

TROPHIES: Temporal Reconstruction of Places, Humans, and Cameras from Multi-view Videos

Supplementary Material

In this supplementary material, we provide additional details and experiments not included in the main paper due to limitations in space.

- Section 1: More evaluation details about metrics and datasets.
- Section 2: More details about method of TROPHIES.
- Section 3: Benefiting from the flexible structure of TROPHIES, this section demonstrates the quantitative and qualitative results of the model when receiving single view inputs.
- Section 4: More qualitative and quantitative results of TROPHIES in diverse environments.
- Section 5: Limitation of TROPHIES.

1. Evaluation Details

1.1. Metrics

We report both human pose accuracy and camera geometry consistency using the following metrics.

W-MPJPE [17] computes the mean per-joint error in world coordinates after applying a single rigid alignment to the first two frames treated jointly. This alignment anchors the motion at the sequence start, and the error is accumulated across all subsequent frames.

WA-MPJPE [17] extends this idea by performing a single rigid alignment over the entire sequence, treating all frames as one unified block. This measures global consistency of the predicted trajectories over long temporal windows.

PA-MPJPE [17] reports the per-joint position error after performing a Sim(3) Procrustes alignment for each frame individually. By normalizing translation, rotation, and scale, this metric isolates articulated pose and body-shape accuracy without being influenced by camera placement or global scene scale.

Accel [16] measures the discrepancy in joint accelerations, averaged over time. As a second-order temporal metric, it reflects the smoothness of reconstructed motion and highlights jitter or discontinuous dynamics.

For camera evaluation, we focus on both absolute trajectory accuracy and pairwise geometric correctness.

TE and **s-TE** quantify camera-center reconstruction error. TE is computed after SE(3) alignment to evaluate metric correctness of camera trajectories, while s-TE uses Sim(3) normalization to remove scale discrepancies and report scale-invariant trajectory error.

RRA@{50,100} (Relative Reconstruction Accuracy)

evaluates the correctness of pairwise camera orientations. For each camera pair, we compute the angular difference between predicted and ground-truth relative rotations and report the proportion of pairs with errors.

CCA@{50,100} (Camera–Scene Consistency Accuracy) measures how closely reconstructed camera centers match ground truth after SE(3) alignment. We report the percentage of cameras whose localization error falls below 50 cm and 100 cm.

s-CCA@{50,100} applies the same evaluation protocol after Sim(3) scaling, providing a scale-invariant measure of spatial agreement between estimated cameras and the scene layout.

1.2. Dataset

EgoHumans: The EgoHumans [4] dataset provides multi-view egocentric–exocentric recordings with challenging human motions, complex camera placements, and diverse indoor/outdoor environments. It offers accurate SMPL annotations and synchronized multi-view imagery, making it suitable for evaluating human–scene reconstruction under realistic, highly dynamic conditions.

To test the methods’ performance with the dynamic camera setting, we filter and select the original dataset to 30-60 frame sequences according to the following metrics: 1) For each sequence, we select one dynamic ego camera and three static cameras. 2) For each sequence, we ensure that the person appears in at least two cameras. We used the above rules to filter the test set from the following sequences:

- *01_tagging: 006*
- *02_lego: 004*
- *03_fencing: 006*
- *04_basketball: 006*
- *05_volleyball: 006*
- *06_badminton: 035*
- *07_tennis: 006*

In the Human Branch, we select some multi-view data from EgoHumans to finetune our multi-view aware module in the following sequences:

- *01_tagging: 001, 002, 003, 004, 005*
- *02_lego: 001, 002, 003, 005, 006*
- *03_fencing: 001, 002, 003, 004, 005*
- *04_basketball: 001, 002, 003, 004, 005*
- *05_volleyball: 001, 002, 003, 004, 005*
- *06_badminton: 002, 003, 013, 014, 019*
- *07_tennis: 001, 002, 003, 004, 005*

