

SemGeoMo: Dynamic Contextual Human Motion Generation with Semantic and Geometric Guidance

Supplementary Material

We present additional details of implement and LLM annotation in Sec 1. Additional experiments are provided in Sec 2, including the user study, generation results after first stage, more ablation studies and visualization results.

1. Additional Details.

1.1. Implement details.

We implemented our model using Pytorch with training on NVIDIA A40 GPU. Batch size $B = 16$. We use AdamW optimizer, and the learning rate is 0.0002. The dimension of human motion $D = 263$, followed MDM [6], and point cloud is downsampled to $N = 1024$. The text feature dimension F_{text} encoded from CLIP [5] and LONGCLIP [8] is 512. We initialize the parameters of both the original MDM and Motion ControlNet from the pretrained MDM weights, freezing the parameters of the original MDM during training. The multi-head attention in transformer for extracting latent features is configured with $n_{head} = 4$.

For the dataset implementation, IMHD² and HoDome are previously released for motion capture tasks. We extend these two human-object interaction datasets for the contextual generation task. Specifically, we divide the original long motion sequences into shorter clips of length $L = 100$, with a frame rate of 30 FPS. Each clip is annotated with a corresponding text description with our LLM annotation module. For IMHD², we establish a new benchmark by dividing the dataset into training and testing sets based on different subjects, with 1000 clips for training and 800 clips for testing. For HoDome, we manually remove parts of the dataset with poor object reconstruction quality. The remaining data serves as unseen samples to evaluate the generalization capabilities of the model.

1.2. LLM annotation details.

We further provide the details of LLM Annotation in Tab. 4 and Tab. 5. We also provide several generated textual descriptions in Fig. 3, illustrating the texts are well-aligned with the corresponding generated motions, enhancing the interpretability and coherence of the results.

2. Additional Experiments

2.1. User study.

We conduct a user study to further evaluate our approach. Specifically, we generated 30 sequences for each method on FullManipulationBody and ask ten participants to rate

Table 1. User study for human motion generation result on Full-BodyManipulation.

	Model	Contact Score	Rationality	Total Score
w/o text	SceneDiff [2]	1.53	2.12	1.83
	OMOMO [4]	4.11	3.63	3.84
w GT text	CHOIS [3]	3.39	3.57	3.39
	MDM-PC [6]	3.07	3.64	3.35
	AffordMotion [7]	1.81	2.71	2.23
w Gen text	SemGeoMo	4.61	4.28	4.45

the results of each interaction, the score range from 1 to 5, higher is better. The evaluation was based on two dimensions: contact score, which measures the feasibility and stability of contact, and physical rationality, which evaluates whether the motions are physically plausible (e.g., avoiding foot slippage or unnatural distortions). Notably, human actions can occasionally exhibit slight distortions even when the contact score is relatively high. Our algorithm addresses this trade-off by incorporating multi-level semantic and geometric constraints and loss guidance to maintain a balance. The results of the study are presented in Tab. 1. Our method outperforms others, achieving a higher contact score and greater rationality, with the help of the integration of semantic and geometric information in our design.

2.2. Additional experiment results.

We further provide the metrics from the first stage SemGeo Hierarchical Guidance Generation. We observe that during interactions, the hand is the primary joint in contact with the object. Therefore, we predict the hand joint and incorporate hand-joint specific guidance in the second stage to enhance the motion generation. To evaluate generation performance in the first stage, we use Hand_JPE, which measures the difference between the predicted hand joint position and the ground truth, and Cosine Similarity measures the quality of the generated affordance map. Experiment result is shown in Tab. 3. The objects in FullBodyManipulation [4] are larger compared to other datasets, resulting in more diverse affordance areas, which reduces the similarity between the generated results and the ground truth. And the interactions in IMHD² [9] are more challenging, making the generation process harder and leading to relatively lower accuracy in joint predictions.

2.3. Ablation studies on other baseline.

We further conduct experiments on OMOMO [4] to validate the effectiveness of incorporating textual information and the impact of loss guidance, denoted as OMOMO-Text and OMOMO-Loss. After integrating language informa-

Table 2. Ablation study on different baselines.

joint constrain	HandJPE↓	MPIPE↓	C_{prec} ↑	C_{rec} ↑	C_{acc} ↑	c%↑	F1↑	FID↓	R-score↑	Diversity↑	FS↓
OMOMO	33.18	18.06	0.77	0.71	0.74	0.61	0.75	1.98	0.38	8.99	0.50
OMOMO-Text	30.52	17.53	0.78	0.72	0.76	0.62	0.76	1.26	0.43	9.18	0.48
OMOMO-Loss	29.16	17.01	0.79	0.73	0.77	0.64	0.76	1.96	0.36	8.96	0.47

Table 3. Joint generation result from SemGeo Hierarchical Guidance Generation.

