A. Detailed Description of Stride Attentions

In general, most of the time complexity of transformers is highly related to the attention operation. We offer stride attentions for efficient correlation learning between discretized tokens. In this section, we describe details of the proposed joint stride attention and temporal stride attention.

A.1. Joint Stride Attention

The number of *pose* tokens (P) can grow dynamically based on the number of joints (R) and people appeared in a scene. Since this is directly related to the amount of computation required for attention, we propose joint stride attention dividing *pose* tokens into several sliding windows. Algorithm A1 describes detailed operation of the proposed joint stride attention. According to the notations used in the Algorithm A1, we can organize comparisons of time complexity between full attention and joint stride attention as described in Table A1.

Table A1: Complexity comparisons between full attention and joint stride attention

Method	Complexity
Full attention	$O(T^2R^2)$
Joint stride attention	$O(T^2wnd^2)$

In joint stride attention, we decompose *pose* tokens using sliding window (wnd) with a $stride$ having halved size of wnd. If *pose* tokens $\mathrm{P} \in \mathbb{R}^{4C\times T\times R}$ are fed to full attention, the time complexity per layer becomes $O(T^2R^2)$ and the equation is as follows:

$$\mathbf{O} = [\mathrm{P}||\mathrm{M}_{pose}||\mathrm{M}^{'}_{CLS}] \tag{A1}$$

$$\begin{aligned}&\mathrm{FullAttention}(\mathbf{OW}_q, \mathbf{OW}_k, \mathbf{OW}_v)\\ &= \sum_t^T \sum_r^R \mathrm{softmax}(\frac{\mathbf{OW}_q^{t,r}\mathbf{OW}_k^{t,r}}{\sqrt{d_h}})\mathbf{OW}_v^{t,r}\end{aligned} \tag{A2}$$

However, when *pose* tokens are fed to joint stride attention, the time complexity per layer becomes $O(T^2wnd^2)$ and the equation is as follows:

$$\begin{aligned}&\mathrm{JointStrideAttention}(\mathbf{OW}_q, \tilde{\mathbf{O}}\mathbf{W}_k, \tilde{\mathbf{O}}\mathbf{W}_v)\\ &= \sum_t^T \sum_w^{wnd} \mathrm{softmax}(\frac{\mathbf{OW}_q^{t,w}\tilde{\mathbf{O}}\mathbf{W}_k^{t,w}}{\sqrt{d_h}})\tilde{\mathbf{O}}\mathbf{W}_v^{t,w}\end{aligned} \tag{A3}$$

where $\mathbf{O}$ and $\tilde{\mathbf{O}}$ from Algorithm A1. In Eq. A3, since wnd is always less than R, the overall time complexity of joint stride attention is smaller than full attention.

Algorithm A1 Joint stride attention

Input: *Pose* tokens P, Memorized *CLS* modal token $\mathrm{M}^{'}_{CLS}$, *Pose* modal token M_{pose}, window size wnd, *query* weight $\mathbf{W}_q$, *key* weight $\mathbf{W}_k$, *value* weight $\mathbf{W}_v$

$stride \leftarrow \lfloor wnd/2 \rfloor$
$\hat{\mathbf{P}}_q \leftarrow \varnothing$ ▷ *query* set
$\hat{\mathbf{P}}_{kv} \leftarrow \varnothing$ ▷ *key* and *value* set

$i \leftarrow 0$
while $i < (R - wnd)$ **do** ▷ Split P into query set $\hat{\mathbf{P}}_q$
  $\hat{\mathrm{P}} \leftarrow \mathrm{P}[:, :, i : i + wnd]$ ▷ $\mathrm{P} \in \mathbb{R}^{4C\times T\times R}$
  $\hat{\mathbf{P}}_q \leftarrow \hat{\mathbf{P}}_q \cup \hat{\mathrm{P}}$ ▷ $\hat{\mathrm{P}} \in \mathbb{R}^{4C\times T\times wnd}$
  $i \leftarrow i + stride$
  if $i > (R - wnd)$ **then**
    $i \leftarrow i - wnd$ ▷ To assure *query* covers entire tokens
  end if
end while

