Muscles in Action

Mia Chiquier and Carl Vondrick
Columbia University
{mia.chiquier,vondrick}@cs.columbia.edu

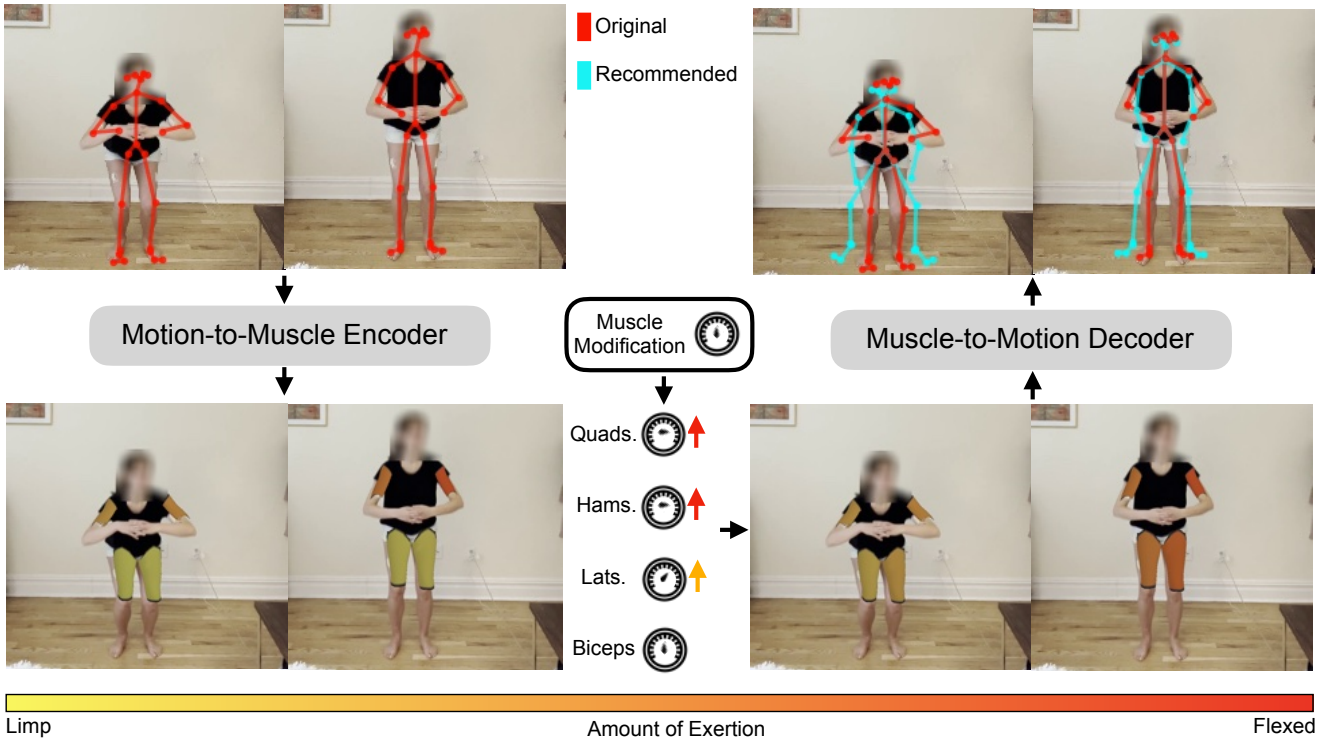

Figure 1: **Bidirectional Mapping between Muscles and Motion**. Visible human motion is created by, and constrained by, our muscles. We learn to predict which muscle groups a person uses during motion (left column), as well as to reconstruct motion from muscle activation (right column). We illustrate how these two mappings can be combined to recommend new motions, similar to the input muscle, this time subject to particular muscle group goals (bottom row).

Abstract

Human motion is created by, and constrained by, our muscles. We take a first step at building computer vision methods that represent the internal muscle activity that causes motion. We present a new dataset, Muscles in Action (MIA), to learn to incorporate muscle activity into human motion representations. The dataset consists of 12.5 hours of synchronized video and surface electromyography (sEMG) data of 10 subjects performing various exercises. Using this dataset, we learn a bidirectional representation that predicts muscle activation from video, and conversely, reconstructs motion from muscle activation. We evaluate our model on in-distribution subjects and exercises, as well as on out-of-distribution subjects and exercises. We demonstrate how advances in modeling both modalities jointly can serve as conditioning for muscularly consistent motion generation. Putting muscles into computer vision systems will enable richer models of virtual humans, with applications in sports, fitness, and AR/VR.

1. Introduction

The vision community has made great progress in modelling and analyzing human motion from video via tasks such as pose estimation [33, 25, 62, 45, 7, 23, 7], action recognition [47, 56, 27, 26, 59], motion transfer [8, 1, 31, 58] and more. However, motion understanding goes beyond

the surface. Human motion is created by and constrained by muscles. Every action is a product of our brain sending electric signals to our nerves, which contract our muscles, in turn moving our joints. Although this process occurs within us, most of us turn to physical therapists and sports instructors for guidance on how to improve our motions to target or avoid particular muscle groups.

In this paper, we take a first step towards building computer vision methods that represent the internal muscle activity that causes human motion. We present a system that, given a video of a person performing an action, learns to infer what muscles a person used. Walking is controlled falling, and any physical motion is a balance between muscle forces and gravity. This interplay leads to an inherent asymmetry: different muscles are engaged in the downward portion of a squat, for instance, than in the upward portion.

Our goal is to learn the complex relationships between physical forces by analyzing synchronized video and muscle activation data. We achieve this by developing a system that can predict muscle activity from motion, and vice versa. One application of this bidirectional system is generating new motions that are similar to an existing motion, while also adhering to specific muscle recruitment targets, illustrated in Figure 1.

The typical method of measuring muscle activity is through the use of electromyography sensors, which exist in an invasive form, as well as an non-invasive form, called surface electromyography (sEMG). We collected a new dataset, which we will release, that consists of over twelve hours of synchronized single-view video and sEMG signals of eight muscles for ten subjects performing fifteen different physical activities. By using commodity cameras and inexpensive sEMG sensors, we make the problem practical and easy for others to build on. The eight muscles recorded are the left and right biceps brachii (biceps), the left and right latissimus dorsi (laterals), both quadriceps (quads), and both biceps femoris (hamstrings), denoted in Figure 2.

