A Stochastic Conditioning Scheme for Diverse Human Motion Prediction* (Supplementary Material)

Sadegh Aliakbarian[†1,2,4], Fatemeh Sadat Saleh[†1,2], Mathieu Salzmann[3], Lars Petersson[1,4], Stephen Gould[1,2]
[1]Australian National University, [2]ACRV, [3]CVLab, EPFL, [4]Data61, CSIRO
{fname.lname}@anu.edu.au, methieu.salzmann@epfl.ch, lars.petersson@data61.csiro.au
https://mix-and-match.github.io/

In this supplementary material, we first provide a qualitative comparison of our approach with other stochastic baselines. We then evaluate the effect of K, i.e., the number of generated motions given a single observation, on the S-MSE metric. We then compare the best of K motions for our approach with the deterministic baselines to show how far the best motion generated by our approach is from the motions generated by state-of-the-art deterministic techniques. We also provide more details on our curriculum learning scheme. We additionally study the effect of α, i.e., the amount of perturbation to the hidden state, on the quality and diversity of the generated motions.

Curriculum Learning of Perturbation

As discussed in the main paper, our approach benefits from curriculum-based perturbation of the hidden state. The parameter α determines the trade-off between the level of determinism and the quality of motions. In Algorithm 1, we provide the pseudo-code for our Mix-and-Match index sampling.

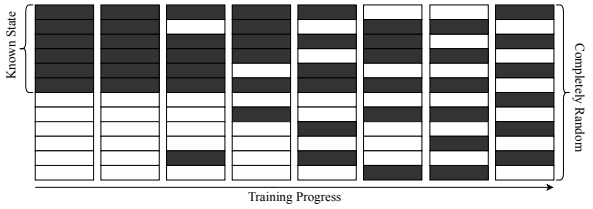

Figure 1. Example of curriculum perturbation of the hidden state. At the beginning of training, the perturbation occurs in a deterministic portion of the hidden state. As training progresses, we gradually, and randomly, spread the perturbation to the rest of the hidden state. This continues until the indices to perturb are uniformly randomly sampled.

*This research was supported by the Australian Government through the Australian Research Council (ARC).

†Equal contribution.

Algorithm 1 Curriculum Index Sampling

Inputs: current epoch `e`, perturbation ratio α, sampling step `step`
Result: Sampled indices for perturbation at epoch `e`, `indices`
s = int($\frac{\texttt{e}}{\texttt{step}}$)
$th = min\{\frac{\alpha L}{2}, \frac{L-\alpha L}{2}\}$
`indices` = []
if $s < th$ **then**
  `indices`.append(sample $\alpha L - s$ indices from known part)
  `indices`.append(sample s indices from the rest)
end
else
  `indices`.append(sample αL indices from hidden state)
end
return `indices`

Evaluating the effect of K

In the main paper, we used $K = 50$ to compare our approach with the state-of-the-art deterministic and stochastic baselines. Here, we provide an ablation study on the effect of K. To this end, we provide results when using $K = 1$ to $K = 500$. In Fig. 2, we plot the results with $K = 50$ as bold black lines, and the shaded area covers the results obtained with $K = 1$ to $K = 500$. While smaller values of K yield large errors, the difference between $K = 50$ and $K = 500$ is very small (barely visible in most cases).

Qualitative Evaluations

In Fig. 3, we provide a qualitative comparison between our approach and existing stochastic motion prediction models. To this end, for four different motions, we generate (i.e., sample) three random motions (not cherry-picked) for our approach and the stochastic baselines. As can be seen in Fig. 3, while being natural and realistic, our approach generates much more diverse motions than the baselines. The motions from `LHP` and `RHP` are very natural, but with very low diversity (almost all identical). While `LPP`'s motions are diverse, they are of considerably lower quality.

Evaluating the effect of α

Our approach depends on the parameter α, which defines the amount of randomness used in our mix-and-match perturbations. In Fig. 4, we report the quality and diversity

