arXiv:2611.32890v1 [cs.CV] 24 Nov 2026

基于生物力学的婴儿视频选择性运动控制与临床结局预测

DivyaJoshi[1, 2], Ruize Yang[1, 3], Colleen Peyton[2, 4], R. James Cotton[1,5][1]仿生医学中心，Shirley Ryan AbilityLab，芝加哥，伊利诺伊州[2]西北大学物理治疗与人体运动科学系，芝加哥，伊利诺伊州[3]西北大学神经科学系，埃文斯顿，伊利诺伊州[4]西北大学儿科系，芝加哥，伊利诺伊州[5]西北大学物理医学与康复系，芝加哥，伊利诺伊州

脑性瘫痪（CP）的早期识别依赖于对婴儿自发运动的观察性评估，例如全身运动评估（GMA）和婴儿观察性选择控制评估（BabyOSCAR）；这些评估需要受训评分者，且评分耗时较长。现有自动化方法从二维关键点中提取手工设计特征，描述的是图像平面上的运动，而非各个关节的三维运动行为。我们提出婴儿生物力学三维身体（BB3D），这是一种适用于婴儿的生物力学处理流程，可从单摄像机视频中估计关节角度轨迹。在对17名婴儿进行的多视角无标记动作捕捉参考基准验证中，BB3D的Procrustes对齐误差中位数在关节中心处为17.5毫米，在标记点位置处为20.4毫米。一个时空变换器先使用来自1k+ 段单摄像机录像的BB3D关节角度轨迹进行预训练，再使用来自350名婴儿的550段视频进行微调；在留出录像上，该模型对每个BabyOSCAR单项的预测平均站点内AUC为0.79，对BabyOSCAR总分的预测与参考评分的斯皮尔曼相关系数为0.86。尽管模型并未针对诊断进行训练，其预测总分仍能区分脑瘫婴儿与非脑瘫婴儿，AUC为0.86，高于复现的二维特征分类器（0.67），但低于临床专家评定者的总分（0.97）。预测分数也包含有关功能能力和身体分布的信息，但信息量少于临床专家评分。这些结果表明，基于生物力学的单目视频表征有助于在婴儿期开展可扩展的选择性运动控制关节层面评估。

日期：2026年11月**通讯作者：**
divyajoshi@northwestern.edu

ISR Intelligence Sensing and Rehabilitation

1 引言

脑性瘫痪（脑瘫）是儿童最常见的运动障碍，早期识别
在发育中的神经系统具有高度可塑性时，早期干预发挥着关键作用（Novak等，2017年；Herskind等，2015年）。临床评估工具侧重于婴儿自发运动，尤其关注早期运动发育过程中出现的运动质量、变异性和选择性。两种值得关注的工具是Prechtl全身运动评估（GMA），该工具评估婴儿在扭动运动阶段和不安定运动阶段自发运动的整体质量（Bosanquet等，2013年；Einspieler和Prechtl，2005年），以及婴儿观察性选择控制评估（BabyOSCAR）（Peyton等，2024年；Sukal-Moulton等，2024年；Barbosa等，2024年；Peyton等，2026年），后者在单个关节轴层面评估运动。

这两种工具都需要受训评分者并耗费大量评分时间，限制了其在临床和研究环境中的可扩展性与可及性。GMA以单一分类的形式对运动质量进行整体评估，而BabyOSCAR则评估32个关节进行独立关节控制的能力，从而能够更精细地揭示独立或非独立的关节运动模式（协同运动、镜像动作、蜿蜒运动）。将这些评估自动化，有望在保留具有临床意义的关节层面信息的同时，为脑瘫（CP）可能性较高的婴儿提供高效、可扩展的筛查。

近期研究已采用计算机视觉和姿态估计，从婴儿视频中预测临床结局（Chambers 等人，2020；Gama 等人，2025；Segado 等人，2026年；Alfano 等人，2026年）。然而，现有方法均使用稀疏的二维关键点表示和人工设计的运动学描述符（Joshi等人，2026年），无法直接编码深度信息或身体的完整三维构型。因此，从这些点得出的运动学信息实际上是对图像平面运动的估计，而非对生物力学姿态的测量。这一局限性对于 BabyOSCAR 尤为重要，因为其临床解读取决于单个关节的相对三维运动行为，以及运动是独立发生还是作为更广泛运动模式的一部分发生。此外，特征工程方法将分析限制在预先人工选定的运动描述符上，可能会阻碍模型从潜在的运动轨迹中发现具有临床意义的模式。最后，既往自动化研究主要聚焦于全身运动评估或脑瘫相关结局；据我们所知，此前尚未有人尝试对 BabyOSCAR 实现自动化。

在本研究中，我们通过使用源自单目视频的生物力学姿态表示来应对这些局限。我们针对婴儿进行适配的流程，将此前建立的跨模型蒸馏（Cotton等人，2026年a）与通过逆运动学拟合训练的生物力学头（Cotton等人，2026年b）相结合，在无需配对的图像到生物力学标签的情况下，估计具有解剖学定义的关节运动。我们将这一组合框架称为婴儿生物力学三维身体（BB3D）。该表示为验证姿态估计值和预测细粒度临床运动评分提供了共同基础。我们探究这一生物力学表示能否同时支持可靠的姿态估计和具有临床相关性的运动评估。

本研究有两个目标。首先，我们在由17名婴儿组成的前瞻性队列中，将单目生物力学估计值与使用已建立的多视角无标记动作捕捉（MMMC）系统（Cotton等人，2023年）获取的记录进行验证，共计33个测评场次、345条记录，并遵循此前所述的数据采集和预处理方案（Joshi等人，2026年）。多相机重建为评估从单摄像机推断出的标记点位置和关节轨迹的准确率提供了参考基准。其次，我们使用来自回顾性队列（574名婴儿、1,284条记录）的单摄像机录像所估计的生物力学轨迹，训练时空变换器（Zhu等人，2023年），以预测具有临床意义的结果，包括BabyOSCAR总分和单项评分、脑瘫诊断、功能能力及身体分布。我们评估了多种模型架构形式，并将所得性能与现有的基于姿态的方法和人工评分方法进行比较。

本研究的主要贡献如下：

1. **婴儿单目生物力学姿态估计的多摄像机验证**。我们将适用于婴儿的单目姿态估计流程（BB3D）与经校准的多摄像机参考基准进行比较，并量化其估计的关节中心和标记点位置的准确率。

