SINC: Spatial Composition of 3D Human Motions for Simultaneous Action Generation Supplementary Material

Nikos Athanasiou[*1] Mathis Petrovich[*1,2] Michael J. Black[1] Gül Varol[2]
[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany
[2]LIGM, École des Ponts, Univ Gustave Eiffel, CNRS, France
sinc.is.tue.mpg.de

This document provides additional details about our method and experiments. In particular, we evaluate our synthetic data approach on a recently proposed diffusion model [2] (Section A), elaborate on our GPT-based body-part annotation method (Section B), our synthetic data creation pipeline (Section C), and our proposed TEMOS score (Section D). We also provide additional quantitative evaluations (Section E).

Supplementary video. Along with this document, we provide a video, available on the project page, which includes visualizations of a sample of generated motions; these are difficult to convey in a static document. (i) We first briefly describe our goal, motivation, and method. (ii) We then introduce baselines and illustrate their failure modes. (iii) We provide qualitative comparisons against baselines, while highlighting limitations of the coordinate-based *APE* metric. (iv) Finally, we demonstrate the ability of our model to generalize to out-of-distribution input combinations, as well as combinations beyond pairs.

A. Additional experiment with diffusion models

To complement our study with the TEMOS model [6], here, we provide an additional experiment by training a more recent state-of-the-art architecture for text-conditioned motion generation. Specifically, we implement Motion Latent Diffusion (MLD) [2] with the same text input pipeline as our method (see Section 3.2). Since MLD applies the diffusion on the latent space, we extract a single latent vector per motion (using the TEMOS model trained on Real-singles as a feature extractor). We train the diffusion model for 1000 epochs on 2 GPUs, with a batch size of 16, and learning rate of 1e-4. Instead of the coordinate-based representation of Guo et al. [4], we directly train on 6D rotation representation (as is done for TEMOS, see Section 3.3). Apart from those adaptations, we use the same architectural choices as in the original paper [2]. In Table A.1, we report the results with and without synthetic data, as we did for TEMOS in the main paper with the rows 10 and 2 of Table 3, respectively. The same conclusion holds for MLD: the model trained on additional synthetic data demonstrates better performance than the one trained only on real data (Real-Pairs and Real-Singles).

Model	Synthetic training	TEMOS Score
MLD [2]	✗	0.612
MLD [2]	✓	**0.638**
TEMOS [6]	✗	0.640
TEMOS [6]	✓	**0.644**

Table A.1. **Additional results with a diffusion model:** We report the performance of MLD [2] with and without adding the synthetic training data. We observe that synthetic data helps for both MLD and TEMOS.

B. Body Part Labeling with GPT-3

BABEL includes 6518 unique language labels for training and validation. We use these raw labels as input in the GPT-3 query. We prompt the public API https://openai.com/api/ for each of the BABEL action labels and automatically retrieve the body parts that are involved in the motion. We experimented with various prompts before deciding on our final prompt template. We observed that GPT-3 outputs are easier to parse and map to our predefined list of body parts if we provide this list, as well as few-shot examples consisting of question-answer pairs. We use the following prompt, to extract the body part annotations for our synthetic data creation, as described in Section 3.1:

```
The instructions for this task are to choose
your answers from the list below:

left arm
right arm
left leg
buttocks
waist
right leg
torso
```

[*] Equal contribution

```
11 neck
12
13 Here are some examples of the question and answer
14 pairs for this task:
15
16 Question: What are the body parts involved in the
17 action of: walk forwards?
18 Answer: right leg
19 left leg
20 buttocks
21
22 Question: What are the body parts involved in the
23 action of: face to the left?
24 Answer: torso
25 neck
26
27 Question: What are the body parts involved in the
28 action of: put headphones over ears?
29 Answer: right arm
30 left arm
31 neck
32
33 Question: What are the body parts involved in the
34 action of: sit down?
35 Answer: right leg
36 left leg
37 buttocks
38 waist
39
40 Question: What are the body parts involved in the
41 action of: [ACTION]?
```

