Subject-driven Text-to-Image Generation via Apprenticeship Learning

Wenhu Chen* **Hexiang Hu*** **Yandong Li** **Nataniel Ruiz**
Xuhui Jia **Ming-Wei Chang** **William W. Cohen**
Google Research
{wenhuchen,hexiang,mingweichang,wcohen}@google.com

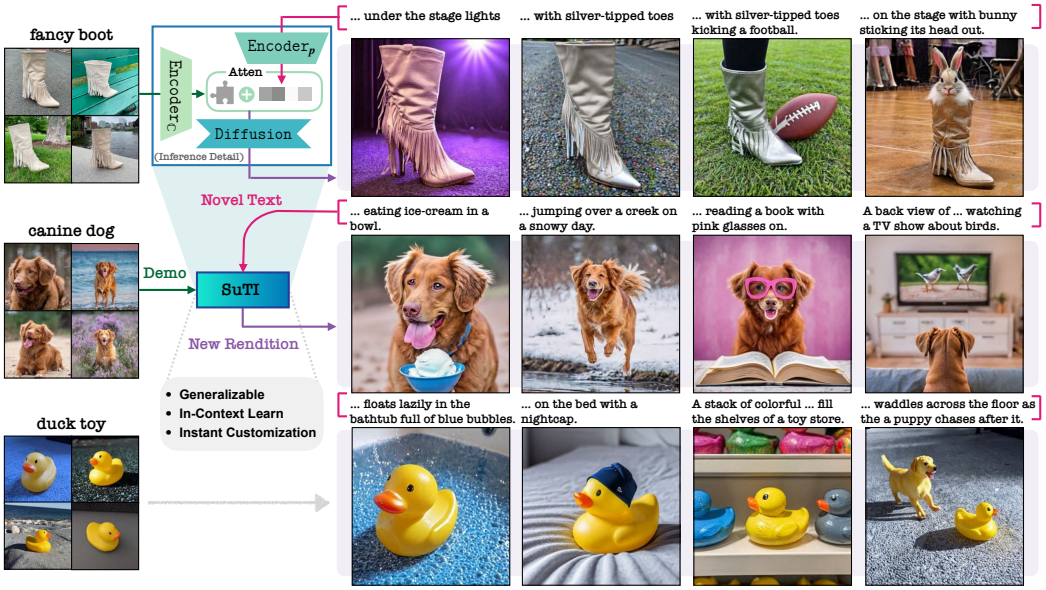

Figure 1: We train *a single* SuTI *model* to generate novel scenes faithfully reflecting given subjects (unseen in training, specified only by 3-5 in-context text→image demonstrations), without any optimization.

Abstract

Recent text-to-image generation models like DreamBooth have made remarkable progress in generating highly customized images of a target subject, by fine-tuning an "expert model" for a given subject from a few examples. However, this process is expensive, since a new expert model must be learned for each subject. In this paper, we present SuTI, a Subject-driven Text-to-Image generator that replaces subject-specific fine tuning with *in-context* learning. Given a few demonstrations of a new subject, SuTI can instantly generate novel renditions of the subject in different scenes, without any subject-specific optimization. SuTI is powered by *apprenticeship learning*, where a single apprentice model is learned from data generated by massive amount of subject-specific expert models. Specifically, we mine millions of image clusters from the Internet, each centered around a specific visual subject. We adopt these clusters to train massive amount of expert models specialized on different subjects. The apprentice model SuTI then learns to mimic the behavior of these experts through the proposed apprenticeship learning algorithm. SuTI can generate high-quality and customized subject-specific images 20x faster than optimization-based SoTA methods. On the challenging DreamBench and DreamBench-v2, our human evaluation shows that SuTI can significantly

*Core Contribution

Preprint. Under review.

outperform existing approaches like InstructPix2Pix, Textual Inversion, Imagic, Prompt2Prompt, Re-Imagen while performing on par with DreamBooth.

1 Introduction

Recent text-to-image generation models [31] have shown great progress in generating highly realistic, accurate, and diverse images from given text prompt. These models are pre-trained on web-crawled image-text pairs like LAION [32] with autoregressive backend models [27, 37] or diffusion backend models [26, 31]. Though achieving unprecedented success in generating highly accurate images, these models are not able to customize to a given subject, like a specific dog, shoe, backpack, etc. Therefore, *subject-driven text-to-image generation*, the task of generating highly customized images with respect to a target subject, has attracted significant attention from the community. Subject-driven image generation is related to text-driven image editing but often needs to perform more sophisticated transformations to source images (e.g., rotating the view, zooming in/out, changing the pose of subject, etc.) so existing image editing methods are generally not suitable for this new task.

Current subject-driven text-to-image generation approaches are slow and expensive. While different approaches like DreamBooth [30], Imagic [18], and Textual Inversion [10] have been proposed, they all require fine-tuning specific models for a given subject on one or a few demonstrated examples, which typically takes at least 10-20 minutes[2] to specialize the text-to-image model checkpoint for the given subjects. These approaches are time-consuming as they require back-propagating gradients over the entire model for hundreds or even thousands of steps per customization. Moreover, they are space-consuming as they require storing a subject-specific checkpoint per subject. To avoid the excessive cost, Re-Imagen [7] proposed a retrieval-augmented text-to-image framework to train a subject-driven generation model in a weakly-supervised fashion. Since the retrieved neighbor images are not guaranteed to contain the same subjects, the model does not perform as good as DreamBooth [30] for the task of subject-driven image generation.

To avoid excessive computation and memory costs, we propose to train a single subject-driven text-to-image generation model that can perform on-the-fly subject customization. Our method is dubbed Subject-driven Text-to-Image generator (`SuTI`), which is trained with a novel *apprenticeship learning* algorithm. Unlike standard apprenticeship learning which only focuses on learning from one expert, our apprentice model imitates the behaviors of a massive number of specialized expert models. After such training, `SuTI` can instantly adapt to unseen subjects and unseen or even compositional descriptions with only 3-5 in-context demonstrations within 30 seconds (on a Cloud TPU v4).