EgoExo4D: The EgoExo4D [2] dataset contains large-

scale synchronized egocentric and exocentric recordings capturing rich human motions, complex interactions, and diverse real-world environments. Its multi-view configuration and high-quality 4D annotations make it a strong benchmark for evaluating holistic human–scene–camera reconstruction. From the full EgoExo4D dataset containing 4,481 video sequences, we restrict our evaluation to the 133 sequences that include EgoPose annotations. These annotated sequences provide synchronized 2D/3D keypoints and camera poses, which are required for our evaluation protocol and for ensuring fair comparison across methods. Sequences without EgoPose labels are excluded, as they lack the necessary supervision for assessing human–scene–camera reconstruction quality.

EMDB: The EMDb [3] is a large-scale benchmark designed for evaluating 3D human motion reconstruction from monocular videos. It provides synchronized RGB videos together with high-quality SMPL ground-truth annotations obtained through multi-view capture setups. EMDb contains diverse indoor and outdoor scenes, varying motion types, and challenging interactions that include self-occlusion, fast movements, and significant depth ambiguities. Its well-calibrated ground truth and realistic imaging conditions make it a widely adopted dataset for assessing both pose accuracy and trajectory consistency in single-view human reconstruction methods. In our experiments, we follow the standard EMDb-2 protocol and report WA-MPJPE, W-MPJPE, and RTE for fair comparison.

2. Details of Methods

Identity association. For quantitative comparisons, we use ground-truth identities to ensure fairness. For in-the-wild video results in the supplementary, we use DEVA for consistent identity associations.

Contact likelihoods. Contact is defined as near zero velocity of hands and feet joints, following GVHMR. In the training stage, contact likelihoods are supervised utilizing ground-truth joint velocities. At inference time, they are predicted by the human branch.

Anchor view selection. During training, the anchor view is randomly sampled from the available views for generalization. During inference, we select the view with the highest human detection confidence.

3. Single View Experiments and Visualization

3.1. Quantitative Results

As summarized in Table 1, TROPHIES achieves state-of-the-art performance on the single-view EMDb-2 benchmark, despite being primarily designed for multi-view reconstruction. Compared with recent single-view methods, TROPHIES reports the lowest W-MPJPE and WA-MPJPE, outperforming strong baselines such as TRAM, GVHMR,

Table 1. Comparison with single-view reconstruction methods on EMDb-2 [3]. TROPHIES, though designed for multi-view reconstruction, generalizes effectively to the single-view setting and achieves state-of-the-art performance. Methods marked with * operate in an online fashion.

Method	EMDB-2 (24)		
	WA-MPJPE ↓	W-MPJPE ↓	RTE ↓
TRACE* [14]	529.0	1702.3	17.7
SLAHMR [10]	326.9	776.1	10.2
GLAMR [18]	280.8	726.6	11.4
JOSH3R* [6]	220.0	661.7	13.1
COIN [5]	152.8	407.3	3.5
Human3R* [1]	112.2	267.9	2.2
WHAM* [13]	135.6	354.8	6.0
GVHMR [12]	110.0	276.5	2.0
TRAM [16]	76.4	222.4	1.4
JOSH [6]	68.9	174.7	1.3
TROPHIES	68.3	172.3	1.3

and Human3R. Notably, TROPHIES obtains an RTE of 1.3, matching or surpassing online systems that explicitly track camera motion. These results highlight the robustness of our human–scene–camera modeling and show that the proposed multi-view formulation generalizes effectively even when only a single input view is available.

3.2. Qualitative results

Figure 1 presents single-view reconstruction results in EMDb-2 [3] obtained with TROPHIES. Despite operating with only one input view, TROPHIES produces stable and temporally coherent human motion trajectories aligned within the reconstructed scene. The SMPL bodies maintain consistent scale and articulation across time, allowing the walking sequence to be visualized as a smooth motion path rather than a collection of disjoint frames. Because our framework leverages a parametric human prior and globally aligned camera–scene estimates, the reconstructed poses remain well-grounded on the floor surface and avoid common failure cases such as foot penetration, floating limbs, or drifting trajectories.

