APPENDIX

In this supplementary, we will first present the implementation details and training parameters in Appendix A. Then, more details about the dataset construction process and statistics are presented in Appendix B. Then we summarize a comprehensive set of related work comparisons in Appendix C. We further provide more information about the test dataset for two-subject evaluation in Appendix D. We will discuss the interpolation results in Appendix E. We will discuss the Additional qualitative results of the ablation studies in Appendix F. We will discuss the failure cases in Appendix G. We will compare our method with methods trained on the Imagen in Appendix H. And finally, more visualization results of our proposed Subject-Diffusion are exhibited in Appendix I.

A IMPLEMENTATION DETAILS

Based on SD v2-base[1], Subject-Diffusion consists of VAE, UNet (with adapter layer), text encoder, and OpenCLIP-ViT-H/14[2] vision encoder, comprising 2.5 billion parameters, out of which a mere 0.7 billion parameters (text encoder, *conv_in* module, adapter layer, and projection matrices $W_K^{(i)}, W_V^{(i)}$) are trainable. The VAE, text encoder, and UNet are initialized from the SD checkpoints, and the CLIP image encoder is loaded from the pretrained OpenCLIP checkpoints. We set the learning rate to 3e-5, the weighting scale hyper-parameter λ_{attn} in Eq. (3) to 0.01, and the balance constant β in the adapter to 1. The entire model is trained on 24 A100 GPUs for 300,000 steps with a batch size of 12 per GPU. The model is trained based on our proposed SDD or OpenImage training set.

B SUBJECT-DIFFUSION DATASET

B.1 DATASET BUILDING STRATEGY

To produce our dataset, all of our training images are sampled from the LAION-Aesthetics V2 5+[3] which is a subset of LAION-5B with an aesthetic score greater than 5. To keep the diversity of images, we only set the filter conditions for resolution, *i.e.,* keep the images with the small side greater than 1024. However, in order to ensure that the images are suitable for our subject-driven image generation task, we apply several filtering rules: (1) We only keep the bounding boxes with an aspect ratio between 0.3 and 3; (2) We only keep images where the subject's bounding box area is between 0.05 and 0.7 of the total image area; (3) We filter out entities with IOU exceeding 0.8; (4) We remove entities that appear more than 5 times in a detection box; (5) We filter out entities with detection scores below 0.2; (6) We remove images where the segmentation mask area is less than 60% of the corresponding detection box area; (7) For the OpenImages training set, we filter out entities that appear in groups and belong to human body parts. After applying these rules, we keep 22 million images for our SDD and 300,000 images for the OpenImages dataset.

B.2 STATISTICS AND COMPARISON

Statistics about our training data are illustrated in Fig. 6 and Table. 5. Among them, Fig. 6 presents a comprehensive analysis of the dataset properties of our training data, which includes a detailed distribution of caption length and bbox number per image. The caption length distribution reveals that the majority of captions fall within a range of 5 to 15 words, with a few outliers exceeding 15 words. On the other hand, the bbox number per image distribution shows that most images contain between 1 and 5 bounding boxes, with a small percentage of images having more than 10 bounding boxes. These statistics provide valuable insights into the nature of our training data and can be used to inform the design of our machine learning models.

In Table. 5, we compare the scale of different well-annotated image datasets with the training data used in the study. The number of images in the datasets ranges from 0.028 million to 11 million, while the number of entities ranges from 0.7 million to 1.1 billion. In Table. 5, we compare the scale

[1] https://huggingface.co/stabilityai/stable-diffusion-2-base

[2] https://github.com/mlfoundations/open_clip

[3] https://huggingface.co/datasets/ChristophSchuhmann/improved_aesthetics_5plus

Table 5: The comparison between well annotated image dataset and our training data. Image # , entity # and class # refer to the number of images, the number of entities and the number of class categories, respectively. SA-1B † does not provide the class label of instances.

