MotionMap: Representing Multimodality in Human Pose Forecasting

Reyhaneh Hosseininejad [*1] Megh Shukla [*1]
Saeed Saadatnejad [1] Mathieu Salzmann [1,2] Alexandre Alahi [1]

[1] École Polytechnique Fédérale de Lausanne (EPFL)
[2] Swiss Data Science Centre (SDSC)
firstname.lastname@epfl.ch

Abstract

Human pose forecasting is inherently multimodal since multiple futures exist for an observed pose sequence. However, evaluating multimodality is challenging since the task is ill-posed. Therefore, we first propose an alternative paradigm to make the task well-posed. Next, while state-of-the-art methods predict multimodality, this requires oversampling a large volume of predictions. This raises key questions: (1) Can we capture multimodality by efficiently sampling a smaller number of predictions? (2) Subsequently, which of the predicted futures is more likely for an observed pose sequence? We address these questions with MotionMap, a simple yet effective heatmap based representation for multimodality. We extend heatmaps to represent a spatial distribution over the space of all possible motions, where different local maxima correspond to different forecasts for a given observation. MotionMap can capture a variable number of modes per observation and provide confidence measures for different modes. Further, MotionMap allows us to introduce the notion of uncertainty and controllability over the forecasted pose sequence. Finally, MotionMap captures rare modes that are non-trivial to evaluate yet critical for safety. We support our claims through multiple qualitative and quantitative experiments using popular 3D human pose datasets: Human3.6M and AMASS, highlighting the strengths and limitations of our proposed method. Project Page (link)

1. Introduction

Human pose forecasting is the task of predicting the future skeletal motion of a person given a set of past skeletal observations. The challenge arises from multimodality since an infinite number of futures with different levels of motion exist for the same observation. Typically, pose forecasting methods make a finite number of predictions to encompass these varied future motions. However, they can never cover all possible modes; one can trivially construct a set of ground-truth futures such that the model predictions have large errors. Indeed, this is reminiscent of the no-free-lunch theorem, which decries the existence of a universal learner. Therefore, one may wonder: Is human pose forecasting truly a solvable problem?

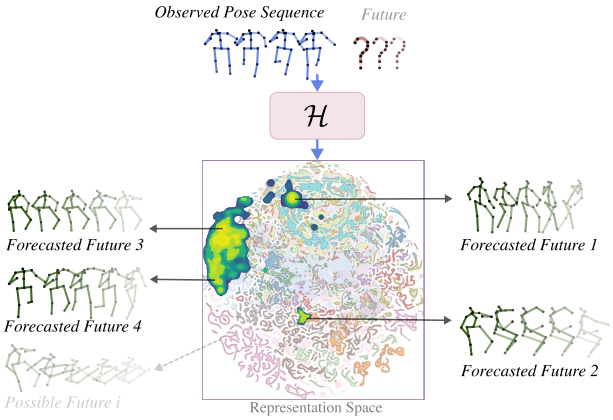

Figure 1. **MotionMap** uses heatmaps to depict a spatial distribution over the space of all possible motions. Local maxima imply that the corresponding motions have a higher likelihood of being a future motion for an observed pose sequence. MotionMap not only predicts a variable number of modes with the corresponding confidence, it explicitly encodes rare modes which could otherwise be averaged out.

We therefore begin by proposing an alternate paradigm to make human pose forecasting well-posed. Instead of attempting to learn an unbounded set of future motions, we encourage the pose forecasting model to explicitly learn future motions present in the observed data. *Specifically, models should use the training set to learn different transitions from input to output pose sequences, and translate them to any unseen test sample.* By doing so the problem is no longer ill-posed, but well-posed since for every input

[*] Equal Contribution

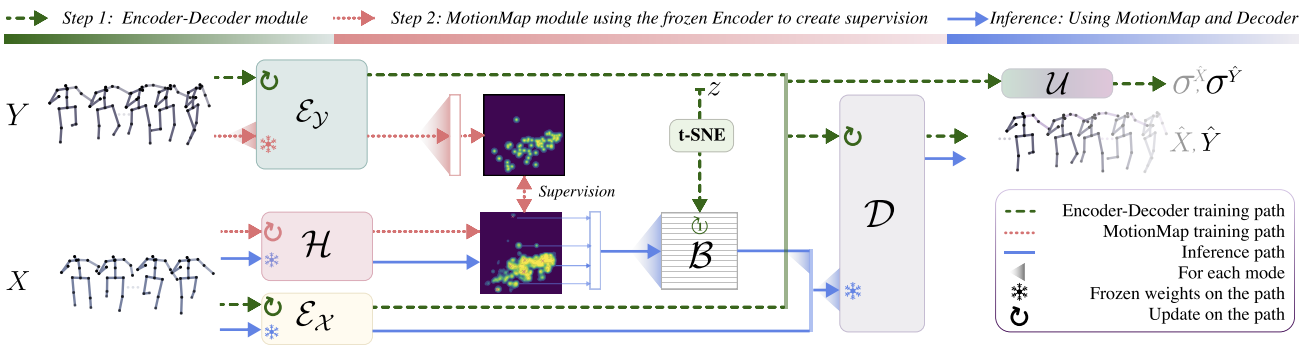

Figure 2. We define a two stage pipeline for human pose forecasting. At first, we train a framework similar to an autoencoder to predict the ground truth and future motion (Sec: 4.3). However, at test time we do not have the future motion and its latent as input. Therefore, we train a heatmap model to predict MotionMap, which along with the codebook encodes the likely motions and their latents as a drop-in replacement (Sec: 4.4). At inference time, we use the predicted MotionMap to obtain latents corresponding to motions with a high likelihood and use it in tandem with the observed pose sequence to predict the future pose sequence (Sec: 4.5)

sequence there exists a fixed number of futures, bounded by the size of the training set. Moreover, explicitly learning these transitions from the training dataset allows us to identify unknown motions at test time.

The question therefore is: How can we explicitly learn the different transitions within the observed data? The challenge is increased by the fact that the number of future motions is variable and dependent on the observed pose sequence. Moreover, state-of-the-art approaches [1–4] model multimodality by oversampling a large number of predictions. We therefore ask: (1) Can we capture multimodality by efficiently sampling a smaller number of predictions? (2) Subsequently, which of the predicted futures is more likely for an observed pose sequence?

In this paper, we propose **MotionMap**, a novel approach that captures different future *motions* for each observation through a heat*map*. Specifically, MotionMap interprets the heatmap as a spatial distribution over the space of all possible motion sequences in two dimensions. Different local maxima on this heatmap correspond to different possible future motions for an observed pose sequence (Fig: 1). This representation has the primary advantage of encoding and predicting a variable number of modes for each observed sequence. Furthermore, we show that this representation allows us to explicitly learn different transitions from the training distribution and translate them for unseen test samples. MotionMap allows us to incorporate two different measures of confidence/uncertainty in our pose forecasting framework. These are the confidence of each mode as well as the uncertainty in the prediction conditioned on the mode. Finally, by design, MotionMap is sample efficient in achieving mode coverage for an observed pose sequence.

We perform experiments on two popular human pose forecasting datasets: Human3.6M and AMASS. Specifically, we compare different methods for (1) sample efficiency, where we compare the metrics for a fixed number of predictions; and (2) the ability to recall transitions present in the observed data. We make two observations: (1) MotionMap has the highest sample efficiency across all methods, and (2) MotionMap can accurately recall transitions from the observed data for unseen test samples. We will make our code publicly available after peer review.

