

# Goal-Driven Human Motion Synthesis in Diverse Tasks

## Supplementary Material

### 1. Preliminaries: Conditional Sampling from Diffusion Model

We describe the diffusion models from which we formulate conditional diffusion to generate adaptive motions with precise control. Diffusion models have shown remarkable performance in learning data distributions and effectively learning or sampling from conditional distributions. These models consist of two processes: a forward diffusion process and a reverse process. Denoting  $p_0(x_0)$  as the original data distribution, the forward process injects i.i.d. Gaussian noise to a data distribution  $x_t = x_0 + \sigma_t \epsilon$  with  $\epsilon \sim \mathcal{N}(0, I)$ , where  $x_0 \in \mathcal{X}$  is an instance within the dataset  $\mathcal{X}$  and  $\sigma_t$  monotonically increases with respect to time  $t \in [0, T]$ . In the reverse process, the model generates data samples from pure Gaussian noise by recursively sampling from a denoising model  $\mathcal{D}_\theta$ . The conditional denoising model  $\mathcal{D}_\theta$  depends on time  $t$  and the additional feature  $c$  in the input data. We use reconstruction loss during training and directly learn to predict the original data  $x_0$ :

$$\mathcal{L} = \mathbb{E}_{x_0, t, c} \|\mathcal{D}_\theta(x_t, t, c) - x_0\|_1. \quad (1)$$

The trained diffusion models can flexibly generate output satisfying the user-specified condition  $y$  with an appropriate guidance function. From the Bayes' rule, the conditional score can be calculated as follows:

$$\nabla_{x_t} \log p(x_t | y) = \nabla_{x_t} \log p(x_t) + \nabla_{x_t} \log p(y | x_t). \quad (2)$$

The first term from the right side can be obtained through a pre-trained denoising model  $\mathcal{D}_\theta$ , while the second term can be acquired by the gradient of an analytic guidance function  $G(x_t, y)$ . The guidance function evaluates how well the diffusion sample  $x$  satisfies the given condition  $y$ , and its gradient is calculated as  $-\nabla_x G(x_t, y)$ .

However, the diffusion sample  $x_t$  inherently contains additive noise, and simply computing the gradient of a function based on  $x_t$  can result in inaccurate gradients. DPS [2] proposes a formulation to find a more meaningful sample point  $\hat{x}_t$  to incorporate the gradient  $\nabla_{x_t} \log p(y | x_t) \simeq \nabla_{\hat{x}_t} \log p(y | \hat{x}_t)$ . We can additionally adopt Monte Carlo sampling in order to update the guidance function  $G$  more accurately [6]. The modified estimate of the guidance function  $G_{MC}$  is computed as follows:

$$G_{MC}(x_t, y) = -\log \left( \frac{1}{n} \sum_{i=1}^n \exp(-G(x^{(i)}, y)) \right), \quad (3)$$

where  $n$  is the number of the samples. And  $x^{(i)}$  are i.i.d. samples from  $\mathcal{N}(\hat{x}_t, r_t^2 I)$ , where  $r_t = \sigma_t / \sqrt{1 + (\sigma_t)^2}$ .

We utilize the guidance sampling method to design an analytic function that satisfies the given constraints, enabling precise control of the generation process. For the main paper, we simply write  $G(x, y)$  to refer to Eq. 3.

### 2. Further Details

#### 2.1. Implementation Details

As stated in the paper, our diffusion network integrates the ControlNet [9] structure into the U-Net architecture proposed in [4]. Our stage1, key joint diffusion model features a lighter architecture with fewer diffusion steps compared to the Stage 2, full-body diffusion model. In the key joint diffusion model, the latent dimension of each block in the U-Net is set to 64, while in the full-body diffusion model, it is set to 256. Also, in Stage 1, we set the diffusion timestep  $T$  as 100, whereas in Stage 2, we set  $T$  as 1000. More hyperparameters are presented in Table 1.

For the *object reaching* scenario, we set  $\lambda_1 = 10, \lambda_2 = 100, \lambda_3 = 20$  for calculating the guidance function during the sampling stage 1. Similarly, for the *rock climbing* scenario, we set  $\lambda_1 = 20, \lambda_2 = 20, \lambda_3 = 3$ . For the *contact aware motion generation scenario*, we set  $\lambda_1 = 0, \lambda_2 = 1, \lambda_3 = 0$ . Lastly, for the *sitting with suggested contact points* scenario, we set  $\lambda_1 = 10, \lambda_2 = 20, \lambda_3 = 0$ .

