Appendix

We make the code available on https://github.com/vita-epfl/MotionMap.

A. Algorithm: Motion Transfer

We use 'motion transfer' to ensure that different labels y have the same skeletal size as that of the input x. This is done to ensure that multimodal ground truths are selected purely on the motion, and not on the size of the person. We also follow this reasoning and use motion transfer during the training and evaluation process.

Algorithm 1: ***Motion Transfer***

```
/* Transfers the motion from pose sequence
   y to skeleton of sequence x                */
// Pose sequence for skeletal reference
Input: x ∈ ℝ^(#frames_x × #joints × 3)
// Pose sequence for motion reference
Input: y ∈ ℝ^(#frames_y × #joints × 3)
// Pose sequence with skeletal size from x
   and motion from y
Output: z ∈ ℝ^(#frames_y × #joints × 3)

// Get last frame
pose = x[-1]
// Get length for each link in pose
ρ, _, _ = cartesian_to_spherical(pose)
// Get motion (angles) for each link in
   pose
_, θ, ϕ = cartesian_to_spherical(y)
// Reconstruct new pose sequence
z = spherical_to_coordinates(ρ, θ, ϕ)
return z
```

B. Implementation Details

We base most of our architecture on those proposed in [1]. Our encoders $\mathcal{E}_X$ and $\mathcal{E}_Y$ are based on gated recurrent units (GRU) with a dimensionality of 128. Our pose forecaster $\mathcal{D}$ is the exact same design as BeLFusion. The major difference is that we predict a concatenation of the input and output sequence. The uncertainty module is a simple multilayer perceptron (MLP) that predicts the uncertainty per joint per time frame. The heatmap model uses a combination of the GRU encoder, and a one layer MLP, and gives 1×1 convolutional layers. The GRU encoder spatio-temporally encodes the last three frames of the incoming pose sequence, which are mapped to the size of the flattened heatmap by the MLP. After reshaping the output of the MLP to match that of the heatmap, we pass this to the convolution layers to get our raw heatmap. The final heatmap is obtained by capping this output with a sigmoid layer. We use OpenTSNE's implementation of t-SNE [37] which also implements the transform function, a feature missing in the original t-SNE variants. Finally, the codebook can be implemented as a tensor or as a dictionary, since the codebook serves as a lookup table where the queries (or keys) are locations on the heatmap of type integer.

C. Dataset: Details

Human3.6M [39] consists of motion-captured poses of seven publicly available subjects performing 15 actions. We follow the protocol proposed by [1]. The first five publicly available subjects (S1, S5, S6, S7, S8) of the dataset are used for training, and the last two (S9, S11) for testing. The dataset consists of 32 keypoints in total, from which 17 are selected. We zero-center them around the pelvis joint, and thus the the remaining 16 joints are forecasted with respect to the pelvis. Videos of this dataset have been recorded at 50 fps, and we take 0.5 seconds (25 frames) as input and forecast the next 2 seconds (100 frames).

AMASS [38] is a collection of various datasets containing 3D human poses. Following [1], we utilize 11 sets (406 subjects) from this collection for training and 7 sets (54 subjects) for testing. The dataset contains videos at 60 fps after downsampling. We use 0.5 seconds (30 frames) as observation and forecast the next 2 seconds (120 frames). We also downsampled the input data of AMASS by increasing the stride to reduce the training time.

D. Additional Quantitative Results

We report our results by restricting the multimodal ground truth to the testing split only. We observe that across both datasets the quantitative results are similar across different methods. While DivSamp is highly diverse, this does not necessarily translate to accurately predicting possible futures. A major observation is that while MotionMap is much more effective in recalling transitions from the test set (Table 1), this does not come at the cost of general performance, as evident by these results (2. 3). Finally, we note that restricting the multimodal ground truth to the testing split limits the diversity of modes in the ground truth. In Figure 14, we demonstrate that the AMASS testing dataset does not adequately represent the training data, with the testing multimodal ground truth missing the majority of modes. Assuming that the test split contains only five samples, each test sample would have between one and five multimodal ground truths. Furthermore, a discrepancy in the distributions of the train and test split means that the multimodal ground truths for the test set share no commonalities with the train set.

E. Additional Qualitative Results

We have provided some examples of generated future forecasts in the format of GIFs which are included in the supplementary materials in a folder called: **GIFs**. In the aforemen-

Table 2. Human3.6M dataset: All baselines are limited to 5 forecasts. Our method, unconstrained by the number of modes, is adjusted to produce an equal number of predictions. Metrics are reported in meters.

