Yuan, Jun Luo, and Ying He. Barehanded music: real-time hand interaction for virtual piano. In *Proceedings of the 20th ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, I3D '16, page 87–94, New York, NY, USA, 2016. Association for Computing Machinery.

[60] Kyeongeun Seo, Hyeonjoong Cho, Daewoong Choi, Sangyub Lee, Jaekyu Lee, and Jaejin Ko. Twohandsmusic: Multitask learning-based egocentric piano-playing gesture recognition system for two hands. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 4614–4618, 2019.

[61] David Johnson, Daniela Damian, and George Tzanetakis. Detecting hand posture in piano playing using depth data. *Computer Music Journal*, 43(1):59–78, 2020.

[62] Erwin Wu, Hayato Nishioka, Shinichi Furuya, and Hideki Koike. Marker-removal networks to collect precise 3d hand data for rgb-based estimation and its application in piano. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2977–2986, January 2023.

[63] Ruocheng Wang, Pei Xu, Haochen Shi, Elizabeth Schumann, and C. Karen Liu. Fürelise: Capturing and physically synthesizing hand motion of piano performance. In *SIGGRAPH Asia 2024 Conference Papers*, SA '24, pages 1–11, New York, NY, USA, 2024. Association for Computing Machinery.

[64] Takanori Oku and Shinichi Furuya. Noncontact and high-precision sensing system for piano keys identified fingerprints of virtuosity. *Sensors*, 22(13), 2022.

[65] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

[66] S Conforto and T D'Alessio. Optimal rejection of artifacts in the processing of surface emg signals for movement analysis. In *Computer Methods in Biomechanics and Biomedical Engineering 2*, pages 799–805. CRC Press, 2020.

[67] S.M. Kay and S.L. Marple. Spectrum analysis—a modern perspective. *Proceedings of the IEEE*, 69(11):1380–1419, 1981.

[68] I.W. Selesnick and C.S. Burrus. Generalized digital butterworth filter design. *IEEE Transactions on Signal Processing*, 46(6):1688–1694, 1998.

[69] Thomas S Buchanan, David G Lloyd, Kurt Manal, and Thor F Besier. Neuromusculoskeletal modeling: estimation of muscle forces and joint moments and movements from measurements of neural command. *Journal of applied biomechanics*, 20(4):367–395, 2004.

[70] Rubana H. Chowdhury, Mamun B. I. Reaz, Mohd Alauddin Bin Mohd Ali, Ashrif A. A. Bakar, Kalaivani Chellappan, and Tae G. Chang. Surface electromyography signal processing and classification techniques. *Sensors*, 13(9):12431–12466, 2013.

[71] Kevin Zakka, Philipp Wu, Laura Smith, Nimrod Gileadi, Taylor Howell, Xue Bin Peng, Sumeet Singh, Yuval Tassa, Pete Florence, Andy Zeng, and Pieter Abbeel. Robopianist: Dexterous piano playing with deep reinforcement learning. *arXiv preprint arXiv:2304.04150*, 2023.

[72] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[73] Afroza Sultana, Farruk Ahmed, and Mohammad Alam. A systematic review on surface electromyography-based classification system for identifying hand and finger movements. *Healthcare Analytics*, 3:100126, November 2022.

[74] Christoph Amma, Thomas Krings, Jonas Böer, and Tanja Schultz. Advancing muscle-computer interfaces with high-density electromyography. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 929–938, New York, NY, USA, 2015. Association for Computing Machinery.

[75] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

[76] Awni Hannun, Ann Lee, Qiantong Xu, and Ronan Collobert. Sequence-to-sequence speech recognition with time-depth separable convolutions. *arXiv preprint arXiv:1904.02619*, 2019.

[77] Ronan Collobert, Christian Puhrsch, and Gabriel Synnaeve. Wav2letter: an end-to-end convnet-based speech recognition system. *arXiv preprint arXiv:1609.03193*, 2016.

[78] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. Convolutional sequence to sequence learning. In *International conference on machine learning*, pages 1243–1252. PMLR, 2017.

[79] Raul C Sîmpetru, Andreas Arkudas, Dominik I Braun, Marius Osswald, Daniela Souza de Oliveira, Bjoern Eskofier, Thomas M Kinfe, and Alessandro Del Vecchio. Sensing the full dynamics of the human hand with a neural interface and deep learning. *BioRxiv*, pages 2022–07, 2022.

[80] Arinobu Niijima and Shoichiro Takeda. Improving putting accuracy with electrical muscle stimulation feedback guided by muscle synergy analysis. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, pages 1–11, New York, NY, USA, 2025. Association for Computing Machinery.