$i \leftarrow stride$
while $i < (R - wnd)$ **do**
  ▷ Split P into *key* and *value* set $\hat{\mathbf{P}}_{kv}$
  $\hat{\mathrm{P}} \leftarrow \mathrm{P}[:, :, i : i + wnd]$ ▷ $\mathrm{P} \in \mathbb{R}^{4C\times T\times R}$
  $\hat{\mathbf{P}}_{kv} \leftarrow \hat{\mathbf{P}}_{kv} \cup \hat{\mathrm{P}}$ ▷ $\hat{\mathrm{P}} \in \mathbb{R}^{4C\times T\times wnd}$
  $i \leftarrow i + stride$
end while

$D_q \leftarrow |\hat{\mathbf{P}}_q|$ ▷ $\hat{\mathbf{P}}_q \in \mathbb{R}^{D_q\times 4C\times T\times wnd}$
$D_{kv} \leftarrow |\hat{\mathbf{P}}_{kv}|$ ▷ $\hat{\mathbf{P}}_{kv} \in \mathbb{R}^{D_{kv}\times 4C\times T\times wnd}$

▷ $\mathrm{M}_{pose}, \mathrm{M}^{'}_{CLS} \in \mathbb{R}^{4C\times T\times 1}$
$\mathbf{O} \leftarrow [\hat{\mathbf{P}}_q||\mathrm{expand}(\mathrm{M}_{pose}, D_q)||\mathrm{expand}(\mathrm{M}^{'}_{CLS}, D_q)]$
$\tilde{\mathbf{O}} \leftarrow [\hat{\mathbf{P}}_{kv}||\mathrm{expand}(\mathrm{M}_{pose}, D_{kv})||\mathrm{expand}(\mathrm{M}^{'}_{CLS}, D_{kv})]$
$\mathbf{O} \leftarrow \mathbf{O} + \mathrm{MSA}(\mathbf{OW}_q, \tilde{\mathbf{O}}\mathbf{W}_k, \tilde{\mathbf{O}}\mathbf{W}_v)$
$\mathbf{O} \leftarrow \mathbf{O} + \mathrm{FFN}(\mathrm{LN}(\mathbf{O}))$

A.2. Temporal Stride Attention

Temporal stride attention is proposed to capture temporal changes between each sequential frame and joint. The overall procedure is described in Algorithm A2. The number of input tokens D_n is sum of the number of *RGB*, *pose* and cross modal tokens. To decompose these tokens into small temporal windows, we apply sliding windows along with the temporal dimension. The complexity comparison result between full attention and temporal stride attention is described in Table A2.

In the case of full attention, the time complexity of attention against the concatenated tokens $\mathbf{N} \in \mathbb{R}^{4C\times T\times D_n}$ becomes $O(T^2D_n^2)$ and the equation is as follows:

Table A2: Complexity comparisons between full attention and temporal stride attention

Method	Complexity
Full attention	$O(T^2 D_n^2)$
Temporal stride attention	$O(wnd^2 D_n^2)$

$$\mathbf{N} = [\mathrm{Z}||\mathrm{P}||\mathrm{M}_{RGB}||\mathrm{M}_{pose}||\mathrm{M}_{CLS}] \tag{A4}$$

$$\begin{aligned}&\mathrm{FullAttention}(\mathbf{NW}_q, \mathbf{NW}_k, \mathbf{NW}_v)\\ &= \sum_t^T \sum_n^{D_n} \mathrm{softmax}(\frac{\mathbf{NW}_q^{t,n} \mathbf{NW}_k^{t,n}}{\sqrt{d_h}}) \mathbf{NW}_v^{t,n}\end{aligned} \tag{A5}$$

On the contrary, when the concatenated tokens are decomposed into several sliding windows, the time complexity per layer becomes $O(wnd^2 D_n^2)$ and the equation is as follows:

$$\begin{aligned}&\mathrm{TemporalStrideAttention}(\hat{\mathbf{N}}_q \mathbf{W}_q, \hat{\mathbf{N}}_{kv} \mathbf{W}_k, \hat{\mathbf{N}}_{kv} \mathbf{W}_v)\\ &= \sum_w^{wnd} \sum_n^{D_n} \mathrm{softmax}(\frac{\hat{\mathbf{N}}_q \mathbf{W}_q^{w,n} \hat{\mathbf{N}}_{kv} \mathbf{W}_k^{w,n}}{\sqrt{d_h}}) \hat{\mathbf{N}}_{kv} \mathbf{W}_v^{w,n}\end{aligned} \tag{A6}$$

where $\hat{\mathbf{N}}_q$ and $\hat{\mathbf{N}}_{kv}$ from Algorithm A2. In this case, because wnd is always less than T, the overall time complexity of temporal stride attention is smaller than full attention.

