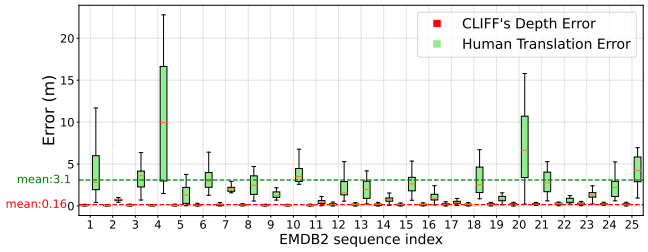
Humans as Checkerboards: Calibrating Camera Motion Scale for World-Coordinate Human Mesh Recovery

Supplementary Material

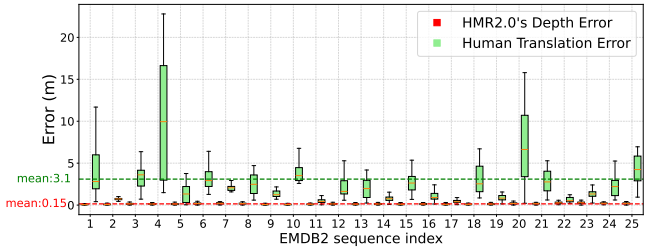
A. Validity of Humans as Checkerboards

In this section, we provide additional details to support the concept of using humans as scale calibration references, as discussed in Sec. 4.1 of the main paper.

Sufficient Small Depth Error. Fig. 2 in the main paper showcased data from only five randomly sampled examples for better clarity and visualization, here in Fig. 7 we expand this analysis by comparing the two errors across the entire EMDB2 [2] dataset. Fig. 7a demonstrates that the depth prediction error of CLIFF model [4] is consistently lower than that of human translation across all samples. Additionally, as illustrated in Fig. 7b, testing with different HMR models such as HMR2.0 [1] has shown a similar trend. These results show that HMR can predict sufficiently accurate depth compared to the global translation error we are addressing in our task [5]. This validates our motivation of using humans as reliable metric cues for scale calibration.



(a) CLIFF's depth prediction error versus global human translation error.



(b) HMR2.0's depth prediction error versus global human translation error.

Figure 7. Quantitative analysis of depth prediction errors from different HMR models, (a) CLIFF and (b) HMR2.0, compared to the global human trajectory error across the entire EMDB2 dataset.

Consistent Bias in Depth Error. Besides the sufficiently small error magnitude, the depth prediction errors from HMR show a consistent pattern within each video sequence. Compared with ground truth, HMR tends to predict either predominantly larger or smaller depth across all frames in a video (as shown in Fig. 8). This consistency is advantageous for calibration purposes: using humans as calibration references can systematically compensate for this bias. Our human-aware calibration effectively adapts to the inherent HMR depth bias, enabling precise overall global human motion recovery, as shown in Sec. 5.3 of the main paper.

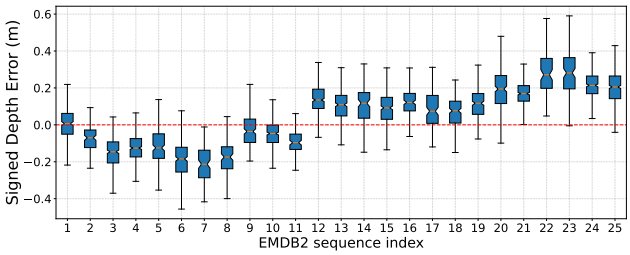


Figure 8. HMR tends to predict either predominantly larger or smaller depth for each video compared with ground truth.

B. Assumptions and Verification

In this section, we clarify the assumptions made by our method and provide corresponding validation experiments.

Firstly, we emphasize what we do *not* assume. For example, we do not assume or require human–scene contact in every single frame (see Sec. G(1)), nor do we assume that the terrain is always a flat ground plane (see Sec. G(2)).