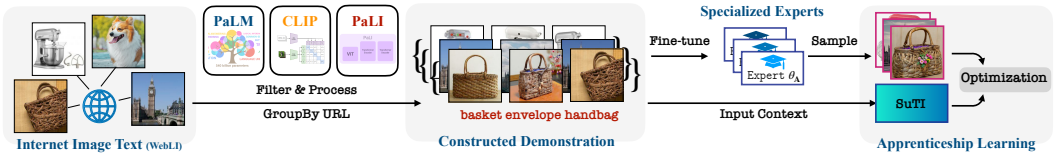

Figure 2: Conceptual Diagram of the Learning Pipeline

Figure 2 presents a conceptual diagram of the learning and data preparation pipeline. We first group the images in WebLI [8] by their source URL form tiny image clusters. As images from the same URL are likely to contain the same subject, sets of images from the same URL form an initial set of proposed subject clusters. We then performed extensive image-to-image and image-to-text similarity filtering to retain image clusters that contain highly similar content. For each subject image cluster, we fine-tuned an expert model to specialize in the given subject. Then, we use the fine-tuned experts to synthesize new images given unseen creative captions proposed by large language models. However, the tuned expert models are not perfect and prone to errors, therefore, we adopt a quality validation metric to filter out a large portion of degraded outputs. The remaining high-quality images are provided as a training signal to teach the apprentice model `SuTI` to perform subject-driven image generation with high fidelity. During inference, the trained `SuTI` can attend to a few in-context demonstrations to synthesize new images on the fly.

[2]Running on A100 according to public colab: https://huggingface.co/sd-dreambooth-library and https://huggingface.co/docs/diffusers/training/text_inversion.

We evaluate SuTI on various tasks such as subject re-contextualization, attribute editing, artistic style transfer, and accessorization. We compare SuTI with existing models on DreamBench [30], which contains diverse subjects from wide categories accompanied by some prompt templates. We compute the CLIP-I/CLIP-T and DINO scores of SuTI's generated images on this dataset and compare them with DreamBooth. The results indicate that SuTI can outperform DreamBooth while having 20x faster inference speed and significantly less memory footprint.

Further, we manually created 220 diverse and compositional prompts regarding the subjects in DreamBench for human evaluation, which is dubbed the DreamBench-v2 dataset. We then comprehensively compare with other baselines like InstructPix2Pix [5], Null-Text Inversion [22], Imagic [18], Textual Inversion [10], Re-Imagen [7], and DreamBooth [30] on DreamBench-v2. Our human evaluation results indicate that SuTI is on par with DreamBooth, while at least 30% better than the best baseline in terms of three metrics: subject fidelity, textual fidelity photorealism.

We summarize our contributions in the following aspects:

- We introduce the SuTI model, a subject-driven text-to-image generator that performs instant and customized generation for a visual subject with few (image, text) exemplars, *all in context*.
- We propose a novel *apprenticeship learning* to train the apprentice SuTI model to imitate half a million fine-tuned subject-specific experts on a large-scale seed dataset, leading to a generator model that generalizes to unseen subjects and unseen compositional descriptions.
- We perform a comprehensive set of automatic and human evaluations to show the capability of our model on generating highly faithful and creative images, on the challenging DreamBench [30] and DreamBench-v2, without optimizing the SuTI at all.

2 Preliminary

In this section, we introduce the key concepts and notations about subject-driven image-text data, then discuss the basics of text-to-image diffusion models.

Diffusion Models. Diffusion models [34] are latent variable models, parameterized by Θ, in the form of $p_\Theta(\boldsymbol{x}_0) := \int p_\Theta(\boldsymbol{x}_{0:T}) d\boldsymbol{x}_{1:T}$, where $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T$ are "noised" latent versions of the input image $\boldsymbol{x}_0 \sim q(\boldsymbol{x}_0)$. Note that the dimensionality of both latents and the image is the same throughout the entire process, with $\boldsymbol{x}_{0:T} \in \mathbb{R}^d$ and d equals the product of <height, width, # of channels>. The process that computes the posterior distribution $q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)$ is also called the forward (or diffusion) process, and is implemented as a predefined Markov chain that gradually adds Gaussian noise to the data according to a schedule β_t:

$$q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) = \prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) \tag{1}$$

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) := \mathcal{N}(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \boldsymbol{I}) \tag{2}$$

Diffusion models are trained to learn the image distribution by reversing the diffusion Markov chain. Theoretically, this reduces to learning to denoise $\boldsymbol{x}_t \sim q(\boldsymbol{x}_t|\boldsymbol{x}_0)$ into $\boldsymbol{x}_0$, with a time re-weighted square error loss—see [14] for the complete proof:

$$\mathbb{E}_{(\boldsymbol{x}_0, \boldsymbol{c}) \sim D}\{\mathbb{E}_{\boldsymbol{\epsilon}, t}[w_t \cdot ||\hat{\boldsymbol{x}}_\theta(\boldsymbol{x}_t, \boldsymbol{c}) - \boldsymbol{x}_0||_2^2]\} \tag{3}$$

where D is the training dataset containing (image, condition) = $(\boldsymbol{x}_0, \boldsymbol{c})$ pairs, the condition normally refers to the input text prompt. In practice, w_t can be simplified as 1 according to to [14, 35].

Subject-Driven Text-to-Image Generation. Existing subject-driven generation models [30, 19, 10] often fine-tune a pre-trained text-to-image diffusion model on a set of provided demonstrations $\mathbb{C}_s$ about a specific subject s. Formally, such demonstration contains a set of text and image pairs $\mathbb{C}_s = \{(\boldsymbol{x}_k, \boldsymbol{c}_k)\}_k^{K_s}$, centered around the subject s. Images $\boldsymbol{x}_k$ contains images of the same subject s, while $\boldsymbol{c}_s$ is a short description of images $\boldsymbol{x}_k$. DreamBooth [30] also requires an additional $\bar{\mathbb{C}}_s$, which contains images about different subjects of the same category as s for prior preservation. To obtain a customized diffusion model $\hat{\boldsymbol{x}}_{\theta_s}(\boldsymbol{x}_t, \boldsymbol{c})$, we need to optimize the following loss function:

$$\theta_s = \arg\min_\theta \mathbb{E}_{(\boldsymbol{x}_0, \boldsymbol{c}) \sim \mathbb{C}_s \cup \bar{\mathbb{C}}_s}\{\mathbb{E}_{\boldsymbol{\epsilon}, t}[||\hat{\boldsymbol{x}}_\theta(\boldsymbol{x}_t, \boldsymbol{c}) - \boldsymbol{x}_0||_2^2]\} \tag{4}$$

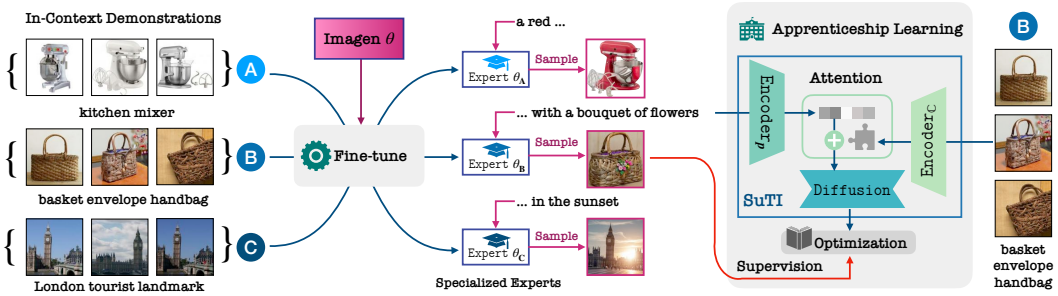

Figure 3: Overview of the apprentice learning pipeline for SuTI. Left part shows the customization procedure for expert models, and the right parts shows the SuTI model that imitates the behaviors of differently customized experts. Note that this framework can cope with expert models of *arbitary architecture and model family*.

