XR 动捕革新

——从稀疏观测到环境感知和多模态的人体运动捕捉

苏卓

ByteDance 字节跳动

1. 研究背景

- XR游戏应用

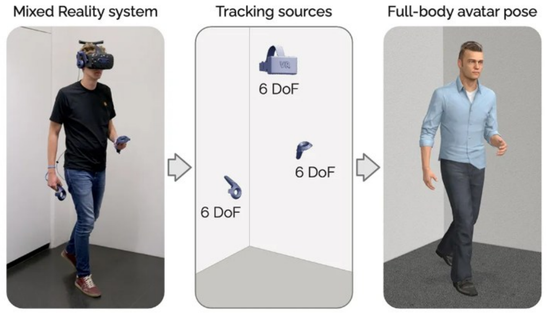
PICO Motion Tracker应用

- 稀疏观测

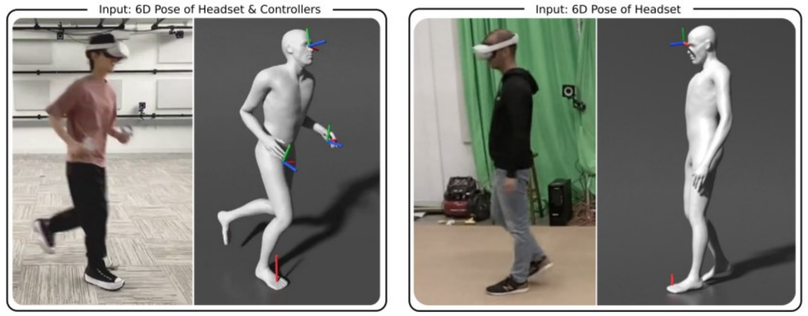
仅用头戴设备和双手柄等稀疏追踪信号重建全身动作

典型XR输入设备

- 现有技术&挑战



AvatarPoser



QuestSim

- **观测严重不足**：大多数部位无直接观测（如躯干、下肢）
- **动作协调性强**：仅局部信息难以准确推理整体姿态
- **易出现伪影**：如漂浮、打滑、穿透地面等物理不合理现象
- **实时性要求高**：往往需后处理/物理仿真，不适用于实时VR场景

[1] AvatarPoser: Articulated Full-Body Pose Tracking from Sparse Motion Sensing, ECCV 2022.
[2] QuestSim: Human Motion Tracking from Sparse Sensors with Simulated Avatars, SIGGRAPH Asia 2022.

