arXiv:2603.00878v1 [cs.CV] 1 Mar 2026

MMTA：用于细粒度卒中康复评估的多成员时序注意力

Halil Ismail Helvaci[1],Justin Huber[2],Jihye Bae[1],和 Sen-ching Samson Cheung[1,3]

[1] 肯塔基大学电气与计算机工程系，美国肯塔基州列克星敦[2] 肯塔基大学医学院，美国肯塔基州列克星敦[3] 加州大学戴维斯分校电气与计算机工程系，美国加利福尼亚州戴维斯 halil.helvaci@uky.edu

摘要。 为了支持个人康复过程中的迭代评估，自动评估个人在日常活动中的能力，需要对治疗视频中的细粒度动作进行时间精度高的分割。现有时序动作分割（TAS）模型难以捕捉亚秒级微动作，同时保留运动情境，因而会模糊快速的阶段转换，并限制对运动功能恢复进行可靠的下游评估。我们提出多成员时序注意力（MMTA），这是一种用于细粒度康复评估的高分辨率时序Transformer。不同于标准时序注意力在每层为每帧分配单一注意力上下文，MMTA使每帧能够在同一层内关注多个经过局部归一化的时序注意力窗口。我们通过特征空间重叠消解融合这些并行的时序视图，在转换附近保留相互竞争的局部上下文，同时借助逐层传播实现更长距离的推理。该方法在不增加网络深度或采用多阶段细化的情况下，提高了边界敏感度。MMTA在统一的单阶段架构中同时支持视频和可穿戴惯性测量单元输入，因此适用于临床与家庭场景。与全局注意力Transformer相比，MMTA在StrokeRehab上始终表现更优，视频和惯性测量单元的编辑分数分别提高 $+1.3$和 $+1.6$，并进一步使50Salads的表现提高 $+3.3$。消融实验验证，性能提升源于多成员时序视图，而非架构复杂度；这为资源受限的康复评估提供了一种实用方案。

关键词： 卒中康复评估 · 时序动作分割 · 细粒度运动分析。

1 引言

《卒中》仍是美国导致长期残疾的主要原因之一，每年影响超过 795,000 人，其中 77.4% 的幸存者出现

上肢功能障碍 [11,18]。手臂功能的恢复与日常生活活动的独立性密切相关，因此，准确且可重复地测量运动表现是康复的核心。然而，常规临床评估往往依赖基于观察的等级量表，这些量表耗时较长，对细微但具有临床意义的变化不够敏感，而且与现实生活中的手臂使用情况相关性较弱 [16,20,5]。这些局限促使人们开发自动化方法，将冗长、连续的治疗记录转化为具有临床可解释性的动作单元和高分辨率定量指标。

时序动作分割（TAS）通过为未剪辑序列中的每个时间步分配一个动作标签来满足这一需求，从而实现自动评估。卒中康复对 TAS 提出了尤为严峻的挑战：动作粒度细微且视觉特征不明显，而具有临床意义的转变可能在不到一秒的时间尺度内仅跨越几帧 [10]。全局自注意力能够捕捉长程时间结构，但随着时间上下文增大，Softmax归一化会将注意力分散到所有帧上，从而削弱局部边界证据；我们将这一固有局限称为时间粒度瓶颈。以往研究通过多阶段细化 [6,9], 分层时间编码器 [22], 或在注意力中施加稀疏性和局部性约束 [2,3,7,14,15,19,23]来缓解边界错误。然而，这些方法在每个层的每一帧上仅生成一次归一化更新，迫使模型在转变附近相互竞争的上下文之间做出次优取舍 [6,8,22,2]。我们认为，保留多重归属上下文对于细粒度 TAS 至关重要：在动作边界附近，帧往往包含跨越多个阶段的模糊证据；标准注意力会将其压缩为单一摘要，而实际上，单帧可以同时保留多个局部上下文视角，这有助于实现稳健的分割。

在本研究中，我们提出了多成员时序注意力（MMTA），这是一种用于细粒度时间动作分割的保边界注意力算子，无需多阶段细化和全局注意力。与标准窗口注意力不同，后者为每个帧生成一个局部归一化输出；MMTA 允许每个帧在每层参与 N 个重叠局部窗口，从而生成多个局部归一化的窗口条件更新，并通过显式的重叠解析规则将其融合。这保留了转变附近相互竞争的上下文证据，并改善了边界定位。我们在卒中康复治疗记录（StrokRehab 视频和惯性测量单元） [10] 以及 50Salads [17],上进行了评估，结果表明，MMTA 在包括编辑分数和动作错误率在内的边界敏感指标上始终优于现有的时间动作分割方法。

