

# SceMoS: Scene-Aware 3D Human Motion Synthesis by Planning with Geometry-Grounded Tokens

## –Appendix–

Anindita Ghosh<sup>1,2,3</sup> Vladislav Golyanik<sup>2,3</sup> Taku Komura<sup>4</sup> Philipp Slusallek<sup>1,3</sup>

Christian Theobalt<sup>2,3</sup> Rishabh Dabral<sup>2,3</sup>

<sup>1</sup>DFKI <sup>2</sup>MPI for Informatics <sup>3</sup>Saarland Informatics Campus <sup>4</sup>The University of Hong Kong

### A. Data Pre-processing

We train and evaluate our model on the TRUMANS dataset [2]. As a pre-processing step, we align the human motion data with the corresponding text annotation data, and then downsample the motion data to 20 fps. Following the original data split, we have  $\sim 1269$  sequences in the training set, and  $\sim 408$  sequences in the test split. Each sequence varies from 655 to 1100 frames depending on the length of motion. Since it is difficult to train on such long sequences, we generate subsequences of 80 frames by using a sliding window on the full sequence during data loading. We then transform the positions and orientations of the person such that their root joint is at the global origin and the body faces the  $Z_+$  direction. The associated scene geometry is transformed accordingly to maintain spatial consistency. We  $Z$ -normalize all the motion features before feeding them to the network. For the bird’s-eye-view (BEV) scene representation, we use Blender to position a virtual camera at a corner of the room with maximal visibility and render a BEV snapshot of the environment.

### B. User Study

**Setup.** To assess the visual quality of our results from a human-centered perspective, we conducted a user study. Participants were presented with a sequence of 20 short motion clips generated by different methods (ours, Trumans [2], TeSMo [5], and Humanise [4]) and rated each clip along two independent axes: Realism and Semantic Alignment. As illustrated in Fig. A.1, the interface presents (i) the text prompt, (ii) a rendered visualization of the generated motion interacting with the scene, and (iii) two 5-point Likert scales. ‘Realism’ measures the physical plausibility of the motion within the scene (*i.e.*, presence of foot skating, penetrations, incorrect contacts, or implausible body configurations), while ‘Semantics’ measures how well the motion matches the provided instruction irrespective of physi-

cal artifacts. The five rating categories were “Poor”, “Bad”, “Average”, “Good”, and “Excellent”.

**Result.** We collected responses from 41 participants, drawn from a mix of graduate researchers and industry practitioners. Each participant evaluated all 20 clips, the interface required both ratings to be filled to prevent incomplete submissions. Tab. A.1 reports the resulting mean realism and semantic scores across all methods. Ground truth achieves the expected upper bound. Among the generative methods, SceMoS yields the best perceptual performance (Realism: 3.41, Semantics: 4.20), outperforming Trumans [2], TeSMo [5], and Humanise [4]. Users consistently rated our motions as semantically faithful, even in scenes with challenging geometric layouts. This trend is further supported by the rating distribution shown in Fig. A.2. SceMoS receives the fewest “Poor / Bad” ratings and over 60% of its ratings fall into the “Good / Excellent” categories, indicating strong perceptual stability across different scene contexts.

Table A.1. User study evaluation across all methods. We report the mean  $\pm$  standard deviation of participant ratings for *Realism* and *Semantics* in a 5-point Likert scale.

Metric	GT	SceMoS	Trumans	TeSMo	Humanise
Realism	4.36 $\pm$ 0.1	3.41 $\pm$ 0.2	3.38 $\pm$ 0.1	2.19 $\pm$ 0.4	1.83 $\pm$ 0.3
Semantics	4.63 $\pm$ 0.2	4.20 $\pm$ 0.6	3.65 $\pm$ 0.7	3.79 $\pm$ 1.3	3.41 $\pm$ 1.4

### C. Additional Quantitative Experiments

**Additional Metrics.** To also assess how well the motions generated from SceMoS adhere to semantic intent and spatial targets, we evaluate R-Precision (Top-1) [1] and Goal Accuracy on the TRUMANS test set. R-Precision evaluates motion-text alignment by measuring the retrieval accuracy of the ground-truth prompt from a set of distractors. Goal Accuracy is defined as the percentage of sequences where the character successfully navigates to within 0.5m

Page 1/7

**Realism:** how well does the motion suit in the scene?  
When evaluating realism, consider artifacts such as object penetrations, improper contacts, and implausible configurations and motions of the virtual character or objects it is interacting with. Please do not consider whether the motion matches the text prompt.

**Semantics:** how well does the motion follow the text instruction?  
When evaluating semantics, consider only the prompt relevance. For example, the motion for "sit on the chair" should have the sit action, and the interacting object should be a chair. Please do not consider motion artifacts, such as object penetrations, improper contacts, and implausible configurations and motions of the virtual character or objects it is interacting with.

