A Supplementary Material

A.1 Extended Related Work

Text-to-Image Generation Previously, different GAN-based models [32, 33, 34, 35] have shown great progress in generating high-quality images. Recently, diffusion-based models models [36, 37, 1, 5, 38, 4] have gained unprecedented popularity to surpass the GAN-based models. These models have shown great progress in generating highly realistic images faithful to the given text control. The progress is mainly driven by diffusion model [14, 15] and auto-regressive backbone [3]. However, these models can only accept text prompt as the input, lacking control from other sources. For example, if we want to generate an image about our own dog or our own backpack in different scenes, it becomes challenging for the existing models [6]. Also, as suggested by [9], the existing generation models are highly biased towards generating frequent subjects while having difficulty generating less common visual entities. These challenges have spawned the new task of 'Subject-Drive Text-to-Image Generation', which is the core task of our paper aims to solve.

A.2 Dataset Construction

To validate the effectiveness, we provide an ablation study to show that higher precision is more important than recall in training the apprentice model. Particularly, when the threshold is set to a lower number (*e.g.*, 0.01 or 0.015), `SuTI` becomes less stable.

As our goal is to collect images of the same subject, we create an initial subject cluster by grouping all (image, alt-text) pairs that come from the same URL ($\sim$45M clustrers), and filter the cluster with less than 3 instances ($\sim$77.8% of the clusters). As a result, it leaves us with $\sim$10M image clusters. We then apply the pre-trained CLIP ViT-L14 model [39] to filter out 81.1% of clusters that has the average intra-cluster visual similarity between 0.82 and 0.98 to ensure the quality of clusters.

Though the mined clusters already contain (image, alt-text) information, the alt-text's noise level is too high. Therefore, we apply the state-of-the-art image captioning model [10] to generate descriptive text captions for every image of all image clusters, which forms the data triples of (image, alt-text, caption). However, current image captioning models tend to generate generic descriptions of the visual scene, which often occlude the detailed entity information about the subject. For example, generic captions like `'a pair of red shoes'` would greatly decrease the expert model's capability to preserve the subject's visual appearance. To increase the specificity of the visual captions, we propose to merge the alt-text, which normally contains specific meta information like brands, names, etc with the model-generated caption. For example, Given an alt-text of `'duggee talking puppet hey duggee chicco 12m'` and a caption of `'a toy on the table'`, we aim to combine them as a more concrete caption: `'Hey duggee toy on the table'`. To achieve this, we prompt the pre-trained large language models [18] to read all (alt-text, caption) pairs inside each image cluster, and output a short descriptive text about the visual subject. These refined captions with the mined images are used as the image-text cluster $\mathbb{C}_s$ w.r.t subject s, which will be used to fine-tune the expert models.

A.3 `SuTI` Skillset

We demonstrate `SuTI`'s skillset in Figure 7.

A.4 Failure Examples

Figure 8 show some failure examples of `SuTI`. We show several types of failure modes: (1) the model 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. For example, the `'bed'` from the demonstration is being brought to the generation, (3) the subject's visual appearance is being modified through, mostly influenced by the context, like the `'candle'` contains non-existing artifacts when contextualized in the `'toilet'`. These three failure modes constitute most of the generation errors. (4) The models are not particularly good at handling compositional prompts like the `'bear plushie'` and `'sunglasses'` example. In the future, we plan to work on how to improve these aspects.

