arXiv:2611.30176v1 [cs.CV] 23 Nov 2026

OmniSenMo：用于泛化人体运动估计、理解与预测的统一多感官语言模型

Cheng Guo[1] QimingCao[2] Pu (Perry) Wang[3] Pu Wang[1] Hongfei Xue[1*]1北卡罗来纳大学夏洛特分校2普渡大学3三菱电机研究实验室 (MERL){cguo3, pwang13, hxue2}@charlotte.edu, cao393@purdue.edu, pwang@merl.com项目主页：https://ch3ngguo.github.io/omnisenmo/

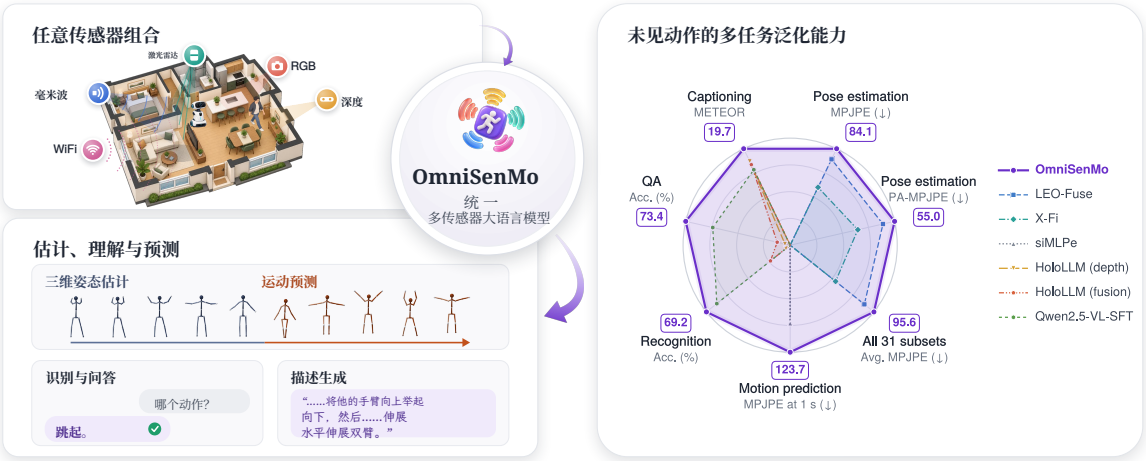


图 1. **一个适用于任意传感器子集、可泛化至未见动作的统一模型**。对于五种传感器中的任意子集，同一个 OmniSenMo 模型在模拟传感器上预训练后，可估计三维人体姿态，通过识别、问答和描述生成来理解动作，并预测未来运动；即使是未包含在真实训练数据中的动作，也能进行预测。右侧：未见过的 MM-Fi 动作；以往方法仅展示了其支持的任务。

摘要

多模态大语言模型（*MLLM*）能够估计、理解和预测人体运动，但主要依赖 *RGB* 相机。在光照不佳、存在遮挡以及对隐私敏感的应用场景中，深度、激光雷达、毫米波和 *WiFi* 可作为 *RGB* 的补充，但现有多传感器模型只涵盖了其中部分感知任务。要仅凭任意传感器子集涵盖所有这些任务颇具挑战，因为异构传感器的令牌在语义上不一致、空间上未对齐，配对多传感器数据稀缺，而且不同任务需要不同的时间粒度。我们提出 *OmniSenMo*，这是一种以人体骨架作为所有传感器共同参考的多传感器语言模型。其以骨架为基础的运动表示将每种传感器的数据转换为关节令牌，每个身体关节对应一个令牌，因此任意传感器子集都能按关节进行融合，并以语言模型从动作捕捉和文本中学到的运动词汇来表达。我们使用多传感器模拟器生成的全部五种传感器数据对其进行预训练，而多粒度观测则为每项任务提供所需的帧。在 *MM-Fi* 的跨动作划分上，采用全部五种传感器的同一模型在每项任务上的表现均优于对比模型，包括姿态估计任务中的 *LEO-Fuse*（*84.1* 对比 *89.0mm MPJPE*）和未见动作识别任务中的 *HoloLLM*（*69.2%* 对比 *42.8%*）。即使不使用 *RGB*，该模型仍优于两种使用 *RGB* 的方法；在 *XRF55* 上进行跨数据集验证时，其识别未见动作的表现也远胜 *HoloLLM*。

1. 引言

家用机器人等具身智能体正越来越多地与人类共同生活在家中 [38, 43]。为了帮助人类，智能体需要知道他们身处何处、正在做什么，以及接下来将如何移动 [59]。基于视觉的多模态大语言模型（MLLM）将视觉感知与语言推理相结合 [48]，以支持人体姿态估计、动作理解和运动预测 [15, 24, 46, 74, 76],；近期模型更是统一了多种运动任务，表现优于各自的单任务对应模型 [74, 76]。然而，仅依赖 RGB 相机会限制这些模型在真实场景中的应用。例如，深度传感器通过直接测量深度来补充 RGB 信息 [67],

*通讯作者。

表1. **相关模型的能力**。 I：RGB，D：深度，L：激光雷达，R：毫米波，W：WiFi。任意子集：任意非空传感器子集。统一：一个模型适用于所有任务。估计：估计，HAR：动作识别，QA：问答，Cap.：动作描述，Pred.：预测。 ○：每个输入仅使用一种传感器，融合仅在其附录中涉及。

	传感器					任务						
方法	I	D	L	R	W	Any 子集	估计	HAR	问答	描述生成	预测	统一
X-Fi [16]	√	√	√	√	√	✓	✓	✓	✗	✗	✗	✗
LEO-Fuse [4]	√	√	√	√	√	✓	✓	✓	✗	✗	✗	✗
HoloLLM [97]	✓	✓	✓	✓	✓	○	✗	✓	✓	✓	✗	✓
mmPred [22]	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✓	✓
UniPose [46]	✓	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✓
Superman [74]	✓	✗	✗	✗	✗	✗	✓	✗	✗	✗	✓	✓
UniMotion [76]	✓	✗	✗	✗	✗	✗	✓	✗	✗	✓	✓	✓
OmniSenMo	√					✓	✓	✓	✓	✓	✓	✓

而激光雷达、毫米波和 WiFi 则能在光照不佳的情况下进行感知 [85, 97]。此外，毫米波和 WiFi 能够穿透遮挡物甚至墙壁进行感知 [44]，并且由于不会记录视觉外观，适用于对隐私敏感的应用 [4, 85]。这些互补能力推动了多传感器人体感知的发展 [16, 85]。

辅助人类需要多传感器模型，既支持姿态估计和运动预测，也能通过动作识别、动作问答（QA）和描述生成来实现动作理解。这些模型还应能泛化到日常生活中遇到、但其真实训练数据中未包含的未见动作。现有的多传感器融合方法，例如 X-Fi [16] 和 LEO-Fuse [4]，支持从任意传感器子集中进行姿态估计和动作识别（表 1）。然而，每项任务都需要单独训练模型，而且动作识别仍局限于预定义标签。传感器-语言模型通过将毫米波 [41, 83, 87]，WiFi [88], 和多种感知模态 [97] 连接到语言模型，实现了更丰富的动作理解，包括问答和描述生成，但无法进行姿态估计或运动预测。此外，现有的多传感器融合模型仅在已见动作上进行评估，其对未见动作的泛化能力尚未得到探索。因此，我们提出了图 1所示的问题： 单个语言模型能否利用任意可用的传感器子集，估计、理解并预测人体运动，包括其真实训练数据中未包含的动作？ 要回答这一问题，必须应对传感器异质性、数据稀缺和多样化任务需求所带来的三项挑战。

首先，传感器异质性会导致传感器标记之间的语义不一致和空间错位。在语义层面，传感器专用编码器会强调运动的不同方面，并将其与依赖于传感器、且与运动无关的模式纠缠在一起，例如图像背景或依赖环境的多径效应。在空间层面，传感器标记的数量可能不同，并且缺乏与相同身体关节的对应关系 [97]。受共享运动—语言表示的启发 [35, 49], , 我们将人体骨架确定为对同一人体运动的异质观测进行对齐的共同参考。现有多模态语言模型接收的是可学习查询 [31, 42] 或经过池化的 CLIP 标记 [62, 97]，与此不同，我们提出一种统一的、以骨架为基础的运动表示，其中每个传感器针对每个关节和帧生成一个关节令牌，并通过该关节的三维位置及其在共享的预训练运动词汇表中的离散运动令牌进行监督。这些共享目标促使不同传感器之间形成一致且聚焦于关节的运动语义，而显式的关节对应关系则提供了空间对齐。由此得到的表示在关节语义上与传感器无关，从而能够对任意可用的传感器子集进行逐关节融合，并复用从动作捕捉和文本中学习到的运动知识。我们进一步对模型进行预训练，使其根据其余传感器恢复被遮蔽的传感器的关节标记，而不是重建图像块 [6] 或通用的模态专属特征 [50, 78],，从而促使模型在传感器不可用时利用互补的关节信息。

其次，配对多传感器数据十分稀缺。采集经校准的同步传感器观测成本高昂且耗时，因此现有涉及射频（RF）信号的多传感器数据集 [13, 36, 73, 85] 所覆盖的动作和受试者远少于 RGB 或动作捕捉数据集 [37, 51]。然而，传感器—语言对齐 [18,31, 97] 与联合多传感器预训练 [6, 7, 53, 55] 都需要大规模配对语料库。为缓解这一稀缺性，我们构建了一个多传感器模拟器，可为多样化的动作捕捉序列生成来自全部五种传感器的同步观测 [51], , 并与三维姿态和语言配对 [26, 60],，从而在适配真实数据之前实现联合预训练。除仿真之外，我们还通过在同步帧上进行特征匹配，将视觉语言模型（VLM）的视觉编码器蒸馏到非 RGB 编码器中。

第三，这些任务需要不同的时间粒度 [77]：估计和预测需要局部运动细节，而理解则需要更广的动作上下文。密集的完整动作输入可以同时提供二者，但令牌数量会随传感器和帧的增加而增长。在有限的帧预算 [45, 52],下，连续观测能够保留局部细节，但可能缺失理解所需的动作上下文。因此，我们提出多粒度观测（MGO），使模型能够以不同的时间粒度处理输入：用连续帧进行估计，用覆盖一个动作的样本进行理解，并用近期历史进行预测。

通过结合模拟的运动多样性、视觉语言模型知识迁移和以骨架为基础的表示，OmniSenMo 能够跨传感器复用运动—语言知识，并泛化到未见动作。我们的贡献如下：

• 我们提出了 OmniSenMo，据我们所知，这是首个能够利用任意传感器子集估计、理解和预测人体运动的统一模型。其以骨架为基础的运

动表示通过人体骨架对齐异构传感器，因为每个传感器的关节标记都对应于相同的关节，并由每个关节的三维位置和运动令牌进行监督。多粒度观测以每项任务的时间粒度进行处理。

- 我们构建了一个多传感器模拟器，从带标注的动作捕捉数据生成涵盖五种传感器的语料库，并配以三维姿态和语言；随后在该语料库上预训练 OmniSenMo，再将其适配到真实数据，从而缓解数据稀缺问题并提升对未见动作的泛化能力。

- 在 MM-Fi 的跨动作划分上，使用全部五种传感器的 OmniSenMo 在每项任务上的表现均优于对比模型；在 XRF55 上进行的跨数据集验证中，它识别未见动作的表现也远胜 HoloLLM。我们将发布代码和模拟语料库。

2. 相关工作

多传感器人体感知。 除了 RGB 相机之外，研究人员还利用深度信息估计人类姿态 [67], 激光雷达 [3, 86],雷达 [21, 94, 95], 和 WiFi [72, 82, 99], , 并利用毫米波预测人体运动 [22]。随着配对多传感器数据集的出现 [13, 36, 73, 85], , X-Fi 能够处理任意传感器子集进行姿态估计和动作识别，而无需针对每种组合分别训练模型 [16], ; LEO-Fuse 将一种通用任务架构应用于多项任务，但为每项任务单独训练模型 [4], ; AWC 则在姿态估计中平衡 RGB、激光雷达、毫米波和 WiFi 的学习 [61]。对于训练集中未出现的动作，零样本识别方法将毫米波、WiFi 或物联网信号与文本对齐 [12, 65, 100], , DeSPITE 则将激光雷达点云与骨架、惯性测量单元和文本嵌入 [39]。HoloLLM 将五种传感器集成到多模态大语言模型中，用于动作识别、问答和描述生成 [97], , 而针对毫米波 [41, 66, 83, 87] 和 WiFi [88] 的语言模型则仅处理单一传感器。OmniSenMo 通过统一模型，针对任意传感器子集估计、理解并预测人体运动。

多模态与运动-语言模型。 多模态大语言模型通过模态专用编码器 [68, 80]或跨模态共享的单个冻结编码器 [31, 93] 引入新模态。对于人体运动，离散运动令牌支持生成 [27, 64, 89], , 语言模型则生成并理解运动 [15, 35, 49]。带时间戳的骨架令牌使模型能够推理动作顺序和持续时间 [52]。ChatPose [24],、UniPose [46], 和 UniMotion [76] 将姿态或运动理解与生成统一起来，并结合文本和图像；Superman [74] 则增加了基于视频的三维姿态估计和运动预测。这些运动模型仅通过 RGB 或动作捕捉观测人类。

数据稀缺与仿真。 成对传感器数据十分稀缺，因此感知模型通常借助 RGB 的监督信号。跨模态蒸馏使用同步 RGB [5, 30, 94], 上的视觉教师模型训练深度、音频或无线电模型，而绑定方法则将更多模态与图像对齐，lan-语言、或者已对齐的模态 [18, 25, 75, 84, 101]。仿真则能创建配对数据。合成人体用于训练视觉人体姿态和体型估计器 [10, 57, 69, 71]。传感器模拟器已将动作捕捉、视频或生成场景转换为深度图像 [67], 惯性信号 [34, 40], WiFi信号 [11], 以及多普勒和毫米波信号 [1, 17, 20, 81, 91] , 以训练感知模型。一些工作将两种或三种传感器结合起来进行模拟，例如激光雷达与惯性传感器 [63] , 或激光雷达与RGB相机 [56]。较新的工作会先在模拟数据上预训练激光雷达和毫米波模型，再将其适配到真实数据 [3, 12, 28], ; 也有工作合成传感器-文本数据，将惯性信号与文本对齐 [92], 并训练用于毫米波的语言模型 [41, 83]。这些模拟器都没有将毫米波或WiFi与相机或激光雷达配对。据我们所知，我们是首个从相同的人体动作模拟全部五种传感器，并在这些数据上联合预训练多传感器模型的研究。

3. 方法

令 $\mathcal{M} = \{\text{I、D、L、R、W}\}$ 表示五种传感器（RGB、深度、激光雷达、毫米波和 WiFi），并令 $S \subseteq \mathcal{M}$ 表示样本中可用的非空传感器子集。每个传感器$m \in S$在 T帧上观测受试者，得到 $\mathbf{x}_1^m$、.....、$\mathbf{x}_T^m$，我们将其称为该传感器的观测。在帧 t ，我们用 J=17 个关节 $\mathbf{X}_t \in \mathbb{R}^{J\times 3}$表示人体，并将其分解为骨盆轨迹 $\mathbf{p}_t$ 和局部姿态 $\mathbf{P}_t = \mathbf{X}_t - \mathbf{1}\mathbf{p}_t^\top$。一个在 AMASS [51] 上训练的冻结向量量化（VQ）分词器 [70] 将 $\mathbf{P}_t$ 的每个关节 j 量化为离散运动令牌$u_{t,\ j}$，而轨迹则保持连续。对于任意 S，OmniSenMo 都能估计每一帧的三维人体姿态 $\mathbf{X}_t$ ，通过动作识别、动作问答和描述生成理解动作，并预测未来运动；所有这些任务均由统一的多感官语言模型完成。如图2所示，OmniSenMo由以骨架为基础的传感器编码器和一个语言模型组成。前者将每种传感器映射为统一表征，使模型能够融合任意传感器子集并恢复缺失传感器（第3.1节）；后者通过 MGO（第3.2节）处理具有不同时间粒度的任务。为缓解数据稀缺问题，多感官仿真到现实预训练先使用完全配对的模拟数据训练整个模型，再将其适配到真实数据（第3.3节）。

3.1. 骨架锚定关节令牌

为了解决传感器令牌的语义不一致和空间错位问题，我们将每个传感器锚定到人体骨架上，而传感器-语言模型则将每个传感器锚定到语言中 [41, 87, 97]。每个传感器在每帧被编码为 J个关节令牌，每个令牌都由其对应关节的三维位置和运动令牌进行监督。不同于原始传感器信号，人体运动可以像一种外语一样由语言模型学习，就像运动—语言模型一样 [35, 46, 49]。

