DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation

Nataniel Ruiz[1,2], Yuanzhen Li[1], Varun Jampani[1], Yael Pritch[1], Michael Rubinstein[1], and Kfir Aberman[1]

[1]*Google Research* [2]*Boston University*

Figure 1: With just a few images (typically 3-5) of a subject (left), *DreamBooth*—our AI-powered photo booth—can generate a myriad of images of the subject in different contexts (right), using the guidance of a text prompt. The results exhibit natural interactions with the environment, as well as novel articulations and variation in lighting conditions, all while maintaining high fidelity to the key visual features of the subject. Image credit (input images): Unsplash.

Abstract

Large text-to-image models achieved a remarkable leap in the evolution of AI, enabling high-quality and diverse synthesis of images from a given text prompt. However, these models lack the ability to mimic the appearance of subjects in a given reference set and synthesize novel renditions of them in different contexts. In this work, we present a new approach for "personalization" of text-to-image diffusion models (specializing them to users' needs). Given as input just a few images of a subject, we fine-tune a pretrained text-to-image model (Imagen, although our method is not limited to a specific model) such that it learns to bind a unique identifier with that specific subject. Once the subject is embedded in the output domain of the model, the unique identifier can then be used to synthesize fully-novel photorealistic images of the subject contextualized in different scenes. By leveraging the semantic prior embedded in the model with a new autogenous class-specific prior preservation loss, our technique enables synthesizing the subject in diverse scenes, poses, views, and lighting conditions that do not appear in the reference images. We apply our technique to several previously-unassailable tasks, including subject recontextualization, text-guided view synthesis, appearance modification, and artistic rendering (all while preserving the subject's key features). Project page: `https://dreambooth.github.io/`

Preprint. Under review.

1 Introduction

Can you imagine your own dog traveling around the world, or your favorite bag displayed in the most exclusive showroom in Paris? What about your parrot being the main character of an illustrated storybook? Rendering such imaginary scenes is a challenging task that requires synthesizing instances of specific subjects (objects, animals, etc.) in new contexts such that they naturally and seamlessly blend into the scene.

Recently developed large text-to-image models achieve a remarkable leap in the evolution of AI, by enabling high-quality and diverse synthesis of images based on a text prompt written in natural language [56, 51]. One of the main advantages of such models is the strong semantic prior learned from a large collection of image-caption pairs. Such a prior learns, for instance, to bind the word "dog" with various instances of dogs that can appear in different poses and contexts in an image. While the synthesis capabilities of these models are unprecedented, they lack the ability to mimic the appearance of subjects in a given reference set, and synthesize novel renditions of those *same subjects* in different contexts. The main reason is that the expressiveness of their output domain is limited; even the most detailed textual description of an object may yield instances with different appearances. Furthermore, even models whose text embedding lies in a shared language-vision space [49] cannot accurately reconstruct the appearance of given subjects but only create variations of the image content (Figure 2).

In this work, we present a new approach for "personalization" of text-to-image diffusion models (adapting them to user-specific image generation needs). Our goal is to expand the language-vision dictionary of the model such that it binds new words with specific subjects the user wants to generate. Once the new dictionary is embedded in the model, it can use these words to synthesize novel photorealistic images of the subject, contextualized in different scenes, while preserving their key identifying features. The effect is akin to a "magic photo booth"—once a few images of the subject are taken, the booth generates photos of the subject in different conditions and scenes, as guided by simple and intuitive text prompts (Figure 1).

More formally, given a few images of a subject ($\sim$3-5), our objective is to implant the subject into the output domain of the model such that it can be synthesized with a *unique identifier*. To that end, we propose techniques to represent a given subject with rare token identifiers and fine-tune a pre-trained, diffusion-based text-to-image framework that operates in two steps; generating a low-resolution image from text and subsequently applying super-resolution (SR) diffusion models. We first fine-tune the low-resolution text-to-image model with the input images and text prompts containing a unique identifier followed by the class name of the subject (e.g., "A [V] dog"). In order to prevent *overfitting* and *language drift* [35, 40] that cause the model to associate the class name (e.g., "dog") with the specific instance, we propose an *autogenous, class-specific prior preservation loss*, which leverages the semantic prior on the class that is embedded in the model, and encourages it to generate diverse instances of the same class as our subject. In the second step, we fine-tune the super-resolution component with pairs of low-resolution and high-resolution versions of the input images. This allows the model to maintain high fidelity to small (but important) details of the subject. We use the pre-trained Imagen model [56] as a base model in our experiments, although our method is not constrained to any specific text-to-image diffusion model.

To the best of our knowledge, ours is the first technique that tackles this challenging problem setting that enables users to take a few casually captured images of a subject and generate novel renditions of them in different contexts while maintaining their key features. Achieving such results with state-of-the-art 3D reconstruction techniques is non-trivial as it may restrict the possible diversity in the input set (e.g., if the input subject is an animal, its pose should be consistent), require more input views, and, even with a good 3D reconstruction, the rendition of the subject in novel articulations and contexts is still an open problem.

In summary, our two main contributions in this work are:

- A new problem: subject-driven generation. Given a few casually captured images of a subject, the goal is to synthesize novel renditions of the subject in different contexts, while maintaining high fidelity to its key visual features.
- A new technique for fine-tuning text-to-image diffusion models in a few-shot setting, while preserving the model's semantic knowledge on the class of the subject.

We discuss various insights related to the suggested mechanism and its analogy to related work, as well as highlight the contribution of each component via ablation studies, and compare our method with alternative baselines. We apply our approach to a myriad of text-based image generation applications including recontextualization of subjects, modification of their properties, original art renditions, and more, paving the way to a new stream of previously unassailable tasks.

Figure 2: **Subject-driven generation with text-to-image diffusion models.** Given a particular clock (shown in the real images on the left), it is very challenging to generate it in different contexts with state-of-the-art text-to-image models, while maintaining high fidelity to its key visual features. Even with dozens of iterations over a text prompt that contains a detailed description of the appearance of the clock (*"retro style yellow alarm clock with a white clock face and a yellow number three on the right part of the clock face in the jungle"*), the Imagen model [56] can't reconstruct its key visual features (third column). Furthermore, even models whose text embedding lies in a shared language-vision space and can create semantic variations of the image, such as DALL-E2 [51], can neither reconstruct the appearance of the given subject nor modify the context (second column). In contrast, our approach (right) can synthesize the clock with high fidelity and in new contexts (*"a [V] clock in the jungle"*). Image credit (input images): Unsplash.

2 Related work

Compositing Objects into Scenes Synthesizing a given subject in different contexts is a challenging task previously tackled in various forms and under different assumptions. For example, image composition techniques [64, 13, 38] aim to clone a given subject into a new background such that the subject melds seamlessly into the scene. However, these techniques are quite limited in their ability to adapt the correct lighting from the background, cast proper shadows, or fit the background content in a semantically aware manner. The challenge is even greater when the requirement is to synthesize the subject in novel articulations and views that were not seen in the few given images of the subject. For that task, even state-of-the-art 3D reconstruction techniques [41, 6, 8, 63] are limited as they usually work on rigid objects and require a larger number of views. In addition, it is extremely challenging to accurately infer material properties and illuminations from sparse image sets, which are required for realistic renderings of a given subject in novel scenes.

Text-Driven Editing One way of preserving object appearance while changing context is image editing. An intuitive approach is through user-provided text prompts. Text-driven image manipulation has recently achieved significant progress using GANs [22, 9, 29–31], which are known for their high-quality generation, in parallel with the advent of CLIP [49], which consists of a semantically rich joint image-text representations. Recent work [46, 21, 65, 1] combines these components to produce highly realistic manipulations using text without needing extra manual labor. Bau et al. [7] further demonstrated how to use masks provided by the user, to localize the text-based editing and restrict the change to a specific spatial region. However, while GAN-based image editing approaches succeed on highly-curated datasets [42], e.g., human faces, they struggle over diverse datasets with many subject types. To obtain more expressive generation capabilities, Crowson et al. [14] use VQ-GAN [18], trained over diverse data, as a backbone. Other works [4, 32] exploit the recent Diffusion models [25, 58, 60, 25, 59, 54, 44, 61, 55, 57], which achieve state-of-the-art generation quality over highly diverse datasets, often surpassing GANs [15]. Kim et al. [32] show how to perform global changes, whereas Avrahami et al. [4] successfully perform local manipulations using user-provided masks for guidance. While most works that require only text (i.e., no masks) are limited to global editing [14, 34], Bar-Tal et al. [5] proposed a text-based localized editing technique without using masks, showing impressive results. While most of these editing approaches allow modification of global properties or local editing of a given image, none of them enables generating novel renditions of a given subject in new contexts.