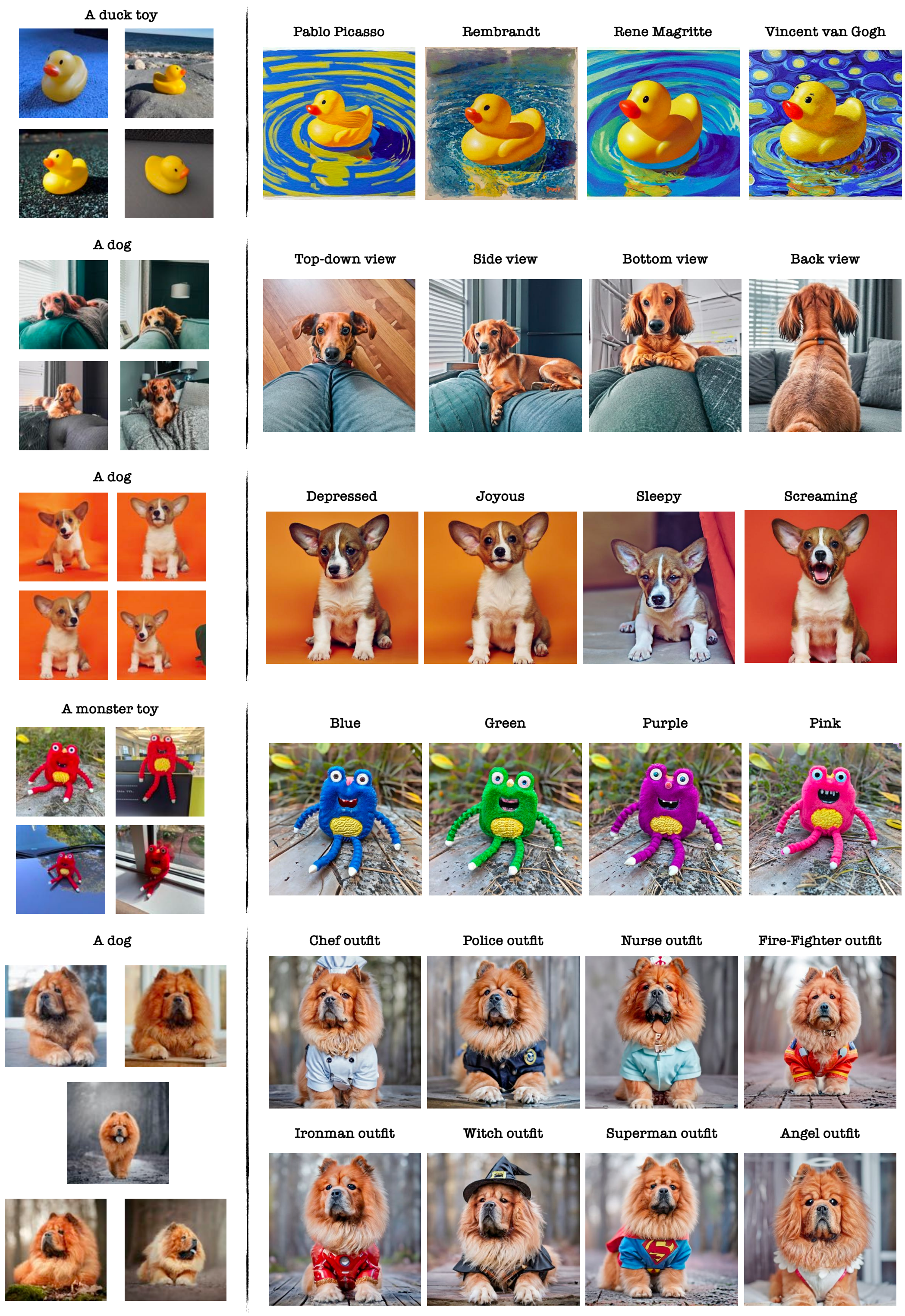

Figure 7: `SuTI`'s in-context generation that demonstrates its skill set. Results generated from *a single model*. First row: art rendition of the subject. Second row: multi-view synthesis of the subject. Third row: modifying expression for the subject. Fourth row: editing the color of the subject. Fifth row: adding accessories to the subject. Subject (image, text) and editing key words are annotated, with detailed template in the Appendix.

Figure 8: `SuTI`'s failure examples on DreamBench-v2.

A.5 Ethical Statement

Subject-driven text-to-image generation has wide downstream applications, like adapting certain given subjects into different contexts. Previously, the process was mostly done manually by experts who are specialized in photo creation software. Such manual modification process is time-consuming. We hope that our model could shed light on how to automate such a process and save huge amount of labors and training. The current model is still highly immature, which can fall into several failure modes as demonstrated in the paper. For example, the model is still prone to certain priors presented in certain subject classes. Some low-level visual details in subjects are not perfectly preserved. However, it could still be used as an intermediate form to help accelerate the creation process. On the flip side, there are risks with such models including misinformation, abuse and bias. See the discussion of broader impacts in [1, 4] for more discussion.