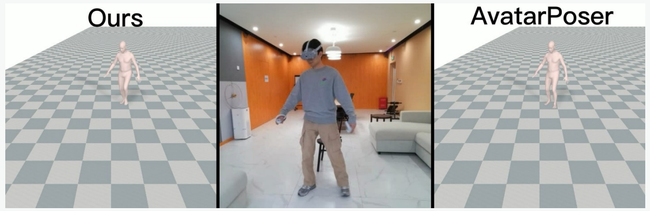
2. 稀疏观测动捕总览

AvatarJLM

ICCV'23

6DoF Head and Hands Input

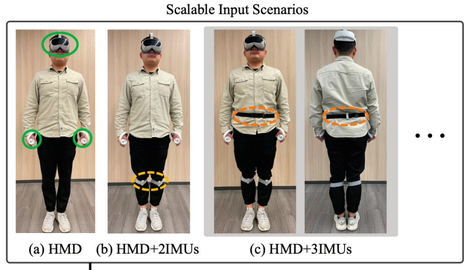
Full-Body Motion Output



6DoF of HMD and hand controllers

HMD-Poser

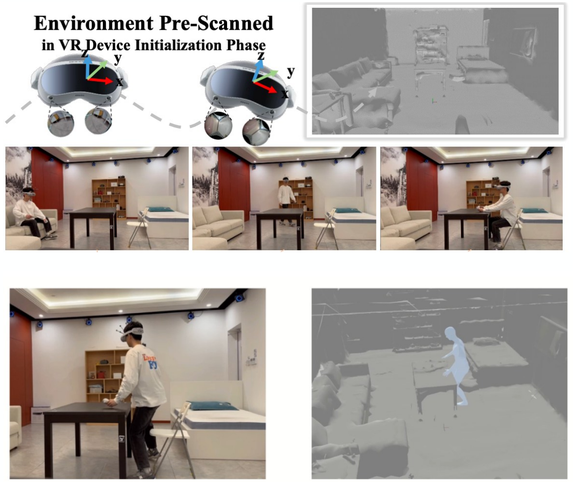
CVPR'24



+ scalable IMUs on legs or pelvis

EnvPoser

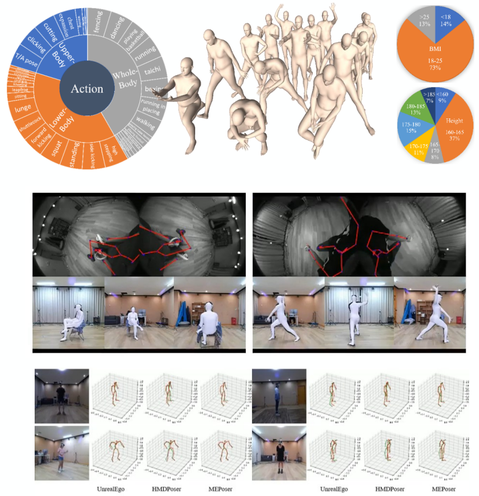
CVPR'25



+ 3D environment from VR device

EMHI (MEPoser)

AAAI'25

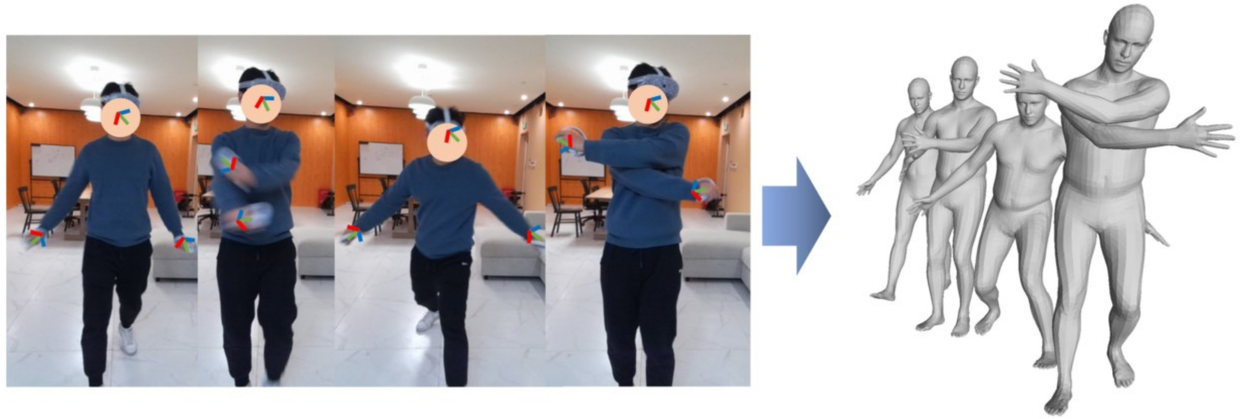


+ egocentric cameras of HMD

[1] Realistic Full-Body Tracking from Sparse Observations via Joint-Level Modeling, *ICCV* 2023.
[2] HMD-Poser: On-Device Real-time Human Motion Tracking from Scalable Sparse Observations, CVPR 2024.
[3] EnvPoser: Environment-aware Realistic Human Motion Estimation from Sparse Observations with Uncertainty Modeling, CVPR 2025.
[4] EMHI: A Multimodal Egocentric Human Motion Dataset with HMD and Body-Worn IMUs, AAAI 2025.

3. AvatarJLM

3.1 Key-insight



输入为头戴和手柄的位姿 | 输出为人体22个关节的姿态

👁 人体动作高度结构化且协调，因此在稀疏观测条件下，仅靠头和手的信号恢复全身动作，显式**建模各关节间的相互关联**，可以有效提升动作预测的合理性与连贯性。

💡 为此，我们提出一个**两阶段关节级建模框架**，先提取关节特征，再通过Transformer建模空间与时间上的依赖关系，从而重建更准确、平滑的全身动作。

[1] Realistic Full-Body Tracking from Sparse Observations via Joint-Level Modeling, *ICCV* 2023.