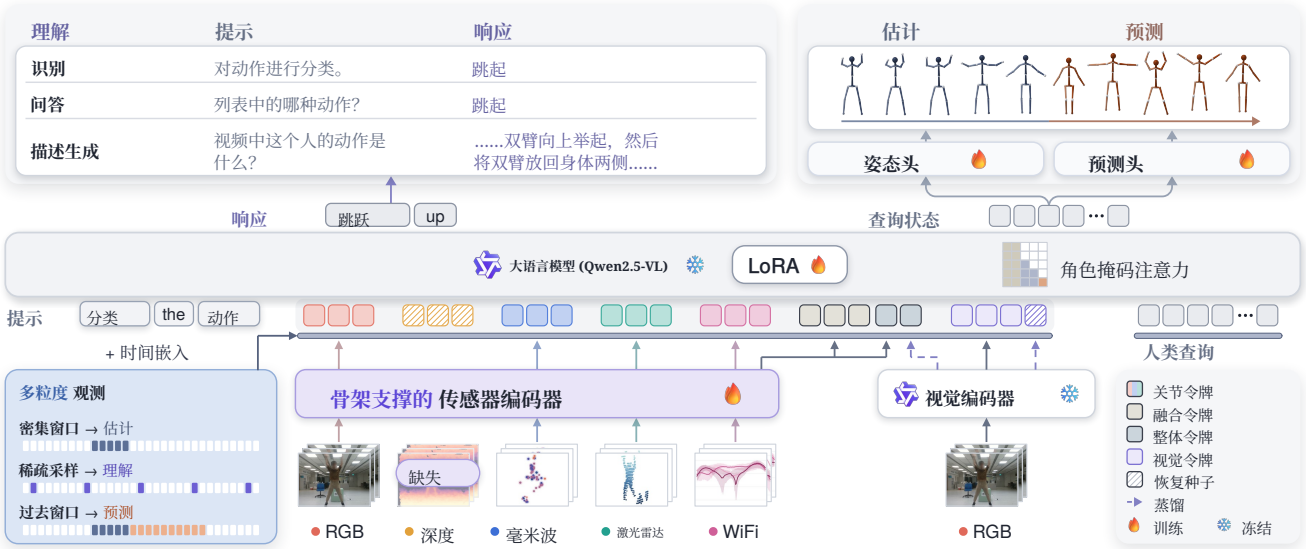


图2。**OmniSenMo 概览**。任意子集中的每种传感器都会经过其骨架锚定编码器，每个关节和帧输出一个关节令牌，并由该关节的三维位置和运动令牌进行监督。学习得到的恢复种子会替代缺失的传感器（此处为深度传感器）。语言模型（采用 LoRA 的 Qwen2.5-VL）的一个上下文包含提示、关节令牌、融合令牌、整体令牌、视觉令牌，以及在角色掩码注意力下的人类查询。在训练过程中，冻结的视觉编码器也充当视觉语言模型教师（虚线所示）。多粒度观测（MGO）设定每项任务使用的帧。查询状态输入姿态预测头和预测头，响应则提供识别、问答和描述文本。估计窗口以蓝色显示，其未来运动以橙色显示。

骨架锚定编码器。 每个编码器都构建于一个在模拟数据上预训练的骨干网络之上，因此在接触真实世界数据之前，它已经具备表征人体运动的能力。每个关节对应的一个可学习嵌入会与每帧的骨干特征进行交叉注意力，时序 Transformer将结果转换为关节特征 $\hat{\mathbf{z}}_{t,j}^{m}$。多层感知机回归每个关节的局部位置 $\mathbf{P}_{t,j}^{m}$，线性分类器则将其运动令牌预测为 j 在向量量化码本上的分布 $\hat{\mathbf{u}}_{t,j}^{m}$。骨盆位置 $\hat{\mathbf{p}}_{t}^{m}$由全局主干特征回归得到后，每个编码器都通过定位损失进行训练

$$\begin{aligned}\mathcal{L}_{\text{skel}}^{m} = \; & \text{CE}(\hat{\mathbf{u}}^{m}, u) + \lambda_{\text{loc}}\, e(\hat{\mathbf{P}}^{m}, \mathbf{P}) \\ & + \lambda_{\text{root}}\, e(\hat{\mathbf{p}}^{m}, \mathbf{p}) + \lambda_{\text{abs}}\, e(\hat{\mathbf{X}}^{m}, \mathbf{X}),\end{aligned} \tag{1}$$

其中，CE 表示交叉熵， e 表示对帧和关节求均值得到的欧几里得距离，$\hat{\mathbf{X}}_{t}^{m} = \hat{\mathbf{P}}_{t}^{m} + \mathbf{1}(\hat{\mathbf{p}}_{t}^{m})^{\top}$，而 λ 表示损失权重。这些共享目标使所有传感器的关节令牌具有一致的运动语义。投影器将每帧映射为 J 个关节令牌，而非 RGB 编码器还会为每次观测汇集两个整体令牌$\mathbf{w}^{m}$。

关节对应关系。 每个编码器都按照相同且固定的关节顺序输出关节令牌，令牌 j 由关节 j监督，因此，所有传感器的第 j个令牌都描述同一个关节，例如左手腕，无论它们来自 RGB 还是 WiFi。这消除了空间错位，使模型能够逐关节融合传感器信息。融合后的令牌会关注每个可用传感器的令牌，并携带按逐关节权重加权平均的逐传感器姿态，这些权重经归一化后——归一化 S。由于传感器的可靠性因关节而异 [4], 这些权重由每个传感器的姿态误差监督。

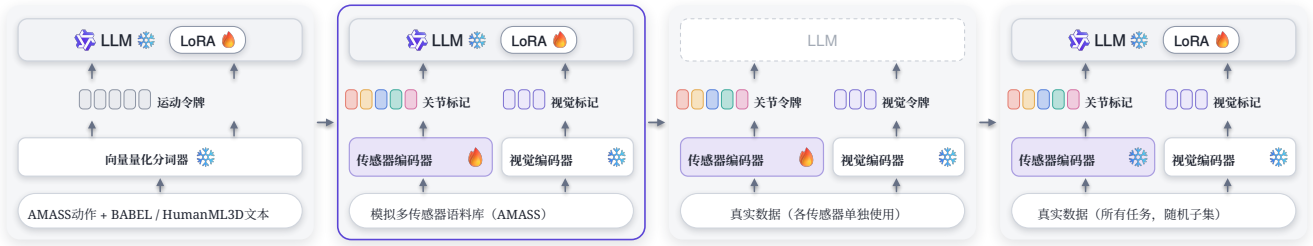
掩码传感器恢复。 由于所有传感器共享一个与骨架对齐的令牌空间，缺失传感器只会在该空间中留下一个缺失块。每个传感器都学习了恢复种子，在语言模型的上下文中，这些种子会取代其关节令牌。我们随机屏蔽传感器，并让被屏蔽传感器的种子 m 回归编码器从被屏蔽的输入中计算出的特征，

$$\begin{aligned}\mathcal{L}_{\text{rec}}^{m} = \; & \ell(\tilde{\mathbf{z}}^{m}, \mathbf{z}^{m}) + \lambda_{g}\, \ell(\tilde{\mathbf{g}}^{m}, \mathbf{g}^{m}) \\ & + \lambda_{p}\, \text{MSE}(\tilde{\mathbf{p}}^{m}, \hat{\mathbf{p}}^{m}),\end{aligned} \tag{2}$$

其中 $\mathbf{z}^{m}$ 是关节特征，$\mathbf{g}^{m}$ 是其在关节上的均值，$\hat{\mathbf{p}}^{m}$ 是编码器的骨盆位置，波浪号表示种子给出的预测结果，而 ℓ 是层归一化特征之间的均方误差。在推理时，同一组种子会取代所有缺失的传感器，如图 2 中深度传感器的示例所示。2。

3.2. 面向多粒度任务的统一语言模型

这些任务的输出和时间粒度各不相同，但 OmniSenMo 使用同一个语言模型和上下文，即可通过任意传感器子集完成所有任务。我们采用 Qwen2.5-VL-7B [8] 作为语言模型，并将其冻结后使用 LoRA 进行适配 [33]。其上下文包含提示、每个传感器的关节令牌（对于缺失的传感器，则使用恢复种子代替）、第3.1节中的融合令牌和整体标记、



(a) 阶段 I：动作-语言预训练　(b) 阶段 II：仿真预训练　(c) 真实数据适配　(d) 第三阶段：多任务微调

图3. **从仿真到真实数据**（第 3.3节）：各阶段中训练（火焰）和冻结（雪花）的部分。虚线部分不参与使用。融合模块、人类查询和预测头（图 2）与 LoRA 一同训练。第二阶段使用模拟多感官语料库进行训练，该语料库由阶段 I 的 AMASS动作构建。在 XRF55 上，第三阶段还会微调每个编码器的一部分（见补充材料）。

冻结的视觉编码器生成的视觉令牌，以及人类查询，最后是响应。给定任务提示后，响应会回答识别和问答问题并生成描述，而人类查询则提供用于估计和预测的姿态。人类查询包含每一帧骨盆和每个关节各一个查询，以及针对整个观测的两个查询。它们按照与关节令牌相同的关节和帧进行索引，因此每个查询都可以关注每个传感器中的对应关节。姿态头回归姿态，预测头预测未来运动。注意力按照 UniPose [46], 中的方式进行角色掩码：人类查询之间采用双向注意力，响应采用因果注意力，且二者都会关注关节令牌。

多粒度观测（MGO）。为处理不同任务在时间粒度上的差异，MGO 使每项任务都能在各自的帧上使用相同的关节标记，这些帧由三种观测协议之一选定，

$$\mathcal{F}_{\mathrm{d}} = \{t, \ldots, t{+}T{-}1\}, \quad \mathcal{F}_{\mathrm{p}} = \{t{-}T{+}1, \ldots, t\}, \\ \mathcal{F}_{\mathrm{s}} = \{f_i \in \mathcal{I}_i\}_{i=1}^{T}. \tag{3}$$

如图2所示，估计使用密集窗口 $\mathcal{F}_{\mathrm{d}}$ 连续帧。理解使用稀疏采样$\mathcal{F}_{\mathrm{s}}$，从 f_i 标注动作区间的每个 T 等长分段$\mathcal{I}_i$ 中各取一帧，与 HoloLLM 中的做法相同 [97]。运动预测使用过去窗口 $\mathcal{F}_{\mathrm{p}}$，并在没有未来传感器数据的情况下预测后续帧。每个关节令牌和每个逐帧人类查询都携带其所在帧的时间嵌入，因此，相同的人类查询既能估计观测帧的姿态，也能预测未来帧的姿态。

3.3. 多传感器仿真到真实预训练

模拟多感官语料库。为缓解配对多传感器数据匮乏的问题，我们从 AMASS [51] 运动的 SMPL-X [58] 网格序列中模拟每种传感器；BABEL [60]和 HumanML3D [26]为这些数据标注动作和文本。我们从网格中光栅化生成 RGB 和深度数据，利用激光雷达向网格发射射线，并将每个网格三角形散射的路径叠加，生成毫米波和 WiFi 信号。虚拟传感器按照 MM-Fi 中的设置，在相同帧上观测每个序列。因此，每段模拟运动都会生成五种同步的传感器流，并与从网格回归得到的 3D 关节以及其动作标签和描述文本配对。

来自视觉语言模型教师的蒸馏。仿真使用三维姿态监督编码器。为迁移视觉语言模型从大量且多样的图像中获得的泛化能力，我们使用视觉语言模型教师对非 RGB 编码器进行蒸馏；该教师是语言模型的冻结视觉编码器，在同一帧的同步 RGB 图像上计算得出。令$\mathbf{y}_t$ 表示帧 t 的归一化教师特征，$\bar{\mathbf{y}}$ 表示其在观测期间的均值；令$\mathbf{s}_t^m$表示编码器 m 的帧特征，$\bar{\mathbf{w}}^m$表示其整体令牌的均值。使用线性投影$\mathbf{A}^m$，

$$\mathcal{L}_{\mathrm{KD}}^{m} = \frac{1}{T}\sum_{t=1}^{T} d\big(\mathbf{A}^m \mathbf{s}_t^m, \mathbf{y}_t\big) + d\big(\mathbf{A}^m \bar{\mathbf{w}}^m, \bar{\mathbf{y}}\big), \tag{4}$$

其中 $d(\mathbf{a},\ \mathbf{b}) = 1 - \cos(\mathbf{a},\ \mathbf{b})$。第一项匹配每一帧，第二项匹配整个观测。RGB 恢复种子预测与 $\tilde{\mathbf{y}}_t$ 相同的教师特征，其中 $\mathcal{L}_{\mathrm{rec}}^{\mathrm{VLM}} = \frac{1}{T}\sum_t d(\tilde{\mathbf{y}}_t,\ \mathbf{y}_t)$。这两种损失都不使用动作标签。

预训练。如图所示，训练过程从运动到模拟传感器，再到真实数据。 3。在 阶段 *I*中，在连接任何传感器之前，运动令牌会扩展语言模型的词汇表，模型通过掩码运动建模学习这些令牌 [14, 27] 在 AMASS 上进行训练，并在 BABEL [60] 和 HumanML3D 上学习运动与文本之间的转换 [26],，从而能够联合处理姿态和文本。 第二阶段 随后在模拟语料库上预训练整个模型。首先单独训练每个编码器，使其学习定位损失 $\mathcal{L}_{\mathrm{skel}}$，对于非 RGB 传感器，还使用 $\mathcal{L}_{\mathrm{KD}}$。然后，整个模型从随机掩码传感器子集中学习。由于每个模拟传感器观测到的都是同一个带标签的姿态，因此任何子集估计出的姿态都以该姿态作为监督信号，而恢复损失$\mathcal{L}_{\mathrm{rec}}$ 则以被遮蔽的传感器本身为目标。RGB 恢复种子通过 $\mathcal{L}_{\mathrm{rec}}^{\mathrm{VLM}}$学习教师特征，而 $\mathcal{L}_{\mathrm{KD}}$ 的第一项则持续蒸馏非 RGB 编码器。模型还会学习预测未来运动，并通过语言模型，利用 BABEL 的动作标签和 HumanML3D 的描述文本来描述运动。

仿真到现实适应。我们将仿真到现实差距视为传感器信号之间的差距，而非运动或其

表2。**未见过的 MM-Fi 动作上的主要结果**，所有任务均使用同一个 OmniSenMo 模型。传感器列于括号中（否则表示全部五种）。(a) Avg_{31}：对全部 31 个传感器子集的 MPJPE 求平均。(b) 校准后 Top-1（%），括号中为未校准结果。(c) 对未来 1 秒内的 MPJPE 求平均（Avg），以及 1 秒时的 MPJPE；预测依据为 OmniSenMo 估计的过去序列（上）或毫米波（下；HumanMAC 使用 mmPred 的估计结果）。粗体：各区块中的最佳结果。–：不可用。 †我们的重新实现。 §加权和融合。

(a) 三维姿态估计（毫米）

方法	MPJPE↓	姿态对齐↓	平均根位置误差↓	平均值$_{31}$↓
X-Fi [16]	105.1	66.7	75.9	123.9
LEO-Fuse† [4]	89.0	58.2	67.3	101.3
OmniSenMo	**84.1**	**55.0**	**64.6**	**95.6**
AWC† [61] (I+L+R+W)	91.8	61.0	66.4	–
OmniSenMo（I+L+R+W）	**86.5**	**55.2**	**65.6**	–
DAPT [3]（L）	90.6	60.8	74.8	–
OmniSenMo (L)	**88.9**	**55.8**	**70.3**	–

(b) 运动理解

方法	HAR↑	QA↑	METEOR↑
HoloLLM§ [97]	40.8(32.3)	39.7 (32.5)	18.1
HoloLLM (I+R+W)§	42.8(32.0)	41.3 (32.2)	18.1
OmniSenMo	**69.2**（61.7）		**19.7**
Qwen2.5-VL-ZS（I）	42.7（36.1）	42.3（25.8）	7.6
Qwen2.5-VL-SFT（I）	64.9（61.4）	**63.9**（62.9）	17.6
HoloLLM（I）	31.5（32.3）	31.1（32.5）	18.4
OmniSenMo (I)	**68.2**(62.4)	5	**19.9**

(c) 运动预测（毫米）

方法	平均↓	ls↓
siMLPe [29]	115.4	143.0
AuxFormer [79]	121.2	150.5
HumanMAC [14]	121.6	151.4
OmniSenMo	**112.8**	1
mmPred [22]（R）	132.3	154.1
HumanMAC(R)	136.3	159.1
OmniSenMo (R)	**1**	

语言之间的差距。因此，真实数据适配 在真实训练集上使用相同的编码器损失分别训练每个编码器，使其将真实信号映射到骨架上，同时冻结投影器、融合模块和语言模型在仿真中学到的内容。由于每个编码器都从仿真数据和真实数据中生成相同的关节标记，因此在仿真中学习到的运动能够迁移到真实数据。第三阶段 随后在真实训练动作上微调模型，同时冻结编码器；每次更新执行一个任务，并为每个样本随机选取一个传感器子集。网络和训练细节见补充材料。

4. 实验

4.1. 设置

数据集与划分。 我们的主要基准是 MM-Fi 的跨动作划分 [85], , 该基准将五种传感器（I：RGB，D：深度，L：激光雷达，R：毫米波，W：WiFi）与三维人体姿态配对。我们按照其官方代码，在其中 21 个动作上进行训练，并在另外 6 个动作上进行测试。XRF55 [73]是一个不包含三维姿态的室内动作识别数据集。在 HoloLLM [97] 所使用的、具有公开 Kinect RGB 数据的子集上，我们使用 RGB、深度、毫米波和 WiFi，在 44 个动作上进行训练，并在另外 11 个动作上进行测试。对于理解任务，我们使用 HoloLLM 在这些跨动作划分下的语言基础基准。

未见动作。 在任一数据集上，若某个动作在训练中没有使用其真实传感器记录，且没有任何模型（包括所有基线方法）以基于该动作记录训练的权重进行初始化，则该动作视为未见动作。因此，姿态估计和运动预测用于测试跨动作泛化，识别和问答则用于测试未见动作识别。

