Multi-Track Timeline Control for Text-Driven 3D Human Motion Generation

Supplementary Material

We encourage the reader to view the supplementary video to observe qualitative results in motion (see Section A for interpretation). This appendix includes additional details about our STMC method (Section B), the creation of the MTT dataset (Section C), and our MDM-SMPL model (Section D). We also present supplementary experiments in Section E.

A. Supplementary Video with Qualitative Results

Besides this appendix, we provide a video on our webpage[1] where we visually explain the method and show qualitative results from STMC, along with a comparison to the baselines. By looking at the generated motions in the video, we can see more clearly the differences between the baselines and STMC. In particular, although SINC w/ Lerp has good metrics overall, some motions do not look natural to the human eye.

B. Spatio-Temporal Motion Collage (STMC)

Resolving unassigned timeframes. This step corresponds to Step 2 of Fig. 3a of the main paper. As shown in Fig. A.1, we first cut the body part timelines so that there are no new texts appearing or disappearing within each cut (left). Then, we apply the SINC [1] heuristic (right) for each cut. The heuristic consists of (1) choosing a "base" text prompt, (2) assigning all the body parts to the base text, and (3) assigning (overriding) the body parts corresponding to the other texts. Note that the other texts are sorted based on the number of body parts involved in decreasing order.

Runtime. We compare the computational complexity of our STMC test-time denoising approach to the independent generation baseline (SINC w/o Lerp) where both methods use our MDM-SMPL backbone. This comparison highlights the overhead introduced in STMC, as the outputs need to be merged (both spatially and temporally) at each diffusion step. We measure the runtime on a single interval consisting of 3 prompts, which includes 3 transitions. On average, the generation time for SINC is approximately 1.14 sec compared to about 1.40 sec for STMC, totalling a 23% increase.

C. Multi-Track Timeline (MTT) Dataset

Full list of texts. As mentioned in Sec. 4.1 of the main paper, to create the MTT dataset, we collect a set of 60 text prompts along with body parts labels for each one. Each of these atomic prompts is shown below, where the relevant body parts are annotated after the # symbol.

```
walk in a circle clockwise # legs
walk in a circle counterclockwise # legs
walk in a quarter circle to the left # legs
walk in a quarter circle to the right # legs
turn 180 degrees to the left on the left foot # legs
turn 180 degrees to the left on the right foot # legs
turn left # legs
turn right # legs
walk forwards # legs
walk backwards # legs
slowly walk forwards # legs
slowly walk backwards # legs
quickly walk forwards # legs
quickly walk backwards # legs
run # legs
jogs forwards # legs
jogs backwards # legs
perform a squat # legs # spine
sit down # legs # spine
low kick with the right foot # legs
low kick with the left foot # legs
high kick with the right foot # legs
high kick with the left foot # legs
applause # left arm # right arm
play the guitar # left arm # right arm
play the violin # left arm # right arm
raise both arms in the air # left arm # right arm
raise the right arm # right arm
raise the left arm # left arm
wave with the right hand # right arm
wave with the left hand # left arm
wave with both hands # left arm # right arm
talk on phone with left hand # left arm # head
talk on phone with right hand # right arm # head
point with his right hand # right arm
point with his left hand # left arm
drink with the left arm # left arm # head
drink with the right arm # right arm # head
eat something with the left arm # left arm # head
eat something with the right arm # right arm # head
look right # head
look left # head
dodge a hit to his head # head # spine
throw something with left hand # left arm
throw something with right hand # right arm
pick something with the left hand # left arm # legs # spine
pick something with the right hand # right arm # legs # spine
bow # spine # head
punch with the left hand # left arm
punch with the right hand # right arm
jump forward # legs # spine
jump backward # legs # spine
hop to the left # legs # spine
hop to the right # legs # spine
play golf # legs # left arm # right arm # head # spine
jumping jacks # legs # left arm # right arm # spine
touches back of head with right hand # right arm # head
touches back of head with left hand # left arm # head
wipe with the left hand # left arm
wipe with the right hand # right arm
```

Listing 1. **Full list of texts:** We use these text prompts as the base "atomic" actions for creating the MTT dataset.

Sampling duration. After choosing the text prompts, we randomly sample durations with a mean of 6.0 seconds and a standard deviation of 1.0 seconds.