Our method relies on the following key assumptions:

1. **Prevalence of Human–Background Contacts:** We assume that human–background contacts are typically observed in the majority of frames in human-centric videos. This assumption is supported by data, as shown in Tab. 7, where 98% of frames in EgoBody dataset and 86% in EMDB2 dataset exhibit such contacts. (GT contact label obtained following HumanML3D [CVPR22])
2. **Contact Joint as the Lowest Joint:** We assume that, within these human–background contacts, the contact joint is typically the lowest human joint along the camera's y-axis. This assumption holds in 90% of cases in EgoBody and 85% in EMDB2 (see second row Tab. 7).

Table 7. Proportion of contact frames in indoor&outdoor datasets.

	EgoBody (indoor)	EMDB 2 (outdoor)
% frames with contact	98%	86%
% where lowest joint is contact	90%	85%

C. Qualitative Results for Scale Calibration

In this section, we present additional qualitative results that demonstrate the effectiveness of our scale calibration, as mentioned in Sec. 5.2 of the main paper. Fig. 9 illustrates the challenge of the unknown scale factor in the initial SLAM output (green trajectory). After our scale calibration, the resulting camera trajectory (blue trajectory) aligns more closely with the ground truth (grey trajectory), demonstrating the effectiveness of HAC.

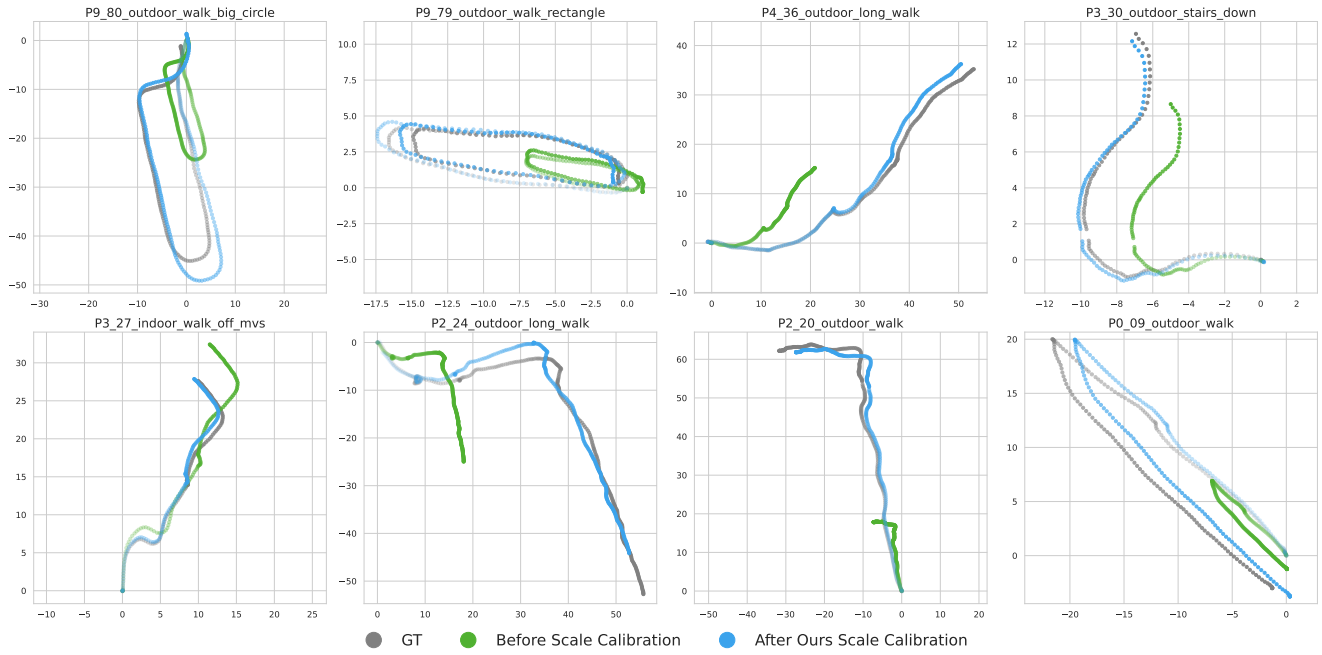


Figure 9. Visualization of camera trajectory before and after our scale calibration. As shown, the original SLAM output is up to an unknown scale. After our scale calibration, it becomes better aligned to ground truth data.