评测方案与指标。 所有评测方案均使用 T=5 帧，帧率为 10Hz。估计报告绝对坐标下的每关节平均位置误差（MPJPE）、经过Procrustes对齐的MPJPE（PA-MPJPE）以及平均根位置误差（MRPE） [54]。预测报告未来十帧（0.1–1秒，如mmPred [22]中所示）单次预测的相对于骨盆的MPJPE。描述生成报告METEOR [9] 和BERTScore [90]。动作识别（HAR）和问答（QA）报告在未见动作上的分类Top-1准确率 [12], , 其中每种方法在验证集 [23, 32, 47, 96, 98]上进行无标签校准后，选择对数似然 [2, 19]最高的候选项。配对95%置信区间根据测试受试者计算；对于消融实验，则根据测试序列计算。

基线方法。 对于估计任务，我们将 X-Fi [16]和 LEO-Fuse [4]，（可融合任意传感器子集）、AWC [61]，（融合固定的四种传感器）以及用于激光雷达的 DAPT [3] 作为对比方法。对于理解任务，我们将 HoloLLM 和 Qwen2.5-VL-7B [8] 作为对比方法；后者使用 RGB 输入，以零样本（ZS）方式运行，或在不使用仿真的情况下，使用相同的真实样本进行微调（SFT）。这些方法与 OmniSenMo 接收相同的五帧。对于运动预测，siMLPe [29], 、AuxFormer [79],和 HumanMAC [14]对 O mniSenMo 估计的历史运动进行外推，而 mmPred [22] 则根据毫米波进行预测。除零样本 Qwen2.5-VL 外，所有基线方法均在我们的数据划分上训练；LEO-Fuse 和 AWC 则为重新实现。

4.2. 主要结果

在未见过的 MM-Fi 动作上（表 2），采用全部五种传感器的统一 OmniSenMo 模型在每项任务上的表现均优于对比模型。

估计。使用全部五种传感器时，OmniSenMo 的姿态误差最低，比 LEO-Fuse 低 5.5%，比 X-Fi 低 20%。在 AWC 融合的固定四种传感器组合上，其表现也优于 AWC；仅使用激光雷达时，也优于专用于激光雷达的模型 DAPT。

理解。在未见动作上，OmniSenMo 的识别和问答表现均比 HoloLLM 和零样本 Qwen2.5-VL 高出 25 个百分点以上；无论是否经过校准，这两种基线方法的得分都低于 43%。OmniSenMo 生成的描述文本也优于 HoloLLM，BERTScore 同样更高（46.3 对比 45.3）。仅提供 RGB 数据时，在相同的带时间戳帧和真实样本上微调的 Qwen2.5-VL 表现与 OmniSenMo 不相上下，但每条观测使用 650 个视觉令牌，而我们仅使用 20 个。使用全部五种传感器时，OmniSenMo 的识别和回答准确度均高于该 Qwen2.5-VL。

预测。 OmniSenMo 的预测结果无论是在平均表现上还是

表3。**按传感器子集划分的结果**：将相同的 OmniSenMo 模型应用于各个子集，并与使用相同子集的模型进行比较。HAR 和问答：校准后 Top-1（%）。粗体：最佳。 †我们的重新实现。§加权和融合。

传感器	MPJPE↓ X-Fi	LEO-Fuse†	我	HAR↑ oloLLM	我	QA↑ oloLLM	我
I	133.6	103.3	**90.1**	31.5	**68.2**	31.1	**55.6**
D	118.3	98.5	**90.9**	25.7	**59.0**	23.2	**53.4**
L	163.6	93.6	**88.9**	33.2	**58.0**	27.1	**57.8**
R	160.6	**153.1**	154.8	25.4	**29.8**	20.1	**29.5**
W	228.8	212.6	**192.1**	10.2	**20.5**	11.7	**27.1**
R+W	157.4	150.7	**144.2**	13.3§	**27.8**	14.6§	**32.4**
D+L+R+W	106.6	90.8	**87.7**	21.6§	**60.5**	20.3§	**64.6**
I+D+L+R+W	105.1	89.0	**84.1**	40.8§	**69.2**	39.7§	**73.4**

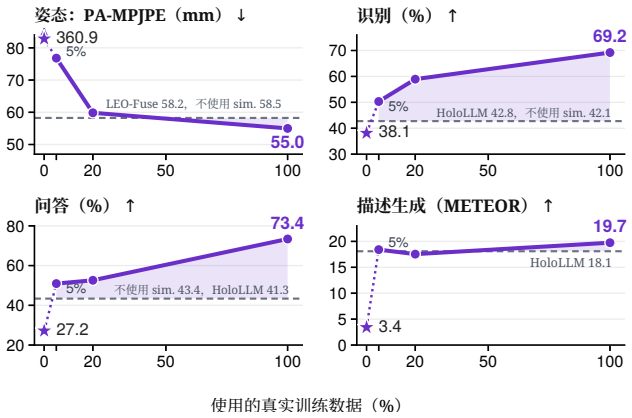


图4. **未见动作上的真实数据效率**，使用全部五种传感器。星号表示仅使用仿真的模型；姿态结果位于坐标轴范围之外；虚线表示使用全部真实数据训练的模型（HoloLLM，使用 I+R+W）。

提前 1 秒的情况下，都比所比较的方法更加准确。在输入相同毫米波数据的情况下，其平均表现也比 mmPred 更准确。

4.3. 传感器子集上的结果

由于缺失的传感器只会将其关节标记转化为恢复种子，同一个 OmniSenMo 模型无需重新训练即可处理 31 种非空传感器子集中的任意一种，因此即使关闭某个传感器，它仍能正常工作。如表3 和补充材料第E节所示，该模型在全部 31 种子集上的姿态误差均低于 X-Fi，在其中 28 种子集上的姿态误差低于 LEO-Fuse。在单传感器设置中，激光雷达的姿态误差最低，而 RGB 的识别准确率最高。使用全部五种传感器时，问答准确率比表现最佳的单个传感器高 15.6 个百分点。

不使用 RGB。 即使不使用 RGB，OmniSenMo 的表现仍优于使用 RGB 的多传感器基线方法。仅凭另外四种传感器，它估计姿态的准确度就高于使用全部五种传感器的 LEO-Fuse 和 X-Fi；识别未见动作的表现优于使用 RGB 的 HoloLLM（60.5% 对比 42.8%），问答准确率也高于仅使用 RGB 的情况（64.6% 对比 55.6%）。使用单一传感器时，其姿态误差平均比 LEO-Fuse 低 6.7%，并且每种传感器上的识别和问答准确率都高于 HoloLLM。

表4。**消融实验**：移除 OmniSenMo 的一个组件或更改关节标记，在五种传感器组合上取平均值，每种传感器数量对应一种组合（I+D+L+R+W、D+L+R+W、I+D+L、R+W 和 I）。括号内为相较于 OmniSenMo 的变化；若配对 95% 置信区间包含零，则以灰色标示。

变体	MPJPE↓	HAR↑	QA↑
OmniSenMo	97.8	59.8	60.0
无关节令牌注意力	97.8 (0.0)	38.8 (↓21.0)	37.9 (↓22.1)
池化令牌	101.2 (↑3.4)	54.6 (↓5.2)	53.8 (↓6.2)
不使用关节对应关系	105.6 (↑7.8)	43.6 (↓16.2)	47.5 (↓12.5)
不使用恢复	97.8 (0.0)	58.4 (↓1.4)	49.1 (↓10.9)
无仿真	101.6 (↑3.8)	37.1 (↓22.7)	38.5 (↓21.5)
无视觉语言模型教师	97.8 (0.0)	56.5 (↓3.3)	51.3 (↓8.7)
单一协议	97.8 (0.0)	49.8 (↓10.0)	43.5 (↓16.5)

4.4. 消融实验

表 4 中的每项消融实验都会移除一个组件，并重新训练模型。仿真和关节令牌的影响最大，尤其体现在理解方面。仿真将传感器与远多于真实数据所含动作的数据配对；关节令牌则将编码器从中学到的信息传递给语言模型。因此，移除任一组件都会使识别和问答准确率下降 20 多个百分点。除去不使用仿真、采用池化令牌以及不使用关节对应关系这几种情况外，其余变体的姿态表现几乎不变，因为它们保留了编码器架构和仿真训练数据。

骨架锚定关节令牌。 不使用关节令牌注意力重新训练模型，使理解模块无法关注关节令牌、融合标记或逐帧人体查询，同时保留整体令牌和视觉令牌。池化令牌用通用令牌替代关节令牌。每帧仍提供 J 个令牌，但这些令牌由该帧的骨干特征经过全局平均池化后形成，类似于 CLIP 和 ViT 风格的编码器，并且其中没有任何令牌与某个关节绑定。不使用关节对应关系保留 J 个关注局部特征的令牌，但不将其中任何令牌与关节绑定，并在没有对应关系的情况下融合传感器。两种变体都从编码器开始重新训练。在相同的模拟数据上训练时，无论是估计还是理解，关节令牌都优于这些通用池化令牌；仅使用毫米波和 WiFi 时，姿态差距平均从 3.4 mm 增大到 13.2 mm。仅移除对应关系会使不同传感器的令牌在空间上错位，平均姿态误差增加 7.8 mm，识别和问答分别下降 16.2 和 12.5 个百分点，这表明关节令牌主要通过与关节的对应关系发挥作用，而非依赖于令牌数量。不使用恢复移除恢复损失，因此不会训练缺失传感器的种子来恢复该传感器。问答平均下降 10.9 个百分点，在使用全部传感器时下降 19.5 个百分点。值得注意的是，即使没有传感器缺失，掩码传感器恢复仍能提升问答性能。

多传感器仿真到真实预训练。无仿真



图5. **未见动作的定性结果**。两行预测均以 OmniSenMo 估计的过去姿态为起点。红框放大显示左脚踝。圆点标示每个模型的传感器，灰色骨架为真实值。正确和错误的描述短语分别以绿色和红色标示。

保留阶段 I，但跳过阶段 II，并使编码器从 ImageNet 或随机权重开始。我们还在 5% 和 20% 的真实训练序列上重复进行真实数据适配和阶段 III，如图 4所示。前 20% 的数据带来了相较于仅使用仿真的大部分增益。使用 5% 的数据时，识别和问答的表现已经超过了在全部真实数据上训练的 HoloLLM 和无仿真的 OmniSenMo。因此，仿真缓解了配对真实数据稀缺的问题。 无视觉语言模型教师 不将冻结的视觉语言模型的 RGB 特征蒸馏到非 RGB 编码器或 RGB 恢复种子中，同时 RGB 仍作为输入。视觉语言模型教师主要使未见动作上的问答受益；移除该教师后，问答性能平均下降 8.7 个百分点，而移除 RGB 后下降 8.6 个百分点。

MGO。单一协议在标注区间中心的密集窗口（连续五帧）上训练并测试理解能力。识别和问答表现均有所下降；使用全部传感器时，两种踢腿动作的识别率都降至 1.6%，因为只有观察完整过程才能识别踢腿动作。对关节标记进行线性探针分析，在五种传感器上取平均，并在训练动作的验证窗口上评估，结果表明，从密集窗口中能更准确地解码姿态（47.5对比52.3 mm MPJPE），而从稀疏样本中能更准确地解码动作（100% 对比89.5%，补充材料第 F.5 节）。因此，让帧与各任务的时间粒度相匹配，可使同一模型同时用于估计和理解。

4.5. 在 XRF55 上进行跨数据集验证

为验证该流程在新硬件和未见动作上的表现，我们使用 XRF55 的硬件参数重新模拟毫米波和 WiFi，因为其雷达输出的是热图而非点云，并将模型适配到训练动作的真实片段。使用相同的传感器（表5），除仅使用深度的传感器子集外，模型在识别未见动作方面均远胜 HoloLLM；在仅使用深度时，两者表现相当。对于相同的 RGB 输入，多传感器 OmniSenMo 的表现可与 Qwen2.5-VL 媲美。

4.6. 定性结果

图5 展示了第4.2节在使用全部五种传感器时对未见动作的结果。在估计中，OmniSenMo跟随了LEO-Fuse遗漏的伸展腿，误差大约只有一半

表5. **XRF55 未见动作的结果**。HAR 和 QA：校准后的 Top-1（%），括号内为未校准结果。粗体：各分组中的最佳结果。 §加权和融合。

传感器	方法	HAR↑	QA↑	METEOR↑
I	Qwen2.5-VL-ZS	65.7 (61.0)	**66.3** (59.7)	7.3
	Qwen2.5-VL-SFT	64.8(61.9)	65.9 (63.2)	**19.6**
	HoloLLM [97]	24.6(17.5)	25.2 (14.2)	18.2
	OmniSenMo	**65.8** (59.8)	(	19.5
D	HoloLLM	**26.7**(18.0)	25.2 (16.5)	**18.5**
	OmniSenMo	26.4 (22.1)	;	**18.5**
W	HoloLLM	13.1(16.4)	12.9 (14.8)	**17.4**
	OmniSenMo	**21.6** (15.6)	;	**17.4**
I+D+W	HoloLLM§	20.3 (19.7)	22.5(16.0)	18.3
	OmniSenMo	**63.4** (56.5)	(	**18.5**
I+D+R+W O		64.3 (56.6)	(	18.4

即LEO-Fuse在左脚踝处的误差。在理解中，它正确指出了一次跳跃，而HoloLLM误认为是鞠躬；其图注描述了手臂运动，而HoloLLM的图注描述的是弓步。在预测中，siMLPe根据同样估计的过去姿态使受试者保持前倾，而OmniSenMo则预测了从鞠躬中起身。

5. 结论

我们提出了 OmniSenMo，这是一种统一多感官语言模型，可利用 RGB、深度、激光雷达、毫米波和 WiFi 中的任意传感器子集来估计、理解和预测人体运动。人体骨架作为通用参考，其关节令牌解决了异构传感器间的语义不一致和空间错位问题，并在语言模型从动作捕捉和文本中学到的运动词汇表中，逐关节融合任意传感器子集。多粒度观测使帧与各任务的时间粒度相匹配，而在模拟传感器上进行预训练，则覆盖了真实传感器数据中缺失的动作。在 MM-Fi 的跨动作划分上，使用全部五种传感器的同一模型在各项任务上的表现均优于对比模型；在 XRF55 上进行跨数据集验证时，该模型识别未见动作的能力也远胜 HoloLLM。这些结果表明，骨架可作为人体感知与语言模型之间的桥梁。

大语言模型使用声明

我们使用 Claude（Anthropic）作为编程和写作助手，编写并调试训练与评估代码、启动实验并根据存储的输出计算指标、制作图和表格、检索并总结相关论文、讨论方法和实验的设计方案，以及起草和编辑文本（包括语法修订）。我们还使用了 OpenAI Codex 编写代码。图 1 中的房屋、机器人和人物插图由 OpenAI 的图像生成工具生成。我们没有使用任何生成式人工智能工具来制作实验数据或结果；Qwen2.5-VL 和 OmniSenMo 中的视觉语言教师模型是方法的组成部分，而非写作辅助工具。所有研究决策均由作者作出或批准；作者核验了每个报告数值与存储输出的一致性，并对本文内容承担全部责任。

致谢

本研究通过美国国家科学基金会资助项目（编号 # 2138259、#2138286、#2138307、#2137603 和 # 2138296），使用了印第安纳大学的 Jetstream2 和国家超级计算应用中心（NCSA）的 Delta 计算资源。该项目由先进网络基础设施协调生态系统：服务与支持（ACCESS）计划分配，分配编号为 CIS262336。

参考文献

[1] Karan Ahuja、Yue Jiang、Mayank Goel 和 Chris Harri- son。Vid2Doppler：从视频合成多普勒雷达数据，用于训练隐私保护活动识别。在*CHI*计算系统人因大会（*CHI*）论文集，2021年。3

[2] Jean-Baptiste Alayrac、Jeff Donahue、Pauline Luc、Antoine Miech、Iain Barr、Yana Hasson、Karel Lenc、Arthur Mensch、Katie Millican、Malcolm Reynolds、Roman Ring、Eliza Rutherford、Serkan Cabi、Tengda Han、Zhitao Gong、Sina Samangooei、Marianne Monteiro、Jacob Menick、Sebastian Borgeaud、Andrew Brock、Aida Nematzadeh、Sahand Sharifzadeh、Mikołaj Bi´nkowski、Ricardo Barreira、Oriol Vinyals、Andrew Zisserman 和 Karen Simonyan。Flamingo：一种用于少样本学习的视觉语言模型。载于 神经信息处理系统进展（*NeurIPS*），2022年。6[3] Xiaoqi An、Lin Zhao、Chen Gong、Jun Li 和 Jian Yang。预训练密度感知姿态变换器，以实现稳健的基于激光雷达的三维人体姿态估计。载于 *AAAI*人工智能会议论文集（*AAAI*），第1755–1763页，2025年。3，6[4] Emrecan Aslan。LEO-Fuse：一种任务通用、模态无关的通用多模态人体感知框架。载于 欧洲计算机视觉会议论文集（*ECCV*），第433–451页，2026年。2，3，4，6[5] Yusuf Aytar、Carl Vondrick 和 Antonio Torralba。SoundNet：从无标注视频中学习声音表征。载于 神经信息处理系统进展（*NeurIPS*），2016年。3