2. Related Work

Early approaches in human pose forecasting employed feed-forward networks [5, 6], and Recurrent Neural Networks (RNNs) to model the temporal aspects of the task [7–9]. Modeling temporal information implicitly is another effective method, commonly achieved by encoding each joint's trajectory with the Discrete Cosine Transformation (DCT), which helps mitigate common failures of auto-regressive models [10]. Subsequent advancements incorporated Graph Convolutional Networks (GCNs) to better capture the spatial dependencies of human poses [11–13].Later, a specific GCN for this task was introduced that combined both temporal and spatial aspects within a single graph framework [14], and another GCN architecture in a two-stage prediction framework was presented, utilizing an initial guess network followed by a formal prediction network [15]. More recently, Transformers have shown effectiveness in capturing spatial and temporal dependencies, needed in human pose forecasting and researchers have incorporated them into their model designs in various configurations, including serial spatial and temporal attention blocks [16], parallel spatial and temporal blocks [17], and hybrid structures [18]. In addition, uncertainty in human pose forecasting has been studied [16], where they showed the impact of modeling homoscedastic uncertainty in the task. While all those approaches have yielded accurate short-term forecasts, they face significant

challenges with relatively long-term predictions due to the inherent multimodality of human motion, where an observed sequence can lead to multiple futures.

Multimodal pose forecasting acknowledges the inherent multimodality of human motion, aiming for a range of possible futures instead of a single outcome. This task has primarily been tackled through stochastic forecasting approaches [1, 2, 19–27] using generative models such as Generative Adversarial Networks (GANs) [19, 20], Variational Auto-Encoders (VAEs) [21–24], probabilistic latent variables [27] and more recently, Diffusion models [1, 2, 25, 26]. To promote diversity across the stochastic samples, DLow [3] introduced a new sampling strategy on top of conditional VAEs by generating multiple Gaussian distributions and then sampling latent codes from different Gaussian priors. GSPS [24] presented a two-stage sampling strategy, addressing the lower and upper body separately. DivSamp [4] proposed using the Gumbel-Softmax sampling strategy from an auxiliary space, and STARS [28] introduced an anchor-based sampling method. SLD [29] expanded this by projecting motion queries into a latent space to allow for more diverse motions, and MDN [30] proposed a transformer-based mechanism that enhances sampling diversity in the latent space. To generate more realistic forecasts, TCD [26] introduced a temporally-cascaded diffusion model that handles both perfect and imperfect observations, and BeliFusion [1] presented a conditional latent diffusion model and recently, CoMusion [2] introduced a single-stage stochastic diffusion-based model using both Transformer and GCN architectures.

While these methods achieved notable results, they rely on stochasticity, learning an implicit distribution over the samples. As a result, these methods rely on large scale sampling for a wider mode coverage. Moreover, the number of samples to draw for different observations is unknown. In contrast, we attempt to determine the number of samples to draw per observation. We also propose a multimodal deterministic pose forecasting approach that explicitly captures all possible modes, ensuring diversity with fewer required samples. This method also enables us to quantify uncertainty and confidence in the predicted motions.

3. Multimodality in Human Pose Forecasting

We ponder upon the question: What is a mode in human pose forecasting? In the absence of a concrete definition, we define a mode as a set of motions corresponding to an action. For instance, walking could represent a class of motions and therefore a mode, and different variations of walking sequences could then be considered samples from this mode. Therefore, one could view multimodality as predicting a diverse set of actions that form a logical transition given an observed pose sequence.

Subsequently, we then wonder: How do we obtain multimodal ground truths for an observed pose sequence? This

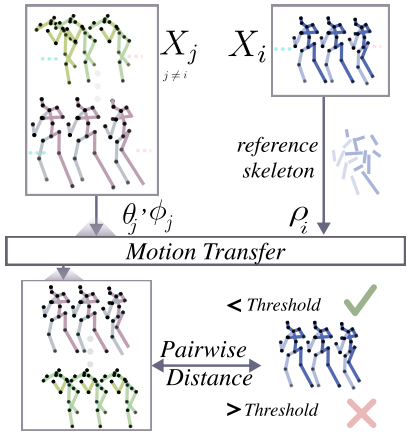

Figure 3. The current approach to finding multimodal ground truths uses only the last frame to measure the similarity between sequences. However, not only does this lose out on motion information, but persons of different sizes with the same motion may not be considered for multimodal ground truth. Hence, we propose computing the ground truths by using the last three frames and scaling the skeleton while retaining the motion. We do this using cartesian to spherical coordinate transformations.

is non-trivial since each input pose sequence is paired with a single output sequence, which is the actual future motion. Therefore, recent literature [1] computes multimodal ground truth by comparing a given input pose sequence with other input pose sequences in the dataset. If two input pose sequences are similar, the corresponding future motions of each sequence can be considered as the multimodal ground truth for each other. This similarity is measured based on a threshold on the distance between the last frame of the two observations.

However, this approach has two key limitations. First, there is no normalization of the skeletal size for different persons, implying that even if two pose sequences have similar motion, they would not be considered as each other's multimodal ground truth if the skeletal sizes are very different. Second, the use of only the last frame from the input sequence may result in abrupt changes in the multimodal ground truth. This is because using one frame carries only positional information, i.e., no information about the motion in the input sequence. Therefore, we modify the definition of multimodal ground truth to (1) involve the last three frames, and (2) scale the skeletal dimensions of all other pose sequences to match that of the given pose sequence (Fig. 3).

We scale the skeleton by converting the pose from cartesian to spherical coordinates. Specifically, we use the spherical coordinate system to represent each joint (e.g. elbow) *w.r.t* its parent (e.g. shoulder). This allows us to maintain the angular positioning but modify the length of the child joint from its parent. By swapping this length with that of the reference skeleton, we obtain a new pose sequence

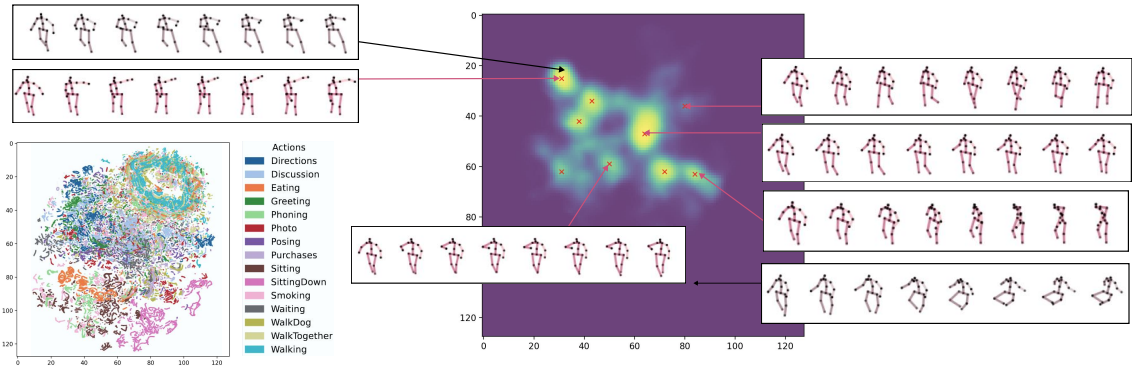

Figure 4. **Controllability.** MotionMap can be used in combination with action labels for controllable pose forecasting. Since each pose sequence is associated with an embedding and action label, a spatial distribution over the space of motions is the same as over the action labels. This allows for the use of MotionMap to select of modes based on the confidence as well as user preference for the forecasted action. We illustrate this distribution over the space of motions $\leftrightarrow$ actions for an example input from the Human3.6M dataset.

the dimensions of which are scaled to match the reference skeleton. We provide the pseudo-code for scaling skeletal dimensions to match the target in the supplementary material (appendix: A).