[81] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transportation distances. *Advances in neural information processing systems*, 26, 2013.

[82] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[83] Yilin Liu, Shijia Zhang, and Mahanth Gowda. Neuropose: 3d hand pose tracking using emg wearables. In *Proceedings of the Web Conference 2021*, WWW '21, page 1471–1482, New York, NY, USA, 2021. Association for Computing Machinery.

[84] Jinbo Xing, Menghan Xia, Yuechen Zhang, Xiaodong Cun, Jue Wang, and Tien-Tsin Wong. Codetalker: Speech-driven 3d facial animation with discrete motion prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12780–12790, 2023.

[85] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020.

[86] Kashif Rasul, Calvin Seward, Ingmar Schuster, and Roland Vollgraf. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In *International conference on machine learning*, pages 8857–8868. PMLR, 2021.

[87] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021.

[88] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*, 2021.

[89] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403, November 2024.

[90] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450. Springer, 2016.

[91] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[92] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in neural information processing systems*, 35:38571–38584, 2022.

[93] Shinichi Furuya and Sayuri Yokota. Temporal exploration in sequential movements shapes efficient neuromuscular control. *Journal of neurophysiology*, 120(1):196–210, April 2018.

[94] Sara A Winges, Shinichi Furuya, Nathaniel J Faber, and Martha Flanders. Patterns of muscle activity for digital coarticulation. *Journal of neurophysiology*, 110(1):230–242, April 2013.

[95] Leonardo G Cohen and Mark Hallett. Hand cramps: clinical features and electromyographic patterns in a focal dystonia. *Neurology*, 38(7):1005–1005, 1988.

[96] Zengyi Qin, Zhenyu Jiang, Jiansheng Chen, Chunhua Hu, and Yu Ma. semg-based tremor severity evaluation for parkinson's disease using a light-weight cnn. *IEEE Signal Processing Letters*, 26(4):637–641, 2019.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Section 1 introduces three main claims, which are respectively achieved through (1) the code published in an anonymous GitHub repository `https://github.com/ruofanliu0129/PianoKPMNet.git`, allowing others to reproduce PianoKPM Net for EMG inference, (2) the multimodal piano performance dataset PianoKPM Dataset also linked on the same repo, and (3) the comprehensive evaluations of the model architecture and generalization articulated in Section 5.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Some of the limitations are discussed in Section 6. Respectively, Appendix B.3 and Appendix C.3 further supplement and analyze the dataset-related and network-related limitations, along with their underlying causes.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This paper does not include theoretical results.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 4.2 and Appendix C fully describe the architecture of PianoKPM Net. Check `https://github.com/ruofanliu0129/PianoKPMNet.git` for more reproduction details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility.

In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide an anonymous Github repo at `https://github.com/ruofanliu0129/PianoKPMNet.git`, containing available code and example dataset.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Dataset splits and reasons are specified in Table 3 and Appendix D.2, also included in configuration JSON files on Github repo. Implementation details are quantified in Appendix C.2, including hyperparameters and computational resources.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Mean and standard deviation across performance segments are reported in Table 2 and Table 3. Error bars regarding the standard deviations are displayed in Figure 5 and Figure 9.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix C.2 includes the compute resources information.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We reviewed the NeurIPS Code of Ethics and all data and code were anonymized.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: Appendix B.3 and Appendix E discuss potential negative and positive ethical, societal, and privacy impacts.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We expect that the piano EMG inference framework and the corresponding dataset have a low risk for misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We check all asset licenses and cite the original papers.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The new PianoKPM Dataset is collected and used with participants' consent and the code is created by the authors. Details of the data collection process are mentioned in Appendix B.1.3.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: Appendix B.1.3, Figure 6, and Figure 7 introduce the collection protocol, compensation, instruction slides, and sheet music given to participants.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

Justification: As mentioned in Appendix B.1.3, the study protocol is approved by a local IRB.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: This research does not involve LLMs as important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

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. [65].