Text-to-Image Synthesis There also exists work on generation of images conditioned on text [16, 24, 62, 36, 37, 47, 48, 52, 67, 14, 19, 54, 27], also known as text-to-image synthesis. Recent large text-to-image models such as Imagen [56], DALL-E2 [51], Parti [66], and CogView2 [17] demonstrated unprecedented semantic generation. These models do not provide fine-grained control over a generated image and use text

guidance only. Specifically, given text descriptions, it is hard or impossible to preserve the identity of a subject consistently across different images, since modifying the context in the prompt also modifies the appearance of the subject. For example sentences such as "a shoe next to a river" and "a shoe on a desk" will display different kinds of shoes. Liu et al. [39] propose a modification to classifier guidance that allows for guidance of diffusion models using images and text, allowing for semantic variations of an image, although the identity of the subject often varies. To overcome subject modification, several works [43, 3] assume that the user provides a mask to restrict the area in which the changes are applied. Recent work of prompt-to-prompt [23] allows for local and global editing without any input mask requirements, but does not address the problem of synthesizing a given subject in novel contexts.

Inversion of diffusion models Inversion of diffusion models, the problem of finding a noise map and a conditioning vector that correspond to a generated image, is a challenging task and a possible avenue of tackling subject-driven novel image generation. Due to the asymmetry between the backward and forward diffusion steps, a naive addition of noise to an image followed by denoising, may yield a completely different image. Choi et al. [12] tackle inversion by conditioning the denoising process on noised low-pass filter data from the target image. Dhariwal et al. [15] show that deterministic DDIM sampling [59] can be inverted to extract a latent noise map that will produce a given real image. Ramesh et al. [51] use this method to generate cross-image interpolations or semantic image editing using CLIP latent vectors. These methods fall short of generating novel samples of a subject while preserving fidelity. The concurrent work of Gal *et al*. [20] proposes a textual inversion method that learns to represent visual concepts, like an object or a style, through new pseudo-words in the embedding space of a frozen text-to-image model. Their approach searches for the optimal embedding that can represent the concept, hence, is limited by the expressiveness of the textual modality and constrained to the original output domain of the model. In contrast, we fine-tune the model in order to embed the subject within the output domain of the model, enabling the generation of novel images of the subject while preserving key visual features that form its identity.

Personalization In recent years, personalization has become a prominent factor in various fields within Machine Learning such as recommendation systems [2], language models [11], and Federated Learning [28]. Within the vision and graphics community, there are few works that tackle the problem of novel synthesis of subjects using GANs. Casanova et al. [10] propose a method to condition GANs on instances, such that variations of an instance can be generated. Generated subjects share features with the conditioning instance, but are not identical and thus cannot solve the problem tackled in our work. Nitzan et al. [45] propose MyStyle to finetune a face synthesis GAN on a specific identity, in order to build a personalized prior. While MyStyle requires around 100 images to learn an adequate prior and it is constrained to the face domain, our approach can reconstruct the identity of different types of subjects (objects, animals, etc.) in new contexts with only 3-5 casually captured images.

3 Preliminaries

Cascaded Text-to-Image Diffusion Models Diffusion models are probabilistic generative models that are trained to learn a data distribution by the gradual denoising of a variable sampled from a Gaussian distribution. Specifically, this corresponds to learning the reverse process of a fixed-length Markovian forward process. In simple terms, a conditional diffusion model $\hat{\mathbf{x}}_\theta$ is trained using a squared error loss to denoise a variably-noised image $\mathbf{z}_t \coloneqq \alpha_t \mathbf{x} + \sigma_t \boldsymbol{\epsilon}$ as follows:

$$\mathbb{E}_{\mathbf{x},\mathbf{c},\boldsymbol{\epsilon},t}\left[w_t \|\hat{\mathbf{x}}_\theta(\alpha_t \mathbf{x} + \sigma_t \boldsymbol{\epsilon}, \mathbf{c}) - \mathbf{x}\|_2^2\right] \tag{1}$$

where $\mathbf{x}$ is the ground-truth image, $\mathbf{c}$ is a conditioning vector (e.g., obtained from a text prompt), $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is a noise term and α_t, σ_t, w_t are terms that control the noise schedule and sample quality, and are functions of the diffusion process time $t \sim \mathcal{U}([0, 1])$. At inference time, the diffusion model is sampled by iteratively denoising $\mathbf{z}_{t_1} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ using either the deterministic DDIM [59] or the stochastic ancestral sampler [25]. Intermediate points $\mathbf{z}_{t_1}, \ldots, \mathbf{z}_{t_T}$, where $1 = t_1 > \cdots > t_T = 0$, are generated, with decreasing noise levels. These points, $\hat{\mathbf{x}}_0^t \coloneqq \hat{\mathbf{x}}_\theta(\mathbf{z}_t, \mathbf{c})$, are functions of the $\mathbf{x}$-predictions.

Recent state-of-the-art text-to-image diffusion models use cascaded diffusion models in order to generate high-resolution images from text [56, 51]. Specifically, [56] uses a base text-to-image model with 64x64 output resolution, and two text-conditional super-resolution (SR) models $64 \times 64 \rightarrow 256 \times 256$ and $256 \times 256 \rightarrow 1024 \times 1024$. Ramesh et al. [51] use a similar configuration, with unconditional SR models. A key component of high-quality sample generations from [56] is the use of noise conditioning augmentation [26] for the two SR modules. This consists in corrupting the intermediate image using noise with specific strength, and then conditioning the SR model on the level of corruption. Saharia et al. [56] select Gaussian noise as the form of augmentation.

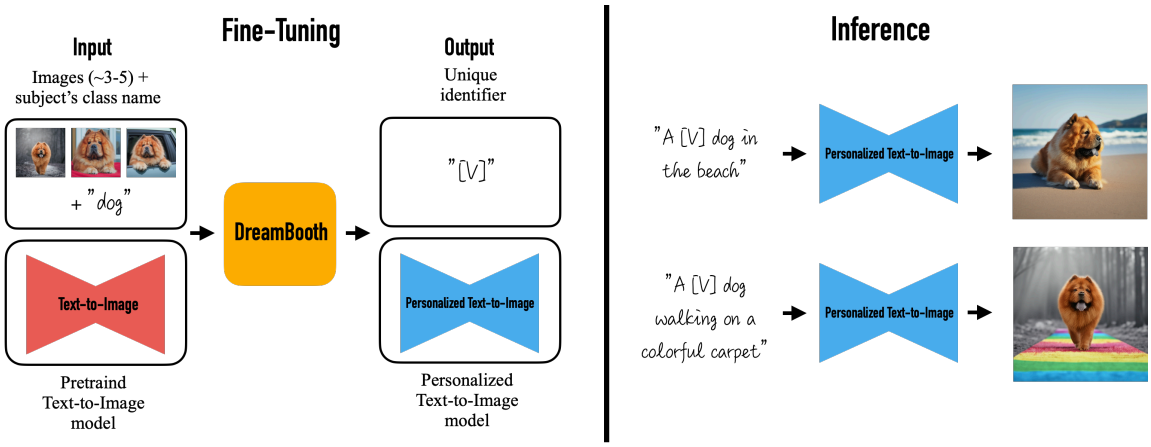

Figure 3: **High-level method overview.** Our method takes as input a few images (typically $3-5$ images suffice, based on our experiments) of a subject (e.g., a specific dog) and the corresponding class name (e.g. "dog"), and returns a fine-tuned/"personalized" text-to-image model that encodes a unique identifier that refers to the subject. Then, at inference, we can implant the unique identifier in different sentences to synthesize the subjects in difference contexts.

Vocabulary Encoding The details of text-conditioning in text-to-image diffusion models are of high importance for visual quality and semantic fidelity. Ramesh et al. [51] use CLIP text embeddings that are translated into image embeddings using a learned prior, while Saharia et al. [56] use a pre-trained T5-XXL language model [50]. In our work, we use the latter. Language models like T5-XXL generate embeddings of a tokenized text prompt, and vocabulary encoding is an important pre-processing step for prompt embedding. In order to transform a text prompt $\mathbf{P}$ into a conditioning embedding $\mathbf{c}$, the text is first tokenized using a tokenizer f using a learned vocabulary. Following [56], we use the SentencePiece tokenizer [33]. After tokenizing a prompt $\mathbf{P}$ using tokenizer f we obtain a fixed-length vector $f(\mathbf{P})$. The language model Γ is conditioned on this token identifier vector to produce an embedding $\mathbf{c} := \Gamma(f(\mathbf{P}))$. Finally, the text-to-image diffusion model is directly conditioned on $\mathbf{c}$.

4 Method

Given only a few (3-5) casually captured images of a specific subject, without any textual description, our objective is to generate new images of the subject with high detail fidelity and with variations guided by text prompts. We do not impose any restrictions on input image capture settings and the subject image can have varying contexts. Examples of output variations include: changing the place where the subject is, changing a property of the subject such as color, species, or shape, and modifying the subject's pose, expression, material, and other semantic modifications. We find that the breadth of these modifications is very large given the powerful prior of these models. A high-level overview of our method is presented in Figure 3.

