Local Dimension Enhancement Representation Learning for Skeleton-Based Action Segmentation (Supplementary Material)

Contents

1. Dimension collapse and local intrinsic dimension measures

1.1. Definition of dimension collapse

Self-supervised learning methods can be easily affected by dimension collapse, where the learned feature space is of low dimensionality and thus lacks the representational capacity for downstream tasks. Formally, given a set of D-dimensional generated features $\mathcal{F} = \{f_i \in \mathbb{R}^D\}_{i=1}^N$, we denote the space spanned by the features in $\mathcal{F}$ as $\text{span}\{\mathcal{F}\}$ and its dimension as $\dim(\text{span}\{\mathcal{F}\})$.

Definition 1 (Dimension Collapse)**.** *Dimension collapse happens in $\mathcal{F}$, if*

$$d \triangleq \dim(\text{span}\{\mathcal{F}\}) < \min\{N, D\}. \tag{1}$$

This kind of dimension collapse is also termed as *global dimension collapse*. The worst case is that $d = 1$, where all the $f_i \in \mathcal{F}$ equal to a constant vector. This is called *complete collapse*. Beyond global dimension collapse, there exists another form of dimension collapse known as *local dimension collapse*.

Definition 2 (Local Dimension Collapse)**.** *Given a location $x \in \mathbb{R}^D$ and a positive number ϵ, the vicinity of x with radius ϵ is denoted as $\mathcal{B}(x, \epsilon)$. Local dimension collapse happens around x, if*

$$d_{(x,\epsilon)} \triangleq \dim(\text{span}\{\mathcal{F} \cap \mathcal{B}(x, \epsilon)\}) < \min\{|\mathcal{F} \cap \mathcal{B}(x, \epsilon)|, D\}.$$

1.2. Local effective rank (LER)

We follow the definition given by Roy *et al*. [6] to calculate the Effective Rank value. Consider a complex-valued non-all-zero matrix $A \in \mathbb{R}^{M\times N}$. Its singular value decomposition (SVD) is given by $A = UDV$ where $U \in \mathbb{R}^{M\times M}$ and $V \in \mathbb{R}^{N\times N}$ are unitary matrices, and $D \in \mathbb{R}^{M\times N}$ is a diagonal matrix whose (real positive) singular values are:

$$\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_Q \geq 0,$$

where $Q = \min\{M, N\}$. We futher define $\sigma = (\sigma_1, \sigma_2, ..., \sigma_Q)^\top$ and the singular value distribution $p = (p_1, p_2, ..., p_Q)^\top$ given by

$$p_k = \frac{\sigma_k}{||\sigma||_1} \quad \text{for } k = 1, 2, ..., Q, \tag{2}$$

where $^\top$ denotes the transpose and $||\cdot||_1$ is the l_1-norm defined as

$$||\sigma||_1 = \sum_{k=1}^{Q} |\sigma_k|. \tag{3}$$

Then the Effective Rank of A, denoted as $erank(A)$, is formulated as

$$erank(A) = e^{H(p)}, \tag{4}$$

where $H(p)$ is the entropy given by

$$H(p) = -\sum_{k=1}^{Q} p_k \log p_k, \tag{5}$$

where the logarithm is to the base e and $0 \log 0 = 0$.

In our experiments, we compute the motion unit-scale Local Effective Rank (LER) for each joint of each sample and report the arithmetic average value. Concretely, for the i-th sample X_i in the cropped skeleton sequence set $\{X_k\}_{k=1}^N$, we denote its latent representation after the encoder as $Z_i \in \mathbb{R}^{T\times V\times C}$, where N is the number of sequences, T is the number of motion units in the temporal dimension, V is the joint number, and C is the dimension of each feature. Then, the matrix corresponding to the feature space of the j-th joint is denoted as $A_{ij} = Z_{i,:,j,:}$. Finally, the average motion unit-scale LER for $\{X_k\}_{k=1}^N$ is given by

$$\frac{1}{N \times V} \sum_{i=1}^{N} \sum_{j=1}^{V} erank(A_{ij}), \tag{6}$$

1.3. Method of moments (MOM)

Following LDReg [3], we adopt the Method of Moments [2] to indicate fractal dimension. Specifically, for a batch of features $B = \{z_k\}_{k=1}^K$, we first choose an anchor feature $z_{anc} \in B$, and compute its Euclidean distances to other features in the batch:

$$d_{anc,k} = ||z_{anc} - z_k||_2, \quad z_k \in B - \{z_{anc}\}. \tag{7}$$

Then we compute the average diatance μ_{anc} and the largest distance ω_{anc}, and the MOM value of the anchor is estimated by

$$\text{MOM}_{anc} = -\frac{\mu_{anc}}{\mu_{anc} - \omega_{anc}} \tag{8}$$

Finally, the MOM value of the batch is given by the arithmetic average of the MOM values calculated with each feature in the batch as the anchor.

