

Supplementary Materials for

Fast-HOI: Fast Human-Object Interaction Synthesis via Distilled Interaction Prior and Physical Constrains

Xiaokang Pan^{1,*} Zhizhong Zhang^{1,*} Yangyuan Liu⁵ Zhuoran Chen¹ Zhiwei Zhang² Bin Ji⁷
Mingang Chen⁴ Yong Xie⁶ Jingyu Gong^{1,4,†} Xuhong Wang³ Xin Tan^{1,3} Yuan Xie¹

¹East China Normal University ²Shanghai Jiao Tong University ³Shanghai AI Lab

⁴Shanghai Key Laboratory of Computer Software Evaluating and Testing

⁵Nanjing University of Information Science and Technology

⁶Nanjing University of Posts and Telecommunications

⁷East China University of Science and Technology

Abstract

The supplementary materials contain three parts. In the first part, we provide additional visual results on the FullBody-Manipulation and 3D-FUTURE datasets. Then, we conduct more ablation studies on the the proposed method. Finally, we discuss about the limitation of the proposed method and future work.

1. Visualization Results

We visualize more experimental results in this part. In Fig. 1, we provide additional visual results for objects from FullBodyManipulation [3]. Besides, we visualize more samples for objects from 3D-FUTURE [2] in Fig. 2.

2. Ablation Studies

To demonstrate the effectiveness of the newly introduced modules in our method, we conducted the following ablation experiments.

2.1. Ablation Study on Foot Correction

We divided the experiments into three groups: **ours** is our proposed method, which is the same as Group C described above; **ours (without \mathcal{L}_{fp})** is the same as ours but without the additional foot loss \mathcal{L}_{fp} during training; **CHOIS** is the baseline method.

Table 2 and Table 1 shows the H_{feet} and FS metrics for each group on two datasets.

Our foot floating and sliding penalties are crucial. As shown in Table 1, after adding the foot floating and sliding

Table 1. $H_{\text{feet}} \downarrow$ for each method on *Fullbody* and *3D-Futue*

Dataset	ours	ours (without L_{fp})	CHOIS
<i>Fullbody</i>	0.0353	0.0535	0.0363
<i>3D-Futue</i>	0.0276	0.0433	0.0319

Table 2. FS \downarrow for each method on *Fullbody* and *3D-Future* datasets

Dataset	ours	ours (without L_{fp})	CHOIS
<i>Fullbody</i>	2.98	4.56	3.28
<i>3D-Future</i>	3.12	4.67	3.70

loss penalties, there is a significant percentage decrease in the average foot height off the ground and foot sliding metrics compared to the version without these penalties and the baseline. Without the floating penalty, the foot height increases, indicating severe foot floating issues. Meanwhile, without the sliding penalty, the average frame-to-frame foot translation during ground contact increases, indicating serious foot sliding problems.

2.2. implementation details

We provide a detailed description of our experimental setup below.

- **Hardware & Software:** All experiments were performed on a single NVIDIA RTX 4090 GPU with 64GB of memory, using an Ubuntu 18.04 operating system.
- **Training Hyperparameters:**
 - Optimizer: AdamW
 - Learning Rate: 1×10^{-4}
 - Batch Size: 32
 - Training Iterations: 12, 000 (for fine-tuning)
 - Data Order: Random Shuffling

¹*Equal Contribution.