The overlay of the SMPL sequence on the static scene further highlights TROPHIES’s ability to disentangle human dynamics from static background geometry. Even in the presence of partial occlusions and weak visual cues, the model recovers physically plausible motion and preserves consistent spatial placement. These visualization results demonstrate that TROPHIES generalizes effectively to the single-view setting, producing clean and interpretable human trajectories within a unified 3D scene.

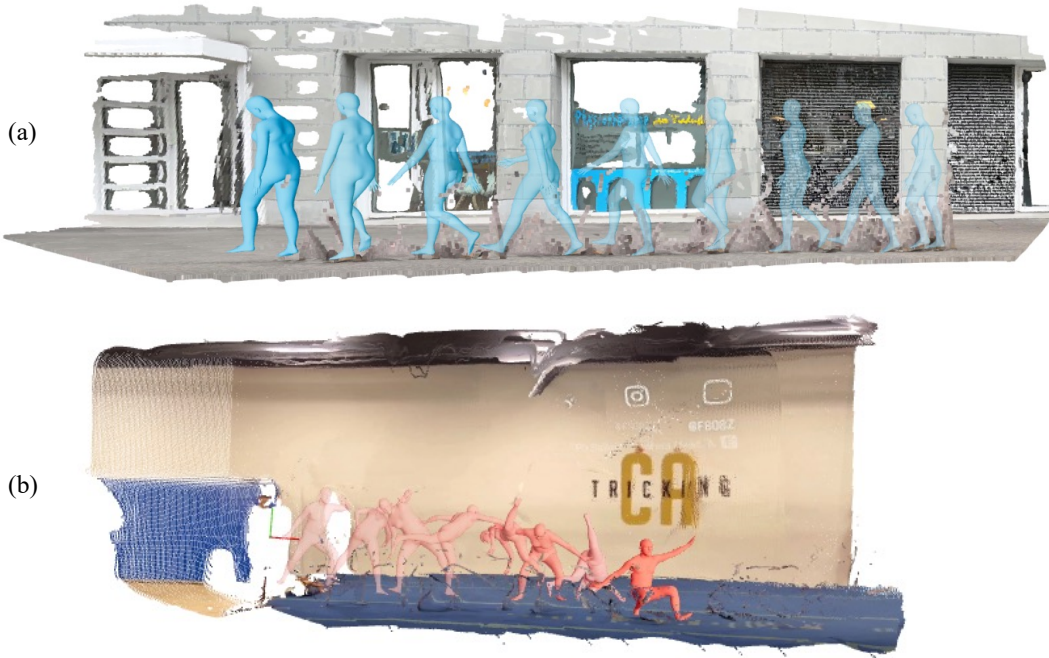


Figure 1. **Single-view reconstruction results.** (a) Visualization on the EMDB dataset, where TROPHIES recovers coherent human motion and scene geometry from a monocular input. (b) In-the-wild single-view video results, demonstrating the robustness of our method under unconstrained real-world conditions.

4. Additional Results

4.1. Quantitative Results

We add more quantitative results as shown in Table 2. First, we finetune the human branch of HSfM [8] on Ego-Humans, shown as HSfM^ψ . Second, we initialize humans with GVHMR and camera geometry with CUT3R [15], then run the same HSfM global alignment and optimization module, shown as HGC (HSfM + GVHMR + CUT3R).