B. Detailed Description of 3D Deformable Attention

In this study, we proposed the 3D deformable attention to adaptively capture not only long-term temporal relations but also spatial relations simultaneously. Inspired by Xia *et al*. [1], we rebuilt a deformable attention transformer (DAT) applicable with various video tasks including action recognition. Our proposed method finds discriminative tokens across 3D space while the DAT leverages only 2D space tokens. Details are described in Algorithm A3.

C. On the fly frames in test phase.

The proposed method showed a good performance on several benchmarks using the suggested modules for capturing temporal changes. According to the Fig. 5 (b) in the paper, it was observed that the performance has a high relevance for the number of input frames in training phase. For that reason, we assumed that if the model well captures spatiotemporal relations on dense frame condition in training,

Algorithm A2 Temporal stride attention

Input: *RGB* tokens Z, *Pose* tokens P, *CLS* modal token M_{CLS}, *RGB* modal token M_{RGB}, *Pose* modal token M_{pose}, window size wnd, *query* weight $\mathbf{W}_q$, *key* weight $\mathbf{W}_k$, *value* weight $\mathbf{W}_v$

$stride \leftarrow \lfloor wnd/2 \rfloor$
$\hat{\mathbf{N}}_q \leftarrow \varnothing$ ▷ *query* set
$\hat{\mathbf{N}}_{kv} \leftarrow \varnothing$ ▷ *key* and *value* set

$\mathrm{N} \leftarrow [\mathrm{Z}||\mathrm{P}||\mathrm{M}_{RGB}||\mathrm{M}_{pose}||\mathrm{M}_{CLS}]$
$D_n \leftarrow (\frac{H}{8} \times \frac{W}{8} + R + 3)$ ▷ The number of tokens

$i \leftarrow 0$
while $i < (T - wnd)$ **do** ▷ Split N into *query* set $\hat{\mathbf{N}}_q$
  $\hat{\mathrm{N}} \leftarrow \mathrm{N}[:, i : i + wnd, :]$ ▷ $\mathrm{N} \in \mathbb{R}^{4C \times T \times D_n}$
  $\hat{\mathbf{N}}_q \leftarrow \hat{\mathbf{N}}_q \cup \hat{\mathrm{N}}$ ▷ $\hat{\mathrm{N}} \in \mathbb{R}^{4C \times wnd \times D_n}$
  $i \leftarrow i + stride$
  if $i > (T - wnd)$ **then**
    $i \leftarrow i - wnd$ ▷ To assure *query* covers entire tokens
  end if
end while

$i \leftarrow stride$
while $i < (T - wnd)$ **do**
  ▷ Split N into *key* and *value* set $\hat{\mathbf{N}}_{kv}$
  $\hat{\mathrm{N}} \leftarrow \mathrm{N}[:, i : i + wnd, :]$ ▷ $\mathrm{N} \in \mathbb{R}^{4C \times T \times D_n}$
  $\hat{\mathbf{N}}_{kv} \leftarrow \hat{\mathbf{N}}_{kv} \cup \hat{\mathrm{N}}$ ▷ $\hat{\mathrm{N}} \in \mathbb{R}^{4C \times wnd \times D_n}$
  $i \leftarrow i + stride$
end while

$\mathbf{N} \leftarrow \hat{\mathbf{N}}_q + \mathrm{MSA}(\hat{\mathbf{N}}_q \mathbf{W}_q, \hat{\mathbf{N}}_{kv} \mathbf{W}_k, \hat{\mathbf{N}}_{kv} \mathbf{W}_v)$
$\mathbf{N} \leftarrow \mathbf{N} + \mathrm{FFN}(\mathrm{LN}(\mathbf{N}))$

