Supplementary Material for
"Narrator: Towards Natural Control of Human-Scene Interaction Generation via Relationship Reasoning"

Haibiao Xuan[1] Xiongzheng Li[1] Jinsong Zhang[1] Hongwen Zhang[2] Yebin Liu[2] Kun Li[1,*]
[1]Tianjin University [2] Tsinghua University
{hbxuan, lxz, jinszhang, lik}@tju.edu.cn {zhanghongwen, liuyebin}@mail.tsinghua.edu.cn

In this document, we provide the following supplementary content:

- Training Loss.
- Perceptual Study.
- Additional Results.
- Failure Cases and Limatations.

We also provide a demo video along with this document.

1. Training Loss

We introduce multiple training losses as supervision for HSI generation: the reconstruction loss, the Kullback-Leibler divergence loss, and the environmental constraint loss. Formally, the entire training loss is formulated as:

$$\mathcal{L} = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}} + \lambda_{\text{env}}\mathcal{L}_{\text{env}}, \tag{1}$$

where λ denotes the loss weight of each term and is not repeated subsequently.

Reconstruction Loss $\mathcal{L}_{\text{rec}}$. This loss is formulated as:

$$\mathcal{L}_{\text{rec}} = \lambda_{\text{mesh}}\mathcal{L}_{\text{mesh}} + \lambda_{\text{para}}\mathcal{L}_{\text{para}}, \tag{2}$$

where $\mathcal{L}_{\text{mesh}}$ and $\mathcal{L}_{\text{para}}$ denote the body mesh reconstruction loss and the SMPL-X parameter reconstruction loss.

$$\begin{aligned}
\mathcal{L}_{\text{mesh}} &= \lambda_v\mathcal{L}_v + \lambda_n\mathcal{L}_n + \lambda_e\mathcal{L}_e + \lambda_c\mathcal{L}_c, \\
\mathcal{L}_v &= \|\hat{V}_b - V_b\|_1, \\
\mathcal{L}_n &= \sum_{f\in F_b}\sum_{(i,j)\in f}\left|\left\langle n_f, \frac{\hat{v}_i - \hat{v}_j}{\|\hat{v}_i - \hat{v}_j\|_2}\right\rangle\right|, \\
\mathcal{L}_e &= \sum_{f\in F_b}\sum_{(i,j)\in f}\left|1 - \frac{\|\hat{v}_i - \hat{v}_j\|_2}{\|v_i - v_j\|_2}\right|, \\
\mathcal{L}_c &= \sum_{f_i,f_j\in F_b, f_i\cap f_j\neq\emptyset} 1 - \left\langle \hat{n}_{f_i}, \hat{n}_{f_j}\right\rangle,
\end{aligned} \tag{3}$$

where V_b and F_b denote the vertices and the faces of the body mesh, n_f denotes the normal of triangle $f \in F_b$ and $\hat{(\cdot)}$ denotes the corresponding reconstruction.

$$\mathcal{L}_{\text{para}} = \lambda_t\mathcal{L}_t + \lambda_r\mathcal{L}_r + \lambda_\beta\mathcal{L}_\beta + \lambda_p\mathcal{L}_p + \lambda_h\mathcal{L}_h, \tag{4}$$

where $\mathcal{L}_t$, $\mathcal{L}_r$, $\mathcal{L}_\beta$, $\mathcal{L}_p$, and $\mathcal{L}_h$ are ℓ_1 distances between the predicted body parameters and the ground-truths (*i.e.*, global translation t, global orientation r, body shape β, body pose p, and hand pose h).

KL Divergence Loss $\mathcal{L}_{\text{KL}}$. We regularize the learned distribution of latent z by encouraging it to be similar to the normal distribution:

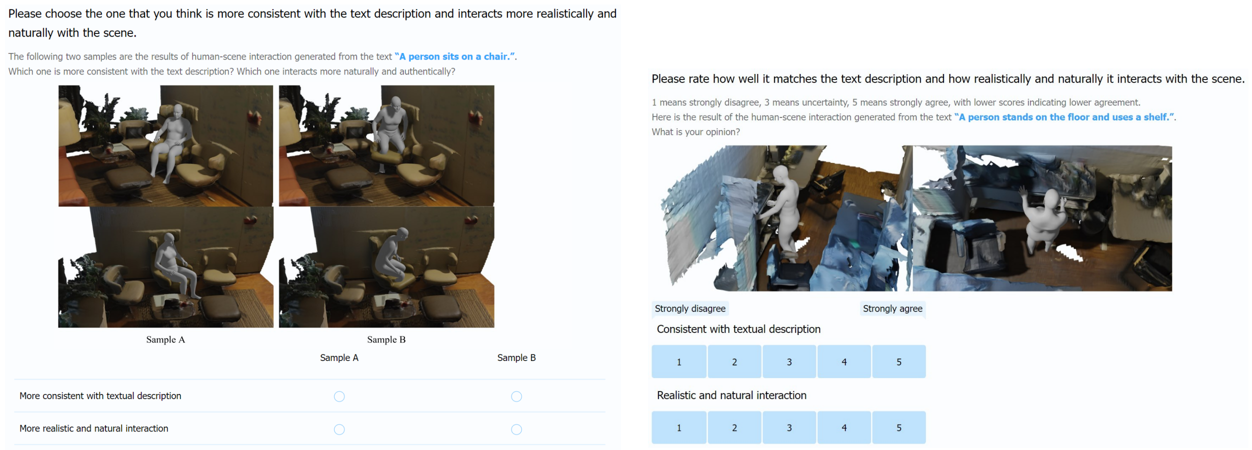
$$\mathcal{L}_{\text{KL}} = D_{\text{KL}}(\mathcal{Q}(z \mid S, W_{1:N}, M)\|\mathcal{N}(0, \mathbf{I})). \tag{5}$$