4. Methodology

Now that we have defined our multimodal ground truths, let us discuss how we encode them through MotionMap, and use MotionMap in multimodal pose forecasting. We begin by describing the notation we use throughout the section.

4.1. Notation

Let X be the sequence of observed poses consisting of T_o frames $X = [x_1, \ldots, x_{T_o}]$. Each pose x is of dimension $(J, 3)$ where J is the number of joints. Similarly, Y is defined as the sequence of future poses with T_f frames, $Y = [x_{T_o+1}, \ldots, x_{T_o+T_f}]$. We define the set of multimodal ground truths as $\boldsymbol{Y_{mm}} : \{Y^i | i : 1, .., M\}$, where M is the number of ground-truth modes. We note that the number of ground-truth modes varies for different samples. During inference, the pose forecaster receives X as input and generates $\hat{\boldsymbol{Y}}_{\boldsymbol{mm}} : \{\hat{Y}^i | i : 1, .., \hat{M}\}$, where $\hat{M}$ is the predicted number of modes. We also note that M and $\hat{M}$ could differ.

4.2. Overview

Our pose forecasting framework, depicted in Figure 2, consists of two trainable modules. We first have an autoencoder consisting of Gated Recurrent Unit encoders $\mathcal{E}_{\mathcal{X}}$, $\mathcal{E}_{\mathcal{Y}}$, a simple multilayer perceptron based uncertainty estimation module $\mathcal{U}$, and a GRUCell based decoder $\mathcal{D}$. We use their architectures from [1], and provide a detailed description in the supplementary material (appendix: B). The autoencoder is not only used for pose forecasting but also to obtain intermediate representations of different output pose sequences, which, as we shall see later, are used to learn the MotionMap.

Our second trainable module is $\mathcal{H}$, which is used to predict the MotionMap for a given observation. Our training process therefore consists of two steps to learn the two modules.

4.3. Step 1: Autoencoder Module

The two encoders $\mathcal{E}_X$ and $\mathcal{E}_Y$ take as inputs X and Y_i respectively. If X has multiple multimodal ground truths $\boldsymbol{Y}_{\boldsymbol{mm}}$, then we select one at random. We expect that different ground truths are selected across different epochs. The outputs of these two networks are arrays z_x and z_y which are concatenated and passed through a simple two layer multilayer perceptron non-linearity giving us $f(z_x \bigoplus z_y)$. Subsequently, this new array is passed to the decoder $\mathcal{D}$ to predict the entire sequence $\widehat{X \bigoplus Y_i}$. The reasoning behind predicting a concatenation of the input and multimodal ground-truth sequence is to avoid discontinuities and predict a smoother transition from the input pose sequence to the predicted pose sequence. This prediction along with the uncertainty predictions from $\mathcal{U}$ are optimized using the negative log-likelihood.

Uncertainty. The uncertainty model uses $f(z_x \bigoplus z_y)$, the input to the decoder, to also predict the corresponding uncertainty of the prediction $\boldsymbol{\sigma}^2 \in \mathbb{R}^{(T_o+T_f)\times J}$. Following Kendall and Gal [31], we use a normal distribution to model the uncertainty where both mean and variance are predicted. If we define the mean square error of the prediction as per joint error (reducing over the joint coordinates), error $= ||\widehat{X \bigoplus Y_i} - X \bigoplus Y_i||_2^2$, the heteroscedastic uncertainty loss function is

$$\mathcal{L} = \frac{\text{error}}{\boldsymbol{\sigma}^2} + \log \boldsymbol{\sigma}^2. \tag{1}$$

We minimize this loss to jointly optimize the uncertainty and decoder, completing the training step for the autoencoder.

4.4. Step 2: MotionMap Module

If we were to use the autoencoder as a pose forecaster at test time, we would be required to know the joint embed-

Figure 5. **Ranking**. Since MotionMap can predict variable number of modes with their associated confidences, our method also allows us to rank predictions. For instance, the highest ranked prediction (top row among $\hat{Y}$) closely matches the ground truth motion. However, rare modes (bottom row) are ranked low since the corresponding mode has a lower confidence.

ding $z_x \bigoplus z_y$, which is the input to the decoder. However, although we can compute z_x since we know the input pose sequence, z_y is nontrivial to obtain. While the literature utilizes various approaches stemming from generative models to obtain this latent, we propose MotionMap, which uses heatmaps and codebooks to represent various motion sequences. Intuitively, MotionMap can be interpreted as a learnt prior over the future motions for each sample, identifying the spread of modes and their associated likelihood. However, as the heatmap represents the spatial spread for two-dimensional coordinates, how can we trace these two dimensional coordinates back to z_y?

4.4.1. Dimensionality Reduction

We implement this by creating a mapping between the embedding z_y and its two dimensional representation h_y which is stored in a codebook $\mathcal{B}$. We obtain this two dimensional embedding by first encoding all future motions Y using $\mathcal{E}_y$ to z_y. Next, we project all embeddings z_y using t-SNE [32] into two dimensions. Subsequently, we scale the two dimensional embeddings such that they span the entire size of the heatmap. Finally, we quantize the embeddings by rounding them to the nearest integer to obtain h_y. Since dimensionality reduction maps from the large volume source domain to a smaller volume target domain, there exist distinct z_ys that are mapped to the same h_y. Therefore, when creating the codebook, we map every h_y to the mean of all z_ys that are mapped to it. To create the heatmap, we iterate over the list of all multimodal ground truths per sample, reduce each of them to their corresponding h_ys, and plot a Gaussian centered around it. Such an approach also ensures that very similar futures are mapped to the same h_y, avoiding duplicity of ground truths. As a result, rare modes are not suppressed in MotionMap.

4.4.2. Training

Training the heatmap predictor $\mathcal{H}$ to obtain MotionMaps for different samples is depicted in Figure 2. For a given sample X, the MotionMap model $\mathcal{H}$ is trained to minimize the pixel wise binary cross entropy loss with the heatmap constructed using ground-truth multimodal samples. To prevent overfitting, the MotionMap model uses a simple GRU encoder which spatio-temporally encodes the last three frames of the observation X. This is followed by a single fully connected layer, which culminates in a series of 1x1 convolutions. The architecture is detailed in the supplementary material.

4.5. Inference

At test time, we use the predicted MotionMap coupled with the codebook to obtain the missing latent z_y. Specifically, given an observation X, we predict the corresponding MotionMap using the model $\mathcal{H}$. We then compute various local maxima *deterministically* which correspond to the likeliest modes of the future sequences. The number of maxima can vary for different observed sequences. Next, we use the codebook B to index a latent vector for the given maxima, giving us the missing latent z_y. We also obtain z_x from $\mathcal{E}_X$, allowing us to pass the updated latent $f(z_x \bigoplus z_y)$ to the decoder for pose forecasting. Finally, we note that this methodology does not rely on stochastic models such as diffusion.

5. Experiments

The goal of this study is to present human pose forecasting as a well posed problem, which in our scenario is to learn and translate different unseen transitions from the observed dataset to unseen test samples. To this effect, we compare our method against multiple state-of-the-art baselines across the AMASS and the Human3.6M datasets.