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-to-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/ruofanliu0129/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

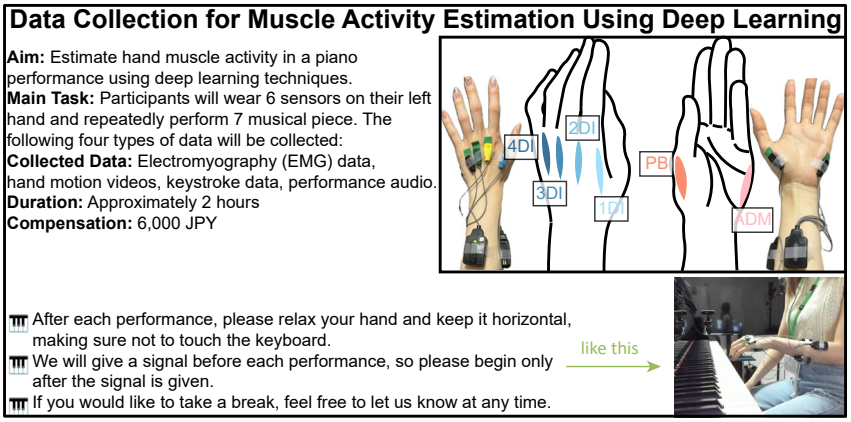

Figure 6: The introduction slide presented to participants before the study. The upper-right corner illustrates the six specified hand muscles and the two types of EMG sensor setups.

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 $\leq$ 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 [64], 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).

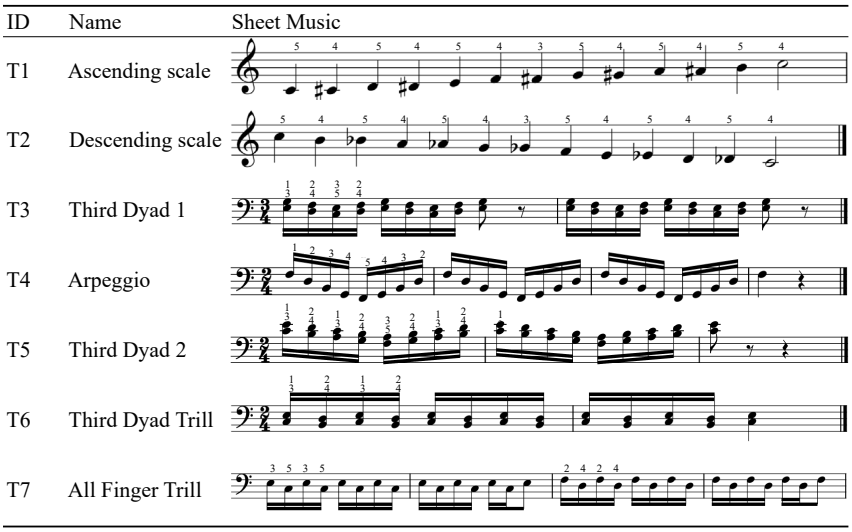

Figure 7: Task descriptions. The name, sheet music, and fingering patterns are provided.

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.

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 $\mathbf{P}_{3D} \in \mathbb{R}^{21\times3}$ and mesh vertices $\mathbf{V}_{3D} \in \mathbb{R}^{778\times3}$ 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 [92] 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 $\mathbf{P}_{2D} \in \mathbb{R}^{V\times J\times2}$ (V: views, J: joints) and customize a fusion pose encoder to extract semantically meaningful 3D pose representations. To address the limitations of single-frame posture estimation algorithms based on visual input, we briefly describe the pipeline for refining and post-processing the extracted hand pose data in Section 4.1.1. Specifically, abnormal frames containing artifacts can be categorized into missing frames, where ViTPose [92] fails to detect human body key points, and invalid frames, where HaMeR [16] does not capture the hand joint positions. During error detection, we first compute the inter-frame differences, their global mean μ, and standard deviation σ , based on which a threshold is predefined as $TS = \mu + 2.5 \cdot \sigma$. During traversal, the current frame and the adjacent valid frame are compared to obtain the corresponding z-score $z_i = (d_i^{valid} - \mu)/\sigma$. Notably, when abnormal frames occur consecutively, the growing temporal gap Δt from the last valid frame can inflate the z-score. Therefore, we apply a dynamic threshold that adapts to the length of this gap by $ts_i = TS \cdot (1 + 0.1\Delta t)$, and a frame is flagged as abnormal when $z_i > ts_i$. In addition, we traverse the sequence in forward and backward directions, and only frames identified as abnormal in their strict intersection are considered anomalous and subsequently corrected via interpolation.

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.

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/ruofanliu0129/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.