Method	Diversity (↓)	ADE (↓)	FDE (↓)	MMADE (↓)	MMFDE (↓)
Zero-Velocity	0.000	0.597	0.884	0.616	0.884
TPK [40]	6.727	0.568	0.757	0.582	0.756
DLow [3]	11.687	0.602	0.818	0.616	0.818
GSPS [41]	14.729	0.584	0.791	0.602	0.791
DivSamp [4]	15.571	0.545	0.782	0.574	0.787
BeLFusion [1]	7.323	0.472	0.656	0.497	0.661
CoMusion [2]	7.624	0.460	0.678	0.505	0.687
MotionMap	8.190	0.491	0.642	0.505	0.643

Table 3. AMASS dataset: All baselines are limited to 6 forecasts. Our method, unconstrained by the number of modes, is adjusted to produce an equal number of predictions. Metrics are reported in meters.

Method	Diversity (↓)	ADE (↓)	FDE (↓)	MMADE (↓)	MMFDE (↓)
Zero-Velocity	0.000	0.755	0.992	0.776	0.998
TPK [40]	9.284	0.762	0.867	0.763	0.864
DLow [3]	13.192	0.739	0.842	0.733	0.846
GSPS [41]	12.472	0.736	0.872	0.741	0.871
DivSamp [4]	24.723	0.795	0.926	0.801	0.928
BeLFusion [1]	9.643	0.620	0.751	0.632	0.751
CoMusion [2]	10.854	0.601	0.768	0.629	0.797
MotionMap	9.483	0.624	0.729	0.643	0.736

tioned visualization, the color blue refers to the input pose sequence, and red to the corresponding future.

E.1. Controllability

Our method enables control over the selection of modes. With the predicted MotionMap and its identified local maxima, we can focus solely on the most probable futures (Figure 9) or, if needed, select a less likely future (using metadata) as required by the application's requirements (Figure 10). To better show this possibility we have provided a demo.

E.2. Uncertainty

We have illustrated the predicted uncertainty plots for all the future predicted poses and the reconstructed past in Figure 13. It is observable that the model is more certain about reconstructing the past since it is encoded as the input. The various trends in uncertainty demonstrate the dependency of the predicted uncertainty on the motion. Furthermore, joints that have greater movement or are further from the pelvis experience higher levels of uncertainty.

E.3. Sampling Comparison

We compare the predicted MotionMaps with the ground truth heatmaps in Figure 11. MotionMap is encouraged to predict a higher number of modes than present in the ground truth to identify rare transitions. Our visualizations confirms that MotionMap identifies other transitions while not missing out on the original ground truth motions. These miscellaneous transitions are learned by the MotionMap model from trends across the dataset.

How well can state-of-the-art baselines predict multimodality without explicitly encoding multimodal transitions? To study this, we collected predictions for each of the baselines for each input pose seqeuence. We then encode these pose forecasts into two dimensions as described in Section 4.2.1. Next, we overlay them on the ground truth heatmap to identify the differences in the predictions and the ground truth. We observe that baselines that rely on anchors although diverse predict transitions which are unlikely for the given pose sequence. This also tallies with our quantitative evaluation. While this effect is reduced for diffusion based baselines, the methods are less diverse and do not capture rare modes. In contrast, MotionMap captures both common and rare mode since they are encoded in the form of local maxima.