4.2. Qualitative Results

We provide extended qualitative results to further demonstrate the effectiveness and robustness of TROPHIES in diverse multi-view settings. As shown in Figures 2, 3, 4, 5, our method consistently reconstructs coherent human motion, accurate camera trajectories, and detailed static scene geometry across a wide variety of indoor and outdoor environments. These additional examples highlight TROPHIES’s ability to maintain multi-view consistency under large baselines, cluttered backgrounds, and complex human interactions. Notably, both human pose sequences and scene reconstructions remain stable and well-aligned even in challenging cases with partial occlusion or rapid motion, illustrating the strong generalization capability of our joint human–scene–camera framework.

Discussion. Reconstructing humans directly as raw ge-

ometry is difficult due to deformable body motion, complex clothing, and frequent self-occlusion. Scene-focused methods that treat humans as unstructured dynamic surfaces often introduce severe artifacts: Figure 6 (a) shows fragmented and noisy human geometry, while Figure 6 (b) illustrates common misclassifications where articulated limbs blend into the background. These issues arise because raw geometry lacks strong anthropomorphic priors, making it difficult to maintain topological consistency or disentangle human motion from surrounding context.

To better focus on the underlying human motion without being distracted by clothing, loose garments, or appearance variations, TROPHIES adopts the SMPL [7, 9, 11] as the human representation. SMPL provides a compact and structurally constrained space that naturally enforces correct articulation and consistent human topology, ensuring that dynamic body motion remains clearly separated from the surrounding static environment. This design minimizes geometry noise, prevents body parts from blending into the background, and supports more stable global alignment and contact-aware optimization.

Importantly, SMPL abstracts away surface-level complexities that often obscure the true motion patterns, allowing TROPHIES to concentrate on recovering clean, interpretable, and physically meaningful human dynamics. By removing the confounding factors introduced by clothing

Table 2. Additional results on Egohuman dataset.

Results on EgoHumans	W-MPJPE ↓	WA-MPJPE ↓	PA-MPJPE ↓	Accel ↓	TE ↓	s-TE ↓	RRA@50 ↑	RRA@100 ↑	CCA@50 ↑	CCA@100 ↑	s-CCA@50 ↑	s-CCA@100 ↑
HSFM ^ψ	201.26	139.31	21.27	49.72	1.81	1.22	0.52	0.70	0.13	0.44	0.28	0.51
HGC	147.25	97.38	24.19	39.61	1.74	1.14	0.48	0.69	0.15	0.45	0.31	0.53
TROPHIES (CUT3R)	97.54	73.23	20.71	14.23	1.03	0.86	0.55	0.78	0.26	0.59	0.40	0.63

