VQ-MyoPose: Movement tokenization improves decoding of hand kinematics from surface EMG wristbands

Rossana Lovecchio[1, 4] Pranav Mamidanna [2, 3] Bart Jansen[1, 5] Dario Farina[2] Tom Verstraten[1, 4]
[1]Robotics and Multibody Mechanics Research Group, Vrije Universiteit Brussel
[2]Department of Bioengineering, Imperial College London, London, UK
[3]I-X Center for AI in Science, Imperial College London, London, UK
[4]Flanders Make, 1050 Brussels, Belgium [5]IMEC, 3001 Leuven, Belgium

Abstract

Surface electromyography (sEMG) is an attractive and easily available signal for decoding hand movement, yet models that map sEMG to fine-grained hand pose often struggle to generalize across users, electrode placements, and task contexts. We propose a two-stage model that first learns a tokenized representation of hand kinematics through a VQ-VAE, and subsequently maps EMG signals to the learned hand movement tokens. Evaluated on the emg2pose benchmark – 16-channel sEMG at 2 kHz synchronized with 60 Hz joint angles – our approach improves over strong regression baselines and generalizes better on cross-task and cross-user settings, measured by mean absolute joint angle error and fingertip landmark distance. These results were evaluated on a subset of the original dataset, which included 12 subjects and 27 movement repertoires. On the held-out user set, our model, VQ-MyoPose, achieves 10.2° MAE and 14.7mm fingertip error, outperforming vemg2pose (12.2°, 15.8mm). Discrete tokens exhibit concentrated codebook usage and correlations with task structure, offering interpretability and compression. These results support movement tokenization as a compact, transferable intermediate for estimating hand kinematics from sEMG signals.

1. Introduction

Decoding hand movement from sEMG opens significant new possibilities for multiple fields, such as human-computer interfaces for virtual environments [4, 12], robust control for prosthetics [3], and movement neuroscience [5]. However, consistently decoding hand movement across different gestures (cross-task) and building interfaces that adapt to novel users (cross-user generalization) remains difficult: anatomy, electrode placement, and task context alter the mapping from muscle activity to hand movement. Previous work relies largely on end-to-end regression of *continuous* hand kinematics (e.g., joint angles, positions, or velocities) [8, 13, 15, 16]. While effective, such models often lack interpretability and may fail to capture recurring structure in movement that might aid generalization.

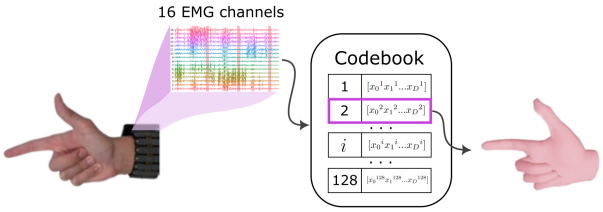

Figure 1. **VQ-MyoPose: From EMG to discrete movement tokens.** Surface electromyography (sEMG) recorded at the wrist is mapped to discrete latent codes drawn from a learned codebook of movement tokens. By predicting these codes, rather than continuous trajectories directly, the model provides a compact and interpretable representation of hand pose.

An alternative is to learn discrete latent representations of movement, in line with neuroscience theories of 'movement primitives' [6, 18], where dexterous behaviours are thought to be composed of a finite set of recurring movement motifs. To achieve this, we employ Vector-Quantized Variational Autoencoders (VQ-VAE) [11], which have shown remarkable success in speech, vision, and control tasks in mapping high-dimensional signals into sequence of tokens that capture recurring motifs. Such a learned codebook of movement tokens provides a novel inductive bias that (i) improves cross-task and cross-user generalisation by reusing shared tokens, (ii) provide a transferable intermediate representation for decoding hand kinematics from EMG, and (iii) yields a more interpretable model of EMG to movement decoding.

2. Background and related work

The task of decoding hand kinematics has traditionally received attention in computer vision and propelled by remark-

able progress in motion capture and hand modelling [2]. However, performance often degrades under occlusion, limited field-of-view, or poor lighting [9]. By contrast, sEMG offers an ever present modality that directly measures the electrical activity of muscles that drives movement in a wearable form factor, making it an attractive complementary modality for robust hand tracking [7, 13].