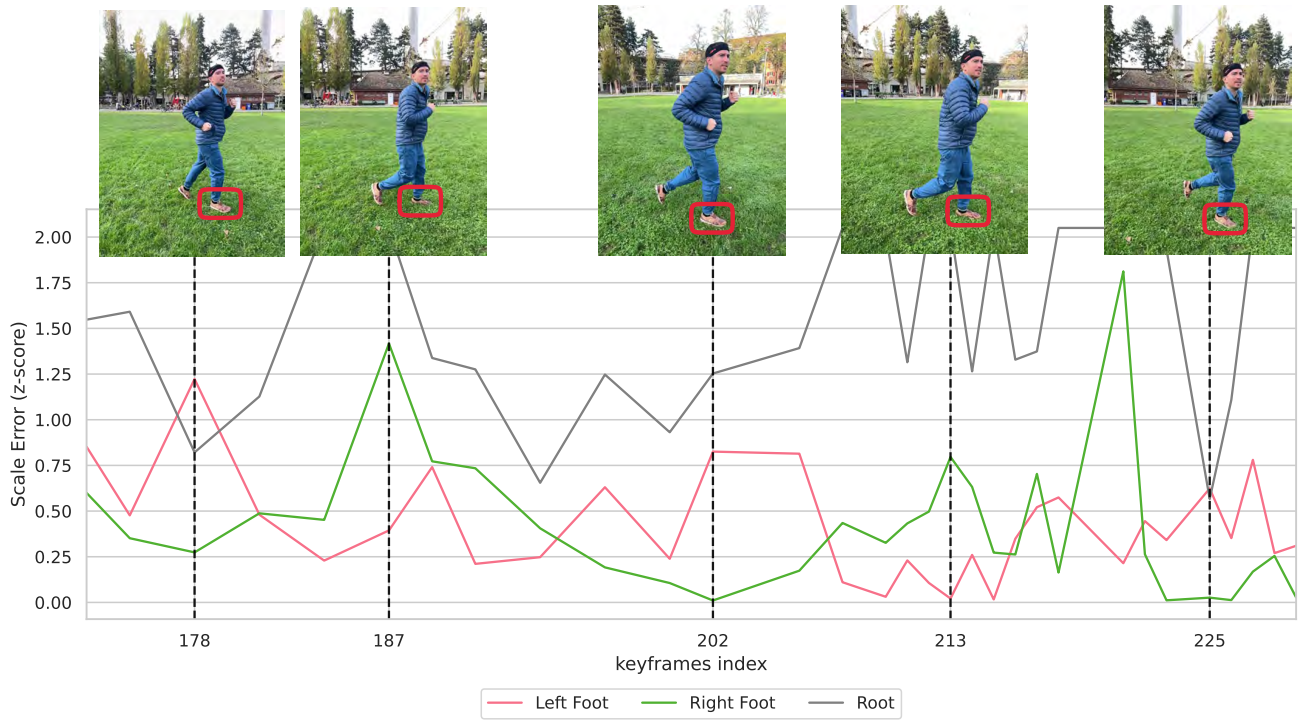


Figure 10. Scale error across some keyframes of the video. Results show a low scale error when the foot is in contact with the ground. Additionally, an inverse correlation between the left and right foot scale errors corresponds with the alternating pattern of footfalls during locomotion.

D. More Insights on Reference Joint Selection

In this section, we delve deeper into how the selection of reference joints impacts the performance of scale calibration, as discussed in Sec. 5.2 of the main paper. Fig. 10 demonstrates pelvis joints show a larger scale error than foot joints, aligning with the findings in Tab. 1 of the main paper. Furthermore, an inverse correlation exists between scale errors of left and right feet, with growth in the left feet corresponding to a contraction in the right. This scale error trend is also consistent with the contact feet as shown in the corresponding image frames, a foot will have less scale error when it contact with the ground. This further verified our motivation for choosing the human-scene contact joint as the reference point.