Environmental Constraint Loss $\mathcal{L}_{env}$. We also consider the contact and penetration issues between the human and the scene, and propose environmental constraint loss as follows:

$$\begin{aligned}
\mathcal{L}_{\text{env}} &= \lambda_{\text{cont}}\mathcal{L}_{\text{cont}} + \lambda_{\text{coll}}\mathcal{L}_{\text{coll}} + \lambda_{\text{IBS}}\mathcal{L}_{\text{IBS}}, \\
\mathcal{L}_{\text{cont}} &= \sum_{v_c\in C(M)} \min_{v_s\in V_S} \rho(|v_c - v_s|), \\
\mathcal{L}_{\text{coll}} &= \sum_i \Psi_S(v_i), \\
\mathcal{L}_{\text{IBS}} &= \sum_{v^p\in V} d_s^p,
\end{aligned} \tag{6}$$

where $C(\cdot)$ the set of all body mesh vertices with contact labels, V_s denotes the vertices of the scene mesh, and Ψ_S denotes the signed distance of body vertex v_i to the scene. In addition, V denotes the set of points that either satisfy penetration or correspond to body vertices with contact labels from the IBS [2] point set. d_s^p indicates the distance from point v^p to the scene.

[*] Corresponding author



Please choose the one that you think is more consistent with the text description and interacts more realistically and naturally with the scene.

The following two samples are the results of human-scene interaction generated from the text "A person sits on a chair.".
Which one is more consistent with the text description? Which one interacts more naturally and authentically?

Sample A Sample B

	Sample A	Sample B
More consistent with textual description	○	○
More realistic and natural interaction	○	○

(a) Binary Perceptual Study

Please rate how well it matches the text description and how realistically and naturally it interacts with the scene.

1 means strongly disagree, 3 means uncertainty, 5 means strongly agree, with lower scores indicating lower agreement.
Here is the result of the human-scene interaction generated from the text "A person stands on the floor and uses a shelf.".
What is your opinion?

Strongly disagree Strongly agree

Consistent with textual description

1 2 3 4 5

Realistic and natural interaction

1 2 3 4 5

(b) Scoring Perceptual Study

Figure 1: Perceptual Study.

2. Perceptual Study

To better evaluate the proposed approach, we conduct perceptual studies to evaluate the accuracy (*i.e.*, how well it matches the text) and the realism (*i.e.*, how natural and plausible the interaction with the scene is) of the interactions, including binary perceptual studies and scoring perceptual studies. The study interface of these two perception studies is shown in Fig. 1. For the binary perception study, we select textual descriptions of several different interactions and generate random samples using our approach and three baselines, respectively. We render each interaction with two different views and compare our approach to the three baselines. During the binary perception study, respondents are instructed to select one of the two samples generated with different methods from two perspectives: more accurate and more realistic, respectively. For the scoring perceptual studies, we sample and render the interactions generated by each method conditioned on text descriptions. These interaction samples are shown to the respondents along with the corresponding textural descriptions, and the respondents are instructed to score the accuracy and the realism from 1 (strongly disagree) to 5 (strongly agree), respectively. We have collected answers from 246 respondents, including 121 females and 125 males with different ages (10 users below 18, 202 users between 18 and 40, 29 users between 40 and 60, and 5 users beyond 60).

3. Additional Results

3.1. Comparison

3.1.1 Additional Qualitative Results

We show more qualitative results given different textual descriptions in Fig. 2-4. Fig. 2 provides some generation examples based on simple text descriptions, where our results is physically more plausible (*e.g.*, less penetration) and interactionally more realistic. Fig. 3 provides some generation examples based on textual descriptions involving spatial relationships, which demonstrates that we can accurately find the required generation position in the 3D scene with our scene graph. Fig, 4 provides some generation examples based on textual descriptions involving multiple actions, which indicates that our approach can handle more diverse and more compositional interactions, thus better meeting the natural needs of the users. Overall, compared to the three baselines, our approach can correctly understand natural language descriptions and controllably generate semantically consistent and physically plausible human-scene interactions, benefiting from our relationship reasoning.

