Person in Place: Generating Associative Skeleton-Guidance Maps for Human-Object Interaction Image Editing

Supplementary Material

1. Appendix

The supplementary materials are divided into the following sections.

- Sec. 1.1 shows how we constructed the dataset involving objects from V-COCO [1] dataset. We selected images containing HOIs by utilizing Intersection over Union of a person bounding box and an object bounding box.
- Sec. 1.2 explains the evaluation details of FID [2], KID [3] and CLIP score [4]. We present how FID [2], KID [3] and CLIP score [4] are measured, explaining the reason why the aforementioned evaluation metrics were used.
- Sec. 1.3 provides the details of methodologies and module, shown in Table 2 and Table 3 in the main paper. In Fig. 2, the conceptual image is shown, describing how methodologies and modules are utilized.
- Sec. 1.4 presents the implementation details of our method.
- Sec. 1.5 presents the details of the user study. We designed 3 survey items and validated our outstanding performance through the user study.
- Sec. 1.6 shows additional qualitative results of object-interactive skeletons. Visualizing more skeleton results, we show the effectiveness of our associative attention mechanism.
- Sec. 1.7 shows additional results of HOI image editing. With additional visualization results of HOI image editing, we show the effectiveness of our method in HOI image editing.
- Sec. 1.8 In section, we discuss our limitations and further study of our method. We discussed editing image using generated skeletons which overlap each others. Moreover, we discussed the weakness of our proposed metric; Skeleton Probability distance (SPD). Finally, we attempted the automatic modification of generated skeleton using PoseStylizer [5] and showed results using them.

1.1. Datasets

V-COCO [1] is well-known dataset in HOI field. Different from datasets such as HICO [6] and Bongard-HOI [7], it contains GTs of segmentation, skeletons and a person's bounding box. We made a masked area based on the segmentation GT of a person and filled the mask using LaMA [8]. As shown in Fig. 1, to select images containing HOIs, we collected images of which Intersection over Union (IoU) of a person bounding box and an object bounding box is greater than zero. We used V-COCO [1] protocol for training and testing.

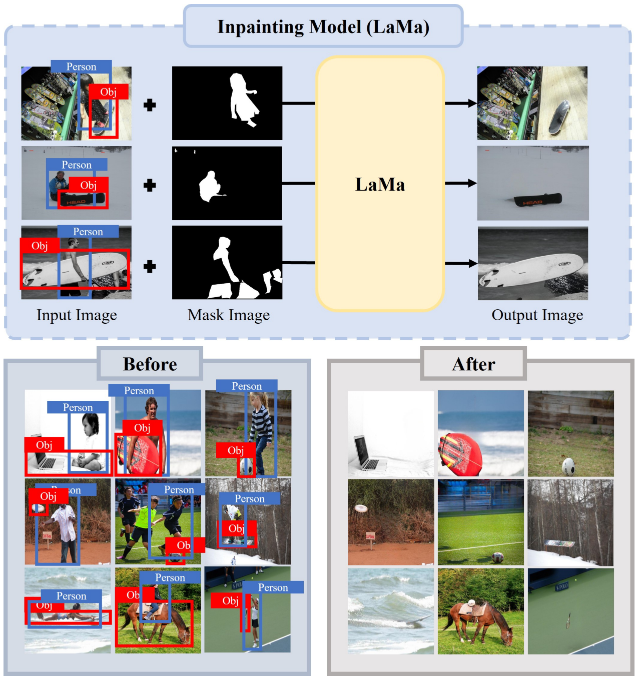


Figure 1. **Our dataset**: This figure shows how we constructed dataset with objects, using LaMa [8]. By merging an input image with an mask image, an output image with objects is generated. In our dataset, various actions (e.g. sitting, holding, kicking) are contained, including actions in V-COCO [1].

1.2. Detailed explanations of FID [2], KID [3], CS

Frèchet Inception Distance (FID [2]) : FID [2] aims to compare the distributions of generated images to the distributions of images from a real dataset. Assuming two datasets follow Gaussian distributions $\mathcal{N}(\mu, \Sigma), \mathcal{N}(\hat{\mu}, \hat{\Sigma})$, FID [2] is defined as:

$$\text{FID}(\mathcal{N}, N) = ||\mu - \hat{\mu}||_2^2 + \text{Tr}(\Sigma + \hat{\Sigma} - 2(\Sigma\hat{\Sigma})^{1/2}) \quad (1)$$

Kernel Inception Distance (KID [3]) : KID [3] measures the squared maximum mean discrepancy (MMD) between the inception network feature map of the real images and generated images using a polynomial kernel. Since it is a non-parametric test, it does not require the strict Gaussian assumption.

CLIP score (CS [4]) : CS [4] measures how well the generated images are aligned with the text conditions. In precise manner, it is a metric that represents the extent to which a text condition matches an images without relying