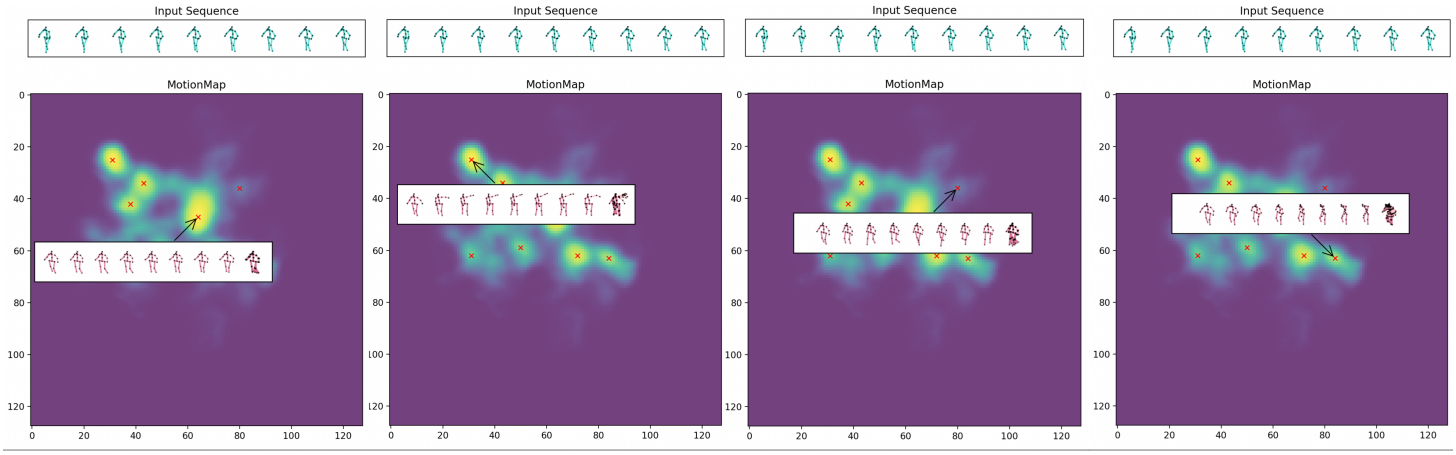

Figure 9. We visualize the modes (in red crosses) predicted by MotionMap. By hovering over the demo tool, we can view the decoded future poses corresponding to the given input pose sequence. We have uniformly selected eight frames in each sequence to demonstrate the motion and **stacked them on top of each other at the end** (the frame on the very right of each visualized sequence) to represent the amount of motion in each sequence.

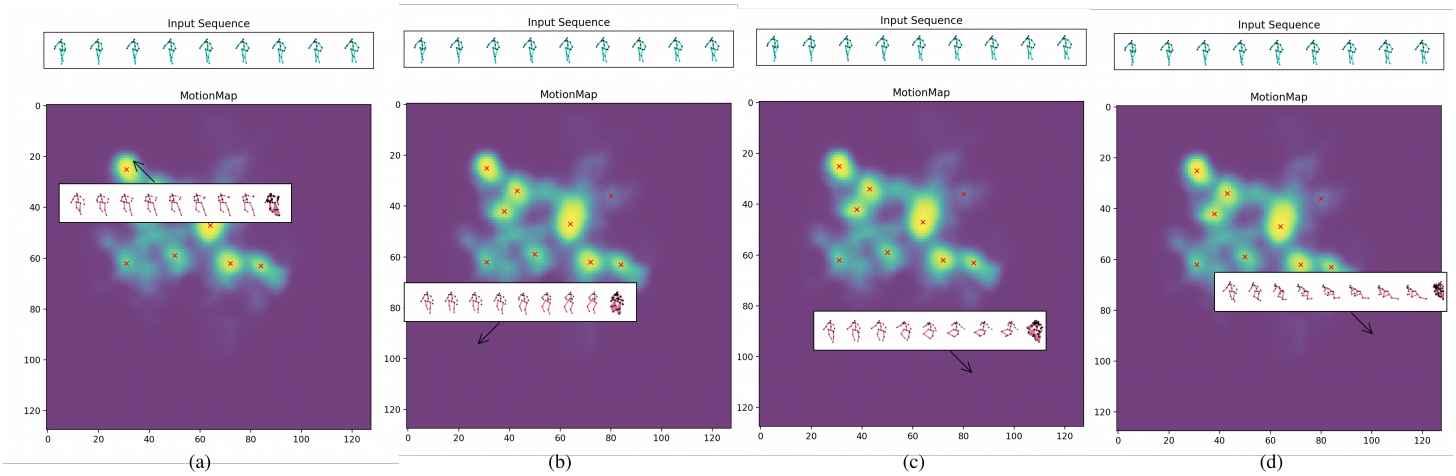

Figure 10. We show different strategies for controlled selection of *non-maxima* forecasts: (a) Selecting samples in the vicinity of a model selected mode. (b,c,d) Based on the distribution of action labels. For instance, we could generate futures for rarer transitions such as sitting down (b) on a chair (c) on the floor, or (d) lying on the floor.