For computing the motion unit-scale MOM, we follow the notation defined in Sec. 1.2, and the feature batch of the j-th joint in the i-th sample is denoted as $\{Z_{i,k,j,:}\}_{k=1}^\top$. After obtaining the MOM value of each batch with Equation 7 and Equation 8, we compute the arithmetic average across all the batches for the final result.

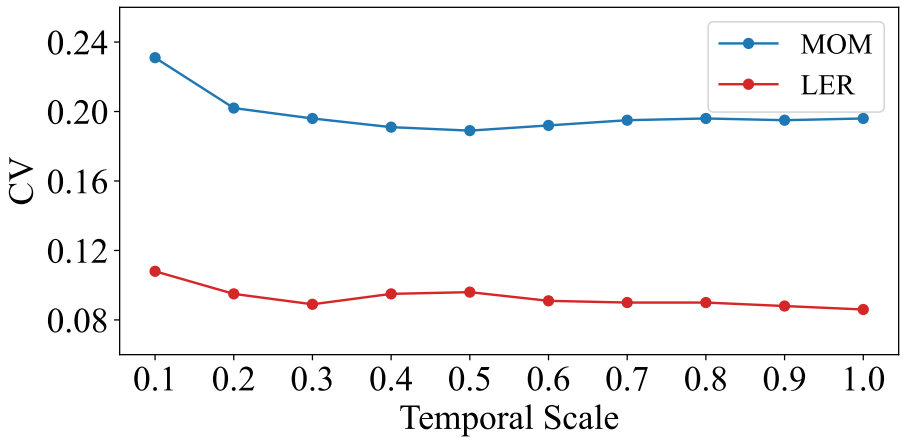

Figure 1. The coefficients of variation comparison between MOM and LER on the representations learned by MAMP. The CV of LER is significantly lower than MOM.

1.4. Stability comparison between LER and MOM

We calculate the *Coefficients of Variation* (CV) of MOM and LER for the representations learned by MAMP [5] as shown in Figure 1 to compare the stability of the two measures. Compared to MOM in Figure 1, LER achieves much lower CV values, and shows a smaller increase in CV at short temporal scales, which is more stable.

We further analyze the stability of the LER loss with respect to hyperparameters.

Batch Size Independence: The LER loss is computed based on the internal correlation of motion units within each sequence, making its formulation theoretically independent of the training batch size. This property ensures consistent optimization behavior regardless of GPU memory constraints.

Convergence and Motion Units: To verify numerical stability regarding the number of motion units, we visualize the training curves of LER loss with different numbers of motion units. As shown in Figure 2, LER loss exhibits an overall stable convergence trend under all settings. This demonstrates that the LER loss is generally robust to the motion unit number.

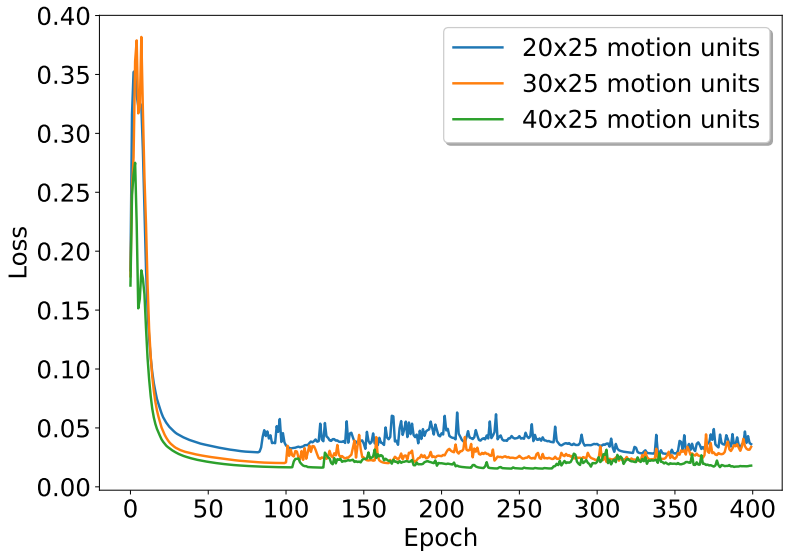

Figure 2. LER loss curves when training with different number of motion units. The number of motion units is labeled as temporal number (20, 30, and 40) $\times$ joint number (25).

