

418 A Supplementary Material

419 A.1 Extended Related Work

420 **Text-to-Image Generation** Previously, different GAN-based models [32, 33, 34, 35] have shown
421 great progress in generating high-quality images. Recently, diffusion-based models [36, 37,
422 1, 5, 38, 4] have gained unprecedented popularity to surpass the GAN-based models. These models
423 have shown great progress in generating highly realistic images faithful to the given text control. The
424 progress is mainly driven by diffusion model [14, 15] and auto-regressive backbone [3]. However,
425 these models can only accept text prompt as the input, lacking control from other sources. For
426 example, if we want to generate an image about our own dog or our own backpack in different scenes,
427 it becomes challenging for the existing models [6]. Also, as suggested by [9], the existing generation
428 models are highly biased towards generating frequent subjects while having difficulty generating less
429 common visual entities. These challenges have spawned the new task of ‘Subject-Drive Text-to-Image
430 Generation’, which is the core task of our paper aims to solve.

431 A.2 Dataset Construction

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

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

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

454 A.3 SuTI Skillset

455 We demonstrate SuTI’s skillset in [Figure 7](#).

456 A.4 Failure Examples

457 [Figure 8](#) show some failure examples of SuTI. We show several types of failure modes: (1) the
458 model has a strong prior about the subject and hallucinates the visual details based on its prior
459 knowledge. For example, the generation model believes ‘teapot’ should contain a ‘lift handle’.
460 (2) some artifacts from the demonstration images are being transferred to the generated images. For
461 example, the ‘bed’ from the demonstration is being brought to the generation, (3) the subject’s visual
462 appearance is being modified through, mostly influenced by the context, like the ‘candle’ contains
463 non-existing artifacts when contextualized in the ‘toilet’. These three failure modes constitute
464 most of the generation errors. (4) The models are not particularly good at handling compositional
465 prompts like the ‘bear plushie’ and ‘sunglasses’ example. In the future, we plan to work on how
466 to improve these aspects.