The primary contribution of this paper is a framework for modeling the association between human motion and internal muscle activity in video, and the rest of the paper will explain this contribution in detail. In section 2, we briefly review related work in human activity analysis, conditional motion generation, multi-modal learning, electromyography, and physics-grounded human motion generation. In section 3, we describe our multimodal dataset in detail and analyze its characteristics. In section 4, we present a method to learn a bidirectional representation between the visual and muscle modalities. Section 5 shows experiments on both in-distribution experiments and subjects, as well as out-of-distribution experiments and subjects. In section 6, we showcase a demo application for learning the bidirectional representation between modalities. By releasing our datasets and models publicly, we hope this paper will spur additional work that models the rich internal structure that drives human activity in video.

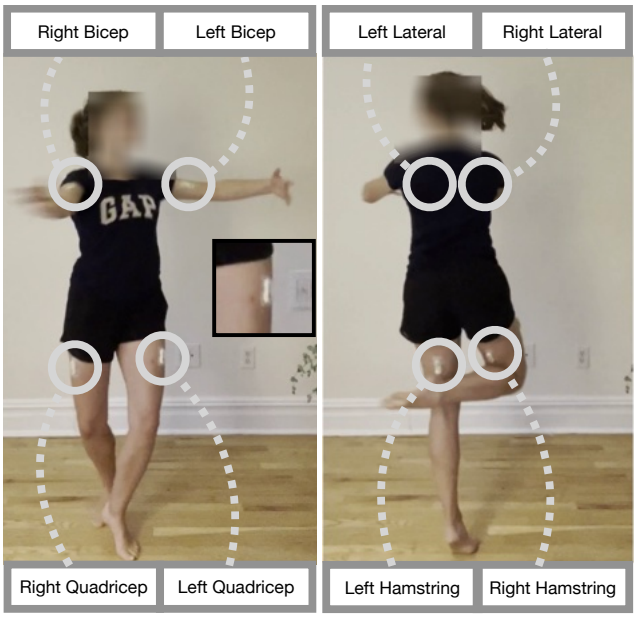

Figure 2: **Sensor Placement.** We illustrate the placement of our 8 sEMG sensors on a subject. We label the 8 measured muscles.

2. Related Work

Human Motion Prediction from Video. The field of computer vision has seen tremendous progress in inferring information about human motion from monocular video. One of the tasks is to regress human pose from video by regressing skeleton key-points and meshes [33, 25, 62, 45, 7, 23, 7]. A related task is action segmentation [28]. Other tasks that span from the pose estimation results include human motion transfer [8, 1, 31, 58] and even pose correction to make a given pose anatomically-correct [20, 48]. Closely related, Park et al. predict the 3D gravity direction from a moving first-person view using inverse dynamics [42].

Conditional Human Motion Generation. The field of conditional human motion generation is well-established, with a diverse set of conditioning mechanisms. There are works that condition based on past frames and/or future target frame(s) [10, 36, 18, 14, 15, 22, 16, 9, 54]. Others condition based on other modalities such as spatial trajectory [19], action class [13, 43, 54], natural language text [3, 44, 54], as well as audio [29, 5]. In this work, our motion generation is conditioned on an input motion, as well as muscle activity constraints.

Multi-Modal Representations. Multi-modal learning with video is a long-standing problem in computer vision. Some works predict other modalities from video, such as

Figure 3: **The MIA Dataset.** We illustrate two canonical frames from our dataset for each of the 15 exercises.

sound [40, 11] by using the natural temporal correspondence between video and sound, as well as video captioning [60]. Beyond prediction, multi-modal learning has been shown to improve video representations, for example from sound [41, 4], and language [60, 52]. The converse has also been explored - leveraging large datasets of unlabelled video has improved representations of other modalities [6, 4, 35, 12]. We are interested in predicting an entirely different modality from monocular video, muscle activation, as well as reconstructing motion from muscle activity.

Electromyography. To measure muscle activation, we use Surface Electromyography (sEMG) sensors, which are attached to electrodes placed on human skin to measure the electrical activity of muscle tissue. Reconstructing parts of human pose from sEMG data is an established task, but only for either arms [32, 2, 38, 46] specifically, or legs [65] separately. In the forward direction, previous work has tackled predicting muscle activation, however, the input modality has not been video. Some works use torque or surface force measurements as input [50, 51, 30], while others use goniometers to track pose [53] or motion capture tracking systems [21, 61, 37]. Our work seeks to predict muscle activation with no additional hardware at test-time besides video. Additionally, certain works predict muscle activation directly from 3D point clouds collected by depth cameras [39, 49]. However, these works rely on seeing the skin deformation on the human subject to infer muscle activity. We infer muscle activity from motion priors, not the visible increase in muscle size. This allows our model to work for clothed humans, or humans exercising at a distance.

Modeling Human Motion with Physics. Recent work has focused on generating motion that respects the physics of motion via physics simulations of human motion dynamics [63, 34, 64]. However, the simulated humanoids are constructed with assumptions about the physics of human motion. Additionally, sEMG studies have shown that for a given motion, different people vary in how they recruit muscle groups to execute that motion [55]. Single humanoid simulations will not capture this diversity in real humans' motion dynamics.

3. The Muscles in Action (MIA) Dataset

To explore the mapping between visual motion and muscle activity, we collected a dataset of synchronized video and sEMG signals. Our dataset contains 15 different exercises, which each of the 10 subjects perform.

3.1. Data Collection

Our dataset consists of 12.5 hours of synchronized video and sEMG signals, for eight muscles. These eight muscles include the left and right biceps brachii (biceps), the left and right latissimus dorsi (laterals), both quadriceps (quads), and both biceps femoris (hamstrings). The collected sEMG values correspond to the neuromuscular junction's total bioelectric energy.

