FaceChain-SuDe: Building Derived Class to Inherit Category Attributes for One-shot Subject-Driven Generation

Supplementary Material

7. Overview

We provide the dataset details in Sec. 8. Besides, we discuss the limitation of our SuDe in Sec. 9. For more empirical results, the details about the baselines' generations are in Sec. 10.2, comparisons with offline method are in Sec. 10.3, more qualitative examples in Sec. 10.4, and the visualizations on more applications are in Sec. 10.5.

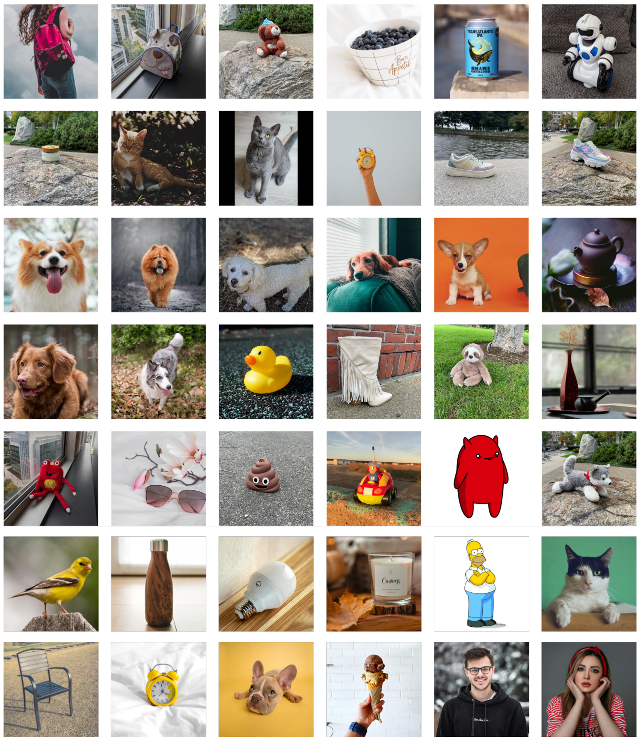


Figure 8. Subject image examples.

8. Dataset Details

8.1. Subject images

For the images from the DreamBench [30], which contains 30 subjects and 5 images for each subject, we only use one image (numbered '00.jpg') for each subject in all our experiments. All the used images are shown in Fig. 8.

8.2. Prompts

We collect 5 attribute-related prompts for all the 30 subjects. The used prompts are shown in Table 2.

9. Limitation

9.1. Inherent failure cases

As in Fig. 9, the text characters on the subject cannot be kept well, for both baselines w/ and w/o SuDe. This is an inherent failure of the stable-diffusion backbone. Our SuDe is designed to inherit the capabilities of the pre-trained model itself and therefore also inherits its shortcomings.

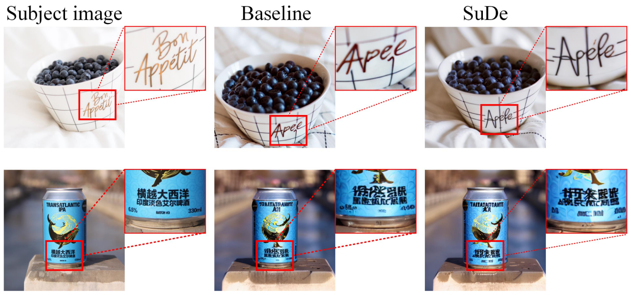


Figure 9. **Reconstruction results of texts.** The baseline here is Dreambooth [30], and the prompt is 'photo of a S^*'.

9.2. Failure cases indirectly related to attributes

As Fig. 10, the baseline model can only generate prompt-matching images with a very low probability (1 out of 5) for the prompt of 'wearing a yellow shirt'. For our SuDe, it performs better but is also not satisfactory enough. This is because 'wearing a shirt' is not a direct attribute of a dog, but is indirectly related to both the dog and the cloth. Hence it cannot be directly inherited from the category attributes, thus our SuDe cannot solve this problem particularly well.