The number of Monte Carlo sampling iterations  $n$  is set to 5 to achieve more accurate gradients.

Hyperparameter	Stage 1	Stage 2
Training iterations	0.3M	1M
Learning rate	1e-4	1e-4
Optimizer	Adam W	Adam W
Weight decay	1e-2	1e-2
Batch size	64	64
Channels dim	64	256
Channel multipliers	[2, 2, 2, 2]	[2, 2, 2, 2]
Variance scheduler	Cosine [5]	Cosine [5]
Diffusion steps	100	1000
Diffusion variance	$\tilde{\beta} = \frac{1-\alpha_t-1}{1-\alpha_t} \beta_t$	$\tilde{\beta} = \frac{1-\alpha_t-1}{1-\alpha_t} \beta_t$
EMA weight ( $\beta$ )	0.9999	0.9999

Table 1. Hyperparameters of each model

#### 2.2. Bounding Box Estimation

We constructed bounding boxes, particularly for the *reaching an object* scenario. To design an upper bounding box, we first connected the joint positions of both shoulders and projected the resulting vector onto a plane to define one axis of the upper bounding box. We set this vector as the direction of one axis, thereby defining the upper body bounding box as the minimal box containing the vertices of the

upper body. Similarly, we connected the joint positions of both feet and utilized the resulting vector as one axis of the lower bounding box, establishing the lower bounding box for the lower body. We predict the shape of the bounding box for each frame based on the positions and 6 DoF poses of the key joints predicted in stage 1, along with the shape parameter  $\beta$ .

### 3. Dataset Description

In the *reaching an object* scenario (Task 1), we utilized the CIRCLE dataset [1]. From the whole dataset, we specifically used the reaching data from the dataset. We augmented the data by reversing the left-hand reaching sequences into right-hand sequences, resulting in a total of 3138 right-hand reaching sequences. Additionally, we processed the data to identify the goal point where the reaching hand moved farthest from its initial position. For the random split experimental setup, 2510 data samples were allocated for training. For the scene split experimental setup, all environments except the media room and closet were designated as the training set, resulting in 2205 data samples for training.

For the *rock climbing* scenario (Task 2), we utilize the dataset from [8], where pose information is accompanied by synchronized RGB images. With this additional information, we manually selected motion sequences starting from when the subject detached from the climbing rock until when they securely reached the climbing rock again, resulting in 156 data samples. Similarly, for the *sitting with suggested contact points* scenario (Task 4), we utilized the dataset from [10]. We manually segmented 160 data samples starting from a stable initial point until the subject sat on the chair.

For the *contact aware motion generation* scenario (Task 3), we utilized the dataset from [3]. We utilized smplx segmentation to convert vertices-level contact into joint-level contact information. If a point within the segmentation of the corresponding part is marked as a contact, the associated joint is designated as a contact joint. We observed that contact usually occurs at the hands and feet, therefore, vertex-level contact was replaced with joint-level contact for the two hands and two feet.

Subsequently, we normalized the entire dataset by setting the face direction at the initial frame to the  $+z$  axis and the initial root position as the origin.

### 4. Additional Task (Task 4): Sitting with Suggested Contact Points

We focus on generating motions sitting on a chair, where contact points on the chair are provided for both hands [10]. We collected 160 samples from the COUCH [10], with 128 sequences used for training. As before, the setting may

Method	Success rate (%)	Dist. to goal (cm)	MJPE (cm)
OmniControl [7]	43.7	16.72	14.02
Ours single-stage	31.2	22.10	15.08
Ours	71.8	10.98	12.84

Table 2. Quantitative evaluation on *sitting with suggested contact points* scenario.

be subject to overfitting and our two-stage composition can provide precise control in unseen conditions. We mark success when the hand positions are within 10 cm of the specified contact points at the final sitting pose. This requirement involves two goal conditions for both hands. Therefore *both hands* are the key joint set of the task.

The scene contains only a chair without a complicated structure and we omitted suggestive-path features and the collision-avoidance guidance term. Our key joint diffusion model still incorporates the trajectory-control guidance, which is sufficient to flexibly adapt the motion sequences to the desired output.

Table 2 shows that our full pipeline outperforms the one-stage pipeline in most of the metrics. Even in the rather simple setting, the single-stage method suffers from overfitting, and the key joint trajectories serve a crucial role in composing the motion with precise control.

### 5. Additional Results

We present additional qualitative results comparing with the baseline for each task, namely the *reaching an object* scenario (Figure 1), *rock climbing* scenario (Figure 2), and *sitting with suggested contact points* scenario (Figure 3). Additionally, we provide supplementary video containing entire motion sequences. Please watch our supplementary video for more results.