²†Corresponding Author. E-mail:jygong@cs.ecnu.edu.cn

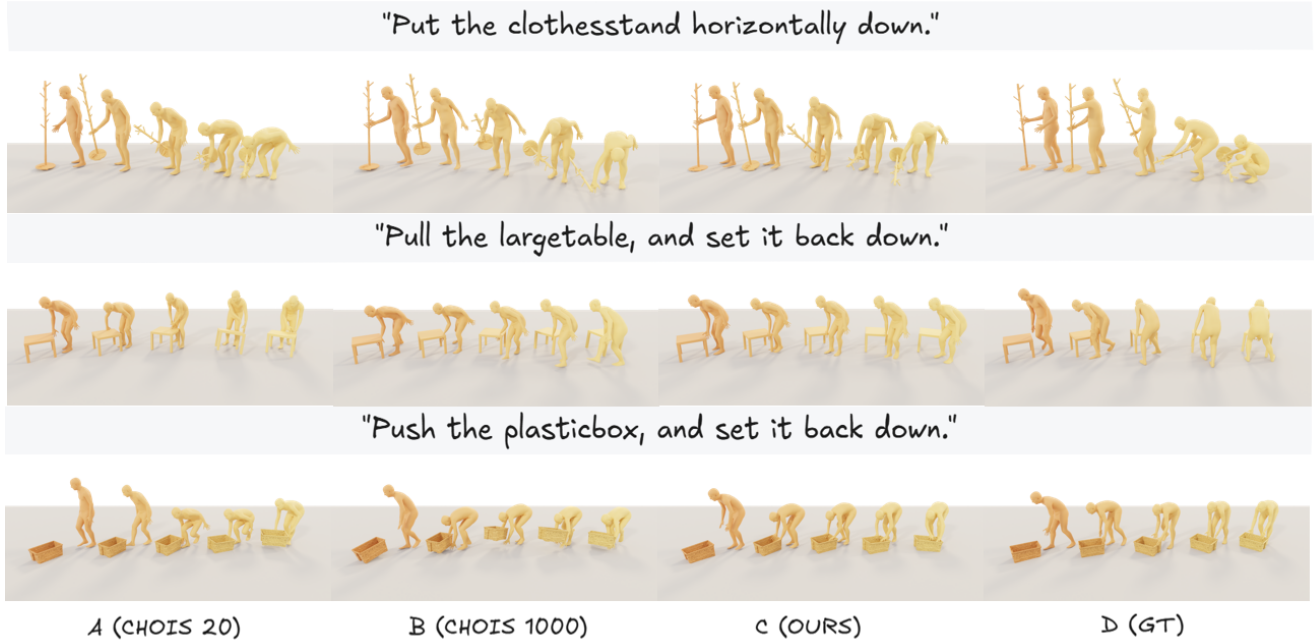


Figure 1. Additional visual results given by competing method for objects from FullBodyManipulation dataset.