The dataset consists of 15 exercises shown in Figure 6. Each subject performed each exercise for 5 minutes, and we asked them to vary the execution's speed, effort, and orientation. There are a total of 10 subjects in the dataset, 5 of which are females and 5 of which are males. We collected 75 minutes of data for each subject, totalling 12.5 hours of data. The subjects varied in body weight

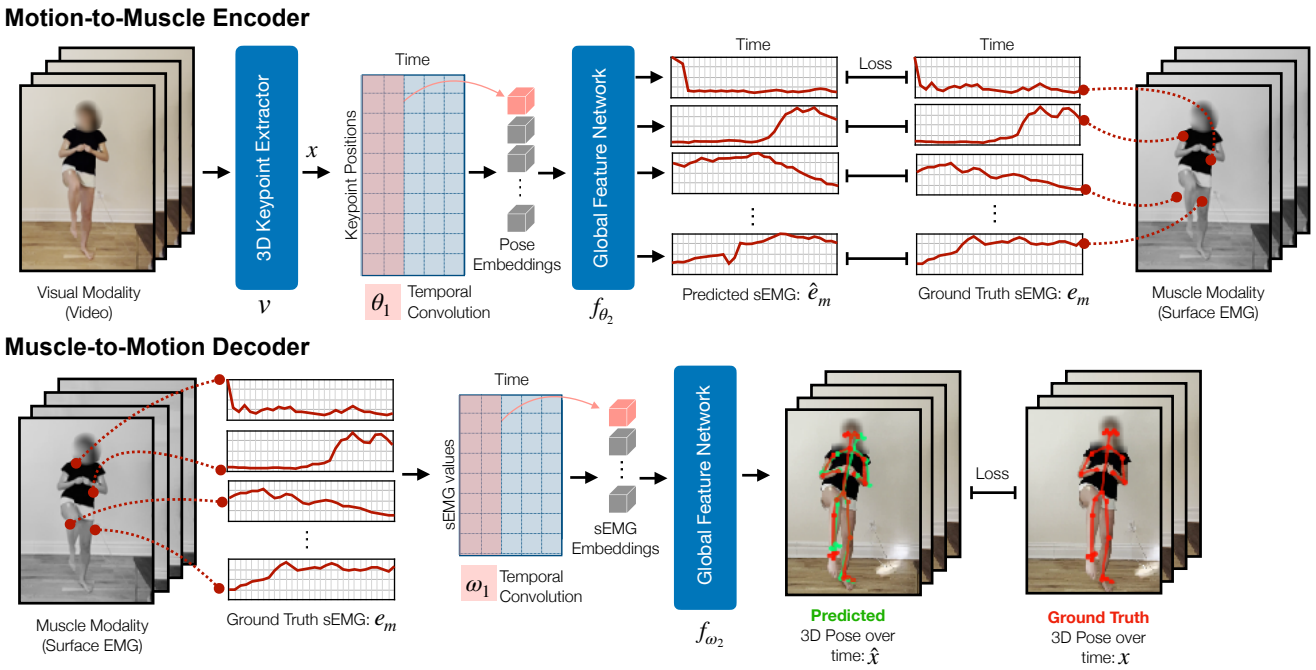

Figure 4: **Encoder and Decoder Architectures.** We illustrate the architectures for our Motion-to-Muscle encoder and our Muscle-to-Motion decoder.

and muscle. To collect the sEMG data, we used eight M40 Muscle Sense bluetooth wireless EMG sensors from ANR Corp. To collect the video, we used a standard iPhone 10 camera. Please see the supplementary for details on the electrode and sensor placement method.

3.2. Data Preprocessing

The sEMG data's sample rate is 10fps, and the each data point from the sensor comes with a timestamp. The iPhone video records at 29.97 fps, and also has a time-stamp. We resample the video to match the frame rate of the sEMG data, and use these time stamps to align the muscle and visual modalities. We explain our exact methodology for this in the supplemental material.

Once the sEMG and the frames are aligned, we extract both 3D keypoints and 2D keypoints with the VIBE model and checkpoints [25]. The 3D keypoints are normalized with respect to a pre-computed bounding box, while the 2D keypoints are absolute with respect to the frame dimensions. For all experiments unless explicitly stated otherwise, the input sequence length is 30 frames and the output sequence length is 30 sEMG values per muscle, corresponding to 3 seconds. Once the dataset was split into intervals of 3 seconds, the train/test split was created by randomly choosing 20% of the 3 second intervals within an exercise per subject to be allocated to the test set, and the remaining 80% was allocated to the training set.

4. Method

Our approach aims to learn the bidirectional mapping between the visual modality and the muscle modality, which allows us to perform three tasks: a) infer muscle activity from video, b) infer pose from muscle activity, and c) provide recommended motions to people that will target certain muscles. In this section, we present this approach.

4.1. Muscle and Motion Mappings

The characteristics of muscle activity make the sEMG signal challenging to analytically process. We aim to overcome these challenges by leveraging the synchronization with the visual modality. By finding the correlations between a person's visible motion and the sEMG signal, we can learn representations that encode muscle activity with respect to motion.

Let $x \in \mathbb{R}^{KD\times T}$ be the human pose of a person, extracted over K keypoints, with dimensionality D, for T frames in a video. Our goal is to predict the muscle activity that created the motion, which we denote as $m \in \mathbb{R}^{M\times T}$ for M individual muscles, as well as to reconstruct motion $x \in \mathbb{R}^{KD\times T}$ from muscle activity m. We aim to learn mappings that transform between these spaces through the functions:

$$\hat{m} = E_\theta(x) \quad \text{and} \quad \hat{x} = D_\omega(m) \tag{1}$$

where $E_\theta(x)$ is an encoder parameterized by θ and D is a decoder parameterized by ω, both of which are neural

networks whose architecture we describe later. We learn the parameters for both models through the supervised learning problem:

$$\min_{\theta,\omega} \mathbb{E}_{(x,m)} \left[\mathcal{L}\left(E_\theta(x), m\right) + \mathcal{L}\left(E_\omega(m), x\right) \right] \quad (2)$$

where we use a mean squared loss function $\mathcal{L}$ to compare predictions to the ground truth in both modalities. We optimize both using stochastic gradient descent with the Adam optimizer [24]. Full implementation details are provided in the supplemental material.

4.2. Modification in Muscle Space

In this section, we explain how our bidirectional model can be used to generate new motions based on the edits in the muscle modality. Given a goal to minimize a use of the muscle, or increase the workout of a muscle, we generate a new motion, similar to the input motion, with a modification that adheres to the muscle activity goal. To do so, given a video, our encoder E first predicts muscle activation $\hat{m} \in \mathbb{R}^{M \times T}$, composed of M sequences. Let $\hat{m}^k \in \mathbb{R}^T$ be one muscle sequences in particular that we choose to scale, either up or down, with scalar $s \in \mathbb{R}$:

$$\bar{m}^k = s \cdot \hat{m}^k \quad \text{and} \quad \bar{m}^j = \hat{m}^j \; \forall_{j \neq k} \quad (3)$$

The new matrix $\bar{m} \in \mathbb{R}^{M \times T}$ is the edited $\hat{m}$ matrix. Our decoder D decodes $\bar{m}$ into a recommended motion $\bar{x}$:

$$\bar{x} = D(\bar{m}) \quad (4)$$

This recommended motion $\bar{x}$ will be similar to the predicted reconstruction $\hat{x}$, except the recommended motion is in agreement with the muscle goals dictated by the edited predicted muscle activation $\bar{m}$.

