Synthesis of Compositional Animations from Textual Descriptions - Supplementary materials

Anindita Ghosh *[1,3], Noshaba Cheema[1,2,3], Cennet Oguz[1,3], Christian Theobalt[2,3], and Philipp Slusallek[1,3]

[1]German Research Center for Artificial Intelligence (DFKI)
[2]Max-Planck Institute for Informatics
[3]Saarland Informatics Campus

1. Results and Discussion

We compare our method with the two baseline methods, JL2P [1] and the method of Lin et al. [2], and also with the four ablations of our method: 'w/o BERT', 'w/o JT', 'w/o 2-St', 'w/o Lo', as described in Section 4.4 of our paper. We include more experiments here with two sub-groups of Ablation 3 ('w/o Lo').

- **Ablation 3a: Training the hierarchical two-stream model without the adversarial loss (w/o AdLo).** We discard the adversarial loss terms (L_D, L_G) described in Section 3.2 when training the model.
- **Ablation 3b: Training the hierarchical two-stream model without the Embedding Similarity Loss (w/o EmLo).** We discard the Embedding similarity loss (L_E) introduced in Section 3.2 when training the model.

We show the average positional error (APE) values for individual joints in Table 1. When compared to the ablations of our model, we find that the APE calculated over the mean of all the joints with the global trajectory is marginally better for the ablations compared to our method (best for the ablation 'w/o 2-St', showing an improvement of 1.96% over our method). This is because the motions get averaged out in the ablations, bringing the joint positions closer to the mean. However, it also reduces the relevant joint movements. By contrast, our method has the lowest APE for the root joint, implying that the overall motion quality is better. The additional metric of the average variance error (AVE) for evaluating the variability of the motions further shows that the joint movements are reduced in the ablations. Our method has the lowest AVE for the root joint as shown in Table 2. Our method also performs the best in terms of the content encoding error (CEE) and the style encoding error (SEE) compared to the ablations and the baseline methods as seen in Table 3.

References

[1] C. Ahuja and L. Morency. Language2pose: Natural language grounded pose forecasting. In *2019 International Conference on 3D Vision (3DV)*, pages 719–728, 2019.

[2] Angela S Lin, Lemeng Wu, Rodolfo Corona, Kevin Tai, Qixing Huang, and Raymond J Mooney. Generating animated videos of human activities from natural language descriptions. *Learning*, 2018:2, 2018.

*Corresponding Author: anindita.ghosh@dfki.de.

Table 1: Average Positional Error (APE) in mm for our model compared to JL2P [1], Lin et al. [2], and the ablations of our method described in Section 4.4 of our paper and in Section 1 of the supplementary. Lower values are better. Although the overall APE is lower for our ablations, we find the overall motion quality to be poorer than our final method due to larger errors in the root. Please refer to Section 5.1 of our paper for details.

	JL2P	Lin et al.	w/o BERT	w/o JT	w/o 2-St	w/o Lo	w/o AdLo	w/o EmLo	Ours
Trajectory	4.12	4.52	1.21	1.27	1.22	1.23	**1.16**	1.23	1.22
Root	7.28	7.78	3.23	3.50	3.22	3.23	**3.21**	3.24	**3.21**
Torso	13.18	14.93	5.84	**5.71**	**5.71**	5.91	5.8	5.85	5.90
Pelvis	14.92	16.10	**6.49**	6.54	6.52	6.67	6.51	6.55	6.60
Neck	33.01	36.03	14.88	**14.50**	14.69	15.04	14.80	14.90	15.01
Left Arm	37.37	41.71	16.54	16.79	**16.09**	16.79	16.91	16.89	16.94
Right Arm	37.91	42.33	16.41	16.56	**15.81**	16.25	16.28	16.15	16.40
Left Hip	13.50	14.33	**6.02**	6.12	6.14	6.18	6.04	6.07	6.21
Right Hip	13.39	14.05	**6.00**	6.15	6.15	6.20	6.06	6.12	6.22
Left Foot	38.38	38.84	16.78	16.63	16.84	**16.25**	16.49	16.70	16.97
Right Foot	39.66	40.31	17.12	17.15	17.24	**16.78**	17.01	17.15	17.22
Mean w/o trajectory	24.86	26.64	10.93	10.96	**10.84**	10.93	10.91	10.97	11.07
Mean	22.97	24.63	10.04	10.08	**9.97**	10.05	10.02	10.08	10.17

Table 2: Average Variance Error (AVE) for our model compared to JL2P [1], Lin et al. [2], and the ablations of our method described in Section 4.4 of our paper and in Section 1 of the supplementary. Lower values are better. Our method has the lowest AVE for the root joint as well as the mean of all the joints with and without the global trajectory.

	JL2P	Lin et al.	w/o BERT	w/o JT	w/o 2-St	w/o Lo	w/o AdLo	w/o EmLo	Ours
Trajectory	18.55	19.00	10.87	10.52	11.20	9.75	**8.91**	9.59	10.29
Root	4.70	5.46	2.45	2.42	2.32	2.30	**2.19**	2.22	**2.19**
Torso	21.44	22.61	12.65	12.20	13.22	11.85	**10.38**	11.41	11.87
Pelvis	23.79	24.51	13.66	13.25	13.99	12.73	12.59	12.59	**12.58**
Neck	45.05	36.03	26.24	25.26	27.37	24.78	**24.08**	23.81	24.65
Left Arm	32.66	41.71	16.59	16.42	16.86	15.66	15.00	**14.67**	15.20
Right Arm	29.15	42.34	15.18	14.54	15.05	14.31	13.98	13.95	**13.95**
Left Hip	27.79	28.73	16.01	15.45	15.82	14.35	14.46	**14.04**	14.71
Right Hip	26.73	27.05	14.46	14.13	14.92	**13.31**	13.41	13.40	13.40
Left Foot	48.34	38.84	24.63	24.03	23.67	22.27	21.65	21.61	**21.57**
Right Foot	47.23	40.31	23.04	23.10	22.80	20.72	**19.43**	20.14	20.87
Mean w/o Trajectory	30.69	30.75	16.49	16.08	16.60	15.22	**14.71**	14.78	15.09
Mean	29.58	29.69	15.98	15.57	16.11	14.73	**14.18**	14.31	14.66

Table 3: Content Encoding Error (CEE) and Style Encoding Error (SEE) for our model compared to JL2P [1], Lin et al. [2], and the ablations of our method described in Section 4.4 of our paper and in Section 1 of the supplementary. Lower values are better. Our method has the lowest CEE and SEE.

Method	JL2P	Lin et al.	w/o BERT	w/o JT	w/o 2-St	w/o Lo	w/o AdLo	w/o EmLo	Ours
CEE	1.06	1.92	1.10	0.99	0.67	1.04	0.54	1.03	**0.53**
SEE	0.38	1.13	0.80	0.76	0.46	0.77	0.20	0.72	**0.19**