2 方法

问题设定。给定一个未剪辑的序列（视频或可穿戴传感器数据），其表示为帧级特征向量 $\mathbf{X} = \{\mathbf{x}_t\}_{t=1}^{T}$，时间动作分割（TAS）会为每个时间步 $\mathbf{Y} = \{y_t\}_{t=1}^{T}$ 预测一个动作类别（共 C 个动作类别）。一个动作实例是一个元组$y_i = (s_i, e_i, a_i)$，包含起始点 s_i、结束点 e_i和类别标签 $a_i \in \{1, \ldots, C\}$，其中

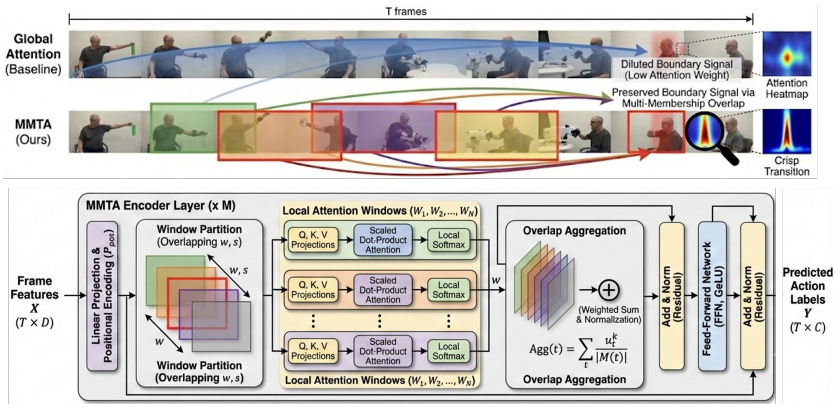

图 1. MMTA 将全局注意力替换为重叠时序窗口上的局部归一化注意力。重叠区域中的帧会接收多次窗口条件更新，并通过重叠消解聚合加以协调。

$1 \le s_i < e_i \le T$。因此，准确的时间动作分割既需要正确分类，也需要精确的边界定位。

概述。 我们采用用于时间动作分割的单阶段 Transformer 编码器骨干网络，并将全局自注意力替换为 MMTA，如图 1 所示。MMTA 将每一帧分配至多个局部重叠窗口，在边界附近产生局部归一化更新，并通过显式重叠消解对帧进行协调，使上下文无需全局 $T \times T$ 注意力即可跨层传播。

时间粒度瓶颈。全局自注意力在整个时间序列 T 上计算注意力。尽管这对捕捉粗粒度活动十分有效，但这种全局归一化会使注意力更均匀地分散到所有帧上。对于标记动作转换的边界帧 t_b，从查询帧 t_b 到邻近键帧 $t_b \pm \delta$ 的注意力权重（其中 δ 较小，例如为 2-5 帧）为：

$$Att(t_b, t_b \pm \delta) = \frac{\exp(s_{t_b, t_b \pm \delta})}{\sum_{j=1}^{T} \exp(s_{t_b, j})}, \qquad s_{i,j} = \frac{\mathbf{q}_i^\top \mathbf{k}_j}{\sqrt{d_k}}, \tag{1}$$

其中$\mathbf{q}_t, \mathbf{k}_j, \mathbf{v}_j \in \mathbb{R}^{d_h}$ 为单个注意力头中的查询、键和值向量，$s_{i,j}$ 表示缩放点积。在细粒度时间动作分割中，边界证据往往短暂且局部化。随着 T 增大，Softmax分母随之增大，分配给任一固定局部邻域的概率质量都会缩小，即使局部相似度很高。

在标准 Softmax 注意力下，预期的局部注意力权重会按 $\mathbb{E}[A(t_b, t_b \pm \delta)] = O(1/T)$ 衰减，从而降低对边界转变的灵敏度。这种注意力稀释会降低对短时转变的灵敏度，并导致时间过度平滑。由于这一效应是由对 T 个元素进行归一化所致，扩大模型深度或宽度并不能直接恢复边界保真度。我们将这一局限称为时间粒度瓶颈。

