## A. Formatted prompt generation

For non-live objects, we use the following prompts to generate words describing shape, color, textures, and background in ChatGPT:

- shape: give me 100 adjective words describing the shape of an object
- color: give me 100 adjective words describing the color of an object
- texture: give me 100 adjective words describing the texture of an object
- background: give me 500 phrases that describe the background, such as "on the table", as diverse as possible.

After removing duplicated ones, there are 85 shapes, 93 colors, 96 textures, and 455 backgrounds.

For live objects, we use the following prompts to generate words describing shape, color, textures, and motion in ChatGPT:

- body: give me 100 adjective words describing the body of an animal
- skin: give me 100 adjective words describing the skin or fur of an animal
- emotion: give me 100 adjective words describing the emotion of an animal
- motion: give me 1000 different short concise sentences that contains a special token "$concept" which stands for a specific animal, which can be a dog, a cat or a human. For example: "a $concept sitting in a temple", "a $concept walking in a supermarket". Keep "a $concept" in the sentences.

After removing duplicated ones, there are 89 bodies, 86 skins/furs, 75 emotions, and 744 motions. For humans, we replace the word "animal" above with "person".

We use

- style: give me 100 image style descriptions, such as "a photo of", and "a painting of".

After removing duplicated ones, there are 99 styles left.

## B. Better Category Naming

We show the name change in Table 3. As shown in Figure 6, there is a notable discrepancy between the utilization of vague class names, such as "toy", and more specific object names, such as "duck toy", on the ground truth images. Notably, the CLIP-T score appears to be significantly influenced by the nomenclature chosen for the object, and thereby potentially undermines its accuracy as an indicator of text-image alignment. To delve deeper into this matter, we calculate the CLIP-T score on the original images and the manually added prompts. Table 1 presents that, when using vague class names, the CLIP-T score for ground truth text-image pairs falls even below the conventional threshold of 0.3, which is typically considered as the threshold for assessing text-image pair compatibility [28]. To rectify this issue, we replace vague class names with highly specific object names, which results in a substantial improvement in the CLIP-T score for ground truth text-image pairs. Additional details regarding the modified nomenclature list can be found in Appendix B.

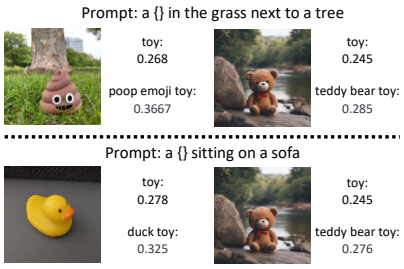


Figure 6. **CLIP-T score on different subject names**. We randomly generate an image with the prompt "a teddy bear toy sitting beside a river". Subsequently, we evaluate prompts "a {} in the grass next to a tree" and "a {} sitting on a sofa". When employing the vague prompt "toy", both the ground truth and the mismatched example are classified as mismatches (<0.3). In contrast, when using specific class names, the ground truth is categorized as a match (>0.3), while the mismatched examples remain classified as mismatches.

| subject name | original class | modified class |
|---|---|---|
| bear_plushie | stuffed animal | bear plushie |
| berry_bowl | bowl | berry bowl |
| can | can | drink can |
| clock | clock | alarm clock |
| duck_toy | toy | duck toy |
| grey_sloth_plushie | stuffed animal | sloth plushie |
| monster_toy | toy | monster toy |
| poop_emoji | toy | poop emoji toy |
| rc_car | toy | racing car toy |
| red_cartoon | cartoon | 2d cartoon devil |
| robot_toy | toy | robot toy |
| wolf_plushie | stuffed animal | wolf plushie |

Table 3. **Name Change**. We change the name for a more reasonable CLIP-T metric and better performance.

## C. Details of User Study

We randomly sampled and paired 300 comparisons of ours(SD) versus DreamBooth, half of which is for the subject alignment and the other half for the text alignment. For subject alignment, we randomly sampled a ground truth image and asked "The foreground object in which image is more similar to the reference?". For text alignment, we asked "Which image better depicts {}?", where {} is replaced by the prompt. We equally divided the questions into 10 groups. Each person randomly received one group. We did the same for ours(SDXL) versus ours(SD). We provide an example of our interface in Figure 7.

