

# MAMMA: Markerless Accurate Multi-person Motion Acquisition

## Supplementary Material

The Supplementary Video shows a side-by-side comparison of our method vs the ground truth data. In the video we intentionally did not label which result is which. The answer is at the end of this document in Section 17. We ask the reader to view the video before checking the answer. Please refer to <https://mamma.is.tue.mpg.de/> for the Supplementary Video.

## 1. Landmark Networks Training Details

Our network, CameraHMR [2], and LookMa\*[1] are trained for 300K iterations using 4 NVIDIA A100 GPUs, with a per-GPU batch size of 24 and gradient accumulation set to 2 steps. The training time took around 3 days. We use the Adam optimizer with 500 warm-up iterations. For MammaNet and CameraHMR, we initialize the transformer backbone with ViTPose-B weights pretrained at a resolution of  $256 \times 192$ . Following Zhang et al. [4], we interpolate the positional embeddings to increase the effective input resolution to  $512 \times 384$ . Hewitt et al. [1] use HRNet-W48 as the backbone. We follow their configuration, except we initialize the network using pose HRNet weights trained on the COCO dataset. The input resolution is set to  $512 \times 384$ , consistent with the transformer-based models. As the other models do not predict contact, we trained a version of our model that only predicts uncertainty and visibility.

## 2. MammaSyn Dataset

Each rendered sequence contains 2–6 synthetic subjects, placed at random positions within the capture volume. They are randomly assigned one of 100 skin textures which are overlaid with one of 1.7K garment textures.

In total, we render 2.8k sequences (5.5 hours). Each sequence is rendered from 8 views, where the first view is randomly selected and the rest via the Farthest Point Sampling (FPS) algorithm on our 32-camera setup, maximizing spatial coverage and projected-pose diversity. For lighting and background variety, we use a pool

of 95 HDR images. We render at 6 fps,  $2056 \times 1504$ px, roughly twice the resolution of BEDLAM, to ensure much finer detail in the hand regions. In Tab. 1 we show the dataset composition by number of images.

In Fig. 1 and Fig. 2, we show more cropped samples of our dataset, and in Fig. 14 and Fig. 15, we show the contact labels.



Figure 1. Crops from MammaSyn-S (Single person).

Subset	# minutes	Datasets	% of images
MammaSyn-S	86	BEDLAM	9.9%
		MOYO	15.9%
MammaSyn-I	211	Harmony4D	17.7%
		Hi4D	8.0%
		Inter-X	14.6%
		InteractionsCouple (ours)	19.4%
		LatinDance (ours)	3.4%
MammaSyn-H	36	InterHand2.6M	10.0%
		SignAvatars	1.0%

Table 1. Composition of the MammaSyn dataset. Percentages are computed over all MammaSyn images.

## 3. Dataset Evaluation

We evaluate our proposed dataset, MammaSyn, on landmark prediction accuracy, particularly in cases in-



Figure 2. Crops from MammaSyn-I (Interactions).

Table 2. **Evaluation of datasets.** Mean 2D Euclidean distance error (in pixels) between ground truth and predicted landmarks; includes visible and invisible landmarks.

Dataset	RICH	Harmony4D	CHI3D	MammaEval-S	MammaEval-D	MOYO
BEDLAM	8.83	18.33	4.36	6.16	7.70	11.04
BEDLAM*	9.35	19.11	4.40	6.49	7.60	11.92
MammaSyn	8.68	19.09	4.74	6.53	7.71	7.05
MammaSyn +B	<b>8.09</b>	<b>17.34</b>	<b>4.09</b>	<b>5.62</b>	<b>6.66</b>	<b>6.95</b>

volving occlusions caused by interactions. We use the best-performing network, MammaNet +masks SAM2, pre-trained on BEDLAM (B), and train it for another 300K iterations on our dataset. We compared against the network trained only on BEDLAM but with an additional 300K iterations (BEDLAM\*). We also train our network combining our dataset and BEDLAM (ALL+B).

Tab. 2 shows that training longer with BEDLAM does not improve performance (in fact, it degrades slightly). However, when we use our datasets along with BEDLAM, the performance of the network improves, especially on challenging poses such as those in MOYO and the interaction sequences.