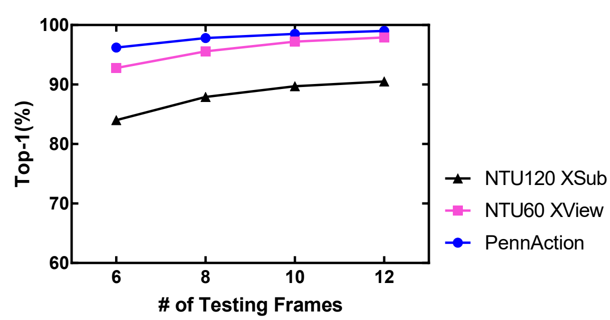


Figure A1: Ablation study with different numbers of frames during test phase against model trained with 12 frames.

then it will be able to defense degradation of performance on sparse test frames. In practical application, some models may have to be run in sparse frames due to environmental

Algorithm A3 3D deformable attention

Input: *RGB* tokens Z, *CLS* modal token M_{CLS}, *RGB* modal token M_{RGB}, *query* weight $\mathbf{W}_q$, *key* weight $\mathbf{W}_k$, *value* weight $\mathbf{W}_v$, kernel size k, 3D conv block f_{off}, bilinear sampling function g, trainable parameter ω
function $\mathrm{3DTS}(\mathrm{Z};\omega)$
  $\mathrm{Z} \leftarrow \mathrm{reshape}(\mathrm{Z})$ ▷ $\mathrm{Z} \in \mathbb{R}^{4C\times T\times \frac{H}{8}\times\frac{W}{8}}$
  $\Delta p \leftarrow \tanh(f_{off}(\mathrm{Z};\omega))$ ▷ $\Delta p \in \mathbb{R}^{3\times\tilde{T}\times\tilde{H}\times\tilde{W}}$
  $p \leftarrow$ reference points from 3D grid ▷ $p \in \mathbb{R}^{3\times\tilde{T}\times\tilde{H}\times\tilde{W}}$
  Initialize $\tilde{\mathrm{Z}} \in \mathbb{R}^{4C\times T\times \frac{H}{8}\times\frac{W}{8}}$
  for $(p_x, p_y, p_z) \in p + \Delta p$ **do**
    $\tilde{\mathrm{z}} \leftarrow 0$
    for $(r_x, r_y, r_z) \in [\{1...\frac{W}{8}\}, \{1...\frac{H}{8}\}, \{1...T\}]$ **do**
      ▷ get spatiotemporal coordinates $r_{\{x,y,z\}}$
      $\phi \leftarrow g(p_x, r_x)g(p_y, r_y)g(p_z, r_z)$
      $\tilde{\mathrm{z}} \leftarrow \tilde{\mathrm{z}} + \phi\mathrm{Z}[:, r_z, r_y, r_x]$
    end for
    $\tilde{\mathrm{Z}}[:, p_z, p_y, p_x] \leftarrow \tilde{\mathrm{z}}$
  end for
  return $\mathrm{flat}(\tilde{\mathrm{Z}})$
end function
Initialize 3D conv. parameter (ω) with k
$\tilde{\mathrm{Z}} \leftarrow \mathrm{3DTS}(\mathrm{Z};\omega)$
$\mathbf{X}, \tilde{\mathbf{X}} \leftarrow [\mathrm{Z}||\mathrm{M}_{RGB}||\mathrm{M}_{CLS}], [\tilde{\mathrm{Z}}||\mathrm{M}_{RGB}||\mathrm{M}_{CLS}]$
$\mathbf{X} \leftarrow \mathbf{X} + \mathrm{MSA}(\mathbf{X}\mathbf{W}_q, \tilde{\mathbf{X}}\mathbf{W}_k, \tilde{\mathbf{X}}\mathbf{W}_v)$
$\mathbf{X} \leftarrow \mathbf{X} + \mathrm{FFN}(\mathrm{LN}(\mathbf{X}))$

limitations. We provided the evaluation results by diversifying the number of input frames in a model trained using 12 frames. The results verify the power of the spatiotemporal feature learning of the proposed method. In Fig. A1, the proposed method shows a uniform performance for various numbers of input frames. Therefore, the proposed method is robust in learning spatiotemporal features, even if the number of testing frames is sparse.