	Left_JPE	Right_JPE	Hand_JPE	Sim
FullBodyManipulation [4]	28.84	27.96	28.40	0.27
Behave [1]	28.94	27.06	28.00	0.15
IMHD ² [9]	30.49	40.26	35.38	0.52

Table 4. Detailed prompting example for LLM Annotation.

Detailed prompting example for LLM Annotation.

Instructions: You are an expert on the interaction between 3D human motion and object. A person will interact with a object, give me a sentence that how the person will interact with this object based on following information.

[start of Given Information]

Coordinate System:

The coordinate system of the 3D scene includes x, y, and z-axes. The person moves on the XOZ plane, and the positive y-axis represents height.

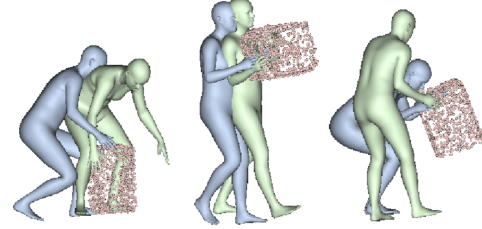
Target Object and category:

The category of the object is CLASS. The size of the CLASS is SIZE.

The interaction with this object will last approximately T seconds. The object center: CENTER.

Possible actions list: ACTION_LIST = ['face', 'flip', 'grab', 'grasp', 'hold', 'kick', 'lift', 'move', 'pick', 'place', 'push', 'pull', 'put', 'release', 'rotate', 'set', 'slide', 'swing', 'tilt', 'turn', 'sit', 'bend', 'shake', 'wave', 'drag']

[End of Given Information]



A person lift the suitcase, move the suitcase, and put down the suitcase.

Figure 1. Visualization compared with ground truth, the potential interaction can be diverse. The person in green is the ground truth and the blue is our generation result.



The person raises the left leg and takes a step forward, with arms hanging at the sides of the clothes. the left leg of the person is bent, and the right leg is straight and standing. the left and right legs of the second rotate in a semi-circle to the right side. The arms slightly swing at the sides of the body. The left leg is bent, and the right leg is straight and standing.



One person holds the other person's left hand with their right hand. the other person places their right hand on one person's waist, while one person rests their left hand on the other person's right shoulder.

Figure 2. Extension on human-human generation.

tion, the FID score significantly improves, indicating that the generation results are more rational and coherent.

2.4. Additional visualization results.

Comparison with ground truth. We provide a sample demonstrating that, given a sequential point cloud, the potential interactions can be diverse. The ground truth interaction is not the only valid solution. While our generated motion may differ from the ground truth, it remains rational and contextually appropriate.

More generation results. We present additional generation results in Figure. 3 and include a detailed comparison and demonstration of the results in our accompanying video.

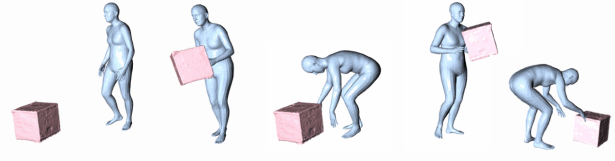
More extension results. We present extension generation results in Figure. 2 and demonstration of the results in our accompanying video.

References

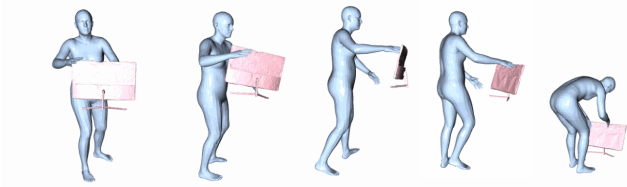
- [1] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15935–15946, 2022. 2
- [2] Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. Diffusion-



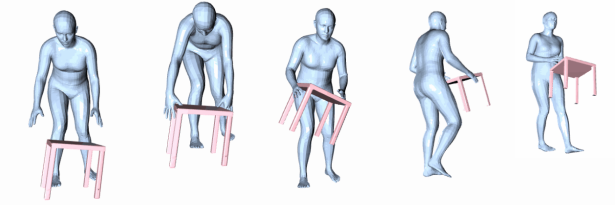
First, the person positions themselves in front of the clothes stand, facing its side. They reach out, placing **the right hand on the clothes stand**. As they bend their knees slightly, they lift the stand off the ground with a straight arm motion. Next, maintaining their grip on the stand, they pivot around its base, using their arms to lift the stand higher as they turn. Finally, they lower the stand, **set it back with the right hand** guiding it down from the **right-top position** as their legs step forward to reposition themselves.