#### 4. MammaEval Dataset

We captured our evaluation datasets (MammaEval-Singles, MammaEval-Dance, and MammaEval-Extra) using a Vicon marker-based motion capture system, synchronized with a multi-view RGB camera system (IO Industries (IOI)). The motion capture setup consists of 54 VICON Vantage V16 cameras. The IOI setup includes 16 Victorem and 17 Volucam cameras. Both systems are mounted on a truss structure with the dimensions of 9m by 7m with a height of 5m. Within

that space, the IOI cameras are positioned to cover a  $3\text{m} \times 3\text{m} \times 3\text{m}$  space with the subject’s full body visible in all views (see Fig. 3). Custom 12K Lux lighting by NORKA provides enough light to avoid shadows and motion blur at high capture frame rates. The lights are not constantly powered but are triggered to only turn on together with the IOI system when the shutter is open during each capture frame. This protects the subjects’ vision from being permanently exposed to high-intensity light and prevents overheating while providing sufficient lighting for the captured images. All systems are synchronized using transistor-transistor Logic (TTL), precision time protocol (PTP), and linear time-code (LTC) to ensure accurate calibration and a frame-to-frame correspondence of captured data. A key factor in creating high-accuracy data across both camera systems is our custom calibration procedure. The VICON calibration wand is utilized to capture a calibration sequence that is processed to determine camera parameters in a joint space with a common origin. All sequences are captured at 30 fps for the IOI, with the framerate for VICON varying from 60 fps for MammaEval-Dance and MammaEval-Extra to 120 fps for MammaEval-Single.

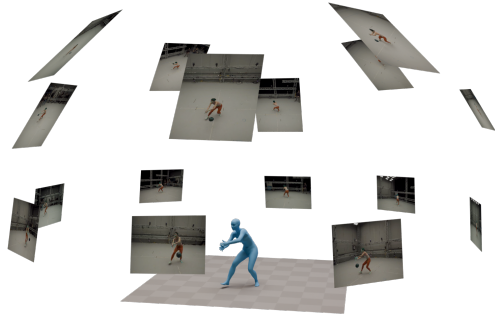
For our evaluation comparing Vicon with MAMMA, we need to know ground truth marker locations on the surface of the SMPL-X body. To that end, we place markers on a subject and then capture their 3D body shape using a 3dMD body scanner. We fit the SMPL-X mesh to this scan to get the true body shape in SMPL-X topology. We then capture the subject moving in the motion capture system and run MoSh++ with the ground 3D shape template. This lets us establish the marker locations on the 3D surface of the body. We use these known locations when evaluating held-out marker error in the comparison section in the main document. Note that we do not use the ground truth body shape when in MAMMA or when running MoSh++ for the comparison with MAMMA. In both cases, the methods estimate the body shape directly from the images or markers, respectively.

#### 5. Comparison with Harmony4D

The closest work to ours is Harmony4D. However, we cannot directly compare with that method since the test set was generated using their own method. Instead, we evaluate the silhouette reprojections against SAM2 [3] mask predictions. For this comparison, we reproject the Harmony4D meshes into each view and use the resulting silhouettes as masks for our network. However, as illustrated in the main paper and Fig. 4, the provided ground-truth meshes are often misaligned or do not perfectly conform to the subject’s body, lead-



(a) MammaEval-Dance



(b) MammaEval-Singles



(c) Sample views from the IOI cameras.

Figure 3. Our MammaEval capture setup.

ing to noisy and imperfect input for our network. We achieve greater mean IoU (mIoU) regardless, 74.28% vs 69.80%. Fig. 4 illustrates that our method overlaps better with the silhouette of the people. We also highlight that Harmony4D uses an off-the-shelf method to obtain the body shapes, meanwhile our network does not have any special initialization or body shape and pose prior. Note that although SAM2 is generally good in extracting masks of the desired person, it sometimes misses parts of the body during video segmentation. This limitation is the main reason why the average IoU is around 70% rather than higher.

## 6. Multiview Matching Evaluation

In the main paper we show that our Multiview algorithm is able to match people across views with a success rate of 100% in the evaluation datasets. The datasets Harmony4D and MammaEval-D have around 20 cameras, whereas CHI3D has only 4.

Here, we evaluate the robustness of our algorithm by reducing the number of cameras from 16 down to 2 in powers of two (16, 8, 4, 2). We used MammaEval-D

dataset. After the evaluation we observe that even with 2 cameras our method is capable to match correctly the people in the scene 100% of the times. This shows that our dense landmarks plus temporal information (from SAM2) provide rich and accurate geometric information that is useful for matching the identity of a person across views. Note that for the single-frame case, our method can fail, as wrong mask predictions or heavy occlusions can confuse the network and lead to incorrect landmark predictions. This is particularly true when the limbs are heavily occluded. In those cases, the network has to “guess” the location of those parts, which can make them not view-consistent.