E. Full runtime comparison including SLAM

We excluded SLAM time in Sec. 5.3 of the main paper since HAC and [3, 6] all use SLAM. Adding it back, HAC is still much faster (see below Tab. 8) since others require post-SLAM optimization.

Table 8. Full runtime comparison (including SLAM).

	SLAHMR	PACE	HAC
Runtime/1000imgs (including SLAM)	402min	10min	**124sec**

F. Video Demos and Point Clouds

In this supplementary folder, we provide the video ***HAC_Qualitative_Examples.mp4***, which illustrates some of our results in video form. Additionally, a folder named ***scene_reconstruction_with_human*** contains corresponding files that show the human and camera meshes along with the scene point clouds.

HAC_Qualitative_Examples.mp4 presents a three-panel side-by-side illustration: (i) the input video, (ii) our results from the current SLAM camera point of view, and (iii) our results from a global static point of view. These visualizations demonstrate that our reconstructed scale is accurate, with smooth and coherent human and camera motion fitting very well within the world coordinates.

scene_reconstruction_with_human **folder** contains the comprehensive human, camera meshes, and scene point cloud corresponding to the aforementioned video. For optimal viewing, we recommend using a 3D visualization tool such as MeshLab. Open **-scene_point_cloud.ply* and **-human_and_camera_meshes.ply* together in MeshLab to explore the human and camera motion within the scene.

G. Discussion

In this section, we provide additional discussion about our method as follows:

1) What if no contact in some frames?
We do not assume or require human-scene contact for every single frame. As described in Sec. 4.4, we determine a unified scale for the entire sequence by taking the median of all frames. Thus, some frames without contact (*e.g.*, running or jumping) are acceptable, provided the majority have contact (98%, 86% of frames for EgoBody, EMDB2 respectively). Even for sequences with less contact (Tab. 9), our estimated mesh outperforms the SOTA method and has virtually no performance drop compared to using GT contact frames for scale (GT contact obtained following HumanML3D [CVPR22]).

Table 9. Robustness evaluation on minimal contact sequences.

Sequence	% frames have contact	WA-MPJPE		
		TRAM	Ours	GT contact
Outdoor run	56%	65.9	63.4	63.3
Stairs up	71%	85.9	60.1	60.0

2) Do we always assume terrain as a flat plane?
We do not make this assumption. Our method supports complex terrains (*e.g.*, staircases and climbing above the ground, see Fig. 11 and supplementary video). We only assume that some out-of-view (OOV) portions of the terrain form a plane with the point of contact, which we see as a reasonable compromise under this challenging setting. Tab. 2 of the main paper shows how our ground plane fitting is simple yet robust.

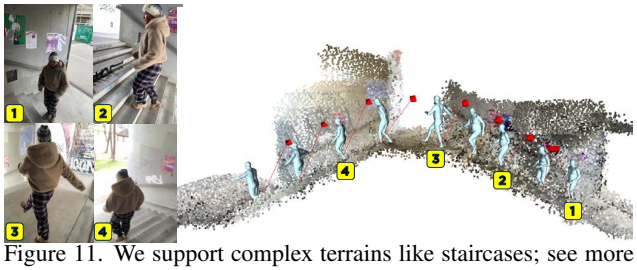


Figure 11. We support complex terrains like staircases; see more examples in the supplementary video.

3) Comparison analysis: ours vs. optimization-based.
HAC offers a significant advantage over previous scale-optimization methods [3, 6, 7] by substantially reducing post-SLAM inference time. As detailed in Sec. 5.3 of the main paper and Sec. E, HAC reduces inference time by approximately 100× by explicitly using the human as the calibration reference. This explicit calibration not only accelerates processing but also enhances robustness. In contrast, global optimization methods optimize all parameters including scale under complex loss functions, making it challenging to converge to the correct scale factor in ambiguous solution spaces. For instance, as illustrated in Fig. 1(b) of the main paper, SLAHMR's scale estimates can deviate by 100% or more. Consequently, our method consistently outperforms global optimization, achieving a root translation error that is only 11% of SLAHMR's (Tab. 3 of the main paper). The performance gap narrows only in datasets with reduced camera motion (*e.g.*, EgoBody exhibits less than 1/10 of the camera motion observed in EMDB2), as the impact of scale factor diminishes when there is limited camera movement. Only in extreme scenarios discussed in our