2. **自动预测关节级选择性运动控制**。我们开发了一种时空 Transformer，用于根据 BB3D 在单摄像机录像中估计的生物力学关节角度轨迹，预测 32 项逐关节 BabyOSCAR 条目，并从一致性和维度两个方面将其评分与专家评分进行比较。

3. **基于预测评分的临床结局**。我们使用根据预测项目得出的评分来分类脑瘫诊断、功能能力和身体分布，并将其

性能与复现的二维关键点分类器以及根据专家评分计算出的相同评分进行比较。

总之，这些贡献检验了以生物力学为基础的动作表征能否在保留表征选择性运动控制所需的关节级信息的同时，使婴儿临床评估更具可扩展性。

2 方法

2.1 参与者

本研究纳入了来自七个数据集的婴儿，这些数据集采集于临床、实验室和居家环境。队列中既有早产儿，也有足月儿，记录时的矫正胎龄为28–64周。若干数据集包含纵向记录，因此同一婴儿可在不同月龄贡献视频。多摄像头验证队列包含17名婴儿、33个记录测评场次，共评估了345条记录。单摄像头临床预测队列包含574名婴儿和1,284条记录：其中1,163条来自505名婴儿，用于预训练；其中，来自350名婴儿的550条经BabyOSCAR评分的记录用于微调；另有一个留出测试集，包含来自69名婴儿的121条记录。各组的人口统计学和临床特征，包括脑瘫诊断、粗大运动功能分级系统（GMFCS）等级、脑瘫身体分布、性别、出生类型以及记录时的矫正胎龄，汇总于表1。

交叉验证折按婴儿分组，因此没有婴儿同时出现在训练折和验证折中。22名同时有训练记录的婴儿被排除在测试集之外，因此没有测试婴儿的记录用于预训练或微调。在其余69名测试婴儿中，68名在出生后月经龄50–60周时有记录，且专家BabyOSCAR评分、脑瘫诊断、GMFCS和身体分布信息完整，因此被纳入临床结局分析（图 4）。

涉及人类参与者的所有程序均经各参与机构相关机构审查委员会或研究伦理委员会批准（西北大学、卢瑞儿童医院、芝加哥大学科默儿童医院和Shirley Ryan AbilityLab）。根据具体情况，已取得父母或法定监护人的知情同意，或经相关机构审查委员会批准免除知情同意。

2.2 视频采集与预处理

2.2.1 多摄像头记录

我们使用MMMC系统采集多摄像头记录，该系统从八个视角捕捉婴儿自发运动（Joshi等人，2026年；Cotton等人，2023年）。婴儿在平静、清醒状态下仰卧位躺卧；每个测评场次包含8–10次试验，每次持续50–60秒。标定、同步和跟踪均遵循我们此前的方法（Joshi等人，2026年；Cotton等人，2023年）。

使用 Sapiens2（Khirodkar等人，2026年）在每个视角中检测二维关键点。Sapiens2 是 Sapiens 的后继模型，在我们先前的比较中，其婴儿密集关键点检测表现最佳（Joshi等人，2026年；Khirodkar等人，2024年）。除俯视摄像机外，其余七台摄像机的关键点均经过稳健的三角测量（Roy等人，2022；Joshi等人，2026年）；置信度低于 0.5 的点被视为未观测点，从而确保参考基准独立于被评估视角。随后将三角测量得到的关键点重新投影到俯视摄像机视图中，作为该摄像机单目估计的参考基准。对每段俯视摄像机录像应用 BB3D（Cotton等人，2026b；Cotton等人，2026a），以估计身体姿态、肢段尺度和标记点偏移量，再通过正向运动学将其转换为三维关节中心和标记点位置。

表1. 训练队列和测试队列的特征。用于Transformer预训练和BabyOSCAR微调的队列的人口统计学和临床特征。训练婴儿在交叉验证期间采用折外方式评估，而测试婴儿则作为留出集。评估时年龄指记录时的矫正胎龄（出生后月经龄，PMA），单位为周。分类变量采用 χ^2检验进行比较，年龄则采用Welch$t$检验。SMD表示微调训练组与测试组之间的标准化均数差；低于0.1的数值表明两组平衡良好（Yang和Dalton，2012年）

特征	预训练	微调	Test	统计量	p	SMD
婴儿，n	505	350	69			
记录，n	1163	550	121			
脑瘫诊断，n				$\chi^2 = 0.59$	0.898	0.10
CP	134	128	22			
可能脑瘫	17	17	4			
NCP	223	152	32			
不太可能脑瘫	53	53	11			
脑瘫患者中的 GMFCS 分级，n				费希尔精确检验 0.186		0.52
I–II	36	34	7			
III–V	44	41	3			
未知	54	53	12			
脑瘫分布，n				$\chi^2 = 0.12$	0.726	0.14
单侧	48	44	9			
双侧	78	78	12			
未知	8	6	1			
性别，n				$\chi^2 = 1.71$	0.191	0.19
女性	184	158	25			
Male	244	188	44			
未知	77	4	0			
出生类型，n				$\chi^2 = 0.08$	0.775	0.06
早产	277	224	46			
Term	155	126	23			
未知	73	0	0			
评估时年龄，矫正胎龄周数，均值 ± 标准差	52.4 ± 4.0	52.2 ± 3.8	52.2 ± 4.1	$t = -0.11$	0.912	0.01

2.2.2 单摄像机录像

单摄像机队列包含来自实验室、临床、新生儿重症监护病房（NICU）和家庭环境的 1–5 分钟视频，因此摄像机位置、图像质量、背景、光照和衣着各不相同。使用以“baby”为提示词的 Segment Anything Model 3（SAM 3）（Carion等人，2026年）跟踪婴儿，并以得分最高的检测结果初始化跟踪，再裁剪图像以进行姿态估计；这样可避免跟踪画面中的照护者或其他人员。裁剪后的图像由适配婴儿的 BB3D流程处理（Cotton等人，2026a；Cotton等人，2026b），其中 Sapiens 提供二维关键点（Khirodkar等人，2024年），BB3D 则提供三维身体构型，并将其转换为关节角度轨迹，作为临床预测模型的输入。