3.1.2 Comparison with COINS

For a fair comparison, we show the quantitative and qualitative comparison results with COINS in Tab. 1 and Fig. 5, respectively. Note that for comparison with COINS, we use the control semantics provided by the PROX-S dataset of COINS [1] (*i.e.*, combinations of actions and objects), where the only difference is that we simply replace them with similar textual descriptions (as shown in the top of Fig. 5) for COINS-Text and our approach. Experimental results show that our modified version does not degrade the performance of COINS, which ensures a fair comparison. In addition, it can also be seen that our results are more natural and reasonable.

3.2. Random Interaction Samples

We show random interaction samples generated by our approach conditioned on the same text description in Fig. 6, which demonstrates that our approach can generate diverse and plausible interactions.

Methods	Physical Plausibility		Diversity	
	Contact	Non-Collision	Entropy	Cluster Size
COINS	0.923	0.931	3.698	1.014
COINS-Text	0.918	0.934	3.701	1.058
Ours	**0.926**	**0.953**	**3.922**	**1.146**

Table 1: Quantitative comparison results with COINS. Contact score and non-collision score are used to evaluate interaction realism and plausibility. Entropy and cluster size are used to evaluate interaction diversity.

3.3. Human Shape Control

Benefiting from body templates with person-dependent shape parameters, our approach allows shape control by varying the SMPL-X body shape parameters during interaction generation. Fig. 7 shows some generated results of one person with the body changing from thin to fat. It can be seen that the human-scene interactions do not appear implausible (*e.g.*, penetration) with the changes in shape, which proves that our approach has a certain generalization ability to the body shape.

4. Failure Cases and Limitations

Although our approach can naturally and controllably generate diverse and complex HSIs in most cases, there are still some limitations and some difficult cases that we have not solved very well. Fig. 8 shows some failure cases of our approach. A common failure mode is body penetration with scenes (even if we design a dedicated penalty), which can happen due to pseudo-ground-truth data and/or the optimization process getting stuck in a local minimum. Apart from this, ambiguous or physically impossible textual descriptions can lead to one-to-many generation mappings and unsatisfactory interaction results, such as the two generation possibilities based on “sits at a table next to a chair” in Fig. 8(c) and simultaneous contact with multiple objects that are far apart. Considering larger scale and higher quality interaction data, we expect to build more expressive interaction generation models and incorporate more diverse and free generative conditions, such as spatial localisation through object properties in the scene graph.

Moreover, Fig. 9 also shows failure cases of our approach for multi-human scene interactions, mainly in interpersonal collisions and inconsistent numbers of people generated and described, which is caused by insufficient space under the generated location. In this aspect, our approach has some limitations. Due to the lack of a multi-human scene interaction dataset, our multi-person generation model currently ignores human-to-human interactions, which are important in the real world. If we incorporate datasets about interpersonal activities into our generation models, it is possible to meet these needs, which would also be an interesting future direction.

References

[1] Kaifeng Zhao, Shaofei Wang, Yan Zhang, Thabo Beeler, and Siyu Tang. Compositional human-scene interaction synthesis with semantic control. *arXiv preprint arXiv:2207.12824*, 2022.

[2] Xi Zhao, He Wang, and Taku Komura. Indexing 3D scenes using the interaction bisector surface. *ACM Trans. Graph.*, 33(3):1–14, 2014.