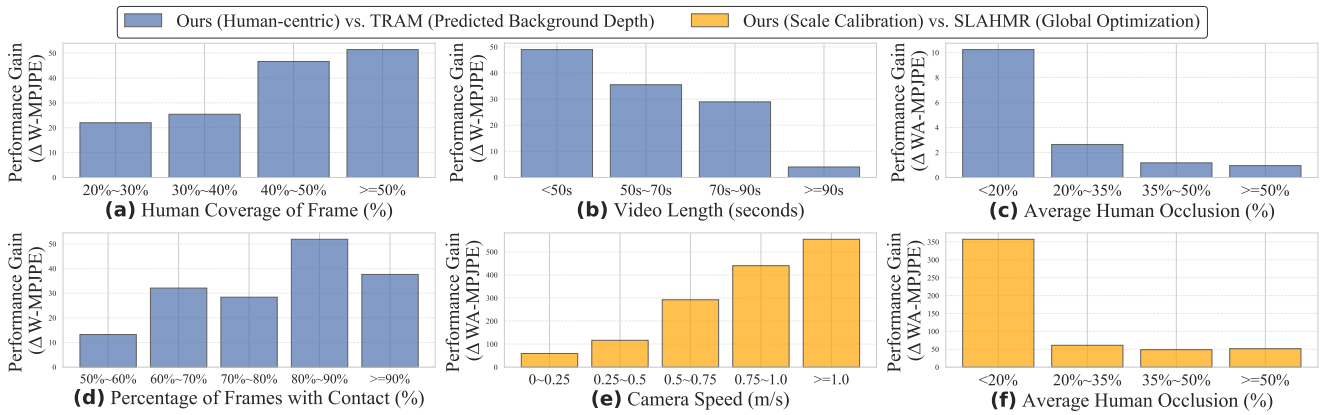


Figure 12. Case analysis of HAC versus TRAM (a–d) and SLAHMR (e–f). Statistics are computed per sequence on EMDB and EgoBody.

limitation section—such as when there is minimal contact (*e.g.*, floating mid-air or swimming)—our method may underperform relative to optimization-based. However, such cases are exceptionally rare for world-coordinate HMR. Moreover, optimization-based methods also face challenges in these scenarios if their motion prior is trained on flat ground, while also incurring substantially longer optimization times.

4) Why human-centric calibration beats TRAM? Human-centric calibration is meaningful in videos where humans dominate the foreground. In such cases, background-based calibration like TRAM struggles, due to: **(a) Scarce background cues**: as humans occupy more of the foreground, the performance gap between TRAM and ours widens (see Fig. 12a). **(b) Unreliable background depth**: motion blur, far distances, and background dynamics degrade TRAM's estimates, whereas we benefit even from multiple moving humans (Fig. 14). The above explains why TRAM requires more redundancy, *i.e.*, longer sequences to be comparable to us (Fig. 12b).

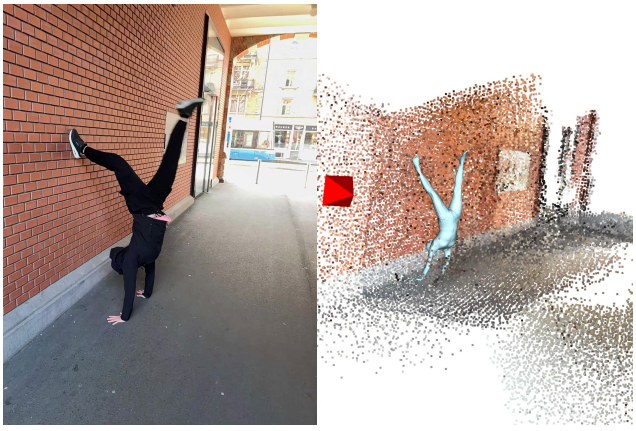


Figure 13. Adaptive contact joint identification for a handstand example, where our method correctly identifies the hands as the contact points, ensuring robust scale calibration for unusual postures where the feet are not the real contacts.