多成员时序注意力（MMTA）。不同于窗口化注意力为每帧生成单个局部归一化表示，MMTA 会在每一层中为每帧生成多个独立归一化的表示，并对其进行显式协调。它通过将归一化限制在局部窗口内来缓解注意力稀释，同时借助重叠保留跨窗口连续性。为建模局部时序依赖并维持片段间的上下文连续性，我们将自注意力操作限制在部分重叠的时序窗口内。给定长度为T的输入序列，我们将其划分为 $N = \lceil T/(w-o) \rceil$ 个窗口，每个窗口覆盖 w 帧，相邻窗口之间重叠 o 帧，步幅为 $s = w - o$。令$\mathcal{W}_i$ 表示第 i^{th} 个窗口中帧的索引集合。由于存在重叠，帧 t 可能属于多个窗口；其归属集合为 $\mathcal{M}(t) = \{i \mid t \in \mathcal{W}_i\}$，且$m(t) = |\mathcal{M}(t)| \geq 1$ 表示归属集合的大小。对于窗口 i，查询、键和值的投影分别记为 $Q_i, K_i, V_i \in \mathbb{R}^{(w)\times d}$。

通过将归一化限制在大小为 $w \ll T$ 的局部邻域内，MMTA 将全局分母 $\sum_{j=1}^{T} \exp(s_{i,j})$ 替换为局部化分母 $\sum_{j=1}^{w} \exp(s_{i,j})$，从而避免局部相似度被稀释，并恢复转变附近清晰的边界注意力。每个窗口内的注意力操作 [1] 遵循如下缩放点积注意力：$\text{Attn}(Q_i, K_i, V_i) = \text{Softmax}\left(\frac{Q_i K_i^\top}{\sqrt{d_k}}\right) V_i$，其中 d_k是用于缩放的键维度，$i = 1, \ldots, N$ 是每个重叠窗口的索引。该公式确保注意力仅在每个重叠局部区域内的帧之间计算，并通过共享的重叠区域实现相邻窗口间的信息交换。

将注意力限制在单个窗口内可能会在窗口边缘引入边界效应。MMTA 利用重叠，在一个层内为同一帧获取多个经过局部归一化的视图：如果 t 属于 $m(t)$ 个窗口，则会产生$m(t)$ 个窗口条件更新。随后，MMTA 应用重叠消解融合，将这些更新整合为每帧一个表示，从而促进跨窗口上下文传递，同时避免全局注意力的稀释效应。

重叠消解聚合。窗口注意力会为 $\mathcal{W}_i$ 内的每个位置生成一个输出。如果某个帧 t 出现在多个窗口中，MMTA 就会生成多个以窗口为条件的输出 $\{\mathbf{u}_t^{(i)}\}_{i\in\mathcal{M}(t)}$。我们通过显式的重叠消解算子来协调多重成员关系。

$$\tilde{\mathbf{h}}_t = \text{Agg}\Big(\{\mathbf{u}_t^{(i)}\}_{i\in\mathcal{M}(t)}\Big) = \frac{1}{m(t)} \sum_{i\in\mathcal{M}(t)} \mathbf{u}_t^{(i)}. \tag{2}$$

与标准窗口注意力不同，后者在每一层为每个帧生成一次归一化更新，MMTA 会构造多个局部归一化更新，并对其进行显式重解。尽管注意力在每个窗口内都是局部的，但重叠使信息能够在相邻窗口之间传递。堆叠 M 个 MMTA 层会将有效感受野扩展至 $w + (M-1)s$ 帧，从而在无需全局注意力的情况下提供长程上下文。

MMTA 算子。设 $\mathbf{H} \in \mathbb{R}^{T\times d}$ 为输入序列，$\{\mathcal{W}_k\}_{k=1}^{N}$ 为重叠时间窗口。对于每个窗口 k，我们计算窗口受限的

注意力 U_k =Attn(**H**$\mathcal{W}_k$)，并通过公式 2 协调多重归属帧，从而得到 MMTA 算子：

$$\mathrm{MMTA}(\mathbf{H}) \;:=\; \left\{ \frac{1}{m(t)} \sum_{k \in \mathcal{M}(t)} \mathbf{u}_t^{(k)} \right\}_{t=1}^{T} \in \mathbb{R}^{T \times d}. \tag{3}$$