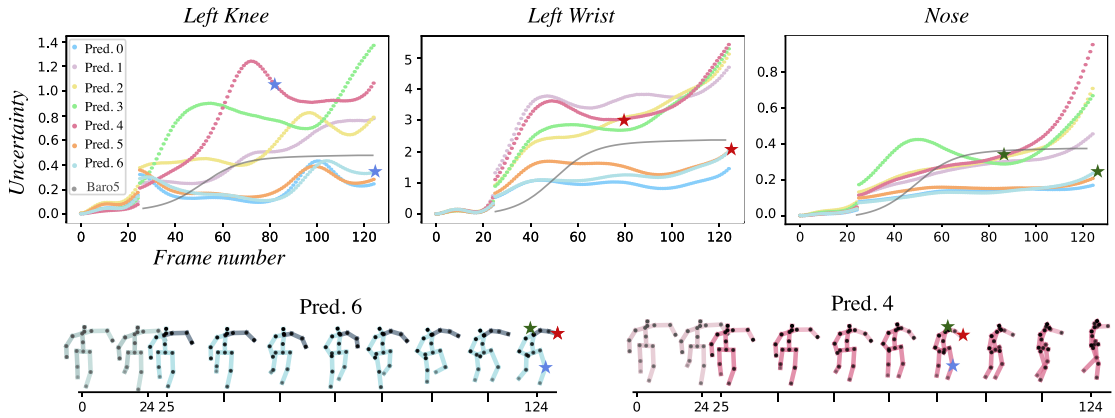

Figure 6. **Uncertainty.** The Baro5 function used in [16] assumes (1) unimodality and a (2) temporally increasing trend in uncertainty. In contrast, we approach multimodality by decomposing the uncertainty into that of the mode and the forecast based on the mode (Sec. 5.3). We observe that this results in semantically richer uncertainty estimates. For instance, prediction 6 has a lower uncertainty in comparison to prediction 4 which involves fast changes in direction. Moreover, the pattern in uncertainty changes based on the motion. Finally, we predict the entire input-output sequence hence the model predicts lower uncertainty in reconstructing the input.

- Human3.6M [33] 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 [34] 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.

5.1. Metrics

We evaluate our results using metrics introduced in prior work [3]. The Average and Final Displacement Error (ADE / FDE) measure the minimum L2 distance among the K predicted pose sequences across all frames and the last frame, respectively. However, this paper focuses on the multimodal versions of these metrics called MMADE and MMFDE, which measure the average ADE and FDE across different multimodal ground truths. A prediction will be penalized if it is not close to any of the test ground truths. We use the standard threshold of 0.5 for Human3.6M and 0.4 for AMASS multimodal ground truth selection.

5.2. Implementation Details

We implemented our methods with PyTorch and ran our experiments on an A100 GPU. The encoding-decoding module was trained for 100 epochs for Human3.6M and 50 epochs for AMASS, with a batch size of 32. We used a learning rate starting from 0.001 and decreasing following a decay on the plateau learning rate schedule. The encoder and decoder are single-layer Gated Recurrent blocks based on [1] and modified for our method. The Heatmap model which predicts the MotionMap is a combination of two simple MLP layers which project the input to the dimensionality of the heatmap, followed by five simple 1x1 convolutional layers.

5.3. Controllability

With Figure 4, we show that the set of pose forecast predictions can be controlled based on user preference. For example, meta data such as action labels can be used to select from within the likeliest pose forecasts based on user preference. Additionally, forecasts similar to a selected one can be sampled by picking the latents corresponding to the nearest neighbors of the local maxima.

5.4. A Tale Of Two Uncertainties

With the proposed approach, we not only learn the aleatoric uncertainty for each prediction but also the confidence of the mode itself. Specifically, suppose Y is a predicted motion, and X is the observed ground truth. In that case, we model the variables as $X \rightarrow Z \rightarrow Y$, where Z is the variable that represents the observed mode. Introducing an intermediate variable allows us to convert a multimodal distribution over Y to an approximate unimodal distribution Figure 6. We observe this difference in modeling by comparing [16] with our approach. In the absence of any additional conditioning

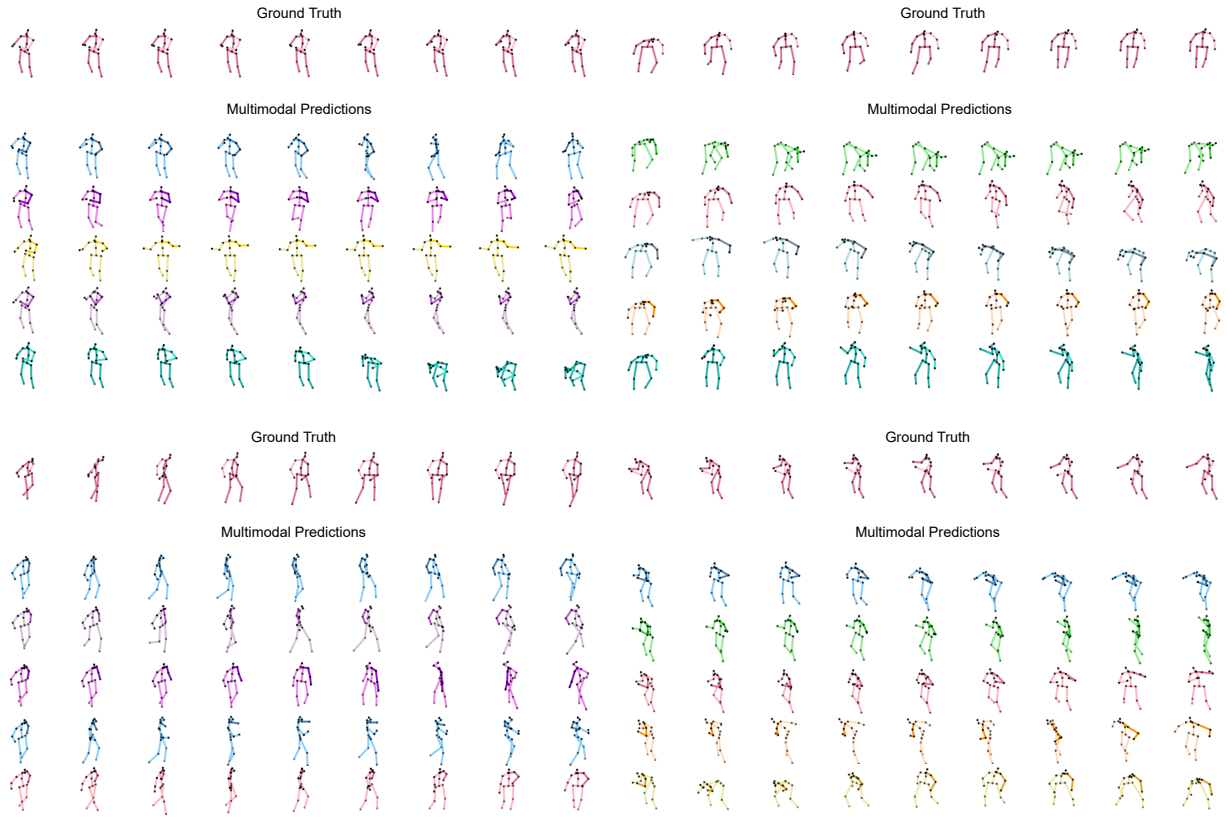

Figure 7. **Diversity**. By virtue of using the same decoder as BeLFusion [1], our method learns realistic yet diverse motions. This is because MotionMap can be decoded to select modes which are sufficiently different yet likely.

information, the predicted variance [16] increases till it saturates at the end of the prediction horizon. This arises due to increased ambiguity in the motion as time increases. In comparison, by conditioning on the mode, we limit the set of future motions, allowing for better modelling of the uncertainty with time. Moreover, a heatmap represents a spatial distribution over different modes, allowing us to quantify the likelihood of the mode itself. The increased uncertainty during the transition from observation (time frames 25-30) to prediction allows the model to smoothly transition and avoid discontinuities in motion. The uncertainty is observed to change smoothly when the future selected poses are the actual ground truth of the input sequence. It is also observed that the uncertainty in the joints closer to the pelvis is generally lower due to limited mobility. However, joints with a higher degree of mobility tend to have higher uncertainties. We provide more visual trends of uncertainty for more sequences and per all joints in the supplementary material.