## D. Extension: Using BLIP to Generate Captions

We tried to use BLIP [16] to generate more personalized captions for the training example images. BLIP outputs a caption for the input image and can be conditioned on the format. We condition BLIP so that it generates prompts that start with "a [subject]". For instance for the "tortoise plushie" image BLIP generates

- a tortoise plushie on a pillow
- a tortoise plushie
- a tortoise plushie sitting on a piano keyboard
- a tortoise plushie on a desk
- ...

To unify the prompt format, we task ChatGPT to *'Change the following sentence to the format "A <new> tortoise plushie blablabla". The "<new>" is a special token that needs to be inserted before tortoise plushie.'* The result are the following prompts

- a <new> tortoise plushie on a pillow
- a <new> tortoise plushie
- a <new> tortoise plushie sitting on a piano keyboard
- a <new> tortoise plushie on a desk
- ...

The results of this "tortoise plushie" dataset is shown in Figure 15. With this addition of using BLIP, it alleviated writing the prompt examples manually, i.e., it replaced the manual steps in Section 3.

## E. Implementation Details

We opt for the identifier word "olis" instead of the more commonly used "sks". This choice is based on the fact that "olis" corresponds to the least frequently utilized token in the model's vocabulary [1]. Each training batch contains one example from training set and one example from regularization set. For SD, we fine-tune the entire model with a learning rate of 2e-6 and perform inference using 200 steps of DDIM [31]. For SDXL, which has a larger model size, we employ a LoRA with a rank of 32 for both the text encoders and UNet. We also train the text embeddings. We set learning rate to 1e-4. We use 50 steps of DDIM for inference. We show the best number of iterations in Table 4. For simplicity, we use 4000 and 8000 iterations for SD and SDXL, respectively.

| subject name | best #iterations on SD | best #iterations on SDXL |
|---|---|---|
| backpack | 6000-8000 | 8000-10000 |
| backpack_dog | 2000-3000 | 4000-6000 |
| bear_plushie | 2000-4000 | 4000-6000 |
| berry_bowl | 6000-8000 | 8000-10000 |
| can | 6000-8000 | 8000-10000 |
| candle | 4000-6000 | 8000-10000 |
| cat | 1000-3000 | 1000-3000 |
| cat2 | 6000-8000 | 8000-10000 |
| clock | 6000-8000 | 8000-10000 |
| colorful_sneaker | 4000-6000 | 6000-8000 |
| dog | 1000-3000 | 1000-3000 |
| dog2 | 2000-4000 | 4000-6000 |
| dog3 | 2000-4000 | 8000-10000 |
| dog5 | 3000-4000 | 6000-8000 |
| dog6 | 3000-4000 | 6000-8000 |
| dog7 | 3000-4000 | 6000-8000 |
| dog8 | 1000-3000 | 1000-3000 |
| duck_toy | 3000-4000 | 3000-4000 |
| fancy_boot | 3000-4000 | 6000-8000 |
| grey_sloth_plushie | 3000-4000 | 6000-8000 |
| monster_toy | 3000-4000 | 8000-10000 |
| pink_sunglasses | 3000-4000 | 4000-6000 |
| poop_emoji | 3000-4000 | 4000-6000 |
| rc_car | 3000-4000 | 4000-6000 |
| red_cartoon | 6000-8000 | 8000-10000 |
| robot_toy | 3000-4000 | 6000-8000 |
| shiny_sneaker | 3000-4000 | 6000-8000 |
| teapot | 6000-8000 | 8000-10000 |
| vase | 6000-8000 | 8000-10000 |
| wolf_plushie | 3000-4000 | 4000-6000 |

Table 4. **Best #iterations of datasets in DreamBench.** The variation mainly comes from the diversity of the dataset itself.

