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] Kfir Aberman[1]
[1] Google Research [2] Boston University

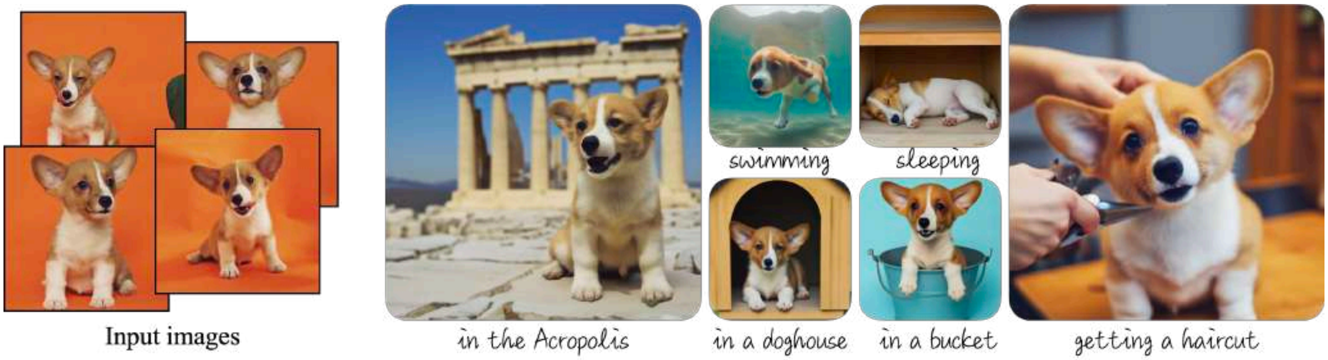


Figure 1. With just a few images (typically 3-5) of a subject (left), *DreamBooth*—our AI-powered photo booth—can generate a myriad of images of the subject in different contexts (right), using the guidance of a text prompt. The results exhibit natural interactions with the environment, as well as novel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

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. Given as input just a few images of a subject, we fine-tune a pretrained text-to-image 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 be used to synthesize 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, and artistic rendering, all while preserving the subject's key features. We also provide a new dataset and evaluation protocol for this new task of subject-driven generation. Project page: `https://dreambooth.github.io/`

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 (e.g., objects, animals) in new contexts such that they naturally and seamlessly blend into the scene.

Recently developed large text-to-image models have shown unprecedented capabilities, by enabling high-quality and diverse synthesis of images based on a text prompt written in natural language [54,61]. 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 the *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.

*This research was performed while Nataniel Ruiz was at Google.

Furthermore, even models whose text embedding lies in a shared language-vision space [52] 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 (∼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 a technique to represent a given subject with rare token identifiers and fine-tune a pre-trained, diffusion-based text-to-image framework.

We fine-tune the 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"). The latter enables the model to use its prior knowledge on the subject class while the class-specific instance is bound with the unique identifier. In order to prevent *language drift* [34, 40] that causes 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.

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. We highlight the contribution of each component in our method via ablation studies, and compare with alternative baselines and related work. We also conduct a user study to evaluate subject and prompt fidelity in our synthesized images, compared to alternative approaches.

To the best of our knowledge, ours is the first technique that tackles this new challenging problem of subject-driven generation, allowing users, from just a few casually captured images of a subject, synthesize novel renditions of the subject in different contexts while maintaining its distinctive features.

To evaluate this new task, we also construct a new dataset that contains various subjects captured in different contexts, and propose a new evaluation protocol that measures the subject fidelity and prompt fidelity of the generated results. We make our dataset and evaluation protocol publicly available on the project webpage.

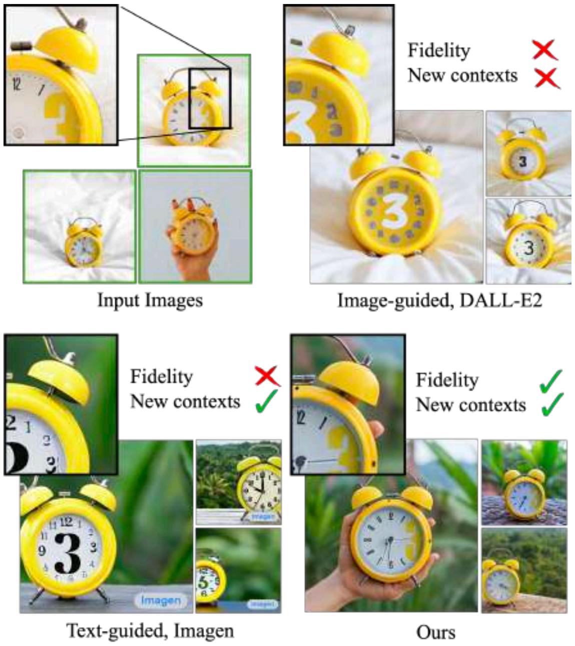


Figure 2. **Subject-driven generation.** Given a particular clock (left), it is hard to generate it while maintaining high fidelity to its key visual features (second and third columns showing DALL-E2 [54] image-guided generation and Imagen [61] text-guided generation; text prompt used for Imagen: *"retro 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"*). Our approach (right) can synthesize the clock with high fidelity and in new contexts (text prompt: *"a [V] clock in the jungle"*).

2. Related work

Image Composition. Image composition techniques [13, 38, 70] aim to clone a given subject into a new background such that the subject melds into the scene. To consider composition in novel poses, one may apply 3D reconstruction techniques [6, 8, 41, 49, 68] which usually works on rigid objects and require a larger number of views. Some drawbacks include scene integration (lighting, shadows, contact) and the inability to generate novel scenes. In contrast, our approach enable generation of subjects in novel poses and new contexts.

Text-to-Image Editing and Synthesis. Text-driven image manipulation has recently achieved significant progress using GANs [9, 22, 28–30] combined with image-text representations such as CLIP [52], yielding realistic manipulations using text [2, 7, 21, 43, 48, 71]. These methods work well on structured scenarios (e.g. human face editing) and can struggle over diverse datasets where subjects are varied. Crowson et al. [14] use VQ-GAN [18] and train over more diverse data to alleviate this concern. Other works [4, 31] exploit the recent diffusion models [25, 25, 45, 58, 60, 62–66], which achieve state-of-the-art generation quality over highly diverse datasets, often surpassing GANs [15]. While most works that require only text are limited to global editing [14, 33], 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 enables generating novel renditions of a given subject in new contexts.

There also exists work on text-to-image synthesis [14, 16, 19, 24, 27, 35, 36, 50, 51, 55, 58, 67, 74]. Recent large text-to-image models such as Imagen [61], DALL-E2 [54], Parti [72], CogView2 [17] and Stable Diffusion [58] demonstrated unprecedented semantic generation. These models do not provide fine-grained control over a generated image and use text guidance only. Specifically, it is challenging or impossible to preserve the identity of a subject consistently across synthesized images.

Controllable Generative Models. There are various approaches to control generative models, where some of them might prove to be viable directions for subject-driven prompt-guided image synthesis. Liu et al. [39] propose a diffusion-based technique allowing for image variations guided by reference image or text. To overcome subject modification, several works [3, 44] assume a user-provided mask to restrict the modified area. Inversion [12, 15, 54] can be used to preserve a subject while modifying context. Prompt-to-prompt [23] allows for local and global editing without an input mask. These methods fall short of identity-preserving novel sample generation of a subject.

In the context of GANs, Pivotal Tuning [57] allows for real image editing by finetuning the model with an inverted latent code anchor, and Nitzan et al. [46] extended this work to GAN finetuning on faces to train a personalized prior, which requires around 100 images and are limited to the face domain. Casanova et al. [11] propose an instance conditioned GAN that can generate variations of an instance, although it can struggle with unique subjects and does not preserve all subject details.

Finally, the concurrent work of Gal *et al.* [20] proposes a method to represent visual concepts, like an object or a style, through new tokens in the embedding space of a frozen text-to-image model, resulting in small personalized token embeddings. While this method is limited by the expressiveness of the frozen diffusion model, our fine-tuning approach enables us to embed the subject within the model's output domain, resulting in the generation of novel images of the subject which preserve its key visual features.

3. Method

Given only a few (typically 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. Example variations include changing the subject location, changing subject properties such as color or shape, modifying the subject's pose, viewpoint, and other semantic modifications. We do not impose any restrictions on input image capture settings and the subject image can have varying contexts. We next provide some background on text-to-image diffusion models (Sec. 3.1), then present our fine-tuning technique to bind a unique identifier with a subject described in a few images (Sec. 3.2), and finally propose a class-specific prior-preservation loss that enables us to overcome language drift in our fine-tuned model (Sec. 3.3).

3.1. 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, we are interested in a pre-trained text-to-image diffusion model $\hat{\mathbf{x}}_\theta$ that, given an initial noise map $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and a conditioning vector $\mathbf{c} = \Gamma(\mathbf{P})$ generated using a text encoder Γ and a text prompt $\mathbf{P}$, generates an image $\mathbf{x}_{\text{gen}} = \hat{\mathbf{x}}_\theta(\boldsymbol{\epsilon}, \mathbf{c})$. They are trained using a squared error loss to denoise a variably-noised image or latent code $\mathbf{z}_t := \alpha_t \mathbf{x} + \sigma_t \boldsymbol{\epsilon}$ as follows:

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

where $\mathbf{x}$ is the ground-truth image, $\mathbf{c}$ is a conditioning vector (e.g., obtained from a text prompt), 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 $t \sim \mathcal{U}([0, 1])$. A more detailed description is given in the supplementary material.

3.2. Personalization of Text-to-Image Models

Our first task is to implant the subject instance into the output domain of the model such that we can query the model for varied novel images of the subject. One natural idea is to fine-tune the model using the few-shot dataset of the subject. Careful care had to be taken when fine-tuning generative models such as GANs in a few-shot scenario as it can cause overfitting and mode-collapse - as well as not capturing the target distribution sufficiently well. There has been research on techniques to avoid these pitfalls [37, 42, 47, 56, 69], although, in contrast to our work, this line of work primarily seeks to generate images that resemble the target distribution but has no requirement of subject preservation. With regards to these pitfalls, we observe the peculiar finding that, given a careful fine-tuning setup

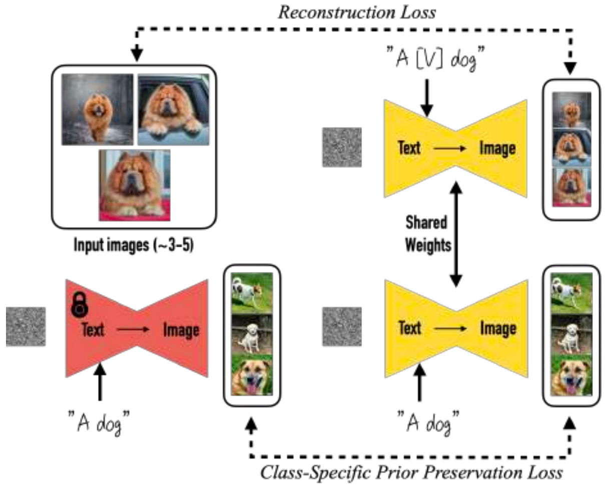


Figure 3. **Fine-tuning.** Given $\sim 3-5$ images of a subject we fine-tune a text-to-image diffusion model with the input images paired with a text prompt containing a unique identifier and the name of the class the subject belongs to (e.g., "A [V] dog"), in parallel, we apply a class-specific prior 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").

using the diffusion loss from Eq 1, large text-to-image diffusion models seem to excel at integrating new information into their domain without forgetting the prior or overfitting to a small set of training images.