Listing 1. **GPT prompt template**

Listing 1 shows the full prompt used to extract the annotations using GPT-3 for composing actions spatially. In Table 1, of the main paper, we quantitatively evaluated the body part labeling performance of this prompt, along with alternative prompts. Here, in Table A.2, we provide qualitative examples to illustrate the behavior of GPT-3 to each of the prompt types. (a) "Free-form" prompt type contains only L40-41 from Listing 1. (b) "Choosing from a list" contains both L1-11, L40-41. (c) "Choosing from a list + Few-shot examples" refers to the full prompt. As shown in Table A.2, using "Free-form" prompting requires a tedious post-processing of GPT-3 responses, since one needs a comprehensive mapping from all possible body part namings to our list. Moreover, the level of details is not consistent across actions (e.g., 'left leg and hips' versus 'deltoid and triceps muscles'). We extract the associated body parts by detecting keywords from a manually constructed lookup table; however, the labeling accuracy based on Table 1 of the main paper is still lower than instructing GPT-3 to choose from a list. We obtain further gains by including few-shot examples in the prompt. This is demonstrated qualitatively in Table A.2 for the label 'rotate shoulders' which GPT-3 includes neck in addition to torso or 'walk backwards with arms attach to the waist' for which arms are mistakenly omitted for the "Choose from a list" prompt. Our final prompt that provides both the list and few-shot examples perform best, while also requiring significantly less post-processing.

We explain the reasoning behind replacing 'global orientation' with 'waist' and 'buttocks' in the list of body parts. In our initial prompts we used 'global orientation' as part of the list. However, we observed that the model frequently returned 'waist' and 'buttocks' even when they were not in the list.

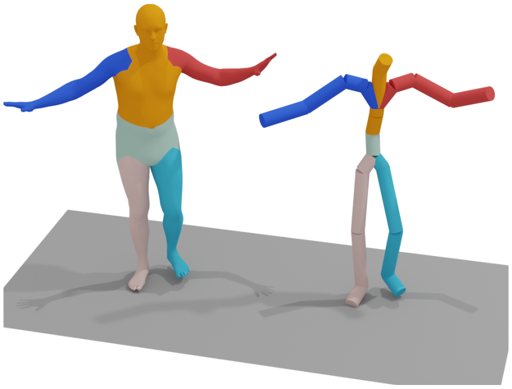

Figure A.1. **Body parts:** Each color indicates a different body part. Vertices (left) and the skeleton (right) are extracted from the SMPL body model.

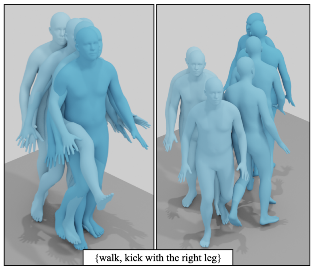

Figure A.2. **Testing incompatible simultaneous actions:** We apply our model SINC on the input{'walk', 'kick with the right leg'} which represents an example of two incompatible actions due to involving the same body part 'right leg'. We display two random generations from our model, once with 2-second duration (left), and once with 4 seconds (right). We observe that SINC generates one of the two actions in each sample ('kick' on the left, 'walk' on the right).

Furthermore, GPT-3 responses included 'global orientation' even in cases when it was not necessary e.g., 'lift arm', 'raise leg'. Consequently, we chose to remove 'global orientation', and add 'waist' and 'buttocks' instead.

Finally, we include the label 'neck' in addition to 'torso', since GPT-3 tends to include 'neck' in its responses, especially when we prompt for the actions: 'look left' / 'look right'.

We use 6 body part labels based on common body segmentation maps [5], which we show in Figure A.1. Since most of the AMASS dataset does not contain fine-grained hand motions, we do not include hands as a separate body part. Furthermore, we observe that GPT-3 behavior may become inconsistent if we provide a long list of fine-grained parts instead of few coarse labels. The main body parts include the right/left extremities, the torso-neck part and the pelvis denoted as global. Global, except for the pelvis rotation, includes global translation of the body and it is used when either the waist or buttocks are included in GPT-3's response.

Action	Prompt Type	GPT-3 Response
move right arm in circular motion	Free-form	The person’s right arm, shoulder, and possibly the upper part of their body.
	Choosing from a list	right arm
	Choosing from a list + Few-shot examples	right arm
make large circles with left leg in front of body	Free-form	The left leg and the hips
	Choosing from a list	left leg
	Choosing from a list + Few-shot examples	left leg
overhead throw	Free-form	The deltoid muscle in the shoulder and the triceps muscle in the arm are moving when someone is doing an overhead throw.
	Choosing from a list	left arm right arm
	Choosing from a list + Few-shot examples	left arm right arm torso
walk backwards with arms attach to the waist	Free-form	The body parts involved in the action of walking backwards with arms attached to the waist are the legs, arms, back, and abdomen.
	Choosing from a list	right leg left leg buttocks
	Choosing from a list + Few-shot examples	left arm right arm left leg right leg waist
put down bottle with left hand	Free-form	Left arm Left hand Fingers
	Choosing from a list	left arm
	Choosing from a list + Few-shot examples	left arm torso
rotate shoulders	Free-form	The body parts involved in the action of rotating the shoulders are the neck, shoulders, arms, and back.
	Choosing from a list	left arm right arm arm torso neck
	Choosing from a list + Few-shot examples	left arm right arm arm torso