## F. More Results

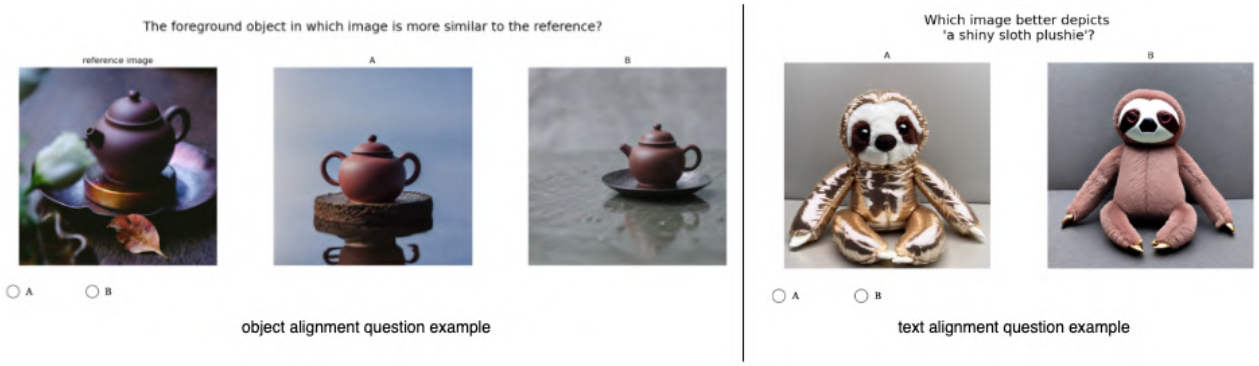


Figure 7. An example of the user study interface. The left is an example for question of object alignment, and the right is an example for question of text alignment. Each user was asked to answer 30 questions about subject alignment (left) and 30 questions about text alignment (right). The examples in the questions are randomly sampled from a large pool.