Dataset	LVIS v1	COCO	ADE20K	Open Images	SA-1B †	SDD (ours)
Image #	0.120M	0.123M	0.028M	1M	11M	76M
Entity #	1.5M	0.9M	0.7M	2.7M	1.1B	222M
Class #	1200	91	2693	600	N/A	162K

of different annotated image datasets to the training data used in our study. The number of images in these datasets ranges from 28,000 to 11 million, with the entity count ranging from 700,000 to 1.1 billion. Although SA-1B (Kirillov et al., 2023) offers the highest entity count of 1.1 billion, it lacks annotated entity categories and tends to include small-sized masks, which is unsuitable for our image generation purposes. In contrast, the training dataset employed in this study comprises 76 million images and 220 million entities, making it the largest-scale dataset available. Furthermore, it is important to note that our study not only provides the number of entity classes but also highlights the superior diversity of our training data compared to other datasets. This diversity is crucial in enabling our model to comprehend and identify a wide range of reference objects in the open world. Our training data includes a vast array of entities, *i.e.* 162K kinds of entities, ranging from common objects such as animals and plants to more complex entities such as vehicles and buildings. This comprehensive dataset ensures that our model is equipped with the necessary knowledge to accurately identify and classify any reference object it encounters. Additionally, our study also takes into account the varying contexts in which these entities may appear, further enhancing the robustness and adaptability of our model. Overall, our research provides a comprehensive and diverse training dataset that enables our model to effectively understand and generate reference objects in the open world.

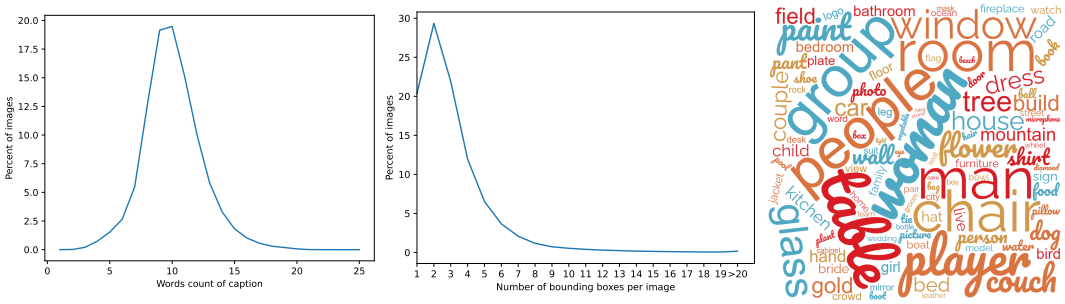


Figure 6: Dataset properties. Left: word count distribution of captions in SDD; Middle: bounding box count distribution of images in SSD; Right: Word cloud diagram of SDD. We can observe that the most frequent entities in our SDD are man, woman, people, table, room, *etc.*

B.3 Discussion on Quality of the Data

We collected 1000 data samples for statistics, and some of the figures are presented in Fig. 7. We also conducted an analysis of four columns of sample data, where the first three columns on the left are the data we selected for training after rule-based filtering, and the column on the right represents the data excluded by the filtering rules. The first column on the left shows high-quality data selected subjectively by the annotators, with filtering criteria consistent with our rule-based filtering motivations. The second column on the left shows low-quality results with low recall, i.e., many subject entities are not detected by the bounding box, possibly due to the generation of corresponding entities being incomplete in BLIP2 or insufficient recall by DINO. The third column on the left corresponds to other low-quality situations, which may include errors in subject identification, i.e., low accuracy, or situations missed by the rule-based filtering. Finally, we conducted a simple analysis of 1000 samples, as shown in Table 6, Subjectively high-quality results only accounted for 35% of

