NoiseCollage: A Layout-Aware Text-to-Image Diffusion Model Based on Noise Cropping and Merging

Supplementary Material

This supplementary material shows additional results generated by NoiseCollage. Each of the following figures shows the layout condition L and the text conditions (S, s_*), the generated image x_0, and N individual object images cropped from x_0 by $l_1, \ldots, l_N$, from top to bottom. These cropped images not only show the detailed appearance of the individual objects but also show whether the objects are generated at their right place.

8. Total Inference step and Time efficiency

NoiseCollage (also Collage Diffusion [31]) requires $O(NT)$-times noise estimations, whereas Paint-with-words requires $O(T)$-times where N and T denote the number of objects in a layout and total denoising step, respectively. Therefore, in NoiseCollage, the total inference step becomes $(N+1)*T$ including noise estimation for the whole image then the number of total inference step increases with the number of objects in a layout. In fact, NoiseCollage needs 25.7s to generate a single image with $N = 5$, whereas Paint-with-words needs 7.42s on a single Nvidia A100GPU.

However, from an optimistic perspective, NoiseCollage can be refined through parallelization, as the $O(N)$-times noise estimations at each time step t are entirely independent. Consequently, it can be executed with $O(T)$ computations.

9. Good and bad cases in the results of NoiseCollage

Figs. 8 and 9 show the images generated by NoiseCollage on the MD30 dataset. The former shows images with good scores, while the latter shows images with bad scores, according to the evaluation metric (multimodal similarity between the n-th object image and its text condition s_n) of Sec. 4.5. The good cases of Fig. 8 show the accurate correspondence between s_n and l_n, even the layout l_n is specified by a bounding box.

Even in the worst cases of Fig. 9, large objects (specified by l_1 and s_1) are generated at the right place with a correct appearance (except for the bicycle image). Except for the remote control that disappears from the generated image, small objects are generated in misaligned locations and still look good.

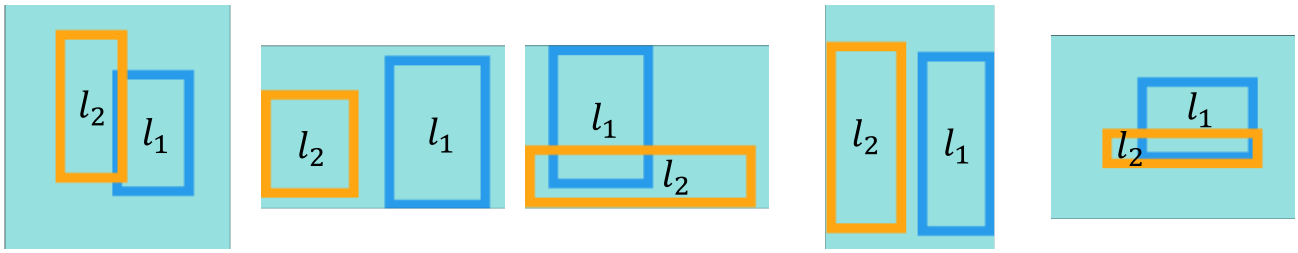
10. More Results of NoiseCollage with ControlNet

While we already showed several results of NoiseCollage with ControlNet [40] in Fig. 6, we show more results in Figs. 10, 11, and 12, which uses edge images, sketches, and pose skeletons as additional constraints, respectively. Like the results in Fig. 6, the additional results also show how the conditions for ControlNet guide the output images accurately. Note that bounding boxes specify the layout conditions, whereas polygons are used in Fig. 6. We can confirm that bounding boxes are also easy but appropriate layout conditions.

11. Results of more crowded layouts

As already stated in the "Limitations" section, it is difficult for not only ours but also baselines to generate images under complex layouts with small objects or a large number of objects. This limitation may come from the fact that the common backbone, StableDiffusion [28], uses a low-resolution latent space of 64×64.