Designing Prompts for Few-Shot Personalization Our goal is to "implant" a new (*unique identifier*, subject) pair into the diffusion model's "dictionary" . In order to bypass the overhead of writing detailed image descriptions for a given image set 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 provided by the user or obtained using a classifier. We use a class descriptor in the sentence in order to tether the prior of the class to our unique subject and find that using a wrong class descriptor, or no class descriptor increases training time and language drift while decreasing performance. In essence, we seek to leverage the model's prior of the specific class and entangle it with the embedding of our subject's unique identifier so we can leverage the visual prior to generate new poses and articulations of the subject in different contexts.

Rare-token Identifiers We generally find existing English words (e.g. "unique", "special") suboptimal since the model has to learn to disentangle them from their original meaning and to re-entangle them to reference our subject. 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. We often find that these tokens incur the similar weaknesses as using common English words. Our approach is to find rare tokens in the vocabulary, and then invert these tokens into text space, in order to minimize the probability of the identifier having a strong prior. We 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 $f(\hat{\mathbf{V}})$. The sequence can be of variable length k, and 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 $\hat{\mathbf{V}}$. For Imagen, we find that using uniform random sampling 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.

3.3. Class-specific Prior Preservation Loss

In our experience, the best results for maximum subject fidelity are achieved by fine-tuning all layers of the model. This includes fine-tuning layers that are conditioned on the text embeddings, which gives rise to the problem of *language drift*. Language drift has been an observed problem in language models [34, 40], where a 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. To the best of our knowledge, we are the first to find a similar phenomenon affecting diffusion models, where to model slowly forgets how to generate subjects of the same class as the target subject.

Another problem is the possibility of *reduced output diversity*. Text-to-image diffusion models naturally posses high amounts of output diversity. When fine-tuning on a small set of images we would like to be able to generate the subject in novel viewpoints, poses and articulations. Yet, there is a risk of reducing the amount of variability in the output poses and views of the subject (e.g. snapping to the few-shot views). We observe that this is often the case, especially when the model is trained for too long.

To mitigate the two aforementioned issues, we propose an autogenous class-specific prior preservation loss that encourages diversity and counters language drift. 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. This allows it to generate diverse images of the class prior, as well as retain knowledge about the class prior that it can use in conjunction with knowledge about the subject instance. 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:

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

where the second term is the prior-preservation term that supervises the model with its own generated images, and λ controls for the relative weight of this term. Figure 3 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 encouraging output diversity and in overcoming language-drift. We also find that we can train the model for more iterations without risking overfitting. We find that $\sim$ 1000 iterations with $\lambda = 1$ and learning rate 10^{-5} for Imagen [61] and 5×10^{-6} for Stable Diffusion [59], and with a subject dataset size of 3-5 images is enough to achieve good results. During this process, $\sim$ 1000 "a [class noun]" samples are generated - but less can be used. The training process takes about 5 minutes on one TPUv4 for Imagen, and 5 minutes on a NVIDIA A100 for Stable Diffusion.

4. Experiments

In this section, we show experiments and applications. Our method enables a large expanse of text-guided semantic modifications of our subject instances, including recontextualization, modification of subject properties such as material and species, art rendition, and viewpoint modification. Importantly, across all of these modifications, we are able to **preserve the unique visual features that give the subject its identity and essence**. If the task is recontextualization, then the subject features are unmodified, but appearance (e.g., pose) may change. If the task is a stronger semantic modification, such as crossing between our subject and another species/object, then the key features of the subject are preserved after modification. In this section, we reference the subject's unique identifier using [V]. We include specific Imagen and Stable Diffusion implementation details in the supp. material.

4.1. Dataset and Evaluation

Dataset We collected a dataset of 30 subjects, including unique objects and pets such as backpacks, stuffed animals, dogs, cats, sunglasses, cartoons, etc. We separate each subject into two categories: objects and live subjects/pets. 21 of the 30 subjects are objects, and 9 are live subjects/pets. We provide one sample image for each of the subjects in Figure 5. Images for this dataset were collected by the authors or sourced from Unsplash [1]. We also collected 25 prompts: 20 recontextualization prompts and 5 property modification prompts for objects; 10 recontextualization, 10

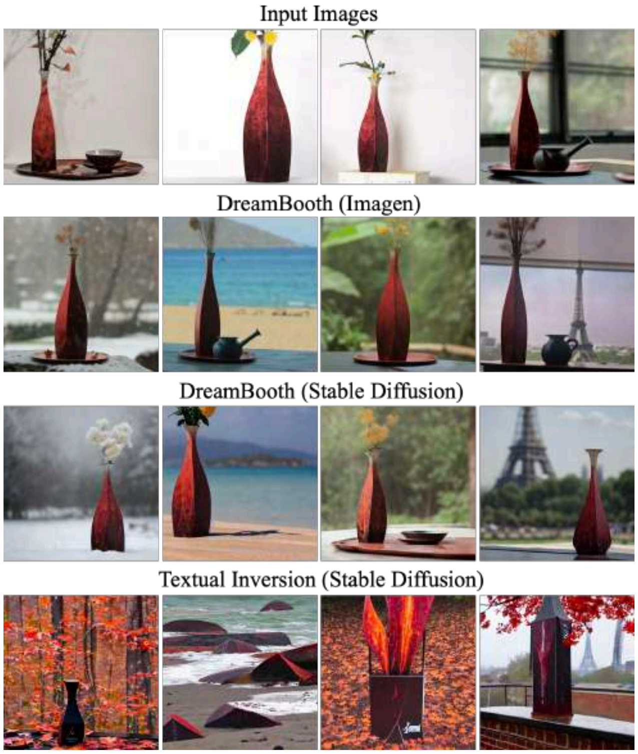

Input Images
DreamBooth (Imagen)
DreamBooth (Stable Diffusion)
Textual Inversion (Stable Diffusion)

Figure 4. **Comparisons with Textual Inversion [20]** Given 4 input images (top row), we compare: DreamBooth Imagen (2nd row), DreamBooth Stable Diffusion (3rd row), Textual Inversion (bottom row). Output images were created with the following prompts (left to right): "a [V] vase in the snow", "a [V] vase on the beach", "a [V] vase in the jungle", "a [V] vase with the Eiffel Tower in the background". DreamBooth is stronger in both subject and prompt fidelity.

Method	DINO ↑	CLIP-I ↑	CLIP-T ↑
Real Images	0.774	0.885	N/A
DreamBooth (Imagen)	**0.696**	**0.812**	**0.306**
DreamBooth (Stable Diffusion)	0.668	0.803	0.305
Textual Inversion (Stable Diffusion)	0.569	0.780	0.255

Table 1. Subject fidelity (DINO, CLIP-I) and prompt fidelity (CLIP-T, CLIP-T-L) quantitative metric comparison.

Method	Subject Fidelity ↑	Prompt Fidelity ↑
DreamBooth (Stable Diffusion)	**68%**	**81%**
Textual Inversion (Stable Diffusion)	22%	12%
Undecided	10%	7%

Table 2. Subject fidelity and prompt fidelity user preference.

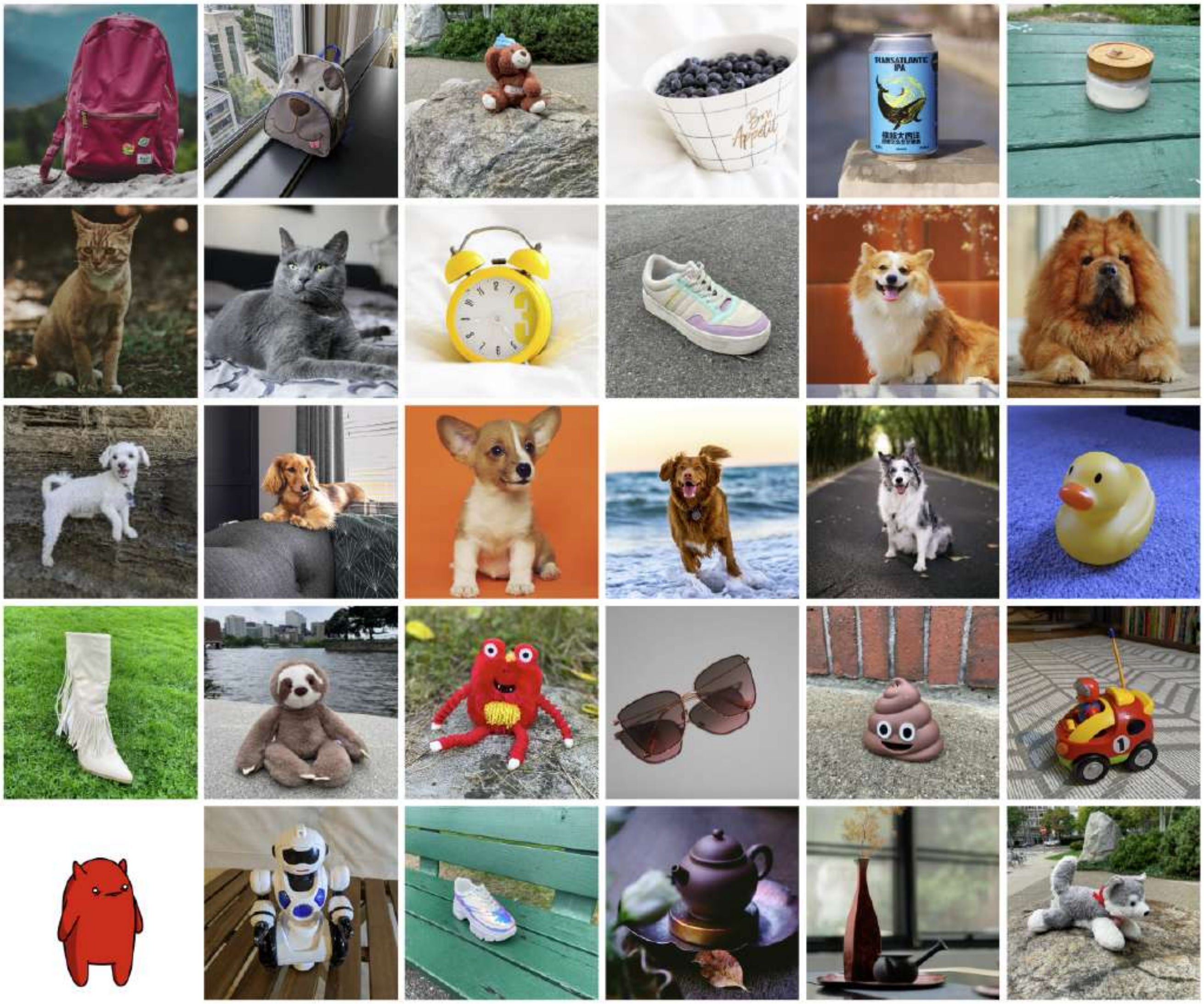


Figure 5. **Dataset**. Example images for each subject in our proposed dataset.

accessorization, and 5 property modification prompts for live subjects/pets. The full list of prompts can be found in the supplementary material.