2.3 生物力学姿态估计的验证

将俯视摄像机拍摄的 BB3D 估计结果与三角测量参考值进行比较。该摄像机的视角最接近临床和家庭记录中使用的俯视视角。比较分两个层面进行：(1) 解剖学关节中心（双侧的肩部、肘部、腕部、髋部、膝部和踝部；共 12 个点），并与相应的 Sapiens2关节关键点比较；(2) 标记点位置（身体、面部和指尖共 34 处），每处均与相应的 Sapiens2 关键点配对（图 1A）。由于标记点位置和图像关键点的定义方式不同（分别位于骨性标志上方的皮肤处和学习得到的图像标志点），因此配对表示最接近的等效点，而非完全相同的点。

2.3.1 Procrustes 对齐误差计算

单个视角无法确定身体的绝对尺寸，因此，对每一帧中的预测点应用 Procrustes 对齐变换（旋转、平移和统一缩放），使其与参考基准对齐；该变换针对每个层面分别拟合，并基于该帧中可观测到的点对（评估至少需要六对）。点误差定义为对齐后该点与参考基准之间的欧氏距离，并在一段记录的所有帧上取平均值。为区分一致性偏差和变异性偏差，残差以躯干固定坐标系表示，该坐标系由参考髋部和肩部定义；系统性偏移是指一段记录中平均残差向量的大小，离散度则是残差相对于该向量的均方根值。误差以所有记录的中位数和四分位距汇总，并分别统计所有点以及各解剖位置（合并左右侧）的结果。

2.3.2 标记点位置偏移标定

在两个阶段评估标记点位置，这两个阶段均基于相同的 BB3D 姿态和肢段尺度计算。在标准 阶段，标记点位置设在模型标记配置中的默认位置，关节中心设在各自的身体段原点，并舍弃 BB3D 预测的标记点偏移量。在标定 阶段，每个位置都沿其所在肢段的未缩放坐标系偏移一个静态偏移量，该偏移量在所有帧和记录中保持一致。我们通过交替进行逐帧 Procrustes 对齐和偏移量的闭式最小二乘更新（Huber 损失，30 mm；向标准位置施加较弱的岭惩罚），将偏移量拟合到三角测量得到的 Sapi- ens2 关键点；标记点位置和关节中心分别拟合，并采用留一参与者交叉验证。由于姿态保持不变，误差的任何变化都反映了位置设定而非姿态估计的影响。作为参考，我们还报告了 BB3D Body 使用其自身预测偏移量时的输出。使用参与者层面的中位数，通过 Wilcoxon 符号秩检验对阶段间变化进行检验。

2.4 临床评分与结局预测

临床预测使用单摄像机录像，并比较了基于不同运动表征构建的两种模型：一种是使用 Sapiens 估计的二维关键点的基线分类器（Khirodkar等人，2024年），另一种是我们使用 BB3D Body 估计的生物力学关节角度（`qpos`）轨迹构建的 Transformer（Cotton等人，2026b；Cotton等人，2026a）。我们预测了32个二元 BabyOSCAR 条目及其衍生的汇总分数，以及临床结局：脑瘫诊断、功能能力（GMFCS 分级）和脑瘫身体分布。

2.4.1 数据划分

所有模型共用一个以参与者为单位的80/20训练集/测试集划分，并按脑瘫诊断分层，因此每名婴儿的所有记录都被分配到同一侧。训练集部分按参与者和重复记录组划分为五个交叉验证折。这些折用于：(1) 每折排除该折参与者，预训练一个 Transformer 模型；(2) 使用其余四折对每折训练一个模型，并利用留出折选择停止轮次；(3) 对训练集部分的每条记录，使用未曾在该记录上训练的模型生成折外预测；(4) 以五个折模型的预测均值作为测试集预测；(5) 对临床结局的回归读数进行交叉拟合。测试集仅用于最终评估。

2.4.2 基线分类器

为与既有研究进行比较，我们使用二维 Sapiens 关键点计算了 Segado等人（Segado 等人，2026年）报告的 38 项运动特征。使用自动_`sklearn2` 自动机器学习软件包（Feurer 等人，2015）（搜索时长一小时）对这组特征进行分类；模型在训练集上训练，以区分患有脑瘫和未患脑瘫的婴儿，并在测试集上评估一次。每段录像的

输出按照与 Transformer 相同的方式取平均值，汇总到婴儿层面，因此两种方法都在同一组测试婴儿上进行评估。

2.4.3 使用预训练时空变换器预测 BabyOSCAR

我们使用了 44 个 BB3D 关节角（qpos）通道，频率为 30 Hz，并排除了根部、手指、跖趾和膝部旋转通道。该 Transformer 在随机采样的 10 秒窗口上以掩码自编码器的方式进行预训练。每个窗口被划分为 30 个、每个包含 10 帧的图块，生成 1,320 个通道–图块标记。该双层、256 维、四头编码器采用了 MotionBERT 的因式分解时空注意力（Zhu 等，2023年）。在整个窗口内，70% 的通道被掩码，并通过单层解码器对其进行重建。预训练使用所有符合条件的已处理记录进行，共运行 200 个周期，但排除了测试参与者和留出折。

在微调阶段，我们移除了解码器，并使用六个均匀间隔的 10 秒窗口表示 550 条经过评分的训练记录。编码器处理每个窗口时显示所有通道。对时间平均的通道嵌入进行归一化，并在各窗口之间取平均，从而为每个通道生成一个 256 维嵌入。每个 BabyOSCAR 条目都使用这些嵌入的学习加权组合，其初始化依据为该条目的关节、邻近和对侧通道；没有对应输入通道的条目则使用所有通道进行初始化。随后，将条目表示与学习得到的关节类型和身体侧别标签结合，并通过共享的多层感知机预测该条目的对数几率。微调过程中，使用跨 32 个条目的梯度范数平衡二元交叉熵，更新编码器的最终层和预测组件。我们以 0.5 的概率应用左右镜像增强，使用留出的训练折选择检查点，训练三个随机种子，并对其测试集概率取平均。

派生的BabyOSCAR评分。 总分是32项独立关节运动概率之和，包括肢体总分、右侧和左侧关节总分，以及根据各项目计算的侧化指数（右侧减左侧），计算方式与人工评估相同（Peyton et al.,2024; Sukal-Moultonet al., 2024; Barbosa et al., 2024; Peyton etal., 2026）。为比较预测评分与专家评分的结构，我们分别针对专家评分和预测概率，对留出录像（交叉验证折外数据和测试集）中所有项目均有评分的32项相关矩阵进行了主成分分析。我们报告了累积方差、达到80%和90%方差所需的成分数，以及参与比——根据特征值分布计算的有效维度数（若一个成分承载全部方差，则为1；若方差均匀分布，则为32）。区间带为按参与者聚类的自助法95%区间（重采样200次）。