In order to accomplish this, our first task is to implant the subject instance into the output domain of the model and to bind the subject with a *unique identifier*. We present our method to design the identifier below as well as a new approach to supervise a fine-tuning process of the model such that it re-uses its prior for our subject instance. A key problem is that fine-tuning on a small set of images showing our subject is prone to overfitting on the given images. In addition, *language drift* [35, 40] is a common problem in language models, and manifests itself in text-to-image diffusion models as well: the model can forget how to generate other subjects of the same class, and lose the embedded knowledge on the diversity and natural variations of instances belonging to that class. For this, we present an *autogenous class-specific prior preservation loss*, where we alleviate overfitting and prevent language drift by encouraging the diffusion model to keep generating diverse instances of the same class as our subject.

To enhance the details preservation, we find that the super-resolution components of the model should also be fine-tuned. However, if they are fine-tuned to generate our subject instance in a naive manner they are not able to replicate important details of the instance. We present insights that allow us to train and test these SR modules in order to better conserve subject details in novel generations - achieving unprecedented perfor-

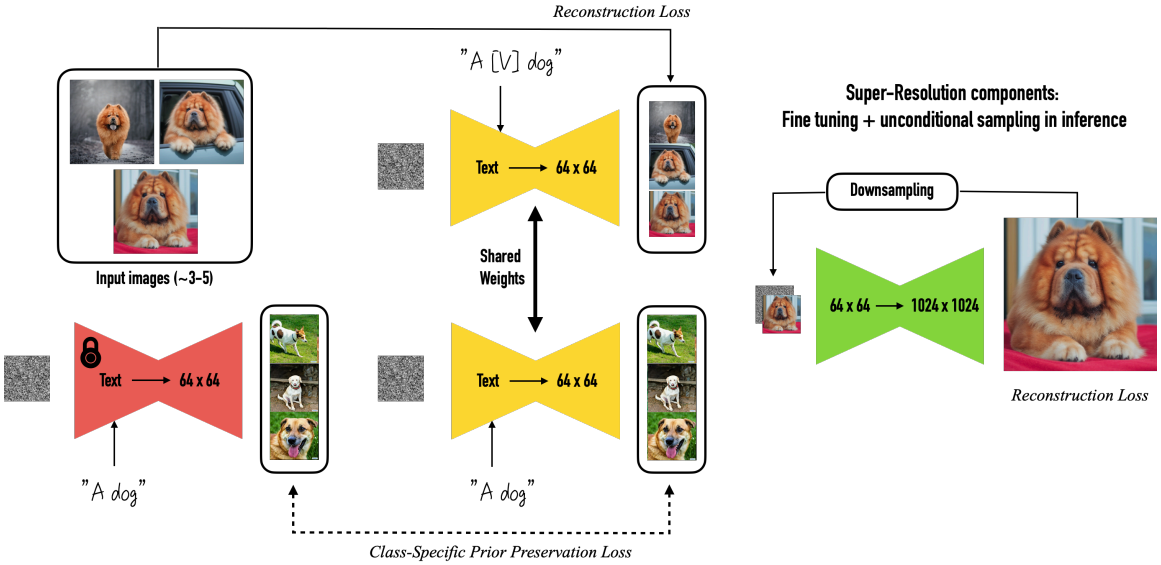

Figure 4: **Fine-tuning.** Given $\sim 3-5$ images of a subject we fine tune a text-to-image diffusion in two steps: (a) fine tuning the low-resolution text-to-image model with the input images paired with a text prompt containing a unique identifier and the name of the class the subject belongs to (e.g., "A [V] dog"), in parallel, we apply a class-specific prior preservation loss, which leverages the semantic prior that the model has on the class and encourages it to generate diverse instances belong to the subject's class using the class name in a text prompt (e.g., "A dog"). (b) fine-tuning the super resolution components with pairs of low-resolution and high-resolution images taken from our input images set, which enables us to maintain high-fidelity to small details of the subject.

mance in recontextualization. A detailed sketch of our proposed training procedure is shown in Figure 4. In this work we use the pre-trained Imagen model as the base model [56].

4.1 Representing the Subject with a Rare-token Identifier

Designing Prompts for Few-Shot Personalization Our goal is to "implant" a new (key, value) pair into the diffusion model's "dictionary" such that, given the key for our subject, we are able to generate fully-novel images of this specific subject with meaningful semantic modifications guided by a text prompt. One way is by few-shot fine-tuning of the model. The question becomes: how to supervise this process? Generally, text prompts are written by humans and sourced from large online datasets. The main limitations include the overhead of writing detailed image descriptions for a given image set, and the high variance and subjectivity of human-written captions.

We opt for a simpler approach and label all input images of the subject "a [identifier] [class noun]", where [identifier] is a unique identifier linked to the subject and [class noun] is a coarse class descriptor of the subject (e.g. cat, dog, watch, etc.). The class descriptor can be obtained using a classifier. We specifically use a class descriptor in the sentence in order to tether the prior of the class to our unique subject. We found that using only a unique identifier, without a class noun, as a key for our subject increased training time and decreased performance. In essence, we want to leverage the diffusion model's prior of the specific class and entangle it with the embedding of our subject's unique identifier. In this way, it can leverage the visual prior to generate new poses and articulations of the subject in different contexts.

A naive way of constructing an identifier for our subject is to use an existing word. For example, using the words like "unique" or "special". One problem is that existing English words tend to have a stronger prior due to occurrence in the training set of text-to-image diffusion models. We generally find increased training time and decreased performance when using such generic words to index our subject, since the model has to both learn to disentangle them from their original meaning and to re-entangle them to reference our subject. This approach can also fail by entangling the meaning of the word with the appearance of our object, for example in the extreme case if the identifier chosen is the adjective "blue" and our subject is grey, colors will be entangled at inference and we will sample a mix of grey and blue subjects (as well as mixes of both). This

motivates the need for an identifier that has a weak prior in both the language model and the diffusion model. A hazardous way of doing this is to select random characters in the English language and concatenate them to generate a rare identifier (e.g. "xxy5syt00"). In reality, the tokenizer might tokenize each letter separately, and the prior for the diffusion model is strong for these letters. Specifically, if we sample the model with such an identifier before fine-tuning we will get pictorial depictions of the letters or concepts that are linked to those letters. We often find that these tokens incur the same weaknesses as using common English words to index the subject.

Rare-token Identifiers In a nutshell, our approach is to find relatively rare tokens in the vocabulary, and then invert these rare tokens into text space. In order to do this, we first perform a rare-token lookup in the vocabulary and obtain a sequence of rare token identifiers $f(\hat{\mathbf{V}})$, where f is a tokenizer; a function that maps character sequences to tokens and $\hat{\mathbf{V}}$ is the decoded text stemming from the tokens $f(\hat{\mathbf{V}})$. This sequence can be of variable length k with k being a hyperparameter of our method. We find that relatively short sequences of $k = \{1, ..., 3\}$ work well. Then, by inverting the vocabulary using the de-tokenizer on $f(\hat{\mathbf{V}})$ we obtain a sequence of characters that define our unique identifier $\hat{\mathbf{V}}$. We observe that using uniform random sampling without replacement of tokens that correspond to 3 or fewer Unicode characters (without spaces) and using tokens in the T5-XXL tokenizer range of $\{5000, ..., 10000\}$ works well.

4.2 Class-specific Prior Preservation Loss

Few-shot Personalization of a Diffusion Model Given a small set of images depicting the target subject $\mathcal{X}_s := \{\mathbf{x}_s^i; i \in \{0, ..., N\}\}$, and with the same conditioning vector $\mathbf{c}_s$ obtained from the text prompt "a [identifier] [class noun]", we fine-tune the text-to-image model using the classic denoising loss presented in Equation 1, with the same hyperparameters as the original diffusion model. Two key issues arise with such a naive fine-tuning strategy: Overfitting and Language-drift.

Issue-1: Overfitting Since our input image set is quite small, fine-tuning the large image generation models can overfit to both the context and the appearance of the subject in the given input images (e.g., subject pose). Figure 12 (top) shows some sample generated images with naive fine-tuning where we clearly see that both the subject dog's appearance and context are overfitted to those in the input images. There are many techniques that can be used to address these problems, such as regularization or selectively fine-tuning certain parts of the model. There is uncertainty on which parts of the model need to be frozen to both obtain good subject fidelity and semantic modification flexibility. In our experience, the best results that achieve maximum subject fidelity are achieved by fine-tuning all layers of the model. Nevertheless, this includes fine-tuning layers that are conditioned on the text embeddings, which gives rise to the problem of language drift.

Issue-2: Language Drift The phenomenon of *language drift* has been an observed problem in the language model literature [35, 40], where a language model that is pre-trained on a large text corpus and later fine-tuned for a specific task progressively loses syntactic and semantic knowledge of the language as it learns to improve in the target task. To the best of our knowledge, we are the first to find a similar phenomenon affecting diffusion models. Since our text prompt contains both the [identifier] and [class noun], when a diffusion model is fine-tuned on a small set of subject images, we observe that it slowly forgets how to generate subjects of the same class and progressively forgets the class-specific prior and can not generate different instances of the class in question. Figure 13 (middle) shows some sample generated images of "a dog" after fine-tuning the model on specific dog images. The results clearly show that the model loses the capability of generating generic dog images with naive fine-tuning.