### References

- [1] Joao Pedro Araujo, Jiaman Li, Karthik Vetrivel, Rishi Agarwal, Deepak Gopinath, Jiajun Wu, Alexander Clegg, and C. Karen Liu. Circle: Capture in rich contextual environments, 2023. 2
- [2] Hyungjin Chung, Jeongsol Kim, Sehui Kim, and Jong Chul Ye. Parallel diffusion models of operator and image for blind inverse problems. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1
- [3] Chun-Hao P. Huang, Hongwei Yi, Markus Höschle, Matvey Saffroskin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J. Black. Capturing and inferring dense full-body human-scene contact. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 13274–13285, 2022. 2
- [4] Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Guided motion diffusion for

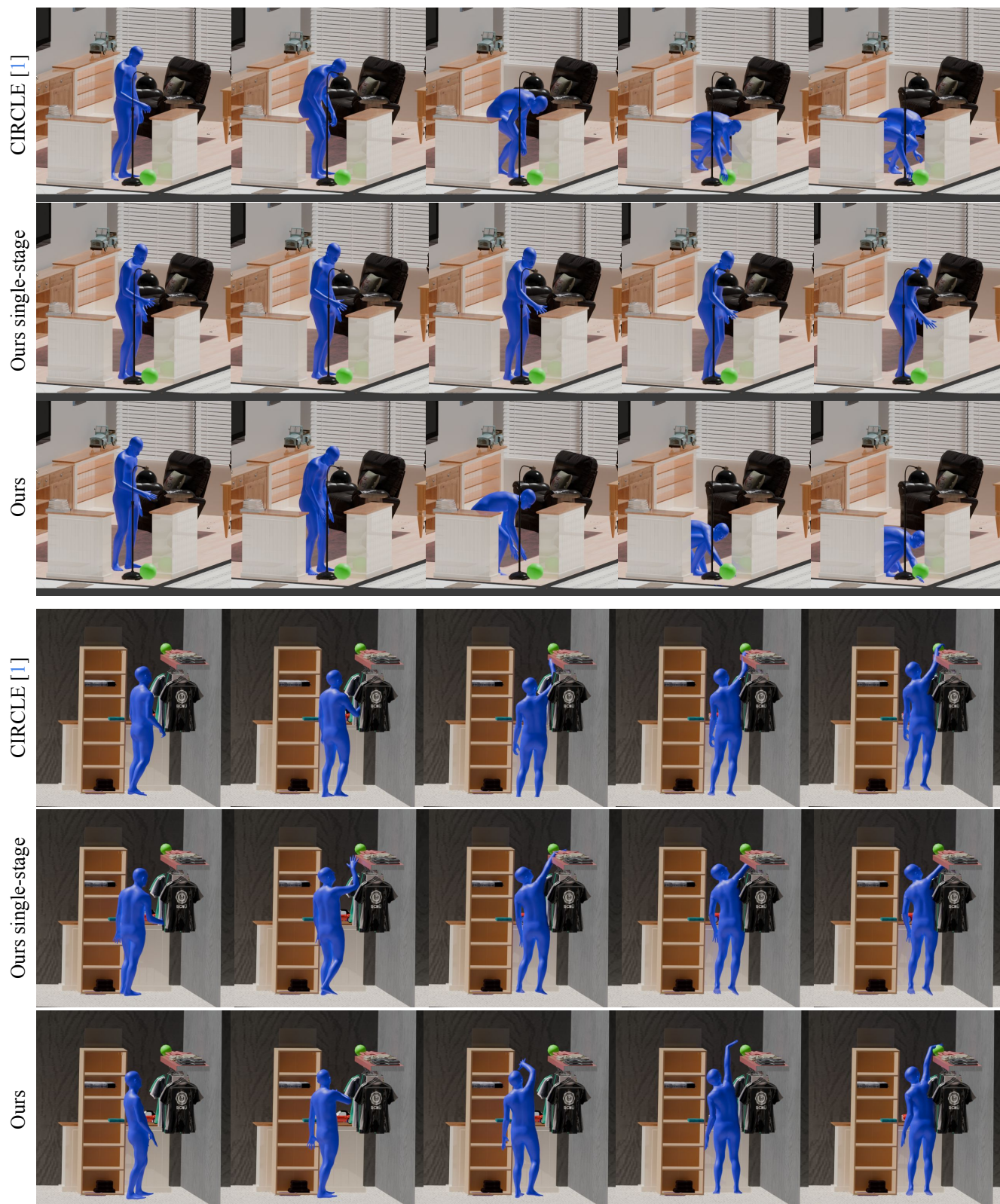
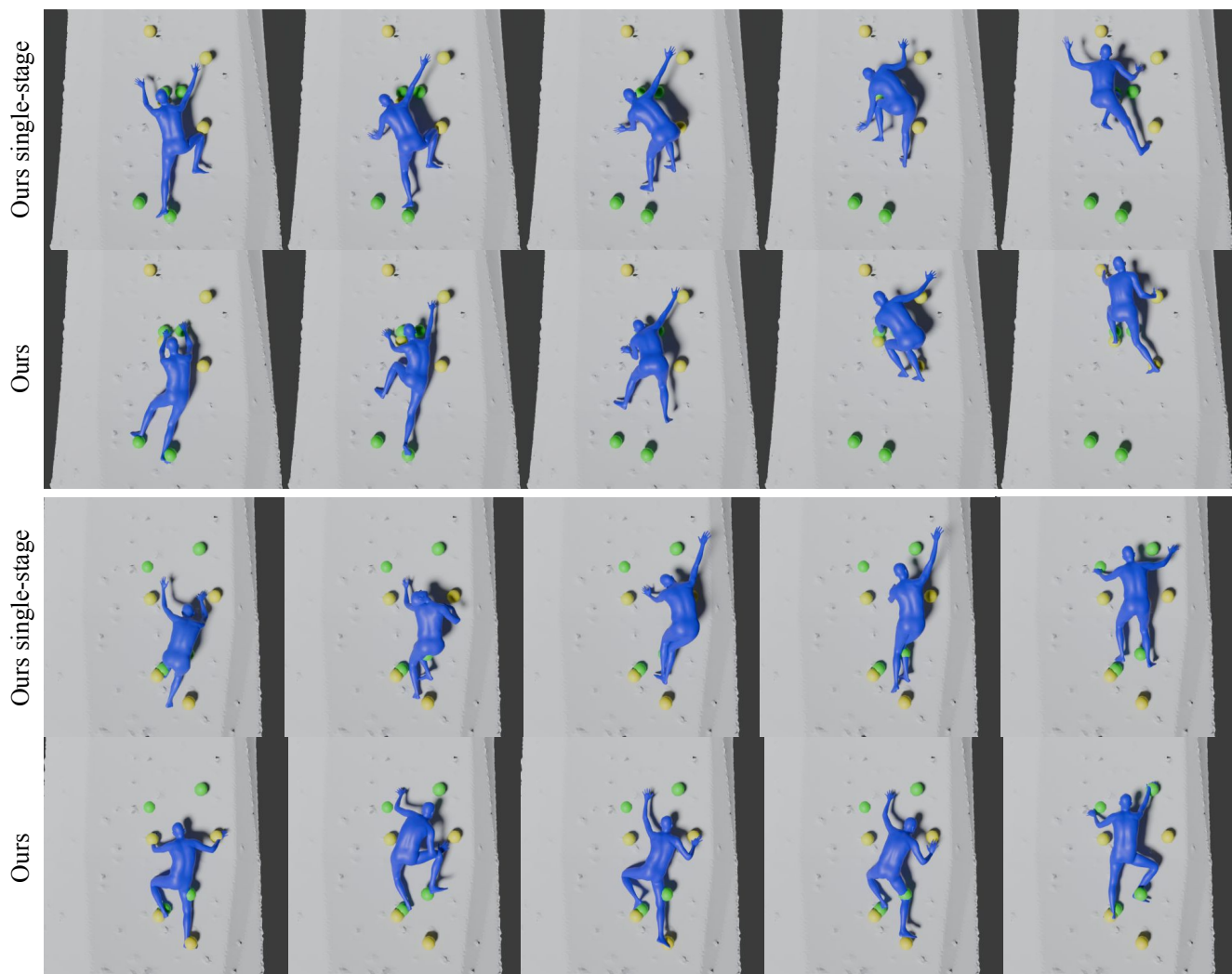


Figure 1. Qualitative results on the *reaching an object* scenario.



Figure 2. Qualitative results on the *rock climbing* scenario.