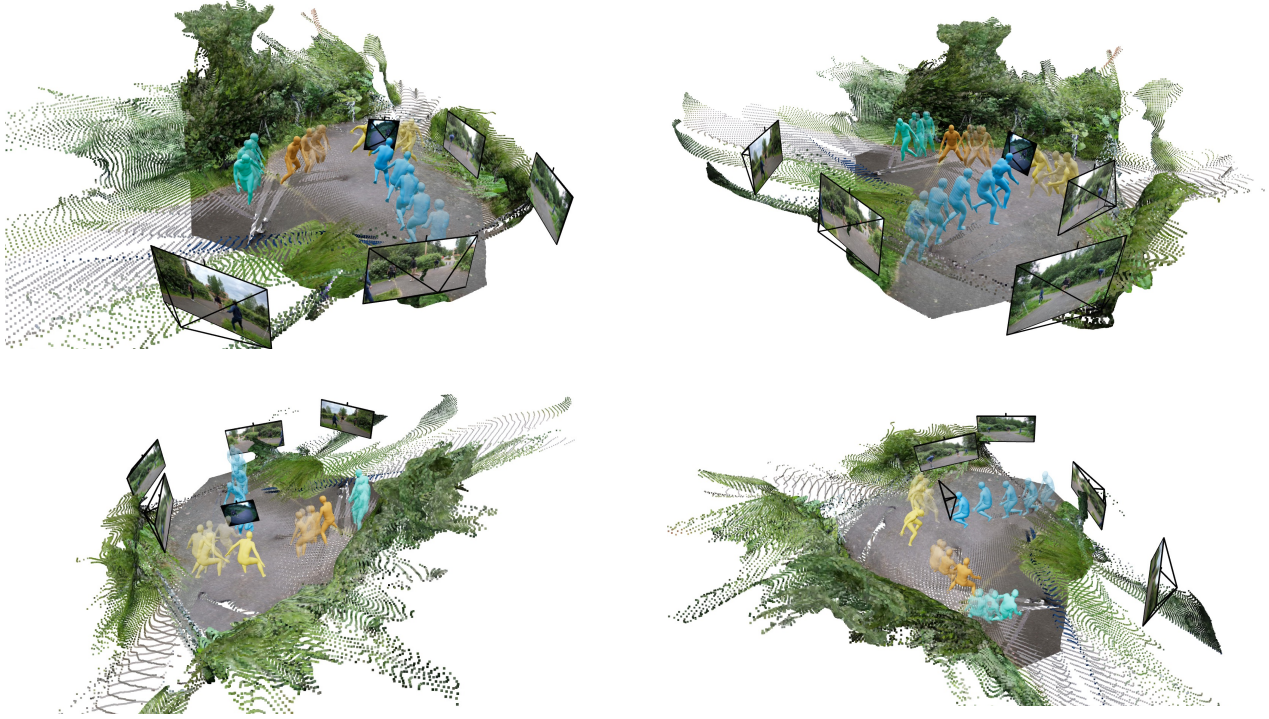


Figure 2. Multi-view outdoor results showing consistent human motion and stable scene geometry across wide baselines.

and visual appearance, SMPL enables a more principled analysis of human movement and facilitates more reliable downstream reasoning about human–scene interactions.

5. Limitation

Ego-view Cameras. In egocentric videos, rapid view-point changes, motion blur, and unstable trajectories can significantly reduce match reliability, leading to degraded scene completeness or temporal inconsistency. Egocentric datasets [4] often contain frames with different image resolutions compared to common camera or strong distortions, which further hinder robust correspondence extraction and affect the accuracy of downstream reconstruction components. Addressing these challenges will require more advanced feature matching under extreme viewpoint shifts, better handling of variable-resolution inputs, and stronger geometric or inertial priors for dynamic camera motion. We believe incorporating such components would further improve the robustness of TROPHIES.

Dynamic Objects. Human-aware attention mechanism is limited to suppressing human motion only and does not

explicitly handle other dynamic objects in the scene. Extending the attention mechanism to general dynamic object categories (e.g., objects with independent motion) is an important future direction, and could be achieved by incorporating more general motion or object segmentation cues.

Clothing. In some cases, clothing may not be segmented cleanly, resulting in artifacts in the generated scene. Body-level masks can be replaced with more complete human instance masks obtained from video segmentation or jointly learned appearance-aware masking, enabling better coverage of clothing and fine-grained human boundaries.

References

- [1] Yue Chen, Xingyu Chen, Yuxuan Xue, Anpei Chen, Yuliang Xiu, and Pons-Moll Gerard. Human3r: Everyone everywhere all at once. *arXiv preprint arXiv:2510.06219*, 2025. 2
- [2] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote,