Prior-Preservation Loss We propose an autogenous class-specific prior-preserving loss to counter both the overfitting and language drift issues. In essence, our method is to supervise the model with its *own generated samples*, in order for it to retain the prior once the few-shot fine-tuning begins. Specifically, we generate data $\mathbf{x}_{\text{pr}} = \hat{\mathbf{x}}(\mathbf{z}_{t_1}, \mathbf{c}_{\text{pr}})$ by using the ancestral sampler on the frozen pre-trained diffusion model with random initial noise $\mathbf{z}_{t_1} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and conditioning vector $\mathbf{c}_{\text{pr}} := \Gamma(f(\text{"a [class noun]"}))$. The loss becomes:

$$\mathbb{E}_{\mathbf{x},\mathbf{c},\boldsymbol{\epsilon},\boldsymbol{\epsilon}',t}\left[w_t \|\hat{\mathbf{x}}_\theta(\alpha_t \mathbf{x} + \sigma_t \boldsymbol{\epsilon}, \mathbf{c}) - \mathbf{x}\|_2^2 + \lambda w_{t'} \|\hat{\mathbf{x}}_\theta(\alpha_{t'} \mathbf{x}_{\text{pr}} + \sigma_{t'} \boldsymbol{\epsilon}', \mathbf{c}_{\text{pr}}) - \mathbf{x}_{\text{pr}}\|_2^2\right], \tag{2}$$

where λ controls for the relative weight of the prior-preservation term. Figure 4 illustrates the model fine-tuning with the class-generated samples and prior-preservation loss. Despite being simple, we find this prior-preservation loss is effective in overcoming the overfitting and language-drift issues. We find that ~ 200 epochs at learning rate 10^{-5} with $\lambda = 1$ is enough to achieve good results. During this process, $\sim 200 \times N$ "a [class noun]" samples are generated, with N being the size of the subject dataset, usually ranging from 3-5 images. The training process takes about 15 minutes on one TPUv4.

4.3 Personalized Instance-Specific Super-Resolution

While the text-to-image diffusion model controls for most visual semantics, the super-resolution (SR) models are essential to achieve photorealistic content and to preserve subject instance details. We find that if SR networks are used without fine-tuning, the generated output can contain artifacts since the SR models might not be familiar with certain details or textures of the subject instance, or the subject instance might have hallucinated incorrect features, or missing details. Figure 14 (bottom row) shows some sample output images with no fine-tuning of SR models, where the model hallucinates some high-frequency details. We find that fine-tuning the $64 \times 64 \rightarrow 256 \times 256$ SR model is essential for most subjects, and fine-tuning the $256 \times 256 \rightarrow 1024 \times 1024$ model can benefit some subject instances with high levels of fine-grained detail.

Low-level Noise Augmentation We find results to be suboptimal if the training recipes and test parameters of Saharia et al. [56] are used to fine-tune the SR models with the given few shots of a subject instance. Specifically, we find that maintaining the original level of noise augmentation used to train the SR networks leads to the blurring of high-frequency patterns of the subject and of the environment. See Figure 14 (middle row) for sample generations. In order to faithfully reproduce the subject instance, we reduce the level of noise augmentation from 10^{-3} to 10^{-5} during fine-tuning of the 256×256 SR model. With this small modification, We are able to recover fine-grained details of the subject instance.

5 Experiments

In this section, we show applications and experimental evaluation of our method. We find a large expanse of potential text-guided semantic modifications of our subject instance, including recontextualization, modification of subject properties such as color, or more complex properties such as combinations of animal species, original art renditions depicting the subject instance, as well as viewpoint and expression modification. Importantly, across all of these varied semantic modifications, some of which were previously insurmountable, we are able to **preserve unique visual features that give the subject its identity or essence**. If the task is recontextualization, then the subject features are unmodified, but appearance might change (e.g., pose changes). If the task has a stronger semantic modification, such as crosses between our subject and other species/objects, then key features of the subject are preserved after modification (e.g. Property Modification experiment below). In this section, we reference the subject's unique identifier using [V]. All experiments are conducted using images from Unsplash [*],

5.1 Applications

Recontextualization Given a customized model $\hat{\mathbf{x}}_\theta$, we can generate novel images for a specific subject instance by prompting the trained model with a sentence containing the unique identifier [V] and the class noun. Specifically, for recontextualization we generally form sentences in the following form: "a [V] [class noun] [context description]". For example, if the subject instance is a clock, we can prompt the model using "a [V] clock on top of snow" to generate our selected clock in a snowy scene. Another idea is to recontextualize our subject in famous landmarks, for example "a [V] clock with the Eiffel Tower in the background". We show examples of this capability in Figure 5. Importantly, as opposed to background modification, we are able to generate the subject in new poses and articulations and with new, previously unseen scene structure. Also, we note the detail in the integration of the subject in the scene, including realistic contact with other objects (e.g. partially planted in the snow, manipulated by human hands, etc.) and realistic shadows and reflections. This shows that our method not only allows for interpolation or recovery of subject details - but for extrapolation, answering questions in the form of: "what would happen if this subject were partially planted in the snow?", given no initial data of the subject in that context.

Art Renditions Given a prompt "a painting of a [V] [class noun] in the style of [famous painter]" or "a statue of a [V] [class noun] in the style of [famous sculptor]" we are able to generate original artistic renditions of our subject instance. In particular, this task does not equate to style transfer in which the semantics of a source scene are preserved and the style of another image is transferred onto the source scene. Instead, we are able to get meaningful changes in the scene, with subject instance details and identity preservation, depending on the artistic style. This means some of the images we generate have the subject in different poses that were not seen in the subject ground truth data and scenes that were also not in the ground truth data. We show examples of original artistic renditions of a personalized model in Figure 6. We use different famous artists to generate the original renditions. The artistic style of the famous artist is well represented in our images,

[*] `https://unsplash.com`

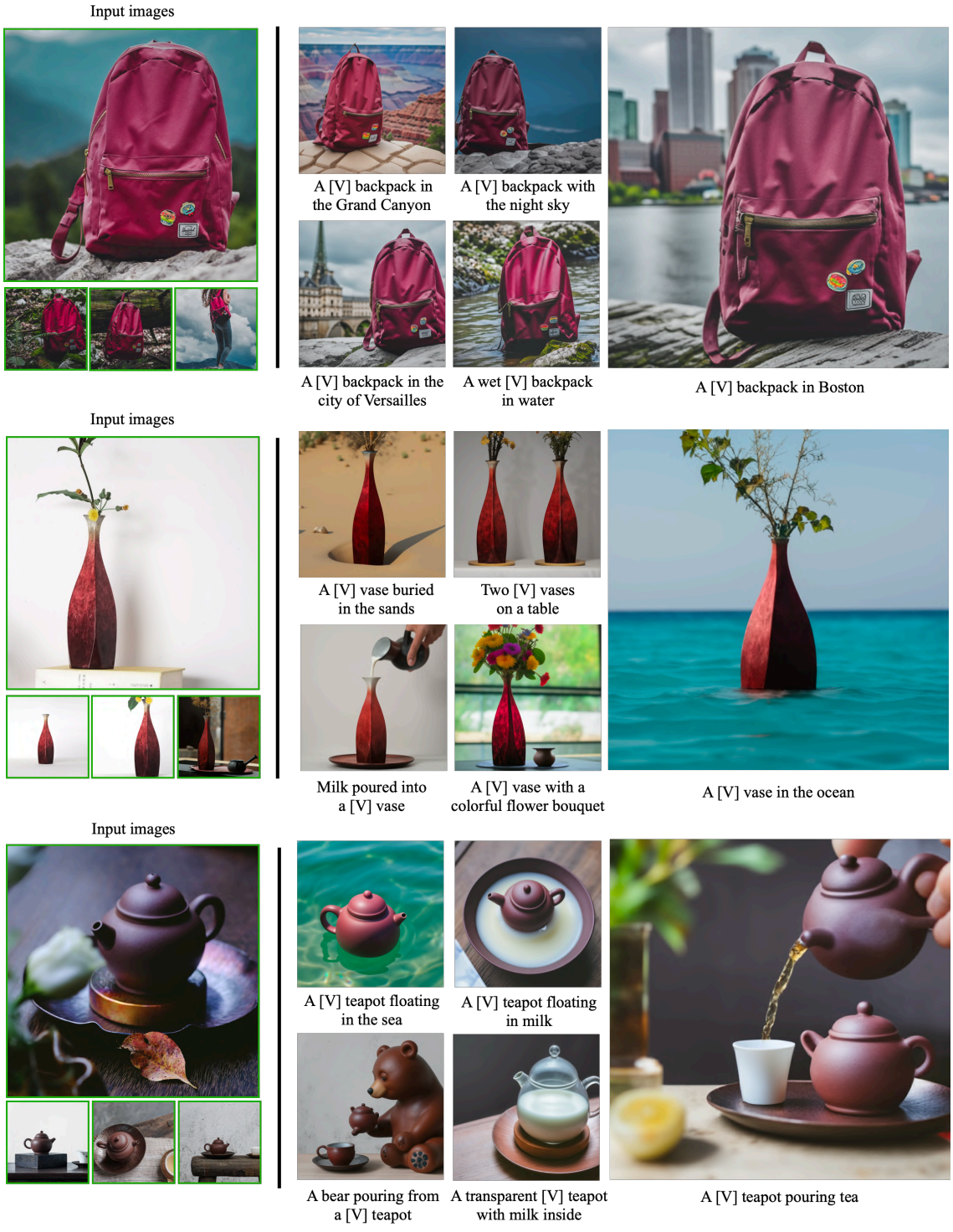

Figure 5: **Recontextualization of a backpack, vase, and teapot subject instances.** By fine-tuning a model with our approach, we are able to generate images of the subject instance in different environments, with high preservation of subject details and realistic interaction between the scene and the subject. We display the conditioning prompts below each image. Image credit (input images): Unsplash.

and there is a certain novelty to the renditions. Note that the dog was never seen in many of the poses shown, for example, there are no real images of this dog in the poses shown in the Michelangelo rendition.