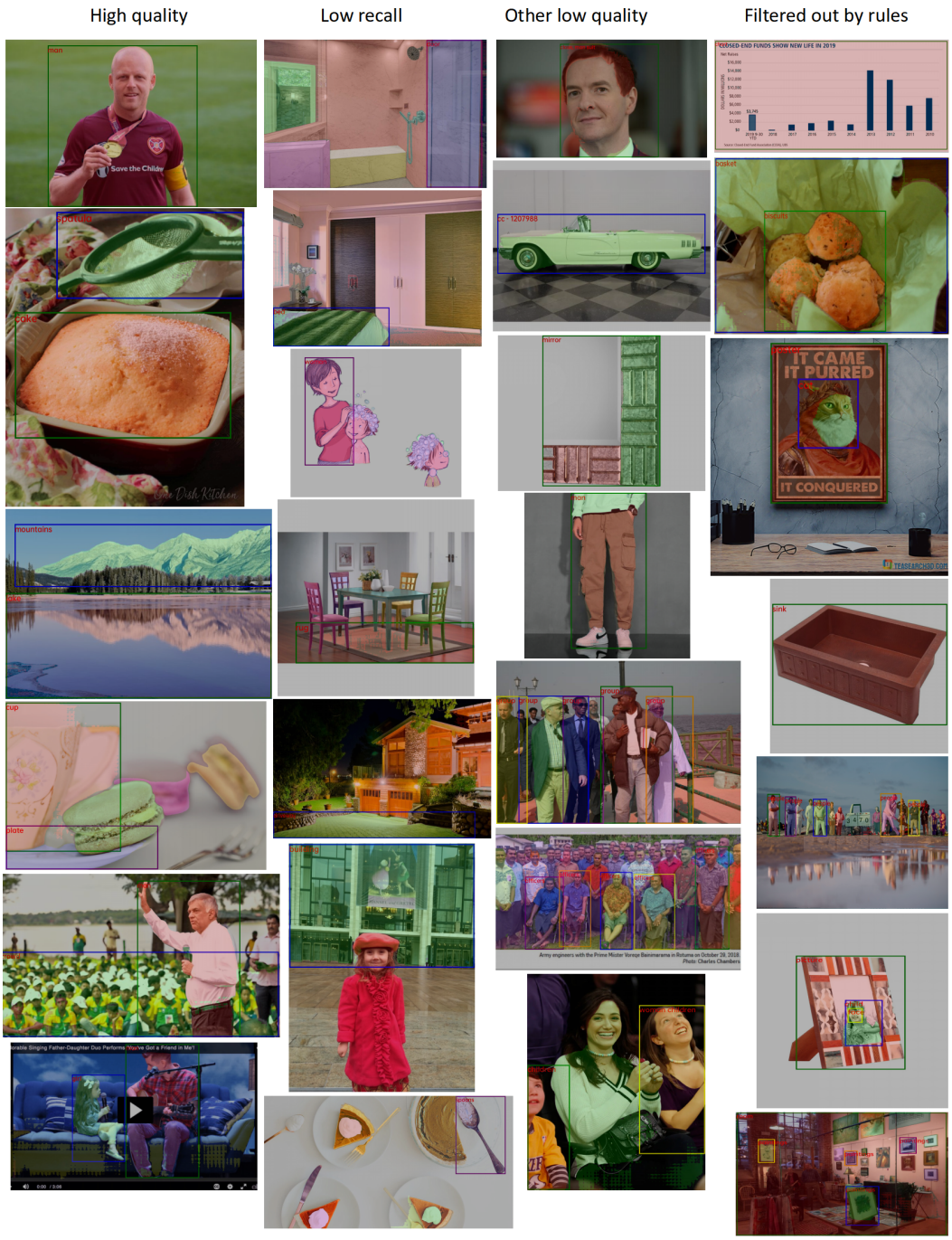


Figure 7: Example of data of different qualities.

the rule-based filtering results. This indicates that there is still a lot of potential to optimize data quality, and we will continue to work hard in this area.

C PERSONALIZATION BASELINES COMPARISON

We carefully survey the personalized image generation papers published in recent years and compile a comprehensive comparison table comparing their support for single reference image, multi-subject

Table 6: Subjective quantitative statistics of data quality.