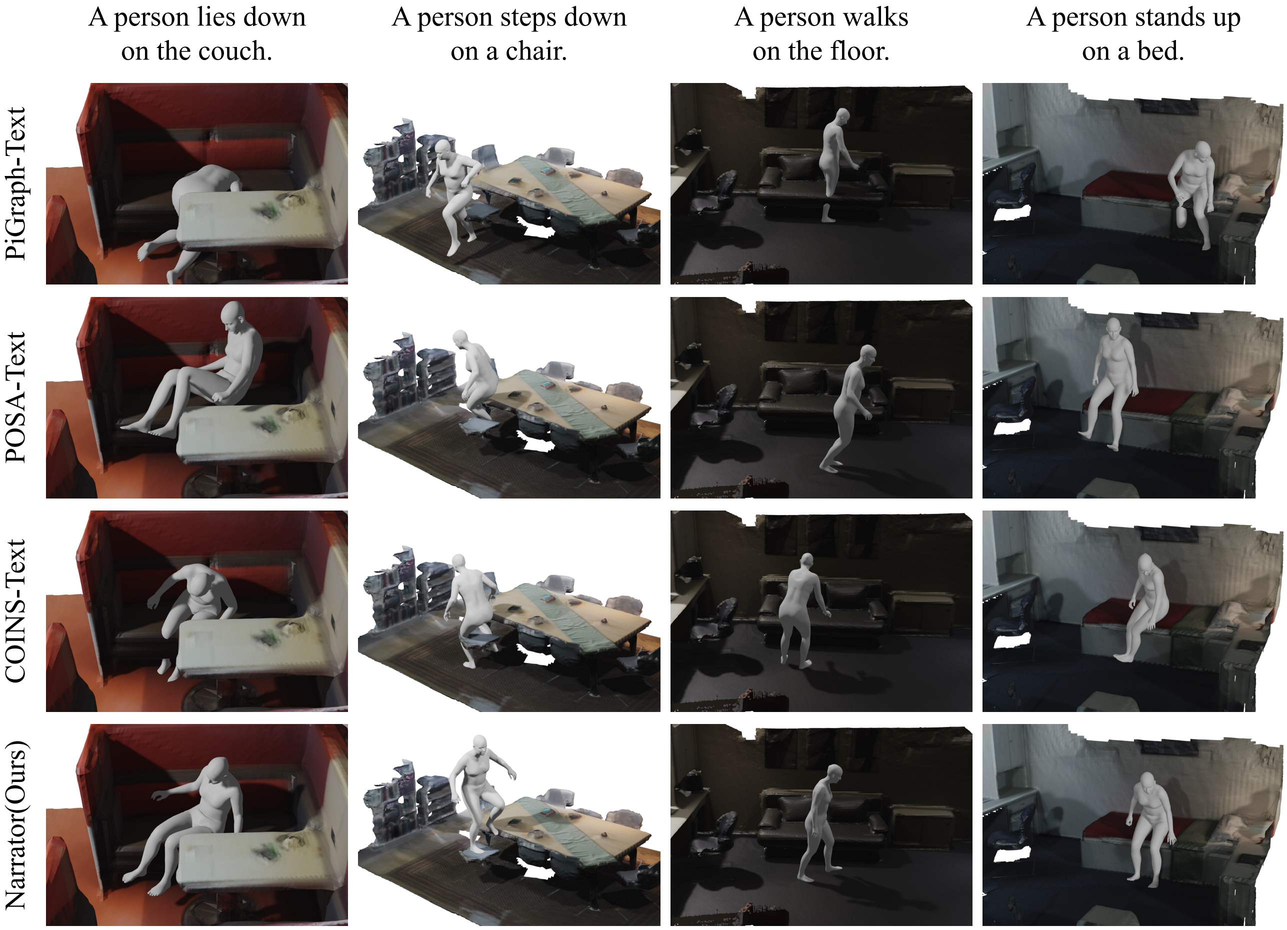

A person lies down on the couch.
A person steps down on a chair.
A person walks on the floor.
A person stands up on a bed.
PiGraph-Text
POSA-Text
COINS-Text
Narrator(Ours)

Figure 2: Additional qualitative comparisons based on simple text descriptions. From top to bottom: PiGraph-Text, POSA-Text, COINS-Text and Narrator(Ours).