2. Theoretical Proofs

2.1. LER's properties for rank estimation

Proposition 1. *The rank of local representations constitutes a tight upper bound for LER, satisfying:*

$$1 \leq \mathrm{LER}(Z) \leq rank(Z) \leq n, \tag{9}$$

where $rank(\cdot)$ denotes the rank operator and $n = \min\{m, d\}$. The equality holds if and only if the non-zero singular values are uniformly distributed, i.e.*,*

$$\sigma_i = \sigma_j \neq 0,\ \forall i, j \leq rank(Z). \tag{10}$$

Proof. Similar to the denotation in Eq. (5), the Shannon Entropy of the spectrum of Z is defined as:

$$H(\boldsymbol{\sigma}) = H((\sigma_1, \sigma_2, \cdots, \sigma_n)) = -\sum_{k=1}^{n} \sigma_k \log \sigma_k, \tag{11}$$

where $\boldsymbol{\sigma}$ is defined as the l_1-normalized spectrum for clarity without loss of generality. The following inequalities naturally holds according to the properties of Shannon Entropy:

$$0 = H(1, 0, 0, \cdots, 0) \leq H(\boldsymbol{\sigma}) \leq H(\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n}) = \log n, \tag{12}$$

from which we can directly derive:

$$1 \leq \mathrm{LER}(Z) \leq n. \tag{13}$$

Denote $rank(Z)$ as q, the entropy of the spectrum degenerates into $H(\sigma_1, \sigma_2, \cdots, \sigma_q, 0, \cdots, 0)$, and the following inequality holds:

$$H(\sigma_1, \sigma_2, \cdots, \sigma_q, 0, \cdots, 0) = -\sum_{k=1}^{q} \sigma_k \log \sigma_k \leq \log q \Rightarrow \mathrm{LER}(Z) \leq q = rank(Z), \tag{14}$$

and it is obvious that when and only when $\sigma_1 = \sigma_2 = \cdots = \sigma_q = \frac{1}{q}$, the equality holds. □

Proposition 2 (Scale Invariance)**.** *For all $c \neq 0$, the following equality holds:*

$$\mathrm{LER}(c \cdot Z) = \mathrm{LER}(Z). \tag{15}$$

Proof. The spectrum of $c \cdot Z$ is $|c| \cdot \boldsymbol{\sigma} = (|c|\sigma_1, |c|\sigma_2, \cdots, |c|\sigma_n)$. After l_1-normalized, the spectrum also turns into $\boldsymbol{\sigma}$, so the equality naturally holds. □

Proposition 3 (Orthogonal Invariance)**.** *For any orthogonal matrix $U \in \mathbb{R}^{d \times d}$, the following equality holds:*

$$\mathrm{LER}(UZ) = \mathrm{LER}(Z). \tag{16}$$

Proof. Consider the singular values as the sqaure roots of the eigenvalues of the corresponding Gram matrix, the eigenvalues are computed as the roots of characteristic polynomial:

$$\det((UZ)^\top (UZ) - \lambda \boldsymbol{I}) = \det(Z^\top U^\top U Z - \lambda \boldsymbol{I}) = \det(Z^\top Z - \lambda \boldsymbol{I}), \tag{17}$$

where $\boldsymbol{I}$ is the identity matrix. It demonstrates that orthogonal transformations do not change the spectrum, and consequently keep the LER value. □

2.2. Coding length of low-rank approximation

Lemma 1. *Denote $Z \in \mathbb{R}^{d\times m}$ as the matrix formed by a set of local representation vectors, with m representing the number of vectors and d denoting the dimensionality of each representation vector. Assume each real value has a fixed coding length of b bits, and the coding length L_r under the optimal rank-r approximation of matrix Z is given by:*

$$L_r = b \cdot r \cdot (d + m + 1). \tag{18}$$

Proof. Let $Z \in \mathbb{R}^{d\times m}$ be a matrix with singular value decomposition (SVD) given by:

$$Z = U\Sigma V^\top = \sum_{i=1}^{n} \sigma_i u_i v_i^\top, \tag{19}$$

where $n = \min\{m, d\}$, $U = [u_1, \ldots, u_d] \in \mathbb{R}^{d\times d}$ and $V = [v_1, \ldots, v_m] \in \mathbb{R}^{m\times m}$ are orthogonal matrices, and $\Sigma \in \mathbb{R}^{d\times m}$ is a diagonal matrix with singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$ on its diagonal.

To perform rank-r approximation, we keep the top-r singular values in Σ and corresponding r columns of U and V unchanged, with values in other positions turned into zero. Formally, the approximation is given by:

$$Z_r = \sum_{i=1}^{r} \sigma_i u_i v_i^\top. \tag{20}$$

Then the coding length can be derived by the kept real values that we need to store:

$$L_r = b \cdot (r + d \cdot r + m \cdot r) = b \cdot r \cdot (d + m + 1). \tag{21}$$

□

The optimality of this approximation is guaranteed by Theorem 2.