4.3. Architectures

We use a common architecture for both the encoder and decoder, with only minimal modifications between them to adapt to their input and output modalities. See Figure 4 for an overview of both architectures. We factorize the architectures:

$$E(x) = f_{\theta_2}\left(g_{\theta_1}\left(x\right)\right) \text{ and } D(m) = f_{\omega_2}\left(g_{\omega_1}\left(m\right)\right) \quad (5)$$

where g is a local feature extractor and f is a global feature network.

The local feature extractor for the Motion-to-Muscle encoder receives keypoints $x \in \mathbb{R}^{KD \times T}$, where K is the number of keypoints, D is their dimensionality, and T is the number of frames. This matrix is then convolved with a filter θ that has c channels:

$$g_\theta(x) = \theta * x \quad (6)$$

The spatial dimension of the kernels θ spans the entirety of the key-point dimension, and the temporal dimension spans roughly a second of time. Each kernel outputs a feature n, where $n \in \mathbb{R}^{1 \times T}$. Since there are c channels, the resulting output from the temporal convolution layer is a sequence of T embeddings $d_1, ..., d_T$, s.t. $d_t \in \mathbb{R}^{128}$.

For the Muscle-to-Motion decoder, the local feature extractor has the same structure, except the input is the muscle activation $m \in \mathbb{R}^{M \times T}$. The convolutional layer's spatial dimension is changed to span the entirety of the muscle dimension, M, and the temporal dimension still spans roughly a second of time. The output dimensionality is thus also a sequence of T embeddings $d_1, ..., d_T$, s.t. $d_t \in \mathbb{R}^{128}$.

The second part of our common architecture, the global feature network, needs to identify long range patterns over time in order to capture the dynamics of the sequence. It is global temporally. We implement it using a Transformer [57] with 4 layers with 8 attention heads and no attention masking. The input to the Transformer is the sequence of embeddings $d_1, ..., d_T$, s.t. $d_t \in \mathbb{R}^{128}$. The output of the Transformer is the sequence of embeddings $o_1, ..., o_T$, s.t. $o_t \in \mathbb{R}^{128}$. For the Motion-to-Muscle encoder, a fully connected layer maps $o_t \in \mathbb{R}^{128}$ to a sequence of embeddings $m_1, ..., m_T$ s.t. $m_t \in \mathbb{R}^M$. For the Motion-to-Muscle decoder, a fully connected layer maps $o_t \in \mathbb{R}^{128}$ to $m_t \in \mathbb{R}^{KD}$.

4.4. Conditioning

For similar motions, different people will vary in muscle activities. This is mostly a product of three factors: a) slight variation in the motion itself b) different muscle recruitment due to personal style [55] c) slight variations in sensor placement and differences in morphology [17]. As such, we construct two additional conditional versions of our encoder and decoder. To accommodate the conditioning, we concatenate a unique tensor $y \in \mathbb{R}^{2 \times T}$ to the sequence of embeddings $d_1, .., d_T$, per subject. Further details can be found in the supplementary.

5. Experiments

The objective of our experiments is to analyze the alignment between the visual modality and the muscle activities underlying motion. We show results across the 15 exercises, using root mean squared error as our metric for both tasks.

5.1. Baselines

Retrieval (Retr.). Our first baseline for solving this problem is to perform nearest neighbor. For the Motion-to-Muscle task, given an example 3D skeleton over time x in the test set X_{test}, we retrieve the nearest neighbor $\bar{x}$ from X_{train}, and assign $\bar{x}$'s muscle activation $\bar{m}$ to $\hat{m}$ as the predicted muscle activation. For the Muscle-to-Motion task, given an example sequence of muscle activity m in the test

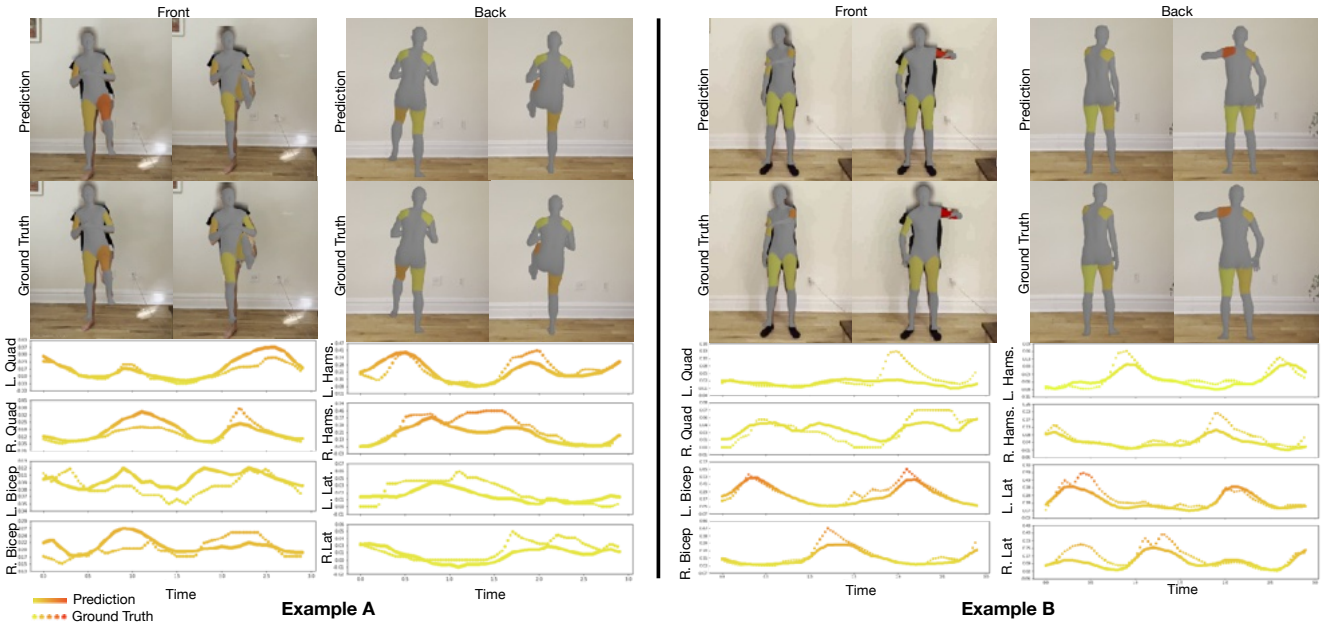

Figure 5: **Motion-to-Muscle Qualitative Results.** We illustrate two separate qualitative results. The first row of frames corresponds to a visualization of the predicted activations, and the second row of frames corresponds to a visualization of the ground truth activations. For the plot beneath the frames, the dotted line corresponds to the ground-truth values, and the solid line corresponds to the predictions. Yellow corresponds to relaxed, and red corresponds to flexed.