Data Quality	**High Quality**	**Low Recall**	**Other Low Quality**	**Filtered Out by Rules**
Proportion	18%	20%	14%	48%

generation, no test-time fine-tuning, and open domain generalization. As delineated in Table 7, the main stream of personalized image generation still considers test-time fine-tuning, which suffers from inference time-consuming ranging from several seconds to more than one hour (Gal et al., 2022; Ruiz et al., 2023a; Kumari et al., 2023; Gal et al., 2023; Han et al., 2023b; Smith et al., 2023; Voynov et al., 2023; Liu et al., 2023b;c; Tewel et al., 2023; Chen et al., 2023a; Avrahami et al., 2023; Alaluf et al., 2023; Gu et al., 2023; Hao et al., 2023; Ruiz et al., 2023b; Arar et al., 2023; Zhou et al., 2023). Only a small portion of papers are dedicated to studying personalized image generation without test-time fine-tuning (Jia et al., 2023; Shi et al., 2023; Xiao et al., 2023; Chen et al., 2023c; 2022; Ma et al., 2023b; Wei et al., 2023; Li et al., 2023a; Chen et al., 2023b). But all of the pioneering works cannot satisfy the four aforementioned requirements, either by being trained on specific domains (Shi et al., 2023; Jia et al., 2023; Xiao et al., 2023), or by supporting only single-concept generation. To the best of our knowledge, our Subject-Diffusion is the first open-domain personalized image generation method that supports multi-concept synthesis and requires only a single reference image for each subject.

Table 7: Survey of recent personalized image generation works in terms of single reference image, multi-subject generation, no test-time fine-tuning and open domain generalization.

Method	Single image	Multi-subject	No fine-tuning	Open domain
Textual Inversion (Gal et al., 2022)	✗	✗	✗	-
Dreambooth (Ruiz et al., 2023a)	✗	✗	✗	-
Custom Diffusion (Kumari et al., 2023)	✗	✓	✗	-
E4T (Gal et al., 2023)	✓	✗	✗	-
SVDiff (Han et al., 2023b)	✓	✓	✗	-
Continual Diffusion (Smith et al., 2023)	✗	✓	✗	-
XTI (Voynov et al., 2023)	✗	✗	✗	-
Cones (Liu et al., 2023b)	✓	✓	✗	-
Cones 2 (Liu et al., 2023c)	✓	✓	✗	-
Perfusion (Tewel et al., 2023)	✗	✓	✗	-
DisenBooth (Chen et al., 2023a)	✓	✗	✗	-
Break-A-Scene (Avrahami et al., 2023)	✓	✓	✗	-
NeTI (Alaluf et al., 2023)	✗	✗	✗	-
Mix-of-Show (Gu et al., 2023)	✗	✓	✗	-
ViCo (Hao et al., 2023)	✗	✗	✗	-
HyperDreamBooth (Ruiz et al., 2023b)	✓	✗	✗	-
Domain-Agnostic (Arar et al., 2023)	✓	✗	✗	-
Regularization-Free (Zhou et al., 2023)	✓	✗	✗	-
Taming (Jia et al., 2023)	✓	✗	✓	✗
InstantBooth (Shi et al., 2023)	✓	✗	✓	✗
PhotoVerse (Chen et al., 2023b)	✓	✗	✓	✗
Face0 (Valevski et al., 2023)	✓	✗	✓	✗
FastComposer (Xiao et al., 2023)	✓	✓	✓	✗
SuTI (Chen et al., 2023c)	✗	✗	✓	✓
Re-Imagen (Chen et al., 2022)	✓	✗	✓	✓
UMM-Diffusion (Ma et al., 2023b)	✓	✗	✓	✓
ELITE (Wei et al., 2023)	✓	✗	✓	✓
Blip-Diffusion (Li et al., 2023a)	✓	✗	✓	✓
Ours (Subject-Diffusion)	✓	✓	✓	✓

D Two-Subject Evaluation Details

We utilize all the objects in DreamBench and randomly select 30 pairs of combinations, out of which 9 pairs belong to live objects. The specific subject pairs are presented in Table 8. For the prompts used in generating images with two subjects, we follow the format outlined in DreamBench, with the two subjects connected using the word "and".