The main task is to learn a mapping from raw sEMG signals to continuous hand kinematics. Formally, given input sequences $X \in \mathbb{R}^{C\times T}$ where C the number of sEMG channels, and T the temporal window length, the objective is to predict corresponding hand kinematics $Y \in \mathbb{R}^{J\times T}$, where J denotes the degrees of freedom of the hand model. This problem is highly non-linear and underdetermined: sEMG encodes muscle activations, which relate more closely to motion derivatives than to static pose [8, 14]. As a result, effective solutions must exploit temporal context and generalize across users and recording conditions. Recent baselines such as vemg2pose [13] and NeuroPose [8] have introduced distinct neural architectures to address this challenge: *vemg2pose* employs a time-depth separable convolutional encoder and autoregressive LSTM decoder to estimate joint angular velocities, while *NeuroPose* leverages a U-Net-style architecture to directly regress joint angles. While these models provide competitive baselines, generalizing from noisy, user-specific signals remains a core challenge.

We base our study on the *emg2pose* dataset [13], the largest publicly available benchmark for hand pose estimation from sEMG. The dataset contains wrist-based recordings from a 16-channel bipolar sEMG device sampled at 2 kHz, alongside synchronized 3D hand joint angle labels captured with a 26-camera motion capture system. In total, it spans 193 users, 370 hours, and 29 kinematic 'stages' (i.e. groups of specific types of gestures) comprising over 80M labeled frames, a scale comparable to leading vision-based hand pose datasets. The collection protocol introduces variability across three main axes: (i) *user anatomy*, (ii) *sensor placement*, and (iii) *movement repertoire* (gestures within each stage).

The scale and diversity of *emg2pose* make it ideally suited to explore modern self-supervised and discrete representation learning methods such as VQ-VAE [11], which have demonstrated success in speech [1] and vision domains [17, 19]. Hence, our goal is to learn a compact and transferable *tokenized motion representations* that disentangle user-specific variability and improve downstream EMG-to-pose decoding.

3. Methods

Our approach consists of two complementary stages that jointly enable discrete representation learning from hand kinematics and subsequent decoding from surface electromyography (sEMG). The overall pipeline is illustrated in

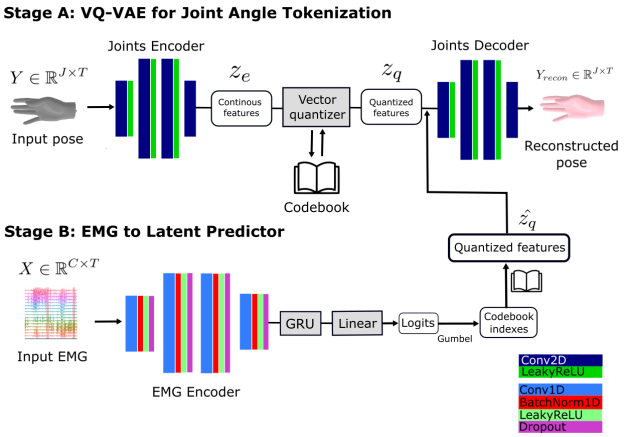

Figure 2. **Overview of the two-stage architecture of VQ-MyoPose.** *Stage A (top):* A VQ-VAE is trained on ground-truth joint angle trajectories $Y \in \mathbb{R}^{J\times T}$ ($J = 20$, joints angles, $T = 2000$, window size) to tokenize movement into a sequence of discrete latent codes. *Stage B (bottom):* An EMG encoder processes input signals $X \in \mathbb{R}^{C\times T}$ ($C = 16$, EMG channels) and uses Gumbel-Softmax sampling to predict code indices $\hat{i}$. These indices select quantized embeddings $\hat{z}_q$ from the frozen codebook of Stage A, which are passed through the frozen decoder to reconstruct joint angles.

Fig. 2.
VQ-VAE for Joint Angle Tokenization. In the first stage, we trained a VQ-VAE [11] on ground-truth joint angle trajectories $x \in \mathbb{R}^{20\times T}$. This module compressed motion sequences into a sequence of discrete latent variables, i.e., movement tokens, drawn from a learned codebook. To exploit structural correlations between channels and the natural configuration of the hand, we reshaped inputs to $x^{(3D)} \in \mathbb{R}^{5\times 4\times T}$. An **encoder** comprising stacked 2D convolutions maps $x^{(3D)}$ to latent features $\hat{z_q} \in \mathbb{R}^{D\times H\times W}$, with strides restricted to the temporal axis to prevent downsampling in height (as per the final configuration, $D = 10, H = 5, W = 200$). A **vector quantizer** mapped each spatial position in z_e to the nearest codeword of a learnable codebook $E \in \mathbb{R}^{K\times D}$, where $K = 128$ is the size of the codebook and $D = 10$ is the size of each codebook vector.