Fig. 13 shows the generated images with more crowded objects ($N = 7, 9$) by NoiseCollage with ControlNet. In these examples, the layout conditions are well reflected in the resulting images. However, if we want to put more objects, say, $N = 20$, it is difficult to expect the accurate reflection of their conditions, as noted above.



s_1 a large brown dog with a green collar on its neck. s_2 a red fire hydrant. s_* a dog sniffing a fire hydrant in a park.	s_1 a black and white dog on a leash sitting on the ground. s_2 a close up of a pink park bench. s_* a dog sitting on a leash next to a bench.	s_1 a young boy in white and black swim suit on a surfboard. s_2 a red surfboard on the water. s_* a man riding a surfboard in the ocean.	s_1 the back of a zebra. s_2 the back of a zebra. s_* two zebras standing next to each other in front of a fence.	s_1 a muscular man on a surfboard. s_2 a white surfboard in the water. s_* A man surfing on a surfboard in the ocean.

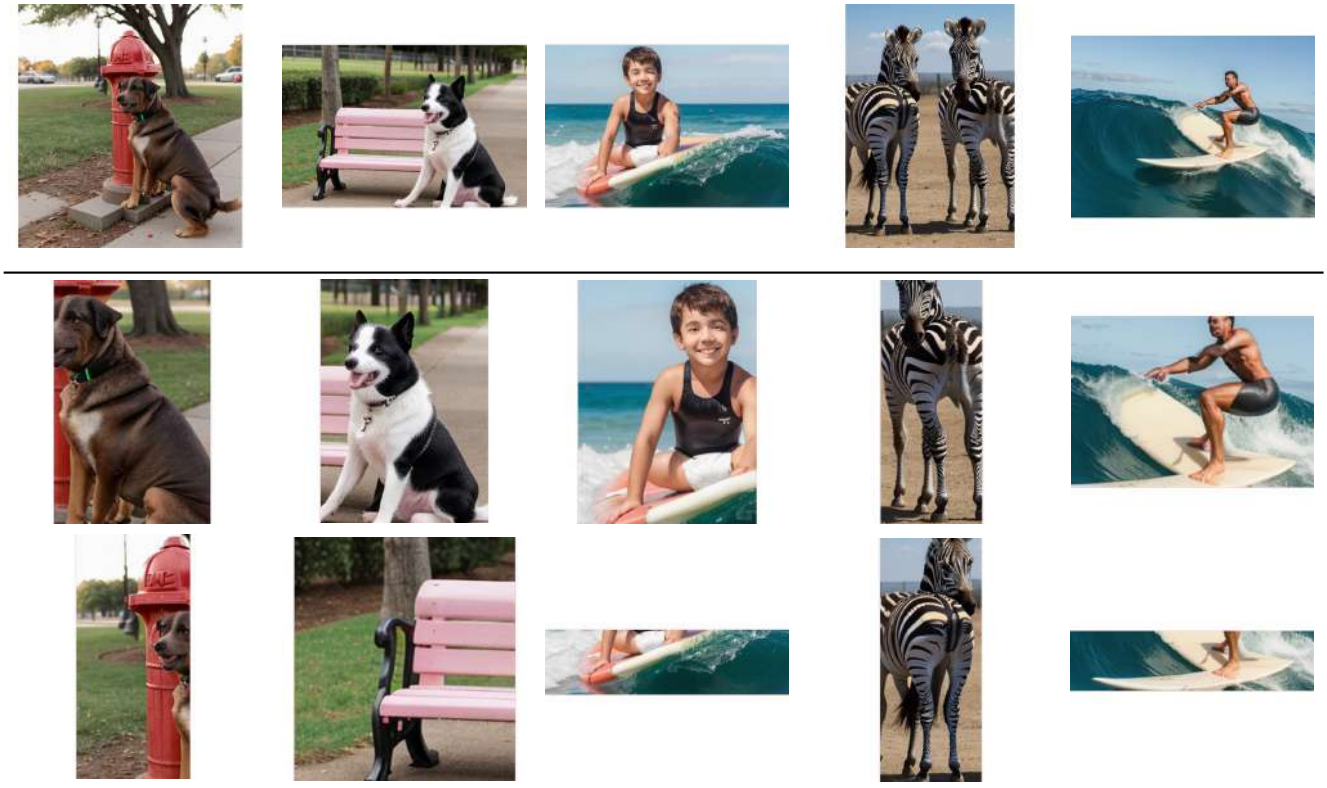
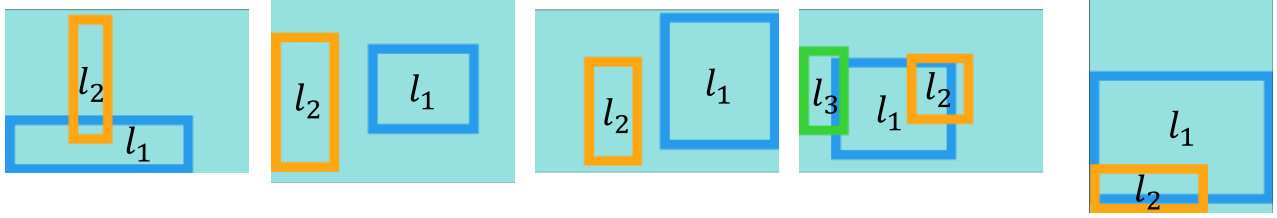