Theorem 2 (Eckart-Young-Mirsky optimal low-rank approximation theorem)**.** *Under the Frobenius norm, the low-rank approximation obtained by retaining the top-r singular values and their corresponding singular vectors yields the optimal rank-r approximation Z_r, and the error is given by:*

$$||Z_r - Z||_F^2 = \sum_{i=r+1}^{n} \sigma_i^2, \tag{22}$$

where $n = \min\{m, d\}$, $\|\cdot\|_F$ represents Frobenius norm, and σ_i is the i_{th} largest singular value.

Proof. We perform singular value decomposition following Eq. (19). For any rank-r matrix $A \in \mathbb{R}^{d\times m}$, we have:

$$\|Z - A\|_F^2 = \|U\Sigma V^\top - A\|_F^2 = \|U^\top(U\Sigma V^\top - A)V\|_F^2 = \|\Sigma - U^\top AV\|_F^2, \tag{23}$$

where we used the unitary invariance of the Frobenius norm.

Let $B = U^\top AV \in \mathbb{R}^{d\times m}$. Since $rank(B) \leq r$, at most r columns of B are linearly independent. The optimal B that minimizes $\|\Sigma - B\|_F^2$ should match Σ on as many entries as possible. Then the best rank-r approximation is achieved by setting:

$$B_r = \text{diag}(\sigma_1, \ldots, \sigma_r, 0, \ldots, 0), \tag{24}$$

which corresponds to:

$$Z_r = UB_rV^\top = \sum_{i=1}^{r} \sigma_i u_i v_i^\top. \tag{25}$$

The approximation error is then:

$$\|Z - Z_r\|_F^2 = \|\Sigma - B_r\|_F^2 = \sum_{i=r+1}^{n} \sigma_i^2. \tag{26}$$

To show this is indeed optimal, consider any other rank-r matrix A' with corresponding $B' = U^\top A'V$. Since $rank(B') \leq r$, B' must have at least $n - r$ zero singular values. The minimal Frobenius norm difference is achieved when B' preserves the largest r singular values of Σ and sets the remaining to zero. Therefore, Z_r is optimal.

Thus, the optimal rank-r approximation under Frobenius norm is obtained by retaining the top-r singular values and vectors, with approximation error given by the sum of squares of the discarded singular values. □

2.3. Inequality of LER's Lower bound

Proposition 4 (Lower Bound of Local Effective Rank)**.** *The lower bound of LER is inversely proportional to the squared Frobenius norm of the similarity matrix of local representations. Specifically, for a matrix $Z \in \mathbb{R}^{d\times m}$ composed of l_2-normalized representation vectors, the following inequality holds:*

$$\mathrm{LER}(Z) \geq \frac{m^2}{\|Z^\top Z\|_F^2}, \tag{27}$$

Equality holds when either all singular values are equal (isotropic case) or there is only one non-zero singular value (rank-1 case).

Proof. We denote the studied vector as the singular value spectrum $\boldsymbol{\sigma} = [\sigma_1, \sigma_2, \cdots, \sigma_n], \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$, and define the function of interest as $\mathrm{LER}(Z) = \mathrm{LER}(\boldsymbol{\sigma}) = \exp(\mathrm{H}(\boldsymbol{\sigma}))$, where $\mathrm{H}(\boldsymbol{\sigma}) = -\sum_{i=1}^n \frac{\sigma_i}{S} \log \frac{\sigma_i}{S}$, $S = \sum_{i=1}^n \sigma_i$. Let $\boldsymbol{\sigma}_2 = [\sigma_1^2, \sigma_2^2, \ldots, \sigma_m^2]$. We first prove the inequality:

$$\mathrm{LER}(\boldsymbol{\sigma}_2) \geq \frac{m^2}{\|Z^\top Z\|_F^2}. \tag{28}$$

Note that $\boldsymbol{\sigma}_2$ corresponds to the singular values of the similarity matrix $Z^\top Z$. We have:

$$\sum_{i=1}^m \sigma_i^2 = \mathrm{tr}(Z^\top Z) = m, \tag{29}$$

where $\mathrm{tr}(\cdot)$ denotes the matrix trace.