5.5. Ranking Predictions and Diversity

Unlike state-of-the-art methods, MotionMap by virtue of predicting the confidence of different modes allows us to sort our predictions in order of confidence. We denote this in Figure 5, where we plot the predicted MotionMap along with the associated pose forecasts, ranked based on the confidence of the mode. We observed that high confidence modes are often correlated with the ground truth multimodal future, with very low confidence modes at times resulting in modes which have a rare occurence.

We have visualized some examples of the generated poses in Figure 7, given test input sequences for qualitative comparison of the generated pose sequences. It can be seen that our generated pose sequences are markedly diverse while also depicting smooth and realistic motions.

5.6. Quantitative Results

We quantitatively evaluate different baselines on their ability to translate multimodality from the observed data for any *unseen test sample*. We implement this by calculating multimodal ground truths closest to the testing sample from within the training labels. Such an evaluation not only makes the problem well posed, but also addresses acute mismatches between the training and testing distributions, which we show in the supplementary material (Fig. 13). Moreover, in such

Table 1. Human3.6M dataset: All baselines (except for zero-velocity) are limited to 7 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	0.617	0.879
TPK [35]	6.533	0.534	0.691	0.559	0.675
DLow [3]	11.770	0.445	0.730	0.576	0.715
GSPS [36]	14.97	0.512	0.684	0.550	0.665
DivSamp [4]	**15.733**	0.480	0.685	0.542	0.671
BeLFusion [1]	7.107	0.441	0.597	0.491	0.586
CoMusion [2]	7.325	**0.426**	0.613	0.531	0.623
MotionMap	7.965	0.472	**0.594**	**0.464**	**0.529**

Table 2. AMASS dataset: All baselines (except for zero-velocity) are limited to 7 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	0.778	0.996
TPK [35]	8.570	0.519	0.634	0.600	0.678
DLow [3]	12.694	0.471	0.594	0.554	0.633
GSPS [36]	13.550	0.501	0.662	0.591	0.688
DivSamp [4]	**25.901**	0.479	0.638	0.623	0.728
BeLFusion [1]	7.917	0.347	0.478	0.488	0.564
CoMusion [2]	7.390	**0.311**	0.460	0.526	0.602
MotionMap	8.821	0.324	**0.447**	**0.448**	**0.510**

scenarios, since the testing multimodal ground truth is not representative of the training data, the testing multimodal ground truth does not encompass a majority of the modes. Let us assume that the testing split has only 5 samples. Then, any sample in the testing data will have at most five and a minimum of one multimodal ground truth.

We compare various baselines in Table 2 and Table 1, corresponding to the Human3.6M and AMASS datasets, respectively. We make two observations: While methods such as DLow and DivSamp are diverse, they do not accurately predict the futures corresponding to the observed motion, since not all anchors in the latent are equally likely. In contrast, by imposing a 'prior' on the latent, MotionMap successfully predicts the likeliest modes while being sample efficient. We additionally note that MotionMap successfully outperforms strong diffusion-based baselines in making correct multimodal transitions learnt from the observed data over all test samples. Finally, we also report the performance using the existing methodology of restricting the multimodal ground truth to the testing split (Tables 3, 4). We observe that while MotionMap is much more in recalling transitions from the test set, this does not come at the cost of general performance for unseen samples.

Limitations. A primary limitation of our method is the lack of fine-grained motion prediction within a mode, such as predicting subtle variations in walking. We particularly observe this effect in the testing split with many of the multimodal ground truths being time-shifted versions of the observation. These ground truths are clumped by MotionMap under one mode. Since MotionMap encodes only one forecast per mode, this may result in higher "errors" since the forecast cannot explain all subtle variations with the samples. However, the choice of clumping similar motions together remains a design choice of MotionMap.

6. Conclusion

In this work, we discussed a paradigm to make the learning and evaluation of human pose forecasting well-posed. Next, we proposed a new representation that successfully encodes various future motions per sample. In addition, the representation allows for quantifying the confidence of different modes, as well as successfully learning diverse transitions between the input and observed pose sequence present in the observed data. Finally, by explicitly predicting the spread of future motions, MotionMap is sample-efficient and does not rely on repeated random sampling to achieve mode coverage. Our experiments showed that MotionMap successfully learns to generalize transitions between pose sequences in the observed data and provides suitable measures of uncertainty and confidence per mode.

Acknowledgment

This research is funded by the Swiss National Science Foundation (SNSF) through the project *Narratives from the Long Tail: Transforming Access to Audiovisual Archives* (Grant: CRSII5_198632). The project description is available on: https://www.futurecinema.live/project/

References

[1] G. Barquero, S. Escalera, and C. Palmero, "Belfusion: Latent diffusion for behavior-driven human motion prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 2317–2327. 2, 3, 4, 6, 7, 8, 11, 12, 16

[2] J. Sun and G. Chowdhary, "Comusion: Towards consistent stochastic human motion prediction via motion diffusion," *Proceedings of the European conference on computer vision (ECCV)*, 2024. 3, 8, 12

[3] Y. Yuan and K. Kitani, "Dlow: Diversifying latent flows for diverse human motion prediction," in *Proceedings of the European conference on computer vision (ECCV)*. Springer, 2020, pp. 346–364. 3, 6, 8, 12

[4] L. Dang, Y. Nie, C. Long, Q. Zhang, and G. Li, "Diverse human motion prediction via gumbel-softmax sampling from an auxiliary space," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 5162–5171. 2, 3, 8, 12

[5] C. Li, Z. Zhang, W. S. Lee, and G. H. Lee, "Convolutional sequence to sequence model for human dynamics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 5226–5234. 2

[6] W. Guo, Y. Du, X. Shen, V. Lepetit, X. Alameda-Pineda, and F. Moreno-Noguer, "Back to mlp: A simple baseline for human motion prediction," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2023, pp. 4809–4819. 2

[7] K. Fragkiadaki, S. Levine, P. Felsen, and J. Malik, "Recurrent network models for human dynamics," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2015, pp. 4346–4354. 2

[8] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena, "Structural-rnn: Deep learning on spatio-temporal graphs," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 5308–5317.

[9] J. Martinez, M. J. Black, and J. Romero, "On human motion prediction using recurrent neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2891–2900. 2

[10] W. Mao, M. Liu, and M. Salzmann, "History repeats itself: Human motion prediction via motion attention," in *Proceedings of the European conference on computer vision (ECCV)*. Springer, 2020, pp. 474–489. 2

[11] W. Mao, M. Liu, M. Salzmann, and H. Li, "Learning trajectory dependencies for human motion prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 9489–9497. 2

[12] Q. Cui, H. Sun, and F. Yang, "Learning dynamic relationships for 3d human motion prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 6519–6527.