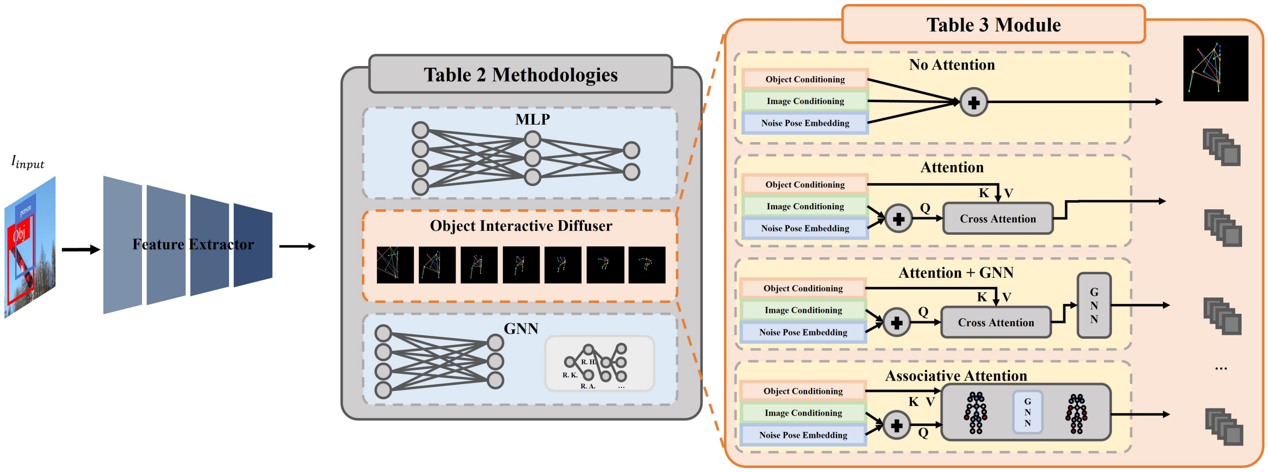


Figure 2. **Details of Table 2 and Table 3**: This figure shows the details of methodologies and object interactive diffuser module. We conducted experiments with these details, as illustrated in Table 2 and Table 3 of the main paper. In **Table 2 Methodologies**, we performed quantitative experiments using a methodology of MLP, GNN and Object interactive diffuser. **Table 3 Module** shows how no attention, attention, attention + GNN work in object interactive diffuser.

on human annotations. Let I be an input image, C be a corresponding text condition, and E_I, E_C be embeddings within the image and text condition, respectively. Then, the CLIP score [4] is defined as follows :

$$\text{CLIPscore}(C, I) = \max(100 \times \cos(E_C, E_I), 0) \quad (2)$$

where the CLIP score [4] is between [0, 100]. FID [2] and KID [3] indicate how realistic generated images are, while CS [4] measures how well a synthesized image is aligned with a prompt which describes interactions. Moreover, to evaluate generated skeletons with HOIs, we define two evaluation metrics.

1.3. More Details of Methodologies & attention module

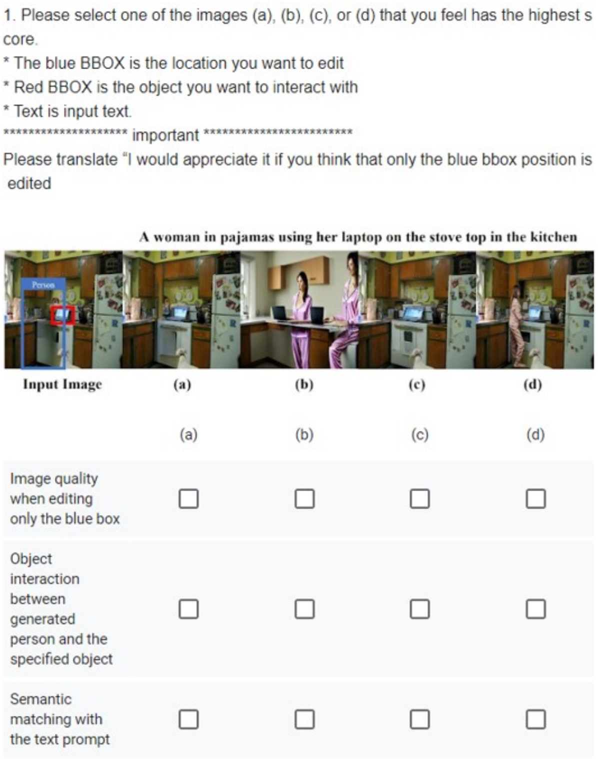
As shown in Fig. 2, we explain the experimental details in Table 2, Table 3 in the main paper. In Table 2, MLP, our object interactive diffuser and GNN are utilized to obtain skeletons. GNN network is implemented after the extracted feature passes through the MLP network. In Table 3, the detailed object interactive diffuser is described with no attention, attention, attention + GNN. Table 3 shows how attention mechanism works during denoiser process. No attention directly passes through denoiser process merging all conditionings, while attention adopts conventional mechanism to utilize conditionings as query. Attention + GNN initially performs conventional mechanism and creates an adjacency matrix corresponding to MSCOCO [9] skeletons, ultimately implementing GNN network.

1.4. Implementation Details

We use Pytorch [10] in training and evaluating our framework. Moreover, we use ImageNet-pretrained ResNet backbone from torchvision [10]. Moreover, we use Adam optimizer [11] in training with batch size 64. Initially, we set learning rate to 10^{-4} and reduce it by $\frac{1}{10}$ at epoch 70 and 120. A single Nvidia RTX-3090 is used to train and inference our framework. Our method employs object interactive diffuser as shown in Table 2 of Fig. 2 and associative attention mechanism as shown in Table 3 of Fig. 2. Moreover, we utilized ControlNet-Inpainting off-the-shelf for skeleton-guided image editing model.