B.2 Dataset Insights

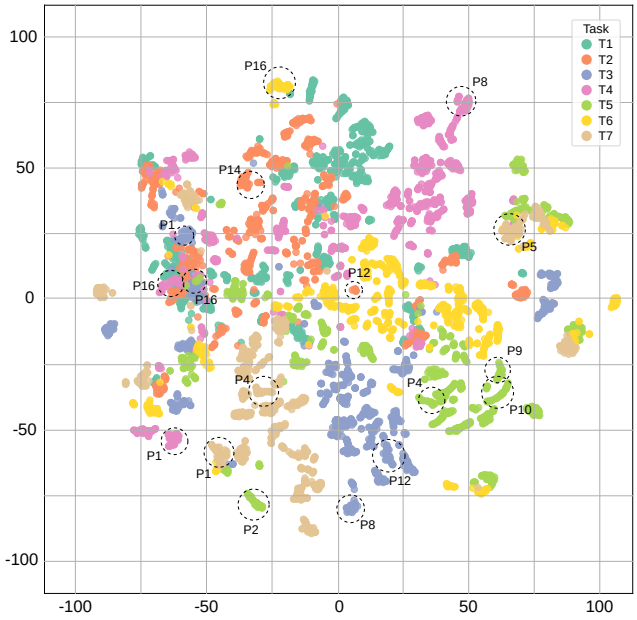

Figure 8: Visualization for EMG features after t-SNE dimensionality reduction. Different colors distinguish different tasks, and dashed circles highlight some example users' EMG data.

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 [72], 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 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 [76]. 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. They map the input channels to a higher-dimensional space, which 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 a 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 [81], 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:

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

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 [89]. 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 [83]: 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 [85], 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 [84]: 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.

(a) *Cross-User-Task Configuration*

	Users	Tasks	Hours
Train, Val	18	6	9.54
Test	Users	Tasks	Hours
	2	1	0.20

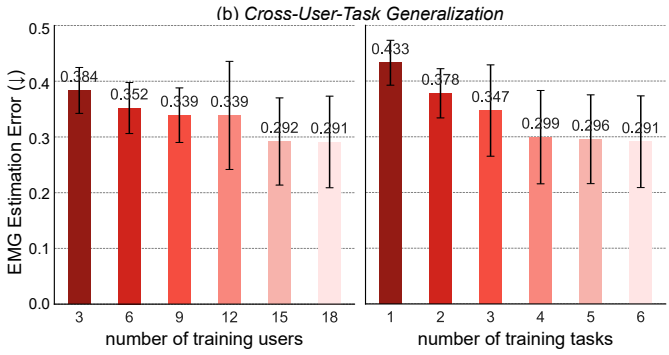

Figure 9: *Cross-User-Task* configuration and results. (a) The test set is split to include both unseen users and task. (b) Model generalization ability across training dataset scale. The bar charts take the same format as in Figure 5.