[13] Z. Liu, P. Su, S. Wu, X. Shen, H. Chen, Y. Hao, and M. Wang, "Motion prediction using trajectory cues," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 13 299–13 308. 2

[14] T. Sofianos, A. Sampieri, L. Franco, and F. Galasso, "Space-time-separable graph convolutional network for pose forecasting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 11 209–11 218. 2

[15] T. Ma, Y. Nie, C. Long, Q. Zhang, and G. Li, "Progressively generating better initial guesses towards next stages for high-quality human motion prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 6437–6446. 2

[16] S. Saadatnejad, M. Mirmohammadi, M. Daghyani, P. Saremi, Y. Z. Benisi, A. Alimohammadi, Z. Tehraninasab, T. Mordan, and A. Alahi, "Toward reliable human pose forecasting with uncertainty," *IEEE Robotics and Automation Letters (RA-L)*, 2024. 2, 6, 7

[17] E. Aksan, M. Kaufmann, P. Cao, and O. Hilliges, "A spatio-temporal transformer for 3d human motion prediction," in *International Conference on 3D Vision (3DV)*. IEEE, 2021, pp. 565–574. 2

[18] W. Zhu, X. Ma, Z. Liu, L. Liu, W. Wu, and Y. Wang, "Motionbert: A unified perspective on learning human motion representations," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2

[19] E. Barsoum, J. Kender, and Z. Liu, "Hp-gan: Probabilistic 3d human motion prediction via gan," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 1418–1427. 3

[20] J. N. Kundu, M. Gor, and R. V. Babu, "Bihmp-gan: Bidirectional 3d human motion prediction gan," in *Pro-*

ceedings of the AAAI conference on artificial intelligence, vol. 33, no. 01, 2019, pp. 8553–8560. 3

[21] J. Walker, K. Marino, A. Gupta, and M. Hebert, "The pose knows: Video forecasting by generating pose futures," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017, pp. 3332–3341. 3

[22] X. Yan, A. Rastogi, R. Villegas, K. Sunkavalli, E. Shechtman, S. Hadap, E. Yumer, and H. Lee, "Mt-vae: Learning motion transformations to generate multimodal human dynamics," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 265–281.

[23] Y. Cai, Y. Wang, Y. Zhu, T.-J. Cham, J. Cai, J. Yuan, J. Liu, C. Zheng, S. Yan, H. Ding *et al.*, "A unified 3d human motion synthesis model via conditional variational auto-encoder," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 11 645–11 655.

[24] W. Mao, M. Liu, and M. Salzmann, "Generating smooth pose sequences for diverse human motion prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 13 309–13 318. 3

[25] L.-H. Chen, J. Zhang, Y. Li, Y. Pang, X. Xia, and T. Liu, "Humanmac: Masked motion completion for human motion prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 9544–9555. 3

[26] S. Saadatnejad, A. Rasekh, M. Mofayezi, Y. Medghalchi, S. Rajabzadeh, T. Mordan, and A. Alahi, "A generic diffusion-based approach for 3d human pose prediction in the wild," in *International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 8246–8253. 3

[27] T. Salzmann, M. Pavone, and M. Ryll, "Motron: Multimodal probabilistic human motion forecasting," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[28] S. Xu, Y.-X. Wang, and L.-Y. Gui, "Diverse human motion prediction guided by multi-level spatial-temporal anchors," in *European Conference on Computer Vision (ECCV)*. Springer, 2022, pp. 251–269. 3

[29] G. Xu, J. Tao, W. Li, and L. Duan, "Learning semantic latent directions for accurate and controllable human motion prediction," *arXiv preprint arXiv:2407.11494*, 2024. 3

[30] H. J. Kim and E. Ohn-Bar, "Motion diversification networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1650–1660. 3

[31] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" *Advances in neural information processing systems*, vol. 30, 2017. 4

[32] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, 2008. [Online]. Available: http://jmlr.org/papers/v9/vandermaaten08a.html 5

[33] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 7, pp. 1325–1339, 2013. 6

[34] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black, "Amass: Archive of motion capture as surface shapes," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 5442–5451. 6

[35] J. Walker, K. Marino, A. Gupta, and M. Hebert, "The pose knows: Video forecasting by generating pose futures," *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017. 8, 12

[36] W. Mao, M. Liu, and M. Salzmann, "Generating smooth pose sequences for diverse human motion prediction," *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 8, 12

[37] [Online]. Available: https://opentsne.readthedocs.io/en/stable/ 11

Apendix

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. 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 (Tables 2, 1), this does not come at the cost of general performance, as evident by these results (3. 4). Finally, we note that restricting the multimodal ground truth to the testing split limits the diversity of modes in the ground truth. In Figure 13, 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.

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

D.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 8) or, if needed, select a less likely future (using metadata) as required by the application's requirements (Figure 9). To better show this possibility we have provided a demo.

D.2. Uncertainty

We have illustrated the predicted uncertainty plots for all the future predicted poses and the reconstructed past in Figure 12. 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.

D.3. Heatmap Comparison

We compare the predicted MotionMaps with the ground truth heatmaps in Figure 10. MotionMap is encouraged to predict

Table 3. 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 [35]	6.727	0.568	0.757	0.582	0.756
DLow [3]	11.687	0.602	0.818	0.616	0.818
GSPS [36]	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	8.308	0.488	0.636	0.502	0.636

Table 4. 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 [35]	9.284	0.762	0.867	0.763	0.864
DLow [3]	13.192	0.739	0.842	0.733	0.846
GSPS [36]	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

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 learn by the MotionMap model from trends across the dataset.

How well can state-of-the-art baselines predict multi-modality without explicitly encoding multimodal transitions? To study this, we collected 50 predictions for each of the baselines for each input pose seqeuence. We then encode these pose forecasts into two dimensions as described in Section 4.4.1. 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.

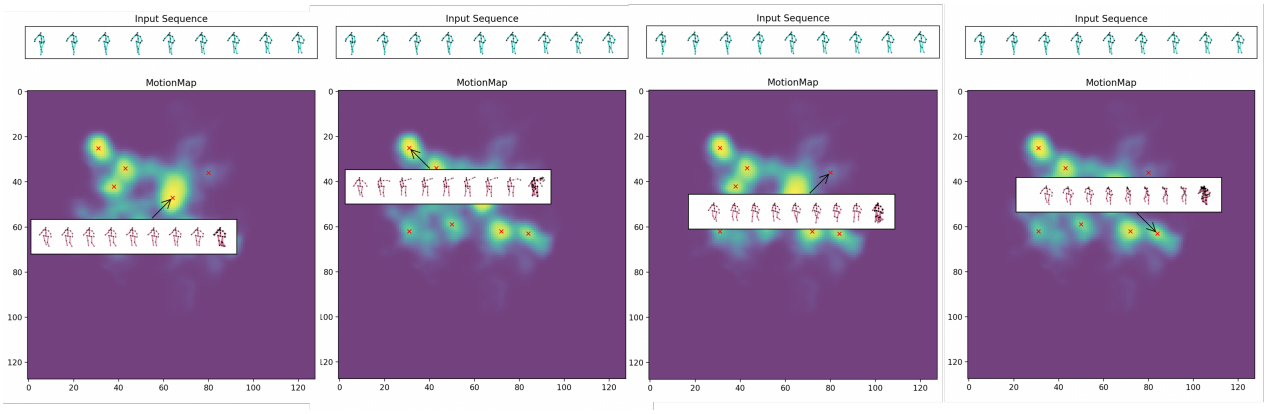

Figure 8. Red crosses mark the modes selected by the model. By hovering over the demo tool, we can view the decoded future poses corresponding to the given input pose sequence. We have uniformly selected eight frames in each sequence to demonstrate the motion and **stacked them on top of each other at the end**(the frame on the very right of each visualized sequence) to represent the amount of motion in each sequence.