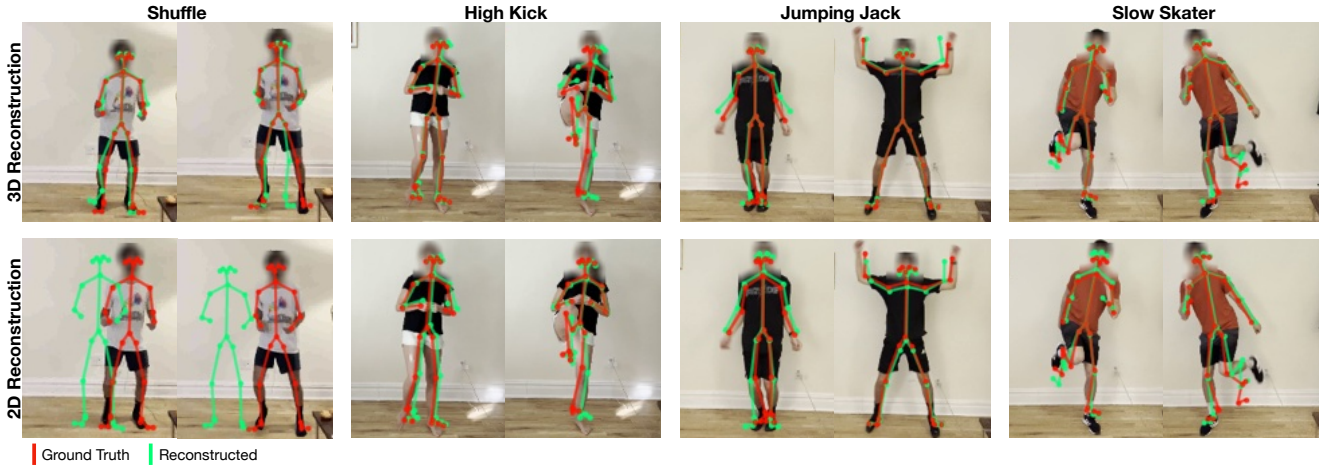

Figure 6: **Muscle-to-Motion Qualitative Results.** We show reconstructed 3D pose and 2D pose from muscle activity, where 3D pose is normalized with respect to a bounding box, while the 2D pose is absolute with respect to the image frame. Since sEMG signals don't contain location information, the 2D model cannot reconstruct the subject in the right location.

set M_{test}, we retrieve the nearest neighbor $\bar{m}$ from M_{train}, and assign $\bar{m}$'s 3D skeleton over time $\bar{x}$ to $\hat{x}$.

Conditional Retrieval (C-Retr.). Our second baseline is conditional retrieval, where we condition on the subject. For the Motion-to-Muscle task, given an example sequence of 3D pose x for subject s, in the test set X^s_{test}, we retrieve the nearest neighbor $\bar{x}$ from X^s_{train}, which only contains data from subject s, and assign $\bar{x}$'s muscle activation sequence $\bar{m}$ to $\hat{m}$ as the predicted muscle activation. We report the average across subjects. The same method is applied to the Muscle-to-Motion task.

5.2. Results

In this section, we report quantitative and qualitative results for both the encoder and the decoder. For the quantitative results, we report the results per exercise. We report these results for our conditional and non-conditional baselines and models.

Motion-to-Muscle Encoder. As seen in Table 1, for all of the 15 exercises, both our conditional and non-conditional model outperform both conditional and non-conditional retrieval baselines. For the out-of-distribution experiments, we retrained 15 models on 15 different datasets, each dataset leaving out one exercise. For each model, we then ran inference on the exercise that was left out and reported it in the last four columns of Table 1, denoted as Out-of-Distribution Encoder. The results indicate that our learning method generalizes better to unseen exercises than the retrieval baselines. Finally, in both the in-distribution and out-of-distribution experiments, and in both our model and baselines, we observe that for most exercises, the conditional model outperforms the non-conditional model. This result confirms the hypothesis presented in Section 4.4.

We show two qualitative examples of predicting muscle activation from motion in Figure 5. The first column shows a subject performing a high kick, and the second column shows a subject performing an elbow punch. The axis in the third and fourth rows are scaled to the range of values per example per muscle, and we denote the absolute value with a gradient from yellow (low activation) to red (high activation). Even when the range of the sEMG signal is small, indicated by a plot that stays mostly one color, we notice that the predictions follow the ground-truth fairly closely. The alignment of the prediction and ground-truth is often close, showing a small error in phase. We show more qualitative results in the supplementary material.

Temporal Analysis. In order to evaluate how crucial the temporal component of our model was, we retrained seperate 7 separate transformer models to predict muscle activation from motion, where both the input and output are n frames for 7 different input/output lengths. We notice that for both the conditional baseline and our conditional model, the performance increases as the model sees examples with longer temporal length. However, the conditional baseline drops in performance from 25 to 30 frames, whereas our conditional model does not.

Exercise	In-Distribution Encoder				Out-of-Distribution Encoder			
	Retr.	C-Retr.	Ours	C-Ours	Retr.	C-Retr.	Ours	C-Ours
ElbowPunch	15.1	15.2	12.0	**12.0**	25.7	29.4	19.8	**19.7**
FrontKick	10.5	9.8	**7.8**	7.9	32.7	54.4	11.0	**11.0**
FrontPunch	10.9	10.7	8.7	**8.6**	27.6	22.8	15.9	**15.5**
HighKick	13.0	12.8	10.1	**10.1**	17.6	17.8	15.8	**15.5**
HookPunch	16.4	16.3	12.5	**12.4**	23.4	23.6	19.3	**18.9**
JumpingJack	25.7	25.4	19.2	**19.2**	47.6	47.0	**37.0**	41.0
KneeKick	11.2	10.7	8.2	**8.0**	16.7	15.5	13.2	**12.8**
KickBack	12.3	11.7	9.3	**9.3**	17.5	19.2	**15.3**	15.6
LegCross	12.0	10.2	8.0	**8.0**	18.1	16.7	**15.2**	15.4
RonddeJambe	23.8	23.7	20.4	**20.3**	36.8	35.0	33.4	**33.2**
Running	15.8	10.6	8.7	**8.6**	27.2	15.6	14.0	**14.0**
Shuffle	13.6	13.2	9.9	**9.9**	22.1	17.0	14.2	**14.0**
SideLunge	17.2	16.5	13.8	**13.7**	27.7	30.1	24.4	**24.0**
SlowSkater	16.8	16.3	13.1	**11.4**	25.2	23.0	21.1	**20.8**
Squat	20.2	19.8	**15.9**	16.0	36.3	34.0	35.2	30.4

Table 1: **RMSE per Exercise for the Encoder**. We report the rMSE per exercise for muscle prediction.

