# Motions as Queries: One-Stage Multi-Person Holistic Human Motion Capture

## Supplementary Material

## A. More Details of Our Proposed Dataset

To address the limited availability of outdoor, multi-person, moving-camera datasets with ground-truth annotations in multi-person motion reconstruction, we constructed our dataset based on BEDLAM [5]'s synthetic dataset generation process. We collected 111 unique outfits and their corresponding textures from BEDLAM, using commercial software to get realistic deformations. The Figure 8 shows more scenes of our proposed dataset, and we also attached many example videos from our dataset, please see in the attached files.

**Motion Selection** To focus on motions with significant translational movement, we selected the SMPL-X [38] motions from BEDLAM based on the magnitude of variation in their translation. Formally,

$$\begin{aligned} D_{\text{total}} &= \sum_{i=2}^{N} \Delta \mathbf{T}_i = \sum_{i=2}^{N} \|\mathbf{T}_i - \mathbf{T}_{i-1}\|_2, \\ k_P &= \lfloor \frac{P}{100} \times M \rfloor, \\ \tau_P &= D_{(k_P)}, \end{aligned} \tag{10}$$

Here, $\mathbf{T}_i$ and $\mathbf{T}_{i-1}$ denote the 3D translation vectors at the $i$-th and $(i-1)$-th frames, respectively. $\Delta \mathbf{T}_i$ represents the frame-to-frame translational difference, computed using the Euclidean norm $\|\cdot\|_2$. The total translational movement across a sequence is given by $D_{\text{total}}$, which aggregates $\Delta \mathbf{T}_i$ over $N$ frames.

$k_P$ is the index corresponding to the top $P\%$ of sequences, calculated as $\lfloor \frac{P}{100} \times M \rfloor$, where $M$ is the total number of sequences. $\tau_P$ denotes the translational movement threshold for the top $P\%$, derived from the sorted $D_{\text{total}}$ values.

Finally, we selected the top 10% of motion sequences, ranked by their translational movement $D_{\text{total}}$, as the motions to include in M3C. These sequences were chosen to ensure significant translational variability, critical for addressing the challenges of outdoor, multi-person, moving-camera scenarios in multi-person motion reconstruction tasks.

**3D Scenes** Although HDRI environments provide more realistic lighting, they cannot perform accurate physical simulations and impose strict limitations on camera perspectives. Therefore, our videos are rendered in photorealistic 3D scenes. To ensure that human figures maintain correct heights on uneven ground, we utilize the ray-casting in Unreal Engine 5 [1] to detect ground elevation. This allows for a rough correction of the figures' heights, making their movements more realistic.

In contrast to existing synthetic datasets, which often avoid placing objects near human figures to prevent unnatural physical collisions, we increase the dataset's complexity and diversity by incorporating scenes with simple objects such as grass and lampposts in the placement areas. Although minor collisions may occasionally occur, this design prioritizes maintaining plausibility in human-scene interactions while enriching the contextual diversity of the environments.

**Comparison of Camera Trajectories** To highlight the richness of camera dynamics in our dataset, we visualized camera trajectories in 3D space in M3C and BEDLAM datasets, as shown in the Figure 5. Notably, the trajectories with higher Z-values in the BEDLAM visualization are a result of coordinate origin discrepancies in the 3D scenes; these trajectories are actually closer to the ground.

From the visualization, it is evident that the M3C dataset exhibits a greater variety and complexity in camera movements compared to BEDLAM. The trajectories in M3C cover a more extensive range of spatial regions, showcasing diverse and dynamic camera motions, including nonlinear paths and variable altitudes. By contrast, the BEDLAM dataset contains trajectories that are more regular and constrained, with fewer variations in motion patterns.

## B. Details of Inference Time Computation

To evaluate the inference time differences across various scenarios, we adopted a 6fps subset of the BEDLAM dataset as the test set. This subset comprises 250 videos, each featuring 3 to 8 individuals. We measured the average inference time per frame for three different methods, grouped by the number of people present in each video. All methods were evaluated on a single NVIDIA 3090 Ti GPU.