- controllable human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2151–2162, 2023. 1
- [5] Alex Nichol and Pratul Dharwal. Improved denoising diffusion probabilistic models, 2021. 1
- [6] Jiaming Song, Qinheng Zhang, Hongxu Yin, Morteza Mardani, Ming-Yu Liu, Jan Kautz, Yongxin Chen, and Arash Vahdat. Loss-guided diffusion models for plug-and-play controllable generation. In *International Conference on Machine Learning (ICML)*, 2023. 1
- [7] Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. Omnicontrol: Control any joint at any time for human motion generation. *arXiv preprint arXiv:2310.08580*, 2023. 2
- [8] Ming Yan, Xin Wang, Yudi Dai, Siqi Shen, Chenglu Wen, Lan Xu, Yuexin Ma, and Cheng Wang. Cimi4d: A large multimodal climbing motion dataset under human-scene interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12977–12988, 2023. 2
- [9] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 1
- [10] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Vladimir Guzov, and Gerard Pons-Moll. Couch: Towards controllable human-chair interactions. 2022. 2

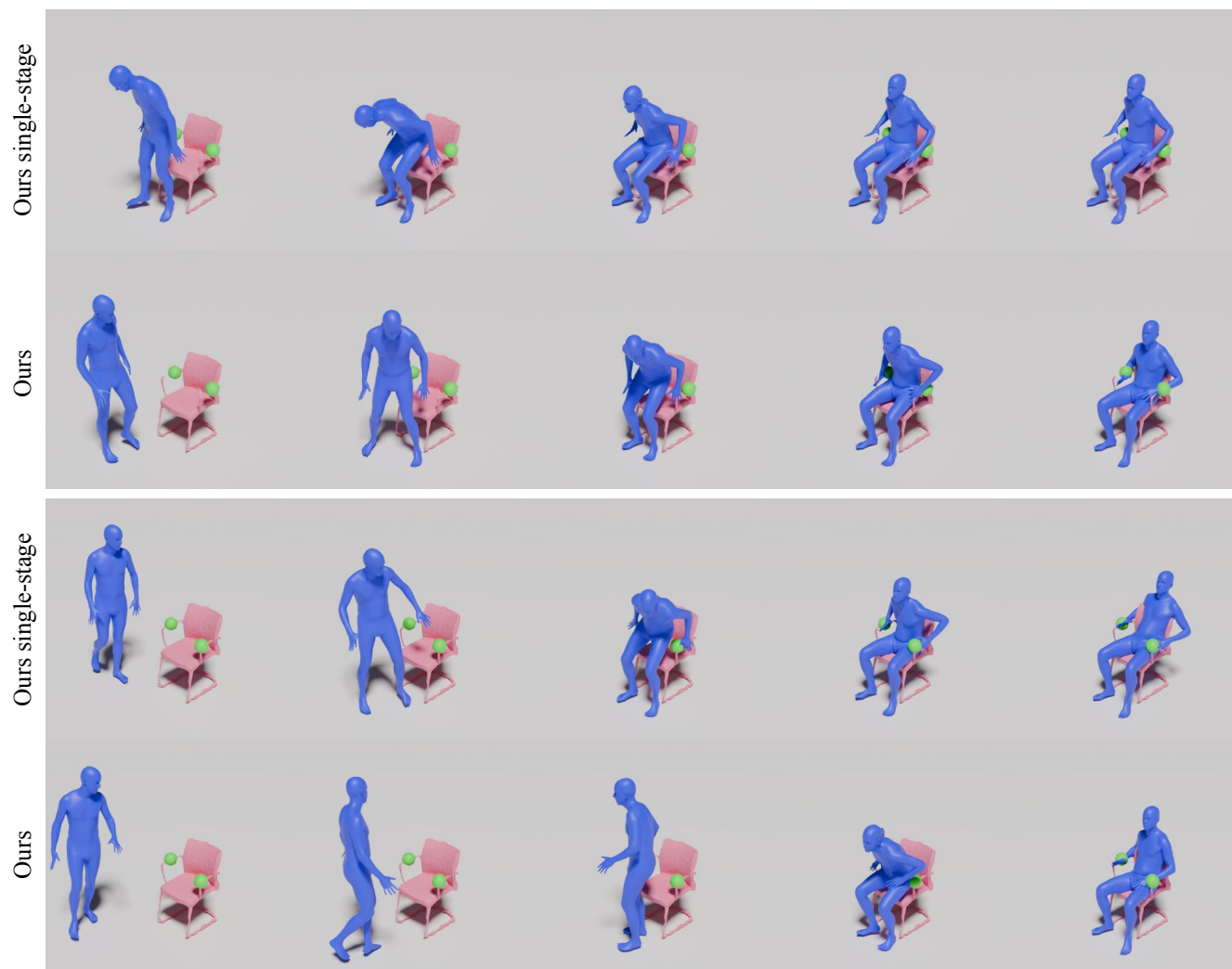


Figure 3. Qualitative results on the *sitting with suggested contact points* scenario.