A.6 More Examples

We demonstrate more examples from DreamBench-v2 in the following:

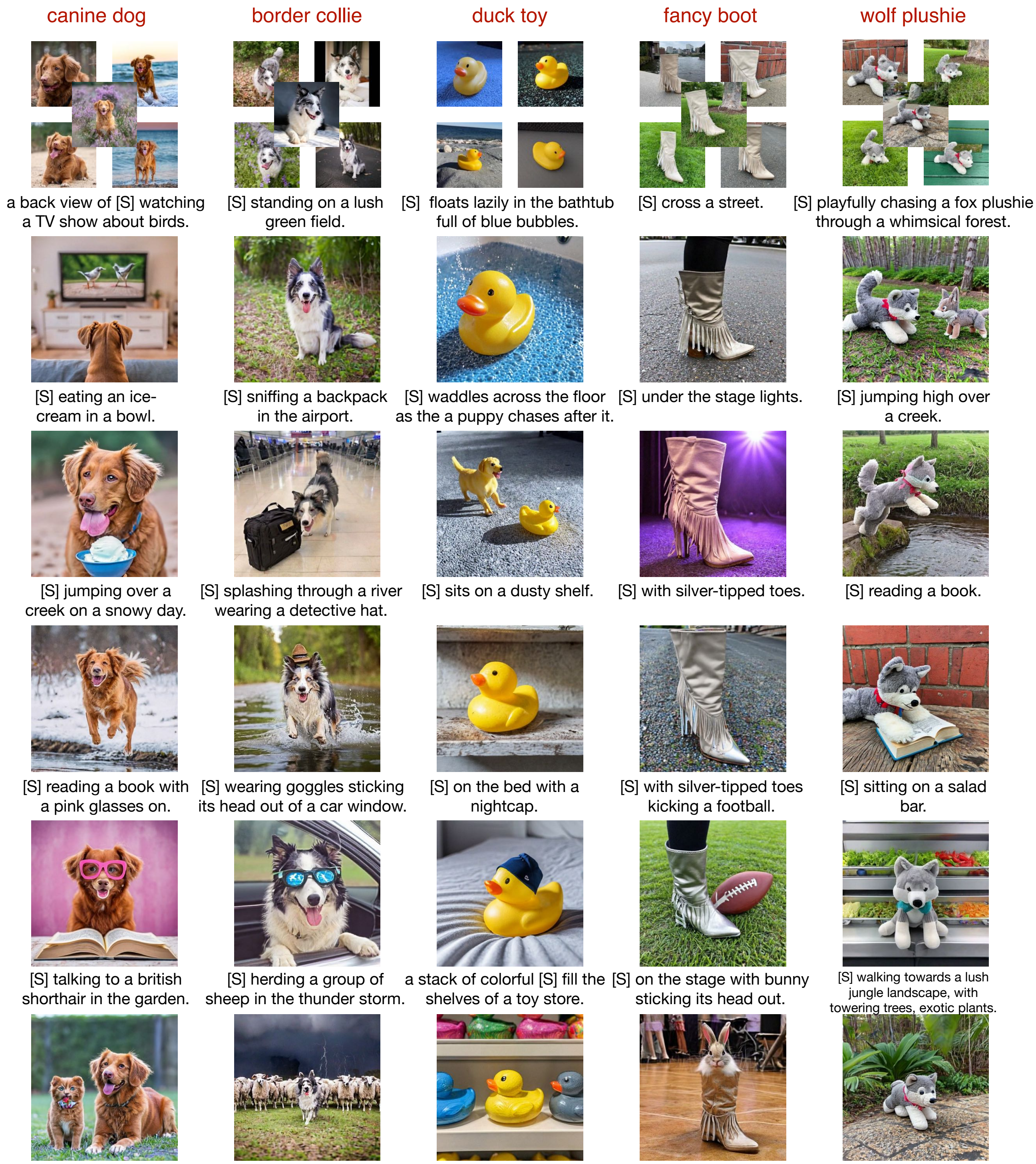

Figure 9: Visualization of SuTI's generation on the DreamBench-v2 (Part 1).