Figure 8. The best five cases by NoiseCollage on MD30. The lower part shows N individual object images cropped from x_0 by $l_1, \ldots, l_N$.



s_1 a black horse with a saddle on its back.
s_2 a standing young man.
s_* a black and white photo of a man riding on the back of a horse.

s_1 a bicycle with a basket on the front.
s_2 a close up photo that a man with helmet is riding on a motorcycle.
s_* a man riding a motor bike and a person is standing next to two parked bikes on a sidewalk.

s_1 a brown horse is standing in a field of grass.
s_2 a woman with green hair, sunglass, red rain boots.
s_* a woman standing next to a horse in a garden.

s_1 a bench.
s_2 a plant in a vase.
s_3 a plant in a vase.
s_* a bench sitting in front of a building.

s_1 a couch with pillows on it in a living room.
s_2 a remote control sitting on top of a round table.
s_* a living room with a couch and a table.

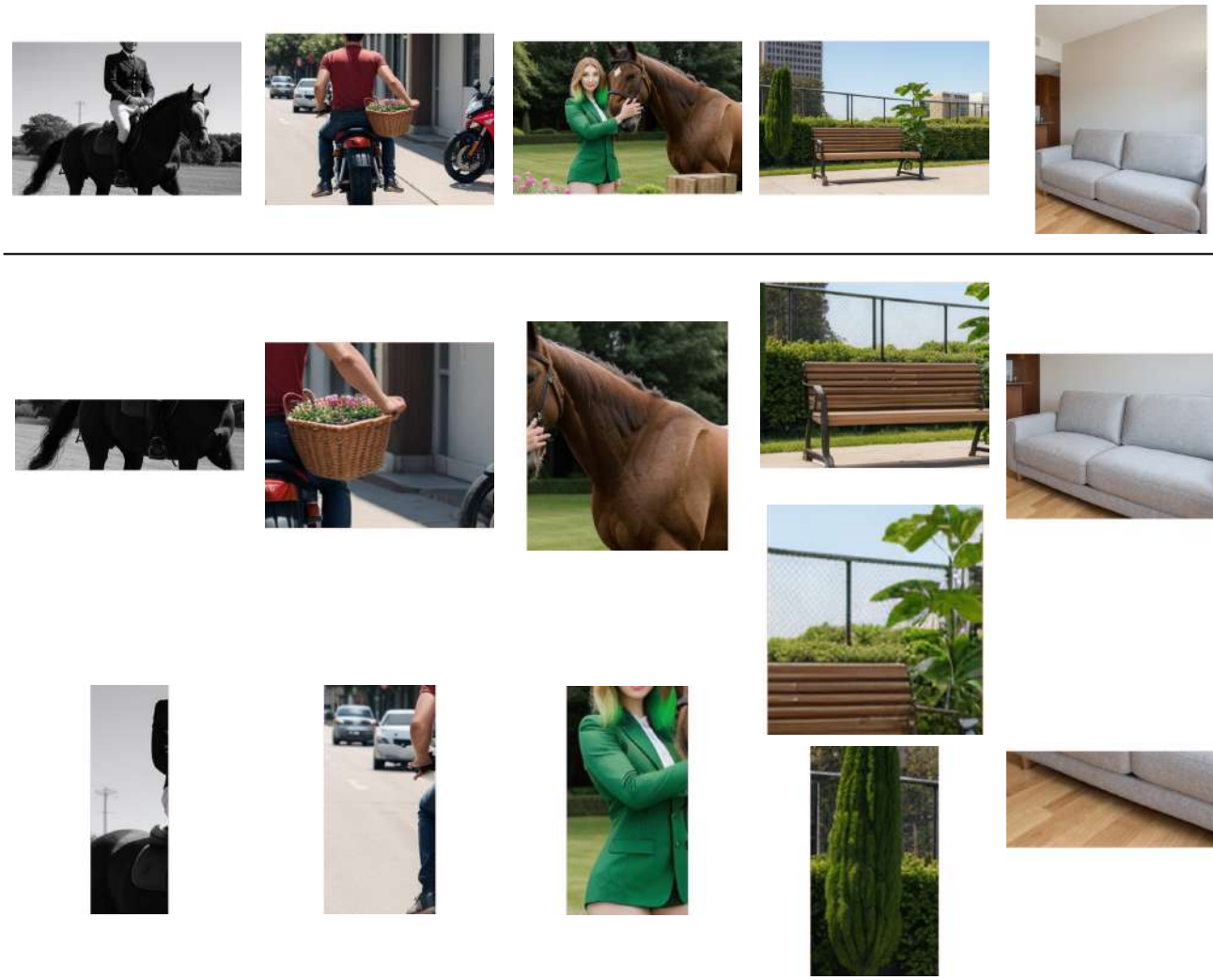
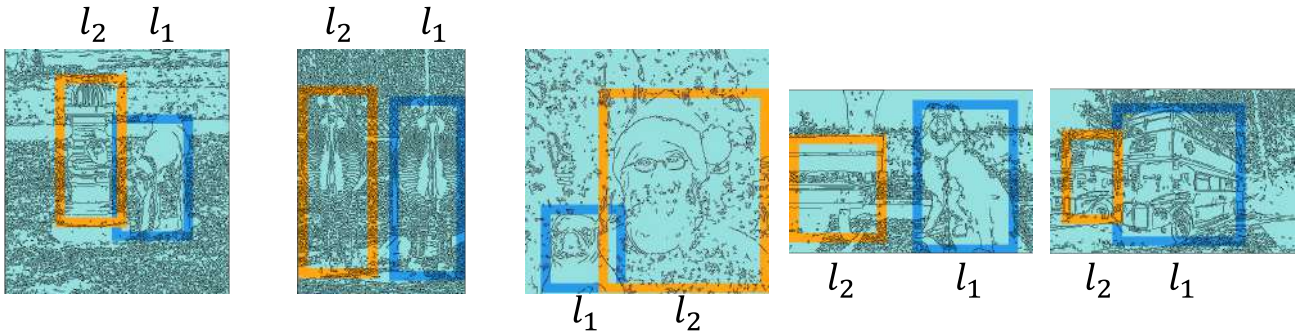


Figure 9. The worst five cases by NoiseCollage on MD30.



s_1 a large brown dog with a green collar on its neck.
s_2 a red fire hydrant.
s_* a dog sniffing a fire hydrant in a park.

s_1 the back of a zebra.
s_2 the back of a zebra.
s_* two zebras standing next to each other in front of a fence.

s_1 a close up of a white dog with red eyes.
s_2 a man with a beard wearing a santa hat.
s_* a man in a santa hat with a dog in front of a christmas tree.

s_1 a black and white dog on a leash sitting on the ground.
s_2 a close up of a pink park bench.
s_* a dog sitting on a leash next to a bench.

s_1 a green double decker bus is parked.
s_2 a red double decker bus is parked.
s_* two double decker buses are parked in front of trees.