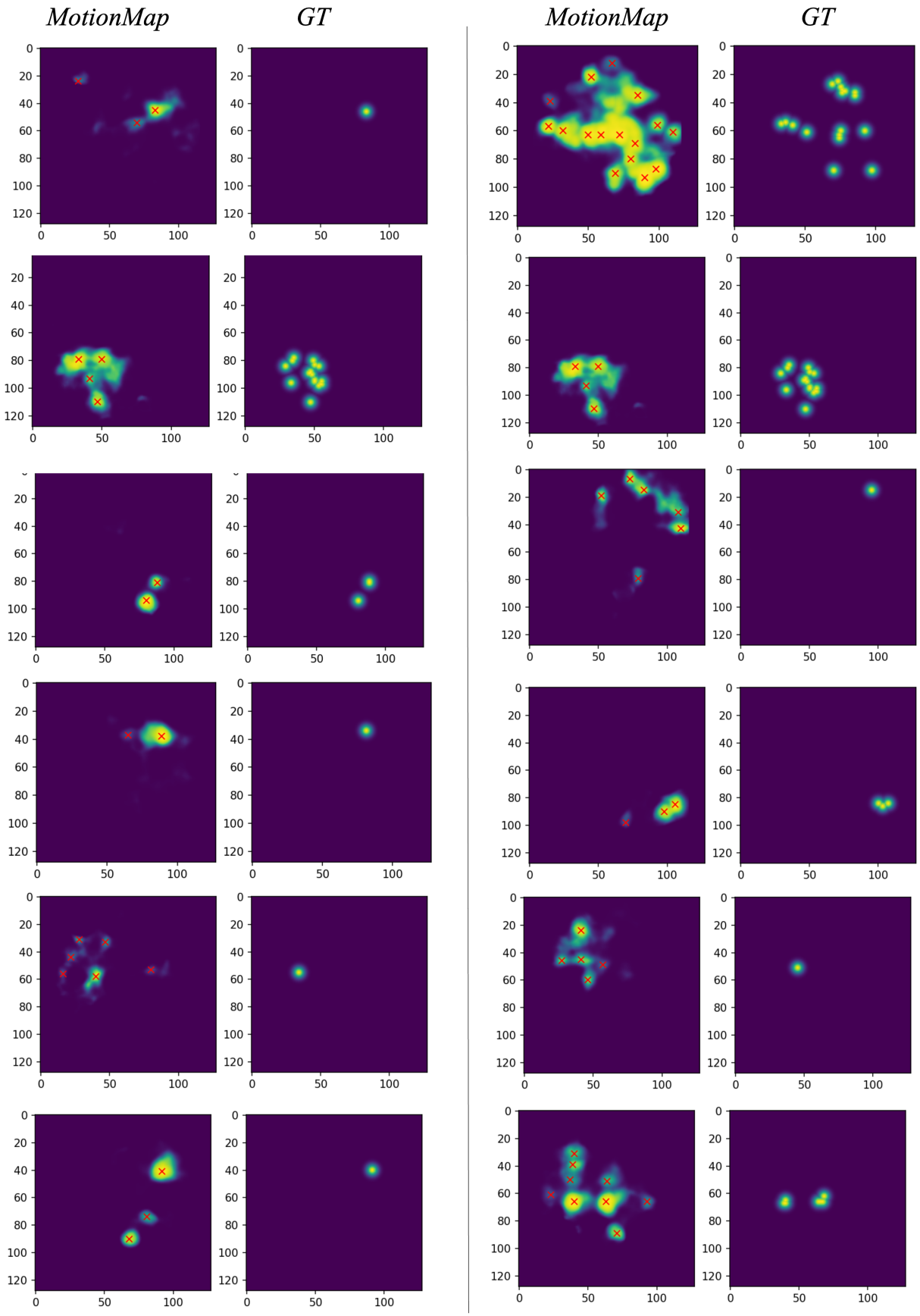

Figure 11. Qualitative comparison between MotionMap and the ground truth multimodal heatmap. Our observations indicate that MotionMap effectively captures the diversity of the modeled scenarios. The presence of a larger number of peaks in MotionMap is a result of learning generalized behaviour across the training split.