For the evaluation suite we generate four images per subject and per prompt, totaling 3,000 images. This allows us to robustly measure performances and generalization capabilities of a method. We make our dataset and evaluation protocol publicly available on the project webpage for future use in evaluating subject-driven generation.

Evaluation Metrics One important aspect to evaluate is subject fidelity: the preservation of subject details in generated images. For this, we compute two metrics: CLIP-I and DINO [10]. CLIP-I is the average pairwise cosine similarity between CLIP [52] embeddings of generated and real images. Although this metric has been used in other work [20], it is not constructed to distinguish between different subjects that could have highly similar text descriptions (e.g. two different yellow clocks). Our proposed DINO metric is the average pairwise cosine similarity between the ViT-S/16 DINO embeddings of generated and real images. This is our preferred metric, since, by construction and in contrast to supervised networks, DINO is not trained to ignore differences between subjects of the same class. Instead, the self-supervised training objective encourages distinction of unique features of a subject or image. The second important aspect to evaluate is prompt fidelity, measured as the average cosine similarity between prompt and image CLIP embeddings. We denote this as CLIP-T.

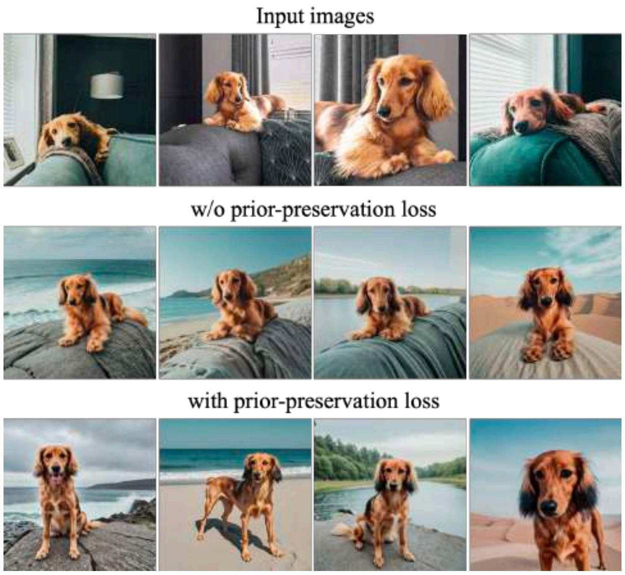


Figure 6. **Encouraging diversity with prior-preservation loss.** Naive fine-tuning can result in overfitting to input image context and subject appearance (e.g. pose). PPL acts as a regularizer that alleviates overfitting and encourages diversity, allowing for more pose variability and appearance diversity.

4.2. Comparisons

We compare our results with Textual Inversion, the recent concurrent work of Gal et al. [20], using the hyperparameters provided in their work. We find that this work is the only comparable work in the literature that is subject-driven, text-guided and generates novel images. We generate images for DreamBooth using Imagen, DreamBooth using Stable Diffusion and Textual Inversion using Stable Diffusion. We compute DINO and CLIP-I subject fidelity metrics and the CLIP-T prompt fidelity metric. In Table 1 we show sizeable gaps in both subject and prompt fidelity metrics for DreamBooth over Textual Inversion. We find that DreamBooth (Imagen) achieves higher scores for both subject and prompt fidelity than DreamBooth (Stable Diffusion), approaching the upper-bound of subject fidelity for real images. We believe that this is due to the larger expressive power and higher output quality of Imagen.

Further, we compare Textual Inversion (Stable Diffusion) and DreamBooth (Stable Diffusion) by conducting a user study. For subject fidelity, we asked 72 users to answer questionnaires of 25 comparative questions (3 users per questionnaire), totaling 1800 answers. Samples are randomly selected from a large pool. Each question shows the set of real images for a subject, and one generated image of that subject by each method (with a random prompt). Users are asked to answer the question: “Which of the two images best reproduces the identity (e.g. item type and details) of the reference item?”, and we include a “Cannot Determine / Both Equally” option. Similarly for prompt fidelity, we ask “Which of the two images is best described by the reference text?”. We average results using majority voting and present them in Table 2. We find an overwhelming preference for DreamBooth for both subject fidelity and prompt fidelity. This shines a light on results in Table 1, where DINO differences of around 0.1 and CLIP-T differences of 0.05 are significant in terms of user preference. Finally, we show qualitative comparisons in Figure 4. We observe that DreamBooth better preserves subject identity, and is more faithful to prompts. We show samples of the user study in the supp. material.

Method	PRES ↓	DIV ↑	DINO ↑	CLIP-I ↑	CLIP-T ↑
DreamBooth (Imagen) w/ PPL	**0.493**	**0.391**	0.684	0.815	**0.308**
DreamBooth (Imagen)	0.664	0.371	**0.712**	**0.828**	0.306

Table 3. Prior preservation loss (PPL) ablation displaying a prior preservation (PRES) metric, diversity metric (DIV) and subject and prompt fidelity metrics.

Method	DINO ↑	CLIP-I ↑
Correct Class	**0.744**	**0.853**
No Class	0.303	0.607
Wrong Class	0.454	0.728

Table 4. Class name ablation with subject fidelity metrics.

4.3. Ablation Studies

Prior Preservation Loss Ablation We fine-tune Imagen on 15 subjects from our dataset, with and without our proposed prior preservation loss (PPL). The prior preservation loss seeks to combat language drift and preserve the prior. We compute a prior preservation metric (PRES) by computing the average pairwise DINO embeddings between generated images of random subjects of the prior class and real images of our specific subject. The higher this metric, the more similar random subjects of the class are to our specific subject, indicating collapse of the prior. We report results in Table 3 and observe that PPL substantially counteracts language drift and helps retain the ability to generate diverse images of the prior class. Additionally, we compute a diversity metric (DIV) using the average LPIPS [73] cosine similarity between generated images of same subject with same prompt. We observe that our model trained with PPL achieves higher diversity (with slightly diminished subject fidelity), which can also be observed qualitatively in Figure 6, where our model trained with PPL overfits less to the environment of the reference images and can generate the dog in more diverse poses and articulations.

Class-Prior Ablation We finetune Imagen on a subset of our dataset subjects (5 subjects) with no class noun, a randomly sampled incorrect class noun, and the correct class noun. With the correct class noun for our subject, we are able to faithfully fit to the subject, take advantage of the

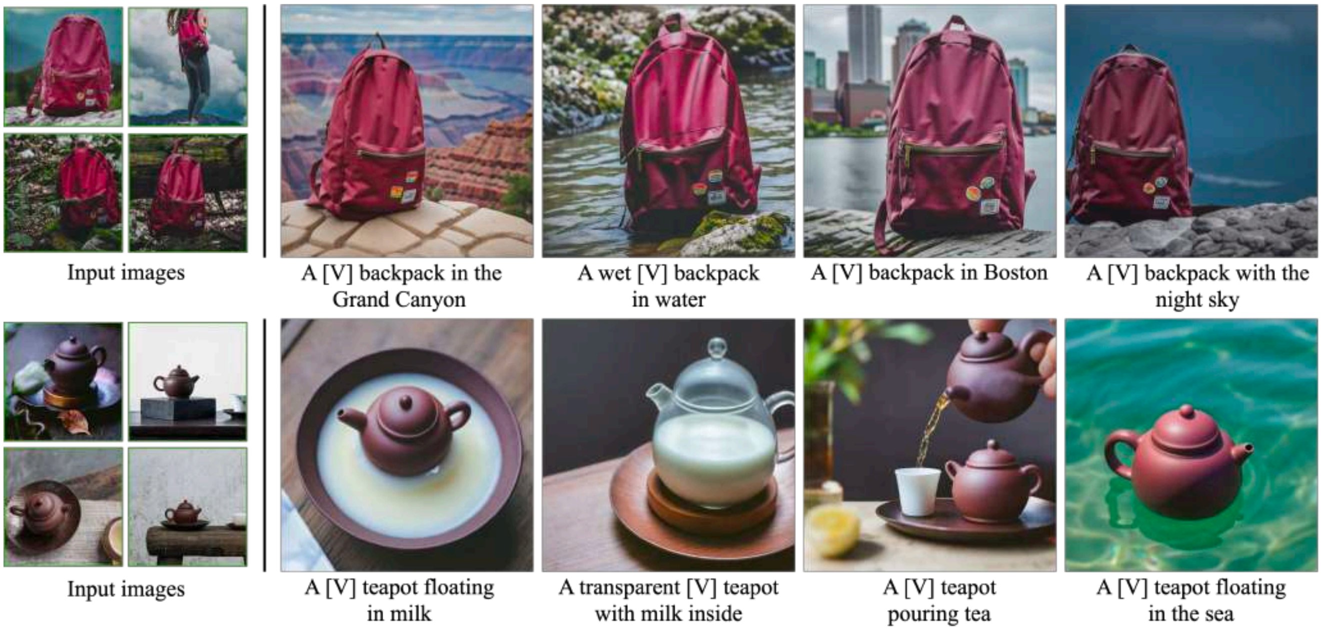

Input images
A [V] backpack in the Grand Canyon
A wet [V] backpack in water
A [V] backpack in Boston
A [V] backpack with the night sky
Input images
A [V] teapot floating in milk
A transparent [V] teapot with milk inside
A [V] teapot pouring tea
A [V] teapot floating in the sea

Figure 7. **Recontextualization.** We generate images of the subjects in different environments, with high preservation of subject details and realistic scene-subject interactions. We show the prompts below each image.

class prior, allowing us to generate our subject in various contexts. When an incorrect class noun (e.g. "can" for a backpack) is used, we run into contention between our subject and and the class prior - sometimes obtaining cylindrical backpacks, or otherwise misshapen subjects. If we train with no class noun, the model does not leverage the class prior, has difficulty learning the subject and converging, and can generate erroneous samples. Subject fidelity results are shown in Table 4, with substantially higher subject fidelity for our proposed approach.

4.4. Applications

Recontextualization We can generate novel images for a specific subject in different contexts (Figure 7) with descriptive prompts ("a [V] [class noun] [context description]"). Importantly, we are able to generate the subject in new poses and articulations, with previously unseen scene structure and realistic integration of the subject in the scene (e.g. contact, shadows, reflections).

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 artistic renditions of our subject. Unlike style transfer, where the source structure is preserved and only the style is transferred, we are able to generate meaningful, novel variations depending on the artistic style, while preserving subject identity. E.g, as shown in Figure 8, "Michelangelo", we generated a pose that is novel and not seen in the input images.

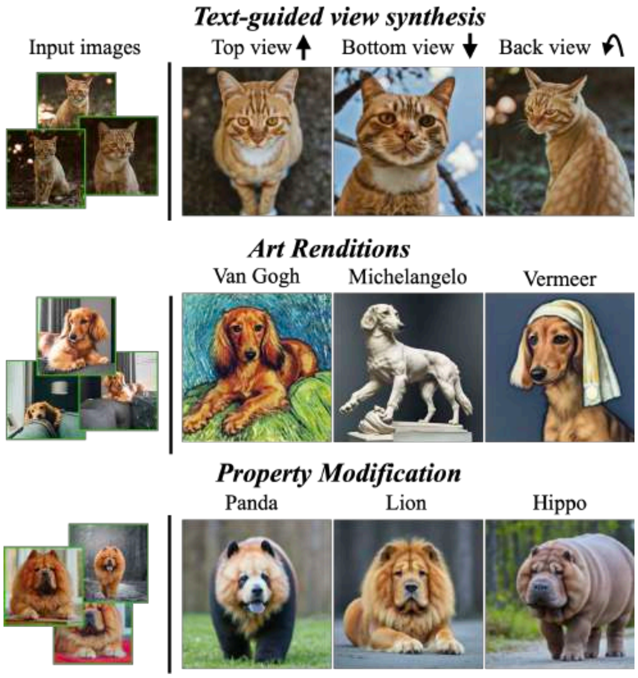


Figure 8. **Novel view synthesis, art renditions, and property modifications**. We are able to generate novel and meaningful images while faithfully preserving subject identity and essence. More applications and examples in the supplementary material.