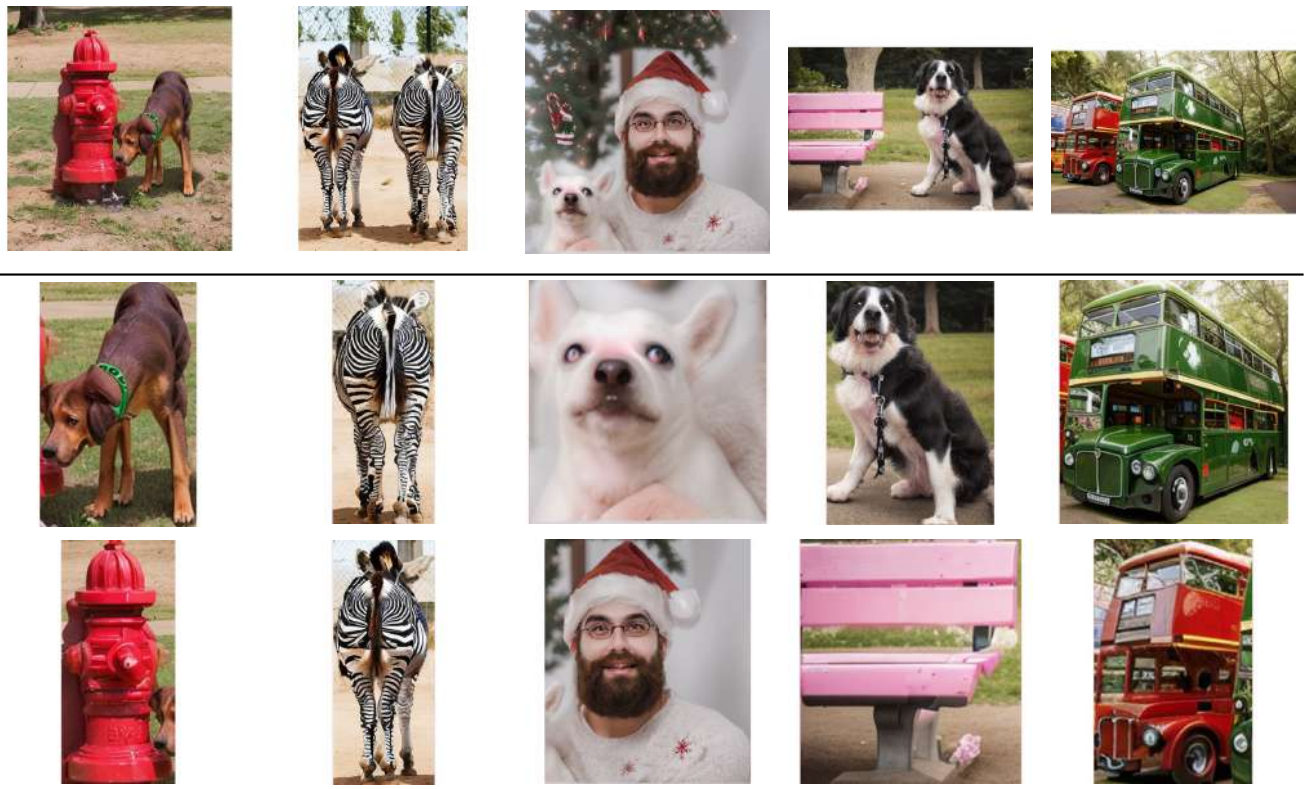
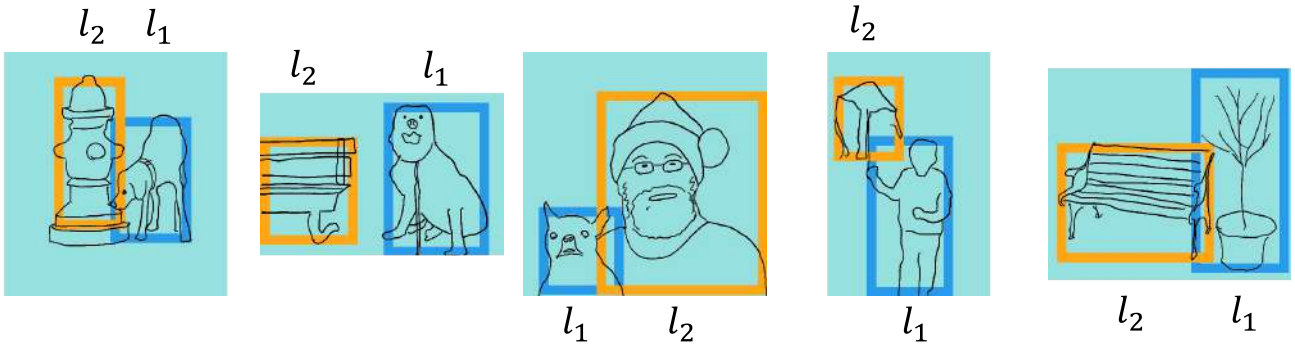