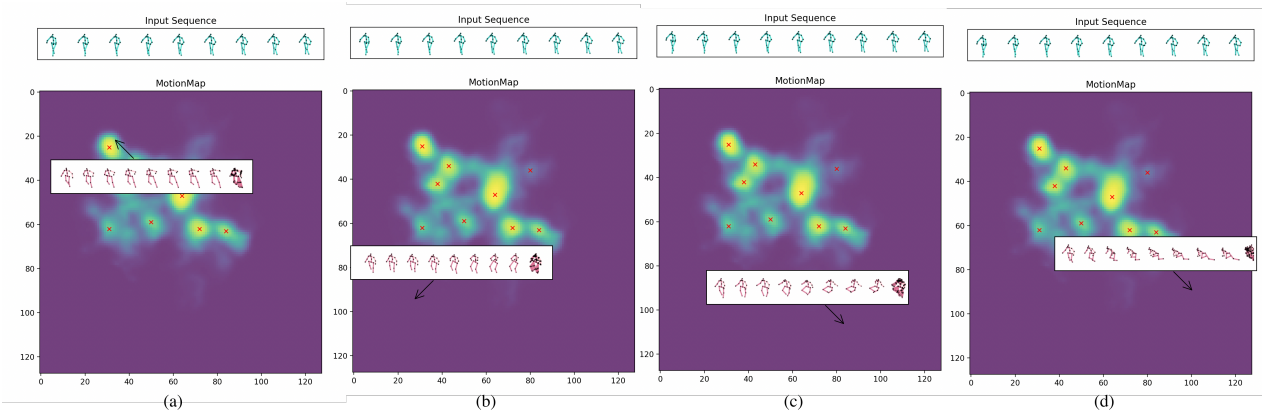

Figure 9. We show different strategies for controlled selection of the other forecasts: (a) Selecting samples in the vicinity of a model selected mode. (b,c,d) Based on the distribution of action labels. For instance, we could generate futures for rarer transitions such as sitting down (b) on a chair (c) on the floor or (d) lying on the floor.

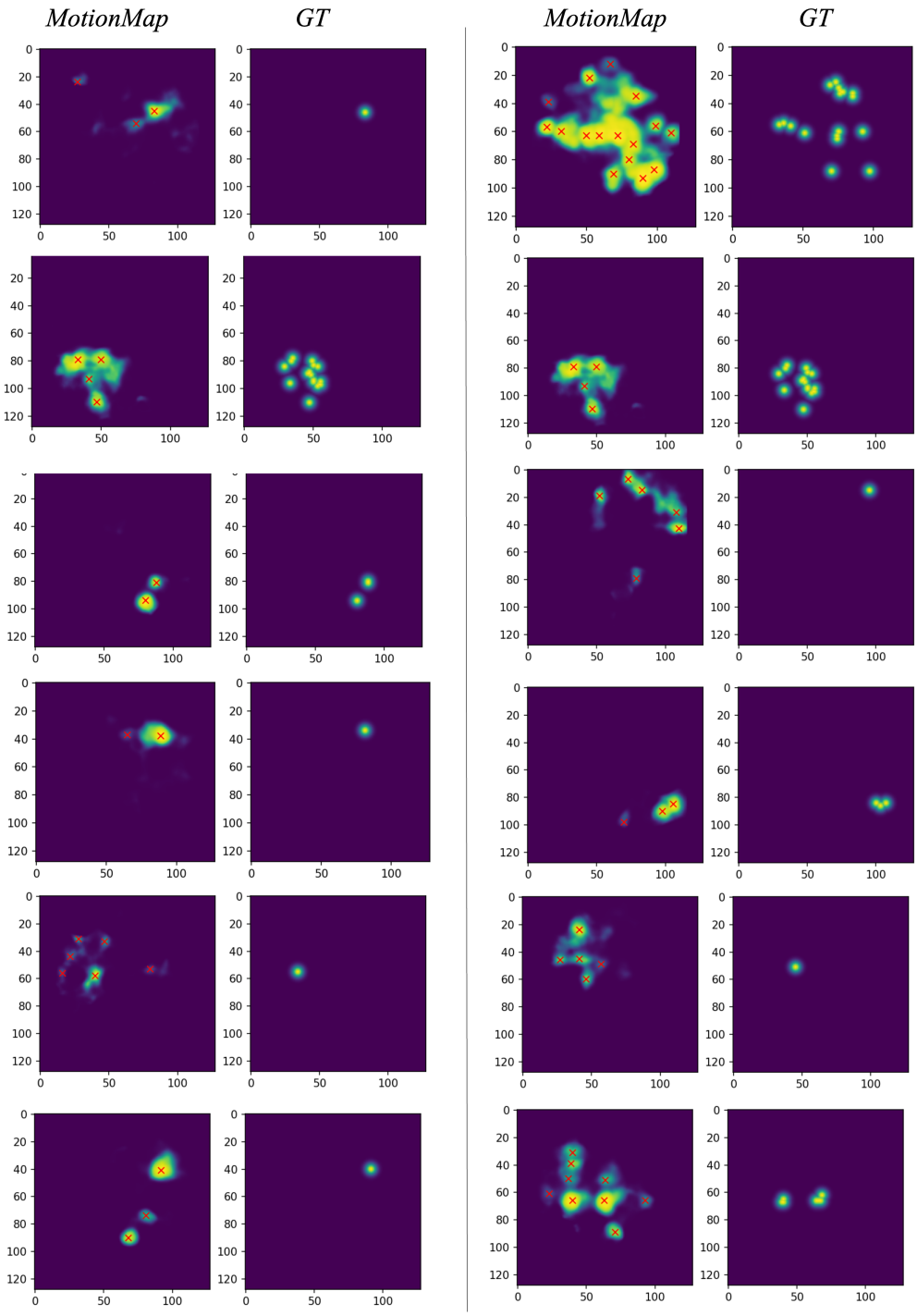

Figure 10. Qualitative comparison between the MotionMap predicted heatmap and the ground truth multimodal heatmap. Our observations indicate that MotionMap effectively captures the diversity of the modeled scenarios. The presence of a larger number of peaks in MotionMap corresponds to the larger diversity of multimodal ground truth found in the train split.