[6] Roman Bachmann、David Mizrahi、Andrei Atanov 和 Amir Zamir。MultiMAE：多模态多任务掩码自动编码器。载于 欧洲计算机视觉会议论文集（*ECCV*），第348–367页，2022年。2[7] Roman Bachmann、O˘guzhan Fatih Kar、David Mizrahi、Ali Garjani、Mingfei Gao、David Griffiths、Jiaming Hu、Afshin Dehghan 和 Amir Za mir。4M-21：一种面向任意输入到任意输出的视觉模型，支持数十种任务和模态。载于 神经信息处理系统进展（*NeurIPS*），第61872–61911页，2024年。2

[8] ShuaiBai、KeqinChen、Xuejing Liu、JialinWang、Wenbin Ge、Sibo Song、Kai Dang、Peng Wang、Shijie Wang、Jun Tang、Humen Zhong、Yuanzhi Zhu、Mingkun Yang、Zhao- hai Li、Jianqiang Wan、Pengfei Wang、Wei Ding、Zheren Fu、Yiheng Xu、Jiabo Ye、Xi Zhang、Tianbao Xie、Zesen Cheng、Hang Zhang、Zhibo Yang、Haiyang Xu 和 Jun- yang Lin。Qwen2.5-VL技术报告。*arXiv*预印本 *arXiv:2502.13923*，2025。4，6

[9] Satanjeev Banerjee和Alon Lavie。METEOR：一种与人工评判机器翻译的指标，与人工判断的相关性更高。载于*ACL*机器翻译和/或摘要内在与外在评估指标研讨会论文集，第65–72页，2005年6–。

[10] Michael J. Black、Priyanka Patel、Joachim Tesch 和 Jin-long Yang。BEDLAM：一个展现细致、逼真的动画运动的人体合成数据集。载于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，第8726–8737页，2023年。3

[11] Hong Cai、Belal Korany、Chitra R Karanam 和 Yasamin Mostofi。教会射频在没有射频训练测量数据的情况下进行感知。*ACM*交互、移动、可穿戴与普适技术汇刊（*IMWUT*）论文集，4(4)：1–22，2020。3

[12] Qiming Cao、Hongfei Xue、Tianci Liu、Xingchen Wang、Haoyu Wang、Xincheng Zhang 和 Lu Su。mmCLIP：通过信号-文本对齐提升基于毫米波的零样本人体活动识别。在 *ACM*嵌入式网络传感器系统会议论文集（*SenSys*），第184–197页，2024年。3，6

[13] Anjun Chen、Xiangyu Wang、Shaohao Zhu、Yanxu Li、Jiming Chen 和 Qi Ye。mmBody 基准：用于毫米波雷达的三维身体重建数据集与分析。在 *ACM* 国际多媒体会议论文集（*MM*），第3501–3510页，2022年。2，3

[14] Ling-Hao Chen、Jiawei Zhang、Yewen Li、Yiren Pang、Xiaobo Xia 和 Tongliang Liu。HumanMAC：用于人体运动预测的掩码动作补全。在 *IEEE/CVF*国际计算机视觉会议论文集（*ICCV*），第9544–9555页，2023年。5，6

[15] Ling-Hao Chen、Shunlin Lu、Ailing Zeng、Hao Zhang、Benyou Wang、Ruimao Zhang 和 Lei Zhang。Motion-LLM：从人体动作和视频中理解人类行为。*IEEE* 模式分析与机器智能汇刊，2025年。1，3

[16] Xinyan Chen 和 Jianfei Yang。X-Fi：用于多模态人体感知的模态不变基础模型。载于国际学习表征会议（ICLR），2025年。2、3、6[17] Xingyu Chen 和 Xinyu Zhang。RF Genesis：通过基于仿真的数据合成和生成式扩散模型，实现毫米波感知的零样本泛化。载于 ACM嵌入式网络传感器系统会议（SenSys）论文集，第28–42页，2023年。3[18] Shenghong Dai、Shiqi Jiang、Yifan Yang、Ting Cao、Mo Li、S uman Banerjee 和 Lili Qiu。Babel：通过可扩展模态对齐实现多模态感知的可扩展预训练模型。载于 ACM嵌入式网络传感器系统会议（SenSys）论文集，第240–253页，2025年。2、3 – 。[19] Wenliang Dai、Junnan Li、Dongxu Li、Anthony Meng Huat Tiong、Junqi Zhao、Weisheng Wang、Boyang Li、Pascale Fung 和 Steven Hoi。InstructBLIP：通过指令微调实现通用视觉-语言模型。载于神经信息处理系统进展（NeurIPS），2023年。6[20] Kaikai Deng、Dong Zhao、Qiaoyue Han、Zihan Zhang、Shuyue Wang、Anfu Zhou 和 Huadong Ma。Midas：从视频生成毫米波雷达数据，用于普适且保护隐私的人体感知任务训练。ACM交互、移动、可穿戴与普适技术汇刊（IMWUT），7(1)：9:1–9:26，2023年。3[21] Junqiao Fan、Jianfei Yang、Yuecong Xu 和 Lihua Xie。扩散模型是基于三维射频视觉进行姿态估计的理想方法。载于欧洲计算机视觉会议（ECCV），2024。3[22] Junqiao Fan、Haocong Rao、Jiarui Zhang、Jianfei Yang 和 Lihua Xie。mmPred：黑暗环境下基于雷达的人体运动预测。载于 AAAI人工智能会议（AAAI）论文集，第3777–3785页，2026年。2、3、6[23] Yu Fei、Yifan Hou、Zeming Chen 和 Antoine Bosselut。缓解上下文学习中的标签偏差。载于计算语言学协会年会（ACL）论文集，第14014–14031页，2023年。6[24] Yao Feng、Jing Lin、Sai Kumar Dwivedi、Yu Sun、Priyanka Patel 和 Michael J. Black。ChatPose：聊聊三维人体姿态。载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第2093–2103页，2024。1、3[25] Rohit Girdhar、Alaaeldin El-Nouby、Zhuang Liu、Mannat Singh、Kalyan Vasudev Alwala、Armand Joulin 和 Ishan Misra。ImageBind：一个嵌入空间绑定所有模态。载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第15180–15190页，2023年。3[26] Chuan Guo、Shihao Zou、Xinxin Zuo、Sen Wang、Wei Ji、Xingyu Li 和 Li Cheng。从文本生成多样且自然的三维人体动作。载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第5152–5161页，2022年。2、5[27] Chuan Guo、Yuxuan Mu、Muhammad Gohar Javed、Sen Wang 和 Li Cheng。MoMask：生成式掩码模型——

三维人体运动。在IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第1900–1910页，2024。3, 5[28] Cheng Guo、Qiming Cao、Shengkai Xu、Haoyu Xie、Kaixiang Su、Pu Wang和Hongfei Xue。mmSimPrior：学习仿真先验，以实现高效数据利用且具有良好泛化能力的真实世界雷达人体运动重建。arXiv预印本 arXiv:2607.22973，2026。3[29] Wen Guo、Yuming Du、Xi Shen、Vincent Lepetit、Xavier Alameda-Pineda和Francesc Moreno-Noguer。回归多层感知机：一种简单的人体运动预测基线方法。载于IEEE/CVF冬季计算机视觉应用会议（WACV）论文集，第4809 4819页，2023 6 – 。[30] Saurabh Gupta、Judy Hoffman和Jitendra Malik。用于监督迁移的跨模态蒸馏。载于IEEE计算机视觉与模式识别会议（CVPR）论文集，第2827–2836页，2016。3[31] Jiaming Han、Kaixiong Gong、Yiyuan Zhang、Jiaqi Wang、Kaipeng Zhang、Dahua Lin、Yu Qiao、Peng Gao和Xi- angyu Yue。OneLLM：一种以语言对齐所有模态的统一框架。载于IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第26584–26595页，2024。2, 3[32] Ari Holtzman、Peter West、Vered Shwartz、Yejin Choi和Luke Zettlemoyer。表面形式竞争：为什么概率最高的答案并不总是正确的。载于自然语言处理实证方法会议（EMNLP）论文集，第

7038 7051, 2021 6 –页。

[33] Edward J. Hu、Yelong Shen、Phillip Wallis、Zeyuan Allen- Zhu、Yuanzhi Li、Shean Wang、Lu Wang和Weizhu Chen。LoRA：大语言模型的低秩适配。载于国际表征学习会议（ICLR），2022。4[34] Yinghao Huang、Manuel Kaufmann、Emre Aksan、Michael J. Black、Otmar Hilliges和Gerard Pons-Moll。Deep inertial poser：学习通过稀疏惯性测量实时重建人体姿态。ACM图形学汇刊（TOG），37(6):1–15，2018。3[35] Biao Jiang、Xin Chen、Wen Liu、Jingyi Yu、Gang Yu和Tao Chen。MotionGPT：将人体运动视为外语。载于神经信息处理系统进展（NeurIPS），第20067–20079页，2023。2, 3[36] Siyang Jiang、Mu Yuan、Xiang Ji、Bufang Yang、Zeyu Liu、Lilin Xu、Yang Li、Yuting He、Liran Dong、Wenrui Lu、Zhenyu Yan、Xiaofan Jiang、Wei Gao、Hongkai Chen和Guoliang Xing。用于人类活动场景理解与推理的大规模多模态数据集和基准测试。载于移动系统、应用与服务年度国际会议（MobiSys）论文集，第352–370页，2026。2, 3[37] Will Kay、Joao Carreira、Karen Simonyan、Brian Zhang、Chloe Hillier、Sudheendra Vijayanarasimhan、Fabio Viola、Tim Green、Trevor Back、Paul Natsev、Mustafa Suleyman和Andrew Zisserman。Kinetics人体动作视频数据集。arXiv预印本 arXiv:1705.06950，2017。2[38] Moo Jin Kim、Karl Pertsch、Siddharth Karamcheti、Ted Xiao、Ashwin Balakrishna、Suraj Nair、Rafael Rafailov、Ethan

Foster、Grace Lam、Pannag Sanketi、Quan Vuong、Thomas Kollar、Benjamin Burchfiel、Russ Tedrake、Dorsa Sadigh、Sergey Levine、Percy Liang和Chelsea Finn。OpenVLA：开源视觉-语言-动作模型。载于机器人学习会议（CoRL）论文集，第2679–2713页，2024。1[39] Thomas Kreutz、Max M¨uhlh¨auser和Alejandro Sa nchez Guinea。DeSPITE：探索对比式深度骨架-点云-IMU-文本嵌入，以提升点云人体活动理解能力。载于IEEE/CVF国际计算机视觉会议（ICCV）论文集，第14633–14643页，2025。3[40] Hyeokhyen Kwon、Catherine Tong、Harish Haresamudram、Yan Gao、Gregory D. Abowd、Nicholas D. Lane和Thomas Pl¨otz。IMUTube：从视频中自动提取虚拟穿戴式身体加速度测量数据，用于人体活动识别。ACM交互、移动、可穿戴与普适技术汇刊（IMWUT），4 29, 2020 3 87 1 87 3 ( ): : – : 。[41] Zengyuan Lai、Jiarui Yang、Songpengcheng Xia、Lizhou Lin、Lan Sun、Renwen Wang、Jianran Liu、Qi Wu和Ling Pei。RadarLLM：赋能大语言模型理解来自毫米波点云序列的人体运动。载于AAAI人工智能会议（AAAI）论文集，第5791–5799页，2026。2, 3 [42] Junnan Li、Dongxu Li、Silvio Savarese和Steven Hoi。BLIP-2：利用冻结图像编码器和大语言模型进行语言-图像预训练。载于国际机器学习会议（ICML）论文集，第19730–19742页，2023。2[43] Manling Li、Shiyu Zhao、Qineng Wang、Kangrui Wang、Yu Zhou、Sanjana Srivastava、Cem Gokmen、Tony Lee、Li Er- ran Li、Ruohan Zhang、Weiyu Liu、Percy Liang、Li Fei-Fei、Jiayuan Mao和Jiajun Wu。具身智能体接口：具身决策大语言模型基准测试。载于神经信息处理系统进展（NeurIPS）数据集与基准赛道论文集，第100428 100534页，2024 1 – 。[44] Tianhong Li、Lijie Fan、Mingmin Zhao、Yingcheng Liu和Dina Katabi。让不可见变得可见：穿墙和遮挡条件下的动作识别。载于IEEE/CVF国际计算机视觉会议（ICCV）论文集，2019。2[45] Yanwei Li、Chengyao Wang和Jiaya Jia。LLaMA-VID：一张图像在大语言模型中值2个词元。载于欧洲计算机视觉会议（ECCV）论文集，第323–340页，2024。2[46] Yiheng Li、Ruibing Hou、Hong Chang、Shiguang Shan和Xilin Chen。UniPose：用于人体姿态理解、生成和编辑的统一多模态框架。载于IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第27805–27815页，2025。1, 2, 3, 5[47] Zhiqiu Lin、Xinyue Chen、Deepak Pathak、Pengchuan Zhang和Deva Ramanan。重新审视语言先验在视觉-语言模型中的作用。载于国际机器学习会议（ICML），2024。6

[48] Haotian Liu、Chunyuan Li、Qingyang Wu 和 Yong Jae Lee。视觉指令调优。载于《神经信息处理系统进展》（NeurIPS），第 34892–34916 页，2023年。1[49] Mingshuang Luo、Ruibing Hou、Zhuo Li、Hong Chang、Zimo Liu、Yaowei Wang 和 Shiguang Shan。M^{3}GPT：用于动作理解与生成的先进多模态多任务框架。载于《神经信息处理系统进展》（NeurIPS），第 28051 28077 页，2024 2, 3 – 。[50] Mengmeng Ma、Jian Ren、Long Zhao、Sergey Tulyakov、Cathy Wu 和 Xi Peng。SMIL：严重模态缺失条件下的多模态学习。载于 AAAI人工智能会议（AAAI）论文集，第 2302–2310 页，2021年。2[51] Naureen Mahmood、Nima Ghorbani、Nikolaus F. Troje、Ger-ard Pons-Moll 和 Michael J. Black。AMASS：作为表面形状的动作捕捉档案。载于 IEEE/CVF国际计算机视觉会议（ICCV）论文集，第 5441–5450 页，2019。2, 3, 5[52] Thomas Markhorst、Zhi-Yi Lin、Jouh Yeong Chew、Jan van Gemert 和 Xucong Zhang。利用语言模型进行细粒度人体运动理解。arXiv 预印本 arXiv:2606.20888，2026年。2, 3[53] David Mizrahi、Roman Bachmann、O˘guzhan Fatih Kar、Teresa Yeo、Mingfei Gao、Afshin Dehghan 和 Amir Zamir。4M：大规模多模态掩码建模。载于《神经信息处理系统进展》（NeurIPS），第 58363–58408 页，2023年。2[54] Gyeongsik Moon、Ju Yong Chang 和 Kyoung Mu Lee。基于相机距离感知的自顶向下方法，从单张 RGB 图像进行三维多人姿态估计。载于 IEEE/CVF国际计算机视觉会议（ICCV）论文集，第 10132–10141 页，2019。6[55] Girish Narayanswamy、Xin Liu、Kumar Ayush、Yuzhe Yang、Xuhai Xu、Shun Liao、Jake Garrison、Shyam A. Tailor、Ja-cob Sunshine、Yun Liu、Tim Althoff、Shrikanth Narayanan、Pushmeet Kohli、Jiening Zhan、Mark Malhotra、Shwetak Patel、Samy Abdel-Ghaffar 和 Daniel McDuff。可穿戴基础模型的规模化。载于国际学习表征会议（ICLR），第 59382 59413 页，2025 2 – 。[56] Zhiyu Pan、Zhicheng Zhong、Wenxuan Guo、Yifan Chen、Jianjiang Feng 和 Jie Zhou。LiCamPose：融合多视角激光雷达与 RGB 相机，实现稳健的单时间戳三维人体姿态估计。载于 IEEE/CVF冬季计算机视觉应用会议（WACV）论文集，第 2484–2494 页，2025。3[57] Priyanka Patel、Chun-Hao P. Huang、Joachim Tesch、David T. Hoffmann、Shashank Tripathi 和 Michael J. Black。AGORA：针对回归分析优化的地理分布中的虚拟化身。载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第 13468–13478 页，2021年。3[58] Georgios Pavlakos、Vasileios Choutas、Nima Ghorbani、Timo Bolkart、Ahmed A. A. Osman、Dimitrios Tzionas 和 Michael J. Black。表现性人体捕捉：三维手部、