2.4.4 基于预测的 BabyOSCAR 分数的临床结局

在出生后月经龄（PMA）为50–60周时，从记录中以婴儿层面读取结局，在每个测评场次内取平均值，再在各场次之间取平均值。读取结果包括固定分数（负总分或不对称性 $\sum |R - L|$），或基于六项汇总指标或全部32个条目的正则化回归结果，并在各折之间进行交叉拟合。任务包括脑瘫（CP）与非脑瘫（NCP）的区分，以及在脑瘫婴儿中区分GMFCS III–V级与I–II级、单侧脑瘫与双侧脑瘫。此外，还将各脑瘫亚组（GMFCS III–V级、GMFCS I–II级、单侧、双侧）分别与NCP组进行比较，使用负总分或不对称性作为指标。从同一批婴儿的专家条目评分中获取相同的读取结果，并通过配对婴儿水平自助法与模型结果进行比较，分别针对测试集和折外训练侧进行分析。

2.4.5 模型实验

每个实验 实验改变了基础模型的一个方面，同时保留其折、编码器和随机种子：

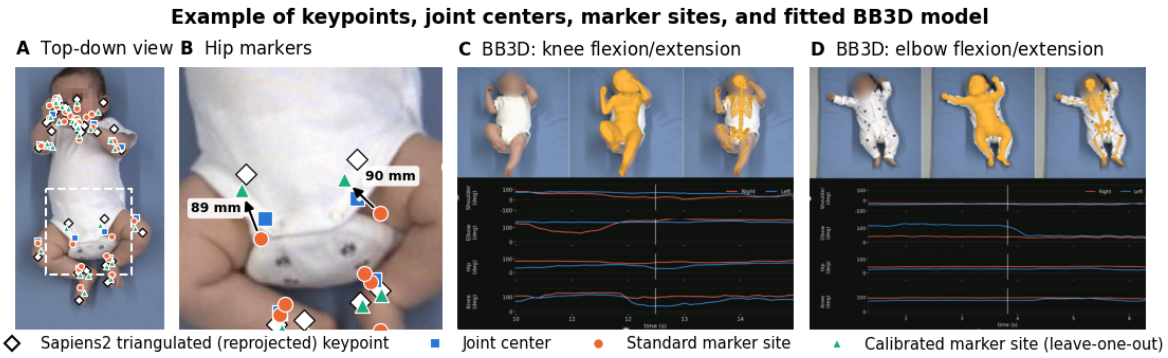


图1. 单个俯视摄像机画面中关键点、关节中心、标记点位置及拟合后的 BB3D 模型示例。(A) 由七台非俯视摄像机进行三角测量并重新投影到俯视图中的 Sapiens2 关键点（白色菱形），以及相应的 BB3D 关节中心（蓝色方块）、标准标记点位置（橙色圆圈）和校准标记点位置（绿色三角形）。虚线框标示了 (B) 中放大的区域。**(B)** 髋部特写。箭头从每个标准髋部标记点位置指向其校准后的位置，并标注以毫米为单位的三维位移。**(C, D)**视频的 BB3D 拟合结果（帧、网格叠加、骨架叠加），以及由此得到的右侧（红色）和左侧（蓝色）关节角度时间序列。(C) 突出显示膝部，(D) 突出显示肘关节屈伸。

- 添加脑瘫诊断和GMA评分头；
- 添加预测总分、肢体总分、侧别总分和侧别的预测头；
- 仅使用出生后月经龄达到40周及以上的记录进行训练，并在相同的 $\geq$40周记录上进行评估；
- 将共享项目网络替换为每个项目各自使用一个网络。

2.4.6 评估

由于各招募数据集的转诊路径和结局患病率不同，主要指标按数据集分别计算，并同时报告汇总值。对于记录数至少为20的数据集，按各项目的ROC-AUC评分，根据不一致对加权，并对通道所测量关节对应的28个项目取平均。总分及衍生分数按数据集内斯皮尔曼相关系数的均值评分，同时报告平均绝对误差和校准斜率。折外预测略显乐观，因为同一折用于选择停止训练的轮次；测试预测则取五个折模型的平均值。模型实验与基础模型在相同记录上进行比较，并使用按参与者聚类的自助法置信区间。

3 结果

3.1 生物力学姿态估计的验证

我们分析了来自 17 名参与者的 345 条记录；留出摄像机的 Sapiens2 检测结果的重投影误差中位数为 5.4 像素（IQR 2.1）。采用标准模型配置时，关节中心和标记点位置的 Procrustes对齐误差中位数分别为 26.9 mm 和 30.5 mm，误差主要是系统性的（偏移分别为 24.8 和 25.9 mm，离散度分别为 10.9 和 16.5 mm），且髋部误差最大（图 1A、B 和 2）。留一参与者校准将误差降至 17.5 和 20.4 mm，并改善了全部 17 名参与者的结果（Wilcoxon $p < 0.0001$；图 1B 和 2），其原因是系统性偏移得到纠正（分别为 13.6 和 14.5 mm），而离散度保持不变。髋部误差降幅最大（图 1B 和 2）。留出参与者的偏移量与使用全部参与者拟合的偏移量几乎完全一致（17.5 对 17.3 mm，20.4 对 20.1 mm）。