The quantized latents z_q were used in place of z_e and optimized with a composite quantization loss:

$$\mathcal{L}_{\text{VQ}} = \underbrace{\|z_q - \text{sg}[z_e]\|_2^2}_{\text{codebook loss}} + \beta \underbrace{\|z_e - \text{sg}[z_q]\|_2^2}_{\text{commitment loss}} + \underbrace{\lambda_{\text{smooth}}\left(\|\Delta_W z_q\|_2^2 + \|\Delta_H z_q\|_2^2\right) - \lambda_{\text{usage}} \frac{H(p)}{\log K}}_{\text{regularizers}}. \quad (1)$$

where $\text{sg}[\cdot]$ denotes the stop-gradient operator, Δ are finite differences along temporal and spatial axes, and $H(p)$ is the entropy of code usage. This loss encourages accurate

reconstruction, encoder commitment, spatial smoothness across tokens, and uniform codebook utilization.

A **decoder** mirrored the encoder with upsampling blocks that expand only along the temporal axis. Reconstructed tensors $\hat{x}^{(3D)} \in \mathbb{R}^{5\times4\times T}$ were reshaped back to $\mathbb{R}^{20\times T}$. Finally, a 1×1 convolutional "mixing" layer implemented a learnable permutation/sign/scale transformation across the 20 channels. Finally, the total loss combined the reconstruction loss (L1 error) calculated on the decoder output and the quantization loss (1):

$$\mathcal{L}_{\text{total}} = \lambda_{\text{recon}}\mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{VQ}} \tag{2}$$

EMG to Latent Predictor In the second stage, we mapped multi-channel sEMG windows to discrete latent code indices (for a fixed VQ-VAE codebook of size K) with a CNN–RNN sequence classifier. Let $x \in \mathbb{R}^{C\times T}$ be a batch of sEMG segments ($C = 16$ channels, $T = 2000$ samples).

A 1-D convolutional *feature encoder* processed x with stacked **Conv1d** layers (kernel/padding/stride: $(10, 5, 3)$, $(6, 2, 2)$, $(4, 1, 1)$, $(3, 0, 1)$), each followed by BatchNorm, LeakyReLU, and dropout ($p = 0.2$). This stage can extract localized myoelectric patterns while reducing temporal resolution in a manner aligned with the Stage A VQ-VAE encoder.

The resulting features were passed to a *bidirectional GRU* and to a 1D convolution with kernel size 31 that acts as an *alignment filter*, initialized as identity but trainable to refine local timing. Finally, a *linear projection* mapped each time step to $H \times K$ logits, reshaped into (H, W, K) and flattened to (L, K) with $L = HW$. These logits were converted to quantized embeddings $\hat{z}_q$ using Gumbel-softmax sampling against the frozen codebook $E \in \mathbb{R}^{K\times D}$ from Stage A. Hence, at each training step, the frozen pose VQ-VAE produced target code indices $i \in [K]^{HW}$ and quantized latents ${z_q}^{tgt} \in \mathbb{R}^{D\times H\times W}$ from ground-truth joint angles. A **cross-entropy** loss supervises token prediction:

$$\mathcal{L}_{\text{latent}} = \text{CE}_{\epsilon}\big(\ell,\, i^{tgt}\big),$$

with smoothing $\epsilon = 0.05$. In addition, the predicted embeddings $\hat{z}_q$ were decoded through the frozen VQ-VAE decoder to produce $\hat{x}$, resulting in a **reconstruction loss**

$$\mathcal{L}_{\text{recon}} = \|\hat{x} - x_{\text{tgt}}\|_2^2.$$

The final training objective is a weighted combination using $\lambda_{latent} = 1.0$ and $\lambda_{recon} = 0.5$

$$\mathcal{L}_{\text{total}} = \lambda_{\text{latent}}\mathcal{L}_{\text{latent}} + \lambda_{\text{recon}}\mathcal{L}_{\text{recon}}. \tag{3}$$

Together, these two stages factorized the sEMG-to-pose inference task into: (i) learning a discrete and interpretable representation of hand motion, and (ii) learning to map noisy muscle activity to this stable latent space. This modular design leverages the strengths of VQ-based representation learning in an attempt to simplify the research problem of joint pose regression from EMG data through the dimensionality reduction of the joint space.

Training setup We trained both models using a batch size of 128 and using 1 second of non-overlapping trajectories. The employed dataset was a subset of the *emg2pose* dataset and included 12 subjects and 27 stages, which was then divided into training, validation, and test sets with a split of 70%, 15%, and 15%. The training was carried out on a NVIDIA GeForce GTX 1080 Ti GPU for 50 epochs.