面部和身体的单图像捕捉。载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第 10975–10985 页，2019。5[59] Xavier Puig、Eric Undersander、Andrew Szot、Mikael Dal-laire Cote、Tsung-Yen Yang、Ruslan Partsey、Ruta Desai、Alexander William Clegg、Michal Hlavac、So Yeon Min、Vladim´ır Vondru˘s、Theophile Gervet、Vincent-Pierre Berges、John M. Turner、Oleksandr Maksymets、Zsolt Kira、Mrinal Kalakrishnan、Jitendra Malik、Devendra Singh Chaplot、Un-nat Jain、Dhruv Batra、Akshara Rai 和 Roozbeh Mottaghi。Habitat 3.0：人类、虚拟化身与机器人的共居环境。载于国际学习表征会议（ICLR），2024。1[60] Abhinanda R. Punnakkal、Arjun Chandrasekaran、Nikos Athanasiou、Alejandra Quiros-Ramirez 和 Michael J. Black。BABEL：带英语标签的人体、动作与行为。载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第 722–731 页，2021年。2, 5[61] Mengshi Qi、Jiaxuan Peng、Xianlin Zhang 和 Huadong Ma。迈向三维人体姿态估计中的平衡多模态学习。载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第 21231–21241 页，2026年。3, 6[62] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark、Gretchen Krueger 和 Ilya Sutskever。通过自然语言监督学习可迁移的视觉模型。载于国际机器学习会议（ICML）论文集，第 8748–8763 页，2021年。2[63] Yiming Ren、Chengfeng Zhao、Yannan He、Peishan Cong、Han Liang、Jingyi Yu、Lan Xu 和 Yuexin Ma。LiDAR辅助惯性姿态估计器：利用稀疏惯性与激光雷达传感器进行大规模人体动作捕捉。《IEEE可视化与计算机图形学汇刊》（TVCG），29(5):2337–2347，2023年。3[64] Foram N. Shah、Parshwa N. Shah、Muhammad Usama Saleem、Ekkasit Pinyoanuntapong、Pu Wang、Hongfei Xue 和 Ahmed Helmy。翩翩起舞前先行走：通过生成式掩码运动先验实现高保真、可编辑的舞蹈合成。载于 AAAI人工智能会议（AAAI）论文集，第 8796–8804 页，2026年。3[65] Yitong Shen、Cheng Guo、Peiliang Wang、Jingzhe Zhang、Yi Sheng、Haopeng Zhang、Hongfei Xue 和 Yili Ren。Zero-Fi：通过对比式信号-语言对齐实现基于 Wi-Fi 的零样本人体活动识别。arXiv 预印本 arXiv:2607.26381，2026年。3[66] Jeongwan Shin、Jaehyeon Kim、Donguk Ko 和 Jaeho Choi。语言模型能理解毫米波数据吗？对基于毫米波雷达的人体理解进行大语言模型基准测试。载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）成果论文集，第 2208–2219 页，2026年。3[67] Jamie Shotton、Andrew Fitzgibbon、Mat Cook、Toby Sharp、Mark Finocchio、Richard Moore、Alex Kipman 和 Andrew Blake。利用单张深度图像实时识别人体姿态部位。载于 IEEE 会议论文集

发表于计算机视觉与模式识别（CVPR），页码 1297–1304，2011年。1, 3[68] Yixuan Su、Tian Lan、Huayang Li、Jialu Xu、Yan Wang 和 Deng Cai。PandaGPT：一个模型，遵循所有指令。载于驾驭大语言模型研讨会（TLLM）论文集，页码 11–23，2023年。3[69] Joachim Tesch、Giorgio Becherini、Prerana Achar、Anas- tasios Yiannakidis、Muhammed Kocabas、Priyanka Patel 和 Michael J. Black。BEDLAM2.0：合成人体与运动中的摄像机。载于神经信息处理系统进展（NeurIPS）数据集与基准赛道，2025年。3[70] Aaron van den Oord、Oriol Vinyals 和 Koray Kavukcuoglu。神经离散表示学习。载于神经信息处理系统进展（NeurIPS），2017年。3[71] G¨ul Varol、Javier Romero、Xavier Martin、Naureen Mah- mood、Michael J. Black、Ivan Laptev 和 Cordelia Schmid。从合成人体中学习。载于IEEE计算机视觉与模式识别会议（CVPR）论文集，页码 4627–4635，2017年。3[72] Fei Wang、Sanping Zhou、Stanislav Panev、Jinsong Han 和 Dong Huang。Person-in-WiFi：使用WiFi进行细粒度人体感知。载于IEEE/CVF国际计算机视觉会议（ICCV）论文集，页码 5451–5460，2019年。3[73] Fei Wang、Yizhe Lv、Mengdie Zhu、Han Ding 和 Jinsong Han。XRF55：用于室内人体动作分析的射频数据集。载于ACM交互、移动、可穿戴与普适技术汇刊（IMWUT），8(1):21:1–21:34，2024年。2, 3, 6[74] Xinshun Wang、Peiming Li、Ziyi Wang、Zhongbin Fang、Zhichao Deng、Songtao Wu、Jason Li 和 Mengyuan Liu。Superman：统一骨架与视觉，实现人体运动感知与生成。载于IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，页码 38335–38344，2026年。1, 2, 3[75] Zehan Wang、Ziang Zhang、Minjie Hong、Hang Zhang、Lup- ing Liu、Rongjie Huang、Xize Cheng、Shengpeng Ji、Tao Jin、Hengshuang Zhao 和 Zhou Zhao。OmniBind：通过绑定空间实现大规模全方位多模态表征。载于国际学习表征会议（ICLR），页码 20831–20851，2025年。3[76] Ziyi Wang、Xinshun Wang、Shuang Chen、Yang Cong 和 Mengyuan Liu。UniMotion：用于运动、文本与视觉理解和生成的统一框架。载于欧洲计算机视觉会议（ECCV）论文集，页码 664–682，2026年。1, 2, 3[77] Yilin Wen、Hao Pan、Lei Yang、Jia Pan、Taku Komura 和 Wenping Wang。用于从第一人称视角RGB视频进行三维手部姿态估计和动作识别的分层时序Transformer。载于IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，页码 21243–21253，2023年。2[78] Sangmin Woo、Sumin Lee、Yeonju Park、Muhammad Adi Nugroho 和 Changick Kim。迈向缺失模态鲁棒动作识别的良好实践。载于AAAI人工智能会议（AAAI）论文集，页码 2776–2784，2023年。2

[79] Chenxin Xu、Robby T. Tan、Yuhong Tan、Siheng Chen、Xin- chao Wang 和 Yanfeng Wang。辅助任务有助于基于三维骨架的人体运动预测。载于IEEE/CVF国际计算机视觉会议（ICCV）论文集，页码 9509–9520，2023年。6[80] Runsen Xu、Xiaolong Wang、Tai Wang、Yilun Chen、Jiang- miao Pang 和 Dahua Lin。PointLLM：赋能大语言模型理解点云。载于欧洲计算机视觉会议（ECCV）论文集，页码 131–147，2024 年。3[81] Hongfei Xue 、 Qiming Cao 、 Chenglin Miao、Yan Ju、Haochen Hu、Aidong Zhang 和 Lu Su。通过信号增强实现泛化的基于毫米波的人体姿态估计。载于移动计算与网络年度国际会议（MobiCom）论文集，页码 1–15，2023年。3[82] Kangwei Yan、Fei Wang、Bo Qian、Han Ding、Jinsong Han 和 Xing Wei。Person-in-WiFi 3D：基于Wi-Fi的端到端多人三维姿态估计。载于IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，页码 969–978，2024年。3[83] Yifan Yan、Shuai Yang、Xiuzhen Guo、Xiangguang Wang、Wei Chow、Yuanchao Shu 和 Shibo He。mmExpert：整合大语言模型，实现全面的毫米波数据合成与理解。载于移动网络与移动计算理论、算法基础及协议设计国际研讨会（MobiHoc）论文集，页码 1–10，2025年。2, 3[84] Fengyu Yang、Chao Feng、Ziyang Chen、Hyoungseob Park、Daniel Wang、Yi ming Dou、Ziyao Zeng、Xien Chen、Rit Gangopadhyay、Andrew Owens 和 Alex Wong。将触觉与万物相连：学习统一的多模态触觉表征。载于IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，页码 26340–26353，2024年。3[85] Jianfei Yang、He Huang、Yunjiao Zhou、Xinyan Chen、Yue- cong Xu、Shenghai Yuan、Han Zou、Chris Xiaoxuan Lu 和 Lihua Xie。MM-Fi：用于多样化无线感知的多模态非侵入式四维人体数据集。载于神经信息处理系统进展（NeurIPS）数据集与基准赛道，页码 18756–18768，2023年。2, 3, 6[86] Dongqiangzi Ye、Yufei Xie、Weijia Chen、Zixiang Zhou、Lingting Ge 和 Hassan Foroosh。LPFormer：具有多任务网络的激光雷达姿态估计变换器。载于IEEE机器人与自动化国际会议（ICRA）论文集，2024年。3[87] Duo Zhang、Zhehui Yin、Zhiyun Yao、Haotong Qin、Xusheng Zhang、Hongliu Yang、Jianyu Sun、Junzhe Wang、Zizhou Fan、Michele Magno 和 Daqing Zhang。教基础模型读取毫米波：用于人类行为理解的姿态引导运动学表征。arXiv预印本 arXiv:2608.04127，2026年。2, 3[88] Haopeng Zhang、Yili Ren、Haohan Yuan、Jingzhe Zhang 和 Yitong Shen。Wi-Chat：由大语言模型驱动的Wi-Fi感知。arXiv预印本 arXiv:2502.12421，2025年。2, 3[89] Jianrong Zhang、Yangsong Zhang、Xiaodong Cun、Yong Zhang、Hongwei Zhao、Hongtao Lu、Xi Shen 和 Ying Shan。利用离散表示从文本描述生成动作。在《IEEE/CVF计算机视觉与模式识别会议》（CVPR）论文集，14730–14740页，2023年。3[90] Tianyi Zhang、Varsha Kishore、Felix Wu、Kilian Q. Weinberger 和 Yoav Artzi。BERTScore：使用 BERT 评估文本生成。在《国际学习表征会议》（ICLR），2020。6[91] Xiaotong Zhang、Zhenjiang Li 和 Jin Zhang。用于人体动作感知的合成毫米波。在《ACM嵌入式网络传感器系统会议》（SenSys）论文集，377–390页，2022年。3[92] Xiyuan Zhang、Diyan Teng、Ranak Roy Chowdhury、Shuheng Li、Dezhi Hong、Rajesh K. Gupta 和 Jingbo Shang。UniMTS：面向运动时间序列的统一预训练。在《神经信息处理系统进展》（NeurIPS），107469–107493页，2024。3[93] Yiyuan Zhang、Kaixiong Gong、Kaipeng Zhang、Hongsheng Li、Yu Qiao、Wanli Ouyang 和 Xiangyu Yue。Meta-Transformer：多模态学习的统一框架。arXiv预印本 arXiv:2307.10802，2023年。3[94] Mingmin Zhao、Tianhong Li、Mohammad Abu Alsheikh、Yonglong Tian、Hang Zhao、Antonio Torralba 和 Dina Katabi。利用无线电信号进行隔墙人体姿态估计。在《IEEE计算机视觉与模式识别会议》（CVPR）论文集，7356–7365页，2018年。3[95] Mingmin Zhao、Yonglong Tian、Hang Zhao、Mohammad Abu Alsheikh、Tianhong Li、Rumen Hristov、Zachary Kabelac、Dina Katabi 和 Antonio Torralba。基于射频的三维骨架。在《ACM数据通信特别兴趣小组会议》（SIGCOMM）论文集，267–281页，2018年。3[96] Zihao Zhao、Eric Wallace、Shi Feng、Dan Klein 和 Sameer Singh。使用前先校准：提升语言模型的少样本性能。在《国际机器学习会议》（ICML）论文集，12697–12706页，2021年。6[97] Chuhao Zhou 和 Jianfei Yang。HoloLLM：用于语言基础的人体感知与推理的多感官基础模型。在《神经信息处理系统进展》（NeurIPS），73543–73568页，2025，2、3、5、6、8[98] Han Zhou、Xingchen Wan、Lev Proleev、Diana Mincu、Jilin Chen、Katherine Heller 和 Subhrajit Roy。批量校准：重新思考上下文学习与提示工程中的校准。在《国际学习表征会议》（ICLR），2024。6[99] Yunjiao Zhou、He Huang、Shenghai Yuan、Han Zou、Lihua Xie 和 Jianfei Yang。MetaFi++：基于 WiFi 的 Transformer 人体姿态估计，用于元宇宙虚拟化身模拟。《IEEE物联网期刊》，10 14136，2023 16 14128 3 ( ): – 。[100] Yunjiao Zhou、Jia nfei Yang、Han Zou 和 Lihua Xie。TENT：将语言模型与物联网传感器连接起来，实现零样本动作识别。《IEEE移动计算汇刊》，25(6):8314–8326，2026年。3[101] Bin Zhu、Bin Lin、Munan Ning、Yang Yan、Jiaxi Cui、HongFa Wang、Yatian Pang、Wenhao Jiang、Junwu Zhang、

Zongwei Li、Wancai Zhang、Zhifeng Li、Wei Liu 和 Li Yuan。LanguageBind：通过基于语言的语义对齐，将视频-语言预训练扩展至 N模态。载于 国际学习表征会议 *(ICLR)*，2024。 3

OmniSenMo：用于泛化人体运动估计、理解与预测的统一多感官语言模型补充材料

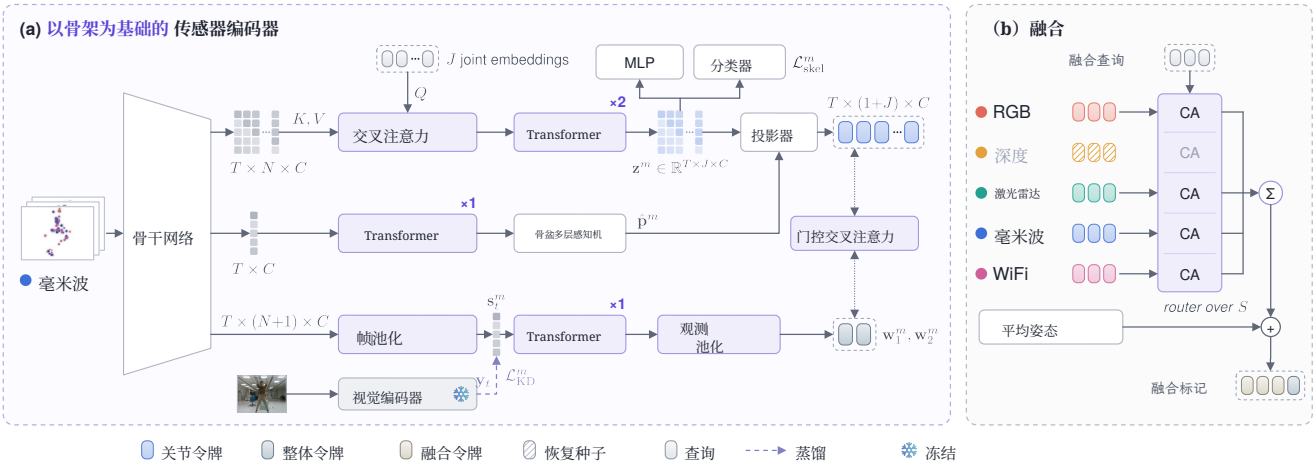


图 1. **每个传感器都被编码为相同的关节**。(a) 单个传感器的编码器 m，以毫米波为例，包含 T=5 帧、每帧 N=32个局部令牌、J=17 个关节和 C=512 个通道。多层感知机和分类器通过关节的 3D 位置和运动令牌来监督关节特征 $\mathbf{z}^m$ 。只有非 RGB 编码器具有整体路径，该路径由同步 RGB 上的冻结视觉编码器蒸馏而来。(b) 融合查询通过各自的交叉注意力（CA）关注每个可用传感器；针对可用传感器 S 的路由器将其融合，并将每个传感器姿态的平均值加到融合令牌中。

A. 方法详情

姿态表示与分词器。姿态采用 MM-FI 骨架 [20]，包含 J=17 个关节，关节顺序遵循 Human3.6M 的定义 [10],，单位为米，每位受试者使用其自身的骨骼长度，并处于保留偏航角的共享 Z 轴朝上坐标系中。一个在 AMASS 上训练的冻结逐帧卷积式 VQ 分词器会将每个关节位置量化为 K=8192 个 512 维编码，其即为运动令牌。其卷积沿单帧的 J 个关节运行，不进行任何下采样或池化，因此每帧会按关节顺序生成 J 个运动令牌，并且每个传感器的第 j 个关节令牌由关节 j的运动令牌监督。骨盆轨迹不进行令牌化：姿态估计预测其度量位置，阶段 I 预测其相对于首个观测帧的位移$\mathbf{r}_t = \mathbf{p}_t - \mathbf{p}^{t_0}$，运动预测则预测其相对于最后一个观测帧的位移。

编码器。骨干网络（用于 RGB 和深度的 HRNet-W32 [14]、用于激光雷达和毫米波的 X-Fi 点云 Transformer [6]、用于 WiFi 的 ResNet-18 [9]）为每帧生成一个全局特征和 32 个局部令牌（图1a）。关节嵌入与每帧的局部令牌进行交叉注意力，并加入关节、帧和时间嵌入；在这些 $T \times J$ 结果上应用一个两层 Transformer，得到关节特征$\mathbf{z}_{t,\,j}^{m}$；而在全局特征上应用一个单层 Transformer，以回归骨盆。投影器将关节特征映射为宽度为 512 的关节令牌，并将其均值与骨盆估计值映射为同样宽度的全局令牌。在非 RGB 编码器的整体路径中，帧特征$\mathbf{s}_t^m$和两个池化查询生成整体令牌$\mathbf{w}_1^m$，$\mathbf{w}_2^m$，并通过门控、停止梯度的交叉注意力与关节令牌路径相连。视觉令牌是冻结的视觉编码器的输出，每帧包含四个 3584 维令牌$\mathbf{v}_{t,\,k}$。在 RGB 编码器中，同样的门控交叉注意力将关节令牌与投影后的视觉令牌相连。

融合与上下文。融合查询中，90 个用于运动（每帧包含一个全局令牌和 J 关节位置），另外两个用于整体令牌；它们通过传感器专属的交叉注意力关注每个可用传感器，并由路由器进行混合，跨 S（图1b）。