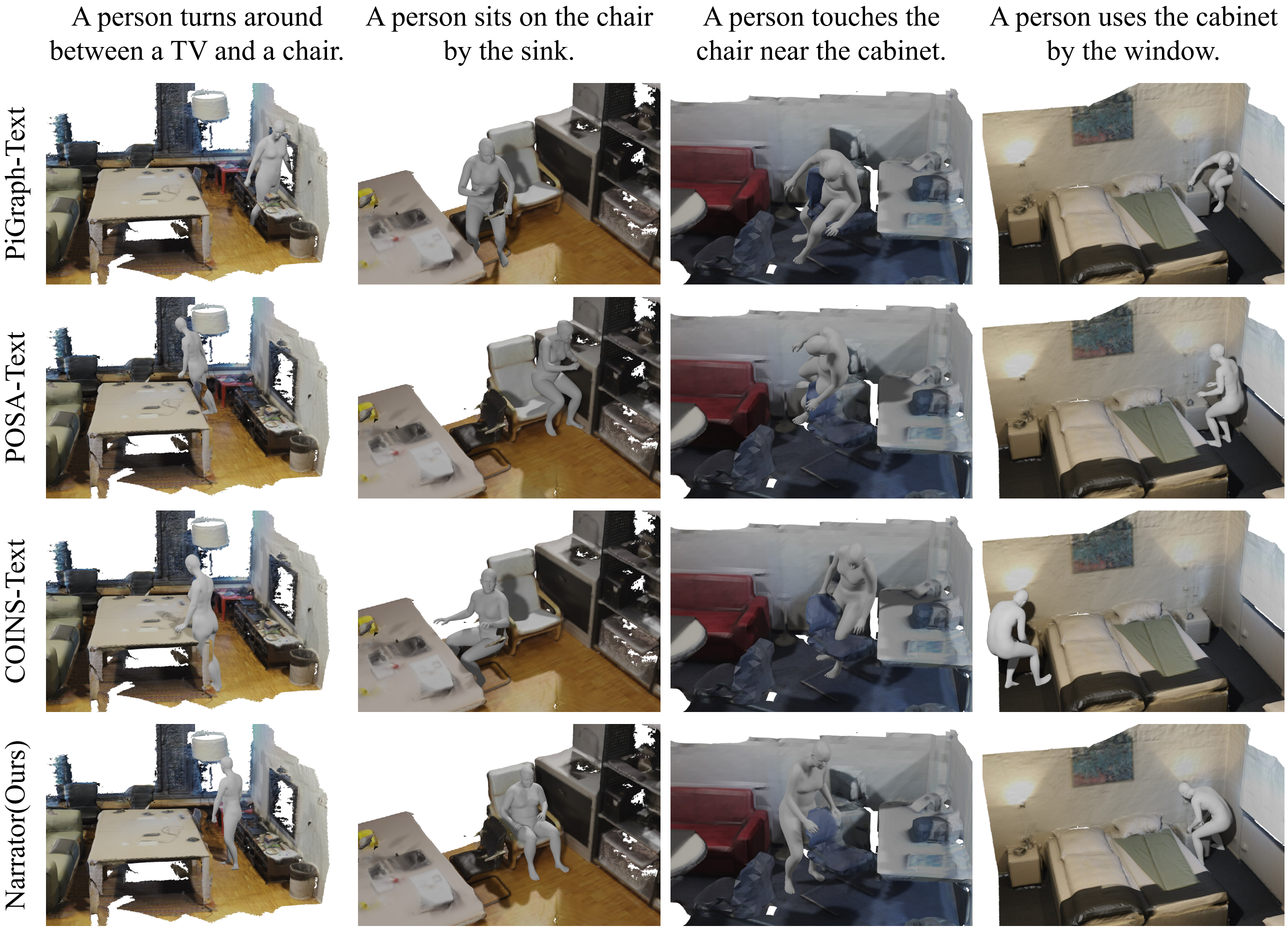

A person turns around between a TV and a chair.
A person sits on the chair by the sink.
A person touches the chair near the cabinet.
A person uses the cabinet by the window.
PiGraph-Text
POSA-Text
COINS-Text
Narrator(Ours)

Figure 3: Additional qualitative comparisons based on textual descriptions involving spatial relationships. From top to bottom: PiGraph-Text, POSA-Text, COINS-Text and Narrator(Ours).