4. Results

Before evaluating EMG-driven prediction, we first assess the reconstruction performance of the VQ-VAE when trained directly on joint angle trajectories (Stage A). The VQ-VAE achieves joint angle errors of $7.1 \pm 0.7°$, $6.7 \pm 0.7°$, and $6.9 \pm 1.1°$ for the *User*, *Stage*, and *User+Stage* splits, respectively, while fingertip landmark errors remain below a centimeter for *User* (8.3 ± 0.8 mm) and *Stage* (7.8 ± 0.7 mm), and rise to 13.3 ± 2.2 mm in the combined *User+Stage* setting. These results confirm that the VQ-VAE can tokenize motion into a compact codebook while maintaining accurate reconstructions across all generalization regimes. This baseline therefore establishes a strong lower bound on achievable MAE, demonstrating that the learned codebook is expressive enough to capture and compose complex hand movements, motivating its use as an intermediate representation for EMG-to-pose decoding.

We next evaluate our full two-stage pipeline, where the EMG encoder predicts discrete tokens from wrist sEMG and the frozen VQ-VAE decoder reconstructs hand kinematics. Compared to *vemg2pose* [13], our approach brings clear gains in the most challenging generalization settings. On held-out *User*, VQ-MyoPose reduces joint angle error from $12.2 \pm 1.3°$ to $10.2 \pm 1.1°$ and fingertip distance from 15.8 ± 1.9 mm to 14.7 ± 2.4 mm, corresponding to a relative improvement of $\sim$16% in angle MAE and $\sim$7% in fingertip accuracy. On the combined *User+Stage* split, errors drop from $15.8 \pm 1.4°$ to $14.6 \pm 0.7°$ and from 21.6 ± 2.0 mm to 18.8 ± 0.9 mm, a gain of $\sim$8% and $\sim$13%, respectively. Performance degrades on the *Stage* split ($21.9 \pm 3.7°$, 27.2 ± 2.6 mm), reflecting the difficulty of generalizing across unseen kinematic conditions without user change, a known limitation of EMG-based decoding. Overall, these results show that tokenizing hand motion with a VQ-VAE provides a more stable latent target than direct regression, especially improving generalization across users and combined user+stage regimes.

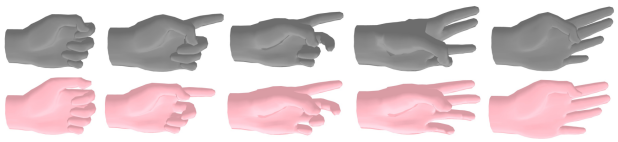

Figure 3. VQ-MyoPose predictions (pink) vs. ground truth (gray) hand poses for a 'finger counting' task.

Table 1. Angle error (MAE, in degrees) under different generalization splits.

Method	User	Stage	User,Stage
vemg2pose [13]	12.2 ± 1.3	15.2 ± 1.6	15.8 ± 1.4
VQ-MyoPose (ours)	$\mathbf{10.2 \pm 1.1}$	21.9 ± 3.7	$\mathbf{14.6 \pm 0.7}$

Table 2. Fingertip landmark distance (mm) under different generalization splits.

Method	User	Stage	User,Stage
vemg2pose [13]	15.8 ± 1.9	20.4 ± 2.2	21.6 ± 2.0
VQ-MyoPose (ours)	$\mathbf{14.7 \pm 2.4}$	27.2 ± 2.6	$\mathbf{18.8 \pm 0.9}$

Codebook analysis The learned codebook exhibits healthy utilization and diversity. In particular, *codebook usage* is high (81.07%), and the elements are well separated as the *mean off-diagonal similarity* is only 0.0128, indicating near-orthogonality between codes. Codebook assignments are reasonably dispersed across entries, as suggested by the *row-wise mean entropy* of 4.7341, and the *spectral decay* of 0.4230 hints that the codes capture mixed-scale structure rather than collapsing to a narrow frequency band. However, there is still room for improvement: the *local smoothness* score (-0.0374) and a modest *silhouette* at k=10 (0.2437) suggest that temporal transitions between tokens could be smoother and clusters more separable.

5. Limitations

While our approach demonstrates promising improvements, several limitations remain. First, residual variability arising from user anatomy and electrode placement indicates the need for personalization strategies or domain adaptation techniques. Second, the current model architecture is relatively simple, leaving substantial room for methodological advancement. Finally, our experiments were conducted on a restricted subset of the *emg2pose* dataset; more definitive results will require training and evaluation at the full scale of the benchmark.