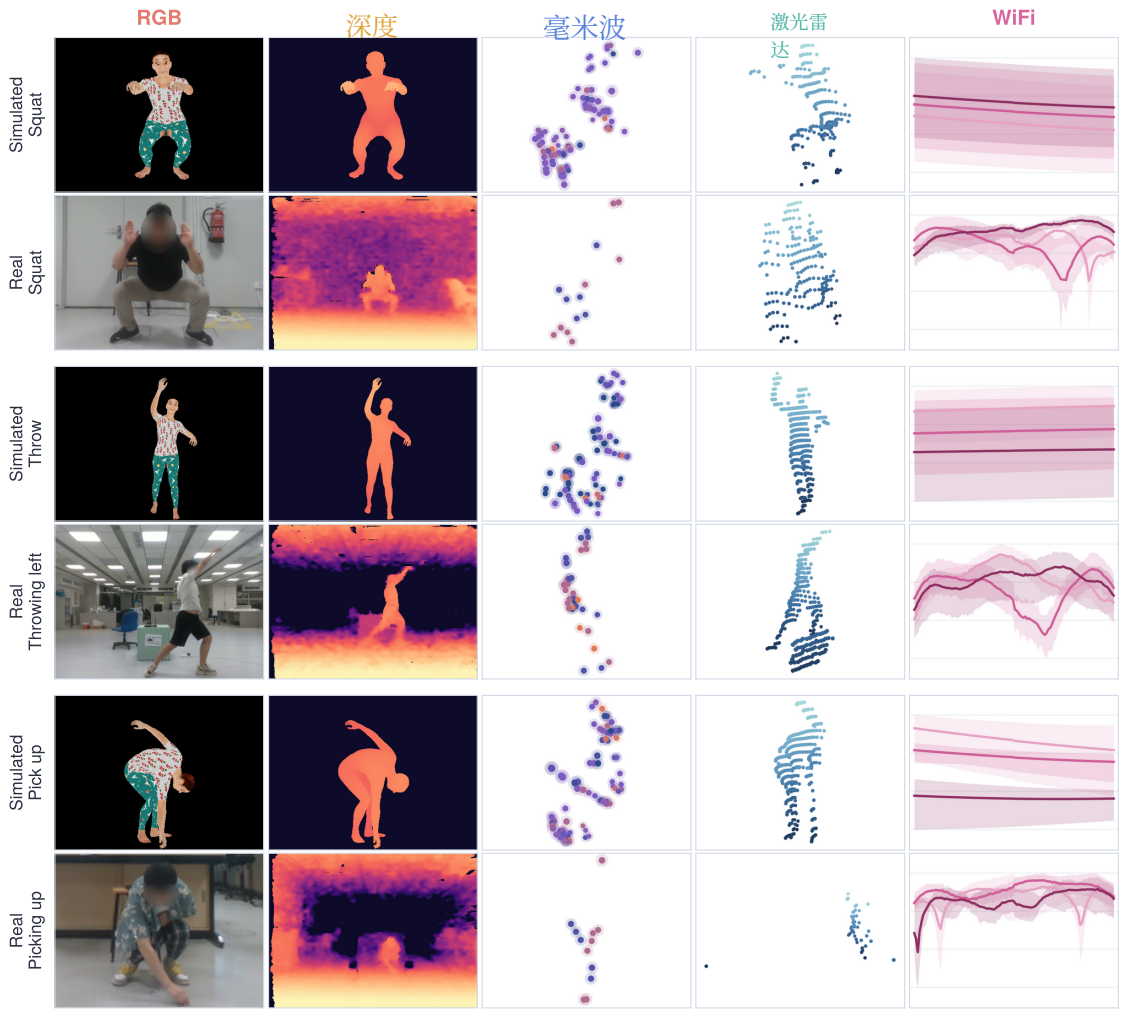


图2. **模拟和真实传感器对相似动作的观测。** 每对的上排：从同一个 AMASS动作在同一帧模拟得到的五种传感器数据。下排：真实相似动作的 MM-Fi 样本。模拟 RGB 和深度中仅包含人。

融合标记携带逐传感器姿态的加权平均值，权重按 S 归一化；局部姿态按关节和帧计算，骨盆则按帧计算。融合模块还根据逐传感器姿态令牌化后的加权投票，给出代码先验 p_{prior}。不进行运动预测时，提示后依照主论文图2的顺序排列 667 个位置：450 个传感器标记（五种传感器和五帧，每个包含一个全局令牌和 J 个关节标记）、92 个融合标记、8 个整体标记、20 个视觉标记、5 个 RGB 恢复种子，以及人类查询，其中包括 90 个运动查询和 2 个整体查询。运动预测会增加 85 个历史标记（将过去五帧的逐传感器姿态取平均，作为运动令牌）和 90 个运动查询，因此运动查询覆盖 10 个目标帧，总计 842 个位置。在注意力掩码中，提示和传感器标记以因果方式进行注意力计算；恢复种子关注它们以及同一传感器的种子；查询关注上述所有内容并彼此关注；响应则关注其前的所有内容。

所有查询都在一次前向传播中解码，图注则采用贪心方式解码。

时间嵌入。 运动查询和预测历史携带嵌入

$$\mathbf{c}_t = f_{\text{time}}(\tau_t, \phi_t, \delta) + \mathbf{e}_{\text{task}} + \mathbf{e}_{\text{obs}} + \mathbb{I}_{\text{hist}}\, f_{\text{root}}(\mathbf{r}_t), \quad (1)$$

其中，时间嵌入 f_{time} 是一个多层感知机，其输入为帧 t 的时间 τ_t（以秒为单位）、该帧在参考时间区间内的相位 ϕ_t 以及该区间的持续时间 δ。任务特定嵌入 $\mathbf{e}_{\text{task}}$ 标记任务，而$\mathbf{e}_{\text{obs}}$ 标记观测帧和目标帧。多层感知机 f_{root} 对骨盆位移 $\mathbf{r}_t$进行嵌入，指示变量 $\mathbb{I}_{\text{hist}}$ 在预测历史对应的位置为 1，在其他位置为 0。稀疏样本的参考时间区间是标注动作；其他情况下则是观测到的 0.5 秒窗口，因此运动预测不会获知动作何时结束。传感器标记、恢复种子和融合标记各自携带时间嵌入，且语言模型的旋转位置编码保持不变。

掩码传感器恢复与蒸馏。每个传感器都有 90 个恢复种子。对于被掩码的传感器，语言模型在其种子位置的输出会被解码为该传感器的关节特征以及包含骨盆信息的全局向量。这些就是主论文公式（2）中的预测结果，其目标来自一个在第二阶段单独训练的冻结编码器副本。对于掩码 RGB，另外五个种子用于预测视觉语言模型教师的特征，质量头则预测这些特征的误差，并据此调整对它们的注意力。对于蒸馏，$\mathbf{y}_t$ 是帧 t的四个视觉令牌经过 ℓ_2归一化后的均值，而 $\mathbf{A}^m$ 将 512 维映射到 3584 维。学生特征取自门控交叉注意力之前，且 RGB 编码器没有 $\mathcal{L}_{\text{KD}}$。在第二阶段，对于每个训练样本，教师都会读取同一窗口的模拟 RGB，之后再读取与之同步的真实 RGB；而学生仅在 RGB 可用时才接收 RGB 输入。

输出头。 姿态头是一个多层感知机，它将运动查询状态与逐传感器姿态的平均值结合起来；语言模型的代码 logits 则通过辅助交叉熵对代码先验进行细化。估计的骨盆位置由逐传感器骨盆位置的平均值，加上全局令牌按相同权重取平均后输入多层感知机得到的结果组成。预测头将局部姿态和骨盆位移预测为相对于最后一个观测估计值的残差。它接收经过 detached 处理的查询状态，因此运动预测仅通过目标帧运动令牌上的交叉熵来训练语言模型。

训练阶段。 阶段 *I* 将 LoRA（秩为 64）添加到 Qwen2.5-VL-7B 的语言层中，同时冻结视觉编码器和基础权重。掩码运动建模会对 AMASS 的 10 帧窗口中的未来片段、中间片段或随机片段进行掩码；文本到运动、运动到文本和运动到动作任务使用 BABEL 与 HumanML3D 区间中的稀疏样本。第二阶段 在编码器初始化后包含三个步骤：首先在模拟语料库上单独训练 RGB、深度、激光雷达和毫米波编码器两个训练轮次，并从 ImageNet 初始化 WiFi。*II-a*单独训练每个编码器，并使用 $\mathcal{L}_{\text{skel}}$；对于非 RGB 传感器，则使用 $\mathcal{L}_{\text{KD}}$。*II-b* 冻结每个编码器的联合标记路径，并使用一个小型 Transformer 替代语言模型，训练融合模块、恢复种子和查询，同时以 0.5 的概率保留每个传感器。*II-c* 通过输入和输出投影器连接阶段 I 的语言模型及其 LoRA，并在无传感器的情况下再次执行阶段 I 的任务。在 II-b 和 II-c 中，运动预测覆盖 0.5 秒；1 秒预测头仅在阶段 III 中训练。真实数据适配 在真实的密集窗口和稀疏样本上单独训练每个编码器；阶段 *III* 使用真实训练动作进行微调，每次更新执行一个任务，并为每个样本随机选择一个传感器子集，同时再次执行阶段 I 的任务。对于识别、问答和描述生成，阶段 III 将一半真实样本从左到右进行镜像，以反映 RGB、深度和激光雷达，同时舍弃毫米波和 WiFi，并交换关节的左右侧以及文本。对于 $\mathcal{L}_{\text{rec}}^{\text{VLM}}$ 的目标，它还会减去每个受试者的均值。阶段 III 在 MM-Fi 上冻结编码器；而在没有三维姿态的 XRF55 上，还会调整投影器，以及每个编码器的整体路径和门控交叉注意力。所有阶段均使用 AdamW 和大小为 32 的全局批次（更新次数和计算成本见表2）。学习率为 1.5×10^{-5}（阶段 I）、10^{-4}（II-b）和 10^{-5}（阶段 III；10^{-4} 用于骨盆分支，3×10^{-4} 用于预测头）。II-a 和 II-c 使用 10^{-4}，但 10^{-5}II-a 中的骨干网络和 II-c 中的 LoRA 除外。WiFi 编码器在整个 II-a 阶段使用 3×10^{-4}。

模拟器。 在 16,920 个模拟的 AMASS 序列中，14,438 个用于训练，1,523 个用于验证，其余序列短于五帧窗口。人体采用中性 SMPL-X 网格，并使用 AMASS 的形状系数；运动被重采样至 10 Hz。每个序列中的传感器装置仅有一种放置方式：高度为 1.15 米，朝向目标，距骨盆 2.7 至 3.4 米，这与 MM-Fi 训练集的范围一致。RGB 和深度将网格与单张纹理一起光栅化到 640×480 适配 MM-Fi 的针孔相机中，背景为空。激光雷达在 $45°$ 垂直视场内向人体和地面平面发射 32 束光。毫米波模拟器合成 60 GHz 雷达的调频连续波信号，该雷达具有三个发射天线和四个接收天线，并将每个网格三角形作为散射体，从最强的 128 个距离-多普勒单元形成点云。WiFi 在一个三天线接收器处，将直达路径与每个网格三角形散射形成的路径相加，使用 114 个子载波，频率为 5.32 GHz。模拟器不添加噪声、杂波或室内反射。只有毫米波编码器的初始化会对模拟点云进行增强，包括点丢弃、抖动和多径鬼影点，并使点数趋近于真实 MM-Fi 帧的点数。图2 比较了模拟传感器与真实传感器。

B. 协议与数据划分

在整个补充材料中，传感器缩写如下：I：RGB，D：深度，L：激光雷达，R：毫米波，W：WiFi。

观测协议。 每个计分密集窗口对应一次前向传播，因此逐帧基线方法与之逐帧匹配，而 OmniSenMo 使用 0.5 秒的上下文。估计任务对 71,280 个测试帧各评分一次，评分来源为互不重叠的密集窗口以及每个序列的最后一个窗口。理解任务则在测试序列中 3,897 个标注动作区间的稀疏样本上评分。每种方法都根据候选项的平均 token 对数似然为每个候选项评分，并减去该候选项在验证集上的平均分；计算时使用相同的检查点和传感器子集。与 HoloLLM 发布的问题一样，识别和问答的训练提示列出了全部 27 个 MM-Fi 动作短语，而训练答案中不包含任何未见过的动作。运动预测以传感器估计为起点，而非

真实姿势。

划分。 六种未见的 MM-Fi 动作 A22–A27 分别是左、右踢腿、左、右身体伸展、跳跃和鞠躬。与 mmCLIP [4] 和 mmExpert [19], 类似，它们使用根据人体运动和文本合成的传感器数据进行训练；如果某个动作没有任何真实传感器数据用于训练，我们就称其为未见动作。此划分采用官方 MM-Fi 工具箱中的示例跨动作划分，未作任何更改。在 MM-Fi 上，验证集包含 21 种训练动作中每种动作的四个完整序列（共 84 个序列），测试集则包含六种未见动作的全部 240 个序列。

XRF55 划分。 我们使用 XRF55 中 18 组受试者—场景录制数据及公开的 Kinect RGB 数据，与 HoloLLM 一致（14 名受试者和 19,800 个片段）。我们的划分将 55 种动作中的 11 种排除在训练集之外，每种动作类别占 20%，最终得到 14,256 个训练片段、1,584 个验证片段和 3,960 个测试片段。验证集由每种已见动作最后两次重复组成。发布的片段中，毫米波和 WiFi 数据与 Kinect 不同步，因此我们在 Kinect 时间加上每个场景和重复区块各自的偏移量处对其进行采样；该偏移量通过对训练片段的运动能量进行互相关估算得出。

XRF55 迁移。 我们使用相同的 AMASS动作和 XRF55 硬件参数重新仿真毫米波与 WiFi，并在该仿真数据上预训练新的毫米波和 WiFi 输入层。其他所有组件均从在 MM-Fi 上训练的模型开始。随后，使用 NLF [13] 根据 RGB 估计的伪真值三维关键点（占帧数的 68%），在真实训练片段上对编码器进行适配；第二阶段在 44 个训练动作上运行。

C. 基线方法

在 MM-Fi 上，没有任何基线方法使用测试动作的录制数据训练权重，因此我们将随机划分所发布的权重替换为使用训练动作重新训练的编码器。X-Fi [6] 和 LEO-Fuse [2] 可根据任意传感器子集逐帧估计姿态。LEO-Fuse 没有公开代码，因此我们基于冻结的 X-Fi 编码器重新实现了它，并复现了其报告的全传感器 MPJPE（见下文）。AWC [12],同样没有公开代码，我们根据其论文在它使用的四种传感器上重新实现了该方法，并将 RGB 转换为二维关键点，使用五帧时序编码器。我们在自己的划分上重新训练了 AWC（50 个训练轮次）和 DAPT [1]（使用官方代码，并以其公开的合成预训练权重初始化）。我们使用附录中的免训练融合方法，在每个划分的训练动作上重新训练 HoloLLM [23]，并采用与 OmniSenMo 相同的方式进行校准。对于 MM-Fi 上的 R+W 和 D+L+R+W，以及 XRF55 上的 D+W，其加权和融合为各传感器分配相同的权重。Qwen2.5-VL-7B [3] 接收每个样本的五帧 RGB 图像。其微调版本还会将每帧的时间、相位以及标注区间的持续时间作为文本输入，这些值与输入 OmniSenMo 时间嵌入的值相同，并且使用与第三阶段相同的 LoRA、真实样本和更新次数。我们还在 mmPred [7] 的毫米波姿态历史上运行 HumanMAC [5] 。UniPose [11]仅在姿态任务上进行比较，使用其发布的权重，因为它仅发布了推理代码，而 MM-Fi 没有可用于训练它的 SMPL 标签。我们不会重新训练 Superman [16] 或 UniMotion [17]。Superman 发布的版本没有语言模型阶段的权重，且仅涵盖 Human3.6M 上的视频到骨架估计。UniMotion 未发布经过训练的权重，而且其运动特征要求使用 MM-Fi 未提供的 22关节 SMPL 骨架。

在 XRF55 上，针对其进行评估的融合模型 [2, 6, 15]从其射频（RF）传感器中对 55 个动作组成的封闭集合进行分类，无法命名未见过的动作。因此，在多传感器模型中，我们将其与 HoloLLM 进行比较；后者没有适用于 XRF55 的毫米波编码器。

LEO-Fuse 重新实现。 LEO-Fuse [2]既未发布代码，也未发布权重。我们根据论文在冻结的 X-Fi [6] 编码器之上重新实现其姿态分支（按说明配置适配器、融合和损失函数），这些编码器在相同划分上训练，仅训练融合模型（21.6M 个参数）。在原论文的 MM-Fi 协议上（包含全部 27 个动作的 P3 协议、采用随机划分的 S1 设置、64,152 个测试帧），我们的版本在使用全部五种传感器时达到 58.04 mm 的 MPJPE 和 39.25 mm 的 PA-MPJPE，而论文报告的数值分别为 58.82 和 38.3。对 31 个传感器子集取均值后，其精度高于论文报告值（68.11 对比75.53 mm），这是因为复用的点云和 WiFi 信道状态信息 X-Fi 编码器强于原论文中的编码器。在跨动作划分上，编码器仅使用训练动作进行训练；在全部 31 个子集上，它均优于 X-Fi（均值为 101.31 对比 123.90 mm）。跨动作划分上的各子集数值见表 3。