这种聚合方式为每一帧生成一个表示，同时借助共享重叠区域实现相邻窗口间的信息传递。**MMTA编码器层**。 在第 ℓ 层，我们采用：

$$\tilde{\mathbf{H}}^{(\ell)} = \mathrm{MMTA}\Big(\mathbf{H}^{(\ell-1)}\Big), \tag{4}$$

$$\mathbf{H}^{(\ell)} = \tilde{\mathbf{H}}^{(\ell)} + \mathrm{FFN}\Big(\mathrm{LayerNorm}(\tilde{\mathbf{H}}^{(\ell)})\Big). \tag{5}$$

复杂度。 全局自注意力每层需要 $O(T^2 d)$ 次运算。MMTA在 $N \approx T/s$ 个窗口内进行注意力计算，每个窗口的计算量为 $O(w^2 d)$，总计为 $O\big(\frac{T}{s} w^2 d\big)$。当 $(w,\ s)$ 固定时，MMTA随序列长度 T 线性扩展，而全局自注意力的计算量则呈二次增长。

3 实验

数据集。 我们在 StrokeRehab [10]（视频和惯性测量单元模态）上评估 MMTA。该临床基础扎实的上肢治疗数据集要求转换时机的准确度达到亚秒级，这对评估运动功能恢复至关重要；此外，我们还在公开的 50Salads 基准 [17] 上进行测试，以评估模型在卒中康复之外的泛化能力。**StrokeRehab（视频/惯性测量单元）**包含来自 51 名卒中患者和 20 名健康受试者的 3,372 次试验，共有 120,891 个经过标注的功能性基本动作，涵盖九种日常活动（例如进食、刷牙）。标签由经过培训的编码人员在专家监督下生成，评定者间信度较高（Cohen's $\kappa \geq 0.96$）。惯性测量单元数据包含来自九个传感器的 76 个运动学通道，采样频率为 100 Hz；视频由两台正交摄像机拍摄（1088×704，60/100 fps）。由于原始视频因隐私原因未公开，本研究使用视频的深度学习 I3D特征 [4] 。该数据集可通过 SimTK 获取：*https://simtk.org/projects/primseq*。**50Salads** 包含 50 段沙拉制作视频（降采样至 15fps），由 25 名受试者完成，涵盖 17 个动作类别（例如切番茄、加盐）。该数据集是用于非临床场景中时序动作分割的标准基准。我们将从视频中提取的 I3D特征作为网络输入。

实现细节。 我们采用与先前工作 [6,9,10,22] 相同的数据划分和标准协议。使用 I3D [4] 提取视频特征。为体现不同模态在时间分辨率和动作时长上的差异，我们针对不同数据集使用特定的窗口长度和步幅（$w,\ s$）：StrokeRehab 视频采用 (200，10)，StrokeRehab惯性测量单元采用 (500，10)，50Salads 采用 (1500， 500)。模型容量按数据集进行调整。StrokeRehab 使用 3 个编码器层、4

表 1. 在 StrokeRehab（上）和 50Sal-ads（下）上与最先进方法的比较。两种 Transformer 基线方法使用相同的骨干网络；全局注意力 使用完整的自注意力，而*MMTA*（本文方法） 则以所提出的模块替代它。

骨干网络	模型	视频		IMU	
		ES ↑	动作错误率 ↓	ES ↑	动作错误率 ↓
TCN	MS-TCN* [6]	60.7	0.408	66.9	0.372
	MS-TCN [6]	62.2	0.392	68.9	0.330
	MS-TCN^{+} [6]	62.7	0.390	68.8	0.317
边界	ASRF* [9]	56.9	0.449	68.2	0.328
	ASRF [9]	58.7	0.436	67.9	0.349
序列到序列	Seg2Seq [10]	67.6	0.322	63.0	0.337
	Raw2Seq [10]	66.6	0.329	68.8	0.305
Transformer	Transformer（全局）	69.8	0.302	68.9	0.311
	MMTA（我们的方法）	**71.1**	**0.289**	**70.5**	**0.295**

骨干网络	模型	50Salads	
		编辑分数 ↑	动作错误率 ↓
TCN	MS-TCN* [6]	68.8	0.47
	MS-TCN [6]	70.8	0.43
	MS-TCN^{+} [6]	76.4	0.32
边界	ASRF* [9]	74.0	0.34
	ASRF [9]	75.2	0.33
序列到序列	Seg2Seq [10]	76.9	0.30
	Raw2Seq [10]	69.4	0.54
扩散	DiffAct [12]	85.0	-
	DiffAct++ [13]	85.8	-
Transformer	ASFormer [22]	79.6	-
	ASPnet [19]	87.5	-
	BaFormer [21]	84.2	-
	LTContext [2]	83.2	-
	Transformer（全局）	85.1	0.149
	MMTA（我们的方法）	**88.4**	**0.116**