1.5. User Study

To validate the capabilities of our method, we conducted an user study, as shown in Fig. 3. The user study includes 20 test samples and 3 questions, surveyed by total 50 computer vision experts. These are 3 survey items we included in the user study. 'Image quality when editing only the blue box', 'Object interaction between generated person and the specified object', 'Semantic matching with the text prompt'. We conducted the user study with aforementioned 3 items (i.e. the image quality, object interaction, text matching). Figure 4 shows results of user study in comparison between ours, SD-Inpainting [12], Instruct-Pix2Pix [13] and CoModGAN [14] in quality of edited image, object interaction and text matching. We can see overwhelmingly positive responses to our method in 3 survey items. The quality of edited image received 74.2 % of positive response, object interaction received 77.3 % and text matching received 78.4 %. On average, we are seeing a positive response rate of 76.7 % .



1. Please select one of the images (a), (b), (c), or (d) that you feel has the highest s core.

* The blue BBOX is the location you want to edit
* Red BBOX is the object you want to interact with
* Text is input text.

******************** important ************************

Please translate “I would appreciate it if you think that only the blue bbox position is edited

A woman in pajamas using her laptop on the stove top in the kitchen
Person
Input Image
(a)
(b)
(c)
(d)

	(a)	(b)	(c)	(d)
Image quality when editing only the blue box	☐	☐	☐	☐
Object interaction between generated person and the specified object	☐	☐	☐	☐
Semantic matching with the text prompt	☐	☐	☐	☐

Figure 3. **The format of user study**: We designed survey items for assessing the visual quality of images and how well they involve object interaction, such as image quality when editing only the blue box. We asked the subjects to select an image between a, b, c, d for 3 survey items. For fair assessment, we randomly shuffled images edited by CoModGAN [14], Instruct-Pix2Pix [13], SD-Inpainting [12], Ours, in total 20 test samples.

1.6. Additional object-interactive skeleton qualitative results

Fig. 7 presents additional visualization results which are not shown in the main paper. In terms of HOI skeleton generation, our method produces improved results compared to conventional attention mechanism. In some cases, conventional attention mechanism fails to perform object interaction or performs it improperly. With associative attention mechanism performing propagation on each joint, our method generates more aligned and natural skeletons. As shown in results, e.g. swinging a tennis racket or typing a laptop, our method produces more object-interactive images than conventional attention mechanism.

1.7. Additional HOI image editing results

Fig. 8, Fig. 9, Fig. 10 and Fig. 11 show additional HOI image editing results. As shown in Fig. 8 and Fig. 9, a shape of person is not generated or a person without object interaction is generated, similar to the results presented in the main paper. As shown in examples of a woman swinging a tennis racket or a surfer in a black body suit, our method performs HOI image editing more naturally without the problems observed in existing methods. As shown in Fig. 10 and Fig. 11, in high-resolution images, a person is not generated using SD-Inpainting [12] and SDXL-Inpainting [15], when a person bounding box is provided in a small size. Compared to existing methods, our method generates natural human images in all results, using a module which generates object-interactive skeletons. The examples of people sitting on the sofa and children standing on the bed show our outstanding performances.

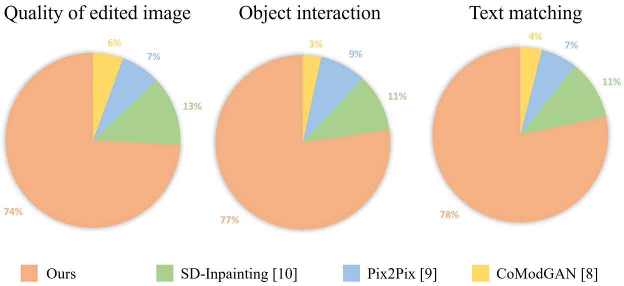


Figure 4. This figure shows visualization results of the user study with 3 survey items. 74 % of the subjects favored our method in quality of edited image, 77 % preferred our method in object interaction, 78 % stated our method outperforms in text matching. The statistical results of the user study show that our results are significant in HOI image editing.

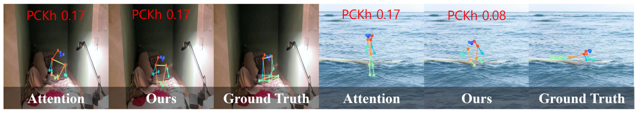


Figure 5. Shown in figure above, PCKh does not represent object interaction. For example, the skeleton generated using attention module and ours have same PCKh but ours interact better than others.

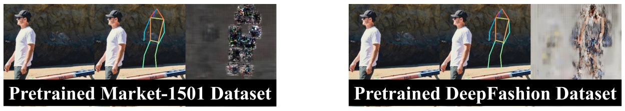

Pretrained Market-1501 Dataset
Pretrained DeepFashion Dataset

Figure 6. Failure case of automatic skeleton modification.

1.8. Discussion

Limitation editing overlapping skeletons : Our method does not fit for editing HOI images that more than two people are overlapped. This is not because we cannot generate overlapping skeletons but the quality of edited images by editing models are not good. The advancement of off-the-shelf image inpainting model would ease this problem.

More Discussions on Skeleton Probability Distance (SPD) : The reason for the inconsistency between the quantitative and qualitative results in Table 3 is that, SPD only assesses how well a skeleton contacts an object. In quantitative results, our associative attention (A.A.) shows a marginally superior result to a standard attention (S.A.) in terms of contact. To the best of our knowledge, there is no metric to assess natural interaction between objects and skeletons. Therefore, to assess the naturalness of the skeleton, we conducted additional user study. Similar to the quantitative results, both A.A. 84.2% and S.A. 73.9% received positive responses in contact. But our A.A. receives highly positive responses 75.3% compared to S.A. 24.7% in naturalness of skeleton. Moreover, there may be questions about the reason why we did not use PCKh for an evaluation metric. However, our goal is not to estimate joints. It is rather generating a skeleton that interacts naturally with objects. As shown Fig. 5, two skeletons have the same PCKh but differ in terms of natural object interaction.