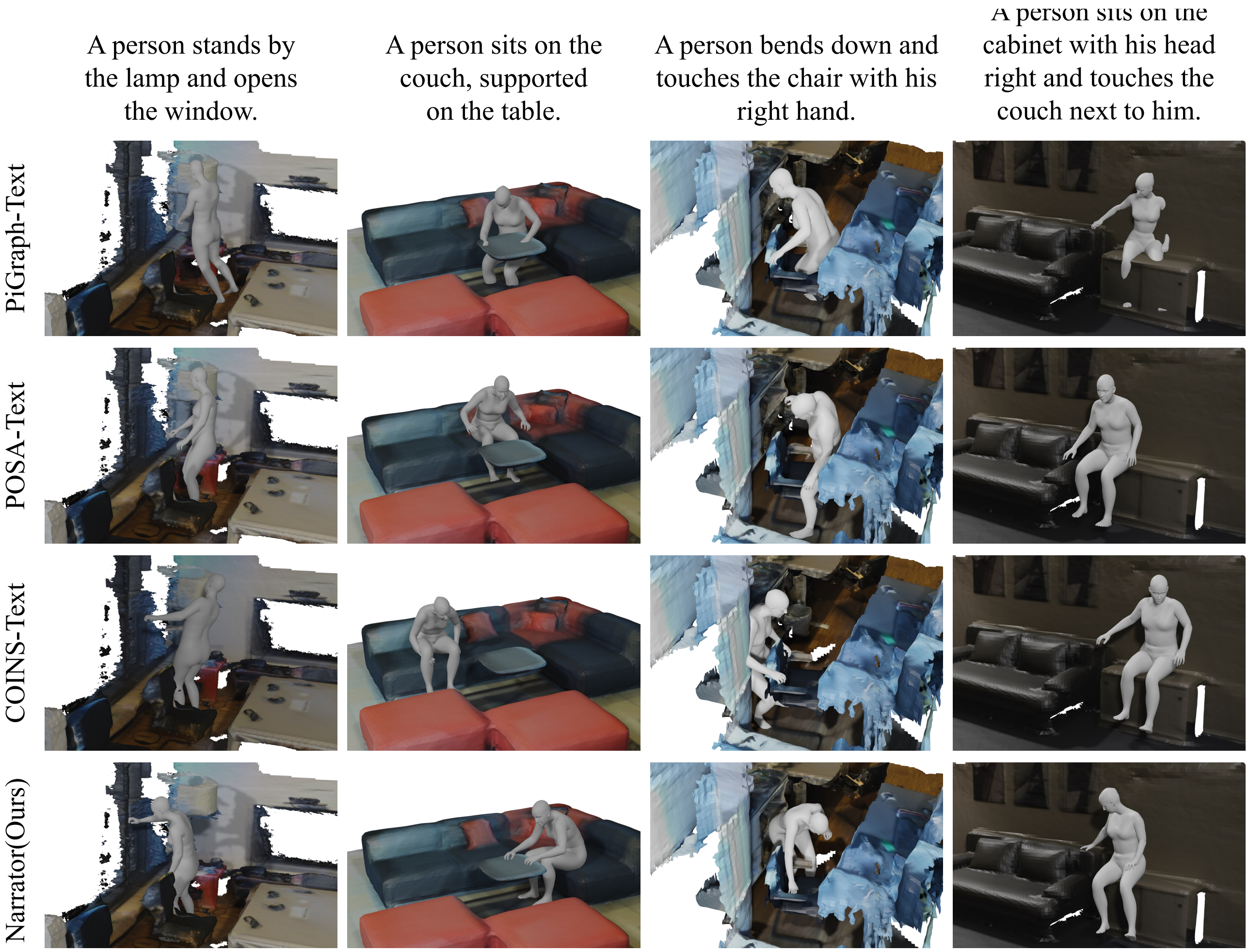

A person stands by the lamp and opens the window.
A person sits on the couch, supported on the table.
A person bends down and touches the chair with his right hand.
A person sits on the cabinet with his head right and touches the couch next to him.
PiGraph-Text
POSA-Text
COINS-Text
Narrator(Ours)

Figure 4: Additional qualitative comparisons based on textual descriptions involving multiple actions. From top to bottom: PiGraph-Text, POSA-Text, COINS-Text and Narrator(Ours).

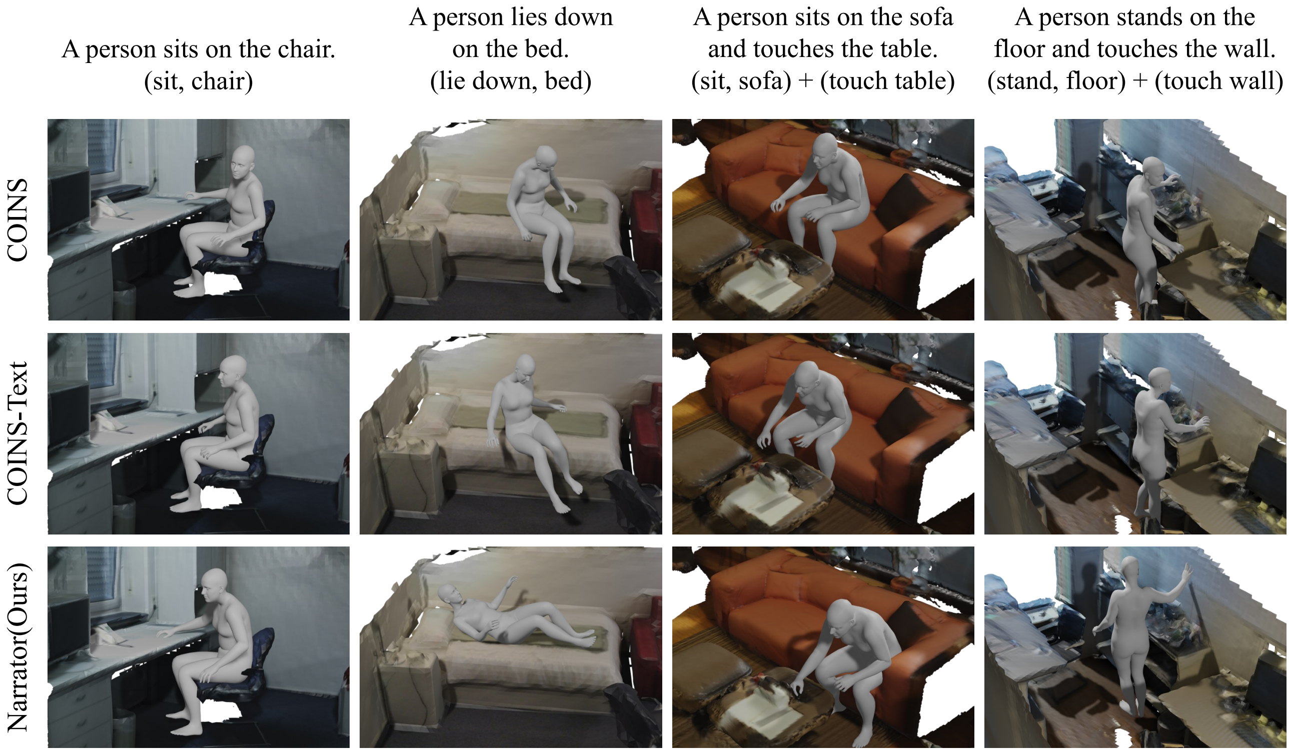

A person sits on the chair.
(sit, chair)
A person lies down on the bed.
(lie down, bed)
A person sits on the sofa and touches the table.
(sit, sofa) + (touch table)
A person stands on the floor and touches the wall.
(stand, floor) + (touch wall)
COINS
COINS-Text
Narrator(Ours)

Figure 5: Qualitative comparison results with COINS using the control semantics provided by the PROX-S dataset of COINS [1] (*i.e.*, combinations of actions and objects). From top to bottom: COINS [1], COINS-Text and Narrator(Ours).