Figure 10. Images generated by NoiseCollage with ControlNet[40] of an edge image condition.



s_1 a large brown dog with a green collar on its neck.
s_2 a red fire hydrant.
s_* a dog sniffing a fire hydrant in a park.

s_1 a black and white dog on a leash sitting on the ground.
s_2 a close up of a pink park bench.
s_* a dog sitting on a leash next to a bench.

s_1 a close up of a white dog with red eyes.
s_2 a man with a beard wearing a santa hat.
s_* a man in a santa hat with a dog in front of a christmas tree.

s_1 a young boy is running.
s_2 a rainbow colored kite flying in the air.
s_* a young boy is flying a kite in a field.

s_1 a potted plant.
s_2 a wooden and metal park bench.
s_* a wooden bench sitting next to a potted plant.

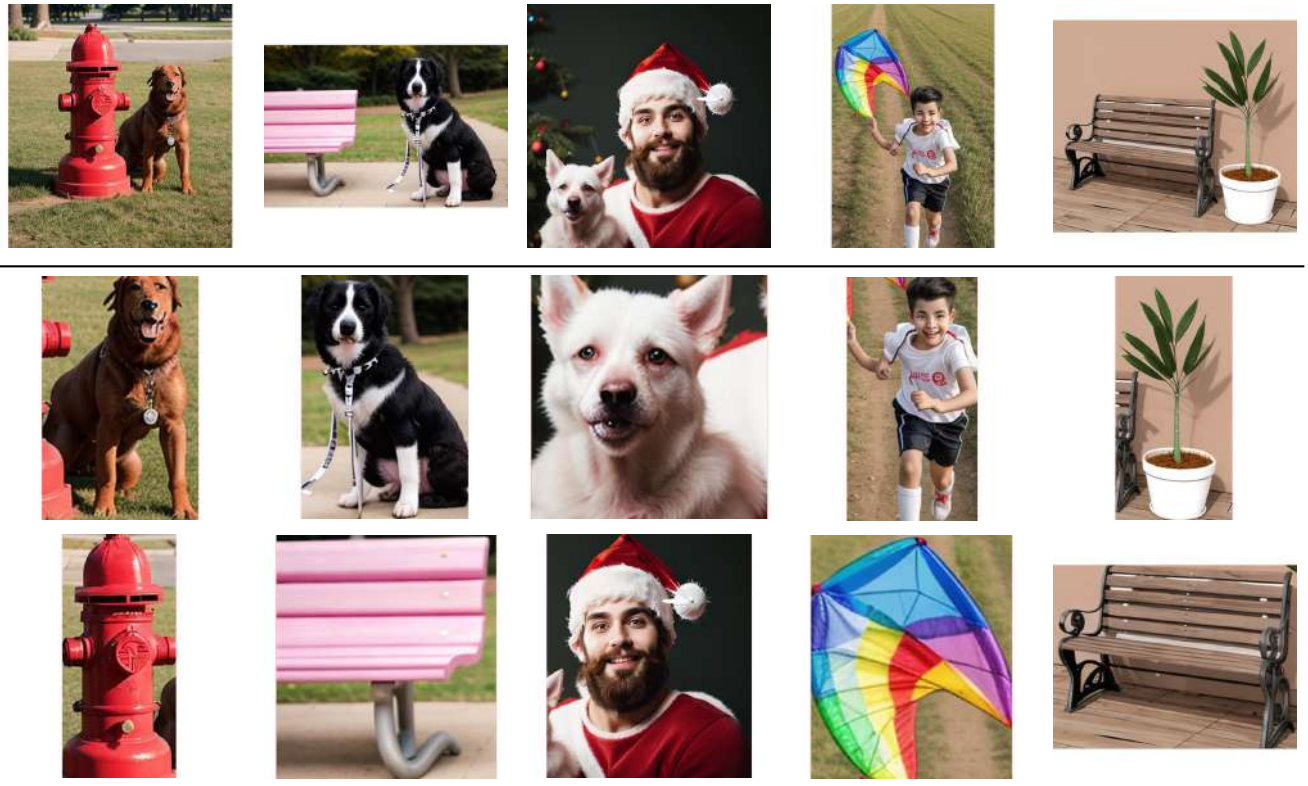
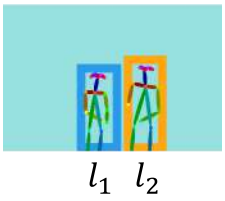
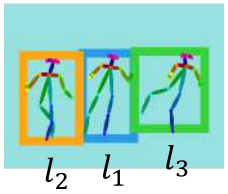