训练动作在验证划分上的表现。表1 评估了 OmniSenMo 和重新训练的基线方法在 84 个验证序列上的表现。这些序列来自 21 个训练动作，每个动作各有 4 个；所列方法均未使用这些序列进行训练。HoloLLM 对这些样本的识别率为 90.5% 至 91.4 %，微调后的 Qwen2.5-VL 为 99.0%（随机猜测为 4.8%）。

姿态是在步长为五帧的窗口所包含的独特帧上取平均值（24,948 帧）。识别和问答在 1,242 个样本上从 21 种训练动作中进行选择，且未进行校准；HoloLLM 使用其自身的稀疏的五帧。未列出 X-Fi 和 LEO-Fuse†，因为它们的跨动作设置会使用 21 种动作的所有序列进行训练，包括这 84 个序列。零样本 Qwen2.5-VL 未经过训练，仅用于说明 21 选1有多难。

D. 实现细节

所有阶段均使用随机种子 3407，在 RTX A6000 GPU 上运行（表2）。阶段 I 采用第 12,288 次更新的结果，该结果是根据一次包含 32,768 次更新的运行在验证集上选出的。阶段 III 不选择检查点，并且

表1。 **验证集划分上的训练动作**，MM-Fi 跨动作，共 84 个验证序列。(a) 姿态误差（毫米）。(b) 21 种训练动作上的 Top-1 准确率（%，随机猜测为 4.8%），未经校准，以及 METEOR。 †我们的重新实现。 §加权和融合。 –：不可用。

(a) 三维姿态估计（毫米）

方法	输入	MPJPE↓	姿态对齐↓	平均根位置误差↓
AWC† [12]	I+L+R+W	55.2	34.8	42.0
OmniSenMo	全部五个	54.7	35.3	41.4

(b) 运动理解

方法	输入	HAR↑	问答↑	METEOR↑
HoloLLM§ [23]	I+R+W	91.4	90.9	28.1
OmniSenMo	全部五个	99.0	98.8	28.4
Qwen2.5-VL-ZS	I	17.4	17.1	–
Qwen2.5-VL-SFT	I	99.0	99.0	28.4
HoloLLM	I	90.5	91.1	28.6
OmniSenMo	I	93.3	91.9	26.3

表2. **仿真和训练成本**（使用 RTX A6000 GPU）。仿真涵盖语料库中的 2.0M 帧。

阶段	更新次数	GPU	GPU小时
I、D、L、R的模拟		8	91.9
W的模拟		3	10.6
阶段 I	12,288	4	52.4
I、D、L、R的初始化	2 个训练轮次	1	63.7
II-a, 仅编码器	49,280	1	69.4
II-b	8,192	4	63.0
II-c	16,384	4	145.4
真实数据适配	28,672	1	33.9
第三阶段	4,096	4	36.7
总计			567.0

OmniSenM 使用其最终更新。RGB 在 MM-Fi 上以每次观测 20 个池化令牌的形式输入语言模型，在 XRF55 上则以人物裁剪图像的 720 个全分辨率令牌的形式输入。OmniSenMo 有 8.83B 个参数，其中 266.9M 个参数在第三阶段训练。

E. 其他结果

E.1. 使用每个传感器子集进行估计

表 3列出了全部 31 种传感器子集的姿态误差，以及各规模子集的均值。OmniSenMo 在 31 种子集中有 30 种的 PA-MPJPE 低于 LEO-Fuse。仅使用毫米波和 WiFi 时，OmniSenMo 的 MPJPE 低于 LEO-Fuse 和 X-Fi（144.2 相比于 150.7 和 157.4 毫米），识别和问答准确率约为 HoloLLM 的两倍（27.8% 和 32.4%相比于 13.3% 和 14.6%）。使用 RGB、激光雷达、毫米波和 WiFi 这四种由 AWC 融合的传感器时，其 MPJPE 比 AWC 低 5.3 毫米（86.5 相比于 91.8 毫米）。

UniPose。 仅使用 RGB 时，OmniSenMo 的 PA-MPJPE 比姿态多模态大语言模型 UniPose [11],以零样本方式运行时低 24.0mm（59.3 *vs*。83.3 mm，在测试受试者上的配对 95% 区间为 22.0 至 25.9 mm）。UniPose 使用其发布的权重，在检测到的人物裁剪图像上逐帧运行；其 SMPL 输出不包含平移信息，因此只有其 PA-MPJPE 具有可比性。

表3。**全部 31 种传感器子集的估计**，MM-Fi 跨动作，MPJPE（毫米），每个测试帧仅计分一次。均值：对传感器数量相同的子集取均值。Ours：OmniSenMo。粗体：每行最低值。 †LEO-Fuse，我们的重新实现。

子集	X-Fi	LEO†	Ours	子集	X-Fi	LEO†	Ours
I	133.6	103.3	**90.1**	I+D+L	108.0	88.9	**83.1**
D	118.3	98.5	**90.9**	I+D+R	106.8	91.8	**85.1**
L	163.6	93.6	**88.9**	I+D+W	108.8	93.4	**86.1**
R	160.6	**153.1**	1	I+L+R	114.6	90.3	**84.8**
W	228.8	212.6	**192.1**	I+L+W	119.9	91.5	**86.2**
Mean	161.0	132.2	**123.4**	I+R+W	120.6	98.6	**91.7**
I+D	110.4	93.0	**85.3**	D+L+R	107.0	90.4	**86.4**
I+L	119.9	90.7	**84.4**	D+L+W	107.2	91.0	**87.1**
I+R	120.2	98.6	**89.6**	D+R+W	107.2	97.5	**93.0**
I+W	128.8	103.8	**92.3**	L+R+W	135.2	**94.1**	96.8
D+L	110.9	90.5	**85.7**	*Mean*	113.5	92.8	**88.0**
D+R	109.4	97.0	**91.2**	I+D+L+R	105.4	88.7	**83.2**
D+W	110.3	98.8	**92.7**	I+D+L+W	106.9	89.3	**83.9**
L+R	137.6	93.5	**93.3**	I+D+R+W	105.6	92.2	**86.1**
L+W	150.6	**94.8**	94.9	I+L+R+W	115.8	90.8	**86.5**
R+W	157.4	150.7	**144.2**	D+L+R+W	106.6	90.8	**87.7**
Mean	125.6	101.1	**95.4**	*Mean*	108.1	90.4	**85.5**
				I+D+L+R+W	105.1	89.0	**84.1**

表4. **未见过的 MM-Fiactions 的描述生成**：主要结果的 METEOR，以及使用其基线重新缩放的 BERTScore F1。粗体：每个分组中的最佳结果。 §加权和融合。

方法	输入	METEOR↑	BERTScore↑
HoloLLM§ [23]	全部五种	18.1	45.3
HoloLLM§	I+R+W	18.1	45.4
OmniSenMo	全部五种	**19.7**	**46.3**
Qwen2.5-VL-SFT	I	17.6	42.5
HoloLLM	I	18.4	45.6
OmniSenMo	I	**19.9**	**46.9**

时间平滑性。使用全部五种传感器时，OmniSenMo 的加速度误差为 56.0 毫米/帧^2, , 而 LEO-Fuse† 为 63.9, X-Fi 为 65.4；其骨骼长度随时间变化的标准差为 11.6 毫米，而另外两者分别为 12.9 和 13.5 毫米。在测试序列上计算的配对95%区间中，这四项差异均不包含零。

E.2. 使用 BERTScore 进行描述生成

表4 增加了 BERTScore [21], , 该指标通过比较每条生成的描述与其参考描述的上下文嵌入来进行评估。使用全部五种传感器时，OmniSenMo 比 HoloLLM 高 0.96 分（在测试序列上的配对95%区间为 0.28 至 1.63），比使用 I+R+W 的 HoloLLM 高 0.86 分（ 0.18 至 1.53 ）。仅使用 RGB 时，OmniSenMo 比微调后的 Qwen2.5-VL 高 4.42 分（3.67 至 5.20），比 HoloLLM 高 1.34 分（0.63 至 2.05）。

表5. **1 秒内的运动预测**在 MM-Fi 跨动作（未见过的动作）上的表现：相对于骨盆的 MPJPE（mm），在 0.1–1 秒内取平均值并报告 1 秒时的值；ADE/FDE（mm）按 mmPred 的定义计算。粗体：各区块中的最佳结果。

过去姿态	方法	Avg↓	1 s↓	平均位移误差↓	最终位移误差↓
OmniSenMo，全部五项	siMLPe [8]	115.4	143.0	626	794
OmniSenMo，全部五项	AuxFormer [18]	121.2	150.5	679	863
OmniSenMo，全部五项	HumanMAC [5]	121.6	151.4	680	872
OmniSenMo ，全部五个	OmniSenMo	**112.8**		**610**	**665**
mmPred ，毫米波	AuxFormer	132.6	**152.2**	726	845
mmPred，毫米波	mmPred [7]	**132.3**	154.1	**713**	**836**
mmPred ，毫米波	HumanMAC	136.3	159.1	746	882
OmniSenMo ，毫米波	siMLPe	130.4	152.7	707	835
OmniSenMo，毫米波	AuxFormer	132.5	152.3	730	851
OmniSenMo，毫米波	HumanMAC	136.1	158.5	749	884
OmniSenMo，毫米波				**678**	**720**

E.3. 不同过去姿态的运动预测

表5 在主要结果的运动预测部分加入了平均位移误差和最终位移误差。运动预测在每个测试序列中每隔五个窗口评估一次，条件是该窗口后接十个未来帧，并另外评估最后一个窗口（共 13,920 个窗口）。在每个区块内，所有方法都外推相同的过去姿态：OmniSenMo 使用全部五种传感器得出的估计、mmPred 的毫米波姿态历史，或 OmniSenMo 仅使用毫米波得出的估计（即主要结果中 OmniSenMo (R) 的输入）。第三个区块中的基线方法沿用第一个区块中的检查点。使用相同的毫米波输入时，OmniSenMo 的预测平均而言比所有基线方法都准确：在 1 秒内的平均误差为 125.9 而非 130.4 至 136.1 毫米；在 1 秒时为 133.8 而非 152.3 至 158.5 毫米。按测试受试者计算的配对95%区间均不包含零。平均位移误差和最终位移误差遵循 mmPred [7]：即预测的 16关节骨盆相对姿态向量与真实向量之间的 L2距离（毫米），取十个未来帧上的平均值（平均位移误差），或取 1 秒时的值（最终位移误差）。

F. 消融实验详述

F.1. 按传感器组合进行的消融实验

表6 列出了主论文消融实验表中用于计算平均值的五种传感器组合。

不使用教师的变体也会移除质量头的损失，并从 II-a 开始重新训练。不使用仿真的变体则使用 ImageNet 权重初始化编码器（RGB 和深度使用 HRNet，WiFi 使用 ResNet-18），或随机初始化（激光雷达、毫米波），并在第三阶段通过 LoRA 联合训练融合模块、恢复种子和查询。

F.2. 关节令牌与关节对应关系

表 6还按传感器组合给出了主论文中的关节令牌消融实验。 无关节对应关系 与 OmniSenMo 的区别仅在于关节对应关系：其编码 J 器保留了可学习查询，每个查询都像 OmniSenMo 中那样关注局部特征，但没有任何查询与特定关节绑定，并且在没有关节对应关系的情况下融合传感器，同时保持令牌数量不变。对五种组合取平均后，池化令牌使 PA-MPJPE 增加 3.2 mm，移除对应关系则使其增加 5.3 mm。使用全部五种传感器时，池化令牌和移除对应关系分别使手腕和脚踝的误差从 113.9 mm 增加 3.8 mm 和 8.0 mm。没有关节令牌注意力时，每种传感器组合的识别和问答准确率都会下降。仅使用毫米波和 WiFi 时，识别率和问答准确率降至 10.7% 和 14.9%，低于六个候选项的 16.7% 随机猜测水平。该变体不增加参数，只改变理解模块关注的令牌，因此它与 OmniSenMo 一样从同一预训练模型开始，且仅重新训练第三阶段。池化令牌变体从传感器编码器开始重新训练，每个编码器增加 2.38M 个参数，融合模块增加 3.2M 个参数。对单独使用的五种传感器取平均后，该变体使姿态误差增加 5.9 mm，其中仅使用激光雷达和毫米波时增幅最大，分别为 13.3 mm 和 15.8 mm。无关节对应关系的变体也从传感器编码器开始重新训练。由于其解码器要将 J 令牌映射到关节，因此每个编码器约增加 3.2M 个参数，融合模块增加 3.2M 个参数。对单独使用的五种传感器取平均后，该变体使姿态误差增加 10.2 mm，其中仅使用毫米波时增幅最大，为 19.3 mm。

F.3. 真实数据效率

在主论文的真实数据效率图中，我们使用 5% 和 20% 的真实 MM-Fi 训练数据，在真实数据适配和第三阶段重新训练 OmniSenMo，并在所有五种传感器下针对未见动作，将其与仅使用仿真数据（0%）的模型以及最终模型（100%）进行比较。这两个子集分别包含 756 个训练序列中的 38 个和 151 个。它们由完整序列构成，按动作和受试者分层抽取，且较小的子集包含在较大的子集中。两个阶段均沿用最终模型的设置和更新次数。0% 对应第二阶段结束时的模型，该模型未经真实数据适配，也未经过第三阶段。表7 列出了所有指标。

真实数据的前 20% 即可带来估计和理解方面的大部分增益。仅使用其中 5% 时，识别率就达到 50.3%（在测试序列上，95 % 区间为 46.0 至 54.7），问答（QA）达到 50.9%，高于 HoloLLM（42.8% 和 41.3%），也高于未使用仿真的 OmniSenMo（42.1% 和 43.4%）；后两者均使用全部真实数据进行训练。使用 20% 时，姿态对齐（PA）-MPJPE 达到 59.8 mm（58.2 至 61.5），与使用全部真实数据的 LEO-Fuse（58.2 mm）和未使用仿真的 OmniSenMo（58.5 mm）相当。

F.4. 在仿真语 语料

相同的模拟语料库对 OmniSenMo 的帮助远胜于对基线方法的帮助（表8）。我们在模拟语料库上预训练 LEO-Fuse 和 HoloLLM：先预训练各自的逐传感器编码器，再预训练融合头或适配器，随后按照第C节的设置，使用真实训练动作对它们进行训练。HoloLLM

表 6. **按传感器组合进行消融实验**，MM-Fi 跨动作：OmniSenMo 的数值以及各项消融实验相对于这些数值的变化；如果配对 95% 置信区间在 240 个测试序列上包含零，则以灰色标示。MPJPE 的单位为毫米，HAR 和 QA 以校准后 Top-1（%）表示。*识别和 QA 在训练时使用的密集窗口上进行评估。

变体	I+D+L+R+W			D+L+R+W			I+D+L			R+W			I		
	MPJPE↓	HAR↑	问答↑	MPJPE↓	HAR↑	问答↑	MPJPE↓	HAR↑	问答↑	MPJPE↓	HAR↑	问答↑	MPJPE↓	HAR↑	问答↑
OmniSenMo	84.1	69.2	73.4	87.7	60.5	64.6	83.1	73.1	73.9	144.2	27.8	32.4	90.1	68.2	55.6
无关节令牌注意力	↓0.1	↓20.1	↓26.7	0.0	↓18.8	↓24.2	0.0	↓25.6	↓28.7	0.0	↓17.1	↓17.5	0.0	↓23.4	↓13.2
池化令牌	↑0.7	↓5.7	↓10.4	↑3.1	↑0.2	↓3.2	↑0.7	↓7.3	↓5.2	↑13.2	↓3.0	↓11.2	↓0.8	↓9.9	↓1.0
不包含关节对应关系	↑4.4	↓14.8	↓16.4	↑8.3	↓10.9	↓16.0	↑5.1	↓16.5	↓15.7	↑16.8	↓17.0	↓8.1	↑4.3	↓21.3	↓6.0
不含恢复	↓0.2	↓0.7	↓19.5	↓0.3	↓6.2	↓16.3	↓0.1	↓4.6	↓15.4	↑0.2	↑10.6	↓4.3	↑0.2	↓5.8	↑1.3
不含仿真	↑6.4	↓27.1	↓30.0	↑9.2	↓23.1	↓26.7	↑5.2	↓35.0	↓32.5	↓4.7	↓0.8	↓4.5	↑2.4	↓27.1	↓13.6
不使用视觉语言模型教师	↓0.2	↓9.0	↓17.0	↓0.4	↑3.5	↓8.6	↓0.1	↓14.2	↓19.5	↑0.5	↑19.0	↑8.5	↑0.2	↓15.7	↓6.8
单一协议*	0.0	↓13.8	↓22.6	0.0	↓4.7	↓13.4	↓0.1	↓15.0	↓21.1	0.0	↑3.8	↓9.0	0.0	↓19.8	↓16.2



(a)



(b)

图3. **未见动作的更多定性结果**。 (a) 右侧身体伸展（估计）、右踢（理解）和鞠躬（预测），红框标出右手腕。(b) 跳跃（估计）、左侧身体伸展（理解）和右侧身体伸展（预测），红框标出左手腕。预测行从 OmniSenMo 估计的过去动作开始。圆点标示各模型的传感器，灰色骨架表示真实值。正确和错误的描述短语分别以绿色和红色显示。