Samples from the MTT dataset. As shown in Fig. A.2, our MTT dataset consists of diverse timelines.

[1] https://mathis.petrovich.fr/stmc

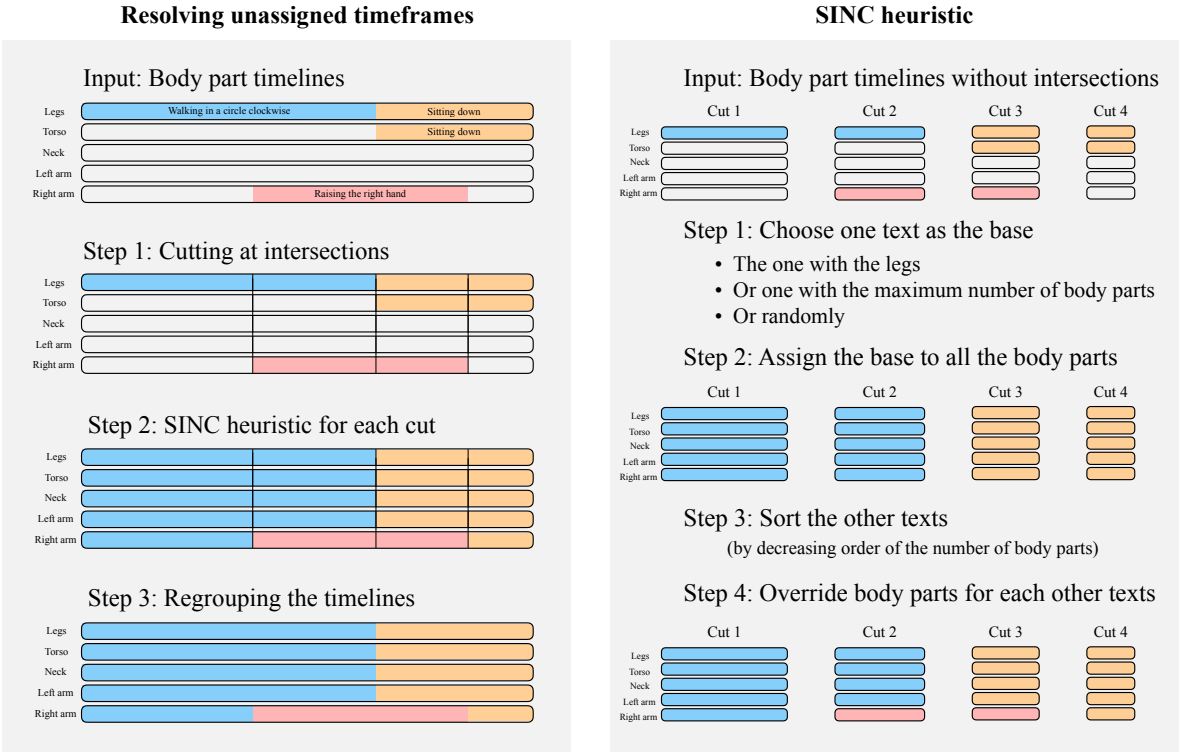

Figure A.1. **Additional details of STMC:** To create the final body parts timeline, we need to "fill the holes" by assigning a text to all locations of the body parts timeline (left). This is done by first splitting the timelines such that there is no intersection with other intervals, and then applying the SINC heuristic for each cut (right). Finally, we regroup the intervals by removing the cuts to obtain full body part timelines.