Automatic skeleton adjustment using PoseStylizer [5] : We argued that using our generated skeletons we may manually adjust skeletons before image editing process. Moreover, we experimented the automation of this manual skeleton adjusting process using off-the-shelf skeleton editing model; PoseStylizer [5]. We expected an image that a person in the reference image interacting with an object to be generated, using the PoseStylizer [5] with our object-interactive skeletons. Unfortunately, shown in the Fig. 6, the inference output is unsuccessful. The absence of training might be the reason, which can be resolved by finetuning. We added this application into our paper.

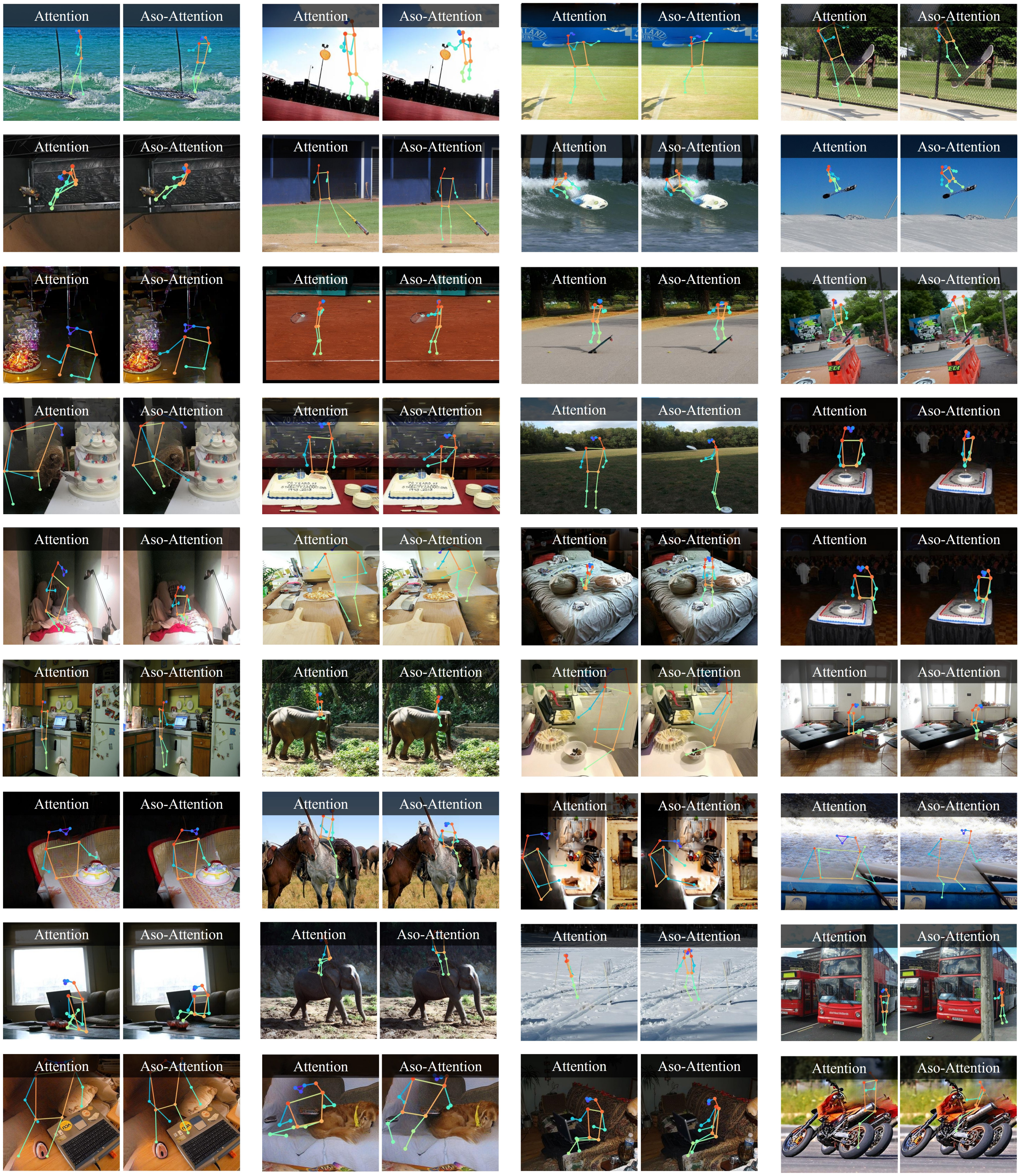


Figure 7. **Comparison between attention mechanism and associative attention mechanism**: By employing our associative attention mechanism, the overall shape of skeleton becomes more natural as shown in images of a baby sitting on the bed or a man sitting on the yacht. Moreover, our associative attention mechanism generates more object-interactive poses, e.g. swinging, sitting, typing, as joints of the skeleton approach the object through propagation process. Moreover, even in scenarios with multiple objects, a natural skeleton is generated while interacting with the specified object.

