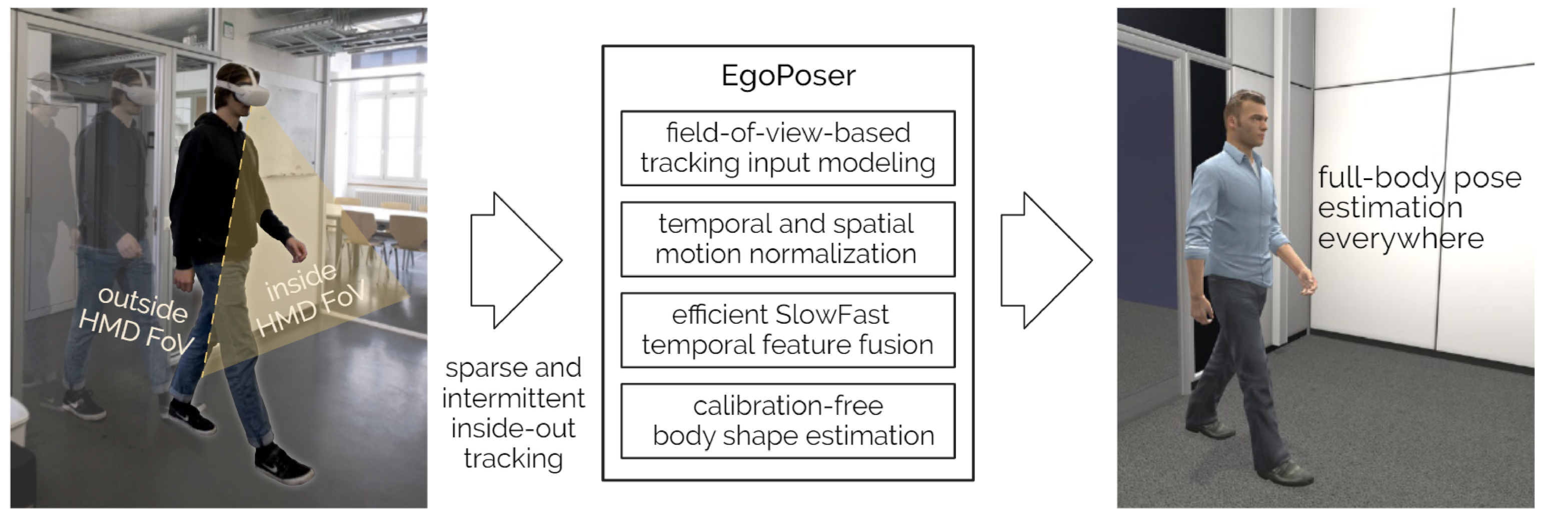
EgoPoser: Robust Real-Time Egocentric Pose Estimation from Sparse and Intermittent Observations Everywhere

Jiaxi Jiang, Paul Streli, Manuel Meier, Christian Holz
Department of Computer Science, ETH Zürich, Switzerland