For inference, we use PNDM scheduler for 50 denoising steps. We use a fixed text guidance scale 3 and image guidance scale 1.5 for all experiments

Table 8: Prompts for a dual-subject personalized image generation testset. The first 21 combinations are still objects, and the last 9 combinations are animals.

backpack-can	bear_plushie-backpack_dog	berry_bowl-vase
duck_toy-can	fancy_boot-shiny_sneaker	grey_sloth_plushie-poop_emoji
teapot-backpack_dog	teapot-berry_bowl	wolf_plushie-backpack_dog
can-bear_plushie	can-candle	can-duck_toy
can-shiny_sneaker	clock-teapot	colorful_sneaker-vase
robot_toy-backpack	shiny_sneaker-duck_toy	shiny_sneaker-poop_emoji
pink_sunglasses-candle	poop_emoji-clock	poop_emoji-shiny_sneaker
cat-dog2	cat-dog5	cat2-dog3
dog2-dog3	dog5-dog6	dog6-dog7
dog6-dog8	dog7-dog8	dog8-dog6

E Text-Image Interpolation

The visualization examples can be found in Fig. 8. We provide this experiment to show that the high-level information of the user-provided images are successfully extracted and rendered in generated images during early backward diffusion stages. Thus we can adjust α to balance image fidelity and editablity according to different prompts.

Figure 8: Text-image interpolation. The prompts are followings: *A man in the rain, the woman is [PH]*; *A dog in the snow, the cat is [PH]*; *A wolf plushie on the beach, the lion is [PH]*.

F ADDITIONAL QUALITATIVE RESULTS OF THE ABLATION STUDIES

In the case of a single subject, Fig. 10 left two columns present two examples that clearly demonstrate the higher fidelity of the generated images without box coordinates. However, these images have lower semantic matching ability and are unable to capture key information from the prompts. On the other hand, images generated without the adapter layer and without image cls feature have slightly lower fidelity. These two strategies aim to enhance the processing of input image information, providing advantages in both objective metrics and subjective evaluation in terms of fidelity.

Regarding the case of two subjects with Fig. 10 right two columns, the conclusions remain consistent with the previous analysis. Images generated without the adapter layer and without image cls feature still exhibit slightly lower fidelity. It is worth mentioning that both the preservation of box coordinates and attention map control have advantages in generating images with multiple subjects, as these conditions help alleviate the issue of generating ambiguous representations of multiple entities.

Figure 9: Example of failure generations.

G FAILURE CASES DISCUSSIONS

We provide an example to address the shortcomings of "editing attributes" and "rendering harmonious images with two subjects". For the "editing attributes" issue, the attributes corresponding to the red-marked prompts in the failed image are highlighted. As for generating images with two subjects, if the source image(s) itself already lacks one or both of the subjects, it may lead to disharmony in the final generated image.The cases are shown in Fig. 9.

H DISCUSSIONS WITH METHODS TRAINED ON THE IMAGEN

From Table 9, We have compared our method with Imagen-based methods, including Re-Imagen and SuTI. Re-Imagen is a retrieval-augmented approach that also achieves personalized image (retrieved reference image) generation. SuTI is a subject-driven text-to-image generator that replaces subject-specific fine tuning with in-context learning. We can see that SuTI has an advantage in all three metrics. However, it may not be fair to make direct comparisons between the two methods based

Figure 10: Additional qualitative results of the ablation studies.

solely on these results.Three issues that need to be discussed are as follows: First, the difference in the base model used, where SuTI is based on the Imagen model structure and Initialization parameters, while our base model is SD2. Second, the image resolution evaluated for SuTI was 1024, while our evaluated image resolution was 512. Third, SuTI provides four demonstration image-text pairs during inference, while we only provide one.

we will compare our results with SuTI in a qualitative side-by-side comparison in Figure 11. We made a simple comparison on the four shortcomings of SuTI:

(1) SuTI has a strong prior about the subject and hallucinates the visual details based on its prior knowledge. For example, the generation model believes 'teapot' should contain a 'lift handle'. (2) Some artifacts from the demonstration images are being transferred to the generated images like second column.Subject Diffusion has advantages in this regard because it removes background input. (3) The subject's visual appearance is being modified through with SuTI, mostly influenced by the context, like last column. Subject Diffusion will be slightly better. (4) SuTI is not particularly good at handling compositional prompts like the 'sunglasses' example like third column. Subject Diffusion will be slightly better.