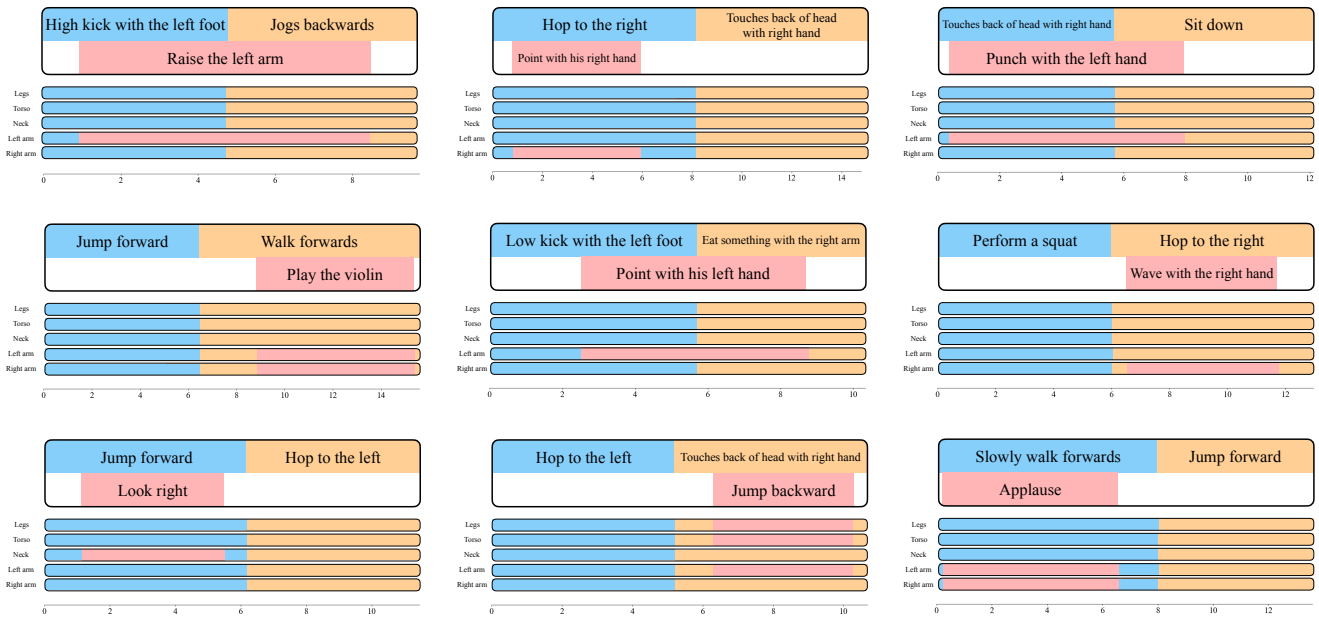

Figure A.2. **Example timelines from MTT dataset:** We display several generated timelines, along with the automatically generated body part timelines. Although each timeline contains only three prompts, the generated timelines are diverse and specify complicated motions.

D. MDM-SMPL Additional Details

Pose representation. As presented in Sec. 3.4 of the main paper, we propose a motion representation for diffusion that includes SMPL pose parameters. We represent a pose $\boldsymbol{x} \in \mathbb{R}^d$ by $\boldsymbol{x} = [r_z, \dot{r_x}, \dot{r_y}, \dot{\alpha}, \boldsymbol{\theta}, \mathbf{j}]$ where r_z is the Z (up) coordinate of the pelvis, $\dot{r_x}$ and $\dot{r_y}$ are the linear velocities of the pelvis, $\dot{\alpha}$ is the angular velocity of the Z angle of the body, $\boldsymbol{\theta}$ are the SMPL [7] pose parameters (encoded with the $6D$ representation [10]), and $\mathbf{j}$ are the joints positions (computed with the SMPL layer). Inspired by Holden et al. [6] and Guo et al. [5], which use a rotation invariant representation, we represent the joints $\mathbf{j}$ in a coordinate system local to the body. To make $\boldsymbol{\theta}$ local to the body, we remove the Z rotation from the SMPL global orientation. This representation enables us to directly extract the SMPL pose parameters, eliminating the need for optimization-based methods typically used to generate a mesh, as in previous work [2, 11].

Architecture and training. We use a similar architecture as MDM [8], but make the following changes (in addition to using the SMPL body):

- We use a cosine schedule as introduced by [3] with 100 steps instead of a linear schedule with 1000 steps.
- After padding to the maximum duration in a batch, we mask the padded area in the Transformer encoder so that the padded area is not used for the computation.

Other minor changes include using two separate tokens for the diffusion step t and the text embedding (instead of one), using two register tokens (introduced in [4]), and pre-computing CLIP embeddings for faster training. We train the model for 10000 epochs with a batch size of 128.

