Supplementary Material

A. Qualitative Examples: Motion-to-Muscle

In the Motion-to-Muscle folder of the supplemental material, we show a qualitative result per exercise, where we predict the muscle activation (sEMG) from motion (3D skeleton). Each video is slowed down by a factor of 3 in order to better visualize the results. The axis in the animated plots are scaled to the min and max of the predicted and ground truth values. The values themselves are normalized per muscle per subject.

B. Qualitative Examples: Muscle-to-Motion

In the Muscle-to-Motion folder of the supplemental material, we show a qualitative result per exercise, where we reconstruct the 3D skeleton from sEMG signals, and visualize it by re-projecting the 3D skeleton onto the ground-truth video frames, utilizing the provided bounding box. The red skeleton is ground-truth, and the green skeleton is our predicted reconstruction. We additionally share results from the same subject performing the same exercise, this time with our 2D decoder, which directly predicts 2D coordinates with respect to the frame.

C. Qualitative Examples of Editing

In the Editing folder, we showcase two types of editing. While in the main paper we only had space to discuss editing via scaling, we also introduce editing via the complete replacement of one or two muscles. We call this second type "Muscle Stitching Editing". We note that for every single one of these examples, the stitching of muscle activations, and corresponding motions, do not exist in the training set. We believe this exemplifies our decoder's ability to generalize decently well. The second type of editing that we share qualitative examples of is "Scaled Editing", whose method is discussed in the main paper.

Scaled Editing. For each video, the top left corner of the video illustrates the ground truth input skeleton in red. The left bottom corner shows the predicted muscle activation for the dorsal muscles, visualized on an SMPL [2] mesh. The right bottom corner shows our edited predicted muscle activation for the same dorsal muscles. Finally, the top right corner visualizes the "recommended" motion, reconstructed from the edited predicted muscle activation, in cyan, on top of the ground truth.

For Example A in the "Scaled Editing" folder, we scale the quads and the hamstrings by a factor of 3. For Example B, we scale the biceps by a factor of 2 and the laterals by a factor of 5. For example C, we scale the hamstrings by a factor of $\frac{1}{5}$. Finally, for Example D, we scale the laterals by a factor of 3.

Muscle Stitching Editing. For each video, the organization of the composed video follows that of the scaled editing videos, as previously explained in 'Scaled Editing'.

For Example A in the "Muscle Stitching Editing" folder, we replace the predicted muscle activation for the quads and hamstrings for a hook punch exercise with the predicted muscle activation for a kick-back exercise.

For Example B, we do the converse, we replace the predicted muscle activation for the biceps and laterals in a kick-back exercise with the biceps and laterals from a hook punch exercise.

For Example C, we replace the predicted muscle activation for the quads and hamstrings for a front punch exercise with the predicted muscle activation for a side lunge exercise.

For Example D, we do the converse, we replace the predicted muscle activation for the biceps and laterals in a side-lunge exercise with the biceps and laterals from a front punch exercise.

D. Architecture Implementation Details

The majority of the architectural details are included in the main paper. The remaining details are mainly with respect to conditioning in the paragraph below. Otherwise, it should be noted that after the convolutional layer, we implement a positional encoding layer. After computing the positional encoding p, where $p \in R^{T \times D}$, we add the positional embedding to the features produced after our convolutional layer.

E. Conditioning Implementation Details

For the conditioning versions of our model, we modify the architecture as follows. The first convolutional layer for both the encoder and decoder have 126 channels. The output is a sequence of embeddings of length T, with each embedding $d_i \in R^{T \times D}$, where D=126. Therefore, each em-

bedding d_i has a channel dimension of 126. We concatenate a unique tensor $y \in \mathbb{R}^{2\times T}$ along the channel axis.

F. Electrode/Sensor Placement

We did our best to position the sensor placement in the vertical middle of each of the following muscles per person: the left and right biceps brachii (biceps), the left and right latissimus dorsi (laterals), both quadriceps (quads), and both biceps femoris (hamstrings). As per standard practice [1], to optimize for low signal noise, we shaved the areas for each subject, wiped the area with alcohol, and patted the area down with paper towel until completely dry, prior to electrode placement.

G. sEMG-Video Alignement (Further Details)

We leverage the timestamps on both video and the sEMG data from the Bluetooth sensors to align the two modalities.

There are two cases. In the first case, if the sEMG recording begins before the video, we simply remove all data prior to the video's origin time, and the first sEMG value that is larger or equal to the starting time gets rounded down to the nearest 10 millisecond interval. The rest of the sEMG data is timestamped by adding 10ms to each data point. In the second case, if the sEMG recording begins after the video, we remove the frames up until we reach a frame for which the timestamp is larger than the starting sEMG value. Then we repeat the processing steps for the first case. To summarize, the maximum alignment error between the video and the sEMG data is less than 10ms.

H. Dataset Statistics

We share two plots to visualize the statistics of the MIA dataset. In Figure 1, we illustrate the box plot of the maximum sEMG value per muscle, across subjects. We notice that there is a great range in maximum sEMG values per muscle across subjects. This is expected, as people vary largely in their morphology, and slight variations in sensor placement can make the range of sEMG data vary largely.