## 7. Optimization: Number of Cameras

We evaluated the effect of camera count on single and two-person cases, MammaEval-(S)ingles and MammaEval-(D)ance respectively. To ensure uniform spatial coverage, we sample [2, 4, 8, 12, 16] cameras using Farthest Point Sampling (FPS). Fig. 5 shows that for both single and two-person cases, our method achieves strong performance with as few as 4 cameras,



Table 3. Full Benchmark 3D fitting errors (mm). We evaluate the error for the full body, only for the body, and only for the hands.

Model	RICH		Harmony4D		CHI3D		MammaEval-S		MammaEval-D		MOYO	
	MPJPE	PVE	MPJPE	PVE	MPJPE	PVE	MPJPE	PVE	MPJPE	PVE	MPJPE	PVE
SMPLify	96.18	71.42	-	-	67.68	51.79	47.15	35.42	53.92	43.08	62.15	44.68
LookMa*	39.52	30.29	59.37	45.6	46.47	39.36	25.97	23.94	27.98	24.89	60.15	53.82
CameraHMR	25.61	21.36	58.59	42.0	40.8	34.61	15.25	18.43	20.41	21.06	33.75	33.74
MAMMA	22.20	19.76	<b>45.26</b>	<b>34.02</b>	38.01	32.84	12.96	17.18	<b>17.71</b>	19.80	22.95	25.48
MAMMA-C	<b>22.20</b>	<b>19.76</b>	45.35	34.05	<b>37.96</b>	<b>32.82</b>	<b>12.96</b>	<b>17.18</b>	17.73	<b>19.78</b>	<b>22.95</b>	<b>25.48</b>
Only body												
SMPLify	71.34	63.99	-	-	40.05	44.21	31.4	30.81	37.69	38.92	47.81	39.66
LookMa*	37.78	28.89	49.09	42.69	33.48	35.18	25.34	22.7	23.81	23.41	62.02	53.57
CameraHMR	28.18	21.28	46.74	38.67	30.58	30.83	18.27	18.5	19.17	20.43	33.0	33.37
MAMMA	27.32	20.44	<b>40.84</b>	<b>32.58</b>	28.36	29.25	17.61	17.88	18.86	19.95	24.72	25.89
MAMMA-C	<b>27.32</b>	<b>20.44</b>	40.87	32.6	<b>28.35</b>	<b>29.25</b>	<b>17.61</b>	<b>17.88</b>	<b>18.82</b>	<b>19.93</b>	<b>24.72</b>	<b>25.89</b>
Only hands												
SMPLify	112.73	93.83	-	-	86.1	74.64	57.65	49.29	64.73	55.62	71.71	59.80
LookMa*	40.67	34.5	66.23	54.37	55.14	51.95	26.38	27.68	30.76	29.34	58.9	54.56
CameraHMR	23.91	21.61	66.5	52.05	47.62	45.98	13.24	18.21	21.24	22.97	34.24	34.87
MAMMA	18.79	17.72	<b>48.2</b>	<b>38.33</b>	44.44	43.65	9.85	15.06	<b>16.94</b>	<b>19.33</b>	21.77	24.26
MAMMA-C	<b>18.79</b>	<b>17.72</b>	48.34	38.41	<b>44.36</b>	<b>43.6</b>	<b>9.85</b>	<b>15.06</b>	17.01	19.34	<b>21.77</b>	<b>24.26</b>



Figure 4. Mesh reprojection comparison between Harmony4D (red) and ours (blue).

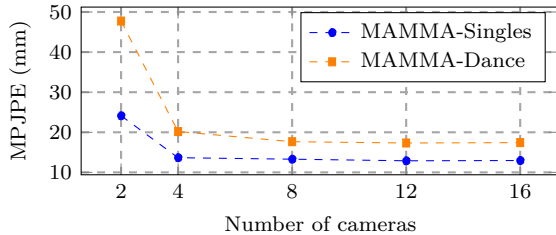


Figure 5. **Camera variation accuracy.**

and reaches optimal accuracy at around 12 cameras after which improvements saturate. The strong performance with as few as 4-8 cameras suggests that our method can be used with low-cost capture setups.

## 8. Optimization: Stages Evaluation

We measured the average runtime of each stage S on MammaEval-D in Fig. 6: S1: global translation and rotation estimation. S2: optimization of pose, shape, and translation. S3: update of uncertainty weights based on reprojection error. S4: incorporation of contact constraints. S2 already provides a good trade-off between accuracy and runtime. S3 and S4 further refine local details and substantially reduce penetrations.