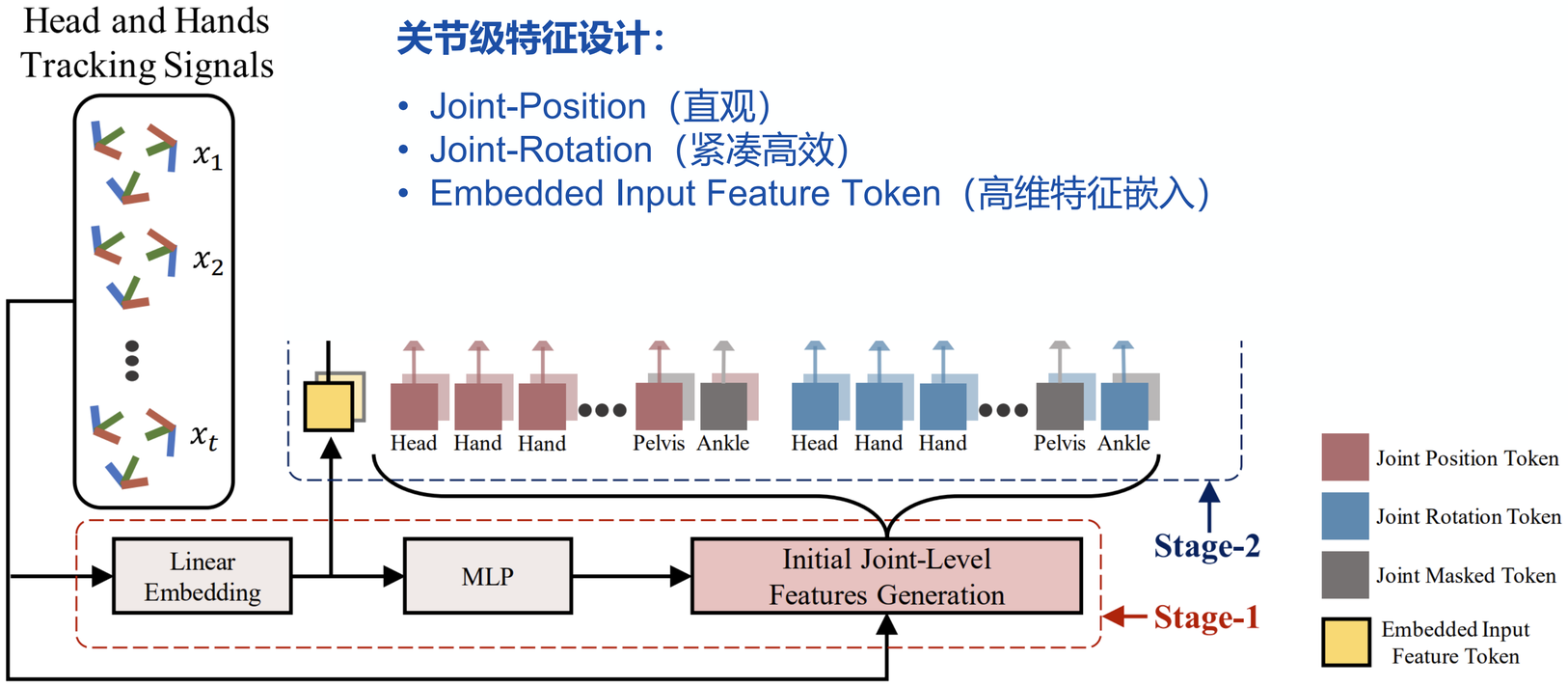
3. AvatarJLM

3.2 两阶段框架

贡献：

✅ 提出两阶段端到端神经网络框架，可仅依赖头和双手的稀疏追踪信号，实现高精度、强时序一致性的全身动作估计，且无需后处理，效果显著优于现有方法。

✅ 精心设计关节级特征提取器，结合旋转、位置与嵌入特征，



关节级特征设计：

- Joint-Position（直观）
- Joint-Rotation（紧凑高效）
- Embedded Input Feature Token（高维特征嵌入）

1. 阶段一：初步生成关节特征（Joint Position + Rotation + EIF）

[1] Realistic Full-Body Tracking from Sparse Observations via Joint-Level Modeling, *ICCV* 2023.