The customized diffusion model $\hat{\boldsymbol{x}}_{\theta_s}(\boldsymbol{x}_t, \boldsymbol{c})$ has shown impressive capabilities to generate highly faithful images of the specified subject s.

3 Apprenticeship Learning from Subject-specific Experts

Notation. Figure 3 presents the concrete workflow of learning. Our method follows apprenticeship learning [1] with two major component, *i.e.*, the expert diffusion models $\hat{\boldsymbol{x}}_{\theta_s}(\boldsymbol{x}_t, \boldsymbol{c})$ parameterized by θ_s regarding subject $s \in \mathbb{S}$ and apprentice diffusion model $\hat{\boldsymbol{x}}_{\Theta}(\boldsymbol{x}_t, \boldsymbol{c}, \mathbb{C}_s)$ parameterized by Θ. The apprentice model takes an additional set of image-text demonstrations $\mathbb{C}_s$ as input. We use $\mathbb{S}$ to denote the superset of subjects we include in the training set.

Dataset. The training set $\mathcal{D}_{\mathbb{S}}$ contains a collection of $\{\mathbb{C}_s, \boldsymbol{p}_s\}_{s\in\mathbb{S}}$, where each entry contains an image-text cluster $\mathbb{C}_s$ accompanied by an unseen prompt $\boldsymbol{p}_s$. The image-text cluster $\mathbb{C}_s$ contains a set of 3-10 image-text pairs. The unseen prompt is an imaginary caption proposed by PaLM [9]. For example, if $\boldsymbol{c}$ is `'a photo of berry bowl'`, then $\boldsymbol{p}_s$ would be an imaginary caption like `'a photo of berry bowl floating on the river'`. We describe the dataset construction process to section 4.

Learning. To obtain an expert $\hat{\boldsymbol{x}}_{\theta_s}(\boldsymbol{x}_t, \boldsymbol{c})$ on a subject s, we fine-tune a pre-trained diffusion models [31] on the image cluster $\mathbb{C}_s$ with the denoising loss as:

$$\theta_s = \arg\min_{\theta} \mathbb{E}_{(\boldsymbol{x}_s, \boldsymbol{c})\sim\mathbb{C}_s}\{\mathbb{E}_{\boldsymbol{\epsilon},t}[||\hat{\boldsymbol{x}}_\theta(\boldsymbol{x}_t, \boldsymbol{c}) - \boldsymbol{x}_s||_2^2]\} \tag{5}$$

where $\boldsymbol{x}_t \sim q(\boldsymbol{x}_t|\boldsymbol{x}_s)$. The training is similar to Eqn. 4 except that we do not have negative examples for prior preservation because finding the negative examples from the same class is expensive.

Once an expert model is trained, we use it to sample images $\boldsymbol{y}_s$ for the unseen text description $\boldsymbol{p}_s$ to guide the apprentice SuTI model. We gather the outputs from the massive amount of expert models and then use CLIP filtering to construct a dataset G. Similarly, we fine-tune the apprentice model $\hat{\boldsymbol{x}}_{\Theta}(\boldsymbol{x}_t, \boldsymbol{p}_s, \mathbb{C}_s)$ with the denoising loss on the pseudo target generated by the expert:

$$\Theta = \arg\min_{\Theta} \mathbb{E}_{(\boldsymbol{y}_s, \boldsymbol{p}_s, \mathbb{C}_s)\sim G}\{\mathbb{E}_{\boldsymbol{\epsilon},t}[||\hat{\boldsymbol{x}}_\Theta(\boldsymbol{x}_t, \boldsymbol{p}_s, \mathbb{C}_s) - \boldsymbol{y}_s||_2^2]\} \tag{6}$$

where $\boldsymbol{x}_t \sim q(\boldsymbol{x}_t|\boldsymbol{y}_s)$, and the training triples $(\boldsymbol{y}_s, \boldsymbol{p}_s, \mathbb{C}_s)$ are drawn from G.

Algorithm. We formally introduce our learning algorithm in the Algorithm 1. To improve the training efficiency, we use distributed training algorithm to accelerate the training process. At each training step, we randomly sample a batch $\{B_{s_i}\}_{i=1}^{\mathtt{K}}$ of size K from the dataset $\mathcal{D}_{\mathbb{S}}$, with $B_{s_i} = (\mathbb{C}_{s_i}, \boldsymbol{p}_{s_i})$. We then fine-tune K expert models separately w.r.t. Eqn. 5 in parallel, across K different TPU cores. For every subject s inside the batch B_s, we use the corresponding expert model θ_s to synthesize the image $\boldsymbol{y}_s$ given the unseen prompt $\boldsymbol{p}_s$. As not all expert models can generate highly faithful images, we introduce a quality assurance step to validate the synthesized images. Particularly, we measure the quality of an expert's generation by the delta CLIP score [20] $\Delta(\boldsymbol{y}_s, \mathbb{C}_s, \boldsymbol{p}_s)$, which is used to decide whether a sample should be included in the dataset G. This ensures the high quality of the text-to-image training signal for SuTI. Specifically, the delta CLIP score is computed as the increment of CLIP score of $\boldsymbol{y}_s$ over the demonstrated images $\boldsymbol{x} \in \mathbb{C}_s$:

$$\Delta(\boldsymbol{y}_s, \mathbb{C}_s, \boldsymbol{p}_s) = \text{CLIP}(\boldsymbol{y}_s, \boldsymbol{p}_s) - \max_{\boldsymbol{x}\in\mathbb{C}_s} \text{CLIP}(\boldsymbol{x}, \boldsymbol{p}_s) \tag{7}$$

Algorithm 1 Apprenticeship Learning from a Large Crowd of Specialized Expert Models