Figure 11. Images generated by NoiseCollage with ControlNet[40] of a sketch condition.

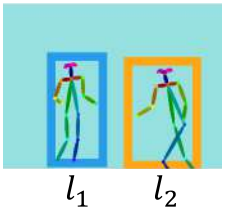
s_1 a boy wearing a jeans, shirt and red tie is jumping in the air.
s_2 a boy wearing a jeans, shirt and blue bow tie is jumping in the air.
s_3 a boy wearing a jeans, shirt and blue tie is standing.
s_* three boys wearing a blue jeans, pink shirts, and ties jumping in the air with different pose.



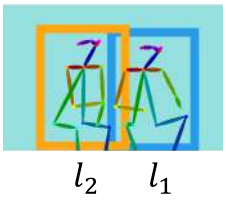
s_1 a woman wearing a dark red apron holding a pizza.
s_2 a man wearing a horizontal-striped shirt and pants holding a pizza in his hand.
s_* a man and woman holding a pizza in a restaurant.



s_1 a woman wearing a black jacket and white pants standing on skis with a smile on her face.
s_2 a woman wearing white jacket and black pants standing on skis with goggles on her face.
s_3 a woman wearing red jacket and black pants standing on skis with goggles on her head.
s_* three people on skis are posing for a picture in font of a forest in the snow.



s_1 a woman wearing blue jacket and white pants on skis posing for a picture on top of a snow.
s_2 a man wearing navy pants and khaki jacket on skis is standing on top of a snow.
s_* a man and woman in ski gear standing in front of a snow mountain.



s_1 a man wearing jeans and green shirt is sitting on a chair with a wii controller.
s_2 a woman wearing pants and red T-shirt with curly hair is sitting on a chair with a wii controller.
s_* a man and woman sitting on a couch playing a game.









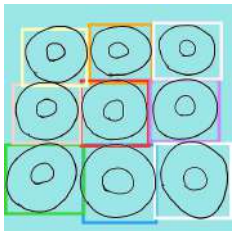




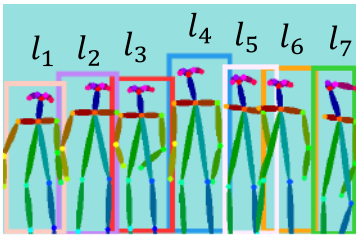




Figure 12. Images generated by NoiseCollage with ControlNet[40] of a pose skeleton condition.



s_1, s_2, s_3 a purple donut
s_4, s_5, s_6 a yellow donut
s_7, s_8, s_9 a chocolate donut
s_* a box of donuts with different color



s_1 a woman in a black dress
s_2 a man in a shirt and tie
s_3 a woman in a white wedding dress
s_4 a man in a navy suit
s_5 a man in a shirt and tie
s_6 a woman in a blue dress
s_7 a young man in a shirt and jeans
s_* a group of people posing for a picture in a room

Figure 13. Generated images with more complex layouts. Ours can control the layout of donuts and the clothes of each person.