project page

EUROPEAN CONFERENCE ON COMPUTER VISION
MILANO 2024

introduction



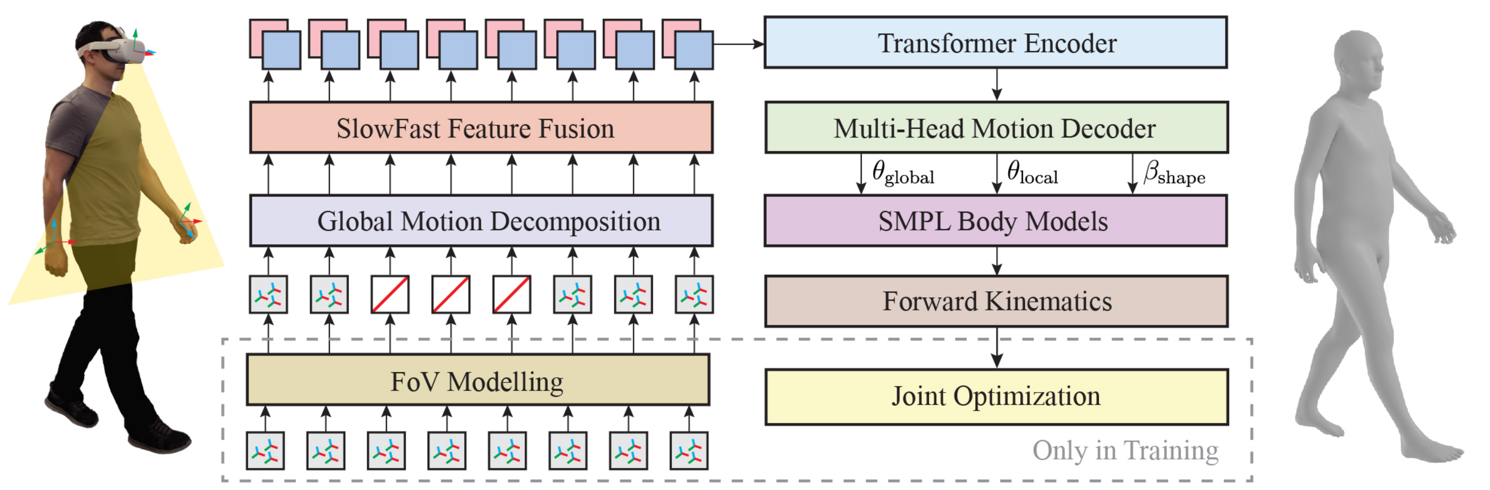
We present EgoPoser, a full-body egocentric pose estimation method that utilizes a single existing Mixed Reality headset. The headset provides global positions and orientations for itself, as well as for the tracked user's hands or controllers.

Limitations of previous work:
- Assumes hand pose data is always available without considering intermittent tracking signals.
- Only works well in limited spaces, failing to provide stable pose estimation in large scenes.
- Captures only short input sequences without utilizing long-term historical information.
- Assumes a mean body shape, without considering the diverse body shapes of different users.

EgoPoser has four main contributions:
- EgoPoser robustly models body pose from intermittent hand tracking signals only when inside a headset's field of view.
- EgoPoser supports robust pose estimation in any position via a novel global motion decomposition method.
- EgoPoser enhances pose estimation by capturing longer motion time series through an efficient SlowFast module design that maintains computational efficiency.
- EgoPoser can jointly predict pose and shape and thus generalizes across various body shapes for different users.

method

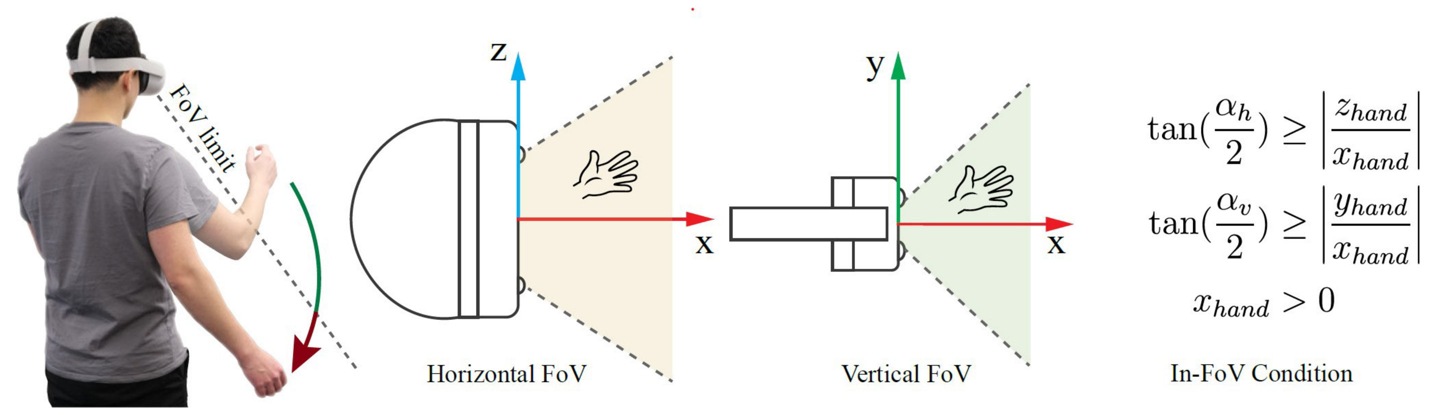
1. Method overview



The architecture of EgoPoser for full-body pose estimation from an MR device. We mask the tracking signals according to realistic FoV modelling during training. The global motion decomposition decomposes global motion from input tracking signals, making the model robust to different user positions. We sample these signals at different rates, capturing both dense nearest information and sparse but longer information, and fuse them through the SlowFast module. The Multi-Head Motion Decoder outputs parameters for global localization, local body pose, and body shape prediction.