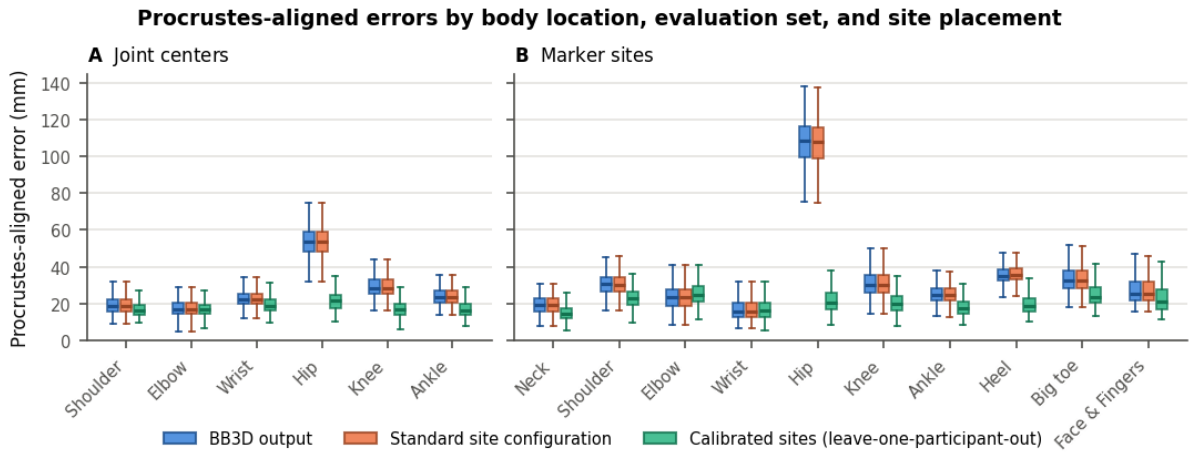


图 2. 按身体部位、评估集和点位布置划分的拟合 BB3D 的 Procrustes对齐误差。误差是相对于由其余七台摄像机进行三角测量得到的 Sapiens2 关键点计算的，并在逐帧进行相似性对齐（旋转、平移和统一缩放）后得出。**（A）**展示 12 个关节中心的结果，**（B）**展示 34 个标记点位置的结果。每个箱线图汇总了 17 名参与者的 345 条记录中的指标：对应身体部位各点位的平均误差，左右侧合并计算。鼻子、眼睛、耳朵、拇指和指尖合并为“面部与手指”。颜色分别表示 BB3D 输出（蓝色）、标准点位配置（橙色）和校准点位（绿色）。校准偏移量根据其余参与者的数据拟合（留一参与者交叉验证）。每组点位均基于自身的点逐帧进行对齐。箱体表示中位数和四分位距；须线延伸至 1.5× 个四分位距；不显示离群值。

3.2 自动化 BabyOSCAR 评分

在来自 69 名婴儿的 121 条留出测试记录中，预测的 BabyOSCAR 总分（即 32 个预测项目概率之和）与专家总分高度一致（数据集内 Spearman $\rho = 0.78$，95% CI [0.65-0.84]；合并后 $\rho = 0.86$），平均绝对误差为 3.5 个条目（图 3A）。这种一致性也体现在各个项目上。在 BB3D 会追踪其关节的 28 个项目中，站点内 AUC 均值为 0.79（0.74-0.82）（图 3B）。项目AUC介于 0.73 至 0.88 之间，右髋内旋/外旋为 0.73，右肩外展/内收为 0.88；在 0.5 阈值下，32 个项目的准确率均高于始终预测最常见分数的准确率。

然而，预测的项目得分维度低于专家的得分。主成分分析显示，专家的 32 项得分中有 18 个成分解释了 80% 的方差（参与比为 6.6），而预测得分仅有 1 个成分（参与比为 1.5）（图 3C）。

3.3 基于预测的 BabyOSCAR 得分得出的临床结局

接下来，我们考察了预测的项目评分是否携带与专家评分相同的临床结局信息。每名婴儿由单一的32项向量表示，即其在50–60矫正胎龄周数各测评场次上预测项目概率的均值。该时间窗对应于不安运动期，在此期间进行全身运动评估（Einspieler和Prechtl，2005），且BabyOSCAR已在此期间被验证可用于结局预测（Barbosa等人，2024年；Sukal-Moulton等人，2024年；Peyton等人，2024年；Peyton等人，2026年）。为避免针对较小的临床分组拟合模型，每项结局均依据项目的固定汇总来判读，判读方式与临床上判读BabyOSCAR一致。总分用于脑瘫诊断和功能能力分类，总分越低，表明对行动支持的需求越大。不对称性，即16项运动的左右绝对差值之和，用于在脑瘫组与非脑瘫组中区分身体分布类型（单侧与双侧），并在脑瘫子集内对身体分布进行分类。专家的项目评分在同一批婴儿上以相同方式汇总，从而为BabyOSCAR本身所含的临床信息提供参考基准。