Table A.2. **GPT response examples for different prompt types:** We show the responses of GPT-3 on some examples that demonstrate the differences between different prompt types (see Table 1 of the main paper). The output of the free-form prompt is non-trivial to parse and map to our list of body parts. On the other hand, providing the list and few-shot examples encourages GPT-3 to follow a more strict format, and to describe the body parts with the same words as in our list.

C. Synthetic Data Creation

We use GPT-3-guided spatial compositions in two parts of this work. First, we use GPT-3 to benchmark how well a single-action baseline can perform, by applying composition as post-processing on independently generated motions (Figure A.3 bottom). Secondly, we use GPT-3 to create synthetic data to train our model. In both cases, we employ the method described in Section 3.1 of the main paper. We use the heuristic of stitching the motion with less body parts (motion B) on top of the other motion (motion A), because the body parts of motion B are more likely to be local (as in “waving the right hand”) and important for keeping the semantic of the motion. On the other hand, motion A is more likely to be a global motion (as in “walking” or “sitting”) and grafting motion B onto motion A usually produces a realistic motion and preserves the semantics of both motions. Note that these heuristics were determined based on visual inspection over several examples, and may not be optimal.

The difference in the case of synthetic data creation is the compatibility test, which makes sure that no body part is involved in both of the motions being composited. Moreover, synthetic data combines existing real motions, and the single-action baseline combines generated motions.

We only apply the compatibility check for the synthetic data generation to avoid composing invalid motions, since a human can physically not perform two actions with the same body part in most cases. This choice was simply to ensure better synthetic data quality, as without it, the composition may be reduced down to one action (e.g., ‘walking’ would overwrite ‘kicking’ as the leg cannot do both). At test time, when we query ‘walk’ and ‘kick with the right leg’ with two different durations, SINC randomly generates one of the two actions, as seen in Figure A.2.

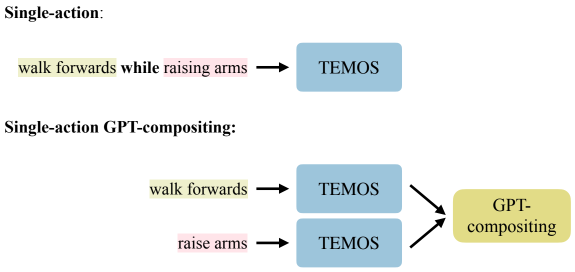

Figure A.3. **Single-action baselines:** For both baselines, TEMOS is trained on Real-Singles of BABEL. On the top, we concatenate the textual inputs by adding the word “while” in between actions. On the bottom, we generate the two actions independently and combine them with the body part guidance from GPT-3.

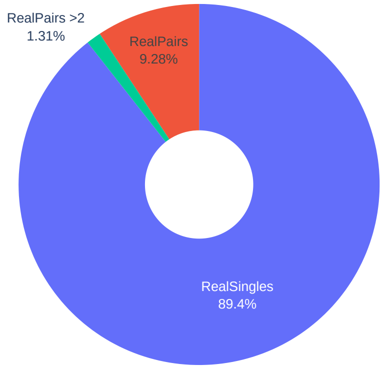

Figure A.4. **Distribution of the training set:** The simultaneous Real-Pairs are the vast minority of the data, highlighting the importance of automatically enriching training data through our synthetic spatial compositions.

D. TEMOS Score

The position-based metrics typically used in prior work [1, 3, 6] compare generated motions with the ground-truth