Novel View Synthesis We are able to render the subject under novel viewpoints. In Figure 8, we generate new im-

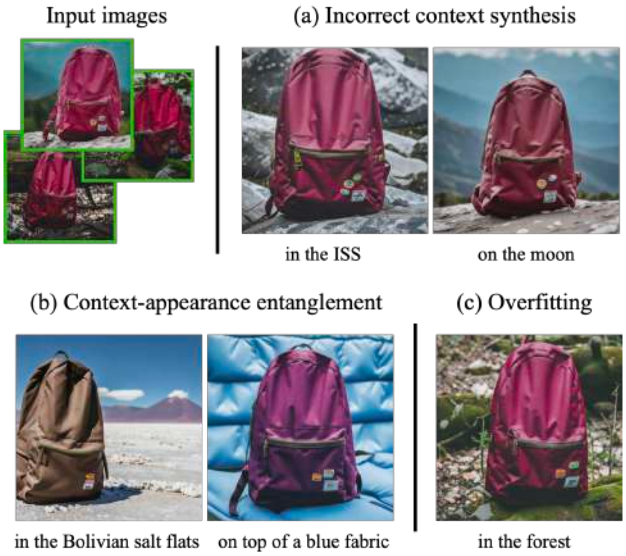

Input images
(a) Incorrect context synthesis
in the ISS
on the moon
(b) Context-appearance entanglement
(c) Overfitting
in the Bolivian salt flats
on top of a blue fabric
in the forest

Figure 9. **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 (b). 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).

ages of the input cat (with consistent complex fur patterns) under new viewpoints. We highlight that the model has not seen this specific cat from behind, below, or above - yet it is able to extrapolate knowledge from the class prior to generate these novel views given only 4 frontal images of the subject.

Property Modification We are able to modify subject properties. For example, we show crosses between a specific Chow Chow dog and different animal species in the bottom row of Figure 8. 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 unique features that are well preserved and melded with the target species. Other property modifications are possible, such as material modification (e.g. “a transparent [V] teapot” in Figure 7). Some are harder than others and depend on the prior of the base generation model.

4.5. Limitations

We illustrate some failure models of our method in Figure 9. The first is related to not being able to accurately generate the prompted context. Possible reasons are a weak prior for these contexts, or difficulty in generating both the subject and specified concept together due to low probability of co-occurrence in the training set. The second is context-appearance entanglement, where the appearance of the subject changes due to the prompted context, exemplified in Figure 9 with color changes of the backpack. Third, we also observe overfitting to the real images that happen when the prompt is similar to the original setting in which the subject was seen.

Other limitations are that some subjects are easier to learn than others (e.g. dogs and cats). Occasionally, with subjects that are rarer, 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 subject features, depending on the strength of the model prior, and the complexity of the semantic modification.

5. Conclusions

We presented an approach for synthesizing novel renditions of a 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. Remarkably - this fine-tuning process can work given only 3-5 subject images, making the technique particularly accessible. We demonstrated a variety of applications with animals and objects in generated photorealistic scenes, in most cases indistinguishable from real images.

6. 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, and to Mohammad Norouzi, Chitwan Saharia and William Chan for providing us with their support and the pretrained Imagen models. Finally, a special thanks to David Salesin for his feedback, advice and for his support for the project.

References

[1] Unsplash. `https://unsplash.com/`.

[2] 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.

[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] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.

[11] 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.

[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] 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.

[29] 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.

[30] 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.

[31] 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.

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

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

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

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

[36] 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.

[37] Yijun Li, Richard Zhang, Jingwan Lu, and Eli Shechtman. Few-shot image generation with elastic weight consolidation. *arXiv preprint arXiv:2012.02780*, 2020.

[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] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Freeze the discriminator: a simple baseline for fine-tuning gans. In *CVPR AI for Content Creation Workshop*, 2020.

[43] 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.

[44] 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.

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

[46] 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.

[47] Utkarsh Ojha, Yijun Li, Jingwan Lu, Alexei A Efros, Yong Jae Lee, Eli Shechtman, and Richard Zhang. Few-shot image generation via cross-domain correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10743–10752, 2021.

[48] 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.

[49] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] Esther Robb, Wen-Sheng Chu, Abhishek Kumar, and Jia-Bin Huang. Few-shot adaptation of generative adversarial networks. *ArXiv*, abs/2010.11943, 2020.

[57] Daniel Roich, Ron Mokady, Amit H. Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Transactions on Graphics (TOG)*, 2022.

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

[59] 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.

[60] 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.

[61] 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.

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

[63] 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.

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

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

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

[67] 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.

[68] 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.

[69] Yaxing Wang, Abel Gonzalez-Garcia, David Berga, Luis Herranz, Fahad Shahbaz Khan, and Joost van de Weijer. Minegan: Effective knowledge transfer from gans to target domains with few images. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[70] 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.

[71] 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.

[72] 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.

[73] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[74] 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.

Supplementary Material

Background

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 $\mathbf{z}_t := \alpha_t \mathbf{x} + \sigma_t \boldsymbol{\epsilon}$ as follows:

$$\mathbb{E}_{\mathbf{x},\mathbf{c},\boldsymbol{\epsilon},t}\big[w_t \|\hat{\mathbf{x}}_\theta(\alpha_t \mathbf{x} + \sigma_t \boldsymbol{\epsilon}, \mathbf{c}) - \mathbf{x}\|_2^2\big] \tag{3}$$

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 [64] 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, $\hat{\mathbf{x}}_0^t := \hat{\mathbf{x}}_\theta(\mathbf{z}_t, \mathbf{c})$, 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 [54, 61]. Specifically, [61] 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. [54] use a similar configuration, with unconditional SR models. A key component of high-quality sample generations from [61] 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. [61] select Gaussian noise as the form of augmentation.

Other recent state-of-the-art text-to-image diffusion models, such as Stable Diffusion [59], use a single diffusion model to generate high-resolution images. Specifically, the forward and backward diffusion processes occur in a lower-dimensional latent space and an encoder-decoder architecture is trained on a large image dataset to translate images into latent codes. At inference time, a random noise latent code goes through the backward diffusion process and the pre-trained decoder is used to generate the final image. Our method can be naturally applied to this scenario, where the U-Net (and possibly the text encoder) are trained, and the decoder is fixed.

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. [54] use CLIP text embeddings that are translated into image embeddings using a learned prior, while Saharia et al. [61] use a pre-trained T5-XXL language model [53]. 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 [61], we use the SentencePiece tokenizer [32]. 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 $\mathbf{c}$.

Dataset

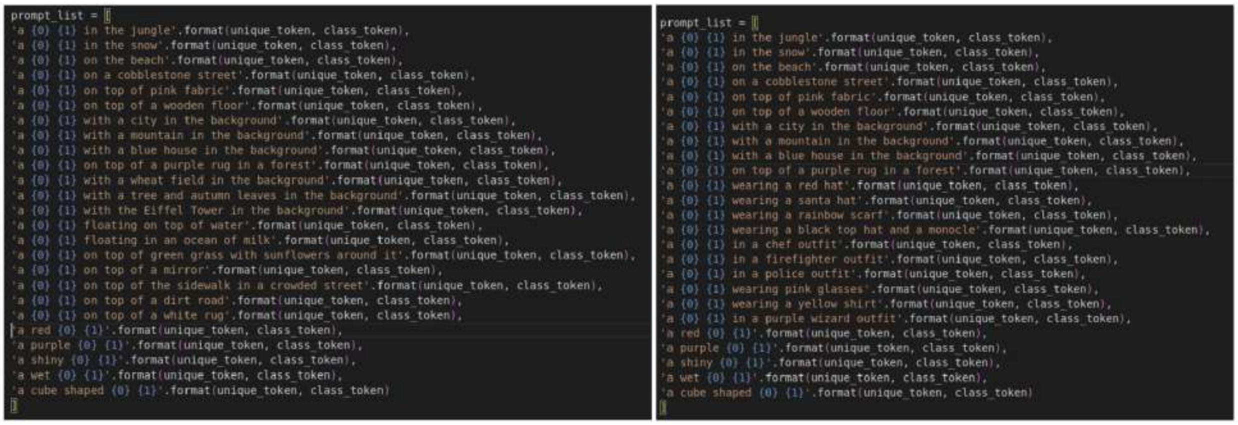
Our dataset includes 30 subjects. We separate each subject into two categories: objects and live subjects/pets. 21 of the 30 subjects are objects, and 9 are live subjects/pets. We provide one sample image for each of the subjects in Figure 5. Images for this dataset were collected by the authors or sourced from Unsplash [1].

We also collected 25 prompts: 20 recontextualization prompts and 5 property modification prompts for objects. 10 recontextualization, 10 accessorization, and 5 property modification prompts for live subjects/pets. Prompts are shown in Figure 10

For the evaluation suite we generate four images per subject and per prompt, totaling 3,000 images. This allows us to robustly measure performances and generalization capabilities of a method. We make our dataset and evaluation protocol publicly available on the project webpage for future use in evaluating subject-driven generation.

Subject Fidelity Metrics

In the main paper we comment on the superiority of our proposed DINO metric in terms of subject fidelity. We hypothesize that this is because DINO is, in essence, trained in a self-supervised manner to distinguish different images from each other modulo data augmentations. This is in contrast to the CLIP-I metric, where CLIP is trained with text-image pairs and encodes more descriptive information about images - but not necessarily fine details that are not present in the text annotations. We give an example in Figure 11, where the first column contains a reference real image, the second column a different real image, the third column a DreamBooth generated image and the last column an image generated using Textual Inversion. We compare the 2nd, 3rd and 4th image to the real reference image using the CLIP-I and DINO metrics. We observe that the



```
prompt_list = [
'a {0} {1} in the jungle'.format(unique_token, class_token),
'a {0} {1} in the snow'.format(unique_token, class_token),
'a {0} {1} on the beach'.format(unique_token, class_token),
'a {0} {1} on a cobblestone street'.format(unique_token, class_token),
'a {0} {1} on top of pink fabric'.format(unique_token, class_token),
'a {0} {1} on top of a wooden floor'.format(unique_token, class_token),
'a {0} {1} with a city in the background'.format(unique_token, class_token),
'a {0} {1} with a mountain in the background'.format(unique_token, class_token),
'a {0} {1} with a blue house in the background'.format(unique_token, class_token),
'a {0} {1} on top of a purple rug in a forest'.format(unique_token, class_token),
'a {0} {1} with a wheat field in the background'.format(unique_token, class_token),
'a {0} {1} with a tree and autumn leaves in the background'.format(unique_token, class_token),
'a {0} {1} with the Eiffel Tower in the background'.format(unique_token, class_token),
'a {0} {1} floating on top of water'.format(unique_token, class_token),
'a {0} {1} floating in an ocean of milk'.format(unique_token, class_token),
'a {0} {1} on top of green grass with sunflowers around it'.format(unique_token, class_token),
'a {0} {1} on top of a mirror'.format(unique_token, class_token),
'a {0} {1} on top of the sidewalk in a crowded street'.format(unique_token, class_token),
'a {0} {1} on top of a dirt road'.format(unique_token, class_token),
'a {0} {1} on top of a white rug'.format(unique_token, class_token),
'a red {0} {1}'.format(unique_token, class_token),
'a purple {0} {1}'.format(unique_token, class_token),
'a shiny {0} {1}'.format(unique_token, class_token),
'a wet {0} {1}'.format(unique_token, class_token),
'a cube shaped {0} {1}'.format(unique_token, class_token)
]
```