3. AvatarJLM

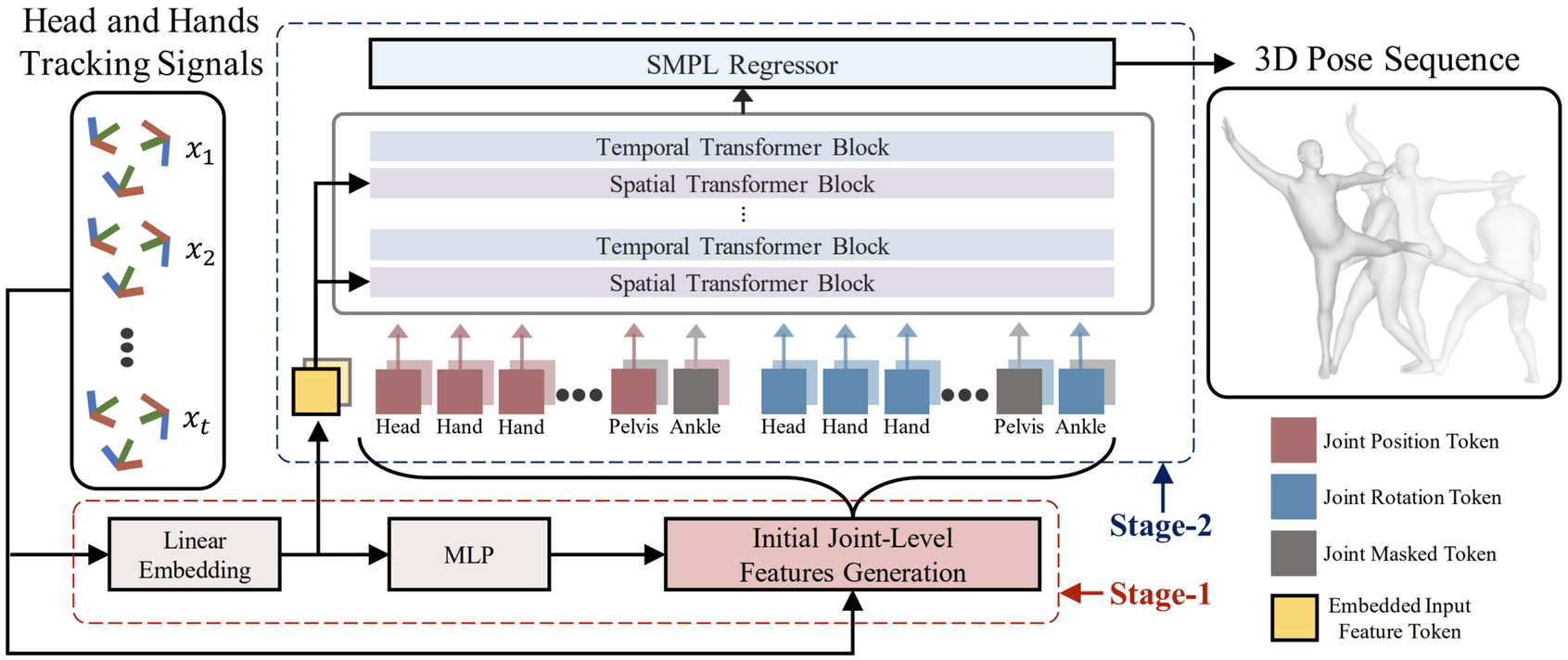
3.2 两阶段框架

贡献：

✅ 提出两阶段端到端神经网络框架，可仅依赖头和双手的稀疏追踪信号，实现高精度、强时序一致性的全身动作估计，且无需后处理，效果显著优于现有方法。

✅ 精心设计关节级特征提取器，结合旋转、位置与嵌入特征，并将其作为时空token输入Transformer，有效建模关节间依赖关系，提升动作重建的连贯性与合理性。

2. 阶段二： 空间Transformer + 时间Transformer交替，建模关节相关性



[1] Realistic Full-Body Tracking from Sparse Observations via Joint-Level Modeling, *ICCV* 2023.

3. AvatarJLM

3.3 损失函数设计

贡献：

✅ 提出两阶段端到端神经网络框架，可仅依赖头和双手的稀疏追踪信号，实现高精度、强时序一致性的全身动作估计，且无需后处理，效果显著优于现有方法。

✅ 精心设计关节级特征提取器，结合旋转、位置与嵌入特征，并将其作为时空token输入Transformer，有效建模关节间依赖关系，提升动作重建的连贯性与合理性。

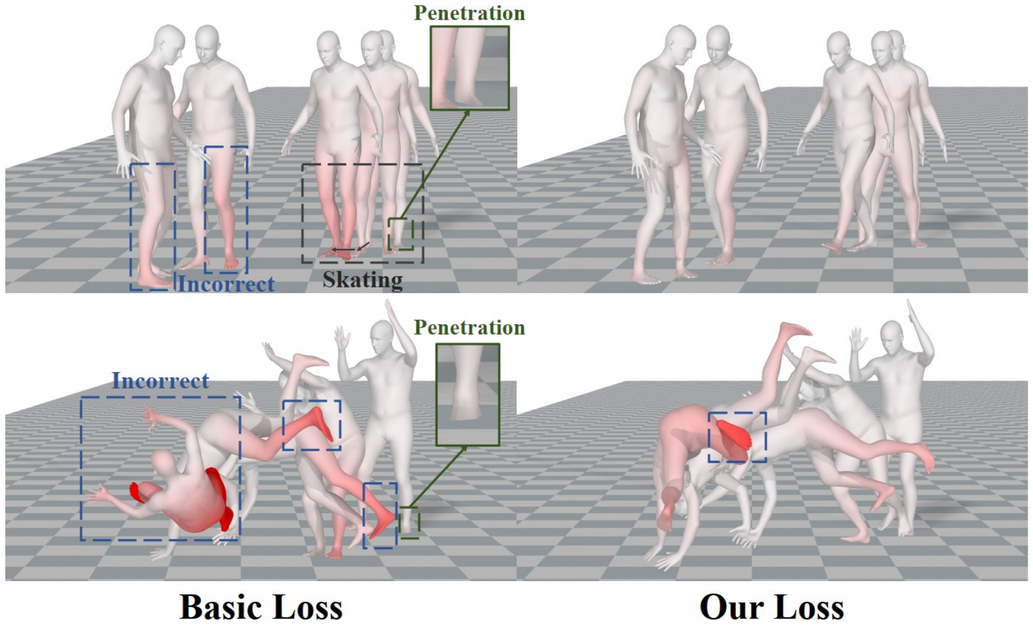
✅引入针对性损失函数组合，包括手部对齐、动作平滑与物理一致性等约束，显著减小漂浮、穿透、打滑等伪影，提高生成动作的物理真实性与质量。