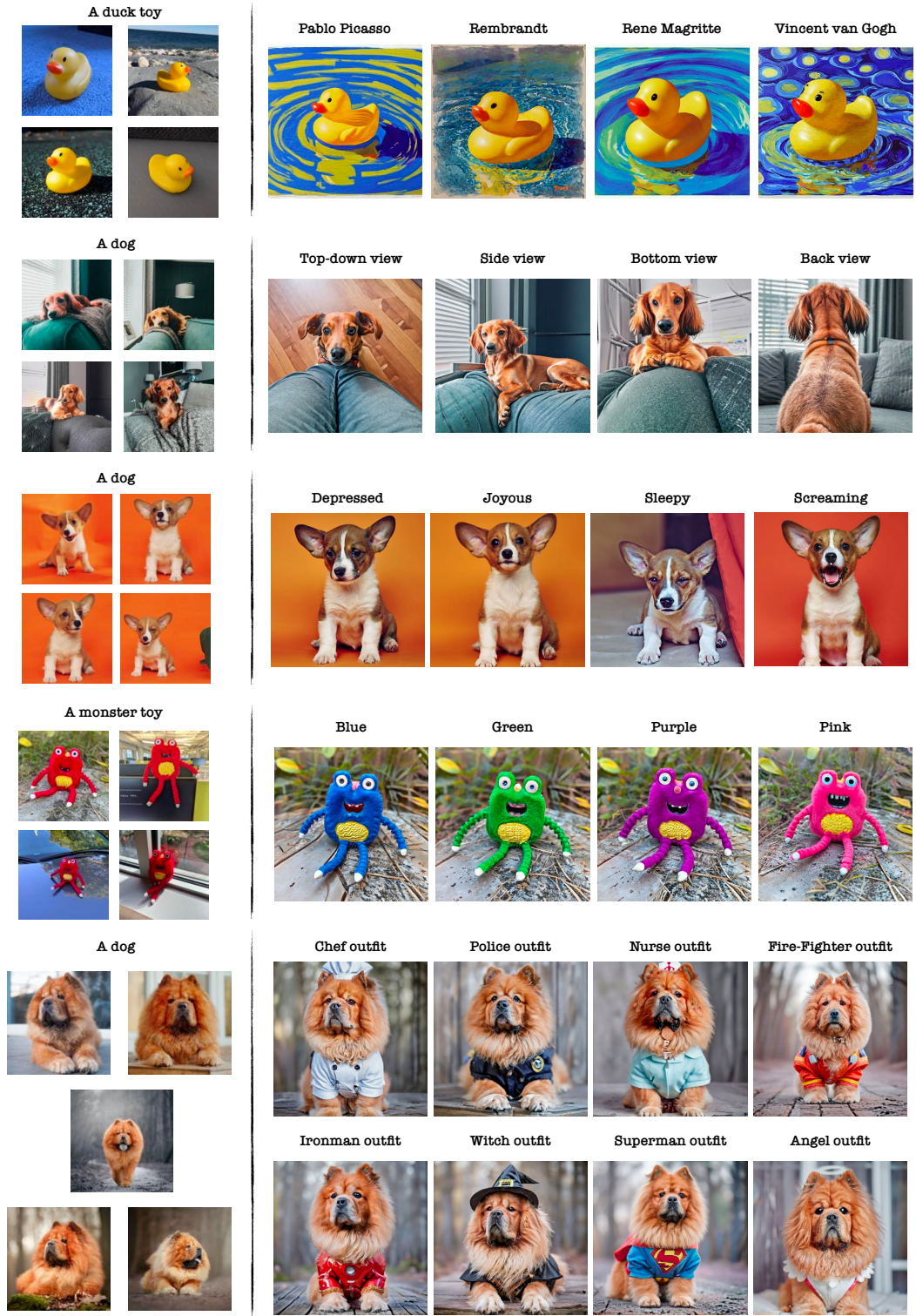


Figure 7: SuTI's in-context generation that demonstrates its skill set. Results generated from *a single model*. First row: art rendition of the subject. Second row: multi-view synthesis of the subject. Third row: modifying expression for the subject. Fourth row: editing the color of the subject. Fifth 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.

467 **A.5 Ethical Statement**

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

478 **A.6 More Examples**

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

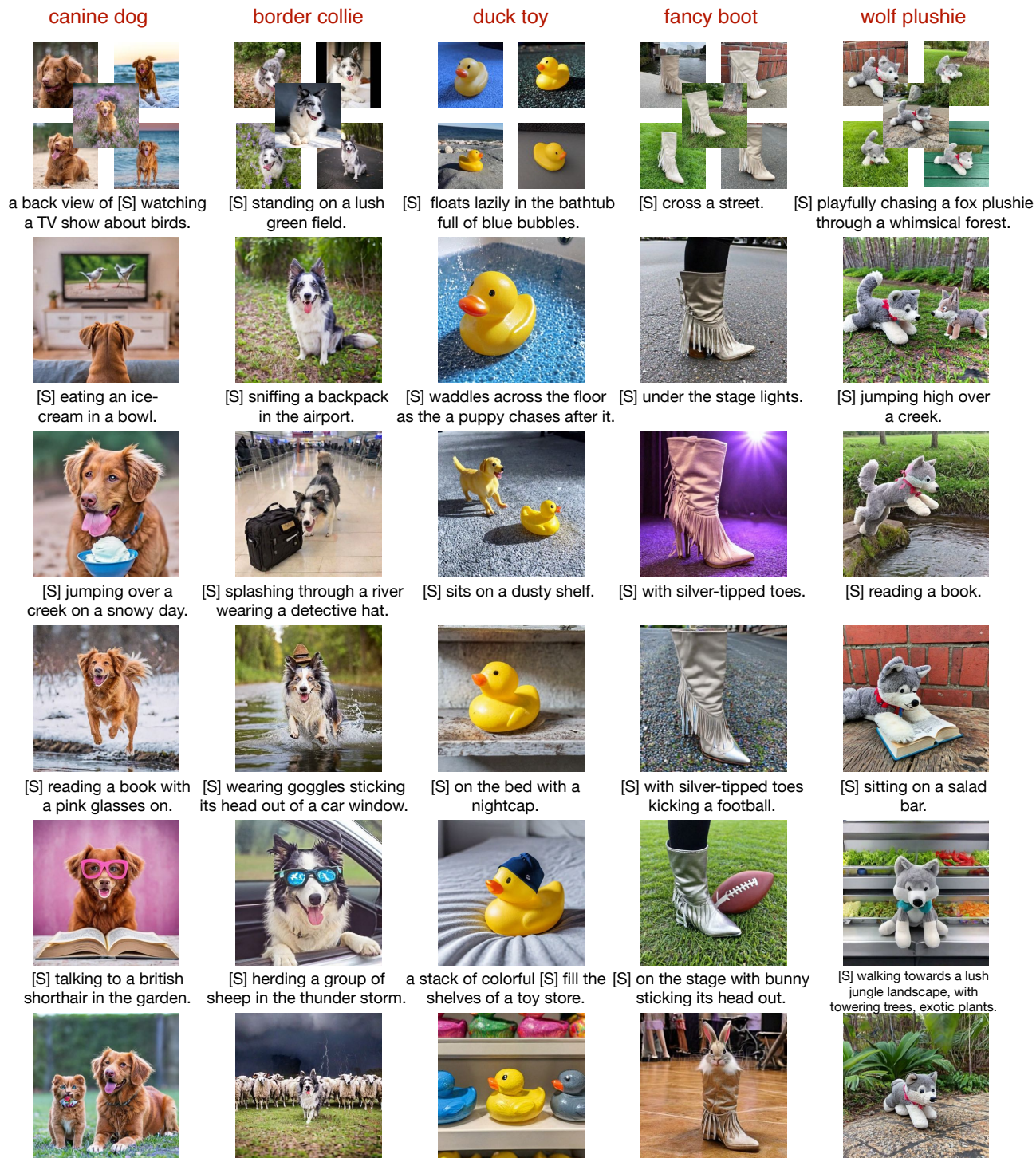


Figure 9: Visualization of SuTI's generation on the DreamBench-v2 (Part 1).