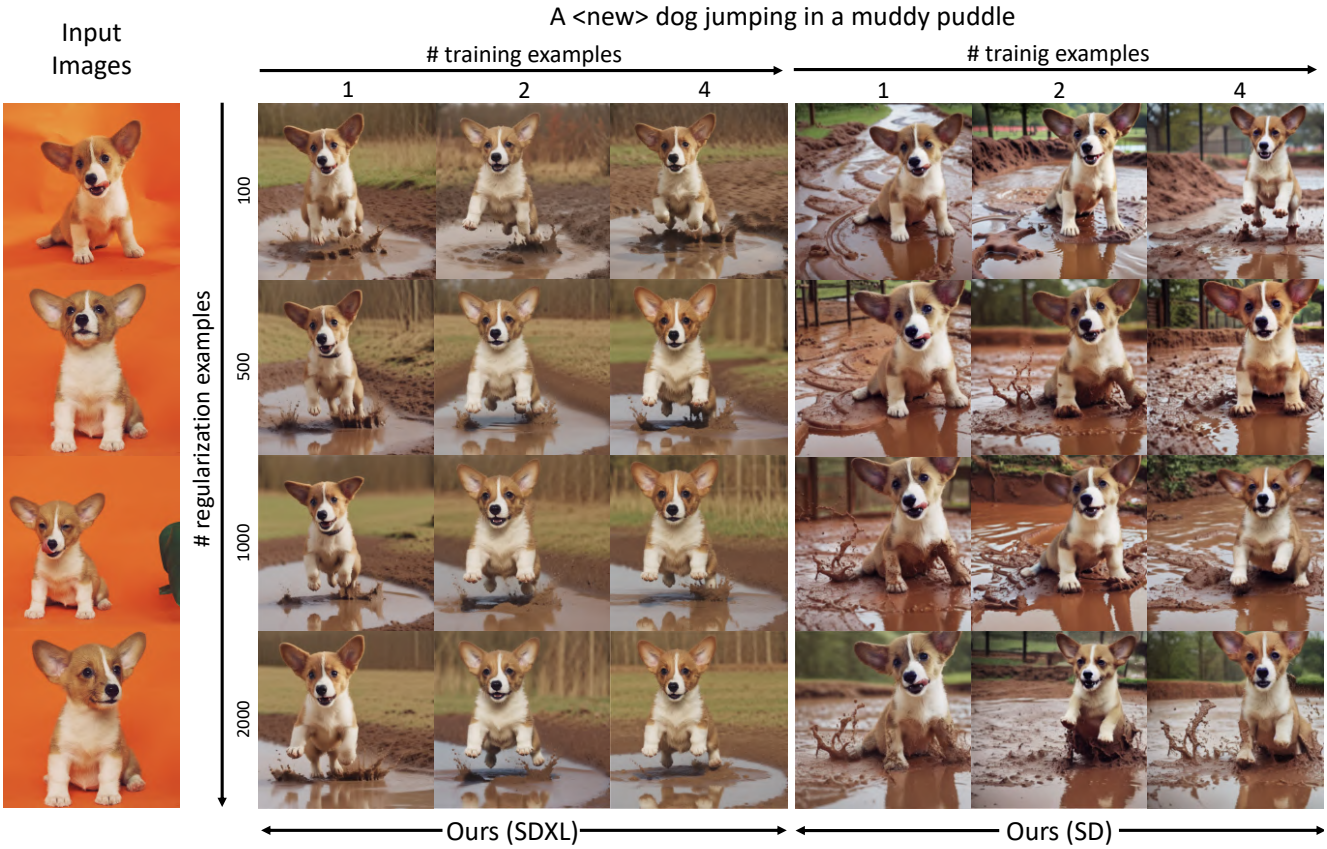


Figure 8. Ablation Tests on number of training examples and regularization dataset size. Even only a very small regularization dataset is given (1̃00 examples), our method still effectively prevent overfitting and preserves the identity.

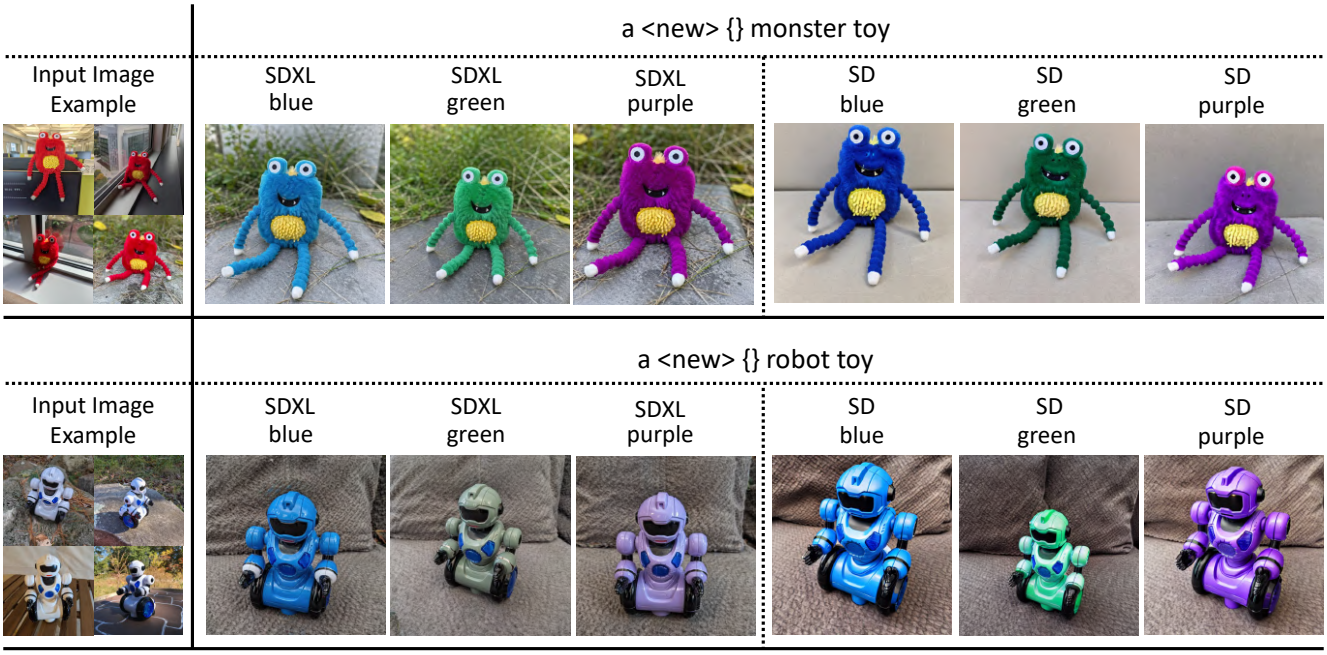


Figure 9. **Color Modification**. Our method can alter the color of the subject. It is important to mention that when modifying the color, using "a <new>[color] [class noun]" is more effective than "a [color] <new>[class noun]".