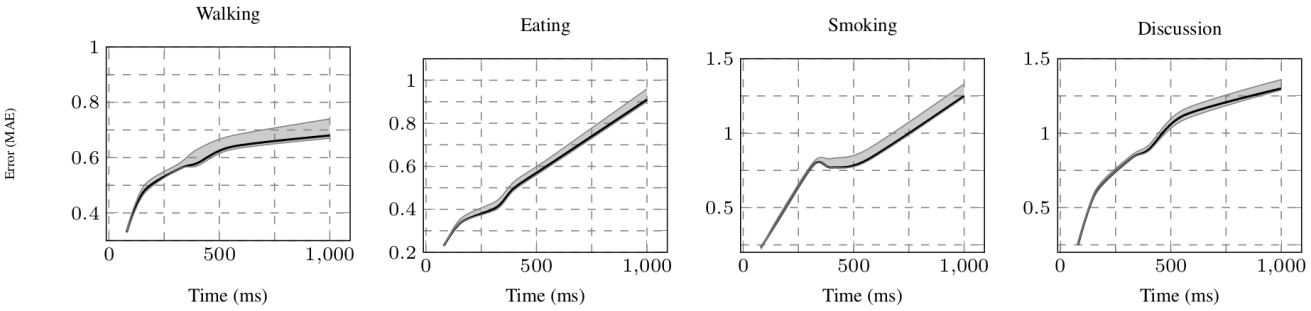

Figure 2. Effect of K on the MAE of our predictions on the Human3.6M dataset. The bold black line is the best of $K = 50$ motions, and the shaded area indicates the region between best of $K = 1$ and $K = 500$.

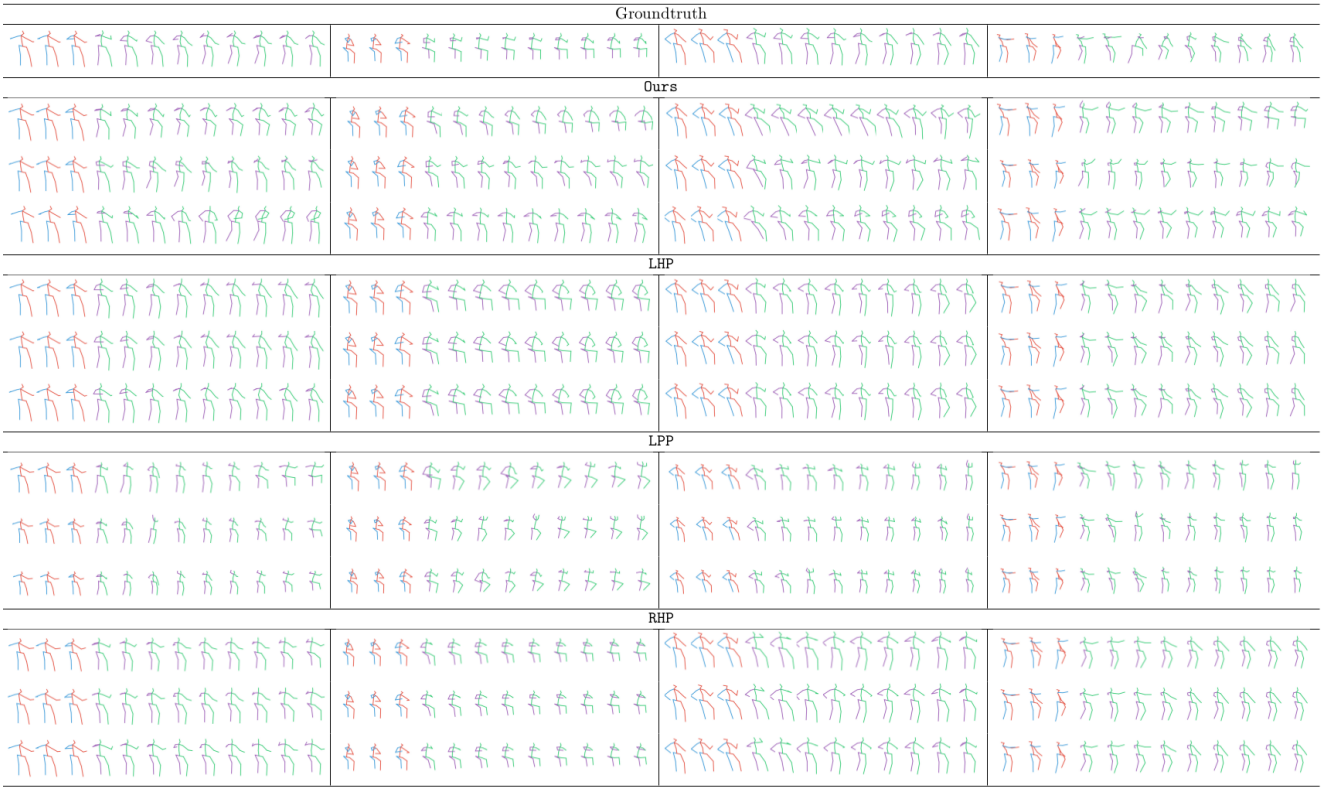

Figure 3. Qualitative comparison to stochastic baselines.