**Video 01**  
"Walk to the dressing table".

**Realism \***

☐ Poor: severe motion artifacts

☐ Bad: mostly severe and only a few minor motion artifacts

☐ Average: mostly minor motion artifacts

☐ Good: very few minor motion artifacts

☐ Excellent: no motion artifacts - as if motion is captured from a real person

**Semantics \***

☐ Poor: motion is not relevant to the prompt at all

☐ Bad: motion is barely relevant to the prompt

☐ Average: motion partially follows the prompt

☐ Good: motion almost entirely follows the prompt

☐ Excellent: motion completely follows the prompt

Figure A.1. Interface of our user study where we ask participants to rank the motion clips based on ‘realism’ and ‘semantics’, in a 5-point Likert scale.

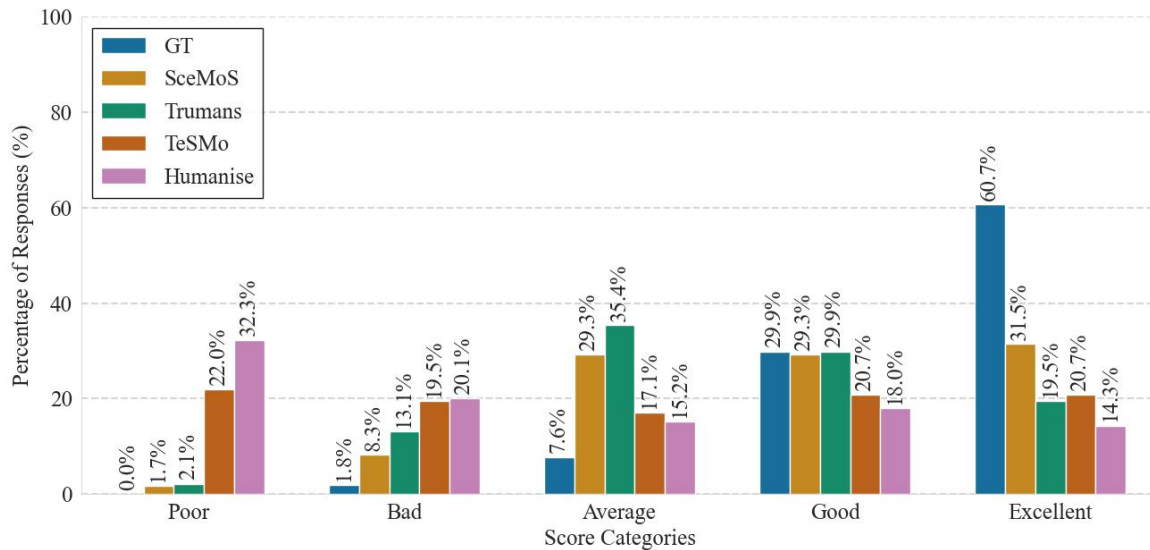


Figure A.2. User rating distribution across score categories. SceMoS receives the fewest “Poor / Bad” ratings and over 60% of its ratings fall into the “Good / Excellent” categories.

of the target object’s location. As shown in Tab. A.2, SceMoS outperforms baselines in both metrics, confirming that our geometry-grounded tokens effectively capture both semantic intent and precise spatial targeting.

Table A.2. **Extended quantitative evaluation.** Left: Task-completion metrics on TRUMANS dataset (Goal Accuracy and R-Precision). Right: Generalization performance on the HUMANISE dataset.

Method	TRUMANS		HUMANISE		
	Goal Acc. $\uparrow$	R-Prec. $\uparrow$	FID $\downarrow$	Div. $\rightarrow$	Cont. $\uparrow$
SceMoS	<b>0.62</b>	<b>0.58</b>	<b>0.95</b>	1.33	<b>0.74</b>
Trumans	0.59	0.57	1.02	1.65	0.74
TeSMo	0.49	0.34	1.17	1.21	0.72
Humanise	0.28	0.31	0.99	<b>1.22</b>	0.72

**Testing on Additional Datasets.** To evaluate the generalization capability of our framework, we tested SceMoS on a subset of the HUMANISE test set [4] without any architectural changes. We observe competitive performance in terms of diversity and contact metrics (see Tab. A.2), demonstrating that our 2D scene representation generalizes effectively across different indoor scene distributions.

**Backbone Compatibility.** We also validated that our model is compatible with newer backbones such as DI-

NOv3 [3]. In an initial convergence analysis limited to 500 iterations, we observed that DINOv3 matches the convergence rate of our default DINOv2 encoder, with the training loss dropping from 10.0 to 3.5 ( $\pm 1.65$ ). This confirms that our framework is compatible with other dense visual backbones and is not strictly tied to the specific DINOv2 architecture.

## References

- [1] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1
- [2] Nan Jiang, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Scaling up dynamic human-scene interaction modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1
- [3] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025. 3
- [4] Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. HUMANISE: Language-conditioned human motion generation in 3d scenes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1, 3
- [5] Hongwei Yi, Justus Thies, Michael J. Black, Xue Bin Peng, and Davis Rempe. Generating human interaction motions in scenes with text control. In *European Conference on Computer Vision (ECCV)*, 2024. 1