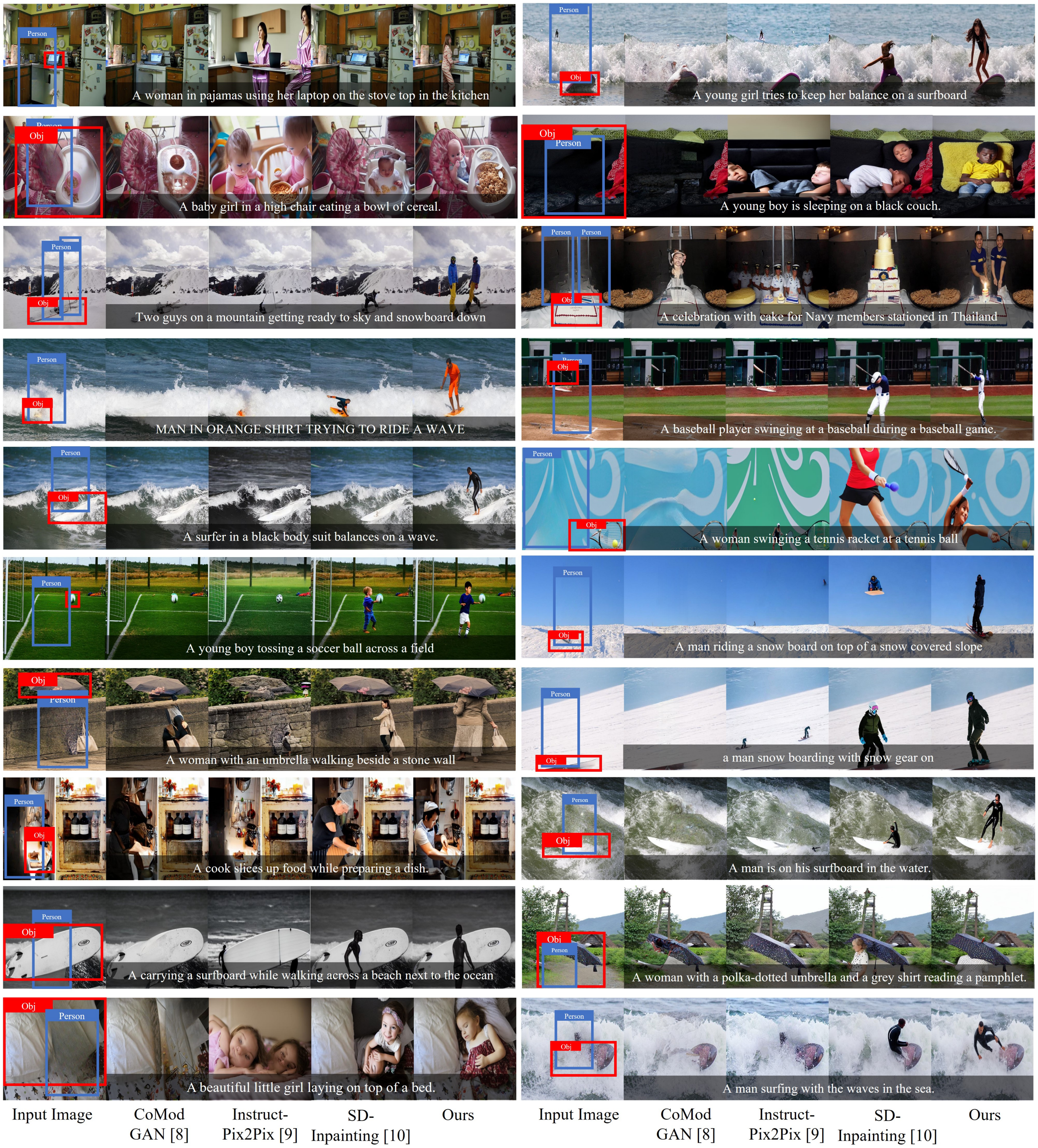

Person
Obj
A woman in pajamas using her laptop on the stove top in the kitchen
A young girl tries to keep her balance on a surfboard
A baby girl in a high chair eating a bowl of cereal.
A young boy is sleeping on a black couch.
Two guys on a mountain getting ready to sky and snowboard down
A celebration with cake for Navy members stationed in Thailand
MAN IN ORANGE SHIRT TRYING TO RIDE A WAVE
A baseball player swinging at a baseball during a baseball game.
A surfer in a black body suit balances on a wave.
A woman swinging a tennis racket at a tennis ball
A young boy tossing a soccer ball across a field
A man riding a snow board on top of a snow covered slope
A woman with an umbrella walking beside a stone wall
a man snow boarding with snow gear on
A cook slices up food while preparing a dish.
A man is on his surfboard in the water.
A carrying a surfboard while walking across a beach next to the ocean
A woman with a polka-dotted umbrella and a grey shirt reading a pamphlet.
A beautiful little girl laying on top of a bed.
A man surfing with the waves in the sea.
Input Image
CoMod GAN [8]
Instruct-Pix2Pix [9]
SD-Inpainting [10]
Ours
Input Image
CoMod GAN [8]
Instruct-Pix2Pix [9]
SD-Inpainting [10]
Ours

Figure 8. **Comparion of CoModGAN [14], Instruct-Pix2Pix [13], SD-Inpainting [12] and Ours**: We use a person bounding box, an object bounding box and a text prompt for HOI image editing. In most cases of the visualization results, CoModGAN [14] and Instruct-Pix2Pix [13] perform poorly in generating natural humans. Our results exhibit more object-interactive images than SD-Inpainting [12], as shown in cases of a woman with wearing a polka-dotted umbrella or a man surfing with the waves. For an example, in the case of 'A woman in pajamas using her laptop on the stove top in the kitchen', CoModGAN [14] and SD-Inpainting [12] did not generate even a human shape. Instruct-Pix2Pix [13] failed to maintain the original image, while our method generated a natural woman that matches the text prompt.