数据集：Amass

损失函数：

$$L = L_{first} + \beta L_{ori} + \gamma L_{rot} + \delta L_{pos} + \epsilon L_h + \zeta L_{mot} + L_{phy},$$

- L1基础损失： 对旋转、位置、姿态参数的监督
- 手对齐损失（Hand Alignment）： 避免使用慢速IK模块
- 运动平滑损失（Motion Smoothness）： 速度/接触约束
- 物理一致性损失（Physical Loss）： 防止地面穿透与漂浮



[1] Realistic Full-Body Tracking from Sparse Observations via Joint-Level Modeling, *ICCV* 2023.

3. AvatarJLM

3.4 消融实验

关节特征消融：

- 使用其他Learnable Token不如显示建模关节特征
- 单独使用Position或Rotation都不如二者结合
- 加入EIF后效果最佳

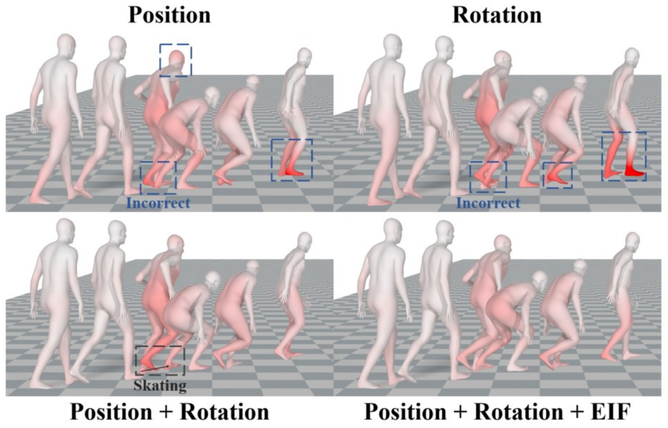


Figure 5. Ablation study for our method with four different generated features for the second stage, in which the errors are color-coded in red.

Stage	Method	MPJRE	MPJPE	MPJVE	Jitter	Ground	Skate	H-PE	U-PE	L-PE
1	Constant Token	6.97	9.00	31.21	3.80	12.02	0.36	1.55	3.71	18.26
	Learnable Token	6.91	8.91	30.82	3.48	11.52	0.36	1.57	3.71	18.01
2	Position	6.15	6.77	25.53	5.76	2.63	0.23	**1.62**	**3.48**	12.51
	Rotation	6.00	7.48	26.41	**3.77**	2.48	0.26	1.92	3.62	14.21
	Rotation + Position	5.90	6.71	23.97	4.42	2.60	0.22	1.71	3.51	12.30
	Rotation + Position + EIF	**5.86**	**6.60**	**23.57**	4.10	**2.46**	**0.21**	1.69	3.52	**12.12**

Table 6. Performance comparisons between our proposed method with different initial joint-level features. The best results are in **bold**, and the second-best results are underlined.

Method	MPJRE	MPJPE	MPJVE	Jitter	Ground	Skate	H-PE	U-PE	L-PE
Ours + Basic Loss	6.09	7.50	32.53	8.98	3.66	0.35	3.11	3.93	13.76
+ Hand	5.87	6.90	29.07	6.88	3.73	0.33	**1.58**	**3.51**	12.85
+ Hand + Motion	**5.81**	6.74	24.25	4.22	3.22	0.22	1.60	3.56	12.32
+ Hand + Motion + Physical	5.86	**6.60**	**23.57**	**4.10**	**2.46**	**0.21**	1.69	3.52	**12.12**

Table 7. Performance comparisons between our proposed method with different loss functions. * denotes our retrained AvatarPoser using their public source code. The best results are in **bold**, and the second-best results are underlined.