D. Additional Qualitative Results

D.1. Additional Joint Stride Attention Visualization

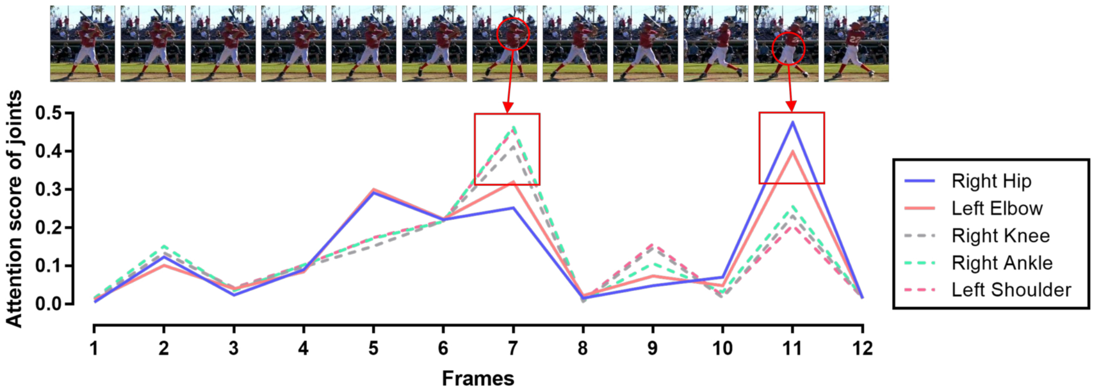
We present the result of analyzing the role of each *pose* token in cross-modal learning. The visualizations of each *pose* token for more diverse actions are shown in Fig. A2.

D.2. Additional 3D Deformable Attention Visualization

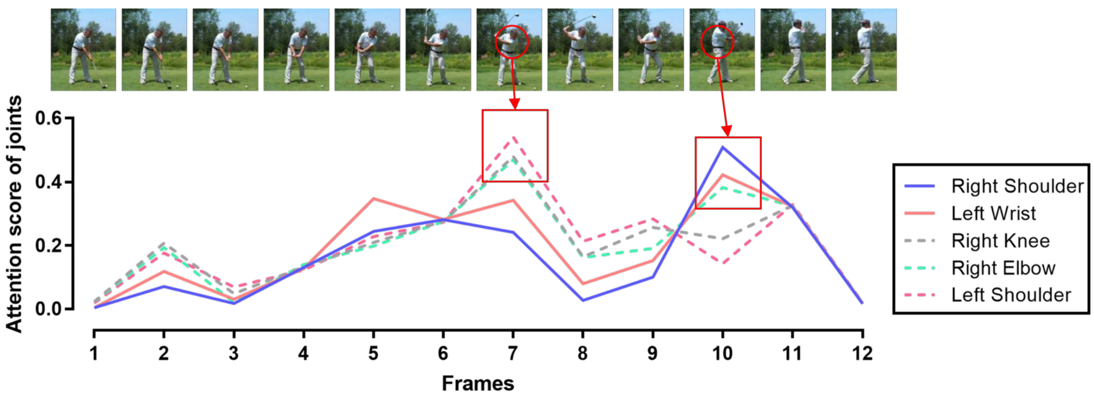
In this section, we provide additional visualization of 3D deformable attention for more diverse actions on each dataset. In the case of PennAction, attention is accurately appeared to the person who is the subject of the action even in complex backgrounds as shown in Fig. A3, and it can be seen that attention is occurring intensively in the frame representing the action. In terms of FineGYM, it consists of fine-grained gymnastic frames with dynamic camera moving. Our proposed 3D deformable attention accurately tracks gymnasts performing dynamic movements as shown in Fig. A4, and clearly understands the differences in each fine-grained actions, even for actions using the same equipment but with different labels. In the case of NTU120, which has a relatively simple background, the proposed method accurately finds key elements for various actions. The interaction between two people is also well tracked, especially in the case of the '*pick up*' action label, the actor's action is important, but the fact that there is a dropped object may be more important to accurately classify this action as shown in Fig. A5.