Conjuction Word	**Seen during training**	**Model**	TEMOS ↑ score	Average Positional Error ↓				Average Variance Error ↓			
				root joint	global traj.	mean local	mean global	root joint	global traj.	mean local	mean global
`while`	✓	Single-action	0.601	0.592	0.551	0.286	0.712	0.076	0.075	0.013	0.083
		SINC	**0.644**	0.493	0.463	0.266	0.616	0.066	0.065	0.012	0.072
`during`	✓	Single-action	0.598	0.629	0.587	0.284	0.752	0.085	0.084	0.013	0.093
		SINC	**0.642**	0.497	0.471	0.261	0.622	0.065	0.063	0.012	0.071
`and ... at the same time`	✓	Single-action	0.599	0.607	0.568	0.283	0.722	0.084	0.083	0.014	0.092
		SINC	**0.643**	0.495	0.468	0.264	0.620	0.065	0.064	0.012	0.072
`in parallel`	✗	Single-action	0.600	0.611	0.570	0.294	0.736	0.081	0.081	0.012	0.089
		SINC	**0.643**	0.583	0.555	0.266	0.704	0.074	0.072	0.012	0.080
`whilst`	✗	Single-action	0.599	0.551	0.511	0.288	0.670	0.073	0.072	0.012	0.080
		SINC	**0.644**	0.491	0.461	0.262	0.614	0.066	0.065	0.012	0.072
`synchronously`	✗	Single-action	0.596	0.520	0.476	0.294	0.644	0.074	0.072	0.013	0.081
		SINC	**0.637**	0.520	0.492	0.261	0.644	0.0644	0.0632	0.011	0.070

Table A.3. **Evaluation using different conjunction words:** In Table 2 of the main paper, we evaluated the models with the conjunction word `while`. Here, we report performance when joining the two actions using other conjunction words, for both seen and unseen conjunction words during training. We observe similar trends for the TEMOS scores and the positional metrics as for using `while` to join the actions. Overall, performance of Single-action methods remains significantly inferior, especially for the TEMOS score. Note that SINC refers to our best model which is trained on both Real Singles, Real Pairs and Synthetic Pairs.

motion in the coordinate space local to the body: they measure differences of positions and do not take into account semantics. Here are four types of examples where the metrics can fail: 1) with a cyclic motion such as "walking", the generation can be out of phase with the ground truth and still be semantically valid; 2) even for a non-cyclic motion such as "throwing an object", the timing can be different and can lead to bad scores on common metrics; 3) if the input text description is ambiguous such as "kick" (where the motion can be done from one leg or the other), the metrics may not reflect the quality of the generated motion; 4) if the motion demonstrates severe foot sliding or body translation artifacts, the error may be dominated by the translation error, effectively ignoring the overall implausibility of the limb motion e.g., feet not moving.

To avoid these issues, we introduce another performance measure called *TEMOS score*. We train a TEMOS model on BABEL Real-Singles for 1000 epochs, freeze its weights, and use its motion encoder component. Then, we extract features by feeding a motion B to the motion encoder, and use the mean of the distribution as the feature vector f. This feature captures the semantics of the motion as the motion space has been trained to explicitly model motion-text matching, i.e., cross-modal embedding space.

To calculate the TEMOS score, we feed the ground truth and the generated motions to the motion encoder, and extract the feature vectors f_{GT} and f_{motion}, respectively. Then we compute the score based on their cosine similarity as follows:

$$\text{TEMOS score}(f_{GT}, f_{motion}) = \frac{1}{2}\left(1 + \frac{f_{GT} \cdot f_{motion}}{||f_{GT}|| \cdot ||f_{motion}||}\right).$$

The range of this score is between 0 and 1, with a maximum at 1, which occurs when the two motions are identical.

	Model used for TEMOS score	
	Single-action	**SINC**
Single-action	0.601	0.594
SINC	0.644	0.637

Table A.4. **TEMOS score with various TEMOS models:** We report performance using different trained models to compute the TEMOS score. While the absolute score slightly differs when measured with a different model (e.g., 0.644 vs 0.637), the relative ranking of the models we compare remains the same.

	Div. →	Multimod. ↑
SINC	1.10	1.13
Real	1.34	-

Table A.5. **Diversity evaluation:** We report the diversity and multimodality metrics of [4] for our SINC model.

E. Additional Quantitative Evaluation

We report quantitative results when evaluating with various conjunction words (Section E.1), when using various TEMOS models to compute the TEMOS score (Section E.2), when evaluating the diversity and multimodality metrics (Section E.3), and, when evaluating on the full validation set for completeness (Section E.4).