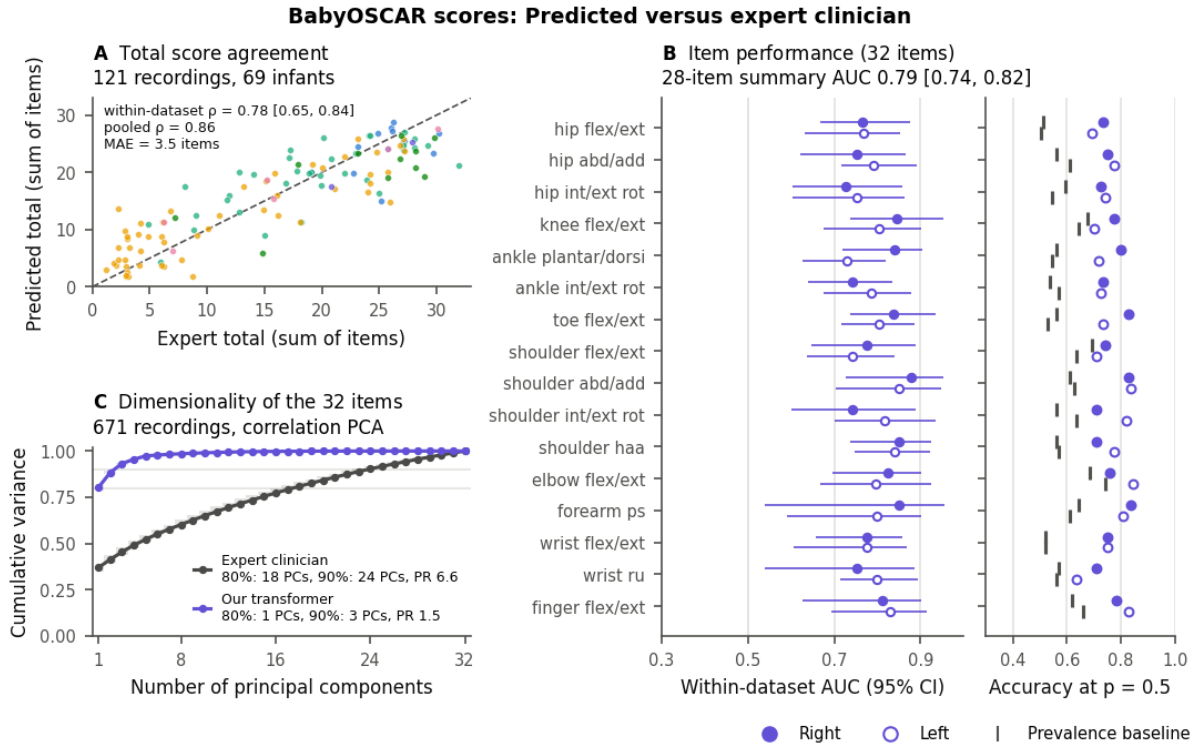


图3。BabyOSCAR 在单目记录上的评分预测。（A）预测总分（32 个预测项目概率之和）与专家总分的对比；每条记录以一个点表示（121 条记录，69 名婴儿），并按招募数据集着色。虚线表示完全一致。报告了数据集内 Spearman ρ 及其 95% CI、汇总 ρ 以及平均绝对误差。**（B）**16 个关节自由度的逐项目表现，右侧（实心）和左侧（空心）。左：数据集内 AUC 及其 95% CI。右：0.5 阈值下的准确率，并列显示患病率基线（始终预测较常见的类别）。标题给出 28 个项目的汇总 AUC。**（C）**32 个项目评分的主成分所解释的累积方差：专家标注（灰色）与预测概率（紫色）的对比。该面板汇总了折外记录和测试记录（671 条记录），因为维度并非留出评分。阴影带表示 95% 的参与者聚类自助法区间。图例给出了达到 80% 和 90% 方差所需的主成分数量以及参与比（PR）。

每项结局均在两组婴儿上评估（图4）。主要分析使用了68名测试集婴儿，任何模型在训练或模型选择过程中均未见到这些婴儿。由于测试集中患脑瘫的婴儿较少，我们还报告了324名验证集婴儿，每名婴儿均由未在其数据上训练过的交叉验证模型进行预测。每个交叉验证模型的微调轮次是根据其验证折上的项目损失选定的，因此训练集估计可能略微偏乐观；测试集估计不受这一选择的影响。

在测试集上，预测总分能够区分后来被诊断为脑瘫的婴儿（n=22）与未患脑瘫的婴儿（n=31），AUC 为 0.86（95% CI 0.74–0.95）；相比之下，专家总分的 AUC 为 0.97（0.92–1.00），而我们实现的 Segado 等人（Segado 等人，2026年）38特征分类器的 AUC 为 0.67（0.52–0.81）。该分类器直接根据脑瘫与非脑瘫进行训练，并在同一婴儿测试集上评估（图4B）。在验证集（115 名脑瘫婴儿、143 名非脑瘫婴儿）上，预测总分和专家总分的 AUC 分别为 0.88（0.83–0.92）和 0.98（0.96–0.99）（图 4A）。

基于功能能力（GMFCS 分级）和身体分布进行分层后，对脑瘫诊断进行分类，结果显示，与非脑瘫组的区分在粗大运动功能分级系统III–V级以及双侧脑瘫患者中最明显（图4C、E）。在验证集中，预测总分区分粗大运动功能分级系统III–V级与非脑瘫组的 AUC 达到 0.97，区分粗大运动功能分级系统I–II级与非脑瘫组的 AUC 为 0.86（两者的专家评分均为 0.99）；预测不对称性区分单侧脑瘫与非脑瘫组的 AUC 为 0.77（专家评分为 0.92），区分双侧脑瘫与非脑瘫组的 AUC 为 0.94（专家评分为 0.99）。测试集呈现相同趋势（粗大运动功能分级系统III–V级为 1.00、I–II级为 0.82、单侧为 0.78、双侧为 0.89；专家评分依次为 1.00、0.97、0.99 和 0.92），相应类别中患有脑瘫的婴儿人数分别为 3、7、9 和 12。