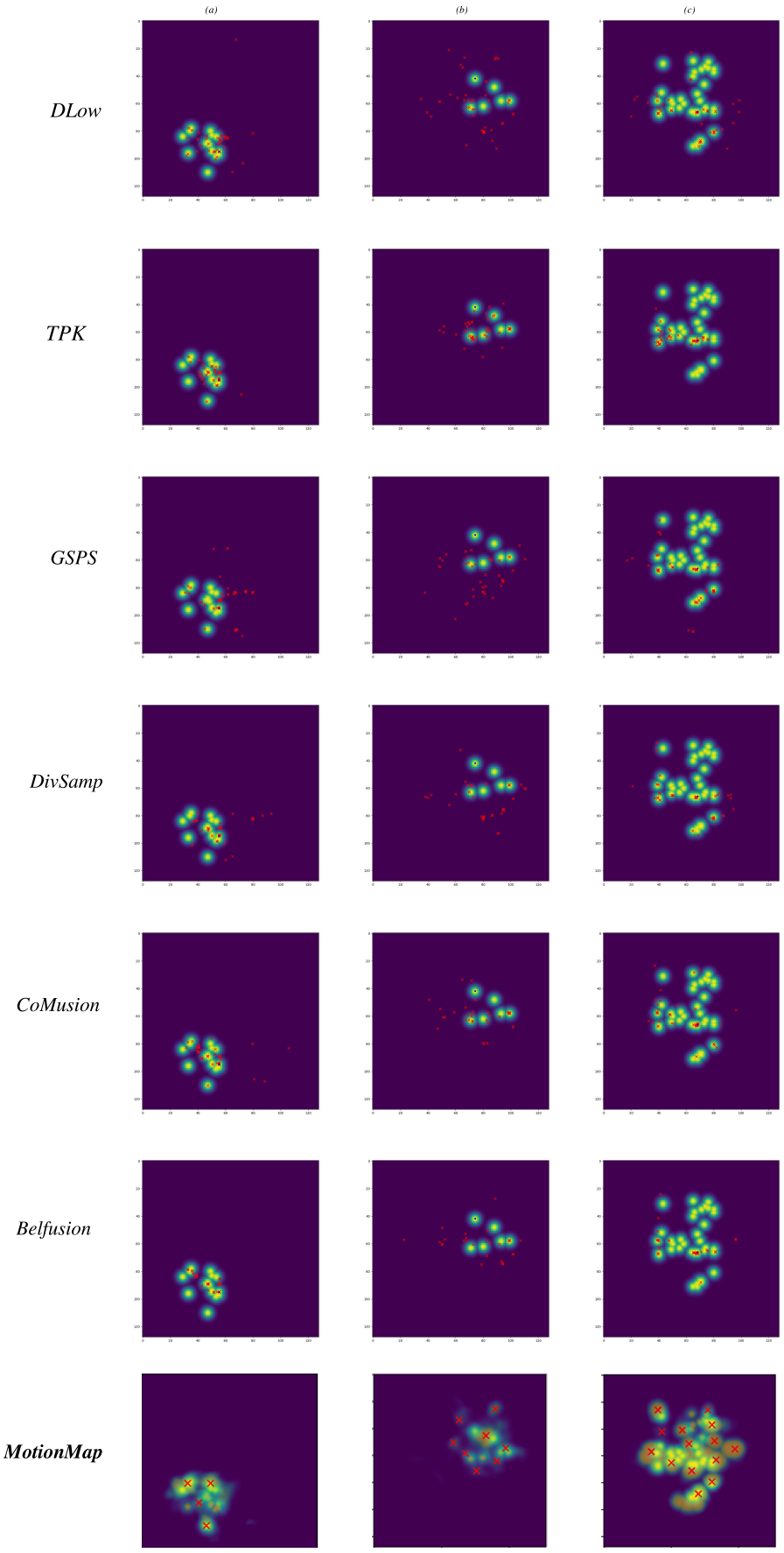

Figure 12. We overlay predictions for each baseline on the ground truth heatmap, for each of the three input pose sequences. The encoding of these predictions is shown as red crosses. For MotionMap, we directly overlay the predicted MotionMap (with crosses for maxima) on the ground truth heatmap. We note that methods are either highly diverse but unrealistic or are less diverse but predict likely futures. In contrast, MotionMap predicts both: common and rare modes since both are explicitly encoded in the training process.