## 9. Perceptual Study

To further evaluate the perceptual quality of our reconstructions, we conducted an Amazon Mechanical Turk



Table 4. IoU comparison between our method and Harmony4D on the Harmony4D test set.

Sequence	Harmony4D	Ours w/ mask
002_hugging	75.72	<b>77.96</b>
025_grappling2	66.26	<b>71.94</b>
028_grappling2	72.26	<b>78.34</b>
030_grappling2	70.26	<b>77.18</b>
031_grappling2	74.35	<b>80.44</b>
032_grappling2	54.79	<b>60.29</b>
033_grappling2	68.47	<b>74.95</b>
034_grappling2	64.16	<b>71.48</b>
035_grappling2	69.42	<b>76.16</b>
036_grappling2	67.82	<b>74.00</b>
037_grappling2	65.71	<b>72.25</b>
038_grappling2	71.93	<b>78.71</b>
039_grappling2	66.44	<b>72.78</b>
040_grappling2	73.41	<b>78.62</b>
041_grappling2	71.59	<b>77.72</b>
042_grappling2	71.62	<b>77.47</b>
043_grappling2	71.94	<b>78.61</b>
009_sword2	68.25	<b>69.19</b>
010_sword2	65.16	<b>66.03</b>
001_sword3	<b>70.70</b>	70.36
002_sword3	71.38	<b>72.52</b>
003_sword3	71.57	<b>72.52</b>
004_sword3	71.39	<b>72.02</b>
005_sword3	71.84	<b>72.30</b>
006_sword3	70.71	<b>71.41</b>
007_ballroom2	67.72	<b>72.11</b>
008_ballroom2	69.51	<b>73.29</b>
009_ballroom2	69.52	<b>72.84</b>
010_ballroom2	69.41	<b>73.11</b>
016_mma4	65.36	<b>76.18</b>
001_mma5	70.69	<b>74.86</b>
002_mma5	72.19	<b>77.27</b>
003_mma5	71.19	<b>75.85</b>
004_mma5	71.86	<b>76.57</b>
005_mma5	71.79	<b>76.77</b>
009_mma5	71.40	<b>75.69</b>
011_mma5	71.40	<b>76.07</b>
013_mma5	72.11	<b>76.09</b>
016_mma5	71.02	<b>74.95</b>
mIoU	69.80	<b>74.28</b>

study in which 33 participants rated the realism of rendered motion from GT and MAMMA-C results using a five-point Likert scale. A one-sided Wilcoxon signed-rank test revealed that our reconstructions were perceived as significantly more realistic than the ground truth on CHI3D ( $p = 2 \times 10^{-6}$ ,  $\Delta = 1.36$ ) and Harmony4D ( $p = 2.4 \times 10^{-4}$ ,  $\Delta = 0.48$ ). For MOYO, the reconstruction also slightly outperformed the GT ( $p \approx 0.0010$ ,  $\Delta = 0.15$ ), although the effect size is