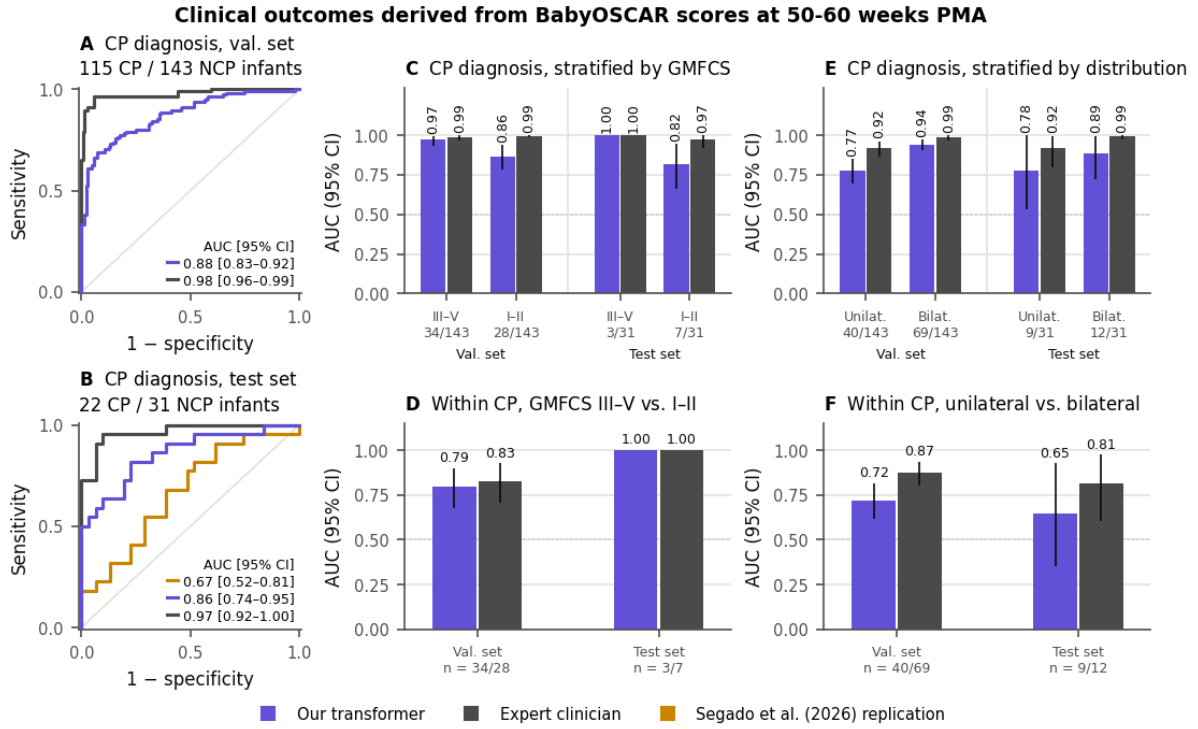


图4. 在出生后月经龄50–60周时对临床结局进行婴儿层面预测。 这些是预先注册的 BabyOSCAR 条目固定读数；没有针对结局拟合模型。我们 Transformer 预测的条目（紫色）和专家的条目评分（灰色）均采用相同方式，在同一批婴儿中读取：(1) 用于脑瘫诊断和 GMFCS 分级的总分负值，以及 (2) 用于侧别的左右不对称性$\Sigma|R-L|$。金色表示我们对 Segado 等 (2026年)方法的实现：使用 38 项运动特征，并通过脑瘫与非脑瘫分类器进行训练，仅在测试集上评分。(**A, B**) 验证集（折外：每名婴儿由未使用该婴儿进行训练的折模型评分）和留出测试集上脑瘫与非脑瘫的 ROC曲线；虚线表示随机水平（0.5）。(**C**) 按 GMFCS 分级分层的脑瘫诊断（GMFCS III–V级与非脑瘫、GMFCS I–II级与非脑瘫）。(**D**)脑瘫组内的 GMFCS 分级。(**E**) 按身体分布分层的脑瘫诊断（单侧与非脑瘫、双侧与非脑瘫）。(**F**) 脑瘫组内的身体分布。柱形表示 AUC，其 95% 置信区间由婴儿聚类自助法计算得出（重抽样 2,000 次）；浅灰线表示随机水平（0.5）。每名婴儿仅计入一次；训练集/测试集划分两侧均有记录的婴儿不纳入分析。

在患有脑瘫的婴儿中，预测总分和专家总分均能以 1.00 的 AUC 区分粗大运动功能分级系统III–Ⅴ级与GMFCS I–II级，测试集两组分别有 3 名和 7 名婴儿；验证集的 AUC 分别为 0.79 和 0.83，两组分别有 34 名和 28 名婴儿（图 4D）。预测不对称性区分单侧脑瘫与双侧脑瘫的 AUC 在测试集中为 0.65（0.35–0.93），两组分别有 9 名和 12 名婴儿；在验证集中为 0.72，两组分别有 40 名和 69 名婴儿。相比之下，专家评分的 AUC 分别为 0.81 和 0.87（图 4F）。

3.4 模型实验

最后，为了检验哪些设计选择有助于提升性能，我们在相同的记录和婴儿数据上，将四种变体与基准模型进行了比较：（1）增加脑性瘫痪和全身运动评估预测头；（2）增加用于总分、肢体、侧别和侧别评分的预测头；（3）仅使用出生后月经龄达到或超过40周的记录进行训练；（4）为每个项目分别设置一个分类器。这些改动均未显著改变任何指标的性能；对于每项指标，与基准模型之间差异的95%置信区间均包含零。因此，我们保留了较为简单的基准设计。

4 讨论

本研究表明，从单目视频中估算得出的三维生物力学信息能够预测婴儿的关节级选择性运动控制和临床结局。与以往基于图像平面关键点的方法（Chambersetal., 2020; Gamaetal., 2025; Segado 等人, 2026年; Alfano 等人, 2026年）不同，我们的表征描述了每个关节在三维空间中的运动方式，而这正是 BabyOSCAR 所定义的层级。

与多摄像机参考基准相比，单摄像机的 BB3D 估计值在标定标记点位置放置后，偏差控制在约 2 cm 以内（关节中心为 17.5 mm，标记点位置为 20.4 mm）。标定主要消除了系统性误差，帧间离散程度保持不变，且结果可迁移至留出婴儿，表明残余误差主要反映的是标志点定义差异，而非姿态差异。

预测的 BabyOSCAR 与专家评分吻合良好（数据集内 $\rho = 0.78$ 总分的一致性；平均项目 AUC 为 0.79）；尽管模型仅使用 BabyOSCAR 条目训练，仍能区分脑瘫与非脑瘫（AUC 为 0.86）。合并数据的一致性高于数据集内一致性（$\rho = 0.86$ 对比 0.78），但这未必反映了站点混杂：Transformer 仅处理尺度归一化关节角轨迹，因此不太可能利用图像层面捷径；此外，每个数据集来源所涵盖的年龄和损伤程度范围都较窄。因此，数据集内一致性是一个保守的下限，而合并数据的一致性则反映了模型对完整发育谱系的敏感性。

对 Segado 等人提出的二维特征分类器进行复现(Segado 等人，2026年)，在不使用深度信息或时间结构的情况下，该分类器在同一批测试婴儿中区分脑瘫与非脑瘫的 AUC 为 0.67，而 Transformer 的 AUC 为 0.86。尽管如此，专家的准确度仍然更高，而评分结构揭示了原因。专家对各项目的评分参与比是预测评分的四倍多（有效维度为 6.6 对 1.5）。模型将 BabyOSCAR 压缩到少数几个共享轴上，且主要由整体运动质量主导；而临床医生在评估每个关节时，会利用模型未能获取的信息：某个动作是孤立出现，还是属于镜像动作或协同模式的一部分；动作的时序；以及整个观察过程的背景（Peyton 等人，2024年；Sukal-Moulton 等人，2024年；Barbosa 等人，2024年；Peyton 等人，2026年）。专家评分维度较高，部分原因也可能是项目层面的评分噪声，因此该比值反映的是模型遗漏信息的上限，而非其确切程度。

这种差异体现在分层结果中。在区分脑瘫 GMFCS III–V级组与非脑瘫组时，Transformer 的表现几乎与专家相当（AUC 分别为 0.97 和 0.99）；这是因为 GMFCS 分级较高的婴儿在多个关节上表现出的独立运动较少，而这一特征可由一个全局轴捕捉。对于 GMFCS I–II级脑瘫婴儿，以及依赖特定关节或身体侧别的区分，二者之间的差距进一步扩大，因为脑瘫婴儿具有更多的独立关节运动和侧别特征。因此，自动评分目前已经能够支持对 GMFCS III–V级婴儿进行可扩展筛查，而 GMFCS I–II级婴儿和具有单侧表现的婴儿仍可从专家评估中受益。