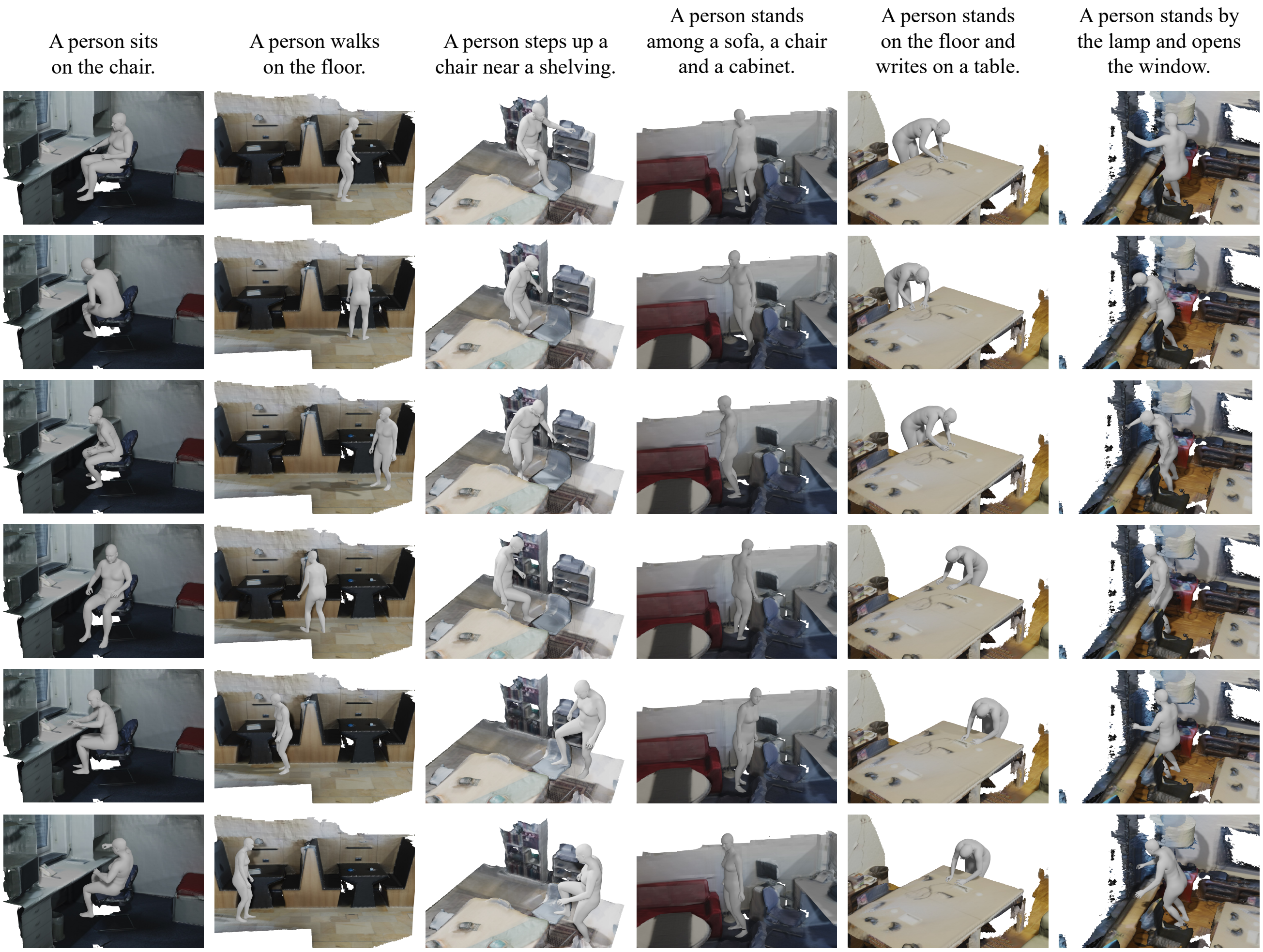

A person sits on the chair.
A person walks on the floor.
A person steps up a chair near a shelving.
A person stands among a sofa, a chair and a cabinet.
A person stands on the floor and writes on a table.
A person stands by the lamp and opens the window.

Figure 6: Randomly sampled interactions from our approach, conditioned on textual descriptions.