Since $-\log x$ is a convex function, by Jensen's inequality:

$$\mathrm{LER}(\boldsymbol{\sigma}_2) = \exp\left(-\sum_{i=1}^m \frac{\sigma_i^2}{m} \log \frac{\sigma_i^2}{m}\right) \geq \exp\left(-\log \sum_{i=1}^m \frac{\sigma_i^4}{m^2}\right) = \frac{m^2}{\sum_{i=1}^m \sigma_i^4} = \frac{m^2}{\|Z^\top Z\|_F^2}.$$

Next, we prove:

$$\mathrm{LER}(\boldsymbol{\sigma}) \geq \mathrm{LER}(\boldsymbol{\sigma}_2). \tag{30}$$

Assume Z has m singular values $\boldsymbol{\sigma} = [\sigma_1, \sigma_2, \ldots, \sigma_m]$. If $m > d$, we pad with zeros to maintain $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_m \geq 0$. Let $S = \sum_{i=1}^m \sigma_i$. We establish:

$$\begin{cases} \forall i \text{ where } \frac{\sigma_i}{m} \geq \frac{1}{S}, & \sum_{k=1}^i \frac{\sigma_k^2}{m} \geq \sum_{k=1}^i \frac{\sigma_k}{S}; \\ \forall i \neq m \text{ where } \frac{\sigma_i}{m} < \frac{1}{S}, & \sum_{k=1}^i \frac{\sigma_k^2}{m} = 1 - \sum_{k=i+1}^m \frac{\sigma_k^2}{m} \geq 1 - \sum_{k=i+1}^m \frac{\sigma_k}{S} = \sum_{k=1}^i \frac{\sigma_k}{S}. \end{cases} \tag{31}$$

This implies the majorization relationship:

$$\left[\frac{\sigma_1}{S}, \frac{\sigma_2}{S}, \ldots, \frac{\sigma_m}{S}\right] \preceq \left[\frac{\sigma_1^2}{m}, \frac{\sigma_2^2}{m}, \ldots, \frac{\sigma_m^2}{m}\right]. \tag{32}$$

We then examine the Schur-concavity of the entropy function $h(\boldsymbol{x}) = -\sum_{i=1}^n x_i \log x_i$:

$$\frac{\partial h}{\partial x_i} = -1 - \log x_i \Rightarrow \forall 1 \leq i, j \leq n,\ (x_i - x_j)\left(\frac{\partial h}{\partial x_i} - \frac{\partial h}{\partial x_j}\right) \leq 0, \tag{33}$$

which means $h(\boldsymbol{x})$ is Schur-concave. Therefore:

$$\mathrm{LER}(Z) = \mathrm{LER}(\boldsymbol{\sigma}) = \exp\left(h\left(\left[\frac{\sigma_1}{S}, \ldots, \frac{\sigma_m}{S}\right]\right)\right) \geq \exp\left(h\left(\left[\frac{\sigma_1^2}{m}, \ldots, \frac{\sigma_m^2}{m}\right]\right)\right) = \mathrm{LER}(\boldsymbol{\sigma}_2). \tag{34}$$

Combining Eq. (28) and Eq. (30) completes the proof of the main inequality.

Equality conditions:

- When all singular values are equal (isotropic case), both sides equal m.
- When only one singular value is non-zero (rank-1 case), both sides equal 1.

□

3. Symbols and pseudo codes

We present the details of our symbol definitions and method designs as follows.

3.1. Local Dimension Enhancement Learning

We formally describe the pre-training process of our LoDE framework in Algorithm 1 for a better understanding, and provide a symbol table for reference in Table 1.

Algorithm 1 The pre-training process of LoDE

Input: A cropped 3D skeleton sequence $X \in \mathbb{R}^{T_0 \times V \times 3}$.
Parameter: Embedding projector Embed$(\cdot)$; Random masking operation $\mathcal{M}(\cdot)$; Encoder $\mathcal{E}(\cdot)$; Decoder $\mathcal{D}(\cdot)$; Sequence-scale, motion unit-scale, frame-scale predicting projectors $\mathcal{P}_{\text{seq}}(\cdot)$, $\mathcal{P}_{\text{motion}}(\cdot)$, $\mathcal{P}_{\text{frame}}(\cdot)$; Sequence-scale and motion unit-scale vector quantization modules $\text{VQ}_{\text{seq}}(\cdot)$, $\text{VQ}_{\text{motion}}(\cdot)$; Encoder spatial and temporal embeddings E^s_{enc}, E^t_{enc}; Decoder spatial and temporal embeddings E^s_{dec}, E^t_{dec}; Sequence-scale, frame-scale and motion unit-scale objective weights λ_s, λ_f, and λ_m; LER regularization weight λ_r.
Begin