Table 9: Quantitative single subject results. † indicates experimental results referenced from SuTI. Boldface indicates the best results of zero shot approaches evaluated in DeramBench.

Methods	Model Base	Testset	DINO	CLIP-I	CLIP-T
Real Images †	-	-	0.774	0.885	-
Re-Imagen †	Imagen	DB	0.600	0.740	0.270
SuTI †	Imagen	DB	**0.741**	**0.819**	**0.304**
Subject-Diffusion	SD	DB	0.711	0.787	0.293

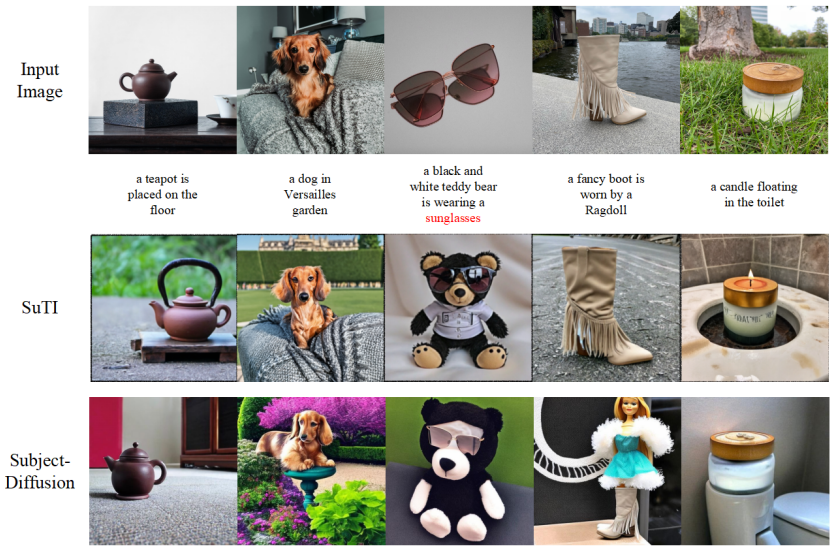


Figure 11: Compare our results with SuTI in qualitative.

I MORE VISUALIZATION RESULTS

In this section, we provide more single-, multi-, and human subject generation visualization examples, as in Fig. 12, Fig. 13 and Fig. 14. Notice that we display 10 generated results for each personal image without carefully cherry-picking, demonstrating the consistent fidelity and generalization ability of our proposed Subject-Diffusion.

Input Image

a backpack in the jungle | a backpack in the snow | a backpack on the beach | a backpack on a cobblestone street | a backpack on top of pink fabric

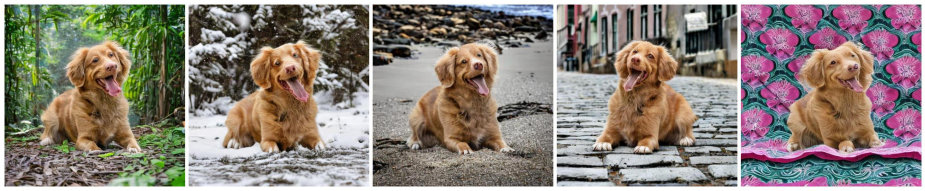
a backpack on top of a wooden floor | a backpack with a city in the background | a backpack with a mountain in the background | a backpack with a blue house in the background | a backpack on top of a purple rug in the forest

Input Image

a cat in the jungle | a cat in the snow | a cat on the beach | a cat on a cobblestone street | a cat on top of pink fabric

a cat on top of a wooden floor | a cat with a city in the background | a cat with a mountain in the background | a cat with a blue house in the background | a cat on top of a purple rug in the forest

Input Image



a dog in the jungle | a dog in the snow | a dog on the beach | a dog on a cobblestone street | a dog on top of pink fabric

a dog on top of a wooden floor | a dog with a city in the background | a dog with a mountain in the background | a dog with a blue house in the background | a dog on top of a purple rug in the forest

Input Image

a toy in the jungle | a toy in the snow | a toy on the beach | a toy on a cobblestone street | a toy on top of pink fabric

a toy on top of a wooden floor | a toy with a city in the background | a toy with a mountain in the background | a toy with a blue house in the background | a toy on top of a purple rug in the forest

Figure 12: More qualitative results for single-subject generation.