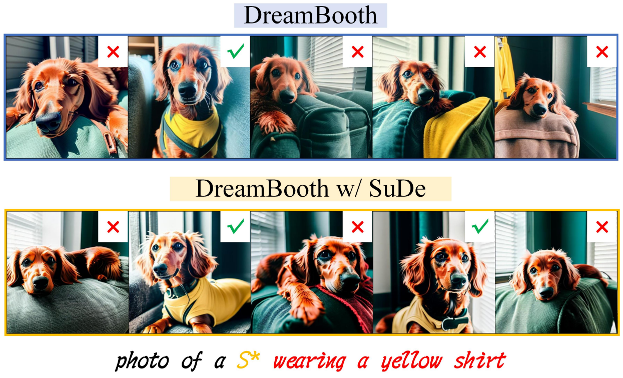
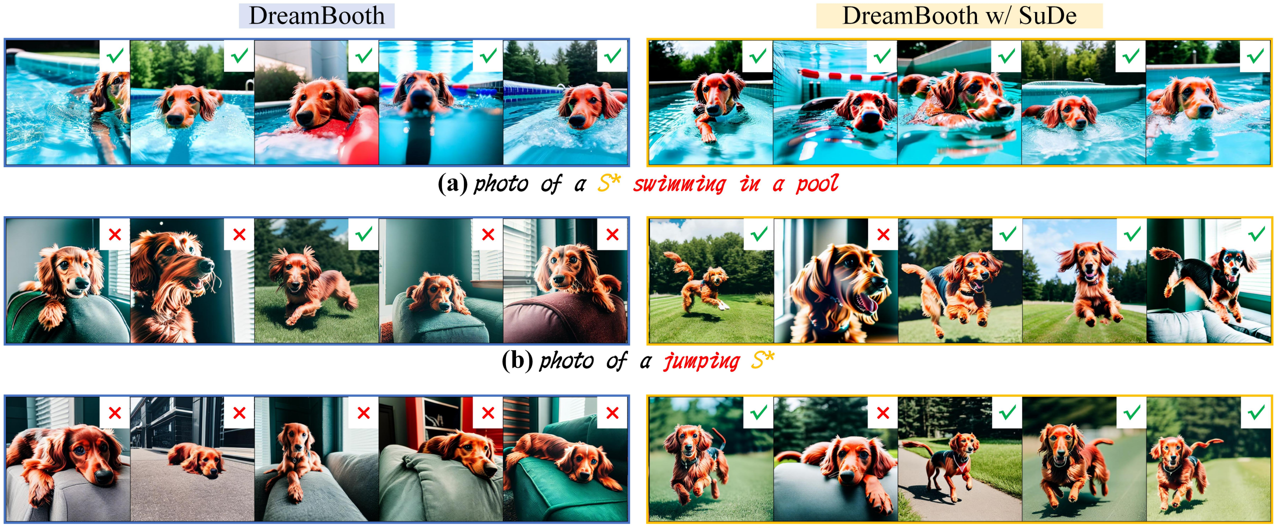


Figure 10. The 5 images are generated with various initial noises.

Table 2. Prompts for each subject.

Class	Backpack	Stuffed animal	Bowl	Can	Candle
Prompt 1	'photo of a blue {}'	'photo of a blue {}'	'photo of a blue {}'	'photo of a blue {}'	'photo of a burning {}'
Prompt 2	'photo of a green {}'	'photo of a green {}'	'photo of a green {}'	'photo of a green {}'	'photo of a cube shaped unburned {}'
Prompt 3	'photo of a yellow {}'	'photo of a yellow {}'	'photo of a metal {}'	'photo of a yellow {}'	'photo of a cube shaped burning {}'
Prompt 4	'photo of a fallen {}'	'photo of a fallen {}'	'photo of a shiny {}'	'photo of a shiny {}'	'photo of a burning {} with blue fire'
Prompt 5	'photo of a dirty {}'	'photo of a wet {}'	'photo of a clear {}'	'photo of a fallen {}'	'photo of a blue{}'
	Cat	Clock	Sneaker	Toy	Dog
	'photo of a running {}'	'photo of a blue {}'	'photo of a blue {}'	'photo of a blue {}'	'photo of a running {}'
	'photo of a jumping {}'	'photo of a green {}'	'photo of a green {}'	'photo of a green {}'	'photo of a jumping {}'
	'photo of a yawning {}'	'photo of a yellow {}'	'photo of a yellow {}'	'photo of a yellow {}'	'photo of a crawling {}'
	'photo of a crawling {}'	'photo of a shiny {}'	'photo of a red {}'	'photo of a shiny {}'	'photo of a {} with open mouth'
	'photo of a {} climbing a tree'	'photo of a fallen {}'	'photo of a white {}'	'photo of a wet {}'	'photo of a {} playing with a ball'
	Teapot	Glasses	Boot	Vase	Cartoon character
	'photo of a blue {}'	'photo of a blue {}'	'photo of a blue {}'	'photo of a blue {}'	'photo of a running {}'
	'photo of a shiny {}'	'photo of a green {}'	'photo of a green {}'	'photo of a green {}'	'photo of a jumping {}'
	'photo of a clear {}'	'photo of a yellow {}'	'photo of a yellow {}'	'photo of a shiny {}'	'photo of a {} swimming in pool'
	'photo of a cube shaped {}'	'photo of a red {}'	'photo of a shiny {}'	'photo of a clear {}'	'photo of a {} sleeping in bed'
	'photo of a pumpkin shaped {}'	'photo of a cube shaped {}'	'photo of a wet {}'	'photo of a cube shaped {}'	'photo of a {} driving a car'



(a) *photo of a* S^* *swimming in a pool*

(b) *photo of a jumping* S^*

(c) *photo of a running* S^*

Figure 11. The subject image here is the dog shown in Fig. 8 line 3 and column 4. These results are generated by various initial noises.