Motionize X into X';
$E \leftarrow \text{Embed}(X') + E^s_{\text{enc}} + E^t_{\text{enc}}$;
// Masked Predicting.
$E_m, M \leftarrow \mathcal{M}(E)$; *// Return visible embeddings and binary mask.*
$Z_m^{\text{motion}} \leftarrow \mathcal{E}(E_m)$;
$Z_m^{\text{seq}} \leftarrow \mathcal{P}_{\text{seq}}(\text{GAP}(Z_m^{\text{motion}}))$; *// Sequence-scale prediction.*
Pad Z_m^{motion} with Z_m^{seq} to get Z_{pad}; *// Sequence-scale conditioning.*
$Z_{\text{pad}} \leftarrow Z_{\text{pad}} + E^s_{\text{dec}} + E^t_{\text{dec}}$;
$Z_{\text{pred}} \leftarrow \mathcal{D}(Z_{\text{pad}})$;
$\hat{Y}^{\text{motion}} \leftarrow \mathcal{P}_{\text{motion}}(Z_{\text{pred}})$; *// Motion unit-scale prediction.*
$\hat{Y}^{\text{frame}} \leftarrow \mathcal{P}_{\text{frame}}(Z_{\text{pred}})$; *// Frame-scale prediction.*
// Target preparation.
// Transformation $\mathcal{T}(\cdot)$, Compute Frame-scale target.
$Y^{\text{frame}}_{1,:,:} \leftarrow 0$;
for $t = 2, 3, ..., T_0$ **do**
  $Y^{\text{frame}}_{t,:,:} \leftarrow X_{t,:,:} - X_{t-1,:,:}$;
end for
$Z_{\text{motion}} \leftarrow \mathcal{E}(E)$;
$Y^{\text{motion}} \leftarrow \text{VQ}_{\text{motion}}(Z^{\text{motion}}).\text{detach}()$; *// Motion unit-scale target. Stop gradient.*
$Y^{\text{seq}} \leftarrow \text{VQ}_{\text{seq}}(\text{GAP}(Z^{\text{motion}})).\text{detach}()$; *// Sequence-scale target. Stop gradient.*
// Compute losses.
$\mathcal{L}_s \leftarrow 1 - \text{cosine\_similarity}(\hat{Y}^{\text{seq}}, Y^{\text{seq}})$;
$\mathcal{L}_m \leftarrow 1 - \text{cosine\_similarity}(\hat{Y}^{\text{motion}}, Y^{\text{motion}}).\text{mean}()$;
$\mathcal{L}_f \leftarrow ||M \odot (\hat{Y}^{\text{frame}} - Y^{\text{frame}})||_2^2 / M.\text{mean}()$;
$\mathcal{L}_r \leftarrow \dfrac{1}{\text{LER}(Z_m^{\text{motion}})}$;
$\mathcal{L} \leftarrow \lambda_m * \mathcal{L}_m + \lambda_s * \mathcal{L}_s + \lambda_f * \mathcal{L}_f + \lambda_r * \mathcal{L}_r$;
Update model parameters by minimizing $\mathcal{L}$;

End

Table 1. Table of Symbols

Symbol	Description
X	Original input skeleton sequence ($\mathbb{R}^{T_0 \times V \times 3}$).

Continued on next page

Table 1 – Continued from previous page

Symbol	Description
T_0	Number of frames in the original sequence.
V	Number of joints per frame.
l	Temporal length (in frames) of each motion unit.
X'	Skeleton sequence segmented into motion units.
T	Total number of motion units in the sequence.
C	Feature dimension of the token embeddings.
E	Flattened sequence of embedded tokens.
r	Ratio of tokens to be masked.
K	Number of visible tokens after masking.
E_{m}	The set of visible tokens from the masked view.
$\mathcal{E}(\cdot)$	The Siamese Transformer encoder function.
Z^{motion}	Motion unit representations from the unmasked view.
$Z_{\text{m}}^{\text{motion}}$	Motion unit representations from the masked view.
$Z_{\text{m}}^{\text{seq}}$	Sequence-scale representation from the masked view.
$\text{GAP}(\cdot)$	Global Average Pooling operation.
$\mathcal{P}_{\text{seq}}(\cdot)$	Projector for generating sequence-scale representations.
Z_{pad}	Input tensor for the Transformer decoder.
$\mathcal{D}(\cdot)$	The Transformer decoder function.
$\hat{Z}_{\text{pad}}$	Reconstructed motion unit representations from the decoder.
$\text{VQ}_{\text{motion}}(\cdot)$	Vector Quantization module for motion units.
Y^{motion}	Target discrete representations for motion units from VQ.
$\text{VQ}_{\text{seq}}(\cdot)$	Vector Quantization module for the sequence representation.
Y^{seq}	Target discrete representation for the sequence from VQ.
$\mathcal{P}_{\text{motion}}(\cdot)$	Linear projector for the motion unit-scale objective.
$\mathcal{P}_{\text{frame}}(\cdot)$	Projector for the frame-scale reconstruction objective.
$\mathcal{T}(\cdot)$	Linear transformation on the input sequence X to prevent shortcuts.
$\mathcal{L}_{\text{m}}$	Loss function for the motion unit-scale learning objective.
$\mathcal{L}_{\text{s}}$	Loss function for the sequence-scale learning objective.
$\mathcal{L}_{\text{f}}$	Loss function for the frame-scale reconstruction objective.
$\text{LER}(\cdot)$	Function to calculate the Local Effective Rank.
$\mathcal{L}_{\text{r}}$	The LER regularization loss term.
$\mathcal{L}$	The overall combined loss function for the model.
$\lambda_{\text{m}}, \lambda_{\text{s}}, \lambda_{\text{f}}, \lambda_{\text{r}}$	Weight coefficients for each component of the overall loss.