个注意力头，隐藏层大小为 512。50Salads 使用 3 层、2 个注意力头，隐藏层大小为 256。

StrokeRehab 使用批量大小为 2、初始学习率为 10^{-3} 进行 25 个 epoch 的训练；如果验证集上的焦点损失连续 5 个 epoch 未改善，则将学习率降低至原来的 0.01 倍。50Salads 使用批量大小为 2 进行 10 个 epoch 的训练。我们使用动量为 0.9、权重衰减为 10^{-4} 的随机梯度下降，随机失活率为 0.2，焦点损失参数为 α=0.25 和 γ=2，并对 StrokeRehab 使用最大范数为 5、对 50Salads 使用最大范数为 6 的梯度裁剪。所有实验均在 NVIDIA RTX A6000 上运行。除非另有明确说明，否则不使用多阶段细化或后处理。

评估指标。我们报告分割编辑分数（ES）和动作错误率（AER），二者均根据真实分段转录 G 与预测分段转录 P[6,22,9,10]之间的 Levenshtein 距离 L计算得出。对于 StrokeRehab，我们还报告逐类别敏感度、特异度和 F1 分数，以补充说明临床上有意义的动作基元分割效果。

与最先进方法的比较。表 1 将 MMTA 与各模型系列中的代表性基线方法进行了比较，所有方法均采用单阶段 Transformer 编码器。基线方法涵盖五种范式：采用时序一维卷积的 TCN（MS-TCN [6]）；显式预测动作边界的边界感知模型（ASRF [9]）；自回归生成分段／标签序列的序列到序列方法（Seg2Seq、Raw2Seq [10]）；用于长程

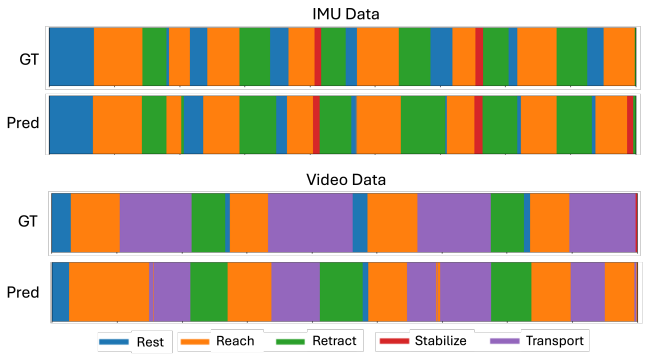

图 2. 在一个示例测试序列上，对比惯性测量单元数据（上）和视频数据（下）的预测（Pred）与真实值（GT）动作标签。

时间建模的 Transformer（ASFormer [22],ASPnet [19],BaFormer [21], 及其变体 [3,2]）；以及将分割视为迭代去噪过程的基于扩散的方法（Dif-fAct [12]）。Transformer（全局）作为我们的直接基线，采用与 MMTA 完全相同的架构，但将多成员时序注意力替换为完整自注意力。为减少帧噪声，可选择对输出 logits 进行平滑处理，随后再进行 $argmax$（+：应用平滑；$*$：根据最佳验证帧准确率选定）。

在 StrokeRehab（视频）上，性能从边界感知方法到时间卷积网络、序列到序列模型和 Transformer 持续提升，这反映出这些方法对细微过渡的长程时序建模能力更强。惯性测量单元的结果更为集中，表明局部运动学线索降低了对长程上下文的依赖，不过 Transformer 仍然表现最佳。在两种模态下，MMTA 的表现始终优于全局注意力，编辑分数高出 1.3–1.6，且动作错误率更低，说明其边界定位能力更强。在 50Salads 上，MMTA 在同时报告两项指标的方法中表现最佳，优于 Transformer（全局），并超过此前的编辑分数和动作错误率结果（如 ASPnet、DiffAct++）。

图 2 展示了预测结果与真实值的对比。MMTA 能更准确地识别边界过渡，减少虚假片段，同时保留动作序列整体的时间结构。错误主要集中在视觉或运动学上难以区分的基元中，例如动作之间的休息阶段。