Figure 10. More Results on inanimate objects.

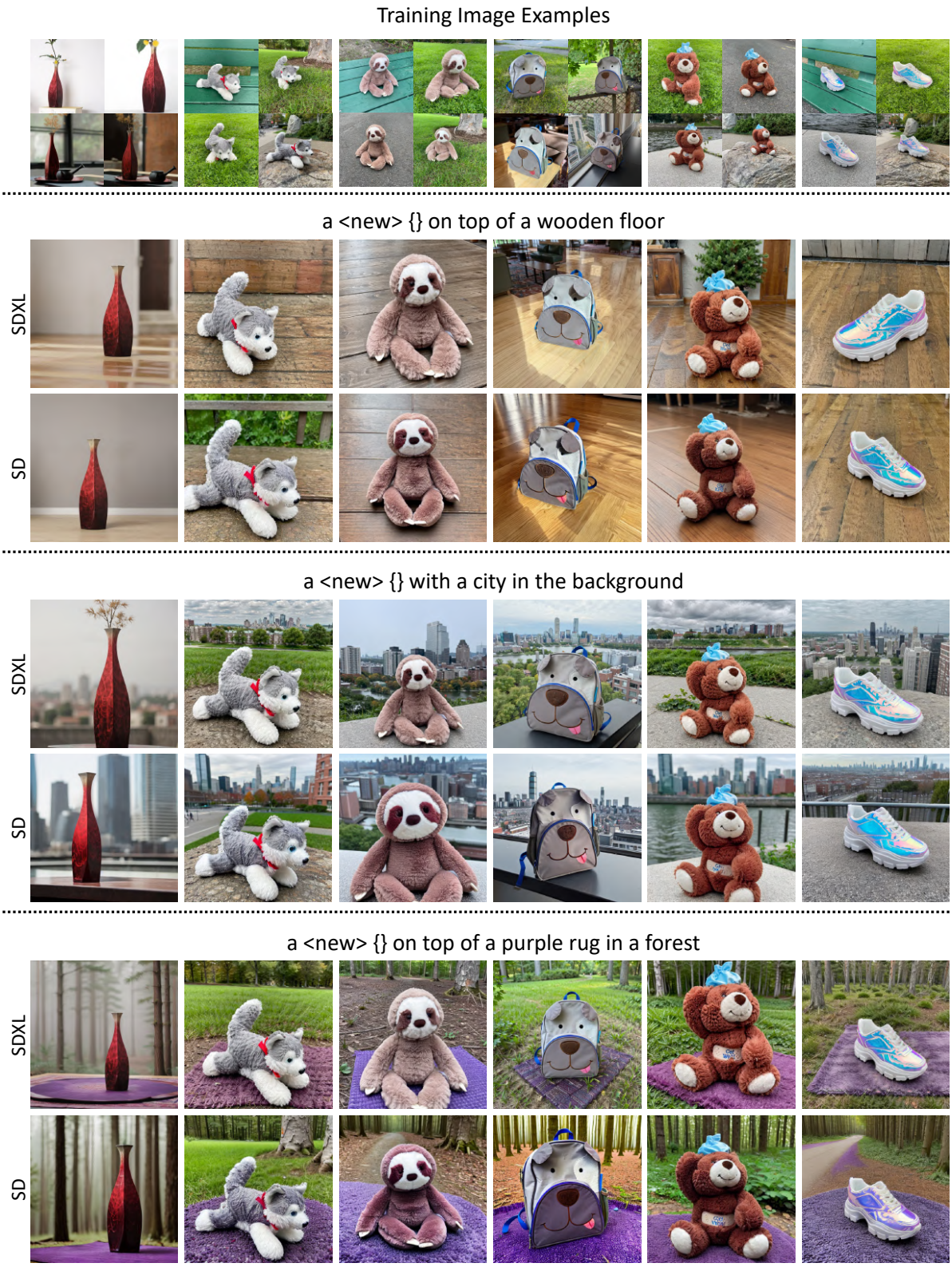


Figure 11. More Results on inanimate objects.