3.2. Vector quantization modules

We perform vector quantization with codebooks for discrete tokenization modules. We update the codebooks by the EMA strategy simultaneously with the vector quantization process, which are integrated into one operation $\text{VQ}_*(\cdot), * \in \{\text{seq}, \text{motion}\}$. The momentum coefficient in EMA is set to 0.999 in our implementation. The operating algorithm is shown in Algorithm 2.

Algorithm 2 The operating process of VQ modules.

Input: A latent feature $Z \in \mathbb{R}^C$, C is the feature dimension.
Parameter: Codebook $B \in \mathbb{R}^{K \times C}$, K is the length of the codebook; α is the momentum coefficient in EMA.
Begin

for $k = 1, 2, ..., K$ **do**
  // Compute cosine similarity with each quantized vector in the codebook.
  $sim_k \leftarrow \text{cosine\_similarity}(Z, B_k)$;
end for
$id \leftarrow \arg\max(sim)$; *// Quantized by the vector with the highest similarity.*
for $k = 1, 2, ..., K$ **do**
  // Update the codebook.
  $B_k \leftarrow \alpha * B_k + (1 - \alpha) * sim_k * Z$;
  $B_k \leftarrow B_k / ||B_k||_2$;
end for
// Return the quantized vector.
return B_{id}

End

4. More ablation study

4.1. Codebook size

We examine the codebook size in the vector quantization module. As shown in Table 2, a small codebook cannot fully capture the semantics-relevant information in representations. On the other hand, an excessively large codebook size falls short in compacting the target representations, which can bring too much noise and semantics-irrelevant information. Therefore, a medium codebook size is optimal, which we set as 256.

Table 2. Ablation study on Codebook size.

Codebook Size	mAP@θ (%)		
	0.1	0.3	0.5
128	86.5	85.1	81.5
192	86.2	85.6	<u>82.7</u>
256	**87.6**	<u>86.4</u>	**83.1**
320	<u>87.5</u>	**86.6**	82.0
384	85.9	85.3	80.9

4.2. EMA momentum coefficient

We examine the momentum coefficient of EMA update in the vector quantization module. As shown in Table 3, a high coefficient helps keep the semantics captured at the early pretraining stage and improves the performance. As a result, we set the momentum coefficient α as 0.999 by default.

4.3. Masking strategies

We explore motion-guided [5] and attention-guided [1] masking, reaching 78.7% and 80.7% mAP@0.5 on PKU-I xsub by linear evaluation, lower than random masking (83.1%). Because our learning framework performs representation alignment in sequence-scale and motion unit-scale learning, advanced masking strategies that mask semantics-rich areas degrade performance instead. Therefore, we adopt the uniformly random strategy to perform masking.

5. More qualitative results of TAS

This section presents more qualitative results of the TAS results in Figure 3 to Figure 6. We use the model trained under the linear evaluation protocol. The predicted and the ground truth action segments of cropped sequences from PKUMMD I [4]

Table 3. Ablation study on EMA momentum coefficient.

Coefficient	mAP@θ (%)		
	0.1	0.3	0.5
0.9	84.8	84.0	80.7
0.99	86.1	85.0	81.1
0.999	**87.6**	**86.4**	**83.1**

Table 4. Ablation study on masking strategies.

Masking Strategy	mAP@θ (%)		
	0.1	0.3	0.5
motion-guided [5]	83.7	82.4	78.7
attention-guided [1]	84.6	84.1	80.7
random	**87.6**	**86.4**	**83.1**

are visualized. Note that *gray* regions represent the segments where no action occurs, and different actions may be mapped to the same color. Generally, compared to the frame-scale learning method, our method achieves higher segmentation quality with more accurate action boundaries and better temporal coherence.