In Figure 2, we perform visual clustering. To do this, given a set of exercises S_i, where i corresponds to a given exercise, we perform nearest neighbor on the entire dataset S_j, $\forall_{j\neq i}$. If the retrieved nearest neighbor is an exercise k, we increment the matrix element $m_{(i,k)}$. We chose an ordering of rows and columns that maximizes self-similarity across the diagonal, in order to visualize which exercises are most similar to one another through clusters.

I. Subject OOD

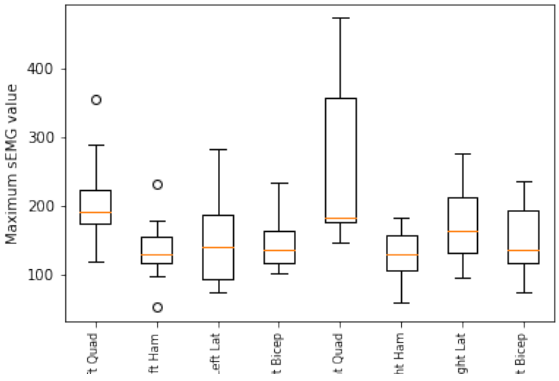

Figure 1: **Box Plot of Maximum sEMG Value per Muscle.** We visualize the maximum sEMG value per muscle, across subjects. We notice that there is a significant range of values per muscle across subjects.

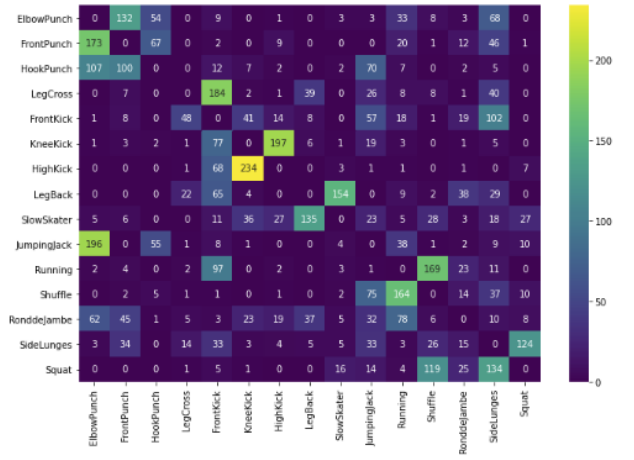

Figure 2: **Nearest Neighbor.** We perform visual clustering by taking advantage of the exercise labels in the dataset as well. We notice that the kicks and punches are clustered together, as well as the aerobic exercises (running and shuffle), and the strength training exercises (side lunges and squats).

	In-Distribution				Out-of-Distribution		
Subject	NN.	C-NN	Ours	C-Ours	NN	Ours	C-Ours
Mean Subject	12.7	12.3	9.9	**9.8**	23.4	15.8	**15.6**

Table 1: **RMSE per Subject for the Encoder.**

References

[1] Hermie J Hermens, Bart Freriks, Catherine Disselhorst-Klug, and Günter Rau. Development of recommendations for semg sensors and sensor placement procedures. *Journal of electromyography and Kinesiology*, 10(5):361–374, 2000.

[2] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard

Subject	In-Distribution				Out-of-Distribution		
	NN.	C-NN	Ours	C-Ours	NN	Ours	C-Ours
Subject 1	12.3	11.7	**9.3**	9.6	22.5	15.8	**15.6**
Subject 2	12.8	12.2	9.3	**9.3**	23.9	18.6	**18.2**
Subject 3	8.4	8.1	6.4	**6.4**	20.6	**12.7**	13.8
Subject 4	10.4	10.7	8.6	**8.6**	21.5	**17.3**	17.6
Subject 5	20.9	20.1	16.0	**16.0**	29.3	**26.7**	27.2
Subject 6	10.9	10.4	**8.7**	8.8	25.5	18.2	**18.1**
Subject 7	10.2	10.3	**7.6**	7.7	21.9	**16.3**	16.4
Subject 8	11.3	11.0	**8.8**	8.9	20.6	16.8	**16.7**
Subject 9	11.9	11.3	8.9	**8.9**	21.2	17.3	**17.1**
Subject 10	18.6	17.4	**14.0**	14.3	27.4	**23.7**	23.9

Table 2: **RMSE per Subject for the Encoder.**

Subject	In-Distribution				Out-of-Distribution		
	NN.	C-NN	Ours	C-Ours	NN	Ours	C-Ours
Subject 1	0.065	0.061	**0.042**	0.043	0.107	0.080	**0.080**
Subject 2	0.062	0.054	**0.038**	0.040	0.113	0.092	**0.089**
Subject 3	0.061	0.060	0.038	**0.038**	0.112	**0.090**	0.092
Subject 4	0.053	0.052	0.038	**0.038**	0.115	0.098	**0.095**
Subject 5	0.075	0.074	0.068	**0.068**	0.109	0.078	**0.075**
Subject 6	0.055	0.050	0.039	**0.039**	0.104	0.083	**0.081**
Subject 7	0.069	0.062	0.048	**0.047**	0.112	0.087	**0.086**
Subject 8	0.054	0.048	**0.036**	0.037	0.107	0.085	**0.083**
Subject 9	0.052	0.046	**0.033**	0.035	0.103	0.085	**0.083**
Subject 10	0.073	0.064	**0.045**	0.046	0.115	0.093	**0.092**

Table 3: **RMSE per Exercise for the Decoder**.

Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015.