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

D.3 More Qualitative Results

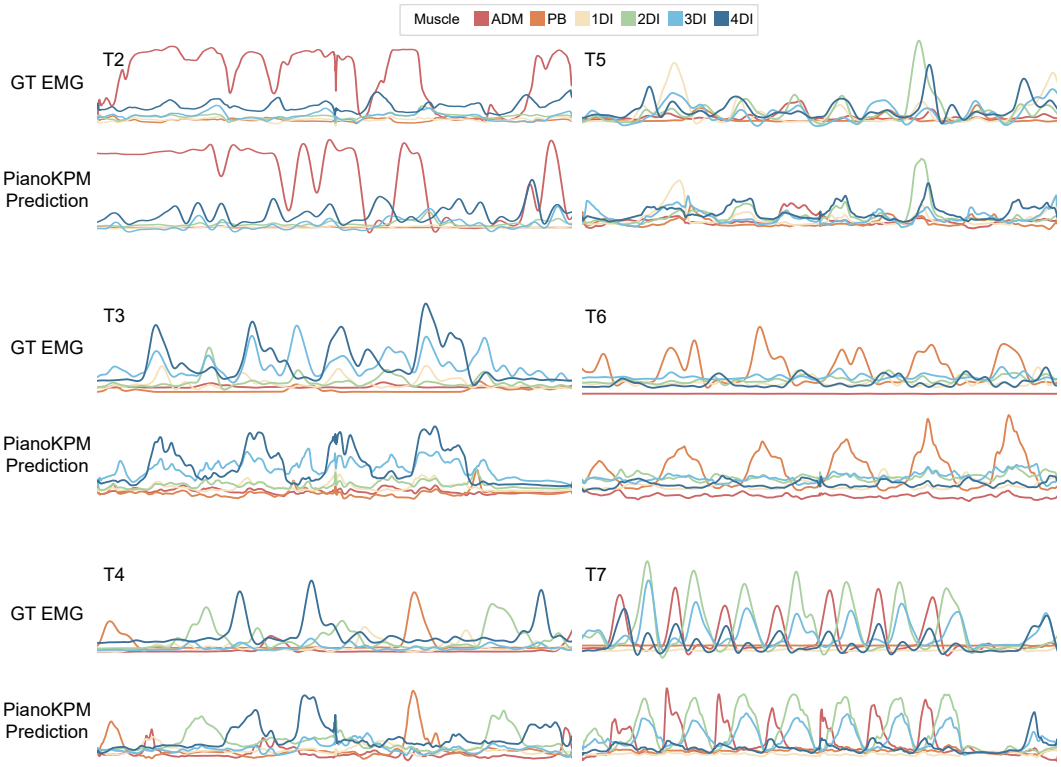

Figure 10: Visualization results for tasks T2 to T7 from several users. Six hand muscle EMGs are represented as line plots. For each task, the first row shows the GT EMG as a reference, while the second row presents the predictions given by PianoKPM Net.

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

Held-Out Evaluation. As shown in Figure 11, the predictions in the left half (*Cross-User* setting) are relatively more accurate. While the predicted EMG signals may not perfectly align with the GT in magnitude, the structures and activation trends of muscles are partially preserved, indicating that the model can capture user-invariant neuromuscular dynamics to some extent. In contrast, the right half (*Cross-Task* setting) shows more discrepancies. For example, in the top-right example, PianoKPM erroneously outputs additional activation for Muscle ADM (red); similarly, in the bottom-right example, Muscle PB (orange) activation is missing. These errors may be due to out-of-distribution tasks exhibiting kinematic patterns not included during training, which leads to incorrect model

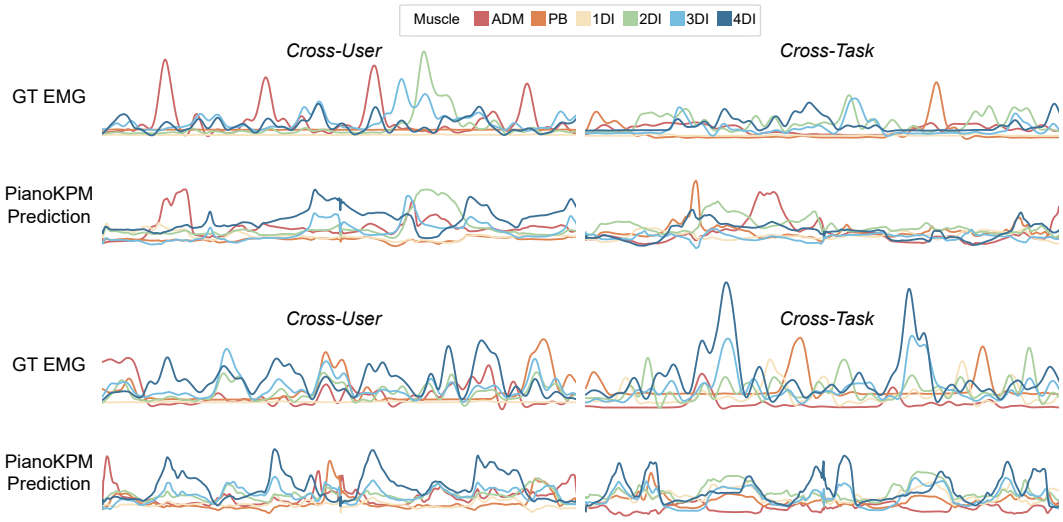

Figure 11: Visualization results under *Cross-User* (Left) and *Cross-Task* (Right) settings.

fitting. On the other hand, variations across users appear to be partially captured or characterized by our advanced network architecture.

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 have potential applications 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.

To further clarify the broader impact and prospects of our work, here we enumerate additional downstream applications. In the domain of piano performance, prior research reveals the neuromuscular control and coarticulation of muscle activity patterns in professional pianists with the use of EMG [93, 94]. Nevertheless, placing numerous sensors on the hand impedes skillful manual movements by expert pianists, which emphasizes the importance of estimating hand muscular activities in a non-contact manner. On the other hand, EMG-based visualization systems are proven to support motor learning in piano training, helping users perceive subtle muscle activations [45]. EMG feedback can boost users' improvisational ability by enhancing their motor control strategies [5]. EMG-based analyses of the finger muscular activities identify aberrant neuromuscular control of patients with neurological disorders such as focal hand dystonia [95]. In general domains, EMG signals serve as key input for prosthesis control, with recent works calling for the further integration of EMG-based input into ubiquitous interactions [1]. EMG is increasingly used in classifying neuromuscular disorders and predicting disease severity [96]. Prior work leverages EMG for analyzing and enhancing motor tasks, such as gait retraining [25].

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.