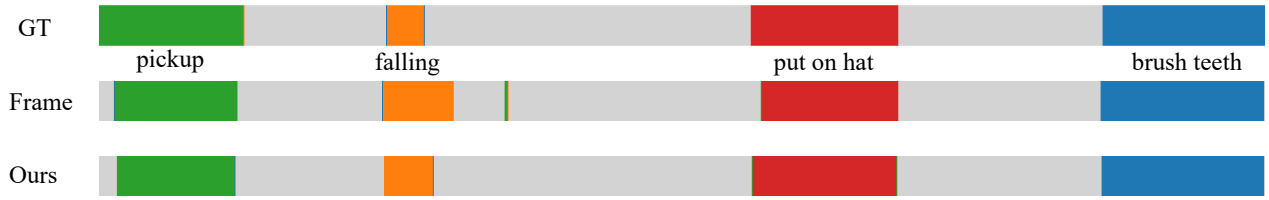

Figure 3. Frame 100 to 1000 from "298-M"

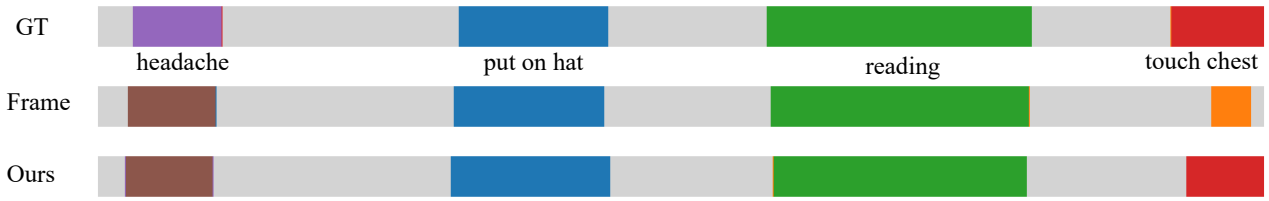

Figure 4. Frame 300 to 1200 from "302-L"

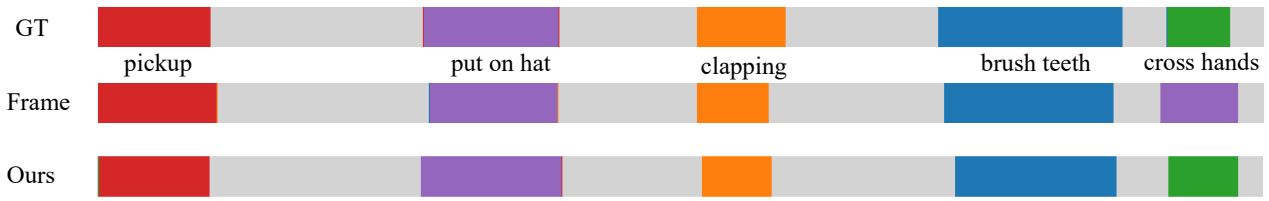

Figure 5. Frame 150 to 1050 from "304-L"

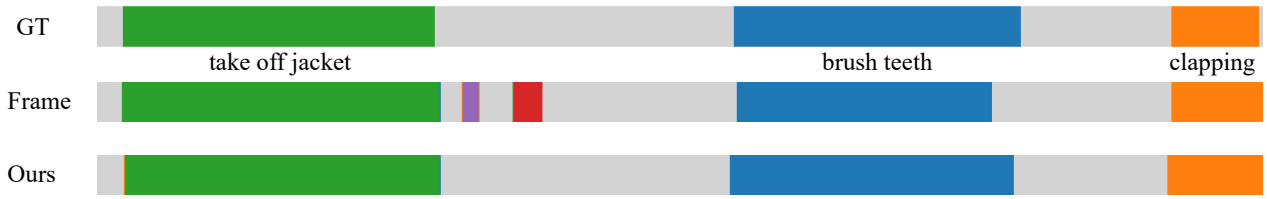

Figure 6. Frame 550 to 1450 from "318-M"

References

[1] Mohamed Abdelfattah, Mariam Hassan, and Alexandre Alahi. MaskCLR: Attention-guided contrastive learning for robust action representation learning. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, 2024.

[2] Laurent Amsaleg, Oussama Chelly, Teddy Furon, Stéphane Girard, Michael E Houle, Ken-ichi Kawarabayashi, and Michael Nett. Extreme-value-theoretic estimation of local intrinsic dimensionality. *Data Mining and Knowledge Discovery*, 32(6):1768–1805, 2018.

[3] Hanxun Huang, Ricardo JGB Campello, Sarah Monazam Erfani, Xingjun Ma, Michael E Houle, and James Bailey. LDReg: Local dimensionality regularized self-supervised learning. In *Proc. International Conference on Learning Representations*, Vienna, Austria, 2024.

[4] Chunhui Liu, Yueyu Hu, Yanghao Li, Sijie Song, and Jiaying Liu. PKU-MMD: A large scale benchmark for continuous multi-modal human action understanding. In *Proc. ACM International Conference on Multimedia Workshop*, Silicon Valley, CA, 2017.

[5] Yunyao Mao, Jiajun Deng, Wengang Zhou, Yao Fang, Wanli Ouyang, and Houqiang Li. Masked motion predictors are strong 3D action representation learners. In *Proc. International Conference on Computer Vision*, Paris, France, 2023.

[6] Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *Proc. European Signal Processing Conference*, Poznan, Poland, 2007.