A grey sloth plushie



[S] climbing a tree.



[S] dangles lazily from a backpack.



[S] reading a paper.



[S] wearing a T-shirt.



An aged [S]



A red monster toy



[S] sitting on a wing chair.



[S] sitting on a wing chair with a teddy bear.



[S] having sushi.



[S] on the book cover.



[S] flying a kite in the desert.



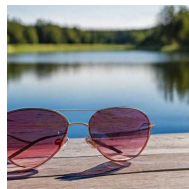
Pink sunglasses



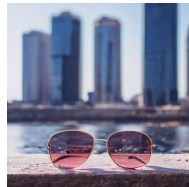
[S] hang on the wall.



[S] on a wooden deck overlooking a lake.



[S] sitting on a river bank facing skyscrapers.



[S] in a yellow sunglasses case.



[S] in the microwave oven.



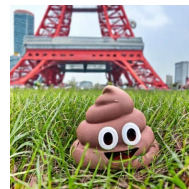
A poop emoji toy



[S] on a clock tower.



[S] under the Tokyo tower.



[S] talking to a red heart emoji toy



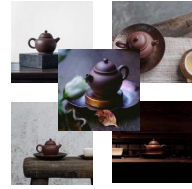
[S] wearing a big nose funny glasses.



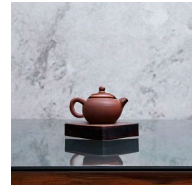
[S] in a hot air balloon in the sunset.



A clay teapot



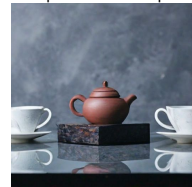
[S] on a glass table.



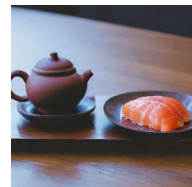
[S] pouring steaming hot water into a teacup.



[S] sitting on a glass table, surrounded by delicate porcelain teacups.



[S] on the wooden table, together with a salmon sushi.



[S] on the floor, surrounded by scattered tea leaves.

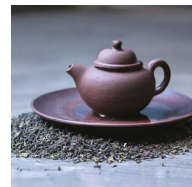


Figure 10: Visualization of SuTI's generation on the DreamBench-v2 (Part 2).

A racing car toy



[S] driven by the super Mario.



[S] zooms past another car toy and arrives at the finish line.



[S] on a railway track.



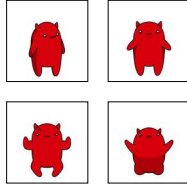
[S] on a railway track facing a train.



[S] on the racing track.



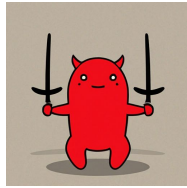
A cartoon devil



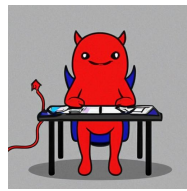
[S] eating a banana in a lush tropical jungle.



[S] playing fencing.



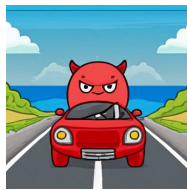
[S] sitting at a desk, typing on multiple keyboards.



[S] playing guitar.



[S] driving a car cruising down a scenic coastal road.



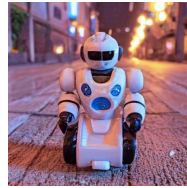
A robot toy



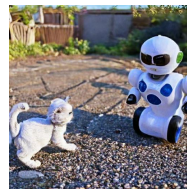
[S] sitting in a comfortable armchair.



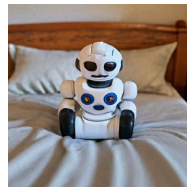
[S] exploring a neon-lit city at night.



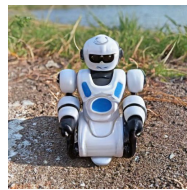
[S] chasing a curious cat through a sunlit garden.



[N] sleeping on the bed.



[S] on the river bank.



A shiny sneaker



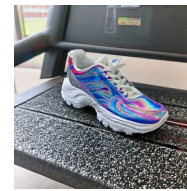
[S] in the shoe box.



[S] in the shoe box at luxury boutique store.



[S] on the treadmill.



[S] on the roof.



[S] perched on the edge of a rooftop, with a panoramic view of a lake.

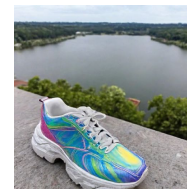


Figure 11: Visualization of SuTI's generation on the DreamBench-v2 (Part 3).

$C_s =$	\emptyset	A Herschel backpack 	A Herschel backpack 	A Herschel backpack 
$p_s =$ A Herschel backpack in Grand Canyon				
$p_s =$ A Herschel backpack in the water				
$C_s =$	\emptyset	A candle 	A candle 	A candle 
$p_s =$ A candle sitting on a Mirror				
$p_s =$ A candle decorated with flowers.				
$C_s =$	\emptyset	A bear plushie 	A bear plushie 	A bear plushie 
$p_s =$ Two bear plushies in the store.				
$p_s =$ A bear plushie in a temple.				

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