Exercise	In-Distribution Decoder				Out-of-Distribution Decoder			
	Retr.	C-Retr.	Ours	C-Ours	Retr.	C-Retr.	Ours	C-Ours
ElbowPunch	0.045	0.43	0.031	**0.031**	0.078	0.078	0.060	**0.060**
FrontKick	0.058	0.052	**0.040**	0.043	0.103	0.099	**0.074**	0.077
FrontPunch	0.047	0.045	**0.032**	0.033	0.075	0.076	0.062	**0.061**
HighKick	0.093	0.090	0.076	**0.074**	0.145	0.139	0.119	**0.119**
HookPunch	0.060	0.055	**0.044**	0.045	0.090	0.087	**0.075**	0.076
JumpingJack	0.071	0.07	**0.062**	0.066	0.143	0.146	0.109	**0.108**
KneeKick	0.079	0.077	**0.061**	0.064	0.119	0.114	0.096	**0.096**
KickBack	0.086	0.082	0.072	**0.069**	0.116	0.114	0.093	**0.093**
LegCross	0.056	0.049	**0.040**	0.042	0.112	0.106	0.087	**0.087**
RonddeJambe	0.074	0.069	**0.055**	0.056	0.119	0.116	**0.092**	0.093
Running	0.047	0.046	0.037	**0.037**	0.074	0.070	0.052	**0.051**
Shuffle	0.058	0.056	**0.043**	0.044	0.077	0.072	**0.056**	0.057
SideLunge	0.07	0.067	0.058	**0.057**	0.127	0.123	0.108	**0.108**
SlowSkater	0.076	0.072	**0.063**	0.065	0.140	0.124	0.109	**0.109**
Squat	0.066	0.064	**0.057**	0.059	0.132	0.126	**0.111**	0.112

Table 2: **RMSE per Exercise for the Decoder.** We report the rMSE per exercise for motion prediction.

Frame Count	1	5	10	15	20	25	30
C-Retr.	19.6	16.3	14.5	14.0	13.5	13.3	14.3
C-Ours	13.3	11.1	10.3	9.6	9.1	8.8	8.8

Table 3: **Temporal Analysis**. We report the root mean squared error for the conditional baseline as well as for our conditional model as we change the length of the sequences.

Muscle-to-Motion Decoder. Similarly, we report the root mean squared error per exercise, for both the conditional decoder and non-conditional decoder in Table 2. For all of the 15 exercises, both our conditional and non-conditional model outperforms both conditional and non-conditional retrieval baselines. For the decoder, there is less of a clear pattern between the conditioned model and the non-conditioned model. We believe that this is explained by the fact that muscle activity already has conditioning embedded within it. Subjects often have trademark muscles, that they use more or less, or with different ranges. As such, explicit conditioning may not be helpful.

We also show four qualitative examples from our Muscle-to-Motion decoder in Figure 6. The first row illustrates results from our main decoder, which regresses to 3D pose over time. The second row illustrates a secondary decoder, which regresses to 2D pose over time. The extracted 3D keypoints from VIBE [25] are normalized with respect to a given bounding box, which we utilize to project the predicted 3D keypoints onto the 2D image. This is why our 3D pose decoder results have the subject in the right location, even for exercises that have high displacement,