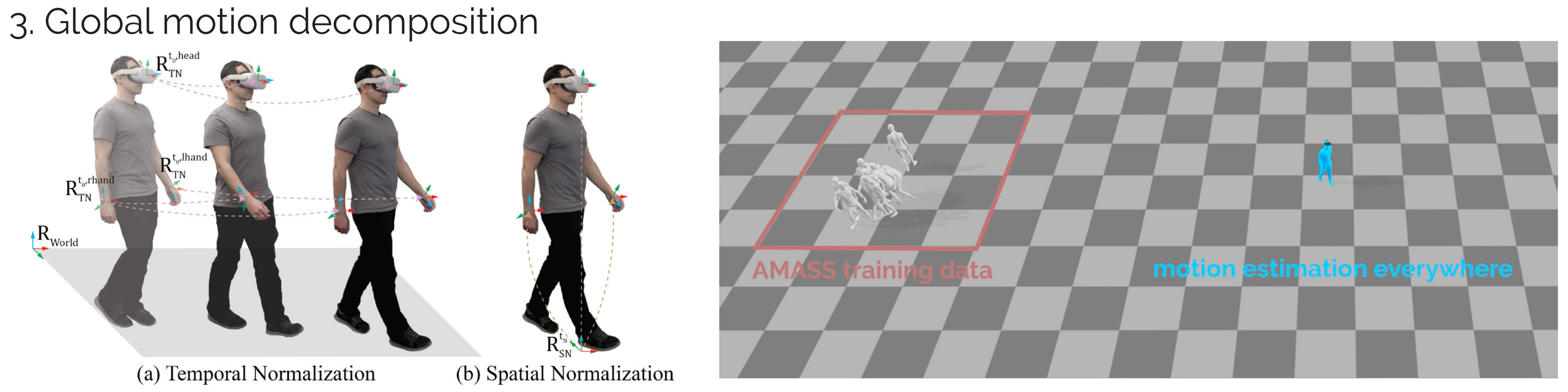
Given N=80 frames as input, we generate the last frame as the full-body representation for each timestamp, facilitating real-time applications.

2. Realistic field-of-view modeling



Based on the head pose, which determines the viewing angle of the headset, and the relative position of the hands, we simulate hand tracking failures for headsets with varying FoVs.

3. Global motion decomposition

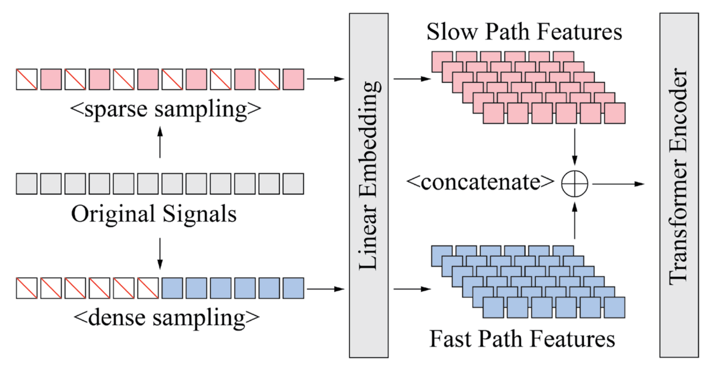


With the global motion decomposition, we can train our model only on the indoor AMASS dataset and have a robust motion prediction everywhere.

Temporal normalization: We subtract the translation of each joint at the first frame from the corresponding joint positions over the temporal window. This operation extracts the relative global trajectory of each joint across the temporal window.

Spatial normalization: we normalize only the horizontal translations relative to the head. The global vertical translation is retained as a crucial feature to encode motion priors.