损失函数消融：

- 加入Motion/Physical Loss后提升显著
- 其他方法加入我们的Loss也能提升效果

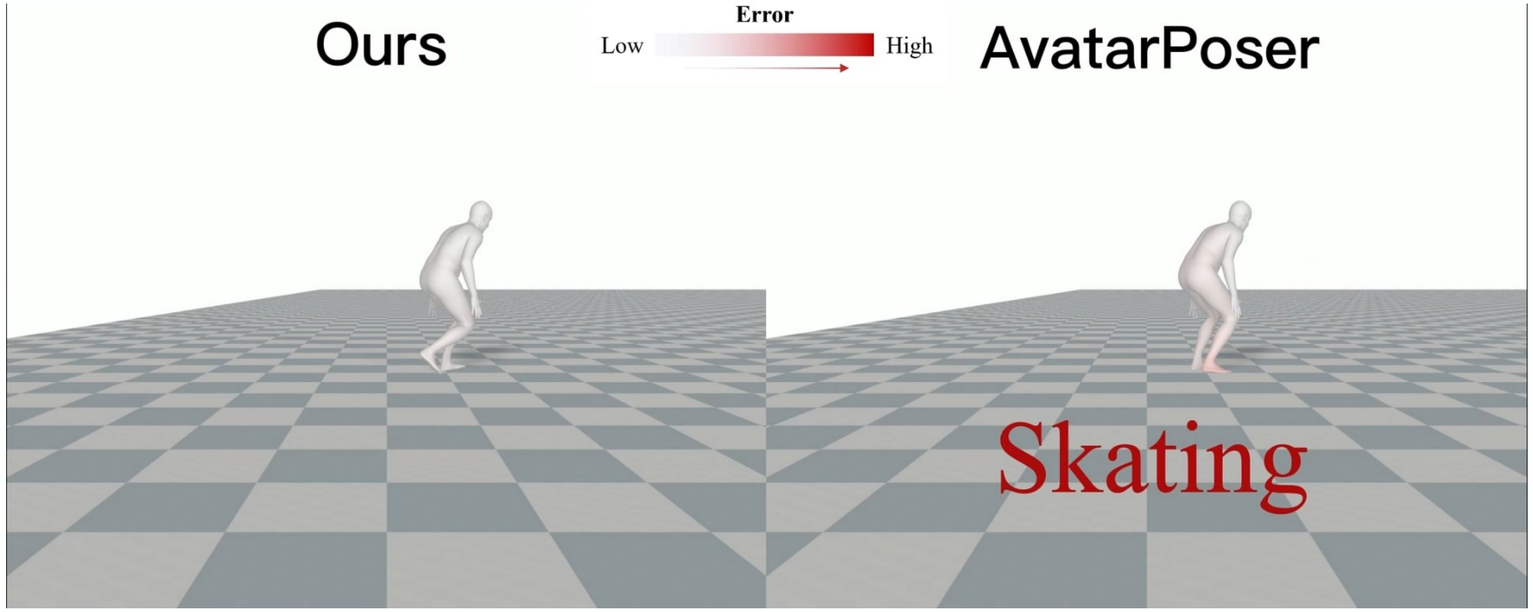
Method	MPJRE	MPJPE	MPJVE	Jitter	Ground	Skate	H-PE	U-PE	L-PE
AvatarPoser	6.39	8.05	30.85	-	-	-	-	-	-
AvatarPoser-L*	5.95	7.80	30.82	6.89	3.83	0.32	4.29	4.19	14.12
AvatarPoser-L* + Our Loss	6.02	7.14	23.92	**3.41**	2.51	**0.21**	1.82	3.54	13.43
Ours + Our Loss	**5.86**	**6.60**	**23.57**	4.10	**2.46**	**0.21**	**1.69**	**3.52**	**12.12**

Table 8. Architecture comparisons with AvatarPoser [16]. * denotes our retrained AvatarPoser using their public source code. AvatarPoser-L denotes a larger version of AvatarPoser.

[1] Realistic Full-Body Tracking from Sparse Observations via Joint-Level Modeling, *ICCV* 2023.

3. AvatarJLM

3.5 对比实验



Dataset	Method	MPJRE	MPJPE	MPJVE
CMU	Final IK	17.80	18.82	56.83
	CoolMoves	9.20	18.77	139.17
	LoBSTr	12.51	12.96	49.94
	VAE-HMD	6.53	13.04	51.69
	AvatarPoser	5.93	8.37	35.76
	Ours	**5.34**	**7.75**	**26.54**
BMLrub	Final IK	15.93	17.58	60.64
	CoolMoves	7.93	13.30	134.77
	LoBSTr	10.79	11.00	60.74
	VAE-HMD	5.34	9.69	51.80
	AvatarPoser	4.92	7.04	43.70
	Ours	**4.71**	**6.49**	**36.96**
HDM05	Final IK	18.64	18.43	62.39
	CoolMoves	9.47	17.90	140.61
	LoBSTr	13.17	11.94	48.26
	VAE-HMD	6.45	10.21	40.07
	AvatarPoser	6.39	8.05	30.85
	Ours	**5.86**	**6.60**	**23.57**

Table 2. Evaluation results under Protocol 2.