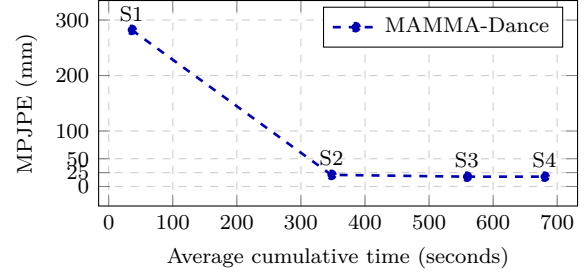


Figure 6. **Optimization (S)tages runtime.**

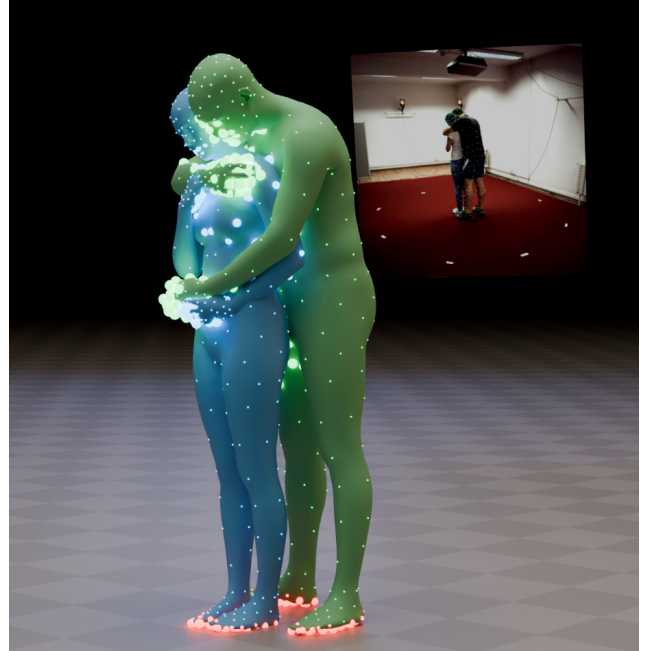


Figure 7. Contact probabilities for the 512 landmarks, averaged over per-view predictions. Brighter and larger points indicate higher contact probability.

smaller given that both versions are already highly realistic (means  $\approx 4.2$ – $4.4$ ). In contrast, no significant advantage was observed for RICH, MammaEval-D, or MammaEval-S, where GT and MAMMA-C results received similarly high ratings (all  $p > 0.3$ ,  $|\Delta| \leq 0.11$ ).

## 10. Visualizing the Contact Probabilities

During optimization, we use contact probabilities averaged over all views. Figure 7 shows a visual example of these averaged values.

## 11. Capture Protocol

Participants in the MammaEval-Single and MammaEval-Extra were selected to have a variety of body shapes and ethnicities, and gender balance. The MammaEval-Dance subjects were chosen based on their proficiency as dancers. All subjects were

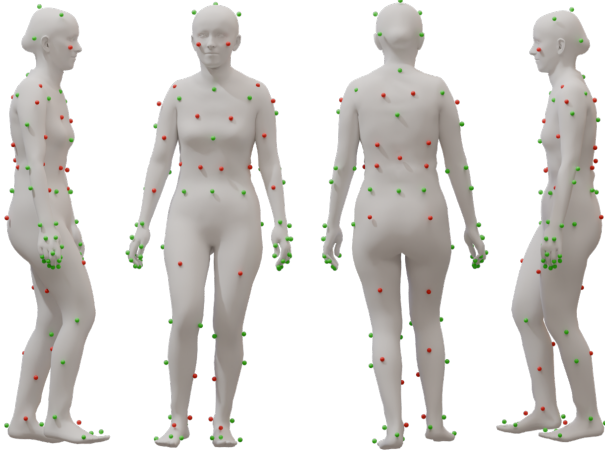


Figure 8. To evaluate MAMMA predictions on the Vicor ground truth, we use the additional held-out 37-marker layout (shown in red) alongside the default 73 Front-Waist10Fingers marker set from Vicor (green).

informed about the details of the capture protocol and the use of their data in advance and gave written informed consent.

The capture procedure and data processing steps have been approved by an ethics committee. All data collection, storage, and processing activities are compliant with privacy regulations. Participants wore tight-fitted clothing in distinctive colors to enable good quality body scanner data and the processing of the multi-view captures. In the body scanner, each subject was recorded performing several sequences and still poses.

For all captures, the *FrontWaist10Fingers* marker set from Vicor was used with a total of 73 markers distributed on the body and hands of the participant. In line with the Vicor data capture process, subjects were calibrated by performing a range-of-motion sequence.

For the MammaEval-Extra, the additional 37 markers were attached after subject calibration to enable both the standard Vicor post-processing and the tracking of added markers that were labelled manually. The positions of the 37 extra markers were selected to be distinct from the standard marker positions, to avoid occlusion, and to include soft-tissue areas that are excluded from standard motion capture templates (Fig. 8). Reference images of the subjects after marker placement were captured to allow for an accurate replacement of fallen markers and support further processing steps.

For the Vicor Marker-based comparison experiment, we excluded one subject due to calibration errors of the Vicor-IOI system and discarded the first 5 frames to account for trigger-light delay.

### 11.1. Computing marker positions

We use MoSh++ to create the GT SMPL-X fits of our MammaEval datasets and to compute the marker positions of the methods used in our Vicor Markers comparison experiment. MoSh++ requires an initialization step that assigns each motion-capture marker to a vertex on the SMPL-X body model. In its first optimization stage, the marker positions are allowed to vary slightly to better align with the input data. While standard marker layouts often come with predefined marker-to-vertex mappings, real-world usage introduces significant variability. Due to human error and the diverse range of body shapes and clothing conditions, the actual marker placements for the same layout can differ substantially across captures, sometimes by several tens of centimeters. MoSh++ is highly sensitive to initialization quality, and an accurate marker-to-vertex mapping is critical for achieving optimal results. To ensure reliable initialization, we capture 3D scans of subjects wearing markers, fit the SMPL-X model to the scans, and manually select the most appropriate vertex for each marker.