Input: Dataset $\mathcal{D}_{\mathbb{S}} = \{(\mathbb{C}_s, \boldsymbol{p}_s)\}_{s \in \mathbb{S}}$ containing subject image cluster $\mathbb{C}_s$ and unseen prompt $\boldsymbol{p}_s$
Input: Pre-trained diffusion model parameterized by θ
Output: Apprentice diffusion model parameterized by Θ
Initialize SuTI parameters θ
Initialize a dataset $G = \varnothing$
while $\mathcal{D}_{\mathbb{S}} \neq \varnothing$ **do**
  $\{B_{s_i}\}_{i=1}^{\mathtt{K}} = \mathtt{Dequeue}(\mathcal{D}_{\mathbb{S}}, \mathtt{K})$, where $B_{s_i} = (\mathbb{C}_{s_i}, \boldsymbol{p}_{s_i})$
  Fine-tune K expert models $\theta_{s_1}, \ldots, \theta_{s_{\mathtt{K}}}$ on $\{B_{s_i}\}_{i=1}^{\mathtt{K}}$ in parallel, based on Eqn. 5
  for $i = 1$ **to** K **do**
    Sample a subject-specific generation $\boldsymbol{y}_{s_i}$ with DDPM using $\hat{\boldsymbol{x}}_{\theta_{s_i}}(\boldsymbol{x}_t, \boldsymbol{p}_{s_i})$
    if $\Delta(\boldsymbol{y}_{s_i}, \mathbb{C}_{s_i}, \boldsymbol{p}_{s_i}) > \lambda$ **then**
      $G = \mathtt{Enqueue}(G, (\boldsymbol{y}_s, \mathbb{C}_{s_i}, \boldsymbol{p}_{s_i}))$
    end if
  end for
end while
Train $\hat{\boldsymbol{x}}_\Theta$ on the generated dataset G, based on the Eqn. 6

We then feed G as a training batch to update the parameter Θ of the apprentice model using Eqn. 6. In all our experiments, we set $\mathtt{K} = 400$, with each TPU core training an expert model.

Inference. To perform subject-driven text-to-image generation, the trained SuTI takes 3-5 image-text pairs as the demonstration to generate new images based on the given text description. No optimization is needed during inference time. The only overhead of SuTI is the cost of encoding these 3-5 image-text pairs and the attention computation, which is more affordable. Our inference speed is roughly in the same order as the original text-to-image generator [31].

4 Mining and Generating Subject-driven Text-to-Image Demonstrations

In this section, we discuss how we created the seed dataset $\mathcal{D}_{\mathbb{S}}$ by mining images and text over the Internet. We construct the seed dataset from a subset of WebLI [8] dataset. We cluster the images by their URL to create the initial image clusters, and then we filter the clusters to ensure high intra-cluster visual similarity. The filtered set of image-text clusters is denoted $\{\mathbb{C}_s\}_{s \in \mathbb{S}}$.

After obtaining the subject-driven image clusters, we further prompt a large language model [9] to generate a description about the subject, with the goal of creating descriptions of plausible imaginary visual scenes. The generating instances of the descriptions will require skills like *subject re-contextualization*, *attribute editing*, *artistic style transfer*, and *accessorization*. We denote the generated unseen captions as p_s. Together with $\mathbb{C}_s$, this forms the final dataset $\mathcal{D}_{\mathbb{S}}$.

The dataset $\mathcal{D}_{\mathbb{S}}$ contains a total of 2M $(\mathbb{C}_s, p_s)$ pairs. Using the aforementioned delta CLIP score filtering (using a high threshold $\lambda = 0.02$), we remove low-quality synthesized images $\boldsymbol{y}_s$ from the expert model, finally obtaining a dataset G with ∼500K $(\mathbb{C}_s, p_s)$ effective training pairs for the following apprenticeship learning.

5 Experiment

In this paper, we only train SuTI on the text → 64x64 diffusion model and retain the original 256x256 and 1024x1024 super-resolution as it is from Imagen [31].

Expert Models. The expert model is initialized from the original 2.1B Imagen 64x64 model. We tune each model on a single TPU core (32 GB) for 500 steps using Adafactor optimizer with a learning rate of 1e-5, which only takes 5 minutes to finish. We use classifier-free guidance to sample new images, where the guidance weight is set to 30. To avoid excessive memory costs, we use fine-tuned experts to sample pseudo-target images and then write the samples as separate files. SuTI will read these files asynchronously to maximize the training speed. Our expert models have a few distinctions

Figure 4: Comparison with other Image Editing and Image Personalization Models.

from the DreamBooth [30]: 1) we adopt Adafactor instead of Adam optimizer, 2) we do not include any class word token like '[DOG] dog' in the prompt. 3) we do not include in-class negatives for prior preservation. Though our expert model is weaker than DreamBooth, such design choices significantly reduce time/space costs to enable us to train millions of experts with reasonable resources.

Apprentice Model. The apprentice model contains 2.5B parameters, which is 400M parameters larger than the original 2.1B Imagen 64x64 model. The added parameters are coming from the extra attention layers over the demonstrated image-text inputs. We initialize our model from Imagen's checkpoint. For the additional attention layers, we use random initialization. The apprentice training is performed on 128 Cloud TPU v4 chips. We train the model for a total of 150K steps. We use an Adafactor optimizer with a learning rate of 1e-4. We use 3 demonstrations during training, while the model can generalize to leverage more demonstrations during inference.

Inference. We normally provide 4 demonstration image-text pairs to `SuTI` during inference. Increasing the number of demonstrations does not improve the generation quality much. We use a lower classifier-free guidance weight of 15 with DDPM [14] sampling strategy.

5.1 Datasets and Metrics

DreamBench. In this paper, we use the DreamBench dataset[3] proposed by DreamBooth [30]. The dataset contains 30 subjects like backpacks, stuffed animals, dogs, cats, clocks, etc. These images are downloaded from Unsplash[4]. The original dataset contains 25 prompt templates covering different skills like recontextualization, property modification, accessorization, etc. In total, there are a total of 750 unique prompts generated by the template. We follow the original paper to generate 4 images for each prompt to form the 3000 images for robust evaluation. We follow DreamBooth to adopt DINO, CLIP-I to evaluate the subject fidelity, and CLIP-T to evaluate the text fidelity.

DreamBench-v2. To further increase the difficulty and diversity of DreamBench, we annotate 220 prompts for the 30 subjects in DreamBench as DreamBench-v2. We gradually increase the compositional levels of the prompt to increase the difficulty, like 'back view of [dog]' $\rightarrow$ 'back view of [dog] watching TV' $\rightarrow$ 'back view of [dog] watching TV about birds'. This enables us to perform a breakdown analysis to understand the model's compositional capabilities.

We use human evaluation to measure the generation quality in DreamBench-v2. Specifically, we aim at measuring the following three aspects: (1) the subject fidelity score s_s measures whether the

[3] https://github.com/google/dreambooth
[4] https://unsplash.com/.

subject is being preserved, (2) the textual fidelity score s_t measures whether it is aligned with the text description, (3) the photorealism score s_p measures whether the image contains artifacts or blurry subjects. These are all binary scores, which are averaged over the entire dataset. We combine them as an overall score $s_o = s_s \wedge s_t \wedge s_p$, which is the most stringent score.