Object Prompts

```
prompt_list = [
'a {0} {1} in the jungle'.format(unique_token, class_token),
'a {0} {1} in the snow'.format(unique_token, class_token),
'a {0} {1} on the beach'.format(unique_token, class_token),
'a {0} {1} on a cobblestone street'.format(unique_token, class_token),
'a {0} {1} on top of pink fabric'.format(unique_token, class_token),
'a {0} {1} on top of a wooden floor'.format(unique_token, class_token),
'a {0} {1} with a city in the background'.format(unique_token, class_token),
'a {0} {1} with a mountain in the background'.format(unique_token, class_token),
'a {0} {1} with a blue house in the background'.format(unique_token, class_token),
'a {0} {1} on top of a purple rug in a forest'.format(unique_token, class_token),
'a {0} {1} wearing a red hat'.format(unique_token, class_token),
'a {0} {1} wearing a santa hat'.format(unique_token, class_token),
'a {0} {1} wearing a rainbow scarf'.format(unique_token, class_token),
'a {0} {1} wearing a black top hat and a monocle'.format(unique_token, class_token),
'a {0} {1} in a chef outfit'.format(unique_token, class_token),
'a {0} {1} in a firefighter outfit'.format(unique_token, class_token),
'a {0} {1} in a police outfit'.format(unique_token, class_token),
'a {0} {1} wearing pink glasses'.format(unique_token, class_token),
'a {0} {1} wearing a yellow shirt'.format(unique_token, class_token),
'a {0} {1} in a purple wizard outfit'.format(unique_token, class_token),
'a red {0} {1}'.format(unique_token, class_token),
'a purple {0} {1}'.format(unique_token, class_token),
'a shiny {0} {1}'.format(unique_token, class_token),
'a wet {0} {1}'.format(unique_token, class_token),
'a cube shaped {0} {1}'.format(unique_token, class_token)
]
```

Live Subject Prompts

Figure 10. **Prompts**. Evaluation prompts for both objects and live subjects.

2nd real image obtains both the highest CLIP-I and DINO scores. The DreamBooth sample looks much more similar to the reference sample than the Textual Inversion sample, yet the CLIP-I score for the Textual Inversion sample is much higher than the DreamBooth sample. However, we can see that the DINO similarity is higher for the DreamBooth sample - which more closely tracks human evaluation of subject fidelity. In order to quantitatively test this, we compute correlations between DINO/CLIP-I scores and normalized human preference scores. DINO has a Pearson correlation coeff. of 0.32 with human preference (vs. 0.27 for the CLIP-I metric used in [20]), with a very low p-value of 9.44×10^{-30}.

User Study

Below we include the full instructions used for our user study. For *subject fidelity*:

- Read the task carefully, inspect the reference items and then inspect the generated items.
- Select which of the two generated items (A or B) reproduces the identity (e.g. item type and details) of the reference item.
- The subject might be wearing accessories (e.g. hats, outfits). These should not affect your answer. Do not take them into account.
- If you're not sure, select Cannot Determine / Both Equally.

For *text fidelity*:

- Read the task carefully, inspect the reference text and then inspect the generated items.
- Select which of the two generated items (A or B) is best described by the reference text.
- If you're not sure, select Cannot Determine / Both Equally.

For each study we asked 72 users to answer questionnaires of 25 comparative questions (3 users per questionnaire), totaling 1800 answers - with 600 image pairs evaluated.

Additional Applications and Examples

Additional Samples We provide a large amount of additional random samples in an annex HTML file. We compare real images, to DreamBooth generated images using Imagen and Stable Diffusion as well as images generated using Textual Inversion on Stable Diffusion.

Recontextualization We show additional high-quality examples of recontextualization in Figure 12.

Art Renditions We show additional examples of original artistic renditions of a personalized model in Figure 13.

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 14. 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.

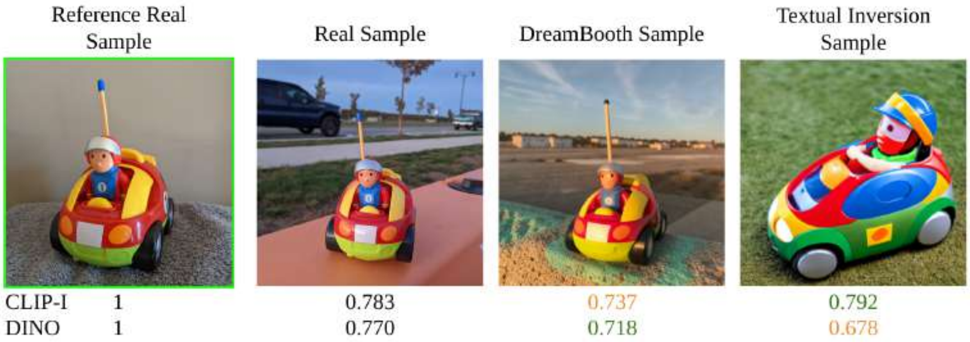


Figure 11. **CLIP-I vs. DINO Metrics.** The DreamBooth CLIP-I similarity to the reference image is lower than that of the Textual Inversion sample, even though the DreamBooth subject looks more similar to the reference subject. The DINO metric more closely tracks human evaluation of subject fidelity here.

Novel View Synthesis We show more viewpoints for novel view synthesis in Figure 15, along with prompts used to generate the samples.

Accessorization An interesting capability stemming from the strong compositional prior of the generation model is the ability to accessorize subjects. In Figure 16 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.

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 17. 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 17. 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.

Comic Book Generation In addition to photorealistic images, our method is able to capture the appearance of drawn media and more. In Figure 18 we present, to the best of our knowledge, the first instance of a full comic with a persistent character generated by a generative model. Each comic frame was generated using a descriptive prompt (e.g “a [V] cartoon grabbing a fork and a knife saying “time to eat””).

Additional Experiments

Prior Preservation Loss

Here we show qualitative examples of how our prior preservation loss (PPL) conserves variability in the prior and show sample results in Figure 19. 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.

Effect of Training Images

Here we run an experiment on the effects of the number of input images for model personalization. Specifically, we train models for two subjects, 5 models per subject with input images ranging from 1 to 5. We generate 4 images for 10 different recontextualization prompts for each subject. We present qualitative results in Figure 20. We can observe that for some subjects that are more common, and lie more strongly in the distribution of the diffusion model, such as the selected Corgi dog, we are able to accurately

Method	1	2	3	4	5
Backpack	0.494	0.515	0.596	**0.604**	0.597
Dog	0.798	0.851	0.871	**0.876**	0.864

Table 5. Effect of the number of input images on subject fidelity (DINO).

Method	1	2	3	4	5
Backpack	0.798	0.851	0.871	**0.876**	0.864
Dog	0.646	0.683	0.734	**0.740**	0.730

Table 6. Effect of the number of input images on prompt fidelity (CLIP-T).

capture the appearance using only two images - and sometimes only one, given careful hyperparameter choice. For objects that are more rare, such as the selected backpack, we need more samples to accurately preserve the subject and to recontextualize it to diverse settings. Our quantitative results support these conclusions - we present the DINO subject fidelity metric in Table 5 and the CLIP-T prompt fidelity metric in Table 6. For both subjects we see that the optimal amount of input images for subject and prompt is 4. This number can vary depending on the subject and we settle on 3-5 images for model personalization.

Personalized Instance-Specific Super-Resolution and Low-level Noise Augmentation for Imagen

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 21 (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.

We find results to be suboptimal if the training recipes and test parameters of Saharia et al. [61] 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 21 (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. We show how using lower noise to train the super-resolution models improves fidelity. Specifically, we show in Figure 21 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.

Comparisons

We include additional qualitative comparisons with Gal et al. [20] in Figure 22. 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.

Next, we show comparisons of recontextualization of a subject clock, with distinctive features using our method and prompt engineering using vanilla Imagen [61] and the public API of DALL-E 2 [54]. 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 the appearance of our subject instance. We show the results in Figure 23, 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.

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.