Figure 6: **Artistic renderings of a dog instance in the style of famous painters**. We remark that many of the generated poses, e.g., the michelangelo renditions, were not seen in the training set. We also note that some renditions seem to have novel compositions and faithfully imitate the style of the painter. Image credit (input images): Unsplash.

Expression Manipulation Our method allows for new image generation of the subject with modified expressions that are not seen in the original set of subject images. We show examples in Figure 7. The range of expressiveness is high, ranging from negative to positive valence emotions and different levels of arousal. In all examples, the uniqueness of the subject dog is preserved - specifically, the asymmetric white streak on its face remains in all generated images.

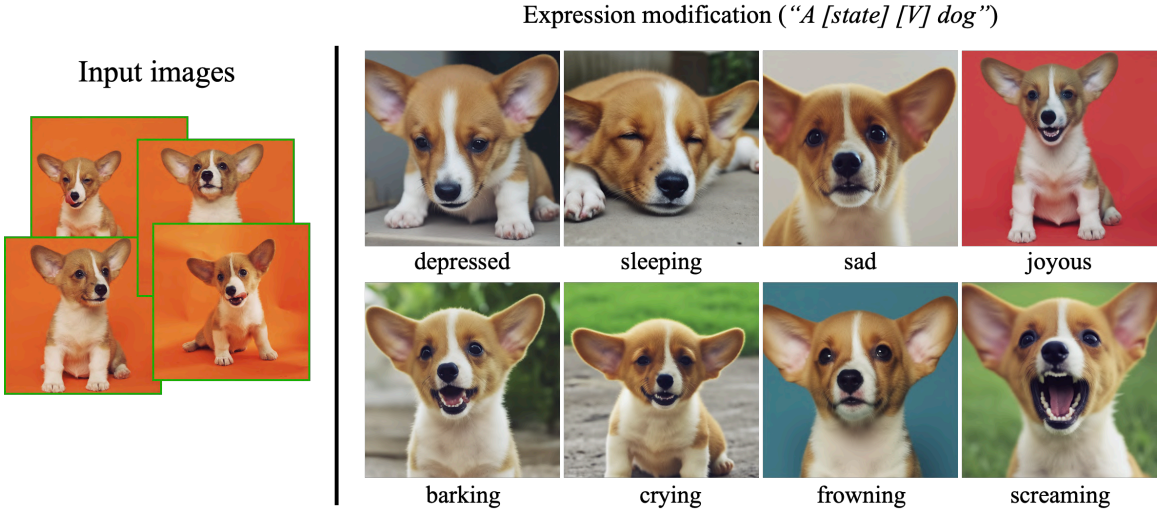

Figure 7: **Expression manipulation of a dog instance.** Our technique can synthesize various expressions that do not appear in the input images, demonstrating the extrapolation power of the model. Note the unique asymmetric white streak on the subject dog's face. Image credit (input images): Unsplash.

Novel View Synthesis We are able to render the subject under different novel viewpoints. In Figure 8 we show all real images of our subject cat - and we are able to generate new images of the same cat (with consistent complex fur patterns) under new camera viewpoints. We highlight that the model has not seen this

specific cat from behind, from below, or from above - yet it is able to extrapolate knowledge from the class prior to generating these novel views given only 4 frontal images of the subject.

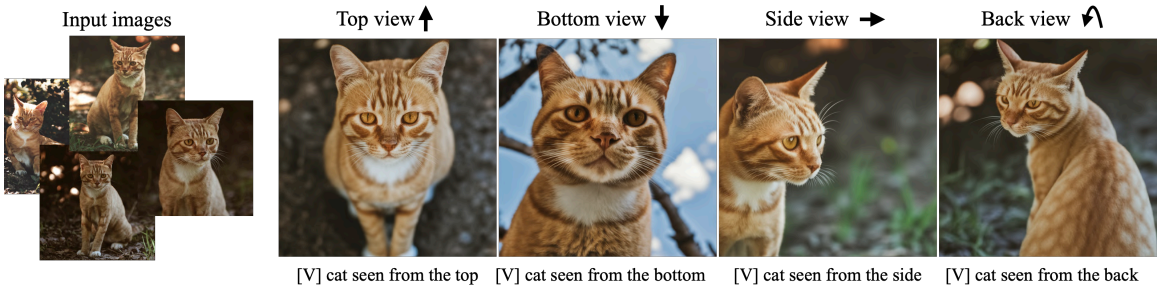

Figure 8: **Text-guided view synthesis**. Our technique can synthesize images with specified viewpoints for a subject cat (left to right: top, bottom, side, and back views). Note that the generated poses are different from the input poses, and the background changes in a realistic manner given a pose change. We also highlight the preservation of complex fur patterns on the subject cat's forehead. Image credit (input images): Unsplash.

Accessorization An interesting capability stemming from the strong compositional prior of the generation model is the ability to accessorize subjects. In Figure 9 we show examples of accessorization of a Chow Chow dog. We prompt the model with a sentence of the form: "a [V] [class noun] wearing [accessory]". In this manner, we are able to fit different accessories onto this dog - with aesthetically pleasing results. Note that the identity of the dog is preserved in all frames, and subject-accessory contact and articulation are realistic.

Figure 9: **Outfitting a dog with accessories**. The identity of the subject is preserved and many different outfits or accessories can be applied to the dog given a prompt of type "a [V] dog wearing a police/chef/witch outfit". We observe a realistic interaction between the subject dog and the outfits or accessories, as well as a large variety of possible options. Image credit (input images): Unsplash.

Property Modification We are able to modify subject instance properties. For example we can include a color adjective in the prompt sentence "a [color adjective] [V] [class noun]". In that way, we can generate novel instances of our subject with different colors. The generated scene can be very similar to the original scene, or it can be changed given a descriptive prompt. We show color changes of a car in the first row of Figure 10. We select similar viewpoints for effect, but we can generate different viewpoints of the car with different colors in different scenarios. This is a simple example of property modification, but more semantically complex property modifications can be achieved using our method. For example, we show crosses between a specific Chow Chow dog and different animal species in the bottom row of Figure 10. We prompt the model with sentences of the following structure: "a cross of a [V] dog and a [target species]". In particular, we can see in this example that the identity of the dog is well preserved even when the species changes - the face of the dog has certain individual properties that are well preserved and melded with the target species. Other property modifications are possible, such as material modification (e.g. a dog made out of stone). Some are harder than others and depend on the prior of the base generation model.

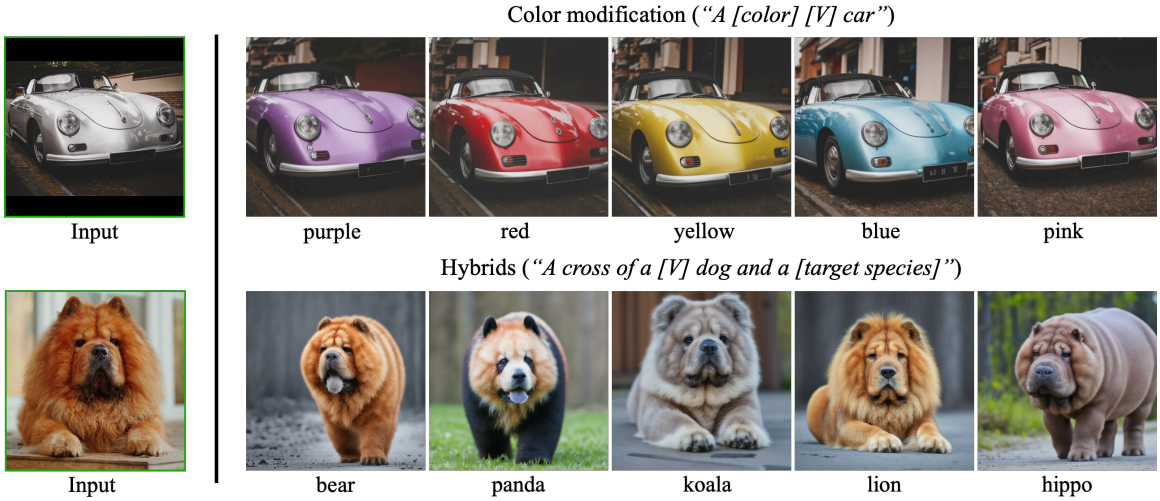

Figure 10: **Modification of subject properties while preserving their key features.** We show color modifications in the first row (using prompts "a [color] [V] car"), and crosses between a specific dog and different animals in the second row (using prompts "a cross of a [V] dog and a [target species]"). We highlight the fact that our method preserves unique visual features that give the subject its identity or essence, while performing the required property modification. Image credit (input images): Unsplash.