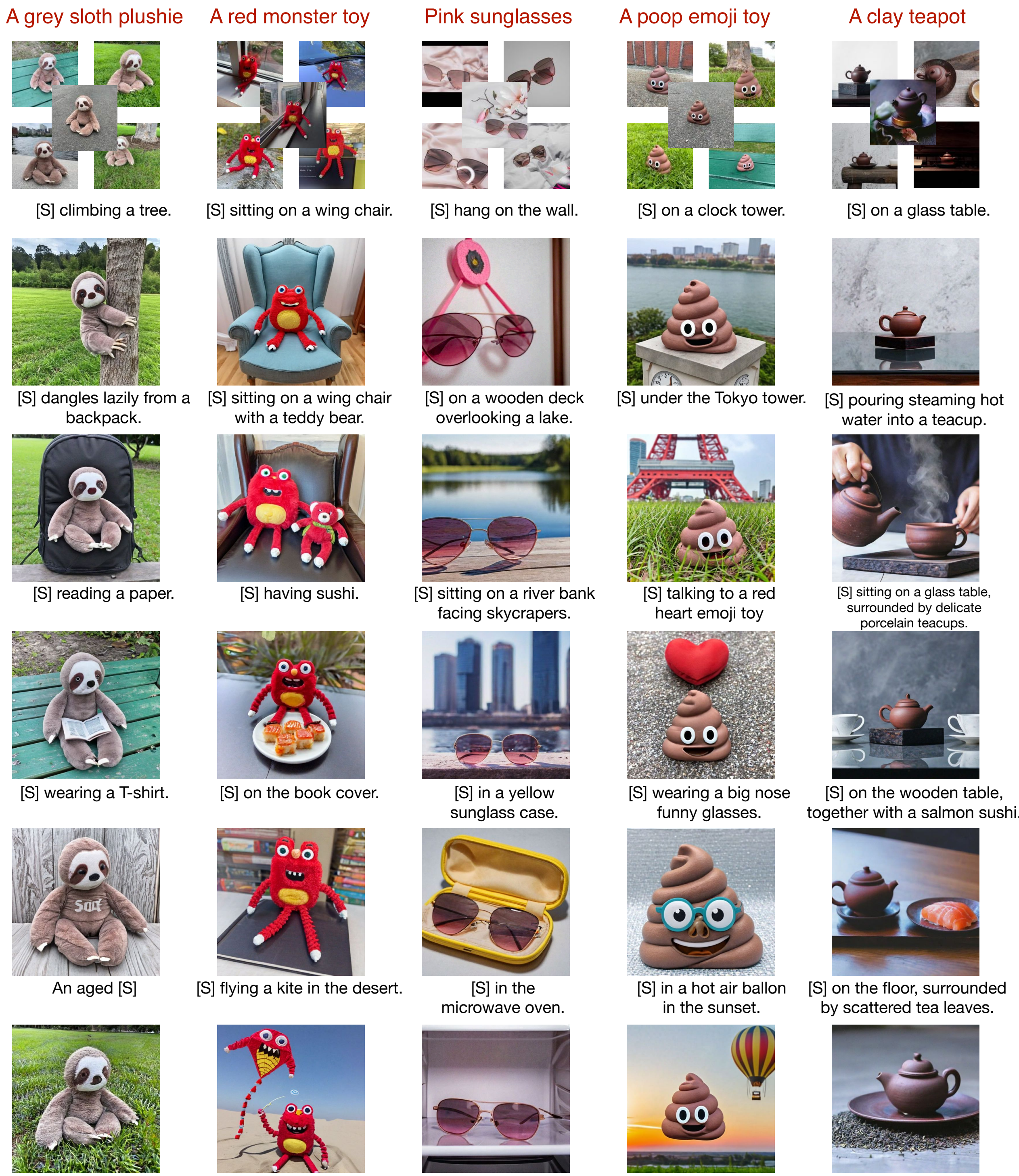

Figure 10: Visualization of `SuTI`'s generation on the DreamBench-v2 (Part 2).