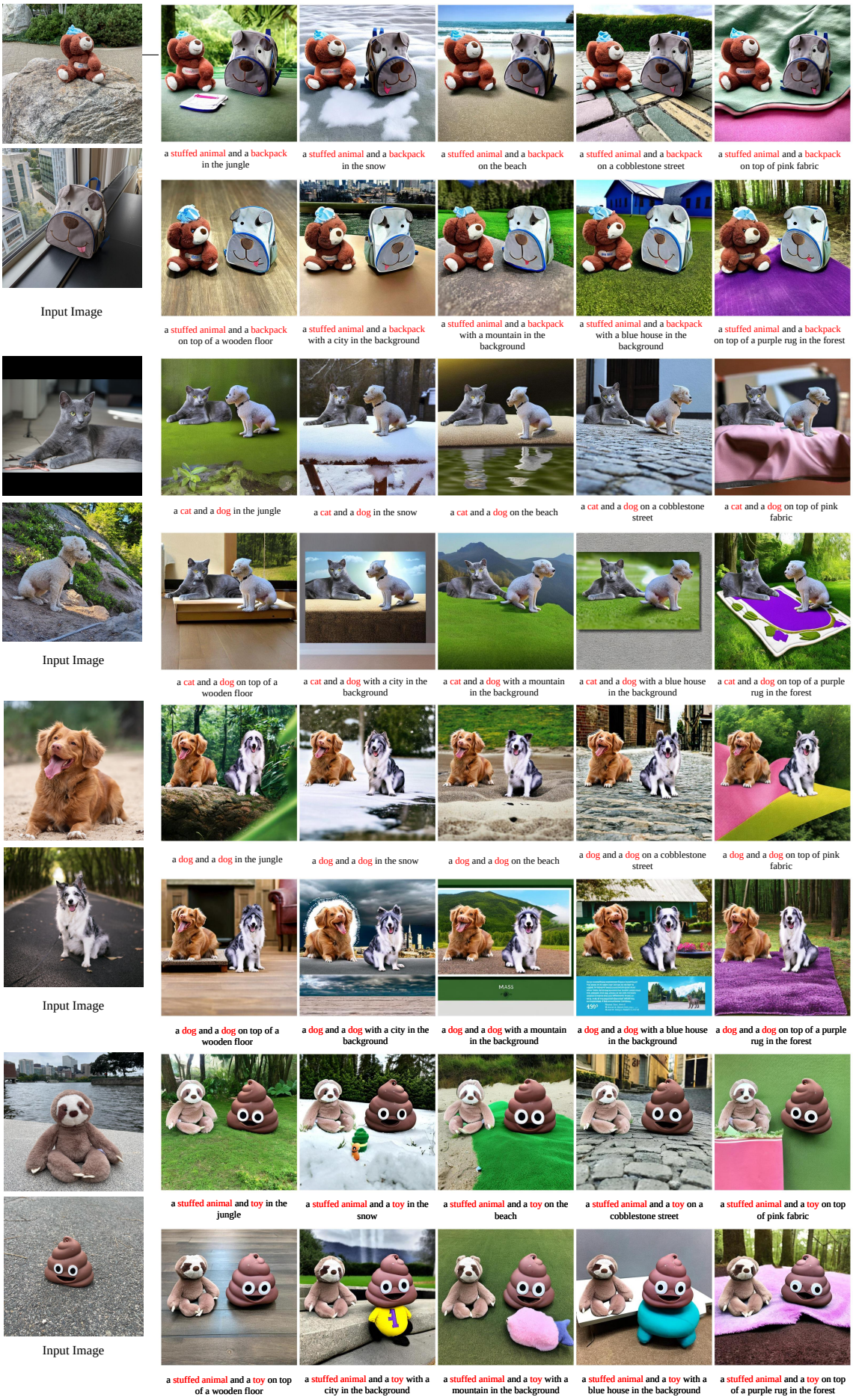


Figure 13: More qualitative results for two-subject generation.