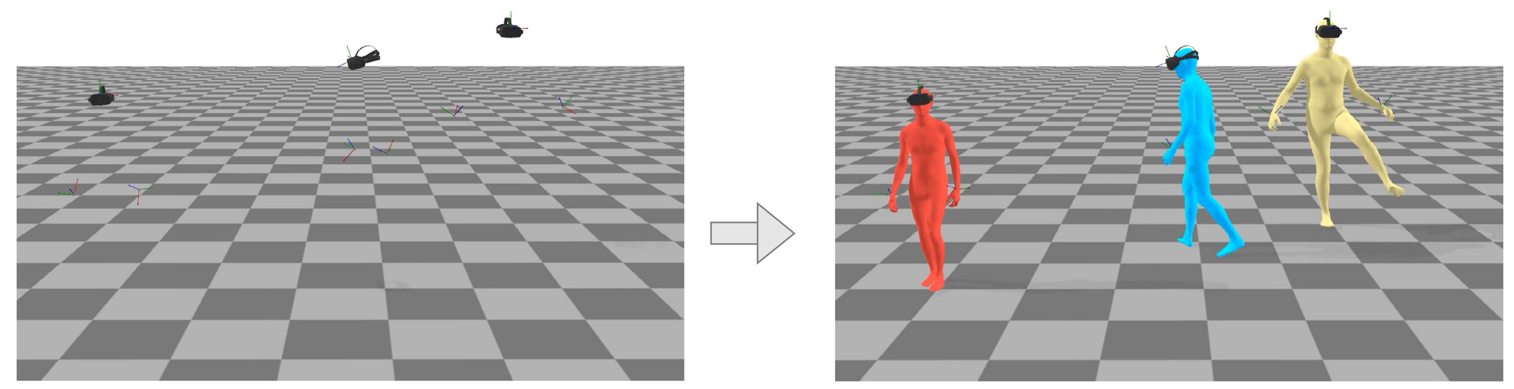
4. SlowFast feature fusion



Original signals are sparsely and densely sampled and then concatenated, thus improving prediction accuracy without increasing the computational burden.

Specifically, given an input window of τ past frames, the SlowFast module concatenates the linear embeddings for the last $\tau/2$ frames (FAST) with $\tau/2$ frames sampled with a stride of 2 over the complete window (SLOW).

5. Shape-aware pose estimation



Our method directly estimates the user's body shape from the poses of the headset and the user's hands. For more details, please refer to our paper.

numerical results

1. Comparisons with SoTA methods on the HPS dataset

Methods	BIB_EG_Tour MPJPE	BIB_EG_Tour MPJVE	MPI_EG MPJPE	MPI_EG MPJVE	Working_Standing MPJPE	Working_Standing MPJVE	UG_Computers MPJPE	UG_Computers MPJVE	Go_Around MPJPE	Go_Around MPJVE
AvatarPoser [21]	22.53	60.25	16.54	36.39	19.08	52.95	23.24	40.65	19.50	59.54
AvatarPoser-Improved	11.48	82.70	13.86	59.66	12.42	77.83	11.42	50.46	12.56	82.42
AGRoL [11]	28.95	166.34	19.41	55.52	17.67	53.97	20.90	109.12	14.16	98.34
AGRoL-Improved	15.04	124.12	13.94	89.42	13.86	89.42	12.71	106.43	13.13	128.42
AvatarJLM [51]	41.27	82.92	12.91	50.44	17.26	69.08	21.31	55.42	11.57	62.18
AvatarJLM-Improved	14.80	79.66	14.72	45.57	13.75	68.98	10.28	45.74	11.19	68.87
EgoPoser (Ours)	**9.55**	**49.39**	**11.05**	**35.60**	**8.70**	**46.49**	**10.25**	**38.29**	**6.90**	**45.10**

2. Ablation on shape estimation

Strategies	MPJPE	Vertex	Height	Arm	GP	FF
Mean Shape	6.36	6.74	7.67	7.42	3.87	5.38
Ours 1 - DA + Calib.	5.26	4.69	**1.36**	**1.24**	**2.06**	1.67
Ours 2 - Shape Est.	**4.79**	**4.08**	1.78	1.66	2.31	**1.64**

3. Ablation on field-of-view modeling

Strategies	FoV = 180° MPJPE	FoV = 180° MPJVE	FoV = 120° MPJPE	FoV = 120° MPJVE	FoV = 90° MPJPE	FoV = 90° MPJVE
Full Visibility [21]	24.75	183.84	38.99	144.42	41.24	95.66
Random Masking [7, 11]	7.09	49.91	13.29	64.09	14.84	58.33
Improved RM	6.52	47.50	11.88	57.44	12.83	52.98
Ours	**5.31**	**39.69**	**6.07**	**46.01**	**6.60**	**48.25**