5.2 Main Results

Baselines. We provide a comprehensive list of baselines to compare with the proposed `SuTI` model:

- *DreamBooth* [30]: a fine-tuning method, which trains the whole model on the given images for 500 steps, and then stores the new checkpoint in the disk. The space consumption is $|M| \times |\mathbb{S}|$ with the model size of $|M|$.
- *Textual Inversion* [10]: a fine-tuning method, which trains the embedding on the given images for 2000 steps, and then stores the trained embedding in the disk. The space consumption is $|E| \times |\mathbb{S}|$ with the embedding size of $|E|$, note that $|E| \ll |M|$.
- *Null-Text Inversion* [22]: a optimization method to find the trace using DDIM inversion [35], it requires storing a null embedding per time step. The space consumption is $|T| \times |E| \times |\mathbb{S}|$, which is $|T|$ times larger than Textual Inversion.
- *Imagic* [18]: a fine-tuning-based method, which requires sequentially optimizing the input text embedding, and then the text-to-image diffusion model, to produce edits on one given image. Therefore, it has the most expensive space consumption among all models as it requires training $|M| \times |\mathbb{S}| \times |\mathbb{P}|$, where $|\mathbb{P}|$ is the number of a text prompt $\mathbb{P} = \{\boldsymbol{p}_s\}$ for the subject set $\mathbb{S}$.
- *InstructPix2Pix* [5]: a non-tuning method, which can generate and edit a given image really fast within a few seconds. There is no additional space consumption.
- *Re-Imagen* [7]: a non-tuning method, which will take a few images as input and then attend to those retrievals to generate a new image. There is no additional space consumption.

Experimental Results. We show our automatic evaluation results on the DreamBench in Table 1. We can observe that `SuTI` can perform better or on par with DreamBooth on all of the metrics. Specifically, `SuTI` outperforms DreamBooth on the DINO score by 5%, which indicates that our method is better at preserving the subject's visual appearance. In terms of the CLIP-T score, our method is almost the same as DreamBooth, indicating an equivalent capability in terms of textual alignment. These results indicate that `SuTI` has achieved promising generalization to a wide variety of visual subjects, without being trained on the exact instances.

Methods	Backbone	DINO ↑	CLIP-I ↑	CLIP-T ↑
Real Image (Oracle)	-	0.774	0.885	-
DreamBooth [30]	Imagen [31]	0.696	0.812	**0.306**
DreamBooth [30]	SD [28]	0.668	0.803	0.305
Textual Inversion [10]	SD [28]	0.569	0.780	0.255
Re-Imagen [7]	Imagen [31]	0.600	0.740	0.270
Ours: `SuTI`	Imagen [31]	**0.741**	**0.819**	0.304

Table 1: Automatic Evaluation on the DreamBench.

We further show our human evaluation results on the DreamBench-V2 in Table 2. It shows the related rankings for the additional storage cost and reported the average inference time measure for inferring on each subject. As can be seen, our model `SuTI` and DreamBooth are both obtaining an almost perfect score. On the overall performance, `SuTI` is around 95% while DreamBooth is around 98%. `SuTI` only fails on a few examples in DreamBench-v2 (shown in Figure 8). In contrast, all the existing baselines are getting much lower human evaluation score.

Comparisons. We compare our generation results with other methods in Figure 4. As can be seen, `SuTI` can generate images highly faithful to the demonstrated subjects. Though `SuTI` is still missing some local textual (words on the bowl gets blurred) or colorization (dog hair color gets darker), the nuance is almost unperceivable for humans. The other baselines like InstructPix2Pix [5], and Null-Text Inversion [22] are not able to perform very sophisticated transformations. Textual Inversion [10] cannot achieve satisfactory results even with 30 minutes of tuning. Re-Imagen [7]

Methods	Backbone	Space	Time	Subject ↑	Text ↑	Photorealism ↑	Overall ↑
Models requiring test-time tuning							
Textual Inversion [10]	SD [28]	$	30 mins	0.22	0.64	0.90	0.14
Null-Text Inversion [22]	Imagen [31]	$$	5 mins	0.20	0.46	0.70	0.10
Imagic [18]	Imagen [31]	$$$$	70 mins	0.78	0.34	0.68	0.28
DreamBooth [30]	SD [28]	$$$	6 mins	0.74	0.53	0.85	0.47
DreamBooth [30]	Imagen [31]	$$$	10 mins	**0.88**	0.82	**0.98**	0.77
Models not requiring test-time tuning							
InstructPix2Pix [5]	SD [28]	-	10 secs	0.14	0.46	0.42	0.10
Re-Imagen [7]	Imagen [31]	-	20 secs	0.70	0.65	0.64	0.42
Ours: SuTI	Imagen [31]	-	30 secs	**0.88**	**0.90**	0.88	**0.80**

Table 2: Human Evaluation on the DreamBench-v2. We report an approximated average inference time (averaged over subjects) and the relative rankings of the space cost (more $: more expensive). Methods that do not fine-tune in test-time requires no additional storage (denoted by -). Time includes both training time and checkpoint saving time.

though gives reasonable outputs, the subject preservation is much weaker than SuTI. Imagic [18] also generates reasonable outputs, however, its failure rate is still much higher than ours. DreamBooth [30] however generates almost perfect images except for the 'blurry' text on the berry bowl. Through the comparison, we can observe remarkable improvement in the output image quality.

Skillset. We provide SuTI's generation to showcase its ability in re-contextualization, novel view synthesis, art rendition, property modification, and accessorization. We demonstrate these different skills in Figure 5. In the first row, we show that SuTI is able to synthesize the subjects with different art styles. In the second row, we show that SuTI is able to synthesize the different view angles of the given subject. In the third row, we show that SuTI can modify subjects' facial expressions like 'sad', 'screaming', etc. In the fourth row, we show that SuTI can alter the color of a given toy. In the last two rows, we show that SuTI can add different accessories (hats, clothes, etc) to the given subjects. Further, we found that SuTI can even compose two skills together to perform highly complex image generation. As depicted in Figure 6, we show that SuTI can combine re-contextualization with editing/accessorization/stylization to generate high-quality images.

5.3 Model Analysis and Ablation Study

We further conducted a set of ablation studies to show factors that impact the performance of SuTI.