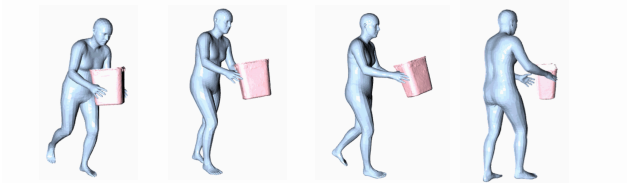
First, the person positions themselves facing the large box. He **reach out with both hands**, extending them from the sides, to grab the box. The person bends slightly at the waist, **lifting the large box** off the ground. Next, the large box is securely held between both hands. Finally, the person lowers the large box back to the ground. They guide the box down using **both hands**, with the left hand applying pressure from the left-top and the right hand providing support from the right-bottom, gradually **lowering the box** until it is back in place.



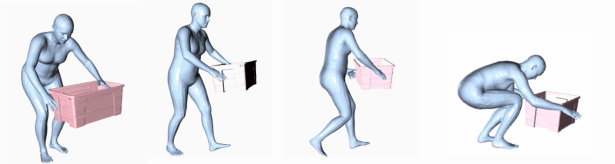
First, the person stands in front of the monitor, grasps it with **both hands from the left-bottom and right-top corners**. As they bend slightly at the waist, they lift the monitor off the surface. Next, maintaining grip, they move the monitor to their desired location, adjusting the height slightly with their hands. Finally, they **place the monitor down**, using their right hand to push from the **right-top** and their **left hand** to steady from the **left-bottom**.



First, the person faces the small table, extending **both arms** from the sides to grasp the edges. They bend their knees slightly, lifting the table off the ground with both hands. Next, maintaining the grip, the person rotates the small table with both hands. Finally, the person carefully puts down the small table, the **right arm supports the right side** of the table from the top, while the **left arm supports from the left side**.



First, the person approaches the trashcan, faces it from the front, and grasps it from the **left-bottom and right-bottom corners with both hands**. Knees bend slightly as they lift the trashcan from the ground. Next, maintaining the grip, the person lifts it slightly higher as they turn. Finally, the person sets the trashcan back down, moving their right arm to the right-top corner.



First, the person faces the plasticbox, grasps it from the left-bottom and right-bottom corners, **lifting** it off the ground with **both arms slightly bent**. Next, maintaining grip, the person moves the plasticbox to the side, keeping it elevated. Finally, the person **sets down** the plasticbox, using the right arm to push from the **right-top corner** and the left arm to steady from the **left-top**, while moving forward with both legs to reposition.

Figure 3. The generation result is aligned with the generated fine-grained text.

Table 5. **Detailed prompting example for fine-grained LLM Annotation.**

Detailed prompting example for LLM Annotation.
Instructions: A person lift the plasticbox, rotate the plasticbox, and set it back down.
You are an expert on the interaction between 3D human motion and object. Given the instruction, give me a sentence that how the person will interact with this object in detailed, including the arm and leg movement in each 3s, make each sentence just include key action.
[start of Given Information]
Coordinate System:
The coordinate system of the 3D scene includes x, y, and z-axes. The person moves on the XOZ plane, and the positive y-axis represents height.
Target Object and category:
The category of the object is CLASS. The size of the CLASS is SIZE.
The object center with hand contact information in total LAST.TIME
At T, object center is CENTER, [no hand contact, single contact hand(left/right), both hand in contact] at position POS.
[End of Given Information]
[Start of Rule]
Divide the total movement into three step.
Inference how their arms and legs move.
Inference the hand-object interaction direction, chosen from:“left”, “right”, “top”, “bottom”,“left-top”, “left-bottom”, “right-top”, and “right-bottom”.
Make sure each sentence includes key action.
[End of Rule]
[Start of Example]
A person lifts the white chair, rotates the white chair, and puts down the white chair.
The fine-grained result:
First, the person faces the back of the white chair, grasps it with both hands from the left-bottom and right-bottom sides, bending slightly at the knees as both arms lift the chair off the ground. Next, maintaining grip, the person rotates the white chair with both hands, lifting the chair slightly higher. Finally, the person puts down the white chair, with the right arm pushing from the right-top and the left arm steadying from the left-top, as both legs move forward to reposition.
[End of Example]

- [4] Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics (TOG)*, 42(6):1–11, 2023. 1, 2
- [5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1
- [6] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2022. 1
- [7] Zan Wang, Yixin Chen, Baoxiong Jia, Puhao Li, Jinlu Zhang, Jingze Zhang, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Move as you say interact as you can: Language-guided human motion generation with scene affordance. In *Proceedings of the IEEE/CVF CVPR*, pages 433–444, 2024. 1
- [8] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. In *ECCV*, pages 310–325. Springer, 2025. 1
- [9] Chengfeng Zhao, Juze Zhang, Jiashen Du, Ziwei Shan, Junye Wang, Jingyi Yu, Jingya Wang, and Lan Xu. I’m hoi: Inertia-aware monocular capture of 3d human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 729–741, 2024. 1, 2

- based generation, optimization, and planning in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16750–16761, 2023. 1
- [3] Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C Karen Liu. Controllable human-object interaction synthesis. *arXiv preprint arXiv:2312.03913*, 2023. 1