4.1 局限性

该模型缺乏事件检测和时序分析，也不会标注解释其评分的运动模式，例如镜像动作或协同动作。不同来源在年龄和功能障碍程度上的差异意味着，汇总指标和数据集内指标只能界定性能范围，而不能对其进行精确测量。GMFCS 分级和身体分布信息往往缺失，因此，脑瘫组内比较和亚组比较仅基于少数婴儿，且置信区间较宽。单目准确度仅在实验室俯视摄像机条件下得到验证，未在临床环境和家庭环境中验证，也未在小于我们验证集所含婴儿年龄的婴儿（例如新生儿重症监护病房中的婴儿）中验证。

4.2 未来方向

未来的模型可以接收临床医生关于每个关节运动模式和事件时序的记录，或显式检测镜像动作和协同动作，从而接近专家评分的维度。我们的数据集还包含出生史信息、临床标签和纵向记录，尽管存在标签缺失，这些信息仍可支持对婴儿运动轨迹进行因果建模。

总之，BB3D 可将婴儿运动的单目三维估计误差控制在约 2 cm 以内。基于这些估计构建的时序模型，在预测选择性运动控制和临床结局方面优于图像平面方法；在识别患有脑瘫且粗大运动功能分级系统III–Ⅴ级的婴儿方面，其表现接近专家水平。这是迈向可扩展的关节级婴儿运动评估的一步。

致谢

我们要感谢 Imani Mann 和 Grace Hoo 协助参与者招募和日程安排，感谢 Shawana Anarwala 和 Kayan Abdou 协助实验设置和数据收集，并感谢 Kunal Shah 协助基础设施建设。本研究由美国国立卫生研究院资助，资助项目为 T32TR005124 和 1R21HD117074。

参考文献

Alfano，L 等（2026年7月）。*5000*名婴儿项目：婴儿脑性瘫痪运动模式的早期检测。Barbosa，VM、C Peyton 和 T Sukal-Moulton（2024年）。“婴儿观察性选择控制评估（BabyOSCA

R):

构念效度与测试性能”。载于：《发育医学与儿童神经病学》。Bosanquet, M, L Copeland, R Ware, and R Boyd (2013). “系统综述：预测幼儿脑性瘫痪的测试”。载于：《发育医学与儿童神经病学》 55.5。Carion, N et al. (2026). “SAM 3：概念驱动的万物分割”。载于：$arXiv$。Chambers, C et al. (Nov. 2020). “利用计算机视觉自动评估婴儿神经运动风险”。载于：《*IEEE*神经系统与康复工程汇刊》 28.11, pp. 2431–2442。Cotton, RJ et al. (Mar. 2023). “无标记动作捕捉与生物力学分析流程”。载于：$ArXiv$。Cotton, RJ, D Joshi, and C Peyton (2026a). “利用未标注婴儿视频对人体姿态基础模型进行跨模型蒸馏，以实现无标记三维姿态估计”。载于：$arXiv$。Cotton, RJ, JD Peiffer, L Williamson, J Leske, and G Pavlakos (2026b). “生物力学三维人体：从三维人体基础模型自监督蒸馏生物力学姿态”。载于：$arXiv$。Einspieler, C and HFR Prechtl (2005). “Prechtl一般运动评估：用于幼年神经系统功能评估的诊断工具”。载于：《智力障碍与发育障碍研究综述》。Feurer, M, A Klein, K Eggensperger, J Springenberg, M Blum, and F Hutter (2015). “高效且稳健的自动化机器学习”。载于：《神经信息处理系统进展》 *28*, pp. 2962–2970。Gama, F et al. (2025). “从视频自动进行婴儿二维姿态估计：比较七种深度神经网络方法”。载于：$ArXiv$。Herskind, A, G Greisen, and JB Nielsen (2015). “脑性瘫痪的早期识别与干预”。载于：《发育医学与儿童神经病学》 57.1。Joshi, D, JD Peiffer, C Peyton, and RJ Cotton (2026). “用于婴儿生物力学全身运动学估计的无标记动作捕捉”。载于：$arXiv$。Khirodkar, R et al. (2024). *Sapiens*：人类视觉模型基础“*Sapiens*”意为与现代人有关或相似。”。技术报告。Meta。Khirodkar, R, H Wen, J Martinez, Y Dong, S Zhaoen, and S Saito (2026). “Sapiens2”。载于：$arXiv$。Novak, I et al. (Sept. 2017). “脑性瘫痪的早期准确诊断与早期干预：诊断与治疗的进展”。载于：《美国医学会杂志·儿科学》 171.9, pp. 897–907。Peyton, C et al. (2024). “婴儿观察性选择控制评估（BabyOSCAR）：3个月时的评分可预测2岁时的功能能力、痉挛型脑性瘫痪分布及诊断结果”。载于：《发育医学与儿童神经病学》。Peyton, C et al. (Feb. 2026). “婴儿早期选择性运动控制的出现：一项针对早产儿（伴或不伴脑性瘫痪）的纵向研究”。载于：《儿科学杂志》 289, p. 114915。Roy, SK, L Citraro, S Honari, and P Fua (2022). “关于将三角测量作为三维人体姿态估计的一种自监督形式”。载于：论文集——*2022*年国际三维视觉会议。

Segado, M, LA Prosser, AF Duncan, MJ Johnson, and KP Kording (2026). “一项预注册的开放流程：基于婴儿视频进行早期脑瘫风险评估”。载于：*GigaScience* 15。Sukal-Moulton, T et al. (2024). “婴儿观察性选择控制评估（BabyOSCAR）：痉挛型脑瘫婴儿与非脑瘫婴儿中的聚合效度、区分效度与信度”。载于：《发育医学与儿童神经病学》。Yang, D and JE Dalton (2012). “使用SAS测量两组间效应量的统一方法”。载于：*SAS*全球论坛*2012*年论文集。佛罗里达州奥兰多：SAS研究所公司，论文335–2012。Zhu, W, X Ma, Z Liu, L Liu, W Wu, and Y Wang (2023). “MotionBERT：学习人体运动表征的统一视角”。载于：*IEEE/CVF*国际计算机视觉会议论文集。