效率。 我们测量了 MMTA 在 50Salads 上的 GPU 显存占用，发现其在不同窗口大小下仅需 422–460 MB，而 MS-TCN 需要 1.7 GB，ASFormer 则需要 3.5 GB。这些方法需要显著更多的显存，因为它们依赖多阶段处理（MS-TCN），或对整个序列进行全局自注意力计算（ASFormer）。

分类性能。除分割质量外，我们还评估了帧级分类性能，以考察各数据集上不同模态的优势。表 2 报告了 StrokeRehab 上关键基元的灵敏度、特异度和 F1 分数。惯性测量单元信号通常在

表 2. StrokeRehab视频和惯性测量单元数据集上的帧级动作分类评估

动作	**视频**			**IMU**		
	灵敏度	特异度	F1 分数	灵敏度	特异度	F1 分数
Rest	0.70	0.92	0.66	0.69	0.93	0.67
伸手	0.52	0.96	0.59	0.59	0.94	0.64
回缩	0.60	0.97	0.62	0.64	0.97	0.69
稳定	0.59	0.91	0.63	0.65	0.90	0.60
运输	0.77	0.85	0.70	0.72	0.88	0.71
宏平均	0.64	0.92	0.64	0.66	0.92	0.66

伸手和回缩等动态基元的灵敏度和 F1 分数更高，体现了其在高采样率下捕捉细粒度运动动态方面的优势。相比之下，视频模态的特异度始终更高，能减少假阳性，尤其是在运输这类视觉特征鲜明的动作上。

消融研究。 表3评估了窗口大小和步幅对 StrokeRehab 的影响。较小的步幅始终能改善边界定位，这是因为转变附近的多重归属帧数量更多；而最佳窗口大小因模态而异：视频在 w=200 时表现最佳，惯性测量单元在 w=500 时表现最佳，这反映了时间分辨率和运动动态的差异。在窗口大小的两个极端，性能都会下降，这证实了模态特定的窗口调优对于平衡局部敏感性和上下文覆盖十分重要。

表3。在 StrokeRehab 视频和惯性测量单元数据集上，对窗口大小和步幅影响的消融研究。

StrokeRehab						
模型	视频		IMU		窗口大小	步幅
	编辑分数 ↑	动作错误率 ↓	编辑分数 ↑	动作错误率 ↓		
MMTA	62.0	0.388	67.6	0.324	1500	500
	66.8	0.332	68.9	0.314	1000	500
	68.2	0.318	66.9	0.331	800	500
	68.7	**0.313**	**70.5**	**0.295**	500	10
	71.1	**0.289**	66.8	0.332	200	10
	68.0	0.320	63.8	0.364	100	10

4 结论

我们提出了 MMTA，这是一种多隶属时序注意力算子，通过允许每帧参与多个重叠的局部窗口，在无需多阶段细化或全局注意力的情况下缓解时间粒度瓶颈，并保留相互竞争的边界证据。与强基线方法相比，MMTA 始终能够提升编辑分数并降低动作错误率，同时降低复杂度。其线性复杂度和较低的内存占用，使 MMTA 成为适用于临床与家庭场景下自动化康复评估的实用方案。MMTA 的一项局限是依赖固定窗口配置，这可能