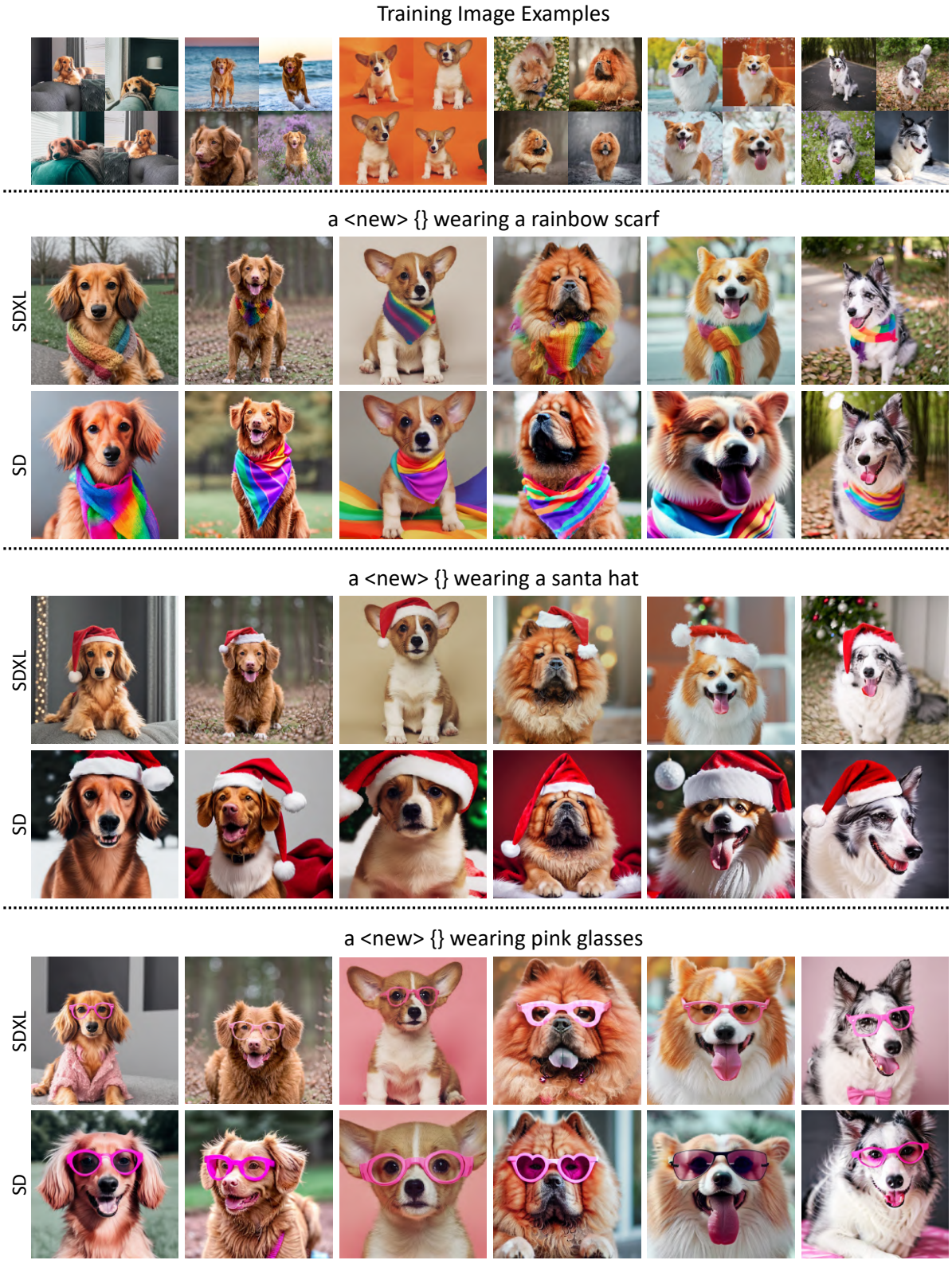


Figure 12. More Results on living entities.