通过自身的动作分类和语言建模损失，HoloLLM 从与 OmniSenMo 相同的模拟样本、动作短语和描述中学习。使用全部五种传感器时，仿真使 LEO-Fuse 的 MPJPE 降低了 0.8 毫米，而使 OmniSenMo 的 MPJPE 降低了 6.4 毫米；OmniSenMo 在不使用仿真时的 MPJPE 为 90.5 毫米（表6）。对于 HoloLLM，仿真反而降低了识别和问答准确率；而对使用全部五种传感器的 OmniSenMo，仿真则使这两项指标分别提高了 27.1 和 30.0 个百分点。仅仅获得相同的带标签模拟数据，并不足以解释 OmniSenMo 在未见动作上的性能提升。我们认为，HoloLLM 的性能下降源于传感器信号中的仿真到现实差距，其中激光雷达的差距最小。模拟 RGB 和深度数据没有背景，而模拟毫米波和 WiFi 信号则缺少真实环境中的多径反射

房间。HoloLLM 将这些信号的编码器特征输入其语言模型，因此差距会延伸到它在仿真中学习的适配器。相比之下，OmniSenMo 将每个传感器都锚定在关节令牌上，而这些令牌仅描述身体，因此适配后的编码器无论使用仿真数据还是真实数据，都能生成这些令牌；掩码传感器恢复预测的也是关节特征，而非信号块。它的视觉令牌来自语言模型的冻结视觉编码器，而仿真不会改变该编码器。因此，仿真带来的收益取决于使用仿真的模型，这与主论文消融实验中仿真和关节令牌的互补作用一致。

表7. **真实数据的前 20% 带来了大部分提升** (MM-Fi 跨动作、未见动作，除非另有说明，均使用全部五种传感器)。误差单位为毫米。HAR 和问答：校准后 Top-1（%），括号内为未校准结果。预测：对未来 1 秒内的 MPJPE 求平均值（Avg），以及 1 秒时的 MPJPE。R+W：仅使用毫米波和 WiFi。 –: 不可用。

真实数据	0%	5%	20%	100%
PA-MPJPE↓	360.9	76.8	59.8	55.0
MPJPE↓	1186.1	134.3	98.5	84.1
平均根位置误差↓	1147.5	93.2	70.9	64.6
HAR↑	38.1 (23.1)	50.3 (39.4)	58.9(57.0)	69.2 (61.7)
QA↑	27.2 (17.1)	50.9 (38.3)	52.6(50.7)	73.4 (56.9)
METEOR↑	3.4	18.4	17.5	19.7
平均预测↓	–	129.9	116.8	112.8
1秒预测↓	–	143.3	130.2	123.7
R+W MPJPE↓	465.1	189.6	161.1	144.2
R+W HAR↑	16.4（14.1）	23.4（21.6）	26.3（24.5）	27.8（29.6）
R+W 问答↑	13.2(14.6)	26.5 (19.1)	26.8(19.6)	32.4 (29.6)

表8。**在模拟语料库上预训练的基线方法**，在 MM-Fi 跨动作（未见动作）上的结果，与主结果中对应行进行对比。(a) 使用全部五种传感器以及对 31 种传感器子集取平均时的姿态误差。(b) 校准后 Top-1，括号中为未校准结果。粗体：每对中较优者。 †我们的重新实现。 §加权和融合。

(a) 三维姿态估计（mm）

方法	MPJPE↓	PA↓	平均根位置误差↓	平均$_{31}$↓
LEO-Fuse†	89.0	58.2	67.3	101.3
LEO-Fuse† +仿真	**88.2**	58.2	**64.5**	**101.0**

(b) 运动理解，I+R+W (%)

方法	HAR↑	QA↑
HoloLLM§	**42.8** (32.0)	**41.3** (32.2)
HoloLLM§ +仿真	31.3 (28.6)	32.3（28.5）

F.5. 关节标记的线性探针

主论文第4.4节中的线性探针固定模型，只改变帧，以展示哪种观测协议能让关节令牌承载姿态，哪种协议能让它们承载动作。

我们冻结 OmniSenMo，并使用 MM-Fi 的 21 种训练动作的全部五种传感器运行模型，处理由五个连续帧组成的密集窗口（与估计任务相同），或处理五帧稀疏样本（每帧取自标注动作区间的五分之一处，与理解任务相同）。对于每个窗口，我们从每个传感器编码器中提取每帧的 J 关节令牌，在五种传感器间取平均值，再将 85 个令牌拼接起来。特征经标准化后，通过主成分分析降至 256 维。使用岭回归解码五帧中相对于骨盆的 J 关节三维位置，并使用逻辑回归在 21 种训练动作中解码动作类别。两个探针、标准化过程和主成分分析均在训练集上拟合（4,431 个密集窗口和 707 个稀疏样本），正则化强度则通过按序列进行的五折交叉验证选定。模型在验证集上评估，验证集包含每种训练动作的四个完整序列（411 个密集窗口和414个稀疏采样）。

表9. **冻结的 OmniSenMo 关节令牌的线性探针**，在 MM-Fi 的21种训练动作的验证窗口上使用全部五种传感器。密集：五个连续帧。稀疏：五个帧，每个取自动作区间的五分之一部分。粗体：各探针表现更好的方案。随机准确率为4.8%。

关节标记	姿态平均关节位置误差（毫米）↓		动作Top-1准确率（%）↑	
	密集	稀疏	密集	稀疏
RGB	**51.0**	54.8	75.7	**98.1**
深度	**49.5**	52.8	81.0	**96.4**
激光雷达	**47.8**	53.2	91.7	**98.8**
毫米波	**76.7**	77.3	84.9	**94.7**
WiFi	**116.2**	124.8	9.7	**14.0**
五者的平均值	**47.5**	52.3	89.5	**100.0**

表10。 **XRF55 在全部 15 个传感器子集上的未见动作**，同一 OmniSenMo 模型与 HoloLLM 的对比 [23]。 $|S|$: 传感器数量。在 11 个未见动作上的 Top-1（%），校准方式同第 B 节，括号内为未校准结果。随机猜测为 9.1%。粗体：准确率较高者。 §加权和融合。–: 无数据。

子集	$\|S\|$	HAR↑		QA↑	
		HoloLLM	OmniSenMo	HoloLLM	OmniSenMo
I	1	24.6 (17.5)	**65.8** (59.8)	25.2 (14.2)	**64.2** (59.3)
D	1	**26.7** (18.0)	26.4 (22.1)	25.2 (16.5)	**28.4** (26.3)
R	1	–	26.3 (18.2)	–	26.7 (19.4)
W	1	13.1 (16.4)	**21.6** (15.6)	12.9 (14.8)	**19.9** (17.3)
I+D	2	20.8§ (18.5)	**63.5** (56.4)	22.4§ (15.8)	**63.0** (55.3)
I+R	2	–	65.9 (59.9)	–	64.8 (59.7)
I+W	2	19.8§ (18.2)	**66.0** (60.9)	22.2§ (15.1)	**65.0** (59.7)
D+R	2	–	27.9 (23.2)	–	29.7 (27.9)
D+W	2	14.6§ (10.7)	**27.1** (22.7)	14.6§ (10.4)	**28.8** (27.1)
R+W	2	–	23.7 (19.6)	–	23.7 (20.7)
I+D+R	3	–	64.5 (56.6)	–	64.1 (56.8)
I+D+W	3	20.3§ (19.7)	**63.4** (56.5)	22.5§ (16.0)	**63.2** (56.1)
I+R+W	3	–	66.3 (59.6)	–	65.0 (59.7)
D+R+W	3	–	28.4 (23.4)	–	30.2 (28.5)
I+D+R+W	4	–	64.3 (56.6)	–	63.9（57.0）

表9 给出了主论文报告的每个传感器关节标记及其平均值上的探测结果。对于每个传感器，密集窗口能够更准确地解码姿态，而稀疏采样能够更准确地解码动作。主论文中的单一协议为每项任务提供连续帧，因为稀疏采样覆盖整个标注区间；对于运动预测，还会包括预测开始之后的帧。

G. XRF55上的跨数据集验证

表 10 列出了使用与主论文相同的OmniSenMo模型，在全部15个传感器子集上针对11个未见XRF55动作进行的未见动作识别和问答结果。

全部四种传感器与仅 RGB。使用全部四种传感器时，OmniSenMo 的识别准确率为 64.3%，回答准确率为 63.9%；仅使用 RGB 时则分别为 65.8% 和 64.2%。在 14 名测试受试者中，这两项差异均不显著（95% 置信区间

$[-3.9, +0.9]$ 和 $[-2.1, +1.7]$ 个百分点）。差距源于深度数据。仅使用深度数据时，模型的回答集中在少数几个候选类别上（在 3,960 个片段中，跑步、开合跳和向左滑动分别出现 1,740、688 和 581 次），而且答错时的置信度与答对时一样高，因此这些回答会削弱 RGB 提供的证据。XRF55 在宽视野下记录深度数据，此时人体仅占编码器输入的 1.4%（在 MM-Fi 上为 5.5%）。

H. 更多定性结果

图 3展示了六段未见动作的额外视频片段，排版与主论文图 5 相同，均由使用全部五种传感器的同一个 OmniSenMo 模型生成。在估计方面，OmniSenMo 能跟踪右手腕完成右侧身体伸展，而 LEO-Fuse 的偏差为 145 至 233 毫米（OmniSenMo 为 60 至 143 毫米）；在所示跳跃的每一帧中，OmniSenMo 对左手腕的跟踪也比 LEO-Fuse 更准确。在理解方面，对于右踢，OmniSenMo 能正确命名，而 HoloLLM 将其回答为跳起和鞠躬；OmniSenMo 的图注将动作描述为朝右前方移动，而 HoloLLM 则描述为向左投掷某物。对于左侧身体伸展，两种模型都能说出动作名称；OmniSenMo 将其描述为左臂向侧面伸展，而 HoloLLM 则描述为向左弓步。在运动预测方面，对于右侧身体伸展，OmniSenMo 在所有展示的预测时域上都比 siMLPe 更接近真实值（1 秒时为 73 ，相比 89 毫米）。对于鞠躬，OmniSenMo 在 1 秒时更接近真实值（164 ，相比 261 毫米）。

I. 局限性

我们的结果存在三项局限性。首先，OmniSenMo 仍需要真实传感器数据来适配其编码器，因为模拟器未包含噪声、杂波和室内反射（第A节）。在使用任何真实数据之前，第二阶段结束时的模型在未见动作上的 PA-MPJPE 达到 360.9 mm（第 F.3节）。其次，仅使用毫米波和 WiFi 的准确率仍远低于仅使用 RGB（MPJPE 为 144.2 对比 90.1 mm，识别率为 27.8% 对比 68.2 %），部分原因在于视觉语言模型教师蒸馏了 RGB 外观信息，而毫米波和 WiFi 无法复现这些信息。第三，理解任务使用了 HoloLLM 发布的问题和描述文本 [23]。其问答会从列表中选择一个动作，而 LLaVA-Video [22] 根据人工编写的种子描述撰写了大多数参考描述，因此，这些任务是在针对部分由模型编写的文本测试对动作层面的理解。

参考文献

[1] Xiaoqi An、Lin Zhao、Chen Gong、Jun Li 和 Jian Yang。为稳健的基于激光雷达的三维人体姿态估计预训练密度感知姿态变换器。载于 *AAAI*人工智能会议论文集（*AAAI*），第 1755–1763 页，2025 年。 18[2] Emrecan Aslan。LEO-Fuse：一种任务通用、模态无关的通用多模态人体感知框架。载于 欧洲计算机视觉会议论文集（*ECCV*），第 433–451 页，2026年。 18

[3] Shuai Bai、Keqin Chen、Xuejing Liu、Jialin Wang、Wenbin Ge、Sibo Song、Kai Dang、Peng Wang、Shijie Wang、Jun Tang、Humen Zhong、Yuanzhi Zhu、Mingkun Yang、Zhaohai Li、Jianqiang Wan、Pengfei Wang、Wei Ding、Zheren Fu、Yiheng Xu、Jiabo Ye、Xi Zhang、Tianbao Xie、Zesen Cheng、Hang Zhang、Zhibo Yang、Haiyang Xu 和 Junyang Lin。Qwen2.5-VL 技术报告。arXiv 预印本 arXiv:2502.13923，2025。18[4] Qiming Cao、Hongfei Xue、Tianci Liu、Xingchen Wang、Haoyu Wang、Xincheng Zhang 和 Lu Su。mmCLIP：通过信号-文本对齐提升基于毫米波的零样本人体活动识别。在 ACM 嵌入式网络传感器系统会议（SenSys）论文集，第184–197页，2024。18[5] Ling-Hao Chen、Jiawei Zhang、Yewen Li、Yiren Pang、Xi-aobo Xia 和 Tongliang Liu。HumanMAC：用于人体运动预测的掩码动作补全。在 IEEE/CVF国际计算机视觉会议（ICCV）论文集，第 9544 9555, 2023 18, 20 –

[6]

页。Xinyan Chen 和 Jianfei Yang。X-Fi：用于多模态人体感知的模态不变基础模型。在国际学习表征会议（ICLR），2025。15、18[7] Junqiao Fan、Haocong Rao、Jiarui Zhang、Jianfei Yang 和 Lihua Xie。mmPred：黑暗环境中的基于雷达的人体运动预测。在 AAAI人工智能会议（AAAI）论文集，第3777–3785页，2026年。18、20[8] Wen Guo、Yuming Du、Xi Shen、Vincent Lepetit、Xavier Alameda-Pineda 和 Francesc Moreno-Noguer。回归多层感知机：一种简单的人体运动预测基线方法。在 IEEE/CVF冬季计算机视觉应用会议（WACV）论文集，第4809–4819页，2023年。20[9] Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun。用于图像识别的深度残差学习。在 IEEE计算机视觉与模式识别会议（CVPR）论文集，第770–778页，2016年。15[10] Catalin Ionescu、Dragos Papava、Vlad Olaru 和 Cristian Sminchisescu。Human3.6M：自然环境中的大规模数据集与三维人体感知预测方法。IEEE模式分析与机器智能汇刊，36(7):1325–1339，2014年。15[11] Yiheng Li、Ruibing Hou、Hong Chang、Shiguang Shan 和 Xilin Chen。UniPose：用于人体姿态理解、生成与编辑的统一多模态框架。在 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第27805–27815页，2025年。18、19[12] Mengshi Qi、Jiaxuan Peng、Xianlin Zhang 和 Huadong Ma。迈向三维人体姿态估计中的平衡多模态学习。在 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第21231–21241页，2026年。18、19[13] Istv´an S´ar´andi 和 Gerard Pons-Moll。用于连续三维人体姿态与形状估计的神经定位场。在神经信息处理系统进展（NeurIPS），第140032–140065页，2024年。18[14] Ke Sun、Bin Xiao、Dong Liu 和 Jingdong Wang。用于人体姿态估计的深度高分辨率表示学习。

在 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第5693 5703页，2019年15 –。[15] Hao Wang、Yanyu Qian、Pengcheng Weng、Zixuan Xia、William Dan、Yangxin Xu 和 Fei Wang。COMPASS：用于缺失模态人体感知的融合匹配监督。arXiv 预印本 arXiv:2604.02056，2026年。18[16] Xinshun Wang、Peiming Li、Ziyi Wang、Zhongbin Fang、Zhichao Deng、Songtao Wu、Jason Li 和 Mengyuan Liu。Superman：统一骨架与视觉的人体运动感知和生成。在 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第38335–38344页，2026年。18[17] Ziyi Wang、Xinshun Wang、Shuang Chen、Yang Cong 和 Mengyuan Liu。UniMotion：用于运动-文本-视觉理解与生成的统一框架。在欧洲计算机视觉会议（ECCV）论文集，第664–682页，2026年。18[18] Chenxin Xu、Robby T. Tan、Yuhong Tan、Siheng Chen、Xinchao Wang 和 Yanfeng Wang。辅助任务有助于基于三维骨架的人体运动预测。在 IEEE/CVF国际计算机视觉会议（ICCV）论文集，第9509–9520页，2023年。20[19] Yifan Yan、Shuai Yang、Xiuzhen Guo、Xiangguang Wang、Wei Chow、Yuanchao Shu 和 Shibo He。mmExpert：整合大语言模型，实现毫米波数据的综合合成与理解。在 ACM移动网络与移动计算国际研讨会（MobiHoc）论文集，第1–10页，2025年。18[20] Jianfei Yang、He Huang、Yunjiao Zhou、Xinyan Chen、Yue-cong Xu、Shenghai Yuan、Han Zou、Chris Xiaoxuan Lu 和 Lihua Xie。MM-Fi：用于多样化无线感知的多模态非侵入式四维人体数据集。在神经信息处理系统进展（NeurIPS）数据集与基准赛道，第18756–18768页，2023年。15[21] Tianyi Zhang、Varsha Kishore、Felix Wu、Kilian Q. Wein-berger 和 Yoav Artzi。BERTScore：使用 BERT 评估文本生成。在国际学习表征会议（ICLR），2020年。19[22] Yuanhan Zhang、Jinming Wu、Wei Li、Bo Li、Zejun Ma、Ziwei Liu 和 Chunyuan Li。LLaVA-Video：使用合成数据进行视频指令微调。机器学习研究汇刊（TMLR），2025年。23[23] Chuhao Zhou 和 Jianfei Yang。HoloLLM：用于语言基础的人体感知与推理的多传感器基础模型。在神经信息处理系统进展（NeurIPS），第73543–73568页，2025年。18、19、22、23