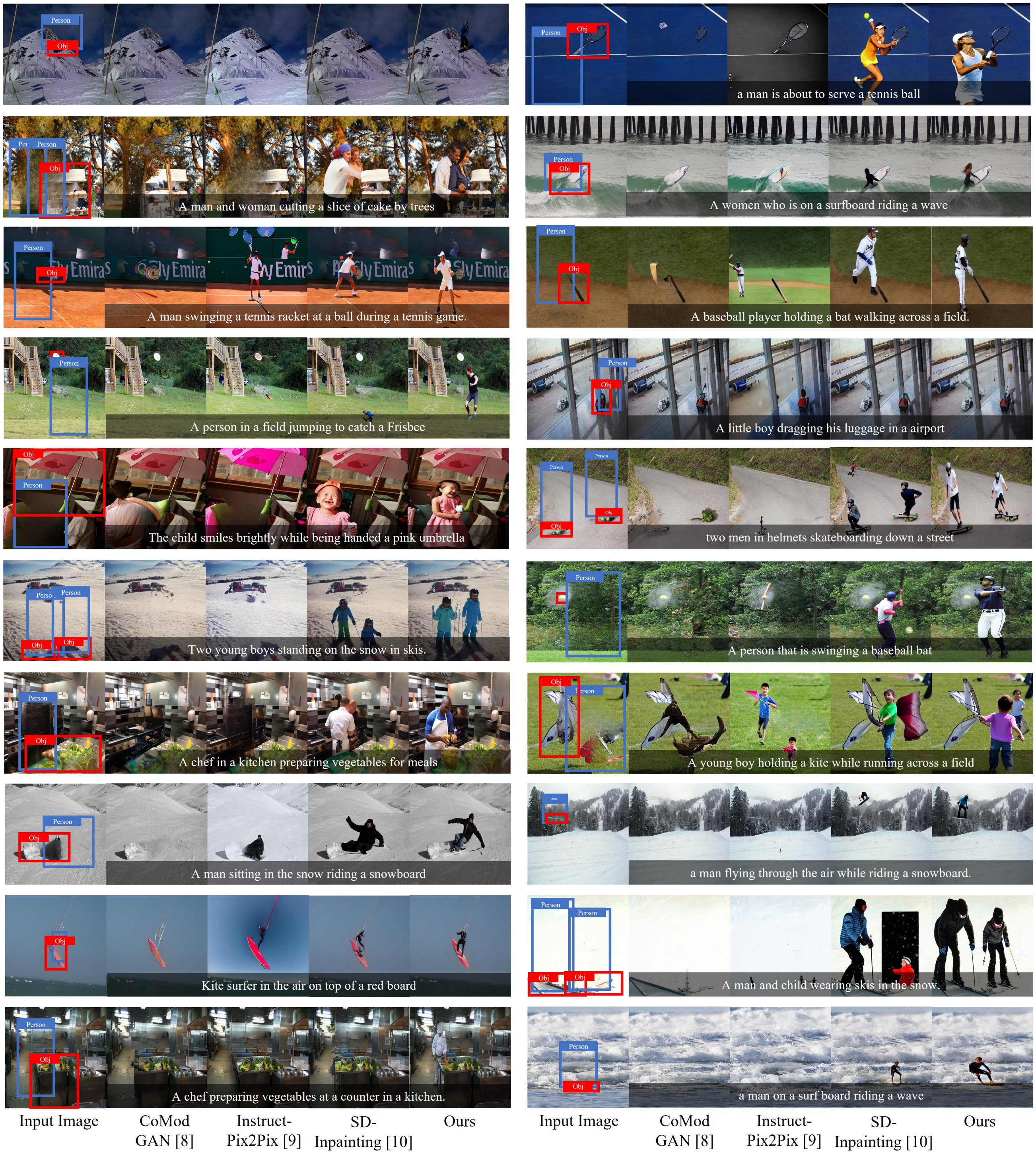

Person
Obj
a man is about to serve a tennis ball
A man and woman cutting a slice of cake by trees
A women who is on a surfboard riding a wave
A man swinging a tennis racket at a ball during a tennis game.
A baseball player holding a bat walking across a field.
A person in a field jumping to catch a Frisbee
A little boy dragging his luggage in a airport
The child smiles brightly while being handed a pink umbrella
two men in helmets skateboarding down a street
Two young boys standing on the snow in skis.
A person that is swinging a baseball bat
A chef in a kitchen preparing vegetables for meals
A young boy holding a kite while running across a field
A man sitting in the snow riding a snowboard
a man flying through the air while riding a snowboard.
Kite surfer in the air on top of a red board
A man and child wearing skis in the snow.
A chef preparing vegetables at a counter in a kitchen.
a man on a surf board riding a wave
Input Image
CoMod GAN [8]
Instruct-Pix2Pix [9]
SD-Inpainting [10]
Ours
Input Image
CoMod GAN [8]
Instruct-Pix2Pix [9]
SD-Inpainting [10]
Ours

Figure 9. **Comparion of CoModGAN [14], Instruct-Pix2Pix [13], SD-Inpainting [12] and Ours**: We use a person bounding box, an object bounding box and a text prompt for HOI image editing. In most cases of the visualization results, CoModGAN [14] and Instruct-Pix2Pix [13] perform poorly in generating natural humans. Our results exhibit more object-interactive images than SD-Inpainting [12], as shown in cases of a person in a field jumping or a baseball player holding a bat. For an example, in the case of 'two men in helmets skateboarding down a street', CoModGAN [14] and Intruct-Pix2Pix [13] failed to generate even a human shape. SD-Inpainting [12] generated two young boys and an additional person which do not match the text prompt, while our method exhibited outstanding results of two men which match the text semantically