4. Ablation on global motion decomposition

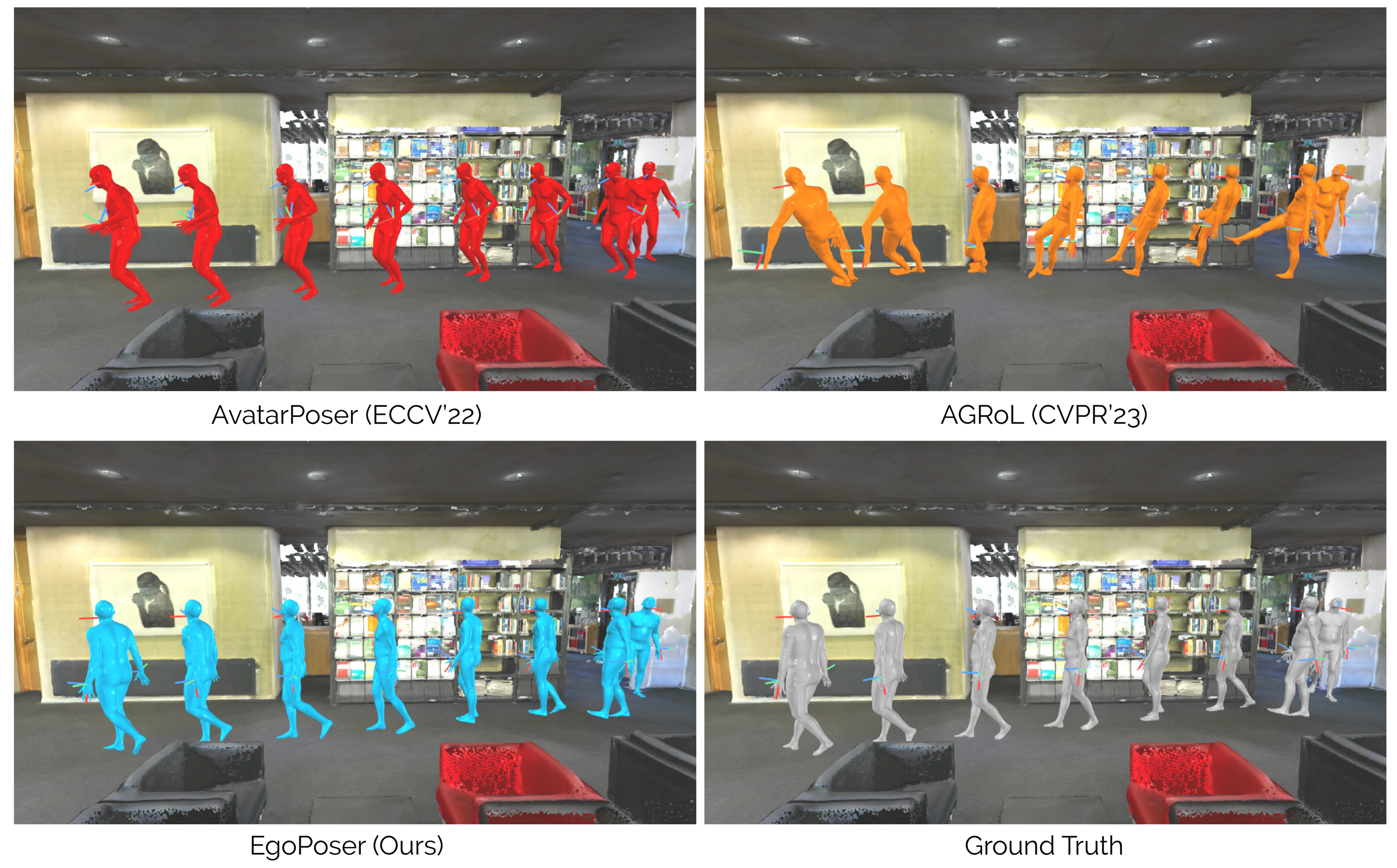
Strategies	MPJPE	MPJVE
Mean Norm. (all features)	6.25	42.69
Mean Norm. (horiz. + vert. pos.)	6.24	42.75
Mean Norm. (horiz. pos.)	6.25	42.87
Spatial Norm. (horiz. + vert. pos.)	4.96	29.59
Spatial Norm. (horiz. pos.)	4.45	27.56
Temporal Norm.	4.58	28.01
Ours	**4.14**	**25.95**