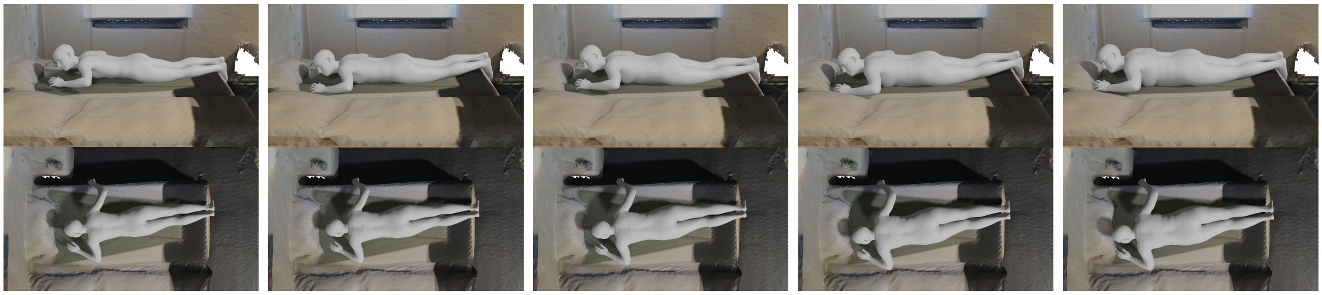


Figure 7: Interaction generation results under shape control. From left to right, we show the results of person changing from thinness to fatness.

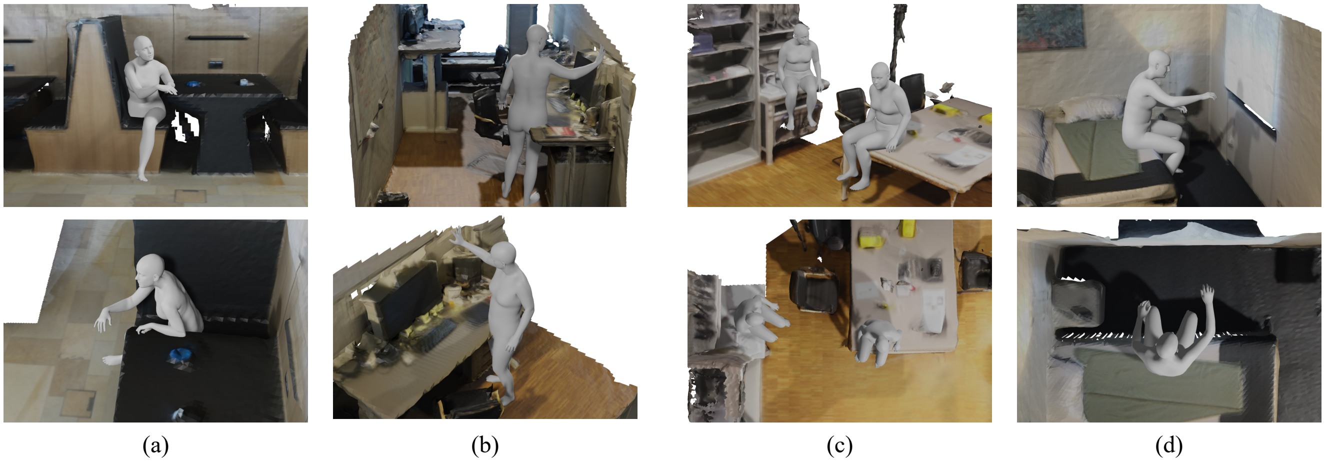


Figure 8: Some examples of failure cases in human-scene interaction. Example (a) shows the penetration caused by incorrect pseudo-ground-truth body fittings. Example (b) shows the penetration caused by being stuck in a local minimum during the optimization process. Example (c) shows unsatisfactory interaction results due to unspecified and ambiguous relationships in the text description (“A person sits at a table next to a chair.”). Example (d) shows an unsatisfactory interaction result due to illogical and unphysical interactive actions in the text description (“A person sits on a bed and touches the window.”)

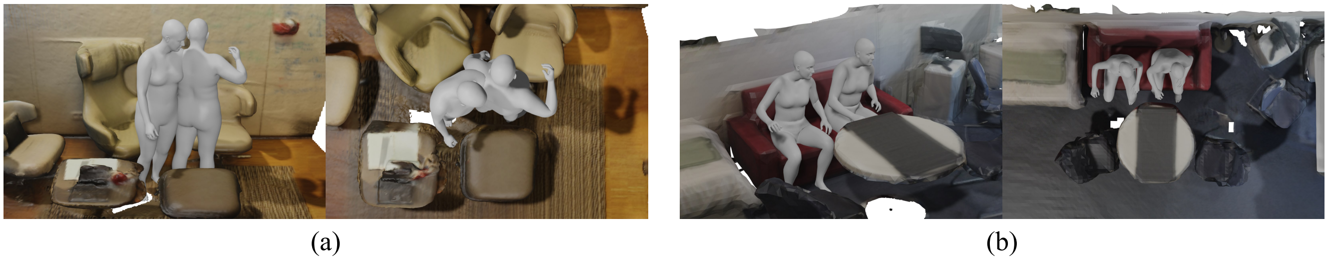


Figure 9: Some examples of failure cases in multi-human scene interaction. Example (a) shows a collision between people. Example (b) shows a failure case where the number of people generated does not match the text description (“Three persons sit together on the sofa.”), due to the small space of the sofa.