Table 3. **The BLIP-T computed with various prompt templates.** The P_0 is the baseline's default prompt of 'photo of a [attribute] {S*}', and P_1 to P_3 are described in Sec. 10.1.

Prompt	P_0	P_1	P_2	P_3
ViCo [13]	39.1	40.8	40.9	41.2
w/ SuDe	43.3 (+4.2)	43.4 (+2.6)	43.1 (+2.2)	42.7 (+1.5)

10. More Experimental Results

10.1. Compare with modifying prompt

Essentially, our SuDe enriches the concept of a subject by the public attributes of its category. A naive alternative to realize this is to provide both the subject token and category token in the text prompt, e.g., 'photo of a {S*} [category]', which is already used in the DreamBooth [30] and Custom Diffusion [17] baselines. The above comparisons on these two baselines show that this kind of prompt cannot tackle the attribute-missing problem well. Here we further evaluate the performances of other prompt projects on the ViCo baseline, since its default prompt only contains the subject token. Specifically, we verify three prompt templates: $\boldsymbol{P_1}$: 'photo of a [attribute] {S*} [category]', $\boldsymbol{P_2}$: 'photo of a [attribute] {S*} and it is a [category]', $\boldsymbol{P_3}$: 'photo of a {S*} and it is a [attribute] [category]'. Referring to works in prompt learning [19, 22, 33, 35], we retained the triggering word structure in these templates, the form of 'photo of a {S*}' that was used in subject-driven finetuning.

As shown in Table 3, a good prompt template can partly alleviate this problem, e.g., $\boldsymbol{P_3}$ gets a BLIP-T of 41.2. But there are still some attributes that cannot be supplied by modifying prompt, e.g., in Fig. 12, $\boldsymbol{P_1}$ to $\boldsymbol{P_3}$ cannot make the dog with ‘open mouth’. This is because they only put both subject and category in the prompt, but ignore modeling their relationships like our SuDe. Besides, our method can also work on these prompt templates, as in Table 3, SuDe further improves all prompts by over 1.5%.

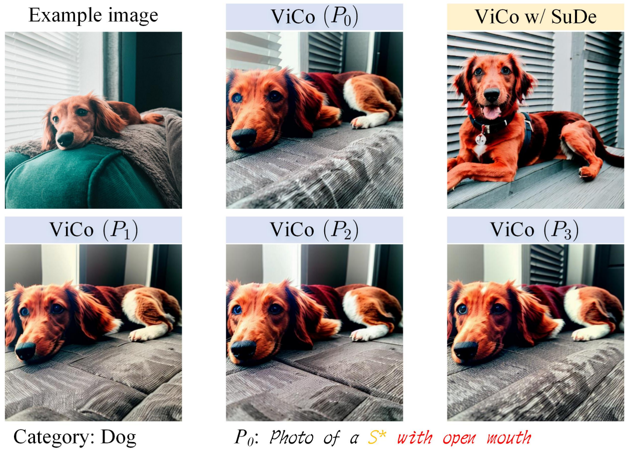


Figure 12. **Generations with various prompts.** The subject is a dog and the attribute we want to edit is ‘open mouth’. $\boldsymbol{P}_0$ is the default prompt, and $\boldsymbol{P}_1$ to $\boldsymbol{P}_3$ are described in Sec. 10.1.

10.2. Details about the generations of baselines

In the figures of the main manuscript, we mainly demonstrate the failure cases of the baseline, and our SuDe improves these cases. In practice, baselines can handle some attribute-related customizations well, as shown in Fig. 11 (a), and our SuDe can preserve the strong ability of the baseline on these good customizations.

For the failures of baselines, they could be divided into two types: **1)** The baseline can only generate prompt-matching images with a very low probability, as Fig. 11 (b). **2)** The baseline cannot generate prompt-matching images, as Fig. 11 (c). Our SuDe can improve both of these two cases, for example, in Fig. 11 (c), 4 out of 5 generated images can match the prompt well.

