

LAMP: Localization Aware Multi-camera People Tracking in Metric 3D World

Supplementary Material

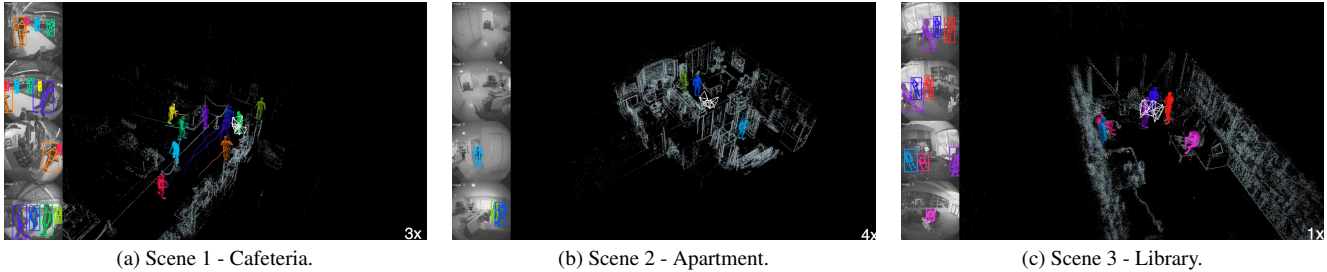


Figure 8. **Real-time real-world demo with Aria Gen 2 [49] headset.** We show 3 scenarios with Aria Gen 2 headset to showcase LAMP in tracking multiple people for casual social activities. Note the algorithm is trained on simulation and tested with real-world data.

6. Supplementary Video

In order to provide better visual assessment of LAMP in both 3D grounding and temporal consistency, we provide videos to visualize the algorithm outputs. We kindly refer the readers to the supplementary video. Following we briefly describe the video content.

6.1. Evaluation on public datasets

We show further comparison of LAMP with the state-of-the-art method PromptHMR [79]. Figure 9 shows the screenshots for the result on the Nymeria dataset [45] and Fig. 10 shows the screenshot for the result on the EMDB dataset [24]. The video rendering uses the same color coding as used in Fig. 5, which corresponds to the Per Vertex Error (PVE) computed in the world coordinate. To provide 3D visual reference, we show the trajectory of observing camera in purple. For Nymeria dataset, the target person also wears a headset, which is localized in the same metric coordinates as the observing cameras. We therefore show their headset trajectories in green to highlight the accuracy of 3D body grounding. If the tracking is accurate, the position of estimated head should match the green headset. In the video, we show that while LAMP performs on par with PromptHMR in estimating the local body poses, LAMP also consistently outperforms PromptHMR in grounding the body motion in metric 3D world. This proves the effectiveness of LAMP in factoring out the headset motion in its formulation. Note that all results from LAMP on EMDB are zero-shot, showing the effectiveness of using multi-view temporal posed 3D rays directly in LAMP-Net training.

6.2. Real-time real-world demo

In addition to evaluation on public benchmark, we show live demos running in real-time with real world data. To this end, we use the Project Aria Gen 2 [49] headset and

collect 3 diverse scenarios where multiple people perform casual activities at home, in the office and at the cafeteria (c.f., Fig. 8 for screenshots of each demo). For real-time demo, we use a lightweight MHR model [16] instead of SMPL [43]. This change only requires minor alter to our algorithm to output MHR parameters. Note LAMP is only trained with simulated Aria Gen 2 data for real-world testing, which is benefit from the ray-based formulation. The demos highlight real-world challenges, with rapid motion, natural occlusions and 2D observations of the same people constant switching across different cameras over time. Leveraging the spatio-temporal posed 3D ray fusion paradigm, LAMP is able to handle these challenges well.

7. Additional Experiments

Camera pose sensitivity In the paper we compare *monocular* LAMP against baselines on EMDB using *GT camera poses*. Here we further perturb GT poses with temporally correlated SE(3) noise (sampled every 10s interval and smoothly ramped within the interval), sweeping translation/rotation from 2–8 cm and 0.02° – 0.2° (Tab. 3), which are in the range of SOTA academic VIO systems like OKVIS2 [35]. We also include the results using poses from DROID-SLAM [71]. As expected, LAMP degrades with noisy poses, but remains superior to PromptHMR (PHMR) even when PHMR uses GT poses. The advantage remains when both methods use DROID-SLAM poses. The results suggest LAMP yields both a higher upper bound with accurate poses and robustness under realistic pose errors. It is important to point out that industrial VIO/SLAM, such as Aria localization algorithm used in our work, are substantially more accurate than academia solutions as reported in LaMAria benchmark [33] and yield order of magnitude lower noise than the value used in Tab. 3. This motivates our design to disentangle camera motion and human motion.