of our results when varying $\alpha \in \{0.1, \ldots, 0.9\}$. Note that these plots show a trade-off between quality and diversity. This is to be expected since, by aiming to increase diversity, the resulting motions will become unrealistic. Nevertheless, our results can be seen to be highly diverse and of high quality for a wide range of values, i.e., by setting $\alpha \in [0.3, 0.7]$. Note that α is the only model-related hyper-parameter of Mix-and-Match. The quality and diversity metrics are monotonic functions of α, thus, one can choose a proper α given a task. Note that, using $\alpha = 0.2$, our method still achieves a SoTA diversity of 2.25 with a higher quality of 45.0%. However, for the sake of fair comparison, we use the default value of $\alpha = 0.5$ for all of the experiments provided here and in the main paper.

Tables of Qualitative Experiments in the Main Paper

In the main paper, we qualitatively evaluated the quality and the diversity of generated motions by our `Mix-and-Match` approach as well as by other stochastic motion prediction techniques [2, 1, 3]. The results of this experiments were illustrated in plots, therefore, here we provide the numbers in the Table 1 to facilitate future comparisons. Similarly, we provide the numbers for the experiment comparing the our quality and diversity metrics

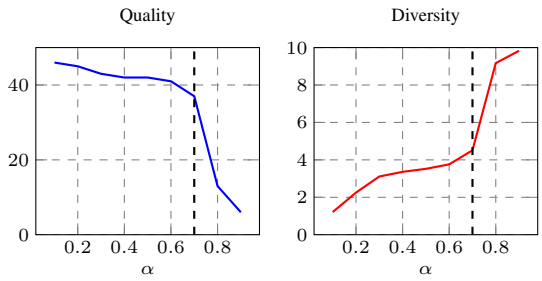

Figure 4. Quality and diversity of the motions generated with our approach as a function of α. Note that with $\alpha > 0.7$, diversity increases significantly, but this diversity is the result of poor-quality motions.

with the human evaluations in Table 2.

References

[1] Emad Barsoum, John Kender, and Zicheng Liu. Hp-gan: Probabilistic 3d human motion prediction via gan. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1418–1427, 2018.

[2] Jacob Walker, Kenneth Marino, Abhinav Gupta, and Martial Hebert. The pose knows: Video forecasting by generating pose futures. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 3352–3361. IEEE, 2017.

[3] Xinchen Yan, Akash Rastogi, Ruben Villegas, Kalyan Sunkavalli, Eli Shechtman, Sunil Hadap, Ersin Yumer, and Honglak Lee. Mt-vae: Learning motion transformations to generate multimodal human dynamics. In *European Conference on Computer Vision*, pages 276–293. Springer, 2018.

	Training Progress									
Baseline	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Yan et al., [3]	1.19/29.0	1.13/37.0	0.28/45.0	0.24/46.0	0.21/46.0	0.47/45.0	0.28/44.0	0.26/45.0	0.24/46.0	0.26/45.0
Barsoum et al., [1]	0.76/27.0	0.65/40.0	0.62/43.0	0.58/45.0	0.54/44.0	0.52/44.0	0.51/45.0	0.50/45.0	0.48/46.0	0.48/47.0
Walker et al., [2]	2.21/03.0	1.35/11.0	1.23/12.0	1.28/12.0	1.43/14.0	1.58/11.0	1.69/12.0	1.74/14.0	1.75/14.0	1.70/13.0
Mix-and-Match	2.02/22.0	2.34/27.0	2.73/30.0	2.95/34.0	3.17/39.0	3.26/40.0	3.49/41.0	3.63/42.0	3.53/42.0	3.52/42.0

Table 1. Quality and diversity of our approach and the stochastic baselines.

Baseline	$\mathrm{Q}_{\mathtt{Cls}}$	$\mathrm{Q}_{\mathtt{Human}}$	Div
Yan et al., [3]	45.0	37.0	0.26
Barsoum et al., [1]	47.0	38.1	0.48
Walker et al., [2]	13.0	15.0	1.70
Real motions (groundtruth)	50.0	50.0	0.00
Mix-and-Match	42.0	40.9	3.52

Table 2. classifier-based and human evaluation of quality for our approach and the baselines. These statistics correspond to evaluation after the models are fully trained.