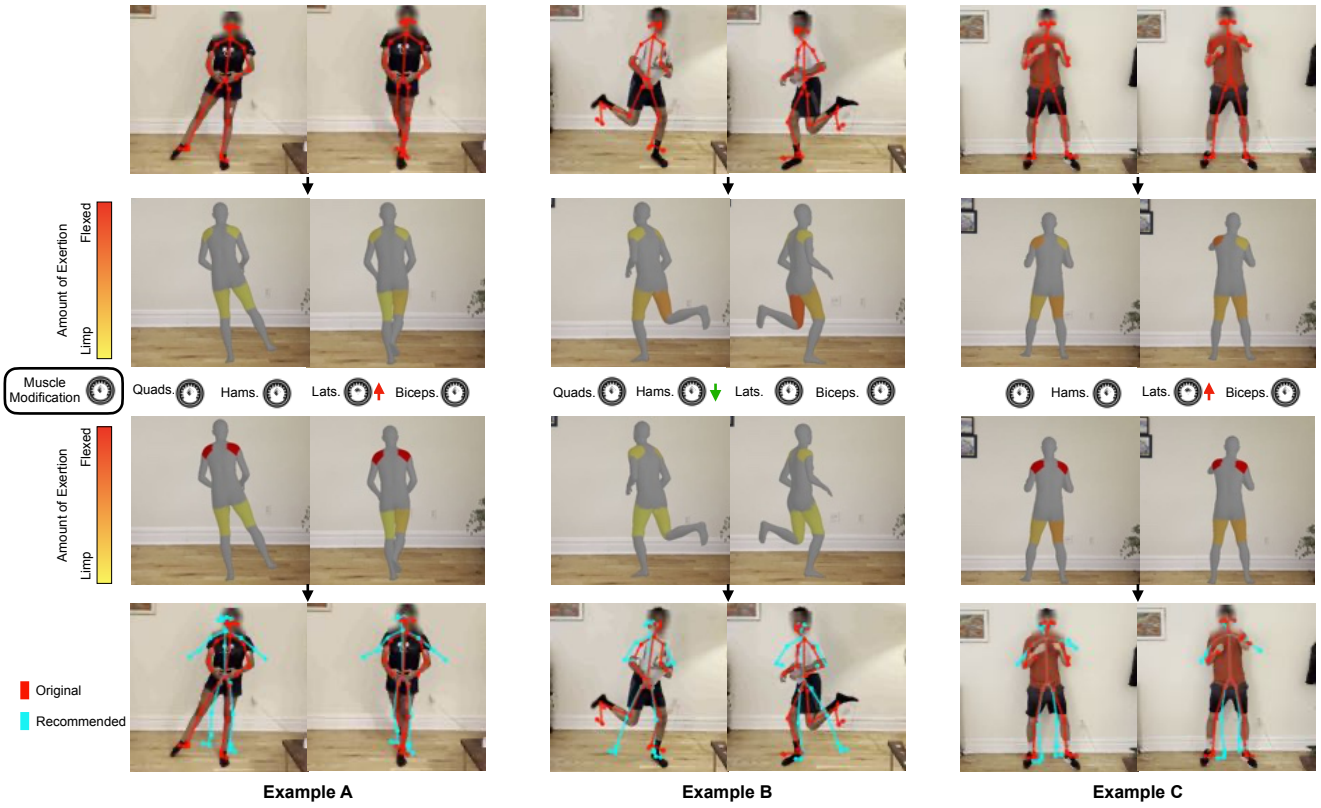

Figure 7: **Qualitative Results.** We illustrate three more qualitative results for the editing task. The first row illustrates the ground truth 3D skeleton projected onto the frame. The second row shows the predicted muscle activation for the dorsal muscles (laterals and hamstrings). Since we are visualizing the back of the person, the meshes are flipped. The third row shows the scaled predicted muscle activations. The fourth row illustrates the recommended motion which the decoder generates from the scaled predicted muscle activations.

such as shuffle. On the contrary, the 2D Decoder, whose ground-truth coordinates are not normalized with respect to a bounding box, is unable to predict the subject's displacement since muscle activation has no information about a subject's location. We show more qualitative results in the supplementary material.

6. Editing

We show examples of how the motion-muscle mappings can be leveraged to generate motion recommendations subject to muscle constraints in Figures 1 and 7. Given a motion, we predict the muscle activation, which we edit one or more muscle predictions, as described in Equations 2 and 3 in Section 4, and decode the edited predicted muscle activation into a recommended motion.

When the modification amplifies the muscle, then it generates a corresponding motion with minimal change that only causes the exercise to engage the target muscle more. For example, in Figure 7A, shows a person performing the Rond de Jambe, however their use of laterals is low. By amplifying the muscular representation, the generated motion lifts the arms up correctly, therefore engaging the laterals more. We see a similar trend in Figure 7C.

There is a converse effect when the modification attenuates the muscle. Figure 7B shows a person performing a slow skater, which due to the non-supporting leg bending backwards, activates the hamstrings significantly. By modifying the muscular representation to attenuate the hamstrings, the generated motion prevents the non-supporting leg from bending backwards, disengaging the hamstrings.

Moreover, the temporal pattern of the recommended motion matches that of the input motion, as the only edit performed is scaling. This is useful for AR/VR applications.

7. Conclusion

This paper presents a new multi-modal dataset, the Muscles in Action (MIA) dataset, for modeling the relationship between muscle activity and motion. We present our framework for learning the bidirectional mapping between the modalities. We also demo how our bidirectional model can be used to generate recommended motions conditioned on muscle activity objectives.

Acknowledgements: We would like to thank our subjects for participating in the dataset. We'd also like to thank Jianbo Shi, Georgia Gkioxari, Huy Ha and Kamyar Ghasemipour for their helpful feedback. This research is based on work partially supported by the NSF NRI Award #1925157. M.C. is supported by the Amazon CAIT PhD fellowship. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the sponsors.

References

[1] Kfir Aberman, Mingyi Shi, Jing Liao, Dani Lischinski, Baoquan Chen, and Daniel Cohen-Or. Deep video-based performance cloning. In *Computer Graphics Forum*, volume 38, pages 219–233. Wiley Online Library, 2019.

[2] Leandro Abraham, Facundo Bromberg, and Raymundo Forradellas. Arm muscular effort estimation from images using computer vision and machine learning. In *Ambient Intelligence for Health*, pages 125–137. Springer, 2015.

[3] Chaitanya Ahuja and Louis-Philippe Morency. Language2pose: Natural language grounded pose forecasting. In *2019 International Conference on 3D Vision (3DV)*, pages 719–728. IEEE, 2019.

[4] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 609–617, 2017.

[5] Andreas Aristidou, Anastasios Yiannakidis, Kfir Aberman, Daniel Cohen-Or, Ariel Shamir, and Yiorgos Chrysanthou. Rhythm is a dancer: Music-driven motion synthesis with global structure. *arXiv preprint arXiv:2111.12159*, 2021.

[6] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. *Advances in neural information processing systems*, 29, 2016.

[7] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.

[8] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5933–5942, 2019.

[9] Yinglin Duan, Tianyang Shi, Zhengxia Zou, Yenan Lin, Zhehui Qian, Bohan Zhang, and Yi Yuan. Single-shot motion completion with transformer. *arXiv preprint arXiv:2103.00776*, 2021.

[10] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *Proceedings of the IEEE international conference on computer vision*, pages 4346–4354, 2015.

[11] Chuang Gan, Deng Huang, Peihao Chen, Joshua B Tenenbaum, and Antonio Torralba. Foley music: Learning to generate music from videos. In *European Conference on Computer Vision*, pages 758–775. Springer, 2020.

[12] Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. Listen to look: Action recognition by previewing audio. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10457–10467, 2020.

[13] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2021–2029, 2020.

[14] Wen Guo, Yuming Du, Xi Shen, Vincent Lepetit, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Back to mlp: A simple baseline for human motion prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4809–4819, 2023.

[15] Félix G Harvey and Christopher Pal. Recurrent transition networks for character locomotion. In *SIGGRAPH Asia 2018 Technical Briefs*, pages 1–4. 2018.

[16] Félix G Harvey, Mike Yurick, Derek Nowrouzezahrai, and Christopher Pal. Robust motion in-betweening. *ACM Transactions on Graphics (TOG)*, 39(4):60–1, 2020.

[17] 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.

[18] Alejandro Hernandez, Jurgen Gall, and Francesc Moreno-Noguer. Human motion prediction via spatio-temporal inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7134–7143, 2019.

[19] Daniel Holden, Jun Saito, and Taku Komura. A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics (TOG)*, 35(4):1–11, 2016.

[20] Joseph HR Isaac, Muniyandi Manivannan, and Balaraman Ravindran. Single shot corrective cnn for anatomically correct 3d hand pose estimation. *Frontiers in Artificial Intelligence*, 5, 2022.

[21] Lise A Johnson and Andrew J Fuglevand. Evaluation of probabilistic methods to predict muscle activity: implications for neuroprosthetics. *Journal of neural engineering*, 6(5):055008, 2009.