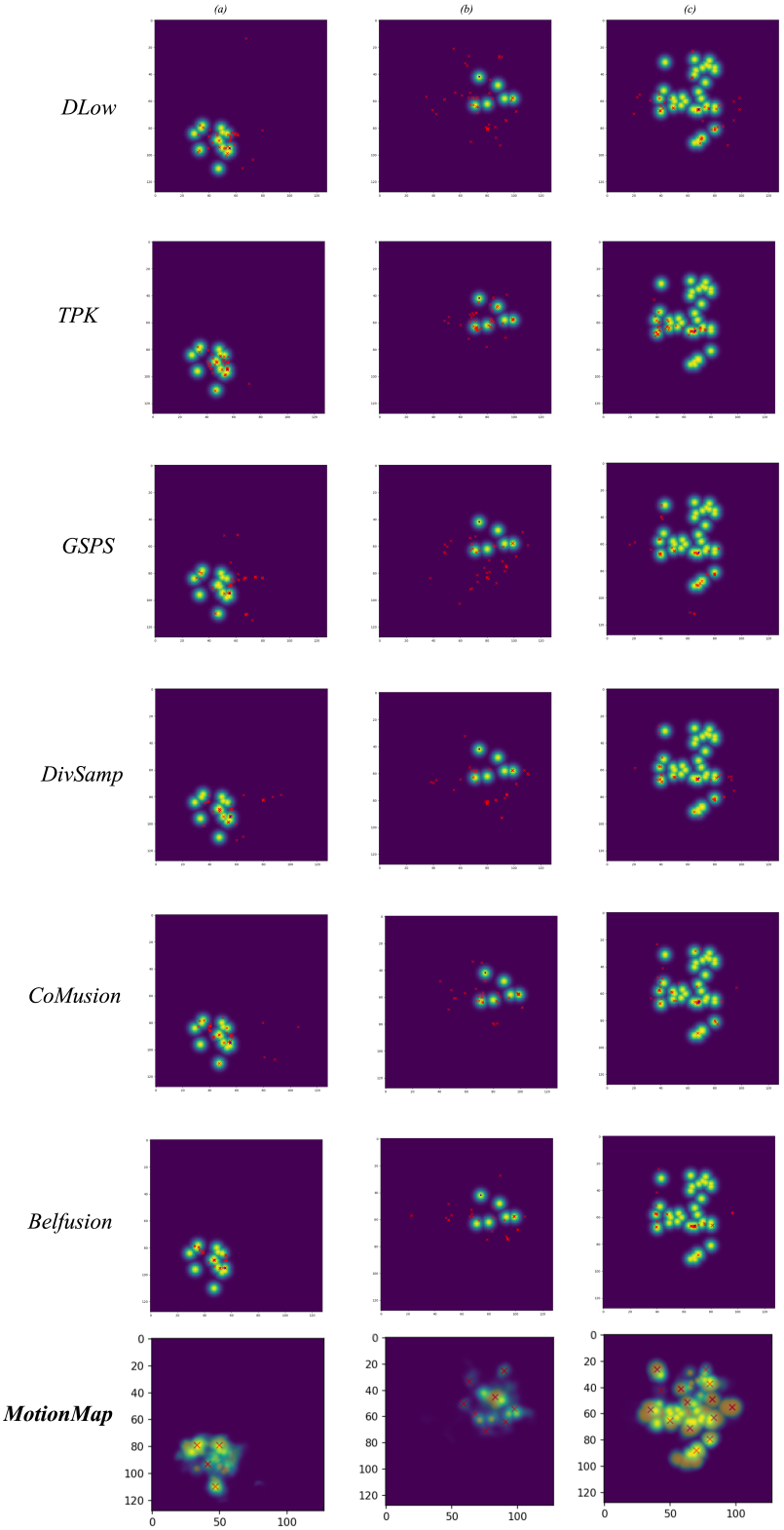

Figure 11. We overlay 50 predictions for each baseline on the ground truth heatmap, for each of the three input pose sequences. The encoding of these 50 predictions is marked with red crosses. For MotionMap, we directly overlay the predicted MotionMap (with crosses for maxima) on the ground truth heatmap. It is observed that methods are either highly diverse but predict unrealistic forecasts or are less diverse but predict likely futures. In contrast, MotionMap predicts both: common and rare modes since both are explictly encoded in the training process.

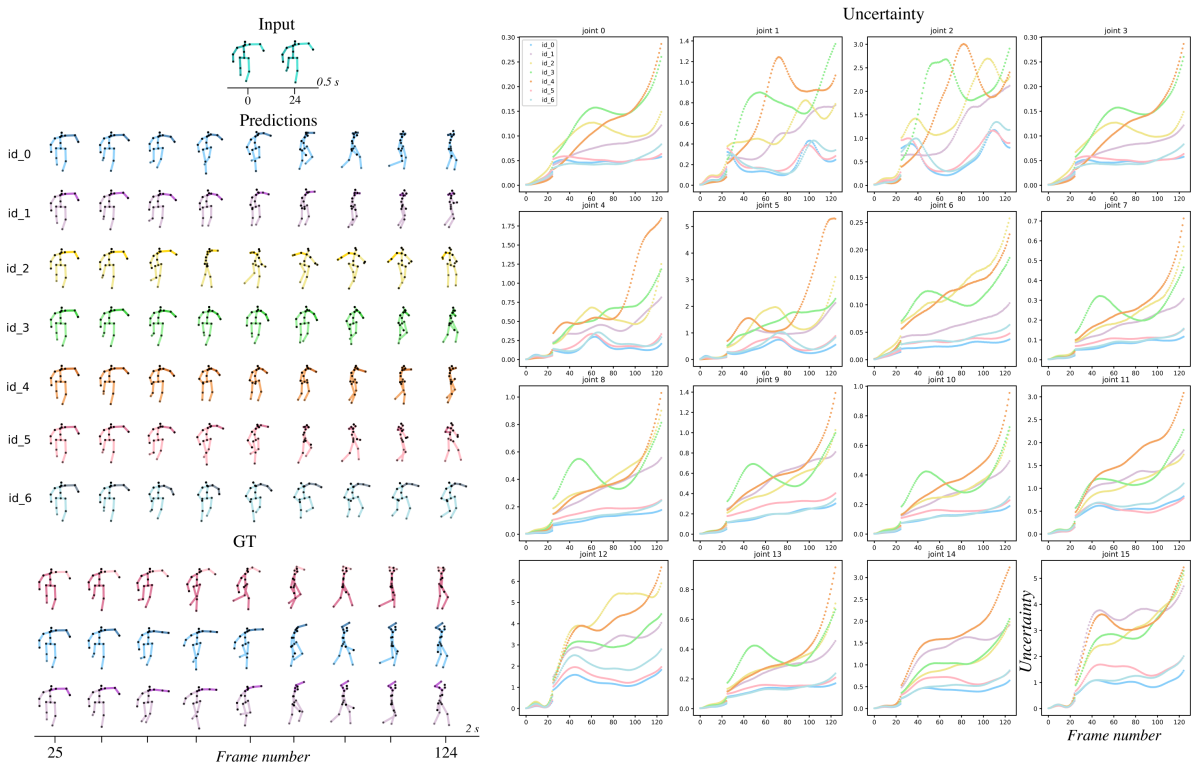

Figure 12. We show additional forecasts along with the predicted uncertainty per joint and time frame.

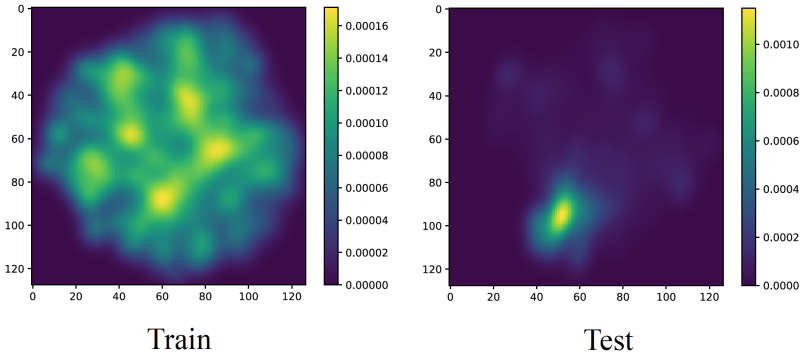

Figure 13. We plot the density map of ground truth sequences Y for the training and testing split of AMASS suggested by [1]. We observe that the splits can be highly imbalanced, and have a significant impact on determining the multimodal ground truth for a sample.