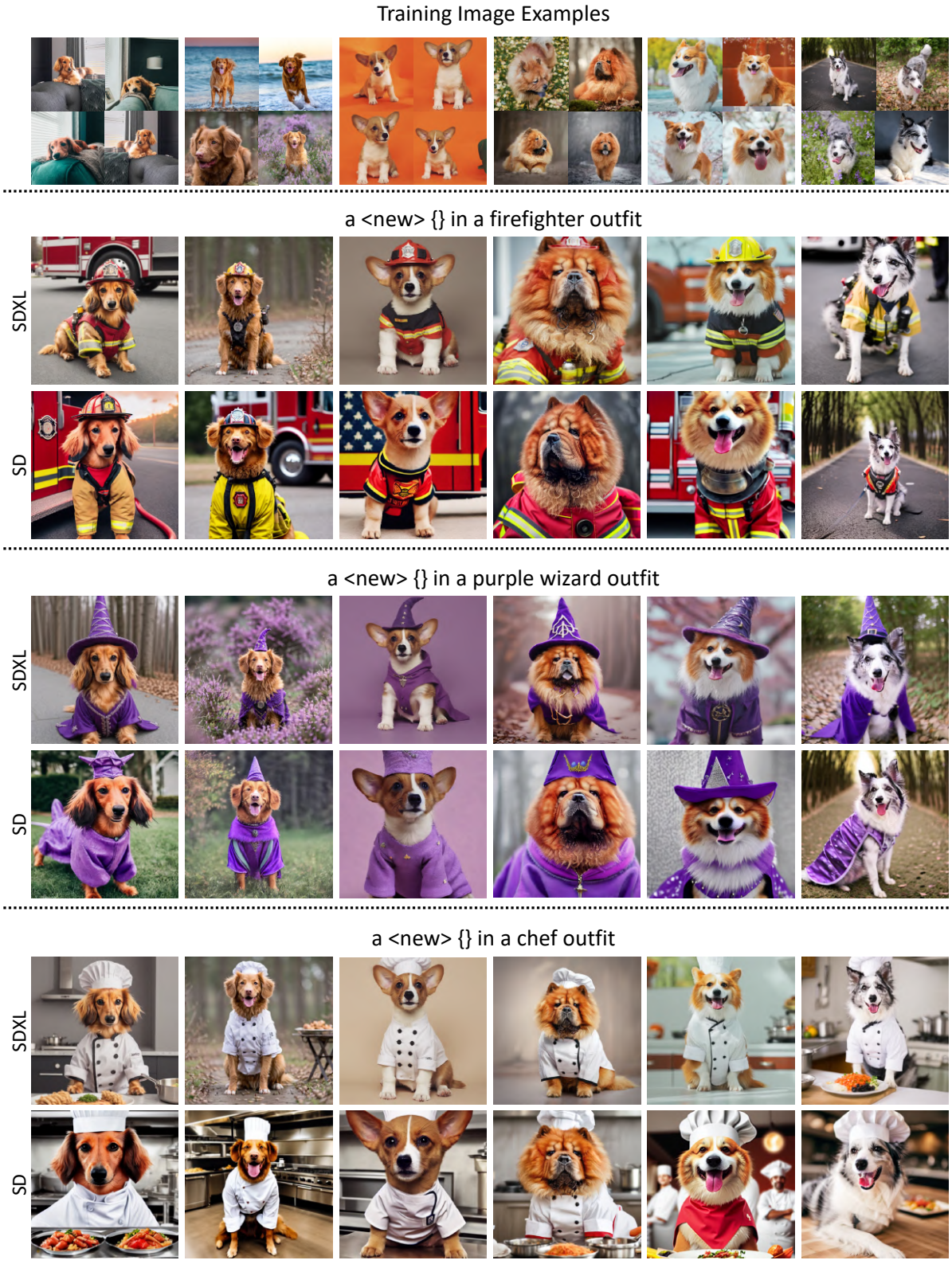


Figure 13. More Results on living entities.