Impact of # Demonstrations. Figure 7 presents the SuTI's in-context generation with respect to an increasing number of subject-specific image examples. Interestingly, we observe a transition in the model's behavior as the number of in-context examples increases. When $\mathbb{C}_s = \varnothing$, SuTI generates images using it prior to the text, similar to traditional text-to-image generation models such as Imagen [31]. When $|\mathbb{C}_s| = 1$, SuTI behaves similarly to an image editing model, attempting to edit the generation while preserving the foreground subject, and avoiding sophisticated transformation. When $|\mathbb{C}_s| = 5$, SuTI unlocks the capability of rendering novel pose and shape of the demonstrated subject naturally in the targeted scene. In addition, we also observe that a bigger $|\mathbb{C}_s|$ would result in a more robust generation of high text and subject alignment, and better photorealism. We also performed a human evaluation on the SuTI's generation with respect to different numbers of demonstrations and visualizes the results in Figure 7 (right). It shows that as the number of demonstrations increases, the human evaluation score first increases drastically and then gradually converges.

Quality of the expert dataset matters. We found that the Delta CLIP score is critical to ensure the quality of synthesized target images. Such a filtering mechanism is highly influential in terms of SuTI's final performance. We evaluated several versions to increase the Δ threshold from None $\rightarrow$ 0.0 $\rightarrow$ 0.01 $\rightarrow$ 0.015 $\rightarrow$ 0.020 $\rightarrow$ 0.025, we observe that the human evaluation (overall score) can increase from 0.54 $\rightarrow$ 0.70 $\rightarrow$ 0.78 $\rightarrow$ 0.84 $\rightarrow$ 0.88 $\rightarrow$ 0.87. Without such intensive filtering, the model's overall human score can go to a very low level (54%). With an increasing Δ, although the size of the dataset G keeps decreasing from 1.8M to around 500K, the model's generation quality keeps improving until saturation. The empirical study indicates that $\Delta = 0.02$ strikes a good balance between the quality and quantity of the expert-generated dataset G.

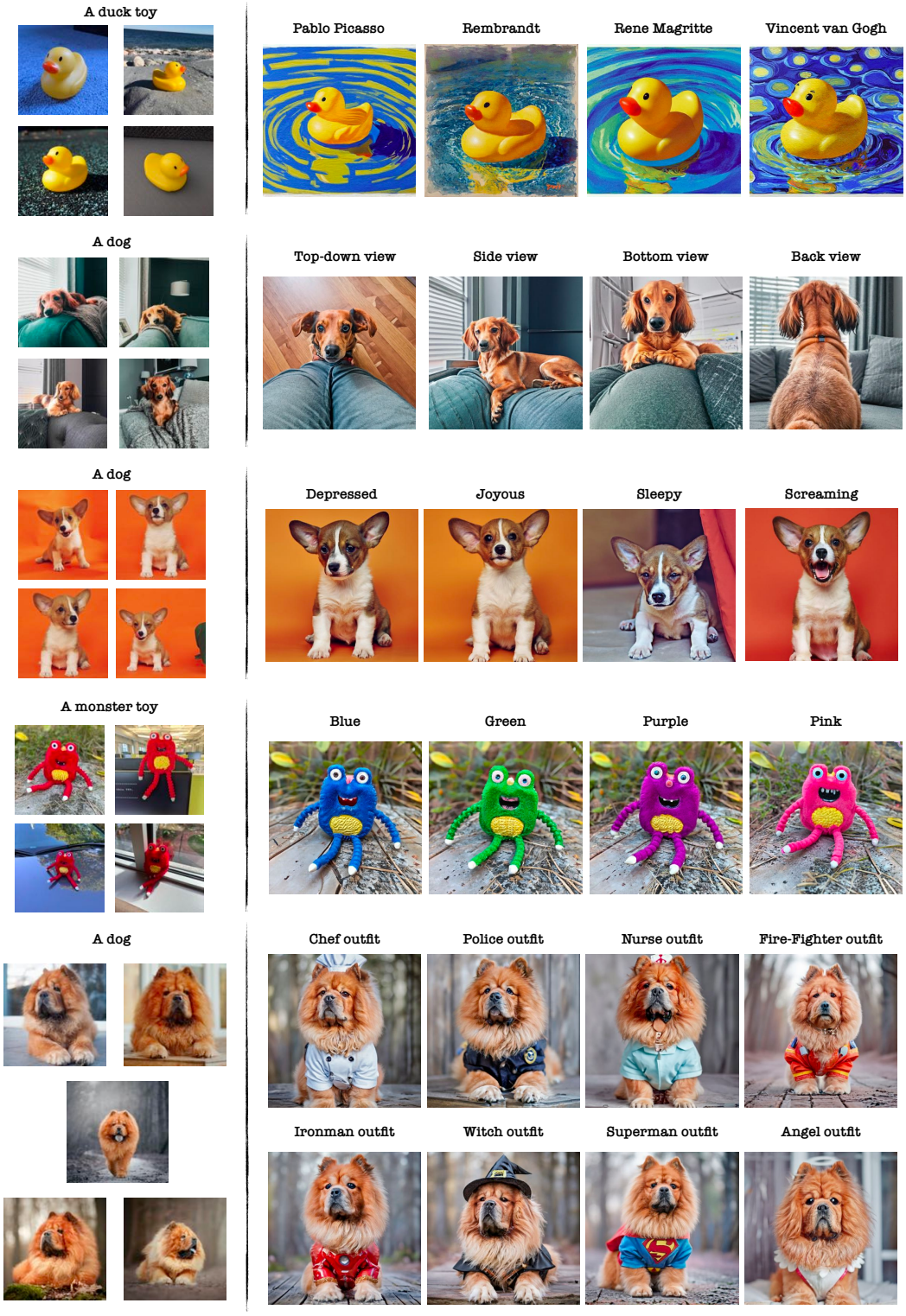

Figure 5: SuTI's in-context generation that demonstrates its skill set. Results generated from *a single model*. First row: art rendition of the subject. Second row: multi-view synthesis of the subject. Third row: modifying expression for the subject. Fourth row: editing the color of the subject. Fifth row: adding accessories to the subject. Subject (image, text) and editing key words are annotated, with detailed template in the Appendix.

Figure 6: SuTI not only re-contextualizes subjects but also composes multiple transformation, all in-context.

Figure 7: (Left) In-context generation by SuTI model, with an increasing # of demonstrations. (Right) Human evaluation score with with respect to the increasing % of demonstrations.

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.

6 Related Work

Text-to-Image Generation Recently, text-to-image generation models [36, 23, 31, 26, 29, 37] have gained unprecedented popularity. 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, 35] and auto-regressive backbone [27]. 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

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

existing models [30]. Also, as suggested by [7], 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.

Text-Guided Image Editing Image has been a long-standing task to study how to preserve object appearance while changing the surrounding context. Recently, different approaches have been proposed to tackle text-driven image editing. Previously, different GAN-based models [12, 16, 17, 15] have shown great progress in generating high-quality images. To decrease manual labor, [25, 10] have demonstrated how to achieve more realistic manipulations over the given image. Recently, more works have focused on using diffusion models to perform image editing.