During the first optimization stage, MoSh++ refines marker positions using a local coordinate system defined as follows: (a) the closest SMPL-X vertex to the marker serves as the origin; (b) the coordinate axes are constructed from two edges of the triangle incident to this vertex, along with their cross product to form an orthonormal basis. The final marker position is stored as a combination of the vertex index and three displacement values, each corresponding to an offset along one of the local coordinate axes. The origin and local coordinate axes obtained during this optimization stage enable us to regress marker positions for any SMPL-X body shape.

### 11.2. MammaEval-Extra Protocol

The process described in Section 11.1, allows us to compute the 37 additional held-out Vicor marker positions from SMPL-X vertices (Fig. 8). This enables us to quantitatively compare markerless to marker-based methods. As a baseline, we use MoSh++, which takes as input only the 73 markers (*FrontWaist10Fingers* marker set). In Tab. 5, we report the mean per-marker distance error of MoSh++ (baseline) and MAMMA, per-sequence, with both methods evaluated without using a GT body shape (i.e. without using the 3D scan).

Similarly, we additionally tested our protocol on the MammaEval-Dance sequences, but this time we held-out 9 markers (from the 73 *FrontWaist10Fingers* marker set) distributed across the body and ran the same protocol. As a baseline, we use MoSh++, which takes as input the remaining 64 markers. The average

Table 5. Vicon Markers comparison experiment: Mean per-marker distance (mm) of MoSh and MAMMA on the Vicon held-out 37 markers of our MammaEval-Extra dataset.

Subject	Action	MoSh	MAMMA
00202	solo_dancing	24.368	25.464
00202	calib_routine	20.081	21.259
00202	walking	23.384	25.841
00202	warmup	21.888	24.108
00219	solo_dancing	23.967	22.889
00219	calib_routine	23.338	22.670
00219	walking	23.507	24.503
00219	warmup	23.898	21.864
00236	solo_dancing	19.864	21.876
00236	calib_routine	17.832	19.050
00236	walking	18.791	21.002
00236	warmup	18.514	19.248
		21.619	22.481

error of our MAMMA is 27.590mm and 26.150mm for MoSh++ (baseline) with a difference of 1.44mm.

## 12. Usability and Cost

In addition to accuracy, the cost of setup and processing is important. We compare the time cost of using our automatic method vs. the marker-based pipeline which requires manually cleaning and labelling data. To estimate the cost of marker labelling, we recorded the performance of 3 marker post-processing technicians who had prior experience cleaning mocap frames, with experience ranging from 118,972 to 149,454 frames, with an average of 134,500 frames. Each technician cleaned 3 mocap sequences of 2 people dancing (West Coast Swing), while timing and documenting their workflow. Using the Vicon Shogun Post tool, the data was inspected frame-by-frame and all missing markers (gaps) and marker swaps were manually corrected. They cleaned 9 sequences with a total of 29,244 frames and a capture duration of 24 minutes and 31 seconds. Cleaning required 46 hours and 51 minutes. The most time-intensive step of the pipeline was fixing the swapped markers located on the fingers, with an average time of 4.95 hours per sequence. While marker swaps on body locations are typically less problematic than those on the fingers, the time to fix them took an average of 15 minutes. Gap filling was performed in Shogun Post and, while manual, the time is minimal compared with the labeling.

After cleaning we ran MoSh++ on the result, which took approximately 25 hours of computation. Ignoring the time to put on and take off markers, the time from capture to SMPL-X fits was approximately 72 hours. In contrast, running MAMMA takes approximately 26

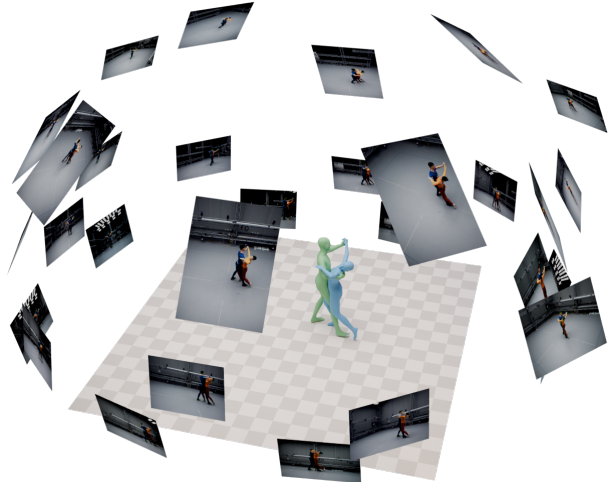


Figure 9. Multi-view setup of the MAMMA Dance sequences captured solely with MAMMA.

hours on a single GTX-4090. SAM2 runs at 2 fps, MammaNet at 12 fps and the optimization on average takes 65 secs. for a sequence of 100 frames.

