Appendix

A. Experimental Details

Model Details We used VIBE [28] for our initial 3D estimate and OpenPose [5] for our 2D psuedo-groundtruth. NeMo used an architecture of 3 hidden layer MLP with 1000 hidden units. The phase networks consists of 100 sigmoids nodes. The instance code dimension was 5. In addition to 2D reprojection loss, during the optimization a regularization w.r.t. the VPoser and GMM pose prior were also used similar to SMPLify [36]. We ran 300 steps using the 3D loss as warmup and 2000 steps using the 2D loss in the second stage. The optimizer used for our optimization was Adam with a learning rate of 0.0001 using the default second order hyperparameters in PyTorch.

"Dynamic Range" Dynamic range is defined as the segment of motion where the maximum joint velocity in 3D is above 2 m/s. A contiguous segment is selected based on the first and last frame of the video that satisfied this criterion.

Metrics For the standard MPJPE and MPVPE, the evaluation is done in a root-centered fashion at the frame-level, meaning that the root translation and orientation were aligned with the groundtruth at every frame. For global MPJPE and MPVPE, since each prediction resides in a different frame-of-reference, results were first aligned using rigid-body transformation (i.e., translation and orientation) with the groundtruth using vertices of the entire sequences. Unlike Procrustes alignment which is sometimes used for HMR studies, we did not perform "scaling".

Penn Action Dataset We annotated each action with a "left-handed" or "right-handed" label, and only put action with the same handedness in the same batch. In the following experiments, for each action we sampled 40 batches of 3 sequences randomly from the training set of each action label. For each sequence, we uniformly sampled 50 frames from the beginning to the end of action. The 2D annotations in the Penn Action dataset is noisy. The joints are sometimes mislabelled. To alleviate this issue, we run OpenPose [5] on the videos. If the OpenPose prediction of a joint and the groundtruth label are more further than a threshold, we drop that keypoint in our optimization. The threshold is set to 10% of the image dimension.

B. Additional Results

MoCap	Method	Baseball Pitch	Baseball Swing	Tennis Serve	Tennis Swing	Golf Swing	Mean
Recon. Err. (↓)	OpenPose [5]	32.74	25.7	50.5	25.46	37.54	34.39
	VIBE [28]	18.17	15.28	20.08	13.88	**14.5**	16.38
	VIBE+SMPLify [28]	18.09	15.46	26.61	13.89	15.04	17.82
	PARE [29]	16.06	15.19	**16.26**	**13.16**	15.45	15.23
	NeMo (Ours)	**15.7**	**14.67**	16.48	13.81	15.12	**15.16**
PCK (↑)	OpenPose [5]	95.77	97.62	94.26	98.12	96.15	96.38
	VIBE [28]	97.64	98.51	96.96	99.55	**99.34**	98.4
	VIBE+SMPLify [28]	97.65	98.56	95.93	99.51	99.25	98.18
	PARE [29]	**99.33**	98.92	**99.23**	99.71	98.51	99.14
	NeMo (Ours)	98.46	**99.61**	98.81	**99.88**	99.16	**99.18**

Table 5. **2D evaluation of our MoCap dataset.**

"2D evaluation on our MoCap dataset." Table 5 shows 2D evaluation on our MoCap dataset. While NeMo still performed the best overall, the trend is not as clear as in using 3D metrics (Table 1). This is because many incorrect 3D poses can be reprojected to the same 2D joint locations. First, it is worth noting that NeMo improves in terms of 3D evaluation while still outperforming baselines in terms of 2D metrics overall. Second, the discrepancy between the 2D and 3D evaluations speaks to the importance of using 3D evaluation for quantitative results, and also checking results visually.

“Number of videos required.” In table 6, we show the average MPJPE on our MoCap dataset with a varied number of videos. NeMo’s accuracy increases with more videos and also works with just one video. NeMo improves over VIBE when N $>$ 1. ‘VS’ stands for VIBE+SMPLify.

N=1	2	3	4	8	VIBE	VS
91.6	86.4	80.1	79.5	75.2	87.4	96.7

Table 6. **Number of videos required.**