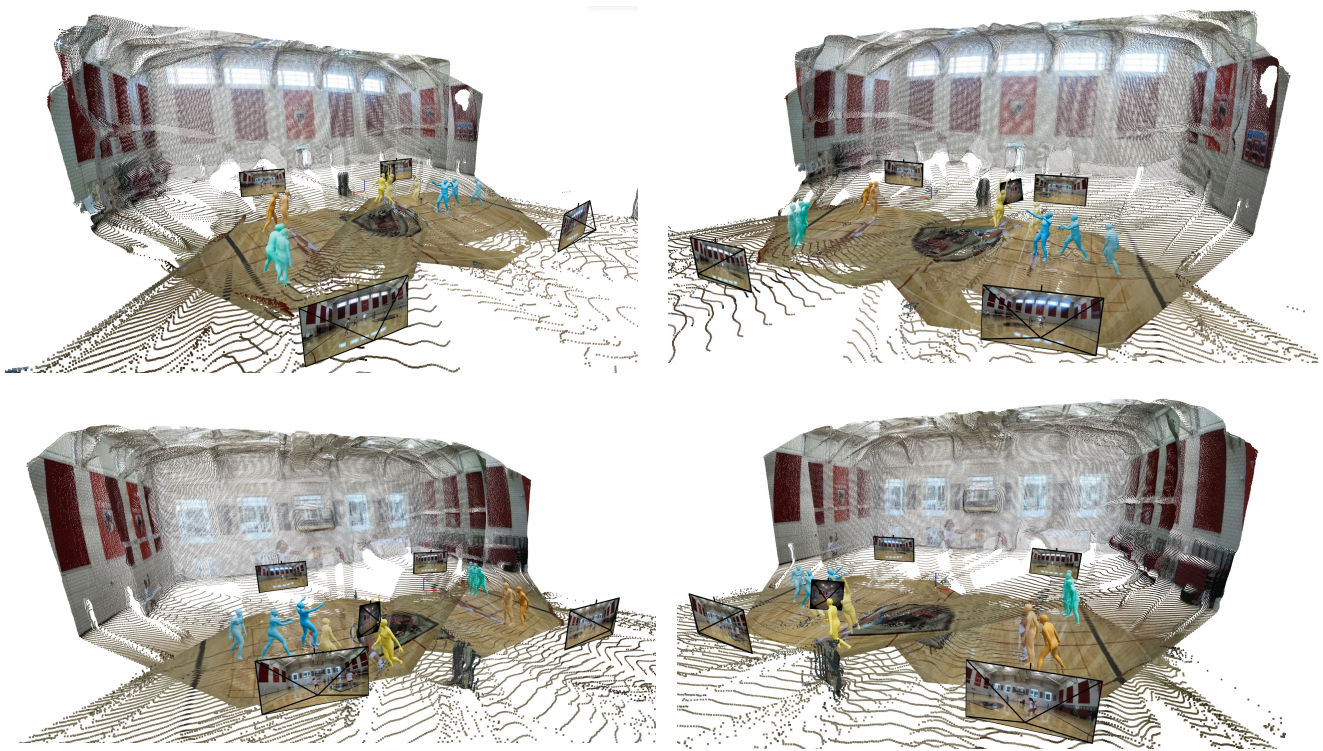


Figure 3. Indoor multi-view reconstructions demonstrating accurate human–scene alignment in large, cluttered environments.