## 13. MAMMA Dance Release

In addition to MammaSyn, MammaEval-Singles, MammaEval-Dance and MammaEval-Extra, we release a new dataset captured using our MAMMA pipeline. The dataset includes Bachata, West Coast Swing, Breakdance and Ballroom, performed by 9 subjects for approximately 1 hour in total. We capture at 30 fps, from 32 synchronized cameras, using the multi-view setup shown in Fig. 9. For the released dataset we used the masks predicted by SAM3.

## 14. More than Two People

We evaluate the performance of our method in scenes involving more than two individuals following the Dance Release configuration. In total, we record 48 sequences: 15, 12, 10, and 11 scenes containing 3, 4, 5, and 6 interacting subjects, respectively; see Fig. 10. The subjects have a diversity in height, body shape, and gender, and wear a wide range of clothing, see Fig. 11.

Among these sequences, 30 exhibit visually accurate results, with physically plausible joint motions throughout the entire sequence. An additional 12 sequences remain largely plausible, with minor artifacts such as occasional joint jitter in a small number of frames. The remaining 6 sequences display unrealistic poses throughout the sequence. These failures are more frequent in scenes with more people, including 3 sequences with 6 subjects and 1 sequence in each of the other groups. We will release the motions we captured.



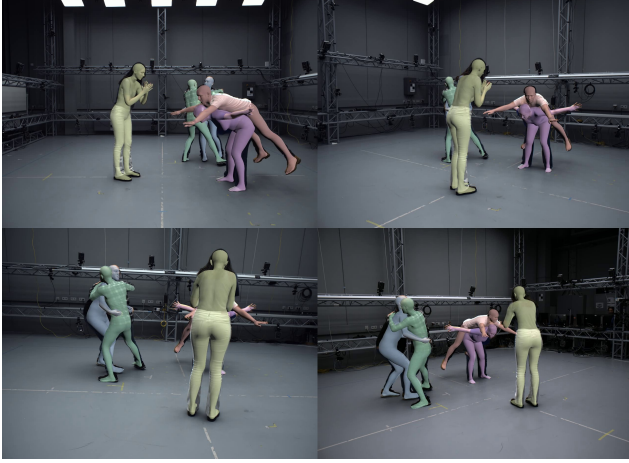


Figure 10. Motion captured by MAMMA with five people interacting in the scene.



Figure 11. Sample showing the diversity in pose, shape and clothes of the six people captured by MAMMA.

We observe that failure cases are primarily caused by inaccurate limb predictions, typically due to severe occlusions or errors in the segmentation masks.

## 15. Consumer-based Cameras

We use four iPhone 17 Pro Max devices synchronized via Blackmagic Genlock hardware. Specifically, each device is connected to a Blackmagic Camera Pro Dock to receive LTC timecode and Genlock signals from an external Ambient LockIt generator. On-device control is handled using the Blackmagic Camera app, enabling one master device to coordinate three slave devices. During recording at 60 fps, we observe occasional synchronization deviations of up to two frames. While our setup relies on synchronization hardware, it can be replaced by simpler alternatives, such as aligning video streams using audio cues (e.g., a clap at the start of recording). Camera calibration is performed using

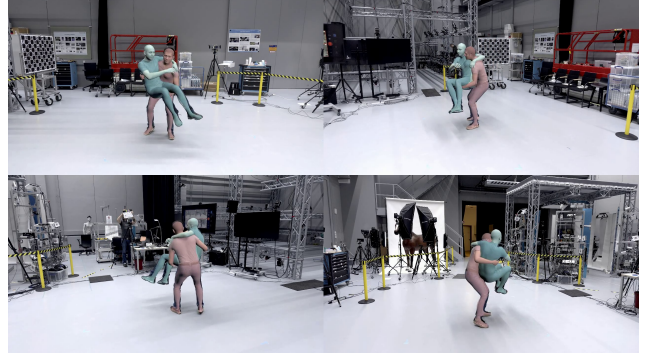


Figure 12. Iphone indoor capture with recovered mesh overlaid on the image.