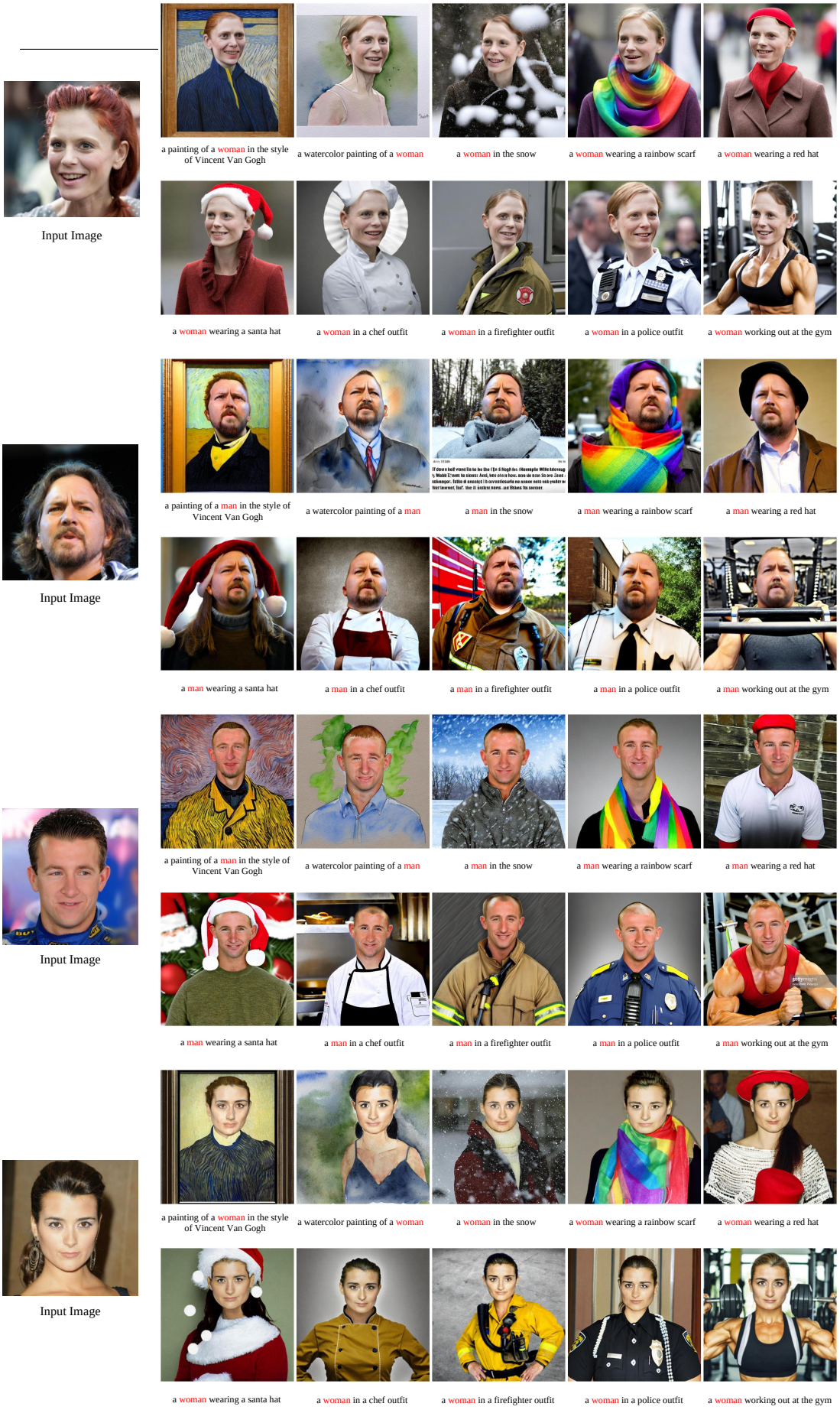


Figure 14: More qualitative results for human image generation.