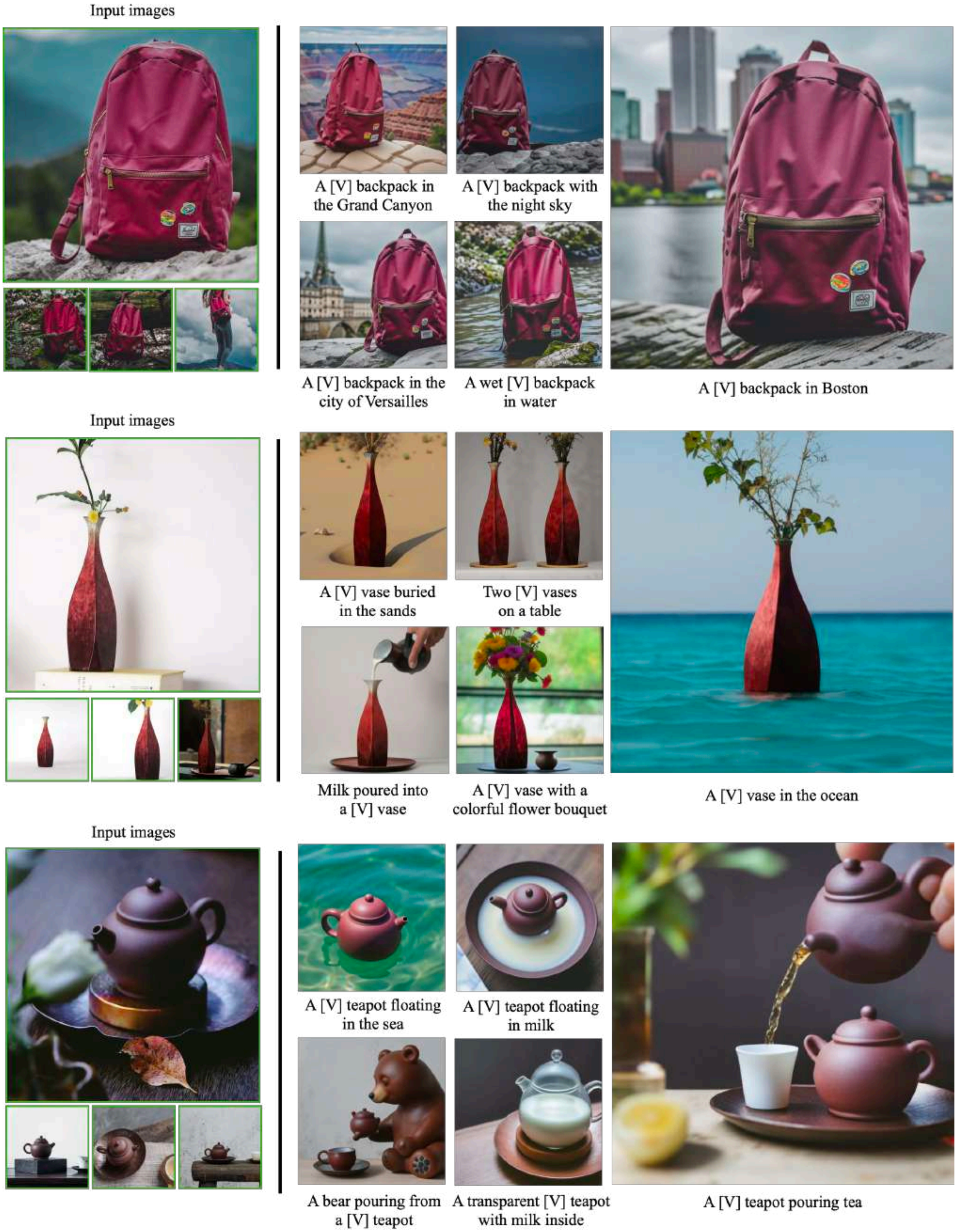

Input images
A [V] backpack in the Grand Canyon
A [V] backpack with the night sky
A [V] backpack in the city of Versailles
A wet [V] backpack in water
A [V] backpack in Boston
Input images
A [V] vase buried in the sands
Two [V] vases on a table
Milk poured into a [V] vase
A [V] vase with a colorful flower bouquet
A [V] vase in the ocean
Input images
A [V] teapot floating in the sea
A [V] teapot floating in milk
A bear pouring from a [V] teapot
A transparent [V] teapot with milk inside
A [V] teapot pouring tea

Figure 12. **Additional recontextualization samples of a backpack, vase, and teapot subject instances.** 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.

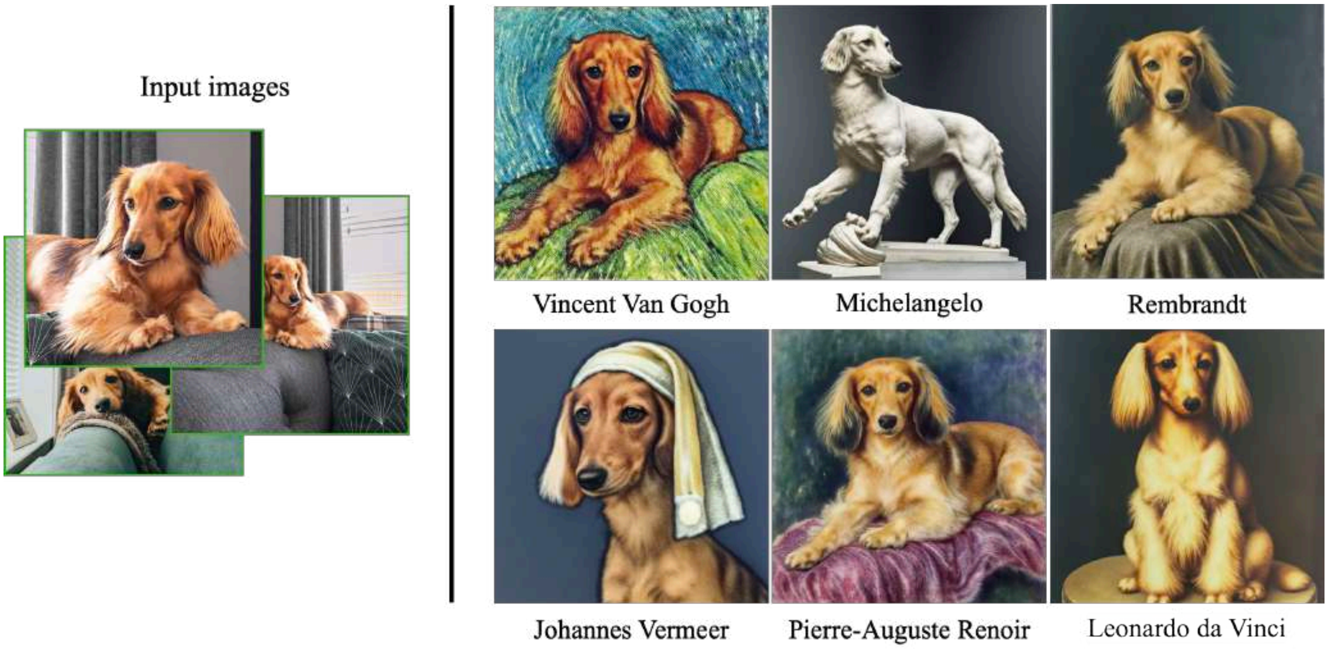


Figure 13. **Additional artistic renderings of a dog instance in the style of famous painters**. We remark that many of the generated poses, e.g., the Michelangelo renditions, were not seen in the training set. We also note that some renditions seem to have novel compositions and faithfully imitate the style of the painter.

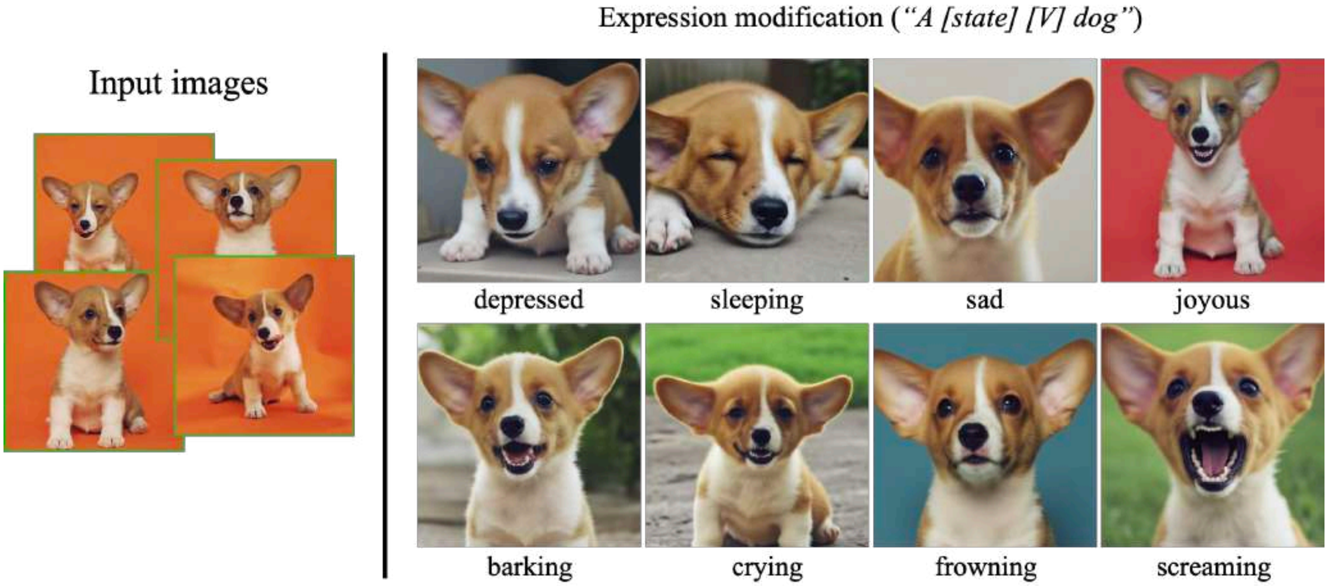


Figure 14. **Expression manipulation of a dog instance.** Our technique can synthesize various expressions that do not appear in the input images, demonstrating the extrapolation power of the model. Note the unique asymmetric white streak on the subject dog's face.

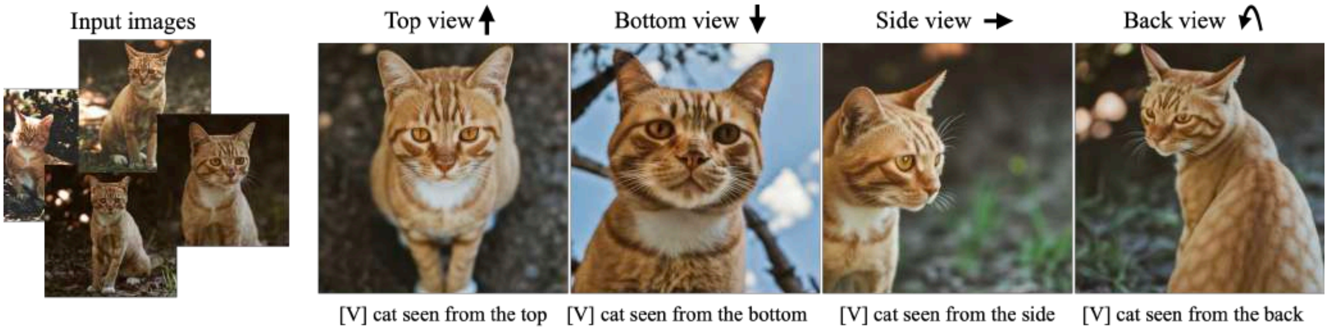

Input images
Top view
Bottom view
Side view
Back view
[V] cat seen from the top
[V] cat seen from the bottom
[V] cat seen from the side
[V] cat seen from the back

Figure 15. **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.

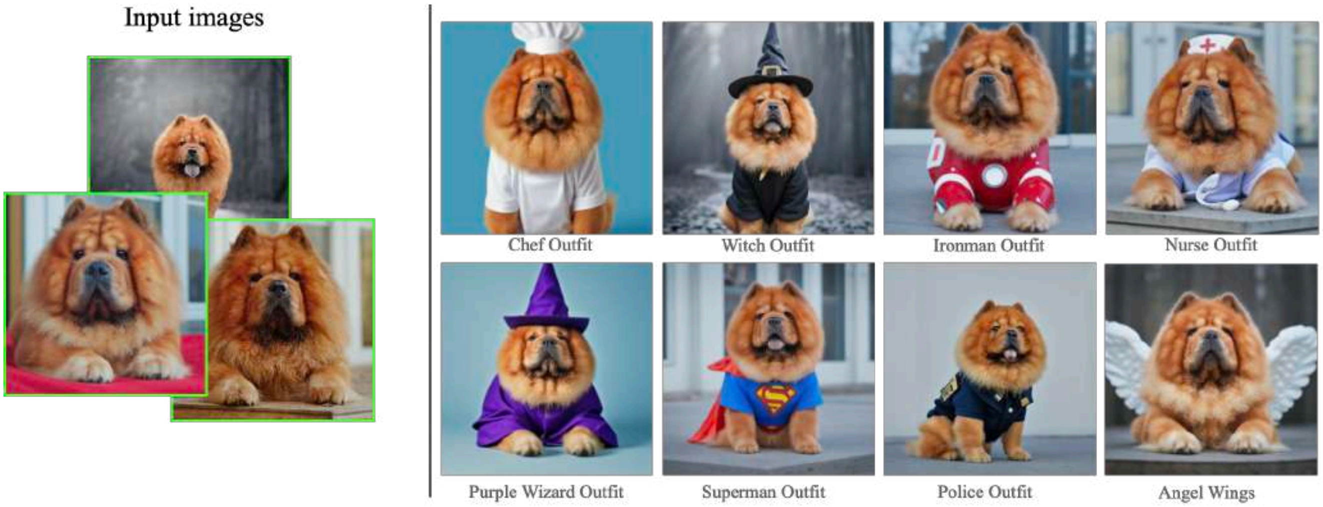


Figure 16. **Outfitting a dog with accessories**. The identity of the subject is preserved and many different outfits or accessories can be applied to the dog given a prompt of type "a [V] dog wearing a police/chef/witch outfit". We observe a realistic interaction between the subject dog and the outfits or accessories, as well as a large variety of possible options.