10.3. Compare with offline method

Here we evaluate the offline method ELITE [41], which encodes a subject image to text embedding directly with an offline-trained encoder. In the inference of ELITE, the mask annotation of the subject is needed. We obtain these masks by Grounding DINO [20]. The results are shown in Table 4, where we see the offline method performs well in attribute alignment (BLIP-T) but poorly in subject fidelity (DINO-I). With our SuDe, the online Dreambooth can also achieve better attribute alignment than ELITE.

10.4. Visualizations for more examples

We provide more attribute-related generations in Fig. 13, where we see that based on the strong generality of the pre-trained diffusion model, our SuDe is applicable to images in various domains, such as objects, animals, cartoons, and human faces. Besides, SuDe also works for a wide range of attributes, like material, shape, action, state, and emotion.

Table 4. Results on stable-diffusion v1.4.

Method	CLIP-I	DINO-I	CLIP-T	DINO-T
ELITE [41]	68.9	41.5	28.5	43.2
Dreambooth [30]	77.4	59.7	29.0	42.1
Dreambooth w/ SuDe	**77.4**	**59.9**	**30.5**	**45.3**

10.5. Visualizations for more applications

In Fig. 14, We present more visualization about using our SuDe in more applications, including recontextualization, art renditions, costume changing, cartoon generation, action editing, and static editing.

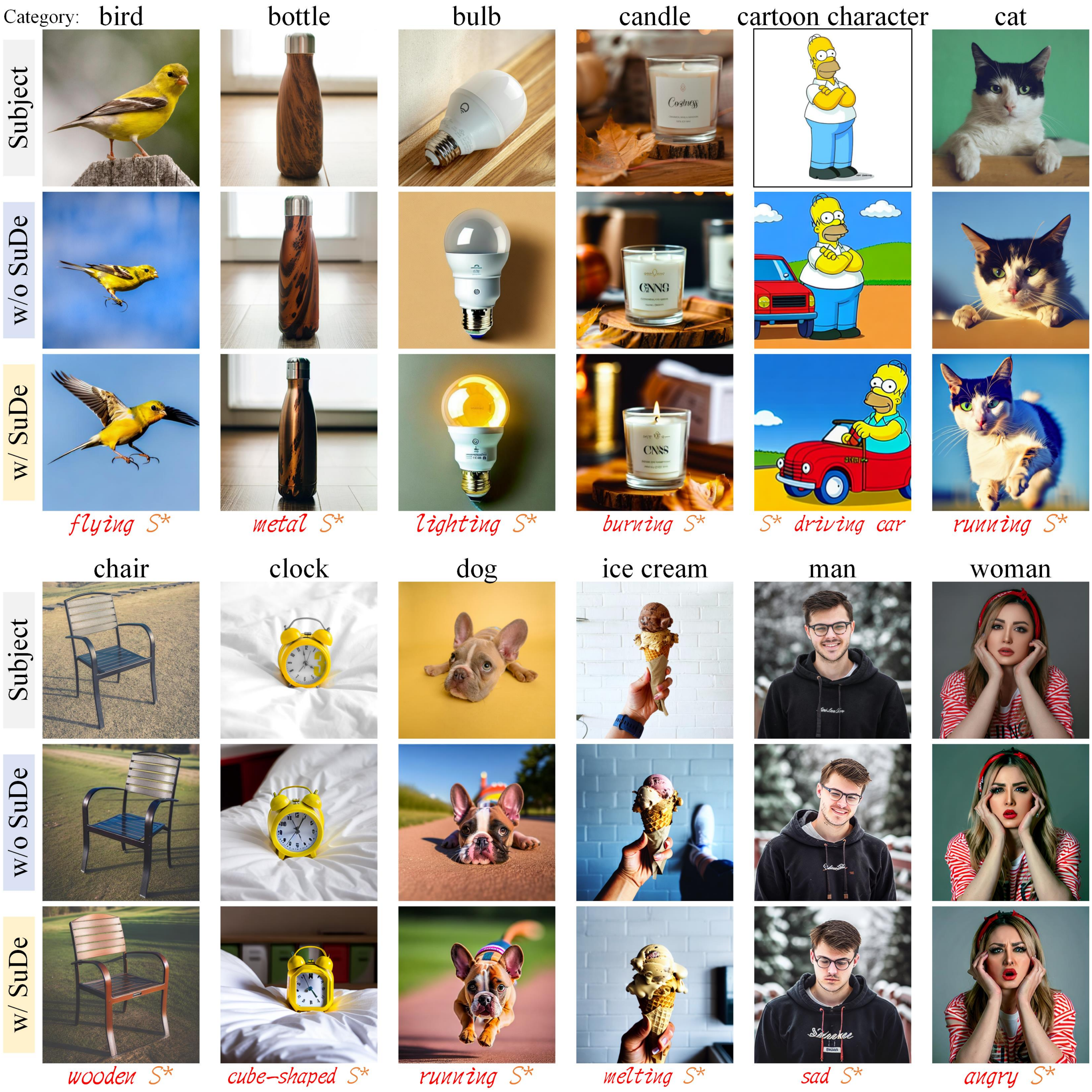


Figure 13. **More examples**. These results are obtained from DreamBooth w/o and w/ SuDe. The subject images are from Unsplash [1].