References

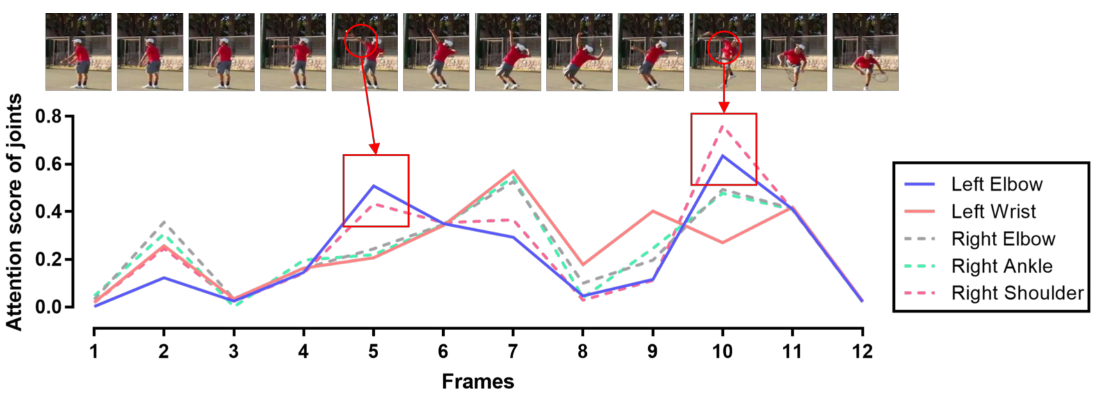
[1] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Vision transformer with deformable attention. In *CVPR*, pages 4794–4803, 2022. 2