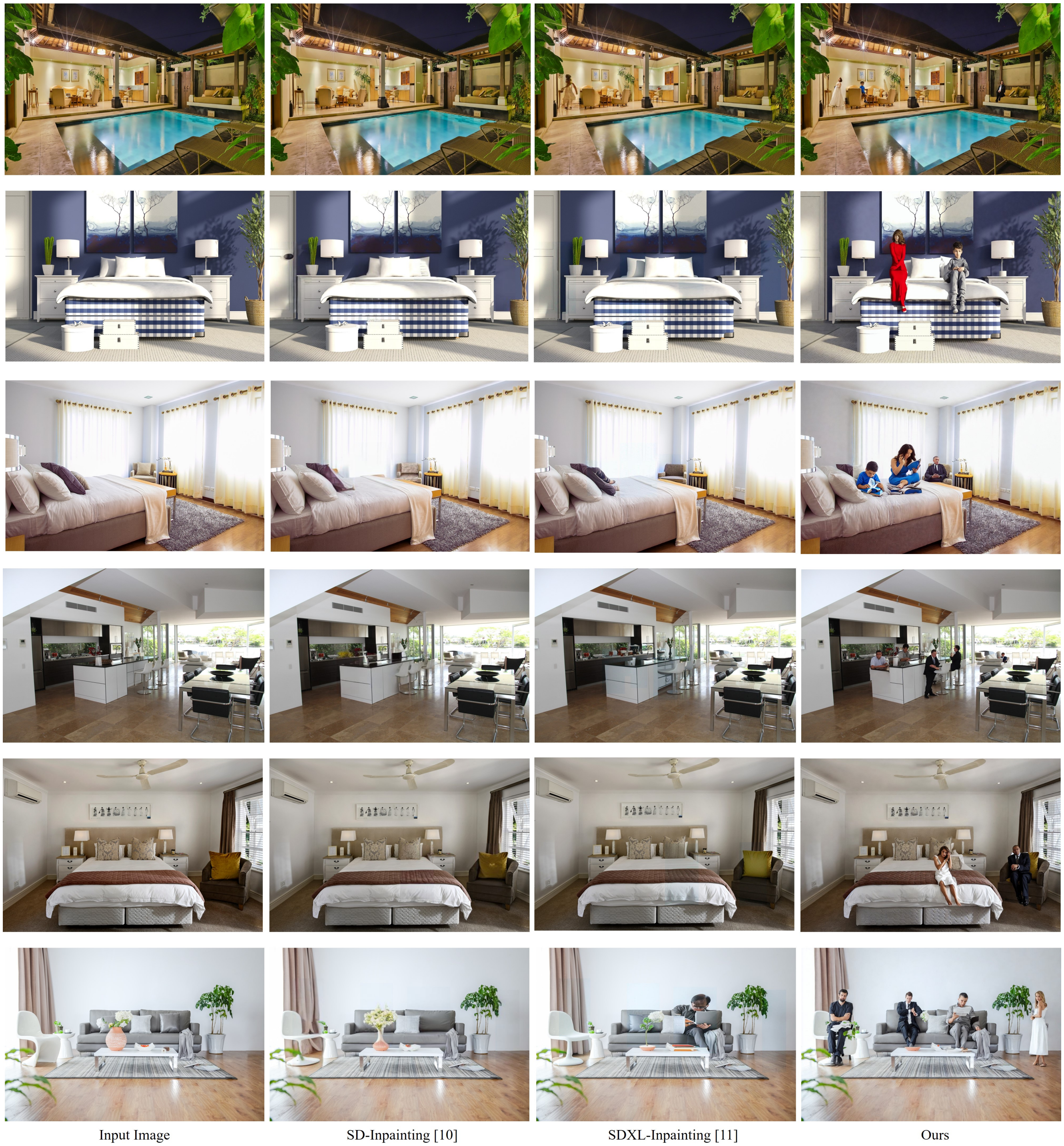


Figure 10. **Performing on multi-people images**: This figure shows HOI-edited images of multiple people using SD-Inpainting [12], SDXL-Inpainting [15] and Ours. In the first and second column, SD-Inpainting [12] and SDXL-Inpainting [15] fail to generate a human when a person bounding box is provided in a small size. On the other hand, our method generates natural HOI images regardless of a size of a person bounding box, since it utilizes object-interactive skeletons. As shown in the fourth column, people sitting on the bed and men sitting on the sofa are generated naturally with our method.