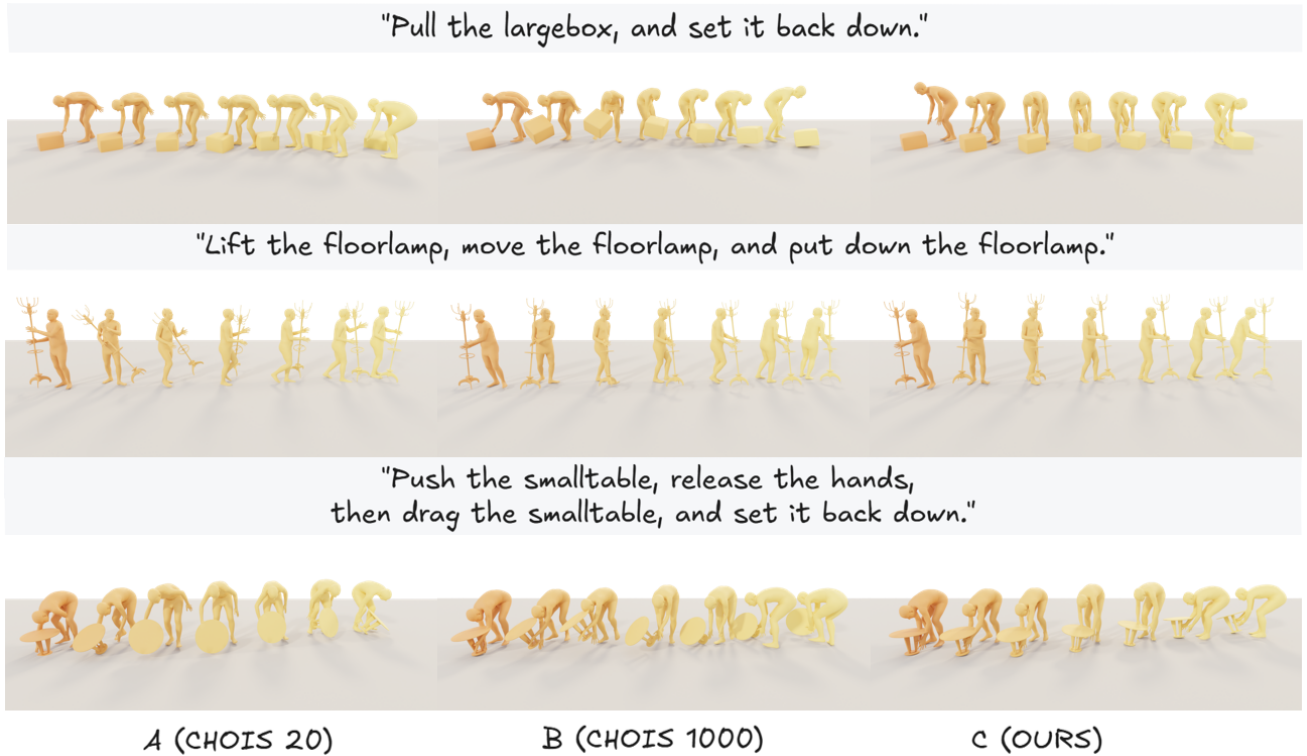


Figure 2. More visual comparison between the proposed method and the SOTA method on objects from 3D-FUTURE dataset.

• **Diffusion Model Configuration:**

- Denoising Steps: 20
- Beta Schedule: ‘cosine’

- Denoising Function: Prediction of \mathbf{x}_0

- **Loss Coefficients:** The composite loss function is defined with the following weights:

- $h^* = 0.195$, $\delta_{\text{toe}} = 0.2$, $\delta_{\text{ankle}} = 0.4$
- $\lambda_{\text{FS}} = 0.5$, $\lambda_{\text{feet}} = 1$, $\lambda_{\text{fk}} = 0.5$
- $\lambda_{\text{obj,kpt}} = 1$, $\lambda_{\text{base}} = 1$, $\lambda_{\text{contact}} = 1$
- **Other Components:**
 - CLIP Version: 'ViT-B/32'

3. Discussion

Although our method improves both efficiency and quality of 3D human motion synthesis, several directions remain worth exploring.

(1) Extension to multimodal conditions: The current framework mainly relies on language descriptions and object meshes as control conditions. Future work may consider integrating multimodal sensory signals, such as incorporating audio rhythms for dance generation and haptic feedback for fine manipulation, to build a richer control system.

(2) Integration of physical simulation: To further enhance the physical plausibility of generated motions, future research can explore combining physical simulation techniques and applying more realistic physical laws to guide human motion generation.

For future work, we aim to integrate physics-based simulation [1] and reinforcement learning [4] to further improve the physical realism and adaptability of HOI motion synthesis. Incorporating differentiable physics engines could allow the model to learn complex interactions and contact dynamics directly, while reinforcement learning could enable the generation of motion sequences that better satisfy task-level goals and constraints [5].

References

- [1] Stelian Coros, Philippe Beaudoin, and Michiel Van de Panne. Generalized biped walking control. *ACM Transactions On Graphics (TOG)*, 29(4):1–9, 2010. 3
- [2] Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3d-future: 3d furniture shape with texture. *International Journal of Computer Vision*, 129:3313–3337, 2021. 1
- [3] Xihui Liu, Dong Huk Park, Samaneh Azadi, Gong Zhang, Arman Chopikyan, Yuxiao Hu, Humphrey Shi, Anna Rohrbach, and Trevor Darrell. More control for free! image synthesis with semantic diffusion guidance. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 289–299, 2023. 1
- [4] Xue Bin Peng, Ze Ma, Pieter Abbeel, Sergey Levine, and Angjoo Kanazawa. Amp: Adversarial motion priors for stylized physics-based character control. *ACM Transactions on Graphics (ToG)*, 40(4):1–20, 2021. 3
- [5] Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew LeFrancq, et al. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018. 3