As illustrated in Figure 7, the top-down methods, represented by SMPLer-X [7], show a linear increase in inference time with the number of individuals. Compared with top-down methods detecting and processing each person separately, bottom-up methods, represented by Multi-HMR [3], process the entire image holistically, maintaining stable inference time regardless of the number of individuals. Our proposed method further enhances efficiency by processing the video sequence as input, achieving significantly reduced per-frame inference time compared to frame-by-frame approaches.

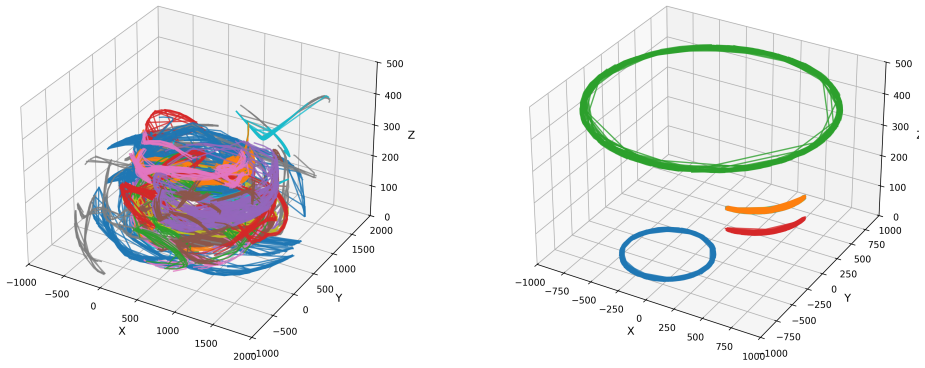


Figure 5. Visualization of camera trajectories in M3C and BEDLAM. Each trajectory is represented using a consistent color within the same subset of each dataset. The axes ($X$, $Y$, $Z$) denote distances from the origin of the 3D scene's human placement region.


Figure 6. More qualitative comparison results on BEDLAM dataset. Our method performs better in occlusion scenarios.

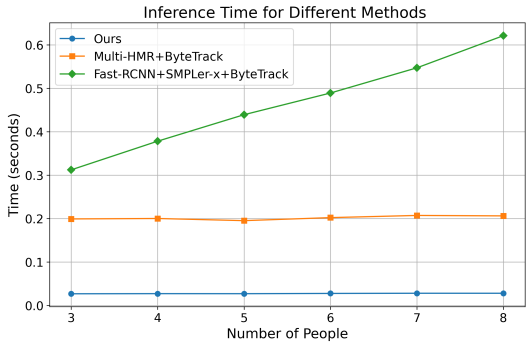


Figure 7. The average inference time per frame of different methods on videos containing varying numbers of people.

## C. More Qualitative Comparisons on the BEDLAM Dataset

We also visualize the results on the BEDLAM dataset. As shown in Figure 6. The advantages of our method can be more clearly seen: the baseline methods are struggling with temporary disappearance of individuals and tend to recognize the same individual with different identities while ours can infer the correct identity information by temporal information. Please see more comparisons in video in the attached files.

## D. More Evaluations

We added combinations of different trackers, detectors, and pose estimators to carry out experiments, as shown in the Tab. 7. The results of TRACE [47] and the body-only 3D

Figure 8. More scenes of our proposed dataset are presented. Our dataset is mainly composed of outdoor scenes with multiple individuals moving in the scene.

| | IDs ↓ | MOTA ↑ | PA-MPJPE ↓ | MPJPE ↓ |
|---|---|---|---|---|
| YOLOX-X + SMPLer-X + BoostTrack++ | 375 | 90.86 | 49.79 | 87.37 |
| Multi-HMR + BoostTrack++ | 466 | 82.66 | 46.45 | 91.15 |
| PHALP (w/ 4D-Human) | 175 | 84.25 | 77.59 | 156.41 |
| TRACE | 516 | 70.26 | 76.36 | 129.98 |
| Ours | **128** | **95.15** | **45.56** | **79.88** |

Table 7. More evaluation on BEDLAM.

human tracker PHALP [39] are also presented.