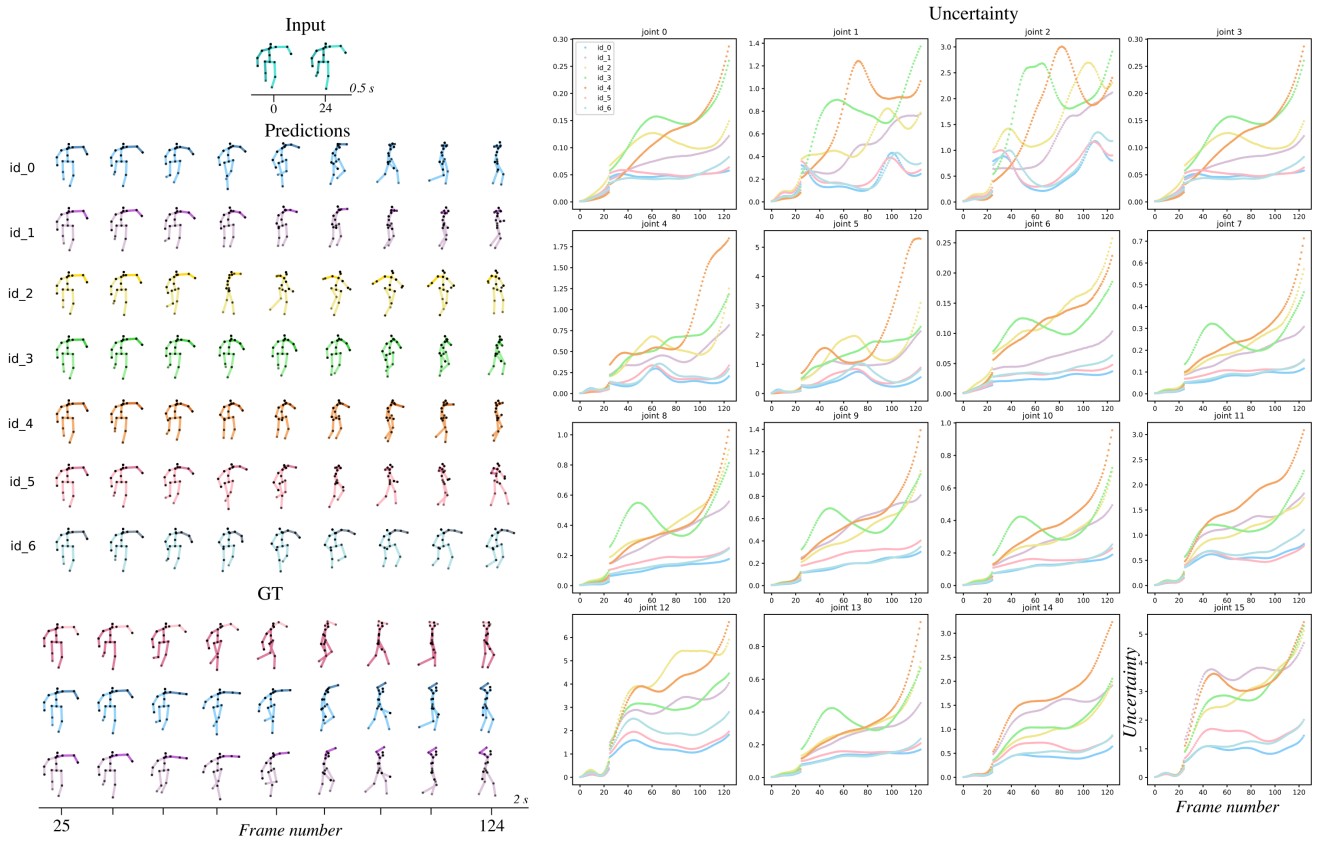

Figure 13. We show additional forecasts along with the predicted uncertainty per joint and time frame.

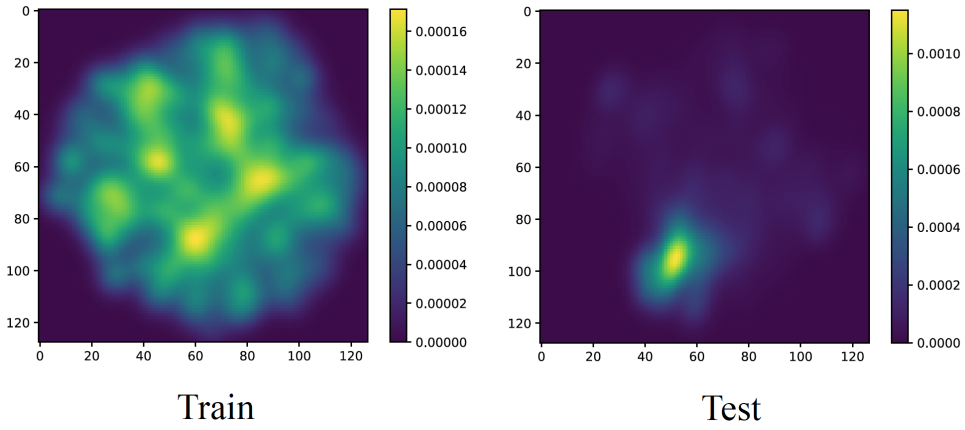

Figure 14. We plot the density map of ground truth sequences Y for the training and testing split of AMASS suggested by [1]. We observe that the splits can be highly imbalanced, and have a significant impact on determining the multimodal ground truth for a sample.