6. Conclusion

In this work, we presented a two-stage framework for decoding hand kinematics from wrist sEMG signals using discrete latent representations. By first training a VQ-VAE [11] to tokenize joint angle trajectories into discrete latent representations and then learning an EMG-to-latent predictor aligned to this discrete codebook, we demonstrated that sEMG-driven pose estimation can benefit from movement-related intermediate representations. Evaluated on a subset of the large-scale *emg2pose* benchmark [13], our approach achieves improved generalization compared to strong regression baselines, particularly under cross-user and cross-stage conditions.

Beyond performance, the use of discrete tokens offers interpretability, compression, and the possibility of applying sequence modeling techniques developed in speech and vision domains (e.g., CPC [10], vq-wav2vec [1]). These findings support the idea that discrete representation learning can serve as a robust interface for biosignal-driven control, spanning prosthetics, teleoperation, and immersive virtual environments. Future work will extend this approach with multimodal fusion (e.g., RGB-D, force sensing) over the learned code sequences.

7. Funding

R.L. is funded from Research Foundation Flanders (FWO) with project number FWOSB163. P.M. is funded by an Eric and Wendy Schmidt Postdoctoral Fellowship in AI for Science.

References

[1] Alexei Baevski, Steffen Schneider, and Michael Auli. Vq-wav2vec: Self-Supervised Learning of Discrete Speech Representations.

[2] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European conference on computer vision*, pages 561–578. Springer, 2016.

[3] Ziming Chen, Huasong Min, Dong Wang, Ziwei Xia, Fuchun Sun, and Bin Fang. A review of myoelectric control for prosthetic hand manipulation. *Biomimetics*, 8(3):328, 2023.

[4] Anany Dwivedi, Yongje Kwon, and Minas Liarokapis. Emg-based decoding of manipulation motions in virtual reality: Towards immersive interfaces. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 3296–3303. IEEE, 2020.

[5] Dario Farina, Roberto Merletti, and Roger M Enoka. The extraction of neural strategies from the surface emg. *Journal of applied physiology*, 96(4):1486–1495, 2004.

[6] Neville Hogan and Dagmar Sternad. Dynamic primitives of motor behavior. *Biological cybernetics*, 106(11):727–739, 2012.

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

[8] Yilin Liu, Shijia Zhang, and Mahanth Gowda. A practical system for 3-d hand pose tracking using emg wearables with applications to prosthetics and user interfaces. *IEEE Internet of Things Journal*, 10(4):3407–3427, 2023.

[9] Franziska Mueller, Dushyant Mehta, Oleksandr Sotnychenko, Srinath Sridhar, Dan Casas, and Christian Theobalt. Real-time hand tracking under occlusion from an egocentric rgb-d sensor. In *Proceedings of the IEEE international conference on computer vision*, pages 1154–1163, 2017.

[10] Aaron Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding, .

[11] Aaron Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural Discrete Representation Learning, .

[12] Cristina Polo-Hortigüela, Miriam Maximo, Carlos A Jara, Jose L Ramon, Gabriel J Garcia, and Andres Ubeda. A comparison of myoelectric control modes for an assistive robotic virtual platform. *Bioengineering*, 11(5):473, 2024.

[13] 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.

[14] 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.

[15] Raul C Sîmpetru, Andreas Arkudas, Dominik I Braun, Marius Osswald, Daniela Souza de Oliveira, Bjoern Eskofier, Thomas M Kinfe, and Alessandro Del Vecchio. Learning a hand model from dynamic movements using high-density emg and convolutional neural networks. *IEEE Transactions on Biomedical Engineering*, 2024.

[16] Viswanath Sivakumar, Jeffrey Seely, Alan Du, Sean Bittner, Adam Berenzweig, Anuoluwapo Bolarinwa, Alex Gramfort, and Michael Mandel. emg2qwerty: A large dataset with baselines for touch typing using surface electromyography. *Advances in Neural Information Processing Systems*, 37:91373–91389, 2024.

[17] Junke Wang, Yi Jiang, Zehuan Yuan, Bingyue Peng, Zuxuan Wu, and Yu-Gang Jiang. Omnitokenizer: A joint image-video tokenizer for visual generation. *Advances in Neural Information Processing Systems*, 37:28281–28295, 2024.

[18] Daniel M Wolpert and Zoubin Ghahramani. Computational principles of movement neuroscience. *Nature neuroscience*, 3(11):1212–1217, 2000.

[19] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021.