Blended-Diffusion [2] and SDEdit [21] propose to blend the noise with the input image to guide the image synthesize process to maintain the original layout. Text2Live [3] generates an edit layer on top of the original image/video input. Prompt-to-Prompt [13] and Null-Text Inversion [22] aims at manipulating the attention map in the diffusion model to maintain the layout of the image while changing certain subjects. Imagic [18] propose an optimization based to achieve significant progress in manipulating visual details in a given image. InstructPix2Pix [5] propose to distill image editing training pairs synthesized from Prompt-to-Prompt into a single diffusion model to perform instruction-driven image editing. Our method resembles InstructPix2Pix in a sense that we are training the model on expert-generated images. However, our synthesized data is generated generated by fine-tuned experts, which are mostly natural images. In contrast, the images from InstructPix2Pix are synthetic images. In the experiment section, we comprehensively compare with these existing models to show the advantage of our model, especially on more challenging prompts.

Subject-Driven Text-to-Image Generation Different from Text-Driven Image Editing, Subject-Driven Image Generation tackles a new challenge. Subject-Driven Image Generation requires the model to understand the visual subject contained in the demonstrations and synthesize totally unseen scene of the demonstrated subjects. Instance-GAN [6] and MyStyle [24] are the pioneers to work on this problem to personalize the image generation model to a particular instance and generate new images. Later on, DreamBooth [30] and Textual Inversion [10, 11] propose optimization-based approach to adapt image generation to a specific unseen subject. However, these two methods are time and space-consuming, which makes them unrealistic in real-world applications. Another line of work adopt retrieval-augmented architecture to enhance the image generation model like KNN-Diffusion [33], Re-Imagen [7] and RetrieveDiffusion [4], however, these methods are trained with weakly-supervised data leading to much worse faithfulness than DreamBooth [30]. In this paper,

we aim at developing an apprenticeship learning paradigm to train the image generation model with stronger supervision demonstrated by fine-tuned experts.

7 Discussion

Our method SuTI has shown strong capabilities to generate personalized images instantly without any further test-time optimization. Through automatic evaluation, our model is already superseding the existing SoTA model, however, we do identify a few weakness of our model. First of all, we identify that SuTI's generations are less diverse than DreamBooth [30], and our model is less inclined to transform the subjects' poses or views in the new image. On the other hand, we also found that SuTI is less faithful to the low-level visual details than DreamBooth [30], especially for more complex and often manufactured subjects such as 'robots' or 'rc cars' where the subjects are contain highly sophisticated visual details that could be arbitrarily different from the examples inside the training dataset. In the future, we plan to investigate how to further improve the low-level visual details.

Broader Impact

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 [31, 37] for more discussion.

References

[1] Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 1, 2004. 4

[2] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022. 11

[3] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, pages 707–723. Springer, 2022. 11

[4] Andreas Blattmann, Robin Rombach, Kaan Oktay, Jonas Müller, and Björn Ommer. Semi-parametric neural image synthesis. In *Advances in Neural Information Processing Systems*, 2022. 11

[5] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022. 3, 7, 8, 11

[6] Arantxa Casanova, Marlene Careil, Jakob Verbeek, Michal Drozdzal, and Adriana Romero Soriano. Instance-conditioned gan. *Advances in Neural Information Processing Systems*, 34:27517–27529, 2021. 11

[7] Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. Re-imagen: Retrieval-augmented text-to-image generator. *International Conference on Learning Representations*, 2023. 2, 3, 7, 8, 11

[8] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022. 2, 5

[9] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. 4, 5

[10] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2, 3, 7, 8, 11

[11] Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Designing an encoder for fast personalization of text-to-image models. *arXiv preprint arXiv:2302.12228*, 2023. 11

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 11

[13] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 11

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 3, 6, 10

[15] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021. 11

[16] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 11

[17] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 11

[18] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. *arXiv preprint arXiv:2210.09276*, 2022. 2, 3, 7, 8, 11

[19] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[20] Manling Li, Ruochen Xu, Shuohang Wang, Luowei Zhou, Xudong Lin, Chenguang Zhu, Michael Zeng, Heng Ji, and Shih-Fu Chang. Clip-event: Connecting text and images with event structures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16420–16429, 2022. 4

[21] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2021. 11

[22] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. *arXiv preprint arXiv:2211.09794*, 2022. 3, 7, 8, 11

[23] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, pages 16784–16804. PMLR, 2022. 10

[24] Yotam Nitzan, Kfir Aberman, Qiurui He, Orly Liba, Michal Yarom, Yossi Gandelsman, Inbar Mosseri, Yael Pritch, and Daniel Cohen-Or. Mystyle: A personalized generative prior. *ACM Transactions on Graphics (TOG)*, 41(6):1–10, 2022. 11

[25] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021. 11

[26] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2, 10

[27] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 2, 10

[28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 7, 8

[29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 10

[30] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022. 2, 3, 6, 7, 8, 11, 12

[31] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, 2022. 2, 4, 5, 7, 8, 10, 12

[32] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 2

[33] Shelly Sheynin, Oron Ashual, Adam Polyak, Uriel Singer, Oran Gafni, Eliya Nachmani, and Yaniv Taigman. Knn-diffusion: Image generation via large-scale retrieval. *arXiv preprint arXiv:2204.02849*, 2022. 11