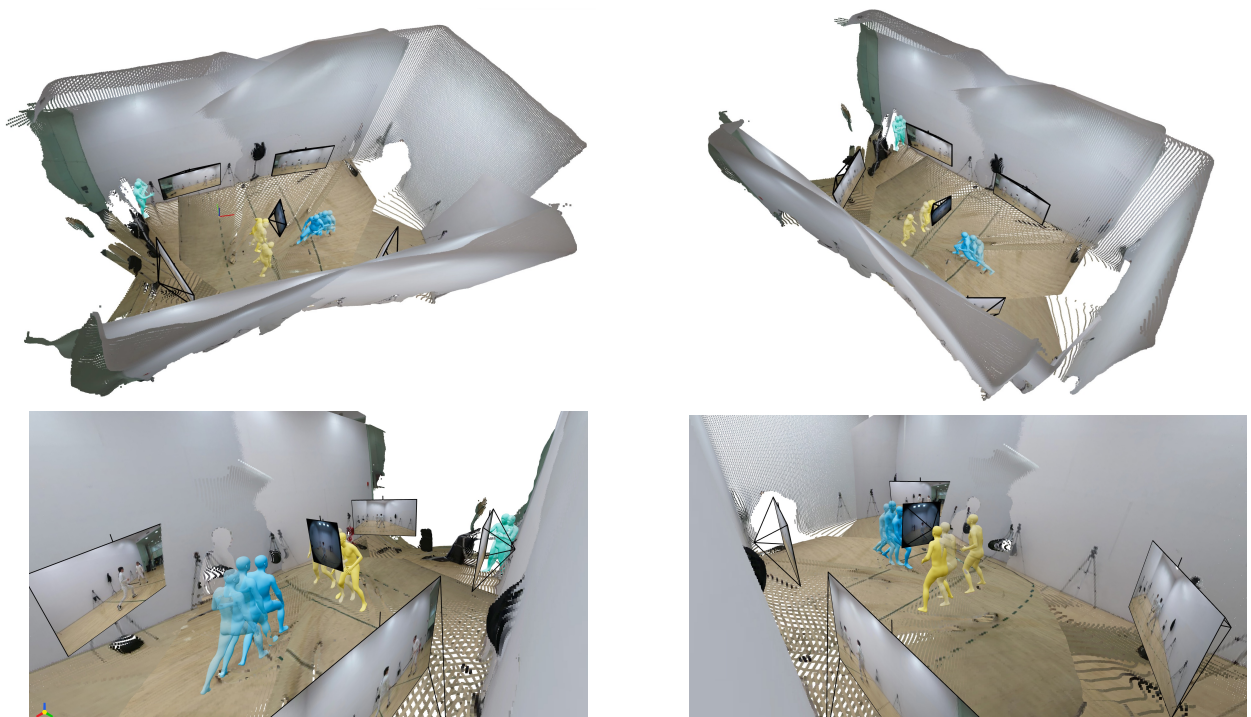


Figure 4. Studio examples illustrating robust multi-view consistency under challenging lighting and viewpoint changes.