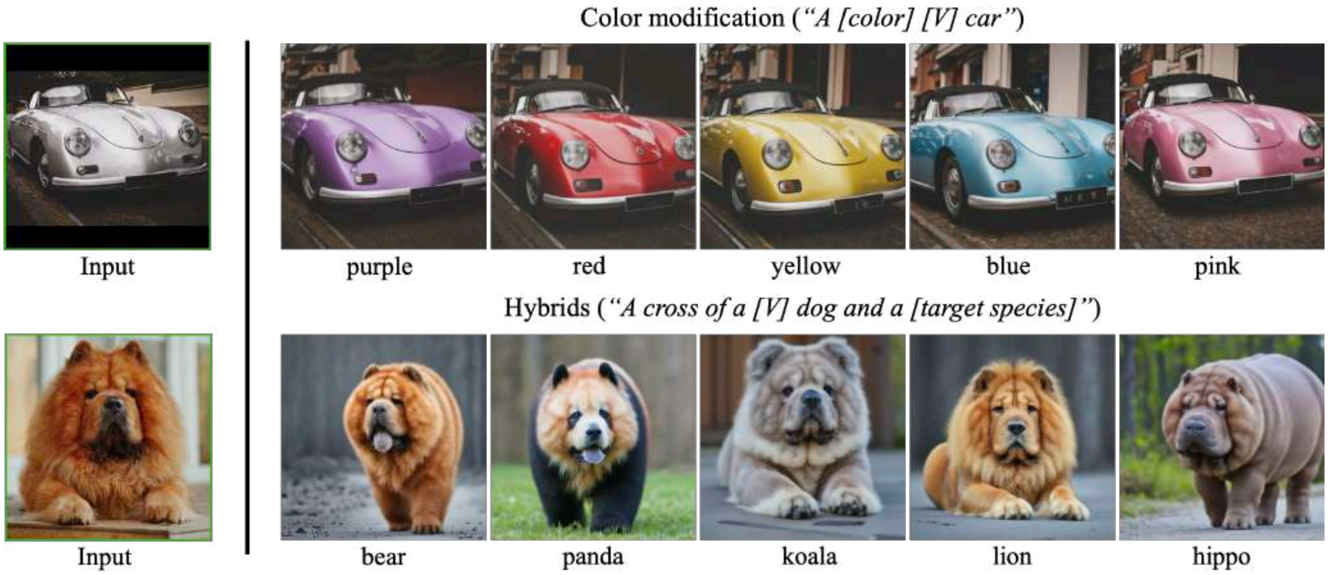

Color modification (*"A [color] [V] car"*)
Input
purple
red
yellow
blue
pink
Hybrids (*"A cross of a [V] dog and a [target species]"*)
Input
bear
panda
koala
lion
hippo

Figure 17. **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.

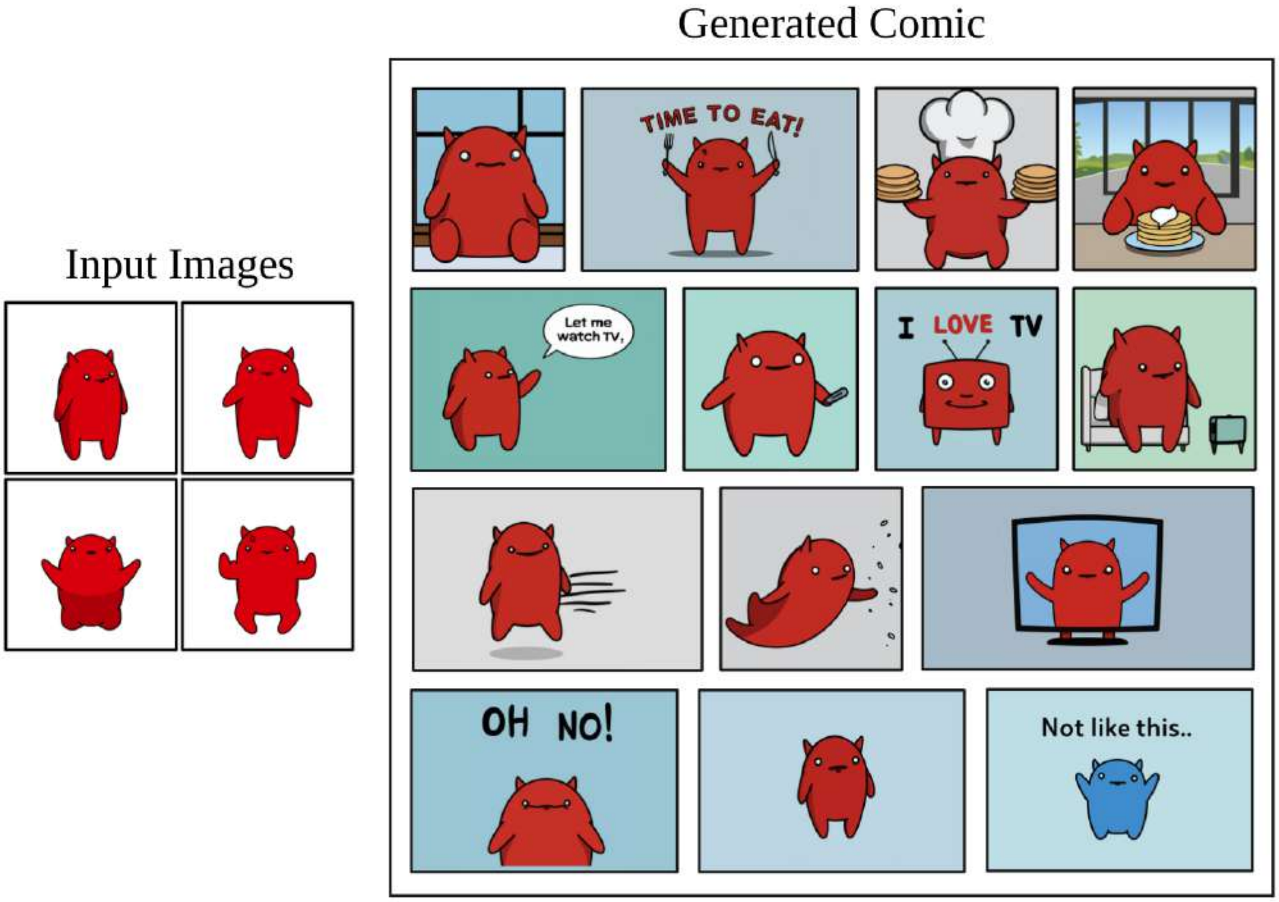


Figure 18. **Generated comic.** We present, to the best of our knowledge, the first comic comic with a persistent character generated by prompting a generative model.

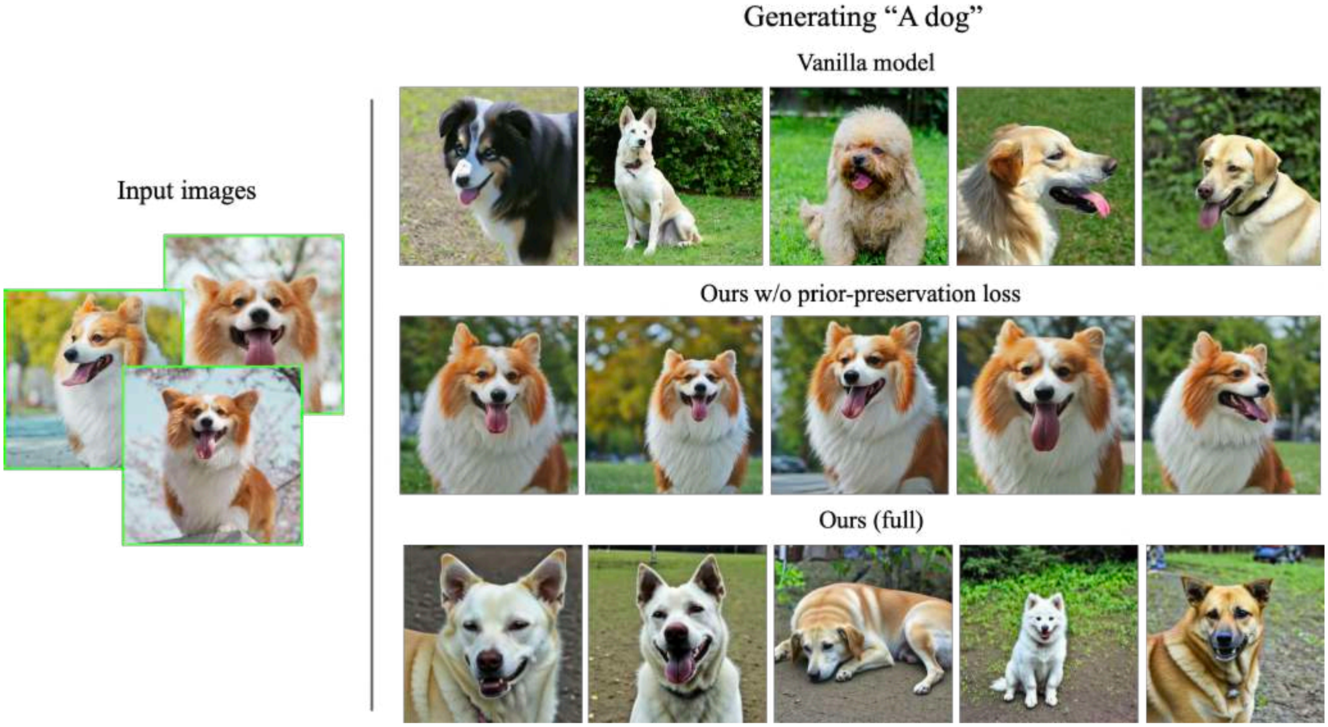


Figure 19. **Preservation of class semantic priors with prior-preservation loss.** Fine-tuning using images of our subject without prior-preservation loss results in language drift and the model loses the capability of generating other members of our subject's class. Using a prior-preservation loss term allows our model to avoid this and to preserve the subject class' prior.