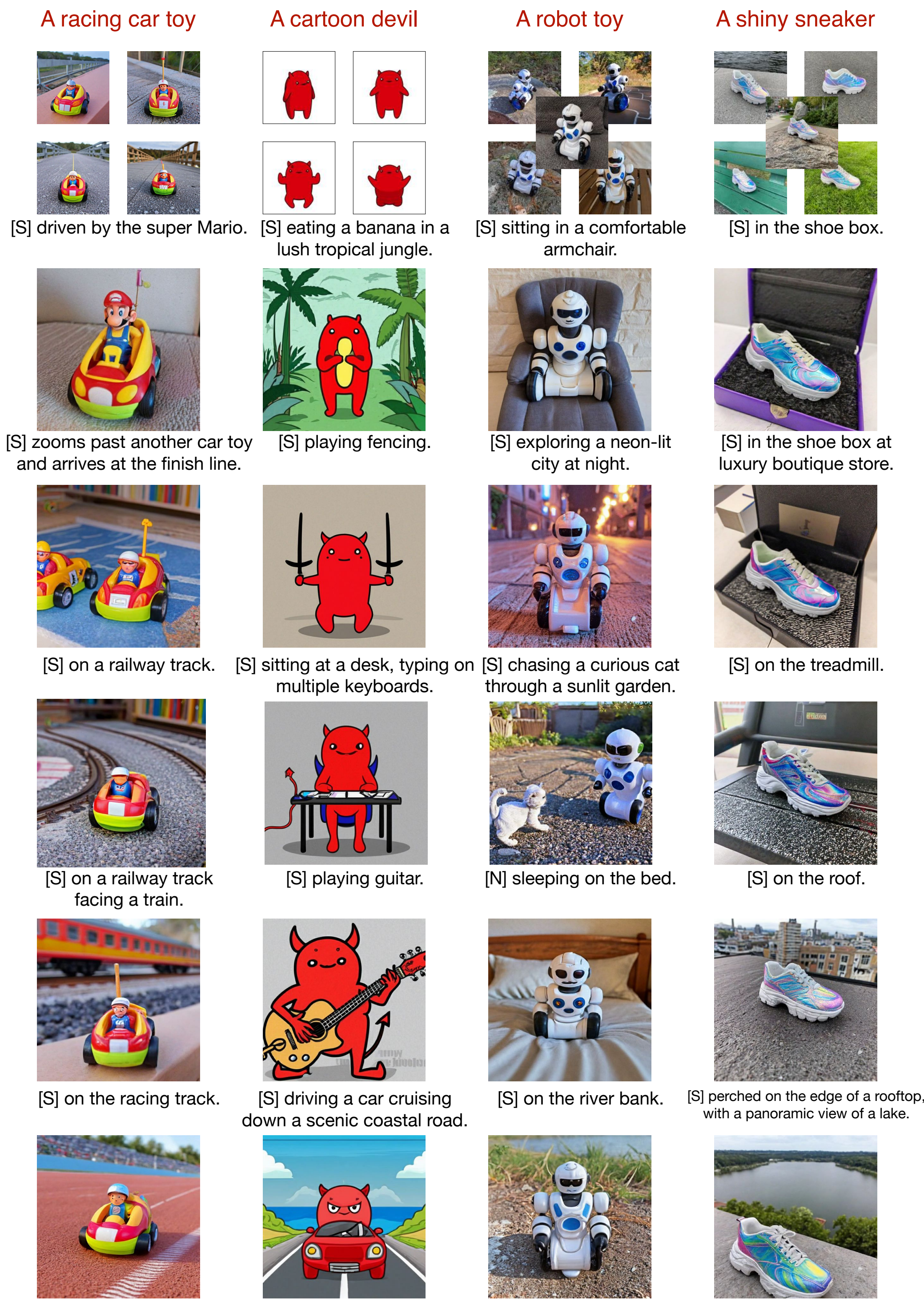

Figure 11: Visualization of `SuTI`'s generation on the DreamBench-v2 (Part 3).

Figure 12: In-context generation by `SuTI` model, with an increasing # of demonstration (More examples).