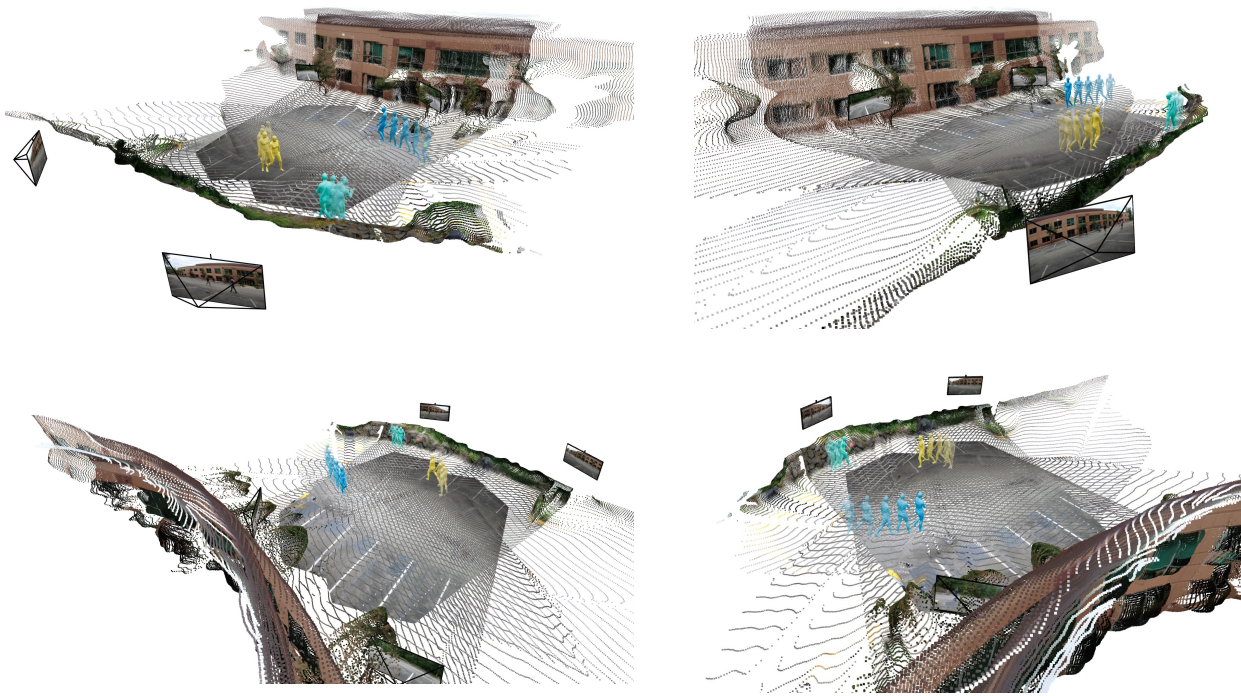


Figure 5. Outdoor scenes highlighting TROPHIES ’s strong generalization in spacious areas and sparse camera setups.



Figure 6. Common failure modes of directly reconstructing humans as raw geometry. (a) Dynamic body motion and complex clothing often lead to fragmented or noisy human surfaces. (b) Articulated limbs may be misclassified as background, causing them to merge into the static scene. These artifacts motivate our use of the SMPL model, which provides a structured human prior and avoids such degeneracies.

et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *CVPR*, 2024. 1

- [3] Manuel Kaufmann, Jie Song, Chen Guo, Kaiyue Shen, Tian-jian Jiang, Chengcheng Tang, Juan José Zárate, and Otmar Hilliges. Emdb: The electromagnetic database of global 3d human pose and shape in the wild. In *ICCV*, 2023. 2
- [4] Rawal Khrodar, Aayush Bansal, Lingni Ma, Richard Newcombe, Minh Vo, and Kris Kitani. Ego-humans: An ego-centric 3d multi-human benchmark. In *ICCV*, 2023. 1, 4

- [5] Jiefeng Li, Ye Yuan, Davis Rempe, Haotian Zhang, Pavlo Molchanov, Cewu Lu, Jan Kautz, and Umar Iqbal. Coin: Control-inpainting diffusion prior for human and camera motion estimation. In *ECCV*, 2024. 2

- [6] Zhizheng Liu, Joe Lin, Wayne Wu, and Bolei Zhou. Joint optimization for 4d human-scene reconstruction in the wild. *arXiv preprint arXiv:2501.02158*, 2025. 2

- [7] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-

- person linear model. *SIGGRAPH Asia*, 2015. [3](#)
- [8] Lea Müller, Hongsuk Choi, Anthony Zhang, Brent Yi, Jitendra Malik, and Angjoo Kanazawa. Reconstructing people, places, and cameras. *arXiv:2412.17806*, 2024. [3](#)
 - [9] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, 2019. [3](#)
 - [10] Davis Rempe and et al. Slahmr: Scale-aware human motion recovery from monocular videos. In *CVPR*, 2023. [2](#)
 - [11] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *SIGGRAPH Asia*, 2017. [3](#)
 - [12] Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. World-grounded human motion recovery via gravity-view coordinates. In *SIGGRAPH Asia*, 2024. [2](#)
 - [13] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J. Black. WHAM: Reconstructing world-grounded humans with accurate 3D motion. In *CVPR*, 2024. [2](#)
 - [14] Yu Sun, Qian Bao, Wu Liu, Tao Mei, and Michael J Black. Trace: 5d temporal regression of avatars with dynamic cameras in 3d environments. In *CVPR*, 2023. [2](#)
 - [15] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *CVPR*, 2025. [3](#)
 - [16] Yufu Wang, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. Tram: Global trajectory and motion of 3d humans from in-the-wild videos. In *ECCV*, 2024. [1](#), [2](#)
 - [17] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *CVPR*, 2023. [1](#)
 - [18] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In *CVPR*, 2022. [2](#)