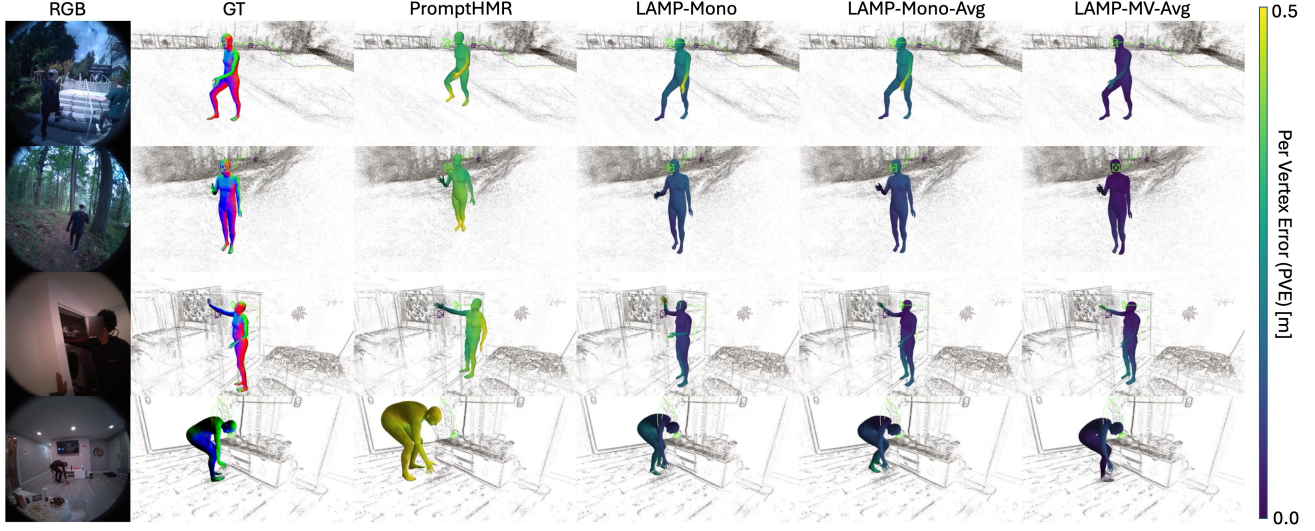


Figure 9. **Qualitative comparisons on Nymeria.** We compare PromptHMR [79] with LAMP variants, and show the benefits of using the temporal averaging and multi-view inputs. The vertices are colored by Per Vertex Error (PVE) in the world coordinate. Please refer to the supplementary video to view the full comparison.