5) Any module that explicitly detects contacts? We are not explicitly detecting the human-scene contact joint. Instead, we assume that the lowest human joint in the camera coordinate system's y-axis is the contact joint, as mentioned in Sec. 4.2 of the main paper. This strategy identifies feet as the contact points for most cases, given their usual position as the lowest joints. However, in cases where an unusual posture (*e.g.*, a handstand as shown in Fig. 13) makes another joint the lowest point (*e.g.*, the hands), our method can adaptively identify that joint as the contact point. We validate this assumption in Sec. B and show the lowest joint to be the contact in 90% in EgoBody and 85% in EMDB2. This accuracy level is sufficient for our purposes, as our calibration only requires the joint to be close enough to the scene surface. Additionally, we utilize the median of scale estimates across the entire sequence which further ensures stability and robustness.

6) Why not use the fitted ground plane for all cases? Using one fitted ground plane as the contact surface is not always feasible, particularly in outdoor scenarios where the terrain can vary significantly, such as stairs or uneven surfaces. Therefore, when human-scene contacts are visible and the scene is well reconstructed, it is more accurate to use the reconstructed scene as the contact surface. The fitted ground plane is utilized as a fallback option for scenarios where contact points are out of view, ensuring robust scale calibration under varying conditions.

7) Multiple humans. HAC naturally supports multiple humans and can benefit from more redundancy. In contrast, TRAM loses more background info, leading to worse scale (see Fig. 14, humans in red box penetrate outside the scene).

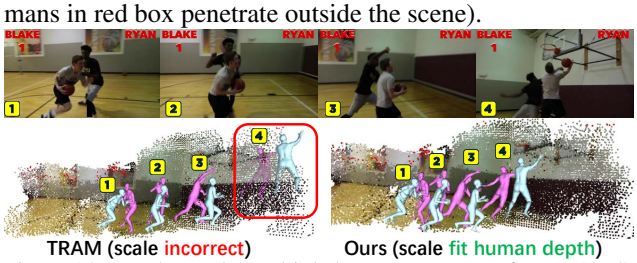


Figure 14. Results on the multiple humans sequence from DAVIS.

8) Case analysis. Fig. 12 shows that our HAC is robust and superior over di-

verse conditions based on sequence length, occlusion, and camera speeds. And HAC works well under: (a) larger camera motions (Fig. 12e, which destabilize optimization); (b) human occlusion (Fig. 12f, our performance gains are reduced but less so than optimization, so we outperform SLAHMR); (c) challenging cases like handstand (Fig. 13; SLAHMR fails since they assume feet on the ground plane).

References

[1] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. In *IEEE International Conference on Computer Vision (ICCV)*, pages 14783–14794, 2023. 1

[2] Manuel Kaufmann, Jie Song, Chen Guo, Kaiyue Shen, Tianjian Jiang, Chengcheng Tang, Juan José Zárate, and Otmar Hilliges. EMDB: The Electromagnetic Database of Global 3D Human Pose and Shape in the Wild. In *IEEE International Conference on Computer Vision (ICCV)*, 2023. 1

[3] Muhammed Kocabas, Ye Yuan, Pavlo Molchanov, Yunrong Guo, Michael J Black, Otmar Hilliges, Jan Kautz, and Umar Iqbal. Pace: Human and camera motion estimation from in-the-wild videos. *International Conference on 3D Vision (3DV)*, 2023. 3

[4] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *European Conference on Computer Vision (ECCV)*, pages 590–606. Springer, 2022. 1

[5] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J Black. Wham: Reconstructing world-grounded humans with accurate 3d motion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2070–2080, 2024. 1

[6] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21222–21232, 2023. 3

[7] Yizhou Zhao, Tuanfeng Yang Wang, Bhiksha Raj, Min Xu, Jimei Yang, and Chun-Hao Paul Huang. Synergistic global-space camera and human reconstruction from videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1216–1226, 2024. 3