相比SOTA方法：
- MPJPE 降低 20~30%
- MPJVE、Jitter 显著下降
- Ground Penetration/Skating等 明显减少

Protocol	Method	MPJRE	MPJPE	MPJVE	Jitter	Ground	Skate	H-PE	U-PE	L-PE
1	AvatarPoser*	3.07	4.15	28.39	16.15	3.80	0.23	2.45	2.00	7.91
	Ours	**2.90**	**3.35**	**20.79**	**8.39**	**3.30**	**0.13**	**1.24**	**1.72**	**6.20**
3	AvatarPoser*	4.56	6.44	34.45	11.15	2.95	0.32	3.70	2.93	12.59
	Ours	**4.30**	**4.93**	**26.17**	**7.19**	**2.17**	**0.21**	**1.45**	**2.27**	**9.59**

Table 4. More metrics comparisons with AvatarPoser [16] under Protocol 1 and Protocol 3. * denotes our retrained AvatarPoser using their public source code.

[1] Realistic Full-Body Tracking from Sparse Observations via Joint-Level Modeling, *ICCV* 2023.

3. AvatarJLM

3.6 实测效果

实测动捕效果可视化：

- 其他方法存在滑动、穿透等问题
- 我们的方法动作更自然、无伪影、接地稳定

Method	MPJRE	MPJPE	MPJVE	Jitter	Ground	Skate	H-PE	U-PE	L-PE
AvatarPoser*	7.28	11.22	31.67	12.87	2.20	0.30	6.60	5.79	20.73
Ours	**6.98**	**9.52**	**25.78**	**10.04**	**0.20**	**0.21**	**5.31**	**5.16**	**17.15**

Table 5. Evaluation results on the real-captured data. * denotes our retrained AvatarPoser using their public source code.

[1] Realistic Full-Body Tracking from Sparse Observations via Joint-Level Modeling, *ICCV* 2023.

3. AvatarJLM

3.7 总结和局限

总结：

✅ 提出基于关节级建模的两阶段网络
✅ 仅靠头+手信号即可恢复真实感全身动作
✅ 性能全面优于三节点SOTA方法，无需后处理即可用于实时VR应用

局限：

❌ 无法还原极端复杂的下半身动作动作（如舞蹈/跌倒重建失败）
❌ 难以仅靠头+手信号即可恢复上半身不动，下半身运动的情形

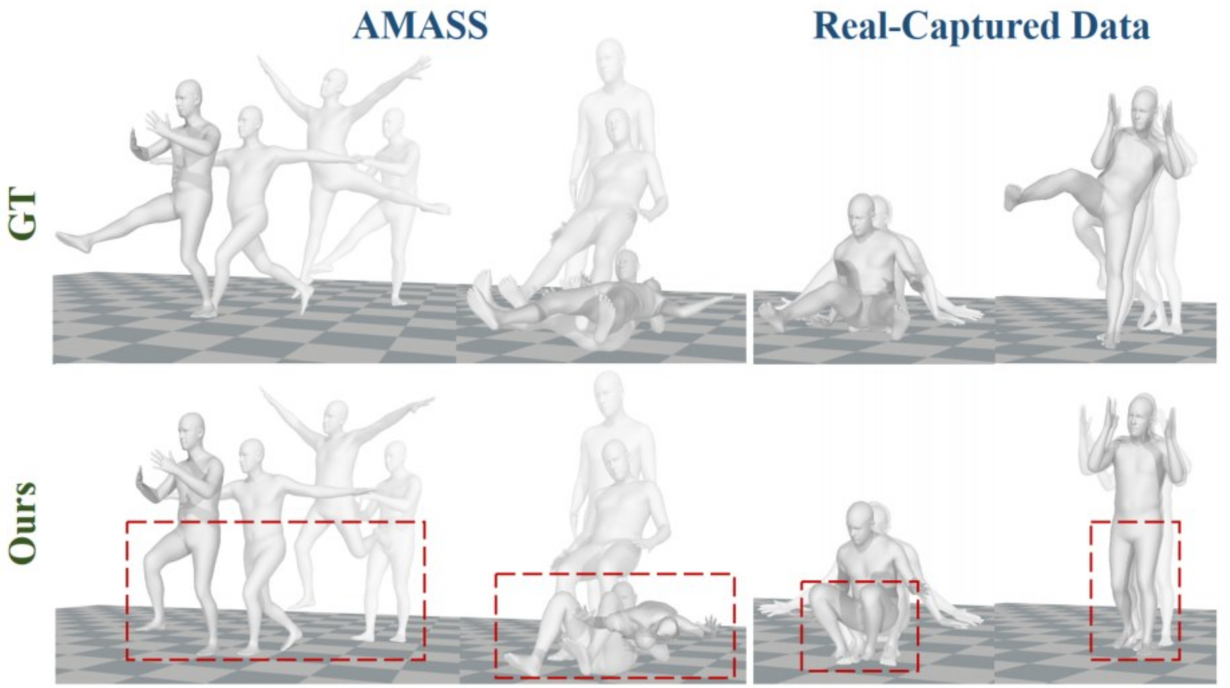


Figure 7. Failure cases on AMASS and real-captured data.