[34] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 3

[35] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020. 3, 7, 10

[36] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019. 10

[37] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022. 2, 10, 12

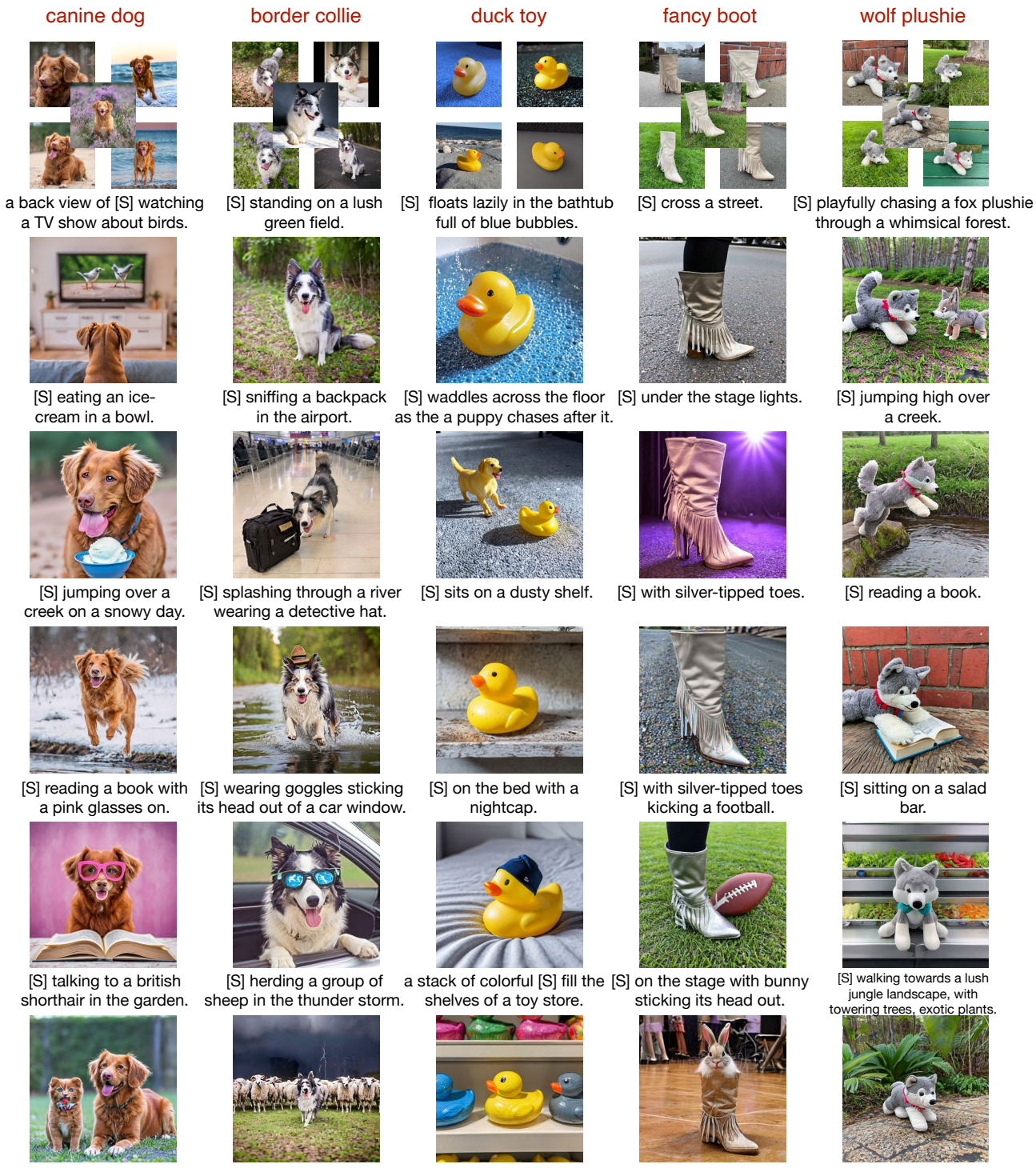

Figure 9: Visualization of SuTI's generation on the DreamBench-v2 (Part 1).

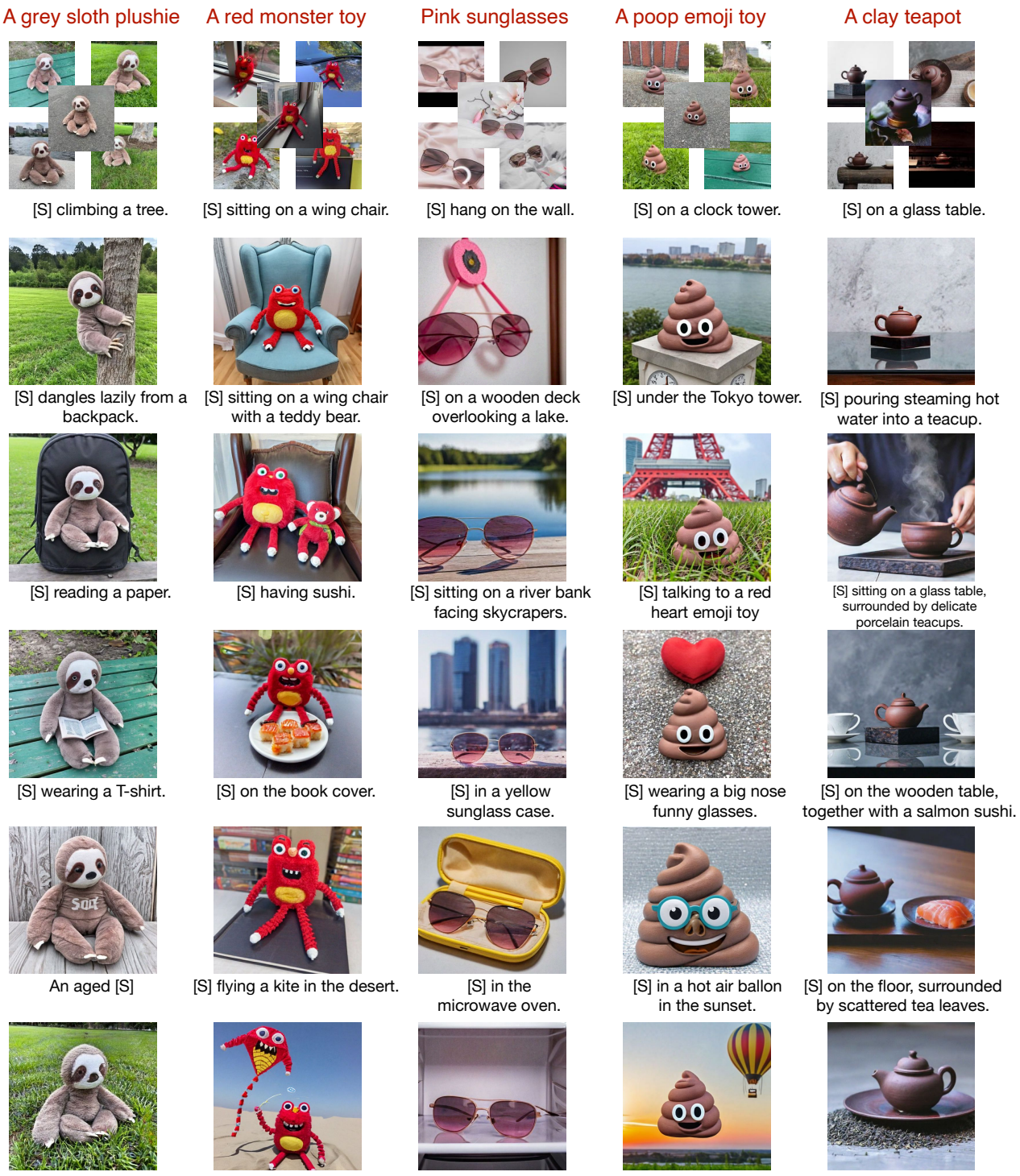

Figure 10: Visualization of SuTI's generation on the DreamBench-v2 (Part 2).

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).