5. Ablation on SlowFast feature fusion

Strategies	MPJPE	MPJVE	FLOPs	#Parameters
length 40	4.36	28.12	0.33G	4.12M
length 80	**4.11**	29.27	0.65G	4.12M
length 80 , s=2	4.13	30.02	0.33G	4.12M
Ours	4.14	**25.95**	0.33G	4.12M

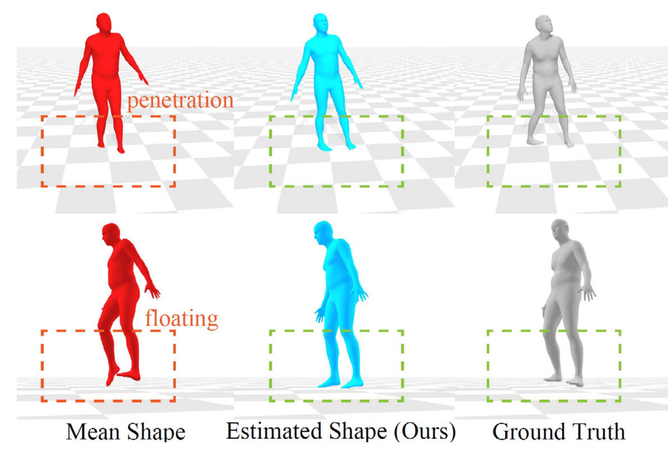
visual results

1. Visual comparisons with SoTA methods on the HPS dataset

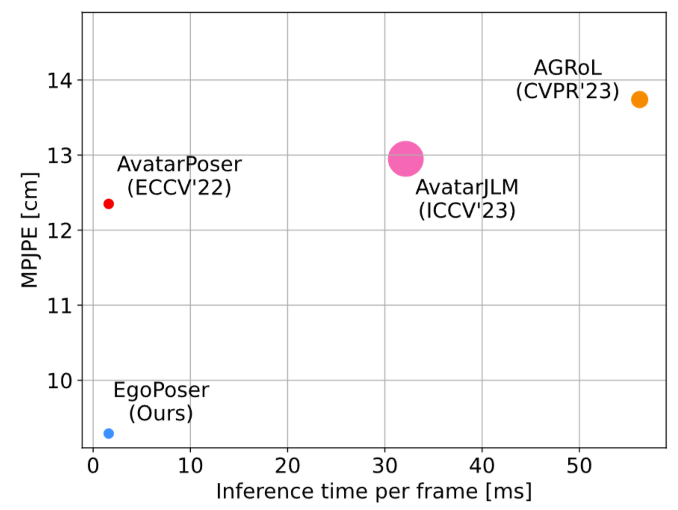


2. Visual results under hand tracking mode

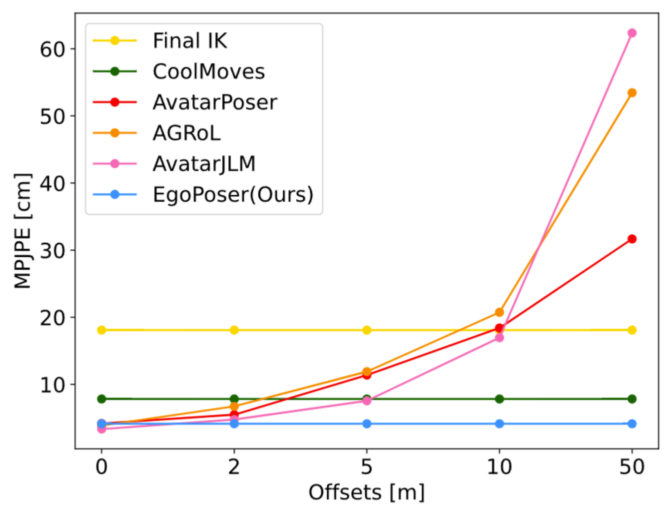
3. Visual comparison of shape estimation



4. Running time comparisons



5. Robustness to the offset of the origin



test on MR devices

EgoPoser works on popular MR systems. We used an Meta Quest 2 as well as two controllers, each providing real-time input with six degrees of freedom (rotation and translation).