Figure 14. More Comparison.


Figure 15. More Comparison.

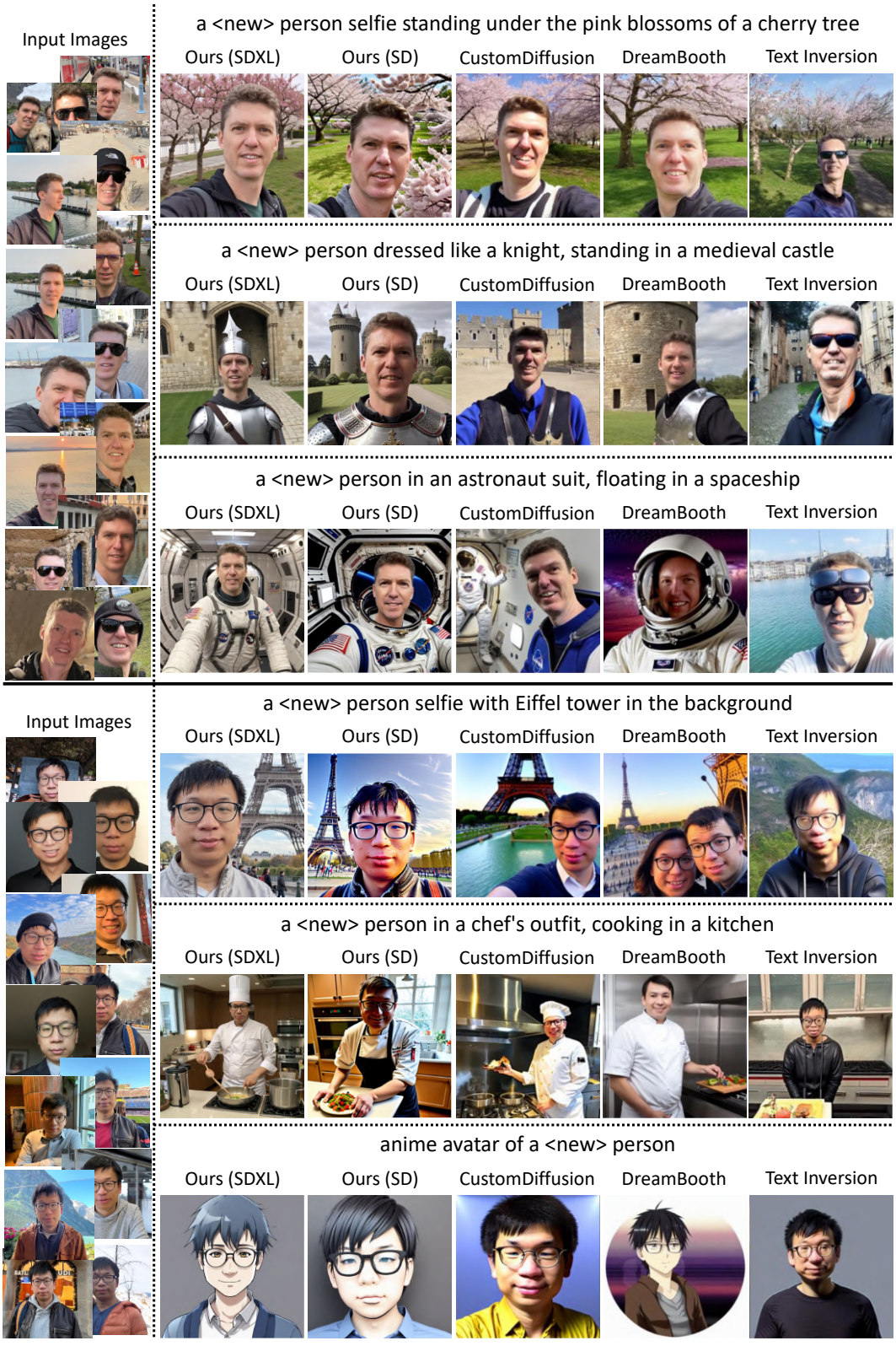


Figure 16. More Comparison.