(a) baseball swing



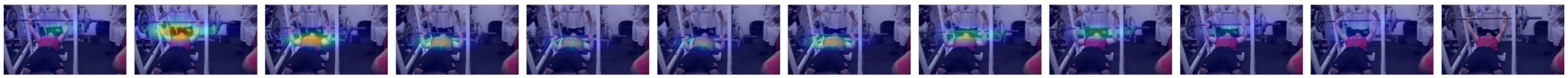
(b) golf swing



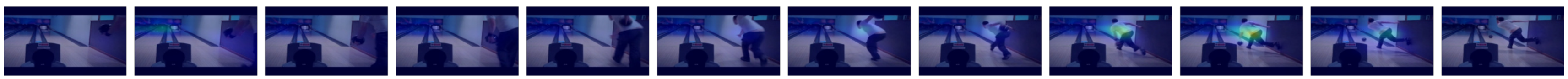
(c) tennis serve

Figure A2: Visualization of joint stride attention on PennAction

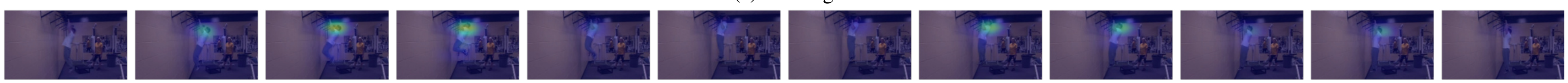
(a) baseball swing



(b) bench press



(c) bowling

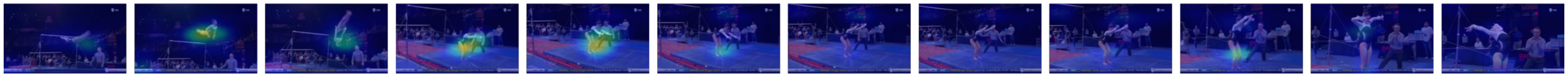


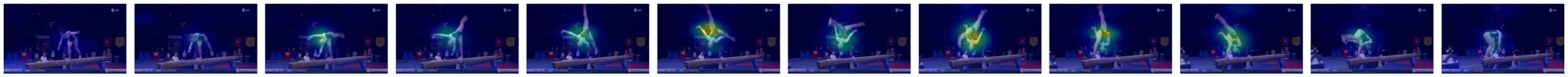
(d) pull ups

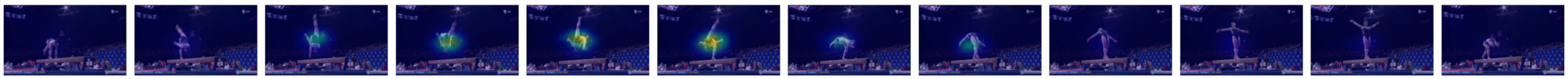
(e) sit ups

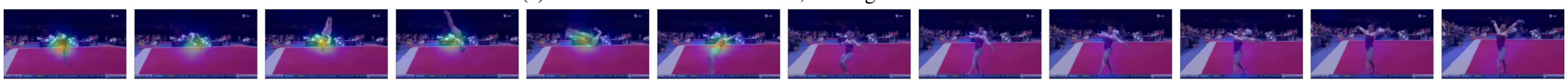
(f) squats

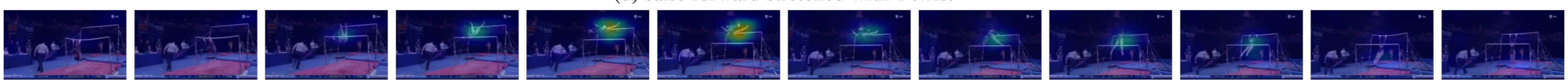
Figure A3: Visualization of 3D deformable attention on PennAction

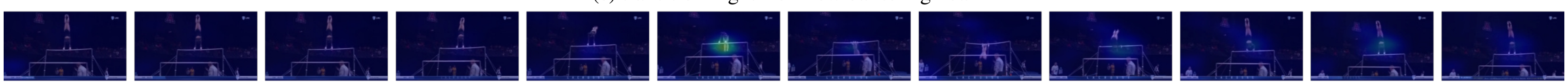

(a) (swing forward) double salto backward stretched

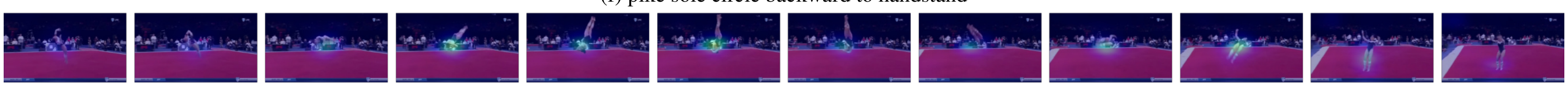

(b) free aerial cartwheel landing in cross position

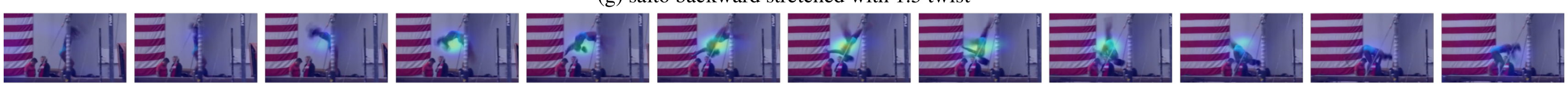

(c) free aerial walkover forward, landing on one or both feet


(d) salto forward stretched with 1 twist


(e) transition flight from low bar to high bar


(f) pike sole circle backward to handstand


(g) salto backward stretched with 1.5 twist


(h) salto backward stretched-step out (feet land successively)

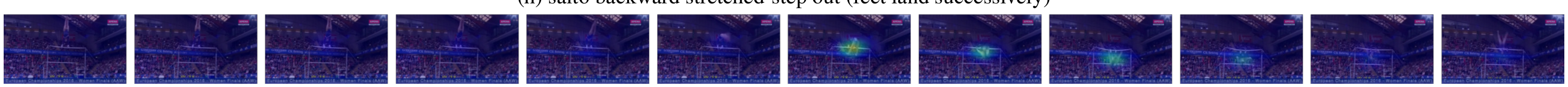
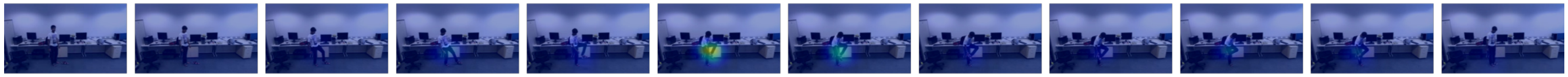
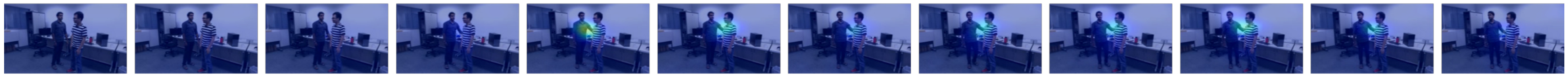