Evaluation on HumanML3D. We evaluate the performance of MDM-SMPL for single text motion generation (on the HumanML3D benchmark [5]). The FID is 0.38 (better than MDM [8] 0.54 and MotionDiffuse [9] 0.63), @R3 is 0.74 (between MDM 0.61 and MotionDiffuse 0.78), and the diversity is 9.67 which is also close the GT (9.5). This suggests that the synthesis quality does not deteriorate when we use MDM-SMPL instead of MDM or MotionDiffuse.

E. Additional Experiments

Varying the overlap size. As outlined in Sec. 4.3 of the main paper, we also experiment with varying the size of the overlap for temporal stitching (corresponding to $2*l$ in the paper) and display the results in Tab. A.1. We find that a smaller overlap size results in a higher transition distance. This means that the transitions may be more noticeable. However, it also leads to a more accurate match of each crop with its corresponding description, as indicated by higher per-crop semantic correctness metrics. With a larger overlap size, the transitions become smoother (i.e., lower transition distance), but this comes at the cost of reduced per-crop semantic correctness metrics.

Total overlap (s)	Per-crop semantic correctness				Realism	
	R@1 ↑	R@3 ↑	TMR-Score ↑ M2T	TMR-Score ↑ M2M	FID ↓	Transition distance ↓
0.25	30.1	51.7	0.675	0.666	0.459	1.0
0.4	29.9	51.1	0.675	0.666	0.459	1.0
0.5	30.5	50.9	0.675	0.665	0.459	0.9
0.6	30.3	50.8	0.674	0.665	0.459	0.9
0.75	28.9	50.4	0.672	0.664	0.460	0.9
1.0	28.5	49.1	0.670	0.662	0.459	0.9
1.25	28.9	48.6	0.668	0.660	0.458	0.9

Table A.1. **Influence of the overlap size**: We report the performance of STMC (with MDM-SMPL) while varying the total overlap size ($2 * l$). We observe that a smaller overlap size leads to a higher transition distance but each crop matches the description better (higher per-crop semantic correctness metrics). We observe the opposite for a larger overlap size.

Evaluation of individual sub-motions. We experiment with generating a motion for each text independently (I), and compare to the crops from STMC generations (S). The *per-crop* FID realism metric is close between S/I: 0.579/0.582 for MDM, 0.451/0.504 for MotionDiffuse, which suggests that the synthesis quality has not deteriorated with STMC.

On the other hand, the semantic correctness results are: (S/I) @R1 25.3/36.9, @R3 45.7/67.3, M2T: 0.639/0.709 T2M: 0.631/0.673. (I) performs better for the retrieval metrics than (S). This is expected since (I) follows only a single text prompt (as opposed to multiple prompts simultaneously in STMC) and there is no need to generate transitions (as in STMC). To give an example of how this may affect semantic metrics for STMC, if we generate "raise the right hand" and "raise the left hand" at the same time, the retrieval metric may end up retrieving "raise both hands", instead of one of the two hands. In the MTT dataset, there is a probability of $2/3$ for a text to have an overlap with another text, therefore, this case happens often.

References

[1] Nikos Athanasiou, Mathis Petrovich, Michael J. Black, and Gül Varol. SINC: Spatial composition of 3D human motions for simultaneous action generation. In *International Conference on Computer Vision (ICCV)*, 2023. 1

[2] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *European Conference on Computer Vision (ECCV)*, 2016. 3

[3] Ting Chen. On the importance of noise scheduling for diffusion models. *arXiv:2301.10972*, 2023. 3

[4] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *arXiv:2309.16588*, 2023. 3

[5] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3D human motions from text. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[6] Daniel Holden, Jun Saito, and Taku Komura. A deep learning

framework for character motion synthesis and editing. *ACM Transactions on Graphics (TOG)*, 2016. 3

[7] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 2015. 3

[8] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *International Conference on Learning Representations (ICLR)*, 2023. 3

[9] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. MotionDiffuse: Text-driven human motion generation with diffusion model. *arXiv:2208.15001*, 2022. 3

[10] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[11] Xinxin Zuo, Sen Wang, Jiangbin Zheng, Weiwei Yu, Minglun Gong, Ruigang Yang, and Li Cheng. Sparsefusion: Dynamic human avatar modeling from sparse rgbd images. *IEEE Transactions on Multimedia*, 2021. 3