5.2 Ablation Studies

Class-Prior Ablation We show the results of using no class noun, a wrong class noun, and the correct class noun for text supervision of our subject images. Results are shown in Figure 11. We observe that the class prior of the wrong class (dog) remains entangled and we are not able to generate new images of our subject when the model is trained in this manner. Given a longer fitting time, we are able to disentangle the prior of the class noun "dog", but at the cost of losing the ability to generate dogs with the model and ultimately with decreased performance. If we are to train without a class noun, the model has difficulty learning the subject instance and does not easily entangle the class prior with the instance. The model takes longer to converge and can generate erroneous samples.

Prior Preservation Loss Ablation Our proposed loss acts as a powerful regularizer. A naively fine-tuned network can quickly overfit to the small set of subject images. In order to explore this phenomenon, we train two models for 200 epochs, one using naive fine-tuning (i.e. using the loss in Equation 1), and a network using our prior-preservation loss shown in Equation 2. We show results for different context captions in Figure 12. We observe that the regularization effect of our loss allows us to capture a wider range of poses for our subject dog without sacrificing subject fidelity. Importantly, we observe that using naive fine-tuning the dog usually lies on a fabric-type material similar to the training images, whereas this is avoided using our method.

Further, we evaluate how our prior preservation loss described in Section 4.2 conserves variability in the prior and show sample results in Figure 13. We verify that a vanilla model is able to generate a large variety of dogs, while a naively fine-tuned model on the subject dog exhibits language drift and generates our subject dog given the prompt "a dog". Our proposed loss preserves the variability of the prior and the model is able to generate new instances of our dog given a prompt of the style "a [V] dog" but also varied instances of dogs given a "a dog" prompt.

Super Resolution with Low-Noise Forward Diffusion We show how using lower noise to train the super-resolution models improves fidelity. Specifically, we show in Figure 14 that if the super-resolution models are not fine-tuned, we observe hallucination of high-frequency patterns on the subject which hurts identity preservation. Further, if we use the ground-truth noise augmentation level used for training the Imagen 256×256 model (10^{-3}), we obtain blurred and non-crisp details. If the noise used to train the SR model is reduced to 10^{-5}, then we conserve a large amount of detail without pattern hallucination or blurring.

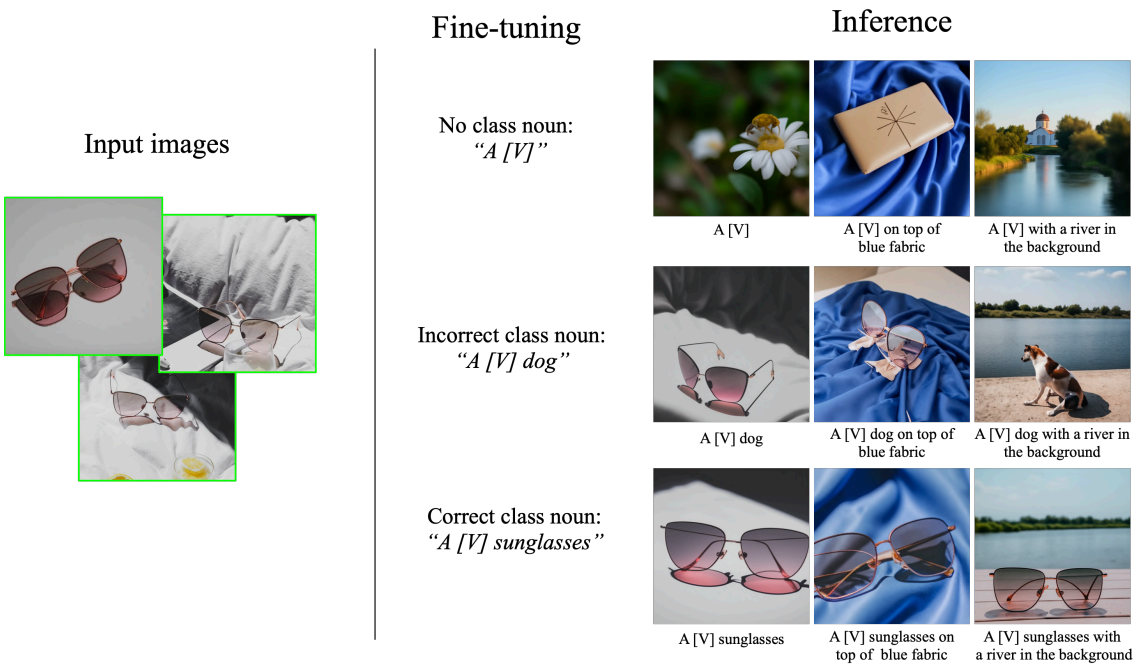

Figure 11: **Ablations with different class nouns in training prompts.** We show the results of fine-tuning the model with the correct class noun ([V] glasses), an incorrect class noun (a [V] dog), and no class noun (a [V]). Each row of generated images corresponds to a different model trained using the captions on its left. The captions under each image show the sentence used in inference. The bottom row shows that with the correct class noun for our subject (glasses), we are able to faithfully fit to the identity of our unique subject, take advantage of what the model knows about glasses, and generate our subject in different environments. The middle row shows that with an incorrect class noun (dog), the model runs into contention between our subject images and what it knows about dogs, and struggles to generate the subject glasses, especially when the contextual text prompt (e.g., "with a river in the background", in the bottom row) is more strongly associated with a dog. The top row shows that with no class noun, we have difficulty taking advantage of the class prior, and the model can generate erroneous samples. Image credit (input images): Unsplash.

5.3 Comparisons

We compare our results with the recent concurrent work of Gal et al. [20] in Figure 15. For this comparison, we train our model on the training images of two objects appear in the teaser of their work (headless sculpture and cat toy) kindly provided by Gal et al. [20], and apply the prompts suggested in their paper. For prompts where they present several generated images, we handpicked their best sample (with the highest image quality and morphological similarity to the subject). We find that our work can generate the same semantic variations of these unique objects, with a high emphasis on preserving the subject identity, as can be seen, for instance, by the detailed patterns of the cat sculpture that are preserved. In addition, we want to highlight that Gal et al. [20] approach is implemented on a different text-to-image model ([53]), hence the comparison might be unfair. That being said, we want to stress that the main technical difference between the approaches is that our approach embeds the subject within the output domain of the model, while their output domain is fixed and limited by expressiveness of the original model. Evidently, most of the results shown in their paper focus on synthesizing different attributes related to the subject, rather than synthesis of the subject itself with high-fidelity details reconstruction.

Next, we show comparisons of recontextualization of a subject clock, with distinctive features using our method and prompt engineering using vanilla Imagen [56] and the public API of DALL-E 2 [51]. After multiple iterations using both models, we settle for the base prompt "retro style yellow alarm clock with a white clock face and a yellow number three on the lower right part of the clock face" to describe all of the important features of the subject clock example. We find that while DALL-E 2 and vanilla Imagen are able to generate retro-style yellow alarm clocks, they struggle to represent a number 3 on the clock face, distinct from the clock face numbers. In general, we find that it is very hard to control fine-grained details of subject appearance, even with exhaustive prompt engineering. Also, we find that context can bleed into

Figure 12: **Avoiding overfitting with prior-preservation loss.** Naive fine-tuning can result in overfitting to both the context and the appearance of the subject in the input images (including pose, lighting, etc.). Our prior-preservation loss acts as a regularizer that alleviates overfitting, allowing pose variability and appearance diversity in a given context. It can be seen that the fabric-type material in the real subject images is conserved in the naively fine-tuning outputs (top row), and subject pose variability is limited. In contrast, with the prior-preservation loss, our results (bottom row) exhibit variation in the poses of the subject within non-entangled contexts (left to right: ocean, beach, river, desert). Image credit (input images): Unsplash.