[22] Manuel Kaufmann, Emre Aksan, Jie Song, Fabrizio Pece, Remo Ziegler, and Otmar Hilliges. Convolutional autoencoders for human motion infilling. In *2020 International Conference on 3D Vision (3DV)*, pages 918–927. IEEE, 2020.

[23] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[25] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5253–5263, 2020.

[26] Yu Kong, Dmitry Kit, and Yun Fu. A discriminative model with multiple temporal scales for action prediction. In *European conference on computer vision*, pages 596–611. Springer, 2014.

[27] Yu Kong, Zhiqiang Tao, and Yun Fu. Deep sequential context networks for action prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1473–1481, 2017.

[28] Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks for action segmentation and detection. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 156–165, 2017.

[29] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13401–13412, 2021.

[30] Zhan Li, David Guiraud, and Mitsuhiro Hayashibe. Inverse estimation of multiple muscle activations from joint moment with muscle synergy extraction. *IEEE journal of biomedical and health informatics*, 19(1):64–73, 2014.

[31] Lingjie Liu, Weipeng Xu, Michael Zollhoefer, Hyeongwoo Kim, Florian Bernard, Marc Habermann, Wenping Wang, and Christian Theobalt. Neural rendering and reenactment of human actor videos. *ACM Transactions on Graphics (TOG)*, 38(5):1–14, 2019.

[32] Yilin Liu, Shijia Zhang, and Mahanth Gowda. Neuropose: 3d hand pose tracking using emg wearables. In *Proceedings of the Web Conference 2021*, pages 1471–1482, 2021.

[33] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015.

[34] Zhengyi Luo, Ryo Hachiuma, Ye Yuan, and Kris Kitani. Dynamics-regulated kinematic policy for egocentric pose estimation. *Advances in Neural Information Processing Systems*, 34:25019–25032, 2021.

[35] Shuang Ma, Zhaoyang Zeng, Daniel McDuff, and Yale Song. Active contrastive learning of audio-visual video representations. *arXiv preprint arXiv:2009.09805*, 2020.

[36] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2891–2900, 2017.

[37] Yoshihiko Nakamura, Katsu Yamane, Yusuke Fujita, and Ichiro Suzuki. Somatosensory computation for man-machine interface from motion-capture data and musculoskeletal human model. *IEEE Transactions on Robotics*, 21(1):58–66, 2005.

[38] Nadia Nasri, Sergio Orts-Escolano, Francisco Gomez-Donoso, and Miguel Cazorla. Inferring static hand poses from a low-cost non-intrusive semg sensor. *Sensors*, 19(2):371, 2019.

[39] Hui Niu, Takahiro Ito, Damien Desclaux, Ko Ayusawa, Yusuke Yoshiyasu, Ryusuke Sagawa, and Eiichi Yoshida. Estimating muscle activity from the deformation of a sequential 3d point cloud. *Journal of Imaging*, 8(6):168, 2022.

[40] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H Adelson, and William T Freeman. Visually indicated sounds. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2405–2413, 2016.

[41] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *European conference on computer vision*, pages 801–816. Springer, 2016.

[42] Hyun Soo Park, Jianbo Shi, et al. Force from motion: decoding physical sensation in a first person video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3834–3842, 2016.

[43] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10985–10995, 2021.

[44] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 480–497. Springer, 2022.

[45] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter V Gehler, and Bernt Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4929–4937, 2016.

[46] Fernando Quivira, Toshiaki Koike-Akino, Ye Wang, and Deniz Erdogmus. Translating semg signals to continuous hand poses using recurrent neural networks. In *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pages 166–169. IEEE, 2018.

[47] Michalis Raptis and Leonid Sigal. Poselet key-framing: A model for human activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2650–2657, 2013.

[48] Ali Rohan, Mohammed Rabah, Tarek Hosny, and Sung-Ho Kim. Human pose estimation-based real-time gait analysis using convolutional neural network. *IEEE Access*, 8:191542–191550, 2020.

[49] Ryusuke Sagawa, Ko Ayusawa, Yusuke Yoshiyasu, and Akihiko Murai. Predicting muscle activity and joint angle from skin shape. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[50] Masashi Sekiya, Sho Sakaino, and Tsuji Toshiaki. Linear logistic regression for estimation of lower limb muscle activations. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(3):523–532, 2019.

[51] Hyungeun Song and Yoichi Hori. Inverse muscle group activity estimation based on neuromusculoskeletal system model. In *TENCON 2015-2015 IEEE Region 10 Conference*, pages 1–5. IEEE, 2015.

[52] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7464–7473, 2019.

[53] Yokhesh K Tamilselvam, Jacky Ganguly, Rajni V Patel, and Mandar Jog. Musculoskeletal model to predict muscle activity during upper limb movement. *IEEE Access*, 9:111472–111485, 2021.

[54] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022.

[55] Elly Trepman, Richard E Gellman, Lyle J Micheli, and CARLO J De Luca. Electromyographic analysis of grand-plié in ballet and modern dancers. *Medicine and science in sports and exercise*, 30(12):1708–1720, 1998.

[56] Arash Vahdat, Bo Gao, Mani Ranjbar, and Greg Mori. A discriminative key pose sequence model for recognizing human interactions. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 1729–1736. IEEE, 2011.

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

[58] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. *arXiv preprint arXiv:1808.06601*, 2018.

[59] Xionghui Wang, Jian-Fang Hu, Jian-Huang Lai, Jianguo Zhang, and Wei-Shi Zheng. Progressive teacher-student learning for early action prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3556–3565, 2019.

[60] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021.

[61] Katsu Yamane, Akihiko Murai, Sadahiro Takaya, and Yoshihiko Nakamura. Muscle tension database for contact-free estimation of human somatosensory information. In *2009 IEEE International Conference on Robotics and Automation*, pages 633–638. IEEE, 2009.

[62] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11038–11049, 2022.

[63] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. *arXiv preprint arXiv:2212.02500*, 2022.

[64] Ye Yuan, Shih-En Wei, Tomas Simon, Kris Kitani, and Jason Saragih. Simpoe: Simulated character control for 3d human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7159–7169, 2021.

[65] Feng Zhang, Pengfeng Li, Zeng-Guang Hou, Zhen Lu, Yixiong Chen, Qingling Li, and Min Tan. semg-based continuous estimation of joint angles of human legs by using bp neural network. *Neurocomputing*, 78(1):139–148, 2012.