[1] Realistic Full-Body Tracking from Sparse Observations via Joint-Level Modeling, *ICCV* 2023.

3. AvatarJLM

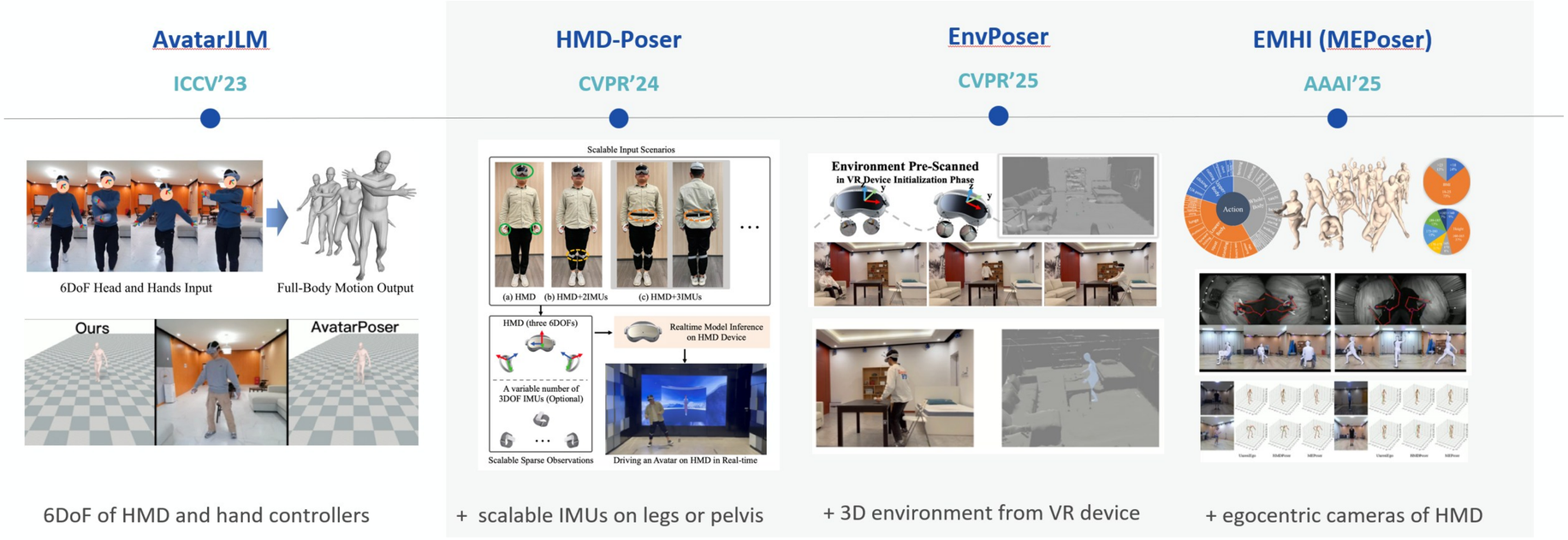
3.8 后续工作

局限：

- ✗ 无法还原极端复杂的下半身动作动作（如舞蹈/跌倒重建失败）
- ✗ 难以仅靠头+手信号即可恢复上半身不动，下半身运动的情形

改进：

- ✅ 引入腿部IMU（HMD-Poser）/ 环境点云 （EnvPoser）/ Egocentric图像（MEPoser）等多模态信息
- ✅ 扩展真实采集数据以提升鲁棒性 (EMHI)



[1] Realistic Full-Body Tracking from Sparse Observations via Joint-Level Modeling, *ICCV* 2023.
[2] HMD-Poser: On-Device Real-time Human Motion Tracking from Scalable Sparse Observations, CVPR 2024.
[3] EnvPoser: Environment-aware Realistic Human Motion Estimation from Sparse Observations with Uncertainty Modeling, CVPR 2025.
[4] EMHI: A Multimodal Egocentric Human Motion Dataset with HMD and Body-Worn IMUs, AAAI 2025.

THANKS

ByteDance字节跳动