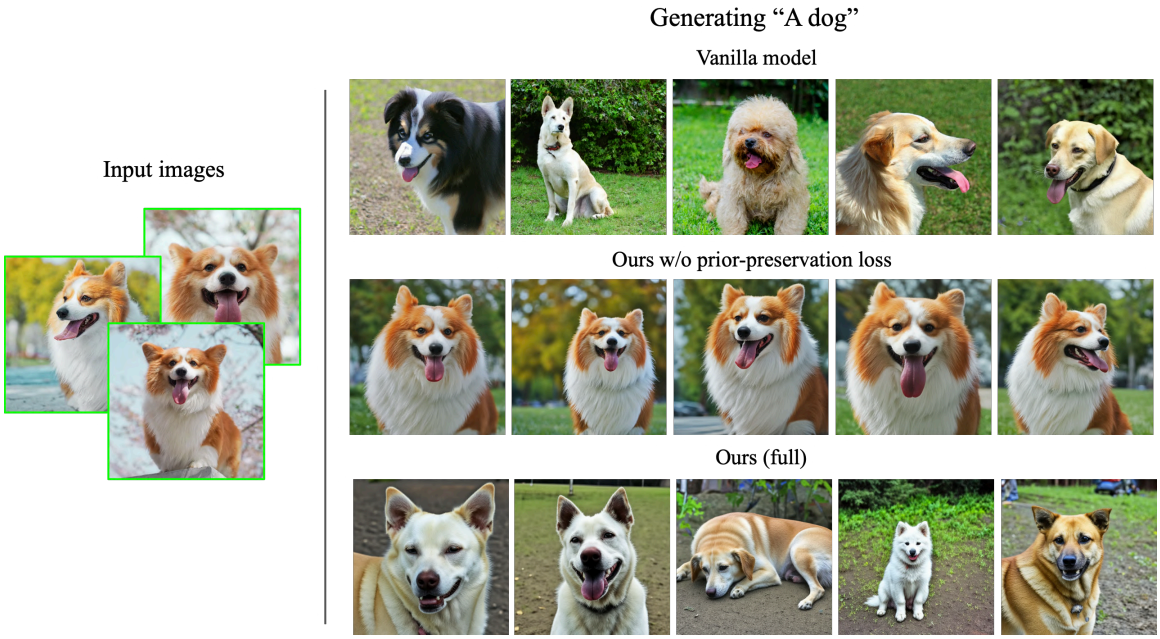

Figure 13: **Preservation of class semantic priors with prior-preservation loss.** Fine-tuning using images of our subject without prior-preservation loss results in language drift and the model loses the capability of generating other members of our subject's class. Using a prior-preservation loss term allows our model to avoid this and to preserve the subject class' prior. Image credit (input images): Unsplash.

the appearance of our subject instance. We show the results in Figure 16, and can observe that our method conserves fine-grained details of the subject instance such as the shape, the clock face font, and the large yellow number three on the clock face, among others.

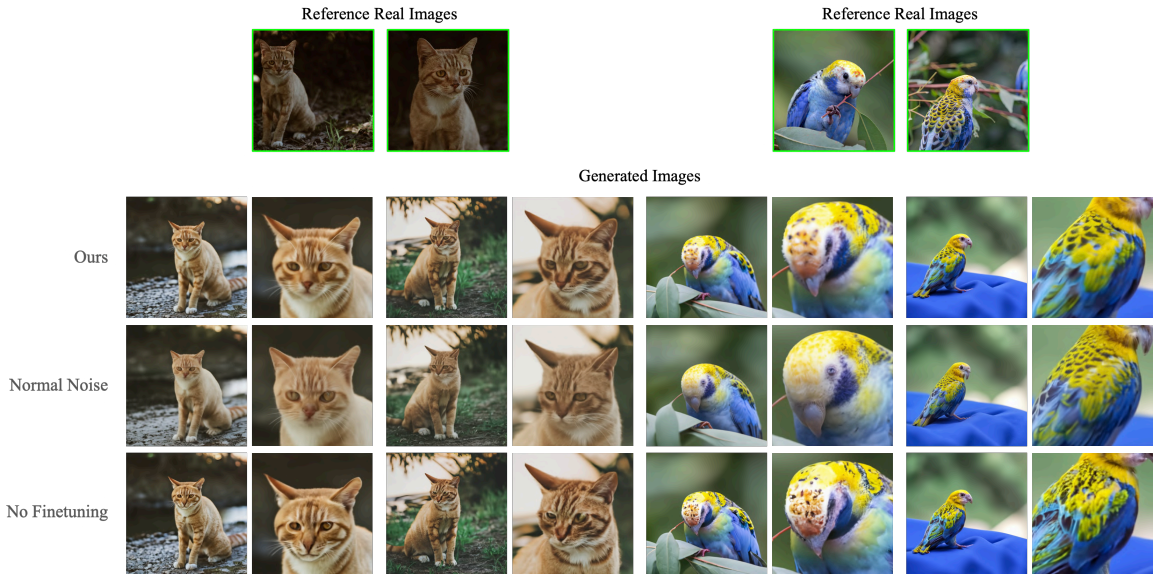

Figure 14: **Ablations with fine-tuning the super-resolution (SR) models.** Using the normal level of noise augmentation of [56] to train the SR models results in blurred high-frequency patterns, while no fine-tuning results in hallucinated high-frequency patterns. Using low-level noise augmentation for SR models improves sample quality and subject fidelity. Image credit (input images): Unsplash.

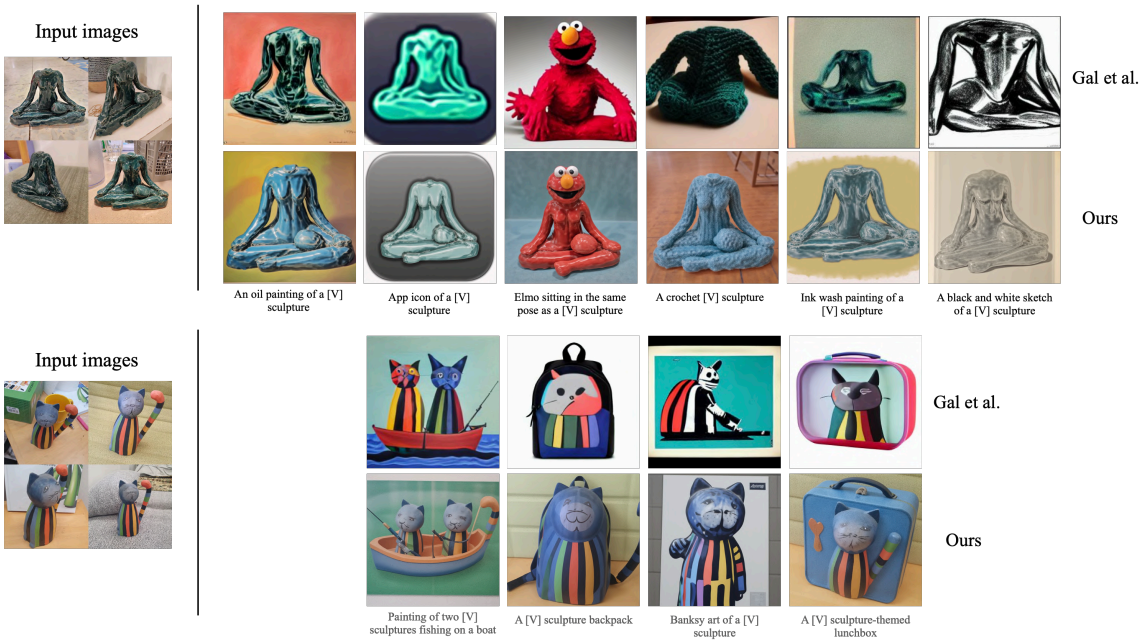

Figure 15: **Comparisons with Gal et al. [20]** using the subjects, images, and prompts from their work. Our approach is able to generate semantically correct variations of unique objects, exhibiting a higher degree of preservation of subject features. Input images provided by Gal et al. [20].

5.4 Limitations

Our method has several limitations, which we demonstrate in Figure 17, grouped into three main failure modes. The first is related to not being able to accurately generate the prompted context. For example, in Figure 17 we observe that when we prompt the model with "a [V] backpack in the ISS" and "a [V] backpack on the moon" it is not able to generate the desired contexts. Possible reasons are that the generative

Figure 16: **Comparison with DALL-E 2 and Imagen with detailed prompt engineering.** After several trial-and-error iterations, the base prompt used to generate DALL-E 2 and Imagen results was *"retro style yellow alarm clock with a white clock face and a yellow number three on the right part of the clock face"*, which is highly descriptive of the subject clock. In general, it is hard to control fine-grained details of subject appearance using prompts, even with large amounts of prompt engineering. Also, we can observe how context cues in the prompt can bleed into subject appearance (e.g. with a blue number 3 on the clock face when the context is "on top of blue fabric"). Image credit (input images): Unsplash.