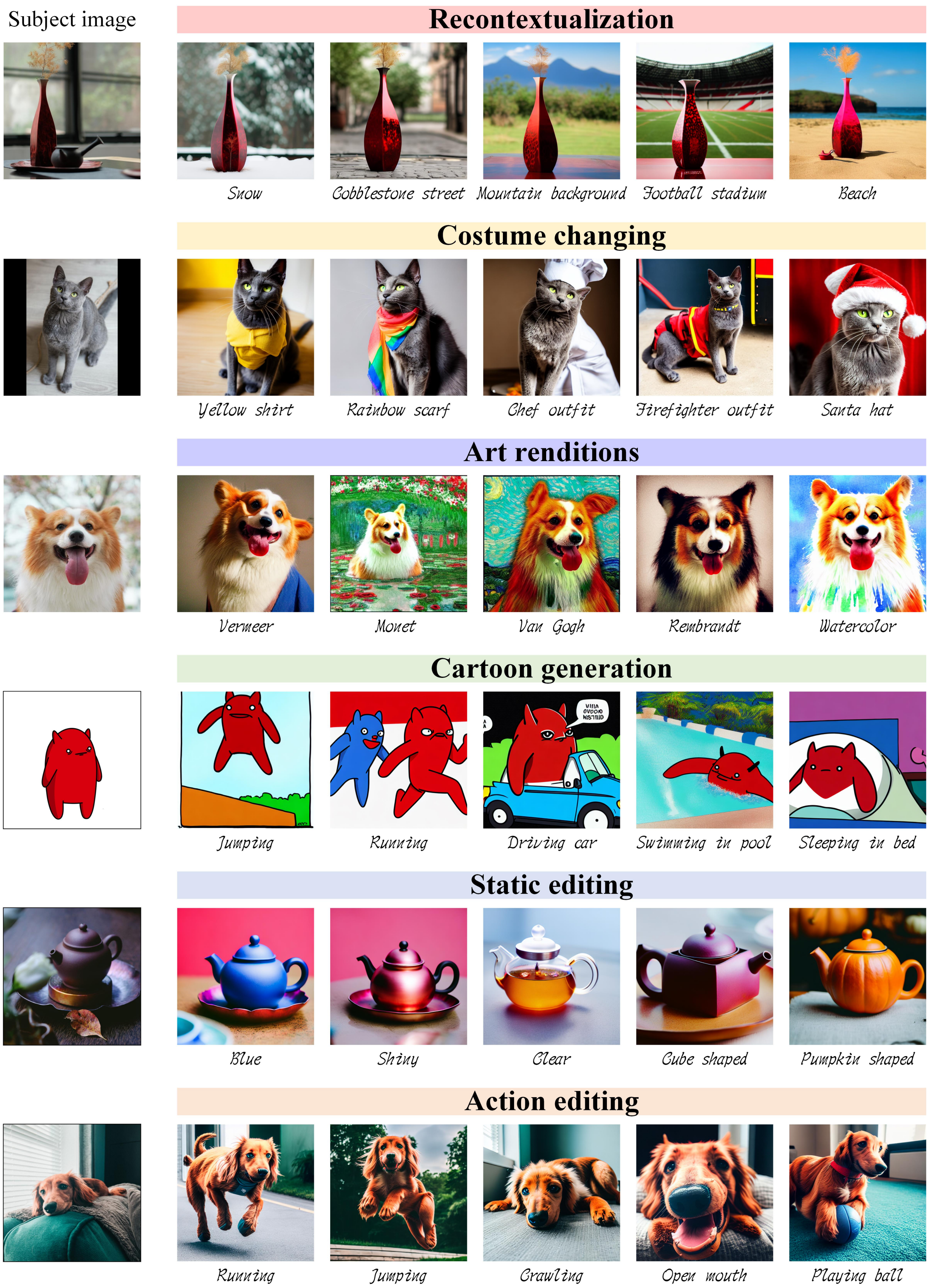


Figure 14. More applications using our SuDe with the Custom Diffusion [17] baseline.