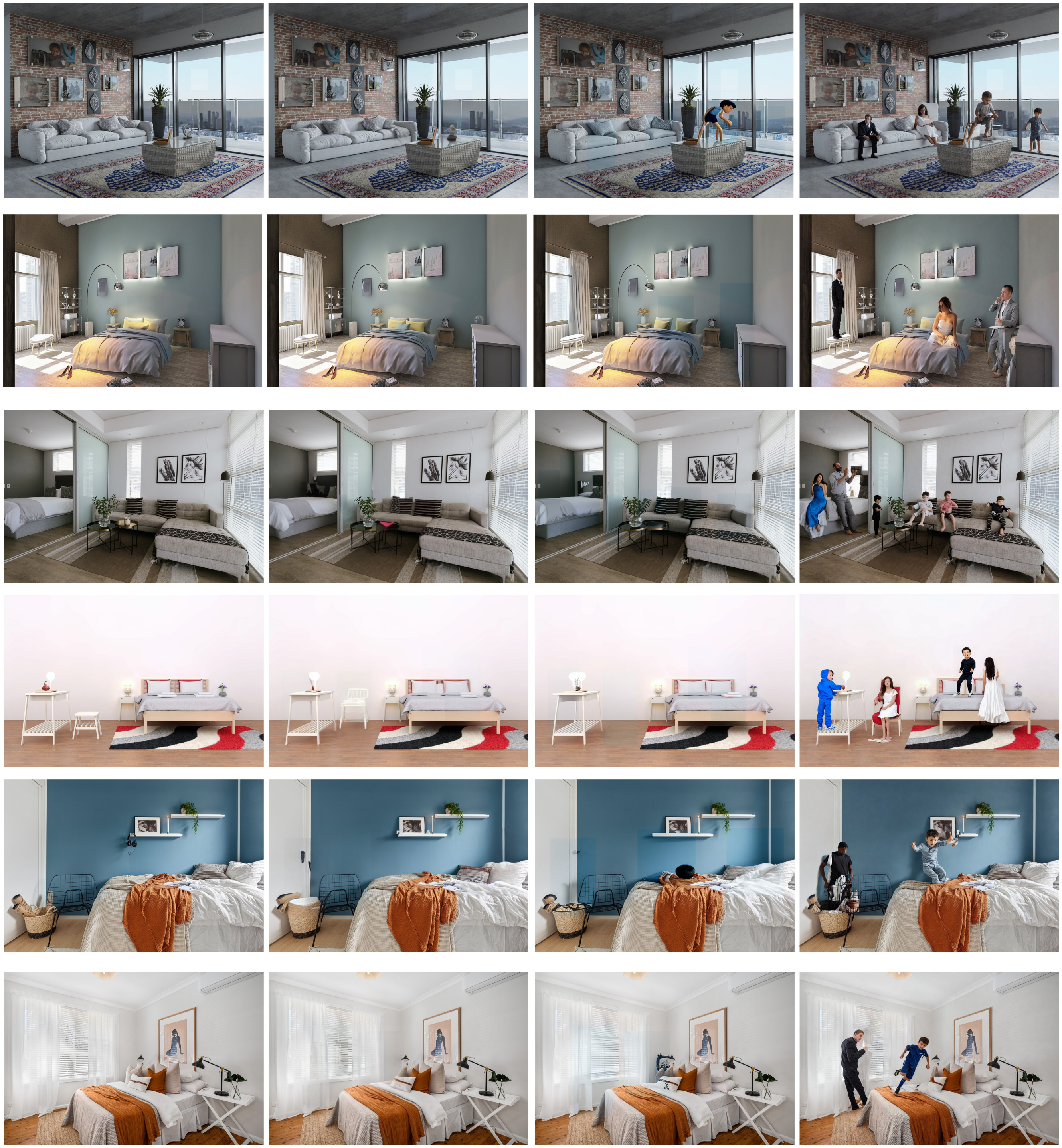


Figure 11. **Performing on multi-people images**: This figure shows HOI-edited images of multiple people using SD-Inpainting [12], SDXL-Inpainting [15] and Ours. In the first and second column, SD-Inpainting [12] and SDXL-Inpainting [15] fail to generate a human when a person bounding box is provided in a small size. On the other hand, our method generates natural HOI images regardless of a size of a person bounding box, since it utilizes object-interactive skeletons. As shown in the fourth column, people sitting on the sofa and children sitting on the sofa are generated naturally with our method.

References

[1] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. 1

[2] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 1, 2

[3] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 1, 2

[4] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: a reference-free evaluation metric for image captioning. In *EMNLP*, 2021. 1, 2

[5] Siyu Huang, Haoyi Xiong, Zhi-Qi Cheng, Qingzhong Wang, Xingran Zhou, Bihan Wen, Jun Huan, and Dejing Dou. Generating person images with appearance-aware pose stylizer. In *IJCAI*, 2020. 1, 4

[6] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. Hico: A benchmark for recognizing human-object interactions in images. In *Proceedings of the IEEE international conference on computer vision*, pages 1017–1025, 2015. 1

[7] Huaizu Jiang, Xiaojian Ma, Weili Nie, Zhiding Yu, Yuke Zhu, and Anima Anandkumar. Bongard-hoi: Benchmarking few-shot visual reasoning for human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19056–19065, 2022. 1

[8] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2149–2159, 2022. 1

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014*, pages 740–755. Springer, 2014. 2

[10] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 2

[11] Zijun Zhang. Improved adam optimizer for deep neural networks. In *2018 IEEE/ACM 26th international symposium on quality of service (IWQoS)*, pages 1–2. Ieee, 2018. 2

[12] Runway. Stable diffusion inpainting. In *https://huggingface.co/runwayml/stable-diffusion-inpainting*, 2022. 2, 3, 6, 7, 8, 9

[13] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 2, 3, 6, 7

[14] Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I Chang, and Yan Xu. Large scale image completion via co-modulated generative adversarial networks. *arXiv preprint arXiv:2103.10428*, 2021. 2, 3, 6, 7

[15] Suraj Patil. Sdxl inpainting. In *https://huggingface.co/spaces/diffusers/stable-diffusion-xl-inpainting/tree/main*, 2022. 3, 8, 9