(i) stalder backward to handstand

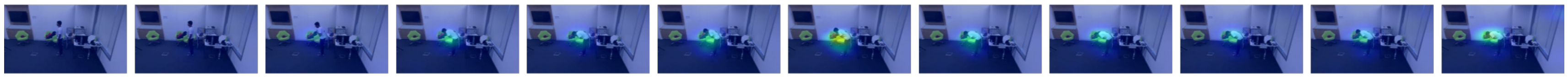
Figure A4: Visualization of 3D deformable attention on FineGYM



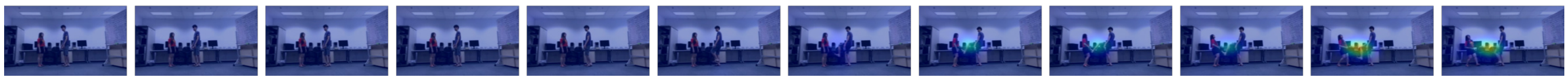
(a) put on a shoe



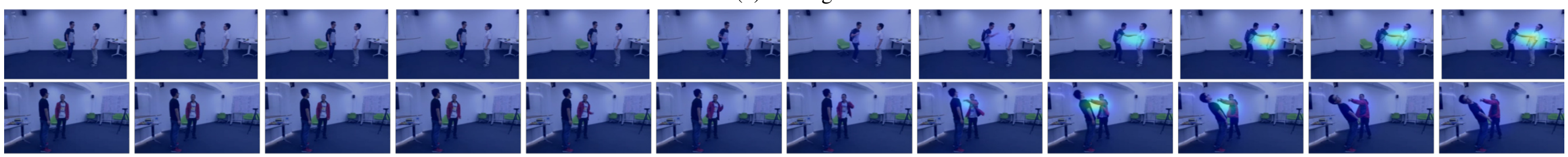
(b) pat on back



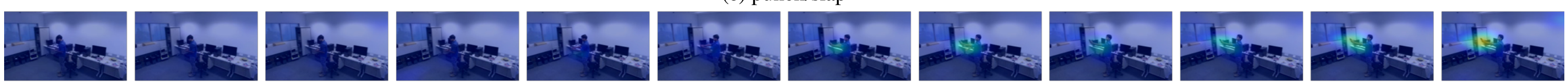
(c) nausea/vomiting



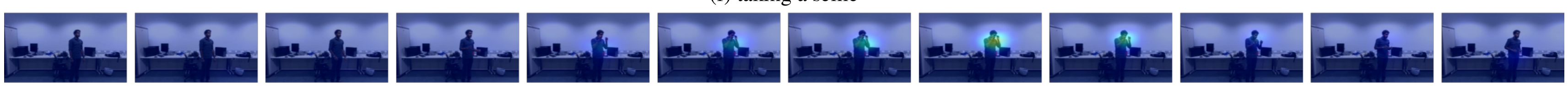
(d) kicking



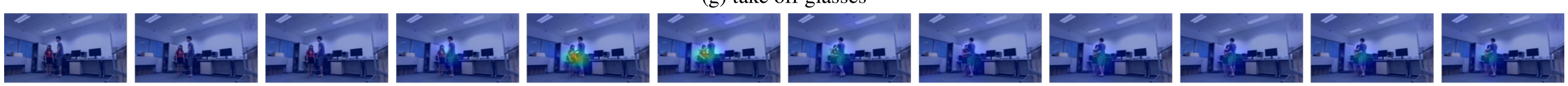
(e) punch/slap



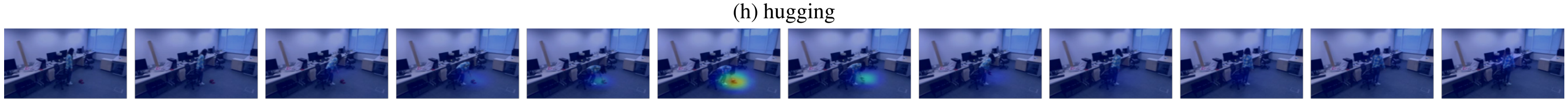
(f) taking a selfie



(g) take off glasses



(h) hugging



(i) pick up

Figure A5: Visualization of 3D deformable attention on NTU120