E.1. More conjunction words

In our main paper experiments, we used `while` as our conjunction word. For completeness, in Table A.3 we evaluate the Single-action method and our best model with other conjunction words at test time. We observe that the differences are minimal and the methods perform similarly across different conjunctions. This is true for all conjunctions both seen and unseen during training. The performance is similar, likely due to

Model	Tr. Data		TEMOS ↑	Average Positional Error ↓				Average Variance Error ↓			
	Real-P	Real-S	score	root joint	global traj.	mean local	mean global	root joint	global traj.	mean local	mean global
Single-action	✗	✓	0.607	0.516	0.483	0.262	0.626	0.067	0.066	0.012	0.073
Single-action GPT-compositing	✗	✓	0.626	0.458	0.431	0.244	0.569	0.068	0.067	0.011	0.074
SINC-STE	✓	✗	0.630	0.502	0.477	0.249	0.616	0.074	0.074	0.010	0.08
SINC	✓	✗	0.634	0.602	0.586	0.243	0.704	0.084	0.083	0.011	0.091
SINC	✓	✓	**0.645**	0.519	0.495	0.248	0.632	0.078	0.077	0.010	0.084

Table A.6. **Baseline comparison on the full validation set of BABEL:** We observe similar trends with the filtered validation set reported in Table 2 of the main paper.

Synthetic data	Training Data			TEMOS ↑	Average Positional Error ↓				Average Variance Error ↓			
	Real-P	Real-S %	Synth-P %	score	root joint	global traj.	mean local	mean global	root joint	global traj.	mean local	mean global
N/A	✓	0	0	0.634	0.602	0.586	0.243	0.704	0.084	0.083	0.011	0.091
	✓	100	0	0.645	0.519	0.495	0.248	0.632	0.078	0.077	0.010	0.084
Random composition	✗	50	50	0.551	0.575	0.534	0.259	0.664	0.072	0.071	0.011	0.078
	✗	0	100	0.552	0.454	0.411	0.263	0.551	0.068	0.067	0.011	0.074
	✓	50	50	0.619	0.396	0.362	0.242	0.504	0.060	0.059	0.010	0.067
	✓	0	100	0.619	0.422	0.390	0.241	0.530	0.062	0.061	0.010	0.068
GPT composition	✗	50	50	0.554	0.641	0.604	0.262	0.731	0.074	0.073	0.011	0.081
	✗	0	100	0.632	0.424	0.405	0.237	0.543	0.055	0.054	0.011	0.062
	✓	50	50	**0.651**	0.418	0.397	0.234	0.533	0.055	0.054	0.010	0.062
	✓	0	100	**0.645**	0.472	0.453	0.237	0.581	0.053	0.053	0.010	0.060

Table A.7. **Contribution of the synthetic data on the full validation set of BABEL:** We complement Table 3 of the main paper, by reporting on the full validation set (without any filtering).

the text embeddings mapping the expressions to similar points.

E.2. TEMOS score with various TEMOS models

As mentioned in Section 4.1 of the main paper, to report the TEMOS score, we use a TEMOS model trained on Real-Singles of BABEL. Here, we analyze whether the choice of the TEMOS model has a large impact on the results when trained on pairs. In Table A.4, we observe that the TEMOS score trend is similar when computed with TEMOS models trained on Real-Singles (Single-action) or on all real and synthetic data (SINC).

E.3. Diversity

Following Guo et al. [4], we report the overall diversity (for all action pairs), and multimodality (i.e., per-action-pair diversity) in Table A.5. We measure the L2 distance between the TEMOS embeddings of two sets of generations. For multimodality we sample 20 generations per description, and for diversity we generate 5 samples per description. Both metrics are computed for 300 random descriptions from the BABEL validation set. Real motions do not contain a sufficient number of motions for each action pair, thus the reason for omitting their multimodality.

E.4. Full validation set

As explained in Section 4.1 of the main paper, we report all the results on a challenging subset of the validation set (i.e., without the action 'stand', and using only unseen examples). Here, we provide the results on the full validation set for completeness. In particular, we repeat the Tables 2 and 3 of the main paper, in Tables A.6 and A.7. As expected, we observe slightly improved results overall on this 'easier' validation set and the conclusions remain similar to the comparison in the main paper.

References

[1] Nikos Athanasiou, Mathis Petrovich, Michael J. Black, and Gül Varol. TEACH: Temporal action compositions for 3D humans. In *International Conference on 3D Vision (3DV)*, 2022. 3

[2] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, Jingyi Yu, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 1

[3] Anindita Ghosh, Noshaba Cheema, Cennet Oguz, Christian Theobalt, and Philipp Slusallek. Synthesis of compositional animations from textual descriptions. In *International Conference on Computer Vision (ICCV)*, 2021. 3

[4] 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. 1, 4, 5

[5] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. In *International Conference on Computer Graphics and Interactive Techniques in Asia (SIGGRAPH ASIA)*, 2015. 2

[6] Mathis Petrovich, Michael J. Black, and Gül Varol. TEMOS: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision (ECCV)*, 2022. 1, 3