Figure 13. Iphone outdoor capture with recovered mesh overlaid on the image.

a ChArUco (chessboard-ArUco) board and the software provided by [calib.io](https://calib.io). We also tested the OpenCV calibration tool and the calibration difference between both tools is minimal. This setup demonstrates that a practical, low-cost, and portable capture system can be realized with our technology.

We record 18 indoor (Fig. 12) and 28 outdoor (Fig. 13) sequences. The indoor subset contains 13 single-person and 8 two-person sequences, while the outdoor subset includes 19 single-person and 9 two-person sequences. The captured actions range from simple walking and running to object interactions, as well as light close-range interactions between subjects. All recorded sequences exhibit plausible human poses and temporally smooth motions. However, 5 sequences show occasional joint flickering in a small number of frames. Consistent with the previous section, these artifacts are primarily caused by inaccurate limb predictions under occlusion. We will release these motions as well.

## 16. Limitations and Future Work

Our system recovers motion sequences of subjects interacting with each other from multiview cameras. Despite its competitive performance, MAMMA still has



Figure 14. Computed floor contact from geometric cues, used in MammaSyn. We show the pose from a side-view and an under-view. SMPL-X vertices that are in contact are shown in red.

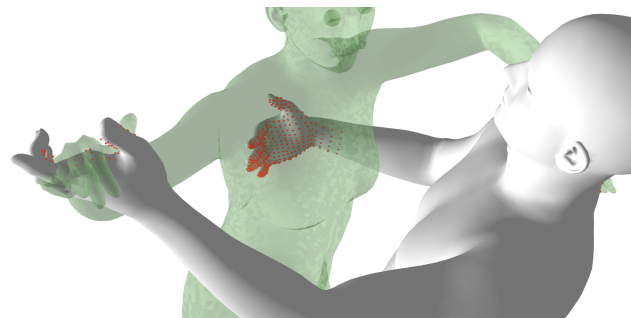


Figure 15. Computed contact between two subjects from geometric cues, used in MammaSyn. SMPL-X vertices of gray subject that are in contact with the green transparent subject are shown in red.

limitations. The predicted landmark uncertainties and visibility probabilities allow the optimizer to reduce the influence of low-confidence or occluded landmarks. As a result, the optimizer may ignore landmarks that are heavily occluded and uncertain across views, leading to flickering artifacts or over-smoothing, especially when smoothing regularization is heavily weighted.

The contact probability prediction of our network is conservative, the highest probability is around 60%. This is due to the single view ambiguity. If two people are interacting, and one occludes the other person, there is a chance that they are really close but not in contact. One fix to make the network more confident is weighting the loss for parts that are in contact, however, this comes at the cost of failing in the previous example.

Foot contact prediction is generally good. However, our network often predicts the foot landmarks based on the height of the shoes. If we then penalize floor contact errors too strongly during optimization, the optimizer may incorrectly pull the person downward to satisfy the contact constraint. In other situations, when the body is already touching the floor, the dense landmarks alone provide sufficient cues, so the addi-

tional contact term has little effect. Therefore, we recommend using the floor contact signal primarily for single-view cases or as a post-processing step to correct foot motion near the floor.

The accuracy of hand-motion recovery can still be improved. It is worth noting that most marker-based captures ignore the hands completely since they are too costly to capture and clean. We believe that image-based methods can be further improved with better training data.

We plan to extend the network to predict landmarks across multiple views jointly to improve inter-view consistency. Similarly, temporal modeling would enhance landmark stability over time. Another potential direction is the integration of diffusion-based human motion priors to refine predictions.

## 17. Supplementary Video Reveal

The videos on the left are the ground truth and MAMMA predictions are on the right.

## References

- [1] Charlie Hewitt, Fatemeh Saleh, Sadegh Aliakbarian, Lohit Petikam, Shideh Rezaeifar, Louis Florentin, Zafirah Hosenie, Thomas J. Cashman, Julien Valentin, Darren Cosker, and Tadas Baltrusaitis. Look Ma, no markers: Holistic performance capture without the hassle. *ACM TOG*, 43(6), 2024. [1](#)
- [2] Priyanka Patel and Michael J. Black. CameraHMR: Aligning people with perspective. Los Alamitos, CA, USA, 2025. IEEE Computer Society. [1](#)
- [3] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollar, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *ICLR*, Red Hook, NY, USA, 2025. Curran Associates, Inc. [2](#)
- [4] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In *ICLR*, Red Hook, NY, USA, 2023. Curran Associates, Inc. [1](#)