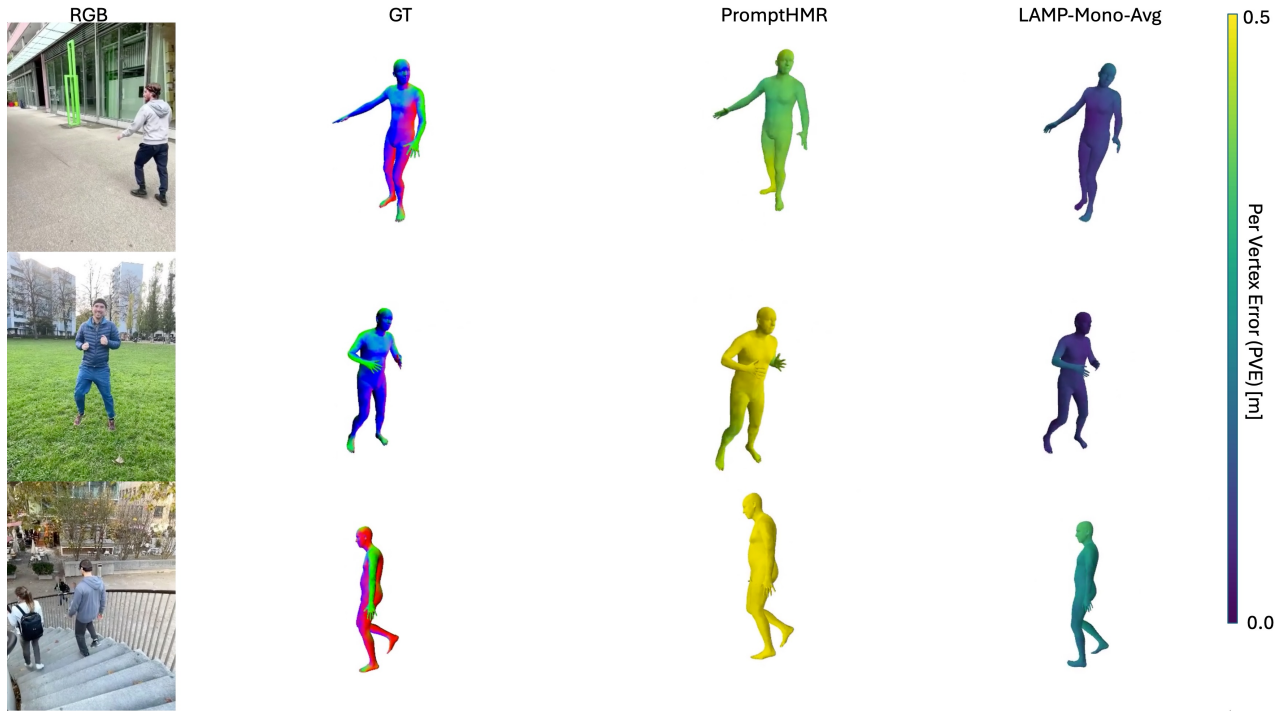


Figure 10. **Qualitative comparison on EMDB.** We compare LAMP-Mono-Avg with PromptHMR [79] using the monocular video input from EMDB. Note the result from LAMP shows the zero-shot generalization without training on EMDB. Please refer to the supplementary video for full assessment.

2D keypoint sensitivity We ablate different ViTPose backbones (S–H) and added Gaussian noise to ViTPose-H on EMDB in Tab. 4. The results show that LAMP degrades minimally under significant noise, confirming tolerance to both limited model capacity and pixel-level jitter. The robust-

ness benefits from the extensive data augmentation during training as described in submission.

Runtime, latency clarification LAMP comprises 3 components: YoloX-S 2D detection, ViTPose-S 2D keypoints and LAMP-Net. Fig. 11 breaks down the runtime on RTX 4090

Method	Cam. Pose	W-MPJPE	Jitter
PHMR	GT	278.1	16.3
LAMP	GT	165.1	4.6
LAMP	2cm/0.02°	165.7	4.7
LAMP	4cm/0.05°	170.2	4.8
LAMP	6cm/0.1°	181.9	4.8
LAMP	8cm/0.2°	212.3	5.0
PHMR	DROID	294.3	17.1
LAMP	DROID	273.9	5.2

Table 3. Ablation on camera poses.

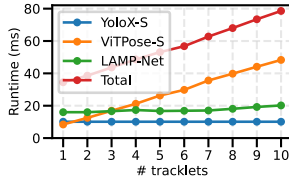


Fig. 11 LAMP System Runtime

2D Kp	W-MPJPE	Jitter
ViTPose-S	170.7	4.7
ViTPose-B	168.2	4.6
ViTPose-L	165.4	4.6
ViTPose-H	165.1	4.6
$\sigma = 1px$	165.1	4.6
$\sigma = 5px$	165.9	4.6
$\sigma = 10px$	166.8	4.8

Table 4. Ablation on 2D KPs.

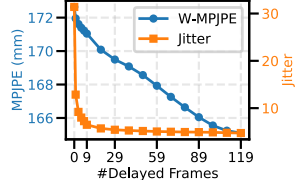


Fig. 12 Temporal Averaging

against number of tracklets. With 10 tracklets, LAMP runs at ~ 12.5 Hz, whereas PHMR runs at ~ 6 Hz with only 1 tracklet. Fig. 12 shows the tradeoff of delay over accuracy/stability with temporal smoothing, allowing applications to choose latency budget.

Tracking Performance We clarify that “tracking” in LAMP emphasizes *world-grounded motion estimation* rather than long-term re-identification. Standard MOT metrics are unsuitable here as our benchmarks provide ground truth for only a single subject. To address tracking concerns, we computed *3D tracking recall* on Nymeria (0.25m threshold at the pelvis, ignoring IDs); LAMP achieves 90.3%, confirming high coverage. Additionally, Fig. 6 quantifies multi-camera benefits via tracking coverage analysis. We further illustrate high recall and stable identity association under rapid camera motion and partial view dropouts in the supplementary video (e.g., crowded cafeteria).

8. Data Augmentations

We conduct extensive data augmentations to improve the robustness of LAMP-Net on the real-world 2D keypoints with two families of augmentations: temporally correlated noise on visible joints and structured masking that removes observations in realistic patterns.

Noise We add temporally correlated Gaussian jitter per joint/keypoint track to model detector behavior. The correlation is high so noise evolves smoothly over time. We also add per-frame noise of which the noise magnitude is smaller. In addition, we also make distal joints $1.5\times$ noisier (e.g., wrists/ankles) for which the detections are usually less stable.

Masking We compose two simple masks to simulate occlusions and view dropouts. First, we random sample time spans with 10 to 20 frames per span. For each span we sample the number of active views, biasing toward multi-view frames. To cover the monocular case, with a small probability, we force the entire input snippet to have exactly one active view. Second, Within each active view we mask joints in contiguous temporal bursts to simulate self-occlusion and tracking drops. Bursts last 10 to 20 frames with 1 to 4 joints. We directly set both the coordinates and confidence to 0 when the points are masked out. In addition, we use a higher probability to mask out feet to simulate real world ego-centric scenarios.

To retain peak accuracy on clean data while gaining robustness, we mix two clip types in training, i.e., clips without noise or masking and clips with light jitter and no masking.