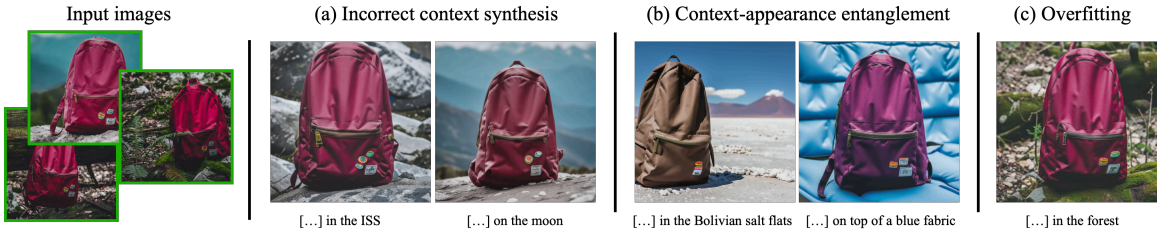

Figure 17: **Failure modes.** Given a rare prompted context the model might fail at generating the correct environment (a). It is possible for context and subject appearance to become entangled, with colors describing the context melding or changing the subject, or the model reverting to its prior with certain rare contexts (b). In this case, generating a brown bag in the rare context of the Bolivian salt flats. Finally, it is possible for the model to overfit and generate images similar to the training set, especially if prompts reflect the original environment of the training set (c). Image credit (input images): Unsplash.

model does not have a strong prior for these contexts, or that representing both the subject and the context together is a difficult task for the model. The second failure mode is context-appearance entanglement, where the appearance of the subject changes due to the prompted context. In Figure 17 we show examples of a backpack that changes colors due to the desired context being rare ("a [V] backpack in the Bolivian Salt Flats") or entangling the color of the context with that of the subject ("a [V] backpack on top of blue fabric"). Third, we also observe overfitting to the real images that happens when the prompt is similar to the original setting in which the subject was seen. An example is shown in Figure 17.

Other limitations observed are that some subjects are much easier to learn than others. For example, the model has a very strong prior for dogs and cats, with many learned variations. Occasionally, with subjects that are rarer or more complex, the model is unable to support as many subject variations. Finally, there is also variability in the fidelity of the subject and some generated images might contain hallucinated features on the subject, depending on the strength of the model prior, and the complexity of the semantic modification.

6 Conclusions

We presented an approach for synthesizing novel renditions of a given subject using a few images of the subject and the guidance of a text prompt. Our key idea is to embed a given subject instance in the output domain of a text-to-image diffusion model by binding the subject to a unique identifier. We achieve this by carefully fine-tuning a pretrained text-to-image model without "forgetting" other visual concepts it had learned during training. Remarkably—especially considering the typical sizes of text-to-image models (hundreds of millions of parameters)—this fine-tuning process can work given only 3-5 casually captured images of the subject, which makes the technique particularly accessible and easy to use from a user perspective. The resulting fine-tuned model is then able to reuse its learned knowledge of the visual world (how the subject typically looks like from different viewpoints, in different poses, how it interacts with the surrounding, etc.) while maintaining the subject's distinctive features. We demonstrated a variety of applications with animals and objects embedded within generated photorealistic scenes, in most cases indistinguishable from real images.

Societal Impact This project aims to provide users with an effective tool for synthesizing personal subjects (animals, objects) in different contexts. While general text-to-image models might be biased towards specific attributes when synthesizing images from text, our approach enables the user to get a better reconstruction of their desirable subjects. On contrary, malicious parties might try to use such images to mislead viewers. This is a common issue, existing in other generative models approaches or content manipulation techniques. Future research in generative modeling, and specifically of personalized generative priors, must continue investigating and revalidating these concerns.

7 Acknowledgement

We thank Rinon Gal, Adi Zicher, Ron Mokady, Bill Freeman, Dilip Krishnan, Huiwen Chang, and Daniel Cohen-Or for their valuable inputs that helped improve this work. We also thanks Mohammad Norouzi, Chitwan Saharia, and William Chan for providing us their pretrained Imagen models.

References

[1] Rameen Abdal, Peihao Zhu, John Femiani, Niloy J Mitra, and Peter Wonka. Clip2stylegan: Unsupervised extraction of stylegan edit directions. *arXiv preprint arXiv:2112.05219*, 2021.

[2] Justin Basilico Ashok Chandrashekar, Fernando Amat and Tony Jebara. Artwork personalization at netflix. `https://netflixtechblog.com/artwork-personalization-c589f074ad76`, 2021. Accessed: January 2022.

[3] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *arXiv preprint arXiv:2206.02779*, 2022.

[4] 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.

[5] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. *arXiv preprint arXiv:2204.02491*, 2022.

[6] Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5855–5864, October 2021.

[7] David Bau, Alex Andonian, Audrey Cui, YeonHwan Park, Ali Jahanian, Aude Oliva, and Antonio Torralba. Paint by word, 2021.

[8] Mark Boss, Andreas Engelhardt, Abhishek Kar, Yuanzhen Li, Deqing Sun, Jonathan T Barron, Hendrik Lensch, and Varun Jampani. Samurai: Shape and material from unconstrained real-world arbitrary image collections. *arXiv preprint arXiv:2205.15768*, 2022.

[9] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[10] 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] Julie Cattiau. A communication tool for people with speech impairments. `https://blog.google/outreach-initiatives/accessibility/project-relate/`, 2021. Accessed: January 2022.

[12] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*, 2021.

[13] Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. Dovenet: Deep image harmonization via domain verification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8394–8403, 2020.

[14] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. Vqgan-clip: Open domain image generation and editing with natural language guidance. *arXiv preprint arXiv:2204.08583*, 2022.

[15] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

[16] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34:19822–19835, 2021.

[17] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *arXiv preprint arXiv:2204.14217*, 2022.

[18] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.

[19] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. *arXiv preprint arXiv:2203.13131*, 2022.

[20] 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.

[21] Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *arXiv preprint arXiv:2108.00946*, 2021.

[22] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[23] 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.

[24] Tobias Hinz, Stefan Heinrich, and Stefan Wermter. Semantic object accuracy for generative text-to-image synthesis. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

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

[26] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.*, 23:47–1, 2022.

[27] Ajay Jain, Ben Mildenhall, Jonathan T. Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. 2022.

[28] Yihan Jiang, Jakub Konečnỳ, Keith Rush, and Sreeram Kannan. Improving federated learning personalization via model agnostic meta learning. *arXiv preprint arXiv:1909.12488*, 2019.

[29] 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.

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

[31] 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.

[32] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2435, 2022.

[33] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *EMNLP (Demonstration)*, 2018.

[34] Gihyun Kwon and Jong Chul Ye. Clipstyler: Image style transfer with a single text condition. *arXiv preprint arXiv:2112.00374*, 2021.

[35] Jason Lee, Kyunghyun Cho, and Douwe Kiela. Countering language drift via visual grounding. In *EMNLP*, 2019.

[36] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip Torr. Controllable text-to-image generation. *Advances in Neural Information Processing Systems*, 32, 2019.

[37] Wenbo Li, Pengchuan Zhang, Lei Zhang, Qiuyuan Huang, Xiaodong He, Siwei Lyu, and Jianfeng Gao. Object-driven text-to-image synthesis via adversarial training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12174–12182, 2019.

[38] Chen-Hsuan Lin, Ersin Yumer, Oliver Wang, Eli Shechtman, and Simon Lucey. St-gan: Spatial transformer generative adversarial networks for image compositing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9455–9464, 2018.

[39] Xihui Liu, Dong Huk Park, Samaneh Azadi, Gong Zhang, Arman Chopikyan, Yuxiao Hu, Humphrey Shi, Anna Rohrbach, and Trevor Darrell. More control for free! image synthesis with semantic diffusion guidance. 2021.

[40] Yuchen Lu, Soumye Singhal, Florian Strub, Aaron Courville, and Olivier Pietquin. Countering language drift with seeded iterated learning. In *International Conference on Machine Learning*, pages 6437–6447. PMLR, 2020.

[41] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, pages 405–421. Springer, 2020.

[42] Ron Mokady, Omer Tov, Michal Yarom, Oran Lang, Inbar Mosseri, Tali Dekel, Daniel Cohen-Or, and Michal Irani. Self-distilled stylegan: Towards generation from internet photos. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings*, pages 1–9, 2022.

[43] Alex 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. *arXiv preprint arXiv:2112.10741*, 2021.

[44] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021.

[45] 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. *arXiv preprint arXiv:2203.17272*, 2022.

[46] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. *arXiv preprint arXiv:2103.17249*, 2021.

[47] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. Learn, imagine and create: Text-to-image generation from prior knowledge. *Advances in neural information processing systems*, 32, 2019.

[48] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. Mirrorgan: Learning text-to-image generation by redescription. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1505–1514, 2019.

[49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[50] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020.

[51] 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.

[52] 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.

[53] 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.

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

[55] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022.

[56] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo-Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

[57] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *arXiv:2104.07636*, 2021.

[58] 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.

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

[60] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019.

[61] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020.

[62] Ming Tao, Hao Tang, Songsong Wu, Nicu Sebe, Xiao-Yuan Jing, Fei Wu, and Bingkun Bao. Df-gan: Deep fusion generative adversarial networks for text-to-image synthesis. *arXiv preprint arXiv:2008.05865*, 2020.

[63] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T. Barron, and Pratul P. Srinivasan. Ref-NeRF: Structured view-dependent appearance for neural radiance fields. *CVPR*, 2022.

[64] Huikai Wu, Shuai Zheng, Junge Zhang, and Kaiqi Huang. Gp-gan: Towards realistic high-resolution image blending. In *Proceedings of the 27th ACM international conference on multimedia*, pages 2487–2495, 2019.

[65] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2256–2265, 2021.

[66] 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.

[67] Zizhao Zhang, Yuanpu Xie, and Lin Yang. Photographic text-to-image synthesis with a hierarchically-nested adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6199–6208, 2018.