无法最优地适应不断变化的时间动态；未来的工作将探索自适应或可学习的窗口化策略。

致谢。 本出版物所报告的研究得到了肯塔基大学启航研究合作计划（IRC）的支持。

利益披露。 作者声明不存在与本文内容相关的利益冲突。

参考文献

1. 注意力就是一切。神经信息处理系统进展 **30**(2017) 2. Bahrami, E., Francesca, G., Gall, J.: 动作分割需要多少长期时间上下文？载于：IEEE/CVF国际计算机视觉大会论文集。第10351–10361页（2023） 3. Beltagy, I., Peters, M.E., Cohan, A.: Longformer：长文档Transformer。arXiv预印本 arXiv:2004.05150（2020） 4. Carreira, J., Zisserman, A.: 动作识别将走向何方？一种新模型和Kinetics数据集。载于：IEEE计算机视觉与模式识别会议论文集。第6299–6308页（2017） 5. Chen, Y.A., Demers, M., Lewthwaite, R., Schweighofer, N., Monterosso, J.R., Fisher, B.E., Winstein, C.: 加速度测量与生态瞬时评估相结合，用于评估卒中后偏瘫臂/手使用：可行性与有效性。《临床医学杂志》**10**(6), 1328（2021） 6. Farha, Y.A., Gall, J.: MS-TCN：用于动作分割的多阶段时间卷积网络。载于：IEEE/CVF计算机视觉与模式识别会议论文集。第3575–3584页（2019） 7. Gulati, A., Qin, J., et al.: Conformer：用于语音识别的卷积增强Transformer。arXiv预印本 arXiv:2005.08100（2020） 8. Helvaci, H.I., Chuah, C.N., Ozonoff, S., Cheung, S.C.S.: 在自闭症谱系障碍婴儿的未剪辑视频中定位动作时刻。载于：2024 IEEE国际图像处理会议（ICIP）。第3841–3847页。IEEE（2024） 9. Ishikawa, Y., Kasai, S., Aoki, Y., Kataoka, H.: 通过检测动作边界缓解过度分割错误。载于：IEEE/CVF冬季计算机视觉应用会议论文集。第2322–2331页（2021） 10. Kaku, A., Liu, K., Parnandi, A., Rajamohan, H.R., Venkataramanan, K., Venkatenatsan, A., Wirtanen, A., Pandit, N., Schambra, H., Fernandez-Granda, C.: Strokerehab：用于亚秒级动作识别的基准数据集。神经信息处理系统进展 NeurIPS **35**，1671–1684（2022） 11. Lawrence, E.S., Coshall, C., Dundas, R., Stewart, J., Rudd, A.G., Howard, R., Wolfe, C.D.: 多民族人群中急性卒中损伤和残疾患病率的估计。《卒中》**32**(6), 1279–1284（2001） 12. Liu, D., Li, Q., Dinh, A.D., Jiang, T., Shah, M., Xu, C.: 扩散动作分割。载于：IEEE/CVF国际计算机视觉大会论文集。第10139–10149页（2023） 13. Liu, D., Li, Q., Dinh, A., Jiang, T., Shah, M., Xu, C.: DiffAct++: 扩散动作分割。《IEEE模式分析与机器智能汇刊》**47**(3), 1644–1659（2025）。https://doi.org/10.1109/TPAMI.2024.3509434

14. Liu, Z., Lin, e.a.: Swin Transformer：利用移位窗口的层次化视觉Transformer。载于：IEEE/CVF国际计算机视觉大会论文集。第10012–10022页（2021） 15. Lu, Z., Elhamifar, E.: FACT：用于高效动作分割的帧-动作交叉注意力时序建模。载于：IEEE/CVF计算机视觉与模式识别会议论文集。第18175–18185页（2024） 16. Murphy, M.A., Willén, C., Sunnerhagen, K.S.: 用于量化卒中后伸手取物及从玻璃杯饮水期间上肢功能表现的运动学变量。神经康复与神经修复 **25(1)**, 71–80（2011） 17. Stein, S., McKenna, S.J.: 将嵌入式加速度计与计算机视觉相结合，以识别食物制备活动。载于：2013年ACM普适与泛在计算国际联合会议论文集。第729–738页（2013） 18. Tsao, C.W., Aday, A.W., Almarzooq, Z.I., Anderson, C.A., Arora, P., Avery, C.L., Baker-Smith, C.M., Beaton, A.Z., Boehme, A.K., Buxton, A.E., et al.: 心脏病与卒中统计数据——2023年更新：美国心脏协会报告。《循环》**147(8)**, e93–e621（2023） 19. Van Amsterdam, B., Kadkhodamohammadi, A., Luengo, I., Stoyanov, D.: ASPnet：通过多数据源的共享-私有表示进行动作分割。载于：IEEE/CVF计算机视觉与模式识别会议论文集。第2384–2393页（2023） 20. Waddell, K.J., Strube, M.J., Bailey, R.R., Klaesner, J.W., Birkenmeier, R.L., Dromerick, A.W., Lang, C.E.: 任务特异性训练能否改善卒中后日常生活中的上肢表现？神经康复与神经修复 **31(3)**, 290–300（2017） 21. Wang, P., Lin, Y., Blasch, E., Wei, J., Ling, H.: 通过边界感知查询投票实现高效的时序动作分割。arXiv预印本 arXiv:2405.15995（2024） 22. Yi, F., Wen, H., Jiang, T.: ASFormer：用于动作分割的Transformer。载于：《英国机器视觉会议》（BMVC）（2021） 23. Zhang, C.L., Wu, J., Li, Y.: ActionFormer：利用Transformer定位动作时刻。载于：欧洲计算机视觉会议。第492–510页。施普林格（2022）