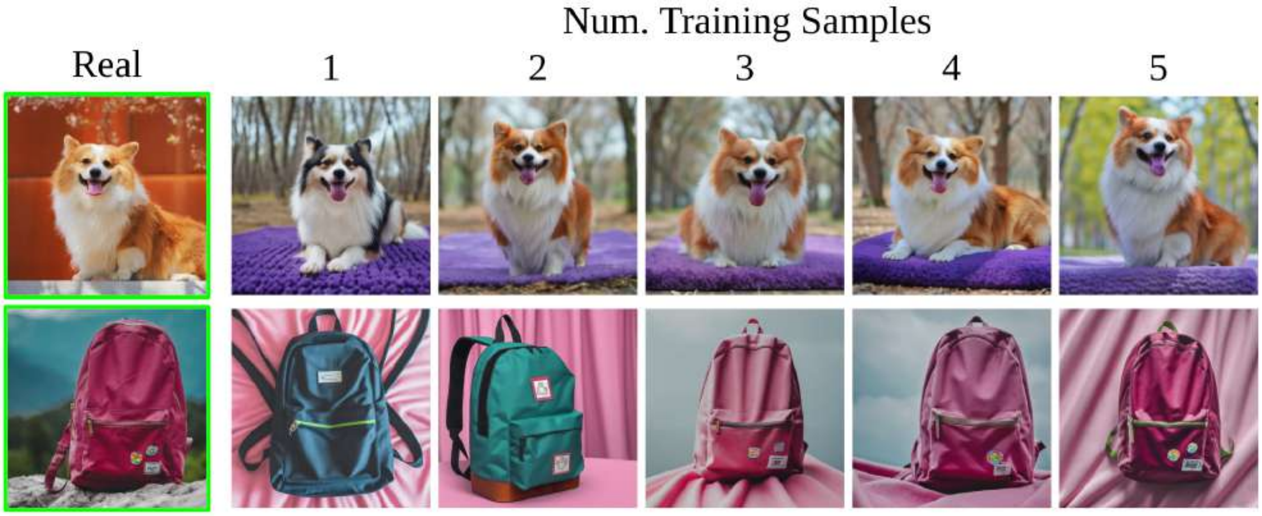


Figure 20. **Impact of number of input images.** We observe that given only one input image, we are close to capture the identity of some subjects (e.g. Corgi dog). More images are usually needed - two images are sufficient to reconstruct the Corgi dog in this example whereas at least 3 are needed for a more rare item such as the backpack.

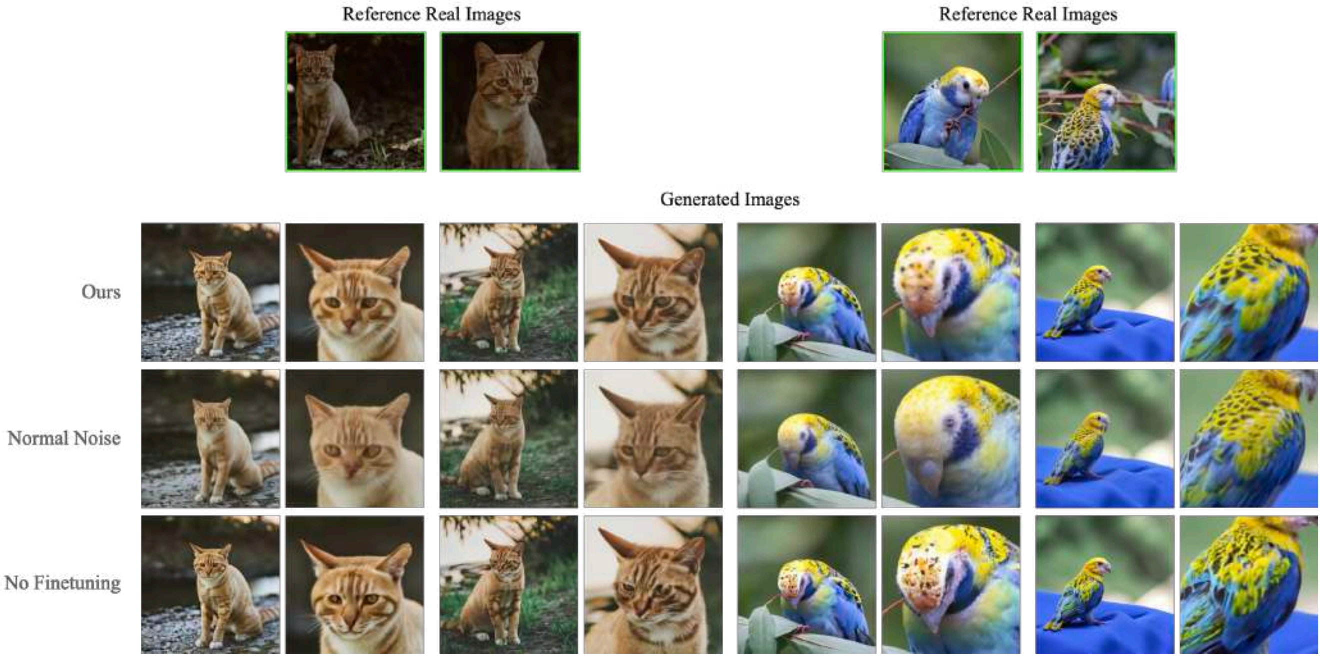


Figure 21. **Ablations with fine-tuning the super-resolution (SR) models.** Using the normal level of noise augmentation of [61] to train the SR models results in blurred high-frequency patterns, while no fine-tuning results in hallucinated high-frequency patterns. Using low-level noise augmentation for SR models improves sample quality and subject fidelity. Image credit (input images): Unsplash.

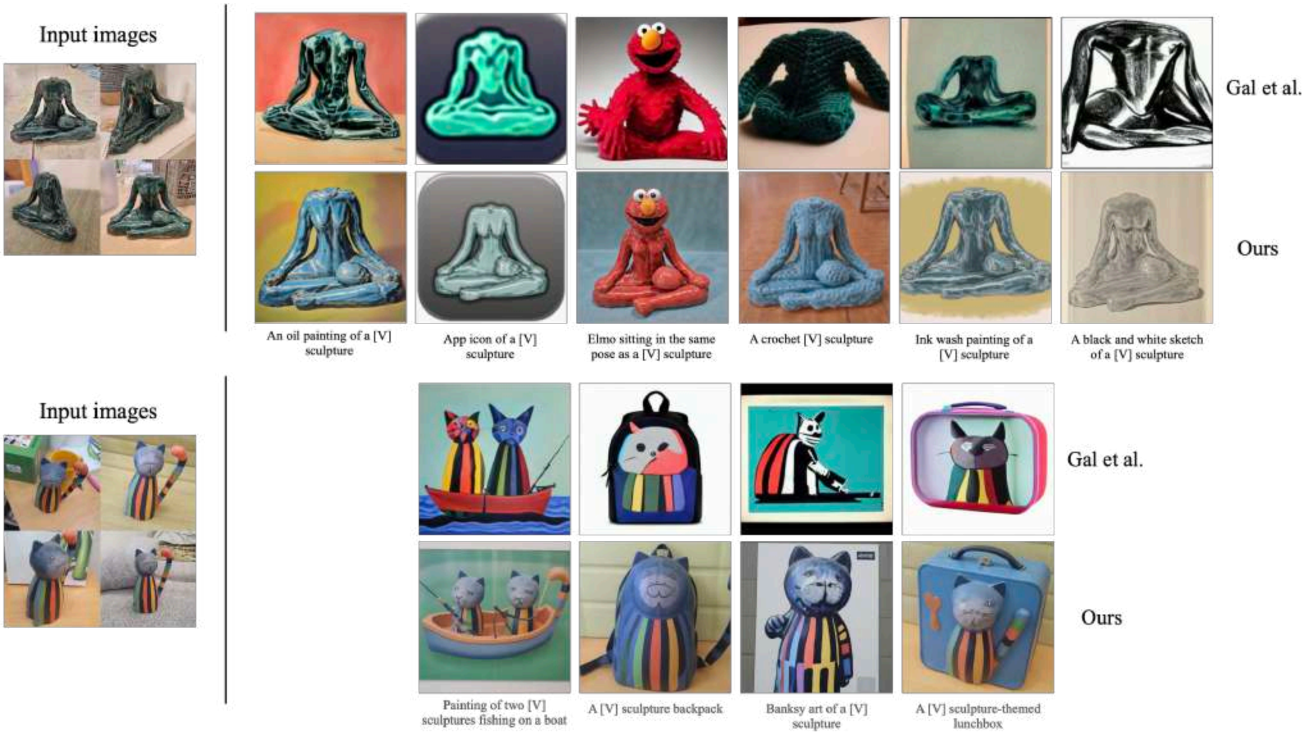

Input images
Gal et al.
Ours
An oil painting of a [V] sculpture
App icon of a [V] sculpture
Elmo sitting in the same pose as a [V] sculpture
A crochet [V] sculpture
Ink wash painting of a [V] sculpture
A black and white sketch of a [V] sculpture
Input images
Gal et al.
Ours
Painting of two [V] sculptures fishing on a boat
A [V] sculpture backpack
Banksy art of a [V] sculpture
A [V] sculpture-themed lunchbox

Figure 22. **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].

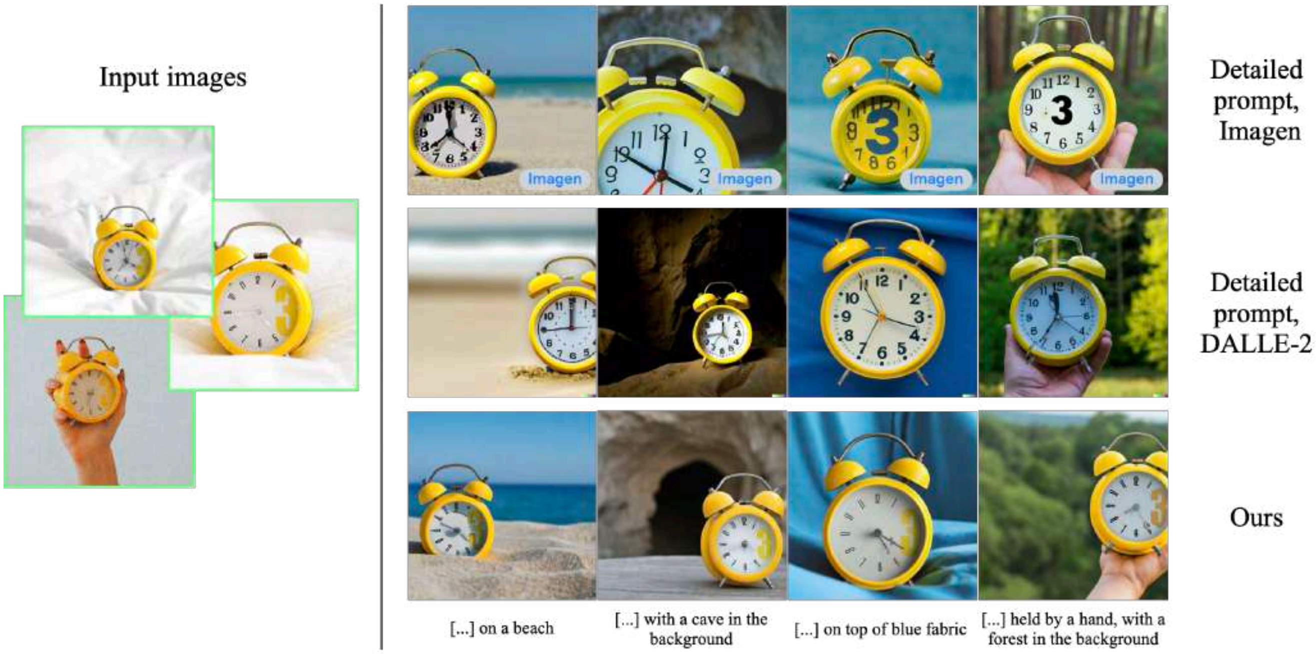

Input images
Detailed prompt, Imagen
Detailed prompt, DALLE-2
Ours
[...] on a beach
[...] with a cave in the background
[